Sparse power-law network model for reliable statistical predictions based on sampled data

A. P. Kartun-Giles,[1] D. Krioukov,[2] J. P. Gleeson,[3] Y. Moreno,[4, 5, 6, 7] and G. Bianconi[1]

[1]*School of Mathematical Sciences, Queen Mary*
University of London, London, United Kingdom
[2]*Departments of Physics, Mathematics,*
and Electrical & Computer Engineering,
Northeastern University, Boston, United States
[3]*MACSI, Department of Mathematics and Statistics,*
University of Limerick, Limerick, Ireland
[4]*Institute for Biocomputation and Physics of Complex Systems (BIFI),*
University of Zaragoza, Zaragoza, Spain
[5]*Department of Theoretical Physics, Faculty of Sciences,*
University of Zaragoza, Zaragoza, Spain
[6]*ISI Foundation, Turin, Italy*
[7]*Complexity Science Hub Vienna, Vienna, Austria*

Abstract

A projective network model is a model that enables predictions to be made based on a subsample of the network data, with the predictions remaining unchanged if a larger sample is taken into consideration. An exchangeable model is a model that does not depend on the order in which nodes are sampled. Despite a large variety of non-equilibrium (growing) and equilibrium (static) sparse complex network models that are widely used in network science, how to reconcile sparseness (constant average degree) with the desired statistical properties of projectivity and exchangeability is currently an outstanding scientific problem. Here we propose a network process with hidden variables which is projective and can generate sparse power-law networks. Despite the model not being exchangeable, it can be closely related to exchangeable uncorrelated networks as indicated by its information theory characterization and its network entropy. The use of the proposed network process as a null model is here tested on real data, indicating that the model offers a promising avenue for statistical network modelling.

I. INTRODUCTION

Network science [1–4] is one of the most rapidly advancing scientific fields of investigation. The success of this field is deeply rooted in its interdisciplinarity. In fact, network science characterizes the underlying structure and dynamics of complex systems ranging from on-line social networks to molecular networks and the brain. Additionally, the theoretical tools and techniques used by network science are coming from different disciplines including statistical mechanics, statistics, machine learning and computer science.

In the last twenty years significant attention has been addressed to modelling framework of complex networks. Since most real networks from the Internet to molecular networks, are sparse, i.e. they have an average degree that does not depend on the network size, statistical mechanics models focus on modelling sparse networks. These statistical mechanics models can be divided between non-equilibrium growing network models [5–13] such as the famous Barabási-Albert model [5] and equilibrium models such as maximum entropy network ensembles [14–19] including Exponential Random Networks [16, 17, 20–22] and block models [23, 24]. The non-equilibrium growing network models have the power to explain the fundamental mechanisms giving rise to emergent properties such as scale-free distributions [5–8], degree correlations [6], communities [9–11] and network geometry [11–13]. On the contrary, maximum network ensembles constitute the least biased models satisfying a given set of constraints. These models are not explanatory but constitute the ideal null hypothesis to which real networks can be compared.

Recently the need to formulate reliable statistical models is receiving significant attention [25]. A reliable statistical model will include projectivity and exchangeability [26–30]. The projectivity of the statistical network model guarantees that the conclusions reached by considering a subsample of the data are consistent with the ones that can be drawn starting from a larger sample of the data. The exchangeability of the nodes implies that the probability of a network does not depend on the specific labels of the nodes. However, how to reconcile these statistical requirements with the sparseness of the networks, i.e., a average degree that is independent of the network size, constitutes a major impasse of network modelling. For instance it has been shown that random uncorrelated networks are only projective if the average degree $\langle k \rangle$ of the network increases linearly with the network size N, i.e., if the network is maximally dense and $\langle k \rangle = O(N)$ [27, 30, 31].

In physical terms the desired projective and exchangeable network process mimicking the subsequent sampling of an increasing portion of the network is a modelling framework that goes beyond the traditional statistical mechanics division between equilibrium and non-equilibrium network modelling approaches. This observation reinforces the belief that actually combining these two properties might be not an easy task.

Already several works have addressed this problem [32–38], using different approaches such as relaxing the condition $\langle k \rangle = O(N)$ but always characterizing models with average degree diverging with the network size N, considering edge exchangeable models or alternatively using an embedding space as a basic mechanism to combine sparsity with projectivity and exchangeability [31, 39].

Here we propose a network process describing a network evolution mimicking the sampling of a network by subsequently expanding the nodes set. Each node is assigned an hidden variable from a hidden variable distribution. This distribution is the key quantity determining the properties of the network process. If the hidden variable is power-law distributed and the network is sufficiently sparse, the degree distribution displays a power-law tail with the same power-law exponent as the hidden variable distribution.

This model is a projective network process but it is not exchangeable. Nevertheless this non-equilibrium network model can be directly related to an equilibrium uncorrelated network ensemble in the sparse regime. In fact, by permuting the order in which nodes are sampled it is possible to calculate the probability that two nodes are connected given their corresponding hidden variables. This connection probability is equal to the connection probability in an uncorrelated exchangeable network ensemble in which the hidden variable of each node is identified with half of its expected degree. The "proximity" between the network process and the uncorrelated network ensemble is here quantified by using information theory tools and comparing the entropy of the two models. In particular, we use the entropy of the two network models [14–17, 40] to evaluate the difference in the information content of the two models, finding that the two models have small relative entropy difference.

Finally we study how well the proposed model can be used as a null model for real power-law network datasets. To this end we identify the hidden variable of each node with half of its observed degree and we run the model by adding the nodes in the network according to a random permutation of the nodes' labels. The degree distribution of the real dataset and the degree distribution of the simulation results are in good agreement when starting

from power-law networks, and the agreement remains good if the network is grown by only considering a subsample of the nodes of the real data. We also compare the correlations of the real dataset with the correlations of the simulation results to show that the simulations are able to generate only weak correlations of the degrees. Therefore a more refined model should be formulated to capture this additional network property.

The paper is structured as described in the following. In Sec. 2 we introduce the definition of the desired statistical properties of network models: *projectivity* and *exchangeability.* In Sec. 3 we discuss major examples of sparse network models (the Barabási-Albert model and the uncorrelated network ensembles) and characterize them with respect to the properties of projectivity and exchangeability. In Sec. 4 we present an account of the difficulties in combining projectivity and exchangeability with the sparseness of networks and we give a brief review of the approaches investigated in the recent literature on the subject. In Sec. 5 we present a network process mimicking a network sampling process. We characterize its structural and dynamical properties relating this non-equilibrum model to equilibrium uncorrelated network ensembles, and we characterize its statistical properties. In Sec. 6 we show the possible use of the proposed network process as a null model for modelling real power-law network datasets. Finally in Sec. 7 we give the conclusions.

II.   STATISTICAL TERMS

Projectivity and exchangeability are two very basic and very natural statistical requirements for reliable statistical network models. In physical terms, projectivity is directly related to the principle of locality, while exchangeability is related to symmetry. In this section, we first discuss projectivity and exchangebility to make clear that they really are "must-have"s in any statistically useful network model, while in the next two sections we will comment on difficulties in combining them both in models of sparse networks, i.e., having average degree independent of the network size N [41].   While projectivity and exchangeability are desired properties of statistically reliable network models, the relevance and of these requirements for any realistic network model is a subject of scientific debate (see for instance contribution of Karthik Bharath in the discussion of the F. Caron and E. Fox paper [33]).In fact it is often observed that most real networks can hardly be exchangeable. Indeed, in a vast majority of real networks nodes are labelled with labels related to some

4

rich metadata and a random permutation of the nodes labels would result in a different network whose probability to be produced by the same stochastic process that produces the real network is certainly not expected to be equal to the probability with which it generates the real network.

In order to investigate the properties of reliable statistical models we consider a network process mimicking the subsequent sampling a network by expanding the set of sampled nodes and detecting all the interactions among this set of nodes.

To this end we consider a set of networks $\{G_t\}_{t=1,2,\dots}$ with $G_t = (V_t, E_t)$ and increasing network size $N_t = |V_t| = t$. The sequence of networks defines a network process, i.e. $G_t = (V_t, E_t)$ is an induced subgraph of the network $G_{t'} = (V_{t'}, E_{t'})$ for all $t < t'$ with node set $V_t \subset V_{t'}$ if $t < t'$. We label the nodes in order of their appearance in the network such that

$$V_t = \{1, 2, \dots, t\}. \tag{1}$$

and assign a probability $P(G_t)$ to each network G_t.

A. Projectivity

Given the set of networks $\{G_t\}_{t=1,2,\dots}$ *projectivity* implies that the statistical properties of the network G_t are directly related to the statistical properties of the network $G_{t'}$ with $t' > t$ by a proper marginalization of the probability of the network $G_{t'}$ over its subgraph G_t.

By definition [26, 27], a projective network model is a model that attributes a given probability $P(G_t)$ to each network G_t of the sequence, such that

$$P(\pi_{t',t}(G_{t'})) = P(G_t),$$

where the projective map $\pi_{t',t}$ maps networks $G_{t'}$ of a larger size $t' > t$ to their subgraph G_t of a smaller size t.

In other words this means that one can first generate a larger graph $G_{t'}$ using the model, then reduce its size to t by throwing out some $t' - t$ nodes according to the projective map specification, and the probability with which the resulting graph G_t is generated using this two-step procedure will be the same as if graph G_t was generated by the model directly.

B. Exchangeability

Exchangeability implies that the order in which two nodes are observed or labelled is not important. Specifically, a network model is exchangeable if, by definition [29, 30], the probability $P(G_t)$ of a network $G_t = (V_t, E_t)$ is independent on the nodes labels, i.e.

$$P(G_t) = P(\tilde{G}_t) \tag{2}$$

where $\tilde{G}_t$ is any network isomorphic to the network G_t, i.e. it is any network obtained from the network G_t by permuting the nodes labels $\{i\}_{i=1,2,\dots,N}$ according to the permutation $\boldsymbol{\sigma}$. If a network model is exchangeable it follows that the marginal the probability p_{ij} of the generic link between node i and node j is unchanged if the node labels are permuted, implying that they are sampled in a different order, i.e.,

$$p_{ij} = p_{\sigma(i),\sigma(j)}. \tag{3}$$

Therefore exchangeability enforces the *symmetry* of the model with respect to the group of graph isomorphisms.

III.   CHARACTERIZATION OF RELEVANT SPARSE NETWORK MODELS FROM THE STATISTICAL PERSPECTIVE

In this section we investigate major examples of non-equilibrium (growing) network models and equilibrium (static) network models widely used to model sparse complex networks. In particular we discuss the Barabási-Albert model [5] and the uncorrelated network ensembles from the statistical perspective. This discussion will reveal that neither of these two very popular frameworks for modelling sparse complex networks display both projectivity and exchangeability, indicating the difficulties in combining these properties with the sparseness of the networks.

A.   Barabási Albert model

The Barabási-Albert model begins with an initial finite network and at each time t a new node enters in the network and is connected to the network by establishing m new links.

Each of these links connect the new node to a node i with degree k_i chosen with probability

$$\tilde{\Pi}_i = \frac{k_i}{\sum_{i'} k_{i'}}. \tag{4}$$

This probability enforces *preferential attachment*, i.e., allows nodes with higher degree to more rapidly acquire new links.

The Barabási-Albert model describes a model that is projective, because as the network grows the network G_t obtained at time t is an induced subgraph of the network $G_{t'}$ obtained at a later time $t' > t$. However the Barabási-Albert model is not exchangeable. The fact that the network is not exchangeable is revealed for instance by the expression for the average number of links $k_i(t, t_i)$ of a node i arrived in the network at time t_i,

$$k_i(t, t_i) = m \left(\frac{t}{t_i} \right)^{1/2}. \tag{5}$$

This expression explicitly indicates that the older nodes are statistically different from the younger nodes, and their degree is much larger than that of younger nodes. Additionally it is possible to observe that the model is not exchangeable because the order of the addition of the nodes, i.e., their time of arrival in the network, is the key property that determines the connection probability [42], i.e.,

$$p_{ij} \simeq \frac{m}{2} \frac{1}{\sqrt{t_i t_j}}. \tag{6}$$

Nevertheless we observe the interesting property that for this model the connection probability p_{ij} between node i and node j can be also expressed as

$$p_{ij} \simeq \frac{k_i(t, t_i) k_j(t, t_j)}{\sum_{i'} k_{i'}(t, t_{i'})}, \tag{7}$$

indicating that actually, although the network process has different statistical properties than the uncorrelated network with the same degree distribution, the expected degree correlations are weak. The relation between the BA model and the uncorrelated network ensemble with the same degree distribution is investigated in detail using information theoretic tools in Ref. [43].

B.   Uncorrelated Network Ensembles

The Barabási-Albert model is projective but not exchangeable. On the contrary the widely used uncorrelated network ensembles are exchangeable models but they are not projective in the sparse regime. In order to show this let us consider an uncorrelated network

model in which each node i has an expected degree θ_i, where the expected degrees of the nodes are consistent with a structural cutoff, i.e.

$$\theta_i \leq \sqrt{\langle \theta \rangle N}. \tag{8}$$

In this case the probability p_{ij} of a link between node i and node j is given by

$$p_{ij} = \frac{\theta_i \theta_j}{\langle \theta \rangle N}, \tag{9}$$

and therefore it only depends on the expected degrees θ_i and θ_j of the nodes i and j and not on the order in which node i and node j have been sampled. The model is therefore exchangeable as long as we consider the simultaneous permutation of the node labels and the expected degrees of the nodes. However if we consider a large sample of the network with $N' > N$ nodes, we see that the model is projective if and only if it is also dense, with the number of links scaling as $L = O(N^2)$. In fact if we assume that in the larger sample the expected degrees of nodes i and j are given by θ_i' and θ_j', the probability that node i and node j are connected in the larger network models including N' nodes is

$$p_{ij}' = \frac{\theta_i' \theta_j'}{\langle \theta' \rangle N'} \tag{10}$$

If we impose projectivity, i.e.

$$p_{ij} = p_{ij}' \tag{11}$$

for $i, j \leq N$, and we assume that the number of nodes $N' > N$ can be written as

$$N' = zN, \tag{12}$$

it is easy to see that we should also have

$$\begin{aligned} \theta_i' &= z\theta_i, \\ \langle \theta_i' \rangle &= z\langle \theta \rangle. \end{aligned} \tag{13}$$

Therefore to guarantee projectivity the expected degree of each node should grow linearly with the network size, resulting in a dense network with the total number of links L scaling with the network size N as $L = O(N^2)$. This implies that the random network $\mathbb{G}(N, p)$ with p independent of N is an exchangeable model whereas the Poisson random network

$\mathbb{G}(N, p)$ with $p = \frac{z}{N}$ and z independent of N is not exchangeable. In fact one cannot throw out $N' - N$ nodes from a network of size N' produced by $\mathbb{G}(N', z/N')$, and hope that the resulting network will have the same probability as in $\mathbb{G}(N, z/N)$, simply because the links in the $\mathbb{G}(N', z/N')$ and $\mathbb{G}(N, z/N)$ ensembles exist with different probabilities z/N' and z/N that depend on the graph size N. Alternatively, if one attempts to formulate $\mathbb{G}(N, z/N)$ as a growing model, then since the edge existence probability depends on N, the addition of a new node affects the probability of existence of edges in the existing network. Since this probability is a decreasing function of N (z/N), upon the addition of a new node all the existing edges must be removed with some probability $(1/N)$. In other words, in such a growing model new node additions must necessarily affect the existing network structure.

IV.   IMPASSE WITH SPARSITY

Surprisingly, combining projectivity and exchangeability with the additional constraint of sparsity, i.e. the requirement that the average degree of the sampled networks is independent of the network size, has been a major impasse. If we exclude spatially embedded networks [31], to the best of our knowledge there exists no model of sparse networks that would be both projective and exchangeable at the same time. This situation is in stark contrast with the case of dense graphs. Dense graphs are known to have well-defined thermodynamic limits known as graphons, and any graphon-based network model is both exchangeable and projective [30].

The thermodynamic limits of sparse graphs are at present quite poorly understood, which appears to be one of the reasons behind the mentioned impasse. Several attempts have been made to understand the limits of sparse graphs, including, for example, sparse L^p graphons [32], which are not projective, or stretched graphons a.k.a. graphexes [33–35]. In the latter case, graphs are sparse, exchangeable and projective, but with two major caveats:

1) the average degree cannot be constant, it must diverge with N (but possibly slower than linearly),

2) exchangeability is completely redefined: it is not with respect to node labels $1, \ldots, N$, but with respect to artificial labels which are positive real numbers.

Another class of attempts suggests to completely give up on the node label exchangeability

requirement, and to consider edge exchangeability instead, e.g., using variations of Pitman-Yor processes [36–38]. It remains unclear at present whether these developments imply that too many network models that were found to be quite useful in practice and that do use node labels $1, \ldots, N$, are statistically hopeless. It seems more likely that further research is needed to understand and resolve this projectivity vs. exchangeability impasse in sparse network models.

A. Proposed solution of the impasse based on network geometry

In [31] it was shown that a generic network model is projective if the probability of edge existence, i.e. the connection probability, does not depend on the network size N. In fact if the connection probability does depend on N, then, the addition of new nodes to the existing network in the growing formulation of the model necessarily affects the existing network structure and the network cannot be projective.

In order to formulate network models in which the connection probability does not depend on the network size N, embedding networks in space can turn out to be very useful. In fact spatially embedded networks can combine projectivity with a constant average degree [31] as their spatial embedding ensures projectivity when the connection probability is *local* and nodes connect typically to nodes that are spatially close. For instance if the nodes are uniformly distributed in $\mathbb{R}^2$ and each node connects only to the nodes with a constant radius r_0, by sampling the network by progressively expanding the spatial region of interest we can build a projective model with constant average degree. This is clearly a realistic scenario in most real networks as it unlikely that a local event in a spatial network causes a global change in the network. For instance in the Internet, the appearance of a new customer of a local Internet provider in Bolivia cannot lead to immediate severance of customers by a local Internet provider in Bhutan.

It turns out that models that are not explicitly constructed from spatial embeddings can also be analysed using geometrical arguments, hence shedding light on their statistical properties. In this vein, it was recently shown that the hypersoft configuration model, which defines maximum-entropy random graphs with a given degree distribution, is sparse and *either* exchangeable *or* projective [39]. Both sparsity and exchangeability definitions are traditional in the model, i.e., the average degree is constant and exchangeability is with

respect to labels $1, \ldots, N$, so that the only caveats are in "either-or" and also in that this "either-or" is achieved only for specific degree distributions (power law with exponent $\gamma = 3$ in [39]).

In the *exchangeable* equilibrium formulation of the model, nodes are points sprinkled at random onto an interval A_N of an N-dependent length L_N, where L_N is a growing function of N, according to a non-uniform point density (if this point density is exponential, then the resulting degree distribution is a power law), and then all pairs of points/nodes i and j, $j > i = 1, \ldots, N$, at sprinkled coordinates x_i and x_j are connected by an edge with the entropy-maximizing Fermi-Dirac connection probability

$$p(x_i, x_j) = \frac{1}{e^{x_i + x_j} + 1} \tag{14}$$

that does not depend on the network size N.

In the *projective* growing formulation of the same model, the interval A_N grows with N, its length growing according to L_n, new node $N + 1$ appears in the interval increment $A_{N+1} \backslash A_N$ of length $L_{N+1} - L_N$, and then connects to existing nodes with the same connection probability as in the exchangeable formulation.

The difficulty of combining projectivity and exchangeability is evident in this example: in the exchangeable formulation, node labels i are random and uncorrelated with their coordinates x_i, while in the projective formulation, nodes are labeled in the increasing order of their coordinates: $i < j \leftrightarrow x_i < x_j$. If nodes are labeled this way, then the projective map $\pi_{N', N}$ trivially throws out nodes with labels $N + 1, \ldots, N'$, and the resulting graph satisfies the projectivity requirement since the connection probability does not depend on N, and since the remaining N nodes lie in A_N. If the node labels are random however, as they are in the exchangeable formulation, then it remains unclear if even an asymptotically correct projective map can be constructed.

V. STATISTICAL MECHANICS MODEL WITH HIDDEN VARIABLES

Our goal is here to reconcile sparseness with a reliable statistical modelling framework without assuming the existence of an embedding geometrical space. In this endeavour we will define a projective network process yielding a sequence of networks growing by the subsequent addition of nodes and links. To each node i we associate a hidden variable

θ_i that is a proxy for the degree that the node will acquire in the model. The statistical properties of the network model when we average over all the possible sequences determining the subsequent addition of the links obey scaling laws and reduce to the uncorrelated network model of any size N in the sparse regime.

Although this model does not ultimately reconcile sparseness with both exchageability and projectivity, we will see in Sec. VI that it provides a very reliable null model for power-law networks also if only a subsample of the original network is considered.

A. The model

The model can be interpreted as a weighted growing network model where we allow multiedges. In the model every node i is assigned a hidden variable θ_i from a hidden variable distribution $\rho(\theta)$.

Starting at $t = 1$ from a single isolated node, at each time $t > 1$ a new node i is added to the network and draws κ_i links to the existing nodes of the network, where κ_i is chosen according to the Poisson distribution with average θ_i, i.e.

$$\hat{P}(\kappa_i|\theta_i) = \frac{1}{\kappa_i!}\theta_i^{\kappa_i}e^{-\theta_i}. \tag{15}$$

Each new link is attached to a node j already present in the network with probability

$$\Pi_j = \frac{\theta_j}{\sum_{r=1}^{t-1}\theta_r}. \tag{16}$$

Note that not all the new links will yield new connections because the nodes i and j might be already connected. Additionally note that this model does not implement preferential attachment as the linking probability is only dependent on the externally attributed hidden variable θ_i and not to the dynamically acquired degree k_i. Whenever a new link connects node i to an already connected node j the multiedge between node i and node j is reinforced, i.e. the weight of the links between node i and node j increases by one.

Here and in the following we will indicate by $\mathbf{a}$ the adjacency matrix of the network, with t_i the time at which node i has been added to the network, with k_i the node degree and with s_i the node strength, i.e., the sum of the weights of the links incident to node i.

B. The strength of a node and its dependence on the hidden variable

The hidden variable θ_i modulates the temporal evolution of the strength of the node i. In fact in the mean-field approach [1, 5, 44], since at each time an average of $\langle\theta\rangle$ links are added and reinforced, the average strength $s_i(t|t_i, \theta_i, \kappa_i)$ of node i given the time t_i of its arrival in the network, its hidden variable θ_i and its initial strength κ_i obeys the equation

$$\frac{ds_i}{dt} = \langle\theta\rangle \frac{\theta_i}{\langle\theta\rangle t} = \frac{\theta_i}{t} \tag{17}$$

with initial condition $s_i(t_i|t_i, \theta_i, \kappa_i) = \kappa_i$. The solution of this equation is

$$s_i(t|\theta_i, \kappa_i) = \theta_i \ln\left(\frac{t}{t_i}\right) + \kappa_i. \tag{18}$$

Therefore in this model the strength depends both on the time of arrival of the node in the network and on its hidden variable. If we average the strength over the nodes with the same hidden variable however, we see that the average strength $\tilde{s}_i(\theta_i)$ of nodes with hidden variable θ_i is given in the large network limit $t \gg 1$ by

$$\tilde{s}_i(\theta_i) = 2\theta_i. \tag{19}$$

In fact we have

$$
\begin{aligned}
\langle\kappa_i|\theta_i\rangle &= \theta_i \\
\tilde{s}_i(\theta_i) &= \frac{1}{t}\int_1^t \theta_i \ln\left(\frac{t}{t_i}\right) dt_i + \langle\kappa_i|\theta_i\rangle \\
&= 2\theta_i + O\left(\frac{\ln t}{t}\right).
\end{aligned} \tag{20}
$$

This implies that if we attribute to a node a hidden variable θ_i and we consider a set of models in which the time of arrival of node i is taken randomly, the strength of node i is (on average over the different network models) determined only by its hidden variable.

C. Strength distribution

The strength distribution of the model is a convolution of exponentials. To find the strength distribution we use the master equation approach [44] under the assumption that the hidden variable distribution has a well defined average value $\langle\theta\rangle$. To this end we write the

equation for $N_\theta^t(s)$, the average number of nodes with hidden variable θ that have strength $s \geq 0$ at time t, as

$$\frac{N_\theta^t(s)}{dt} = \langle\theta\rangle\Pi(\theta)N_\theta^t(s-1)[1-\delta(s,0)] - \langle\theta\rangle\Pi(\theta)N_\theta^t(s) + \rho(\theta)\hat{P}(\kappa=s|\theta), \qquad (21)$$

where $\delta(x,y)$ indicates the Kronecker delta and where we denote by $\Pi(\theta)$ the probability that a node with hidden variable θ is attached to the new node arrived in the network at time t by one of its connections, i.e.

$$\Pi(\theta) = \frac{\theta}{\sum_{\theta'}\theta'\sum_s N_{\theta'}^t(s)} \simeq \frac{\theta}{\langle\theta\rangle t}. \qquad (22)$$

Given the continuous growth of the network asymptotically in time, for $t \gg 1$ it is possible to assume that

$$N_\theta^t(s) \simeq tP_\theta(s), \qquad (23)$$

where $P_\theta(s)$ is the probability that a random node has strength s and hidden variable θ. By inserting this asymptotic expression in the master equation (21) and solving for $P_\theta(s)$ we get

$$P_\theta(s) = \rho(\theta)\sum_{\kappa=0}^{s}\hat{P}(\kappa|\theta)\frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^{s-\kappa}. \qquad (24)$$

Therefore given the value of the hidden variable θ and the initial number of links κ the strength distribution is exponential. The overall strength distribution $P(s)$ of the model determining the probability that a random node has strength s is given by the integral of $P_\theta(s)$ over all possible value of the hidden variable θ, i.e.

$$P(s) = \int d\theta\rho(\theta)\sum_{\kappa=0}^{s}\hat{P}(\kappa|\theta)\frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^{s-\kappa}. \qquad (25)$$

This result reveals that the strength distribution can be different from the distribution of hidden variables. For instance if all the hidden variables are the same, the strength distribution will still allow for fluctuations of the strengths. However for power-law hidden variable distributions

$$\rho(\theta) \simeq C\theta^{-\gamma} \qquad (26)$$

the strength distribution has a power-law tail with the same exponent γ

$$P(s) \simeq \hat{C}s^{-\gamma} \qquad (27)$$

14

for $s \gg 1$. In fact, by inserting the explicit expression of $\hat{P}(\kappa|\theta)$ and of $\rho(\theta)$ in Eq. (25) we get

$$P(s) = C \int d\theta \frac{\theta^{-\gamma}}{1+\theta} \sum_{\kappa=0}^{s} \frac{1}{\kappa!} \frac{\theta^s}{(\theta+1)^{s-k}} e^{-\theta}. \tag{28}$$

For $s \gg 1$ we can approximate the sum over κ with the infinite sum getting

$$P(s) = C \int d\theta \frac{\theta^{-\gamma}}{1+\theta} \left(\frac{\theta}{\theta+1} \right)^s e^{-1} \simeq \hat{C} s^{-\gamma} \tag{29}$$

where the last expression is valid if $s \gg 1$. Therefore, although in general it is not true that the hidden variable distribution is the same as the strength distribution, in the case of power-law distributed hidden variables the strength distribution displays a power-law tail with the same exponent. Note that this is valid for power-law exponents in the range $\gamma \in (2, 3]$ but also in the range $\gamma \in (1, 2]$. Therefore in this case the hidden variables can be used to directly tune the strength distribution.

D. Connection probability

In this section we derive the expression for the connection probability between any two nodes. Let us consider the probability $P(a_{ij} = 1|\theta_i, \theta_j, \kappa_j, t_j > t_i)$ that node i is connected to node j, i.e. $a_{ij} = 1$ given the hidden variables of node i and node j, their time of arrival with $t_j > t_i$ and the initial strength κ_j of node j. This probability is one minus the probability that all of the initial links of node j do not connect to node i, i.e.

$$P(a_{ij} = 1|\theta_i, \theta_j \kappa_j, t_i < t_j) = 1 - \left(1 - \frac{\theta_i}{\sum_r \theta_r} \right)^{\kappa_j}. \tag{30}$$

If we now average over the probability $\hat{P}(\kappa_j|\theta_j)$ we get the closed form expression

$$P(a_{ij} = 1|\theta_i, \theta_j, t_j, t_i < t_j) = \sum_{\kappa_j} P(\kappa_j) \left[1 - \left(1 - \frac{\theta_i}{\sum_{r=1}^{j} \theta_r} \right)^{\kappa_j} \right] = \left[1 - \exp\left(-\frac{\theta_i \theta_j}{\langle \theta \rangle t_j} \right) \right] \tag{31}$$

where we have assumed that the average of the hidden variables $\langle \theta \rangle$ is well defined. Therefore we have found that the connection probability between two nodes depends both on the hidden variables and on their time of arrival in the network. It follows that the model is not expected to be exchangeable, as this would require a connection probability independent of the time of arrival of the two nodes. However the fact that this connection probability does not *only* depend on the time of arrival of the nodes in the network (or the order in which they are sampled) can be a useful characteristic of a reliable statistical model.

15

E. Degree distribution in the sparse regime

Here we derive the degree distribution of the model in the sparse regime, when we can assume that $p_{ij} \ll 1$. We will show that in this regime, each node has a Poisson degree distribution with an expected average degree $\overline{k}_i$ depending both on the value of its hidden variable and on the time of its arrival in the network.

The probability $P(k_i|\theta_i, t_i)$ that a node i arrived in the network at time t_i and, having hidden variable θ_i, has degree k_i can be calculated starting from the connection probabilities p_{ij} given by Eq. (31). Let us indicate with $\mathbf{a}_i = \{a_{ij}|j \in \{1, 2, \ldots, N\}\}$ the elements of the adjacency matrix in the i-th row indicating the connections of node i. Since node i is connected with each node j with probability p_{ij} given by Eq. (31), the probability $\mathcal{P}(\mathbf{a}_i)$ is given by

$$\mathcal{P}(\mathbf{a}_i) = \prod_{j=1}^{N} \left[p_{ij} a_{ij} + (1 - p_{ij})(1 - a_{ij}) \right]. \tag{32}$$

Using this result we can express the probability $P(k_i|\theta_i, t_i)$ that node i has degree k_i as

$$P(k_i|\theta_i, t_i) = \sum_{\mathbf{a}_i} \mathcal{P}_i(\mathbf{a}_i) \delta \left(k_i, \sum_{j=1}^{N} a_{ij} \right) = \sum_{\mathbf{a}_i} \mathcal{P}(\mathbf{a}_i) \int \frac{d\omega}{\sqrt{2\pi}} e^{-i\omega(k_i - \sum_{j=1}^{N} a_{ij})} \tag{33}$$

where we have used the integral representation of the Kronecker delta $\delta(x, y)$. By performing the sum over all the elements of $\mathbf{a}_i$ we get

$$P(k_i|\theta_i, t_i) = \int \frac{d\omega}{\sqrt{2\pi}} e^{-i\omega k_i} \prod_{j=1}^{N} \left[1 - p_{ij}(1 - e^{-i\omega}) \right] = \int \frac{d\omega}{\sqrt{2\pi}} e^{F(\omega)} \tag{34}$$

where

$$F(\omega) = -i\omega k_i + \sum_{j=1}^{N} \ln \left[1 - p_{ij}(1 - e^{-i\omega}) \right]. \tag{35}$$

For $p_{ij} \ll 1$ we can approximate $F(\omega)$ with

$$F(\omega) = i\omega k_i - \sum_{j=1}^{N} p_{ij}(1 - e^{-i\omega}) = i\omega k_i - \overline{k}_i(1 - e^{-i\omega}) \tag{36}$$

where $\overline{k}_i$ is the expected degree of node i given by

$$\overline{k}_i = \sum_{j=1}^{N} p_{ij}. \tag{37}$$

Note here that since the connection probability p_{ij} depends both on the hidden variables of the nodes i and j and on their arrival time in the network, it follows that also the expected degree $\overline{k}_i$ of node i will be both a function of the node's hidden variable and its time of arrival in the network. Using Eqs. (34) and (36) we can derive the explicit expression for $P(k_i|\theta_i, t_i)$. In fact we have

$$P(k_i|\theta_i, t_i) \simeq \int \frac{d\omega}{\sqrt{2\pi}} e^{i\omega k_i - \overline{k}_i(1-e^{-i\omega})} = \sum_{h=0}^{\infty} \frac{1}{h!} \overline{k}_i^h e^{-\overline{k}_i} \int \frac{d\omega}{\sqrt{2\pi}} e^{i\omega(k_i-h)}, \tag{38}$$

and by identifying the last integral with the Kronecker delta $\delta(h, k_i)$ we get the Poisson distribution

$$P(k_i|\theta_i, t_i) = \frac{\overline{k}_i^{k_i}}{k_i!} e^{-\overline{k}_i}. \tag{39}$$

Therefore the probability that node i, which arrived in the network at time t_i with hidden variable θ_i, has degree k_i is given by the Poisson distribution with average $\overline{k}_i$ given by Eq. (37). It follows that the degree distribution $P(k)$ of the network at time t is given by

$$P(k) = \int d\theta \rho(\theta) \frac{1}{t} \sum_{t'=1}^{t} P(k|\theta, t'). \tag{40}$$

Note that for sufficiently sparse networks where each two connected nodes are typically connected by a link of weight one, the degree of a node can be identified with its strength

$$k_i \simeq s_i. \tag{41}$$

It follows that in this case the degree distribution can be approximated by the strength distribution and we have that if the hidden variables are power-law distributed with power-law γ (as described in Eq. (26)) then also the degree distribution has a power-law tail with the same exponent γ, i.e.

$$P(k) = \tilde{C} k^{-\gamma} \tag{42}$$

for $k \gg 1$.

F. Random permutation of the node sequence

Here we investigate whether the described network process can be related to the generation of uncorrelated networks. In this way we aim at reconciling the non-equilibrium growing

nature of the network model, displaying projectivity, with the properties of exchangeable but not projective uncorrelated network models.

We observe that this expression depends both on the hidden variable and on the time of arrival of the nodes i and j in the network. However if we consider several realizations of the model in which the times of arrival of node i and node j are random, but the hidden variables are preserved, we observe that the probability that node i and node j are connected satisfies

$$
\begin{aligned}
P(a_{ij} = 1 | \theta_i, \theta_j, t = N) &= \frac{1}{N^2} \int_1^N dt_i \int_1^N dt_j \int_1^N d\tau \delta(\tau, \min(t_i, t_j)) \left[1 - \exp\left(-\frac{\theta_i \theta_j}{\langle \theta \rangle \tau} \right) \right] \\
&= \frac{2}{N^2} \int_1^N d\tau \tau \left[1 - \exp\left(-\frac{\theta_i \theta_j}{\langle \theta \rangle \tau} \right) \right] = 2\frac{\theta_i \theta_j}{\langle \theta \rangle N} + o\left(\frac{\theta_i \theta_j}{\langle \theta \rangle N} \right).
\end{aligned} \tag{43}
$$

Therefore if the network is sufficiently sparse, i.e.

$$
\frac{\theta_i \theta_j}{\langle \theta \rangle N} \ll 1, \tag{44}
$$

we have that the expected degree $k_i(\theta_i)$ of a random node i of hidden variable θ_i is given by

$$
\tilde{k}_i(\theta_i) = 2\theta_i, \tag{45}
$$

and the probability that a node with hidden variable θ_i is connected with a node with hidden variable θ_j independently of their time of arrival in the network, is given by the uncorrelated network marginal corresponding to the number of nodes in the sample, i.e.

$$
\tilde{p}_{ij} = P(a_{ij} = 1 | \theta_i, \theta_j, t = N) = \frac{\tilde{k}_i(\theta_i) \tilde{k}_j(\theta_j)}{\langle \tilde{k}(\theta) \rangle N}. \tag{46}
$$

Note that in this case if the sample increases in size and includes $N' > N$ nodes, the probability that node i and node j are connected will satisfy

$$
\tilde{p}'_{ij} = P(a_{ij} = 1 | \theta_i, \theta_j, t = N') = \frac{\tilde{k}_i(\theta_i) \tilde{k}_j(\theta_j)}{\langle \tilde{k}(\theta) \rangle N'}. \tag{47}
$$

In this case the network process induces a probability $\tilde{p}_{ij}$ that depends on the network size N and at the same time enforces the sparseness of the network. In fact the expected degrees $\{\tilde{k}_i\}$ of the nodes are only determined the the hidden variable and are independent on the network size.

G. Entropy of the network model

In order to compare our model with hidden variable distribution $\rho(\theta)$ to an uncorrelated network ensemble in which the expected degrees are $\tilde{k}_i = 2\theta_i$, in this section we use information theory tools. Specifically we will compare the entropy of the two ensembles. The entropy of a network model or of a network ensemble [14–17, 40] is a fundamental tool to evaluate the information content in the network model. It indicates the logarithm of the typical number of networks generated by the ensemble and as such evaluates the complexity of the model and can be used in inference problems [40]. Since for our network model the connection probability p_{ij} of any two pair of nodes is i and j is given by Eq. (31), the entropy of the model is given by

$$S = -\sum_{i<j} \left[p_{ij} \ln p_{ij} + (1 - p_{ij}) \ln(1 - p_{ij}) \right]. \tag{48}$$

where in the sparse regime we can approximate p_{ij} with $t_j > t_i$ as

$$p_{ij} \simeq \frac{\theta_i \theta_j}{\langle \theta \rangle t_j}. \tag{49}$$

Similarly for the uncorrelated network ensemble with connection probability $\tilde{p}_{ij}$ the entropy is given by

$$\tilde{S} = -\sum_{i<j} \left[\tilde{p}_{ij} \ln \tilde{p}_{ij} + (1 - \tilde{p}_{ij}) \ln(1 - \tilde{p}_{ij}) \right]. \tag{50}$$

In order to compare these two entropies we use the explicit expression for the connection probability $\tilde{p}_{ij}$ when we put $\tilde{k}_i(\theta_i) = 2\theta_i$ which reads

$$\tilde{p}_{ij} = 2\frac{\theta_i \theta_j}{\langle \theta \rangle N}. \tag{51}$$

By performing a straightforward calculation we find that S is given, up to the linear terms in N, by

$$S = \langle \theta \rangle \ln(N!) - 2N \langle \theta \ln \theta \rangle + N \langle \theta \rangle \ln \langle \theta \rangle + \langle \theta \rangle N \tag{52}$$

and that the entropy S of our model is smaller than the entropy of the uncorrelated network ensemble. In fact, S differs from $\tilde{S}$ only by

$$\Delta S = S - \tilde{S} \simeq \langle \theta \rangle \ln \left(\frac{N! 2^N}{N^N} \right) \simeq \langle \theta \rangle N \ln \left(\frac{2}{e} \right). \tag{53}$$

19

The entropy difference ΔS quantifies the information loss when the proposed network process is approximated with its corresponding uncorrelated network model. We observe here that the uncorrelated network model is obtained when the causal construction of the original network model is disregarded and the only retained information is the probability $\tilde{p}_{ij}$ that two nodes of hidden variables θ_i and θ_j are connected regardless of their time of arrival in the network. Therefore ΔS captures the loss of information when the causal nature of the original model is disregarded. Interestingly in the large network limit $N \gg 1$, $|\Delta S|$ is low when compared to S revealing the proximity between the two models. Additionally ΔS is only dependent on $\langle \theta \rangle$ indicating that the information loss from one model to the other is independent of the particular distribution of the hidden variables $\rho(\theta)$ as long as $\langle \theta \rangle$ is kept constant.

VI. STATISTICAL TESTING OF THE MODEL

In order to study the utility of the proposed model as a null model for sampled data we consider three power-law networks: the arxiv hep-ph (high energy physics phenomenology) citation network [45, 46], the Berkeley-Stanford web network [47] and the Notre Dame web network [48] of network sizes $N = 34,546$, $N = 685,230$, $N = 325,000$ respectively. All data are freely available on the Stanford Network Analysis Project webpage. To each node of the network we assign a different label $i \in 1, 2, \ldots, N$ according to a random permutation of the indices from 1 up to N. We then assign to each node i of the network a hidden variable

$$\theta_i = \frac{1}{2} k_i, \tag{54}$$

where k_i is the observed degree of node i in the dataset. Given our random node labelling and the hidden variables $\{\theta_i\}_{i=1,2,\ldots,N}$ we have generated a random network according to the proposed network process. Interestingly the proposed model preserves to a large extent the degree distribution (see comparison of the real degree distribution with the one generated by the model in Figure 1). Additionally these results are quite stable if we consider a model generated only by adding a subsample of randomly chosen nodes, showing that the model preserves the degree distribution under random sub-sampling of the nodes (see Figure 1).

The generated model however is to be considered mostly as uncorrelated. In fact if we compare the degree correlations of the real datasets with the degree correlations of the

network generated by the model we observe that the model deviates from the real data and displays very weak/marginal degree correlations (see Figure 2). In fact from the results obtained for the three studied network datasets it seems that the model is able to better reproduce weakly assortative behaviour than strongly disassortative bahaviour. In future, modifications of the proposed model could be envisaged to capture also the degree correlations of real datasets.

VII. CONCLUSIONS

In conclusion, we have given a wide overview of the desirability of the projectivity and exchangeability properties in good statistical models and we have emphasized the difficulty in combining these properties with the sparseness of the network. While this problem is a widely discussed subject in statistics of networks and graph theory, here we have proposed a model that provides a trade-off solution. Our model describes a network process in which nodes and links are subsequently added according to a probability dependent on some hidden variables associated to the nodes. As long as the hidden variables are power-law distributed this model generates a scale-free network with the same exponent. This model is projective but not exchangeable. However, the expected probability that two nodes are connected when one considers a random permutation of the sequence in which nodes are added to the network reduces to the expression valid for the marginal of an uncorrelated exchangeable network with the same expected degrees (given by the double of the hidden variables) provided the network is sufficiently sparse. Finally, we tested this model as a statistical null model for scale-free sparse real networks, showing that it can reproduce the degree distribution (but not degree correlations) also if a partial subset of the data is considered.

55867-P).

---

[1] A.-L. Barabási, *Network Science* (Cambridge University Press, 2016).

[2] M. Newman, *Networks: an introduction* (Oxford University Press, 2010).

[3] E. Estrada, *The structure of complex networks: theory and applications* (Oxford University Press, 2012).

[4] V. Latora, V. Nicosia, and G. Russo, *Complex networks: principles, methods and applications* (Cambridge University Press, 2017).

[5] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).

[6] G. Bianconi and A.-L. Barabási, EPL (Europhysics Letters) **54**, 436 (2001).

[7] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Physical Review Letters **85**, 4633 (2000).

[8] P. L. Krapivsky, S. Redner, and F. Leyvraz, Physical Review Letters **85**, 4629 (2000).

[9] G. Bianconi, R. K. Darst, J. Iacovacci, and S. Fortunato, Physical Review E **90**, 042806 (2014).

[10] P. Krapivsky and S. Redner, Journal of Statistical Mechanics: Theory and Experiment **2017**, 073405 (2017).

[11] Z. Wu, G. Menichetti, C. Rahmede, and G. Bianconi, Scientific Reports **5**, 10073 (2015).

[12] G. Bianconi and C. Rahmede, Physical Review E **93**, 032315 (2016).

[13] G. Bianconi and C. Rahmede, Scientific Reports **7**, 41974 (2017).

[14] G. Bianconi, EPL (Europhysics Letters) **81**, 28005 (2007).

[15] G. Bianconi, Physical Review E **79**, 036114 (2009).

[16] K. Anand and G. Bianconi, Physical Review E **80**, 045102 (2009).

[17] K. Anand and G. Bianconi, Physical Review E **82**, 011116 (2010).

[18] O. Sagarra, C. P. Vicente, and A. Dïaz-Guilera, Physical Review E **88**, 062806 (2013).

[19] T. Squartini, J. de Mol, F. den Hollander, and D. Garlaschelli, Physical Review Letters **115**, 268701 (2015).

[20] T. A. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock, Sociological Methodology **36**, 99 (2006).

[21] J. Park and M. E. Newman, Physical Review E **70**, 066117 (2004).

[22] D. Garlaschelli and M. Loffredo, Phys Rev E **78**, 015101 (2008).

[23] T. P. Peixoto, Physical Review X **4**, 011047 (2014).

[24] T. P. Peixoto, Physical Review E **85**, 056122 (2012).

[25] A. Goldenberg, A. X. Zheng, S. E. Fienberg, E. M. Airoldi, et al., Foundations and Trends in Machine Learning **2**, 129 (2010).

[26] O. Kallenberg, *Foundations of modern probability* (Springer, New York, 2002).

[27] C. R. Shalizi and A. Rinaldo, Ann Stat **41**, 508 (2013).

[28] N. Spencer and C. R. Shalizi, arXiv preprint arXiv:1709.09702 (2017).

[29] D. J. Aldous, Journal Multivariate Analysis **11**, 581 (1981).

[30] P. Diaconis and S. Janson, Rendiconti di Matematica e delle sue applicazioni **28**, 33 (2008).

[31] D. Krioukov and M. Ostilli, Physical Review E **88**, 022808 (2013).

[32] C. Borgs, J. T. Chayes, H. Cohn, and Y. Zhao (2014), 1401.2906.

[33] F. Caron and E. B. Fox, Journal of the Royal Statistical Society Series B (Statistical Methodology) **79**, 1295 (2017).

[34] V. Veitch and D. M. Roy (2015), 1512.03099.

[35] C. Borgs, J. T. Chayes, H. Cohn, and N. Holden (2016), 1601.07134.

[36] H. Crane and W. Dempsey (2016), 1603.04571.

[37] D. Cai, T. Campbell, and T. Broderick, in *NIPS* (2016).

[38] S. Janson, Journal Statistical Physics pp. 1–37 (2017).

[39] P. van der Hoorn, G. Lippner, and D. Krioukov, Journal Statistical Physics (2017).

[40] G. Bianconi, P. Pin, and M. Marsili, Proceedings of the National Academy of Sciences **106**, 11433 (2009).

[41] We note here that while in the statistics literature the term *sparse network* refers to a network where the average degree is sublinear in the number of nodes N, i.e. $\langle k \rangle = o(N)$ here we adopt the physics terminology and use the term *sparse network* to indicate networks with average degree independent of N, i.e. $\langle k \rangle = O(1)$. These networks are also indicated in the statistical literature as *ultra-sparse*.

[42] G. Bianconi, Physics Letters A **303**, 166 (2002).

[43] K. Zhao, A. Halu, S. Severini, and G. Bianconi, Physical Review E **84**, 066113 (2011).

[44] S. N. Dorogovtsev and J. F. Mendes, *Evolution of networks: From biological nets to the Internet and WWW* (Oxford University Press, 2013).

[45] J. Leskovec, J. Kleinberg, and C. Faloutsos (ACM, 2005), pp. 177–187.

[46] J. Gehrke, P. Ginsparg, and J. Kleinberg, ACM SIGKDD Explorations Newsletter **5**, 149 (2003).

[47] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, Internet Mathematics **6**, 29 (2009).

[48] R. Albert, H. Jeong, and A.-L. Barabási, Nature **401**, 130 (1999).
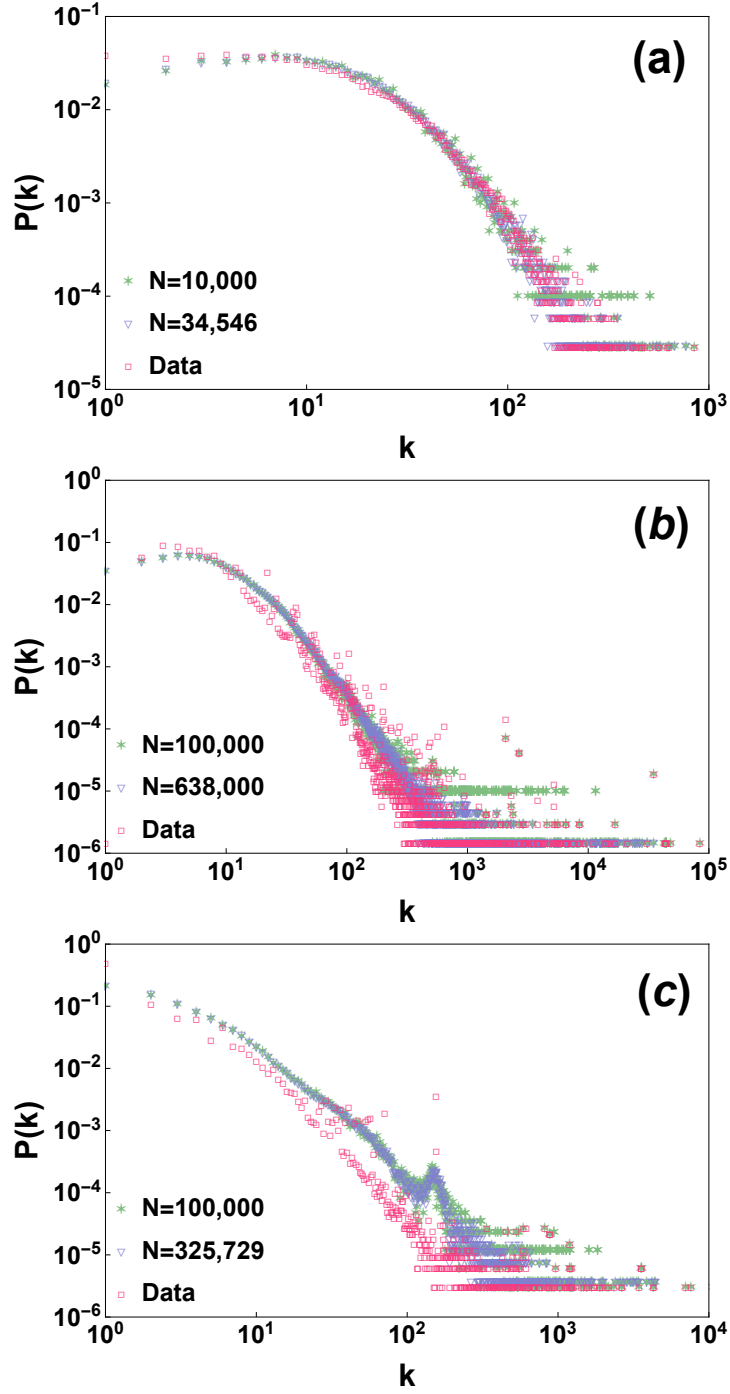
Figure 1: The degree distributions $P(k)$ of the three analysed datasets is compared with the results of the model generated by using all the nodes of the network or with just a subsample of nodes of the network of size N. Panels (a,b,c) display the results for the arxiv hep-ph citation network [45, 46] ($N = 34,546$) the Berkeley-Stanford web network [47] ($N = 685,546$) and the Notre Dame web network [48] ($N = 325,000$) respectively.
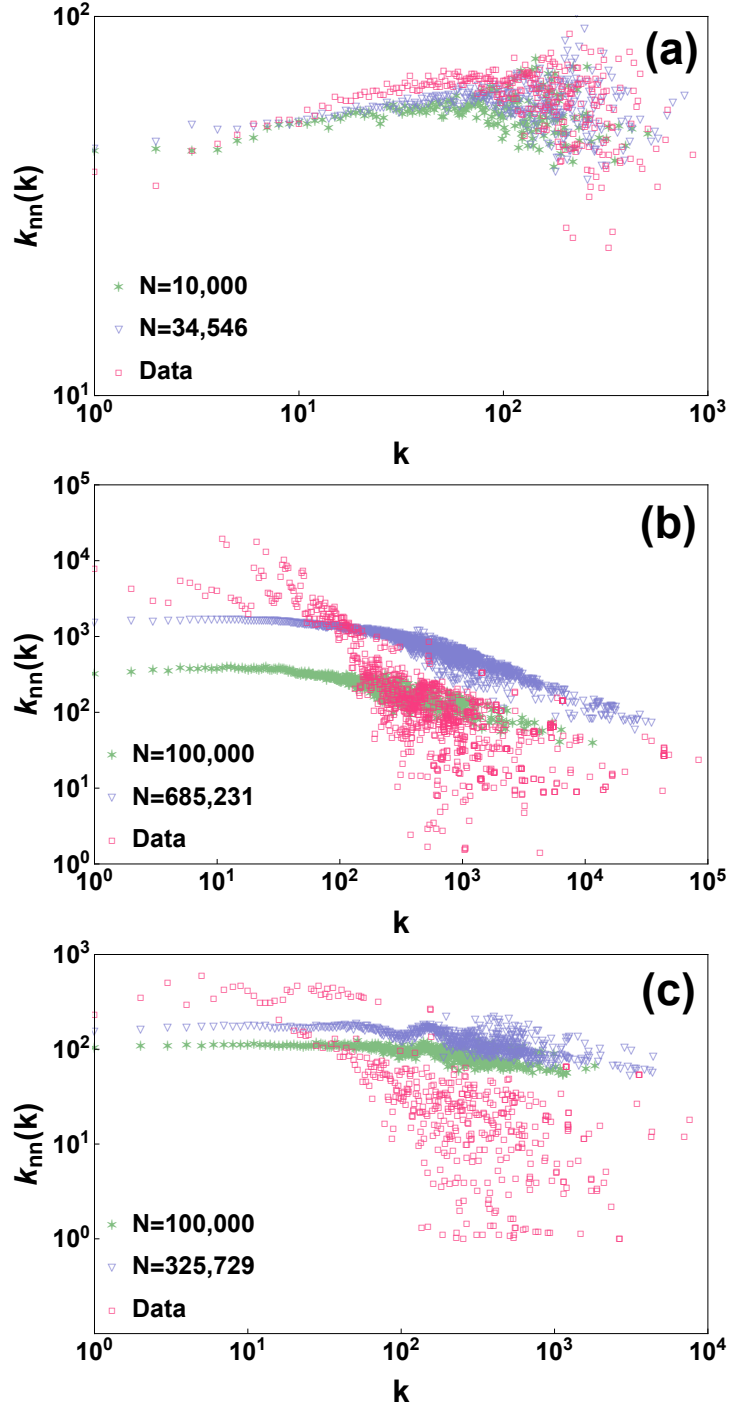
Figure 2: The average degree $k_{nn}(k)$ of the neighbour of a node of degree k of the three analysed datasets is compared with the results of the model generated by using all the nodes of the network or with just a subsample of nodes of the network of size N. Panels (a,b,c) display the results for the arxiv hep-ph citation network [45, 46] ($N = 34,546$) the Berkeley-Stanford web network [47] ($N = 685,546$) and the Notre Dame web network [48] ($N = 325,000$) respectively.