# Evaluating the Natural Variability in Generative Models for Complex Networks

Viplove Arora and Mario Ventresca(✉)

School of Industrial Engineering, Purdue University, West Lafayette IN 47907, USA
`mventresca@purdue.edu`

**Abstract.** Complex networks are used to represent real-world systems using sets of nodes and edges that represent elements and their interactions, respectively. A principled approach to understand these network structures (and the processes that give rise to them) is to formulate generative models and infer their parameters from given data. Ideally, a generative model should be able to synthesize networks that belong to the same population as the observed data, but most models are not designed to accomplish this task. Due to the scarcity of data in the form of populations of networks, generative models are typically formulated to learn parameters from a single network observation, hence ignoring the natural variability of network populations. In this paper, we evaluate four generative models with respect to their ability to synthesize networks that belong to the same population as the observed network. Our empirical analysis quantifying the ability of network models to replicate characteristics of a population of networks highlights the need for rethinking the way we evaluate the goodness of fit of new and existing network models.

**Keywords:** Network models · Network populations · Network analysis

## 1 Introduction

Many natural and artificial systems can be described as networks composed of sets of nodes and edges that represent system elements and their interactions, respectively [20]. The structure of these complex networks can also capture the functional abilities of the system they represent. The analysis and modeling of complex networks has provided transformative perspectives, models and methods in diverse application domains such as computer science, sociology, chemistry, biology, anthropology, psychology, geography, history and engineering [11,20]. In particular, the increasing availability of network data from a wide variety of sources such as the internet, online social networks, citation and collaboration networks, biological networks (brain connectivity, protein-protein interactions), etc. has fueled a great deal of interest in the analysis and modeling of networks.

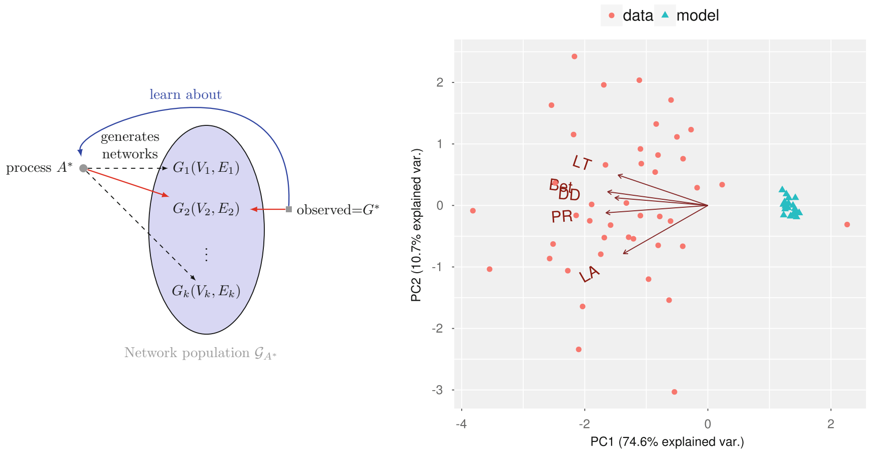
L. M. Aiello et al. (Eds.): COMPLEX NETWORKS 2018, SCI 812, pp. 743–754, 2019.
https://doi.org/10.1007/978-3-030-05411-3_59

A goal of network modeling is to solve the problem of decoding how the observed structure of a network supports its perceived/desired function [1]. As a consequence, a long-standing question in the network science community has been regarding the existence of a model capable of generating synthetic networks that are statistically representative of real networks. Most of the existing models either make assumptions biased by system-specific observations that are not plausible across domains, or focus on replicating a few predefined topological features, such as degree distribution and clustering, at the expense of other potentially more important characteristics. Without any indication that they are either necessary or sufficient as descriptors for the actual network data, these summary quantities can often be highly misleading [11]. Further, even when a model is capable of consistently reproducing a set of target properties, it might fail to capture the naturally occurring stochasticity in those properties [12]. This is depicted in Fig. 1b, which shows the variance in a population of real-world networks (evaluated using five different network properties) compared with the networks synthesized by *dk*-random graphs [21].

**Definition 1 (Network Population).** *Let $G_1(V_1, E_1)$ be a network that has non-zero probability of being generated using the process $A^*$. A set of such realizations $\mathcal{G}_{A^*} = \{G_1(V_1, E_1), \dots, G_k(V_k, E_k)\}$ is called a network population.*

The inability of certain network models to reproduce the naturally occurring variability in networks can be attributed to the fact that they sample each edge independently through Bernoulli distributions [19]. Further, modeling networks based on a single network observation might bias a network model to synthesize networks that over-fit the observed network, thus ignoring the natural variability. In statistical analysis, the goodness of fit of a model is evaluated by measuring the discrepancy between observed values and the values expected under the model in question. Similarly, in the context of networks, we would like to *evaluate the ability of a model to approximate the population of networks that can be created by a process $A^*$ using a single observed network $G^*$* (see Fig. 1a for a pictorial representation). Unfortunately, we generally do not have a population of independent instances of networks that can be used to draw a set of samples [17]. To recap, an ideal generative model $M$ would exactly correspond to the true process $A^*$ that defines the dynamical processes responsible for the observed data $G^*$. That is, if $A^*$ defines a probability distribution $\mathbb{P}_G(A^*)$ $\forall G \in A^*$, then $\mathbb{P}_G(M)$ and $\mathbb{P}_G(A^*)$ would be identical. As stated above, $A^*$ is usually unknown and the number of observed networks in the data $G^*$ are usually small (sometimes only one).

In this work, we study the distributional properties of four competing generative models: Chung-Lu model and *dk*-random graphs, which are designed to match specific properties of the observed network, and exponential random graphs and action-based network generators, which were designed to capture local graph properties (Sect. 2 briefly describes each model). We consider networks drawn from three known processes and three real-world populations. For each model, we learn parameters from a representative sample (see Fig. 1a).



(a) Assuming $G^*$ is not an outlier, how well do existing network models approximate the process A*?

(b) Variance in real-world data of Indian villages versus networks synthesized by *dk*-random graphs.

**Fig. 1.** Evaluation of network models: Fig. 1a depicts the procedure used for evaluating network models in this paper, while Fig. 1b highlights the need for such a procedure. Social networks in Indian villages [4] is the data used in Fig. 1b. The networks are compared using the Kolmogorov-Smirnov distance on five different network properties, namely: degree distribution (DD), local assortativity (LA), local transitivity (LT), PageRank (PR), and Betweenness (Bet). Principal component analysis of the five properties was performed, where the first two principal components were able to account for more than 85% of the variance.

Then, the learned models are used to synthesize networks followed by an investigation of their distributional properties. This evaluation is done by comparing the statistical properties of the synthesized networks with the properties of the corresponding population of networks.

## 2 Background

In this section, we briefly introduce the four generative models that are used in the empirical analysis in Sect. 4. Where applicable, details regarding user-defined inputs have also been provided. Further, as shown in Fig. 1a, each model uses the network $G^*$ to learn a fixed set of parameters.

### 2.1 Chung-Lu Model

In the Chung-Lu model [6,7], a vertex $i$ is assigned a degree $d_i$ from the given degree distribution and an edge is placed between the vertex pair $(i, j)$ with probability proportional to $d_i d_j$, i.e. the probability that an edge exists between nodes $i$ and $j$ is given by

$$\mathbb{P}_{i,j} = \frac{d_i d_j}{\sum_k d_k}$$

The Chung-Lu model is often used as the baseline for comparison owing to its simplicity and ability to synthesize fairly realistic networks [24]. Unfortunately, the Chung-Lu model synthesizes networks with low clustering coefficients making it unsuitable for most real-world applications.

### 2.2 Exponential Random Graphs

One of the most popular statistical network models in the social science literature are the exponential random graph models (ERGM) [29,32]. ERGMs represent probability distributions over networks with an exponential linear model that uses feature counts of local graph properties considered relevant by the modeler (for example, edges, triangles, paths, etc.):

$$P(\mathbf{Y} = G^* | \boldsymbol{\theta}) = \frac{1}{Z} \exp(\boldsymbol{\theta}^T \boldsymbol{\phi}(G^*))$$

where (i) $\boldsymbol{\phi}(G^*)$ are feature counts of $G^*$; (ii) $\boldsymbol{\theta}$ are parameters to be learned; (iii) $Z$ is a normalizing constant. Though ERGMs are the most widely used models for social networks, they are plagued with the degeneracy problem (i.e., the probability distribution is biased towards empty and complete networks), whereas real-world networks are sparse. In our experimental evaluation, the following feature counts $\boldsymbol{\phi}(G^*)$ were used as they are known to be capable of circumventing the degeneracy problem (see [13,28] for more details): (i) total number of edges, (ii) geometrically weighted degree distribution, (iii) geometrically weighted dyadwise shared partner distribution, and (iv) geometrically weighted edgewise shared partner distribution.

### 2.3 *dk*-Random Graphs

In [21] it was observed that fixing some structural properties in a network model to those observed in the given network can lead to the appearance of other statistical properties as a consequence. These observations follow from earlier research on the *dk*-series [16], which defines a series of null models or random graph ensembles [21]. Consequently, *dk*-random graphs [21] model networks as random ensembles, where ensemble size is controlled using *dk*-distributions. *dk*-random graphs rely on ergodic edge-swapping operations to sample networks from an ensemble defined using the chosen *dk*-distributions. Experimental results [2,21,27] have shown that the networks synthesized by *dk*-random graphs have very low dissimilarity to most real-world networks. Despite this fact, the limited inferential capabilities and inability to perform tasks such as compression, extrapolation, etc. limit the utility of *dk*-random graphs. In our empirical analysis, we used the *dk*2.5 variant as it is known to outperform other network models on a variety of measures [2,27].

### 2.4 Action-Based Network Generators

The action-based approach of [2] models networks using local node interactions based on simple link creation processes known as actions. An action is a decision process a node uses to form a link with another node. Given a pre-defined set of actions, the aim of action-based networks is to learn a probability distribution over these actions, such that the resultant model can synthesize networks statistically similar to a given network observation. A synthesis algorithm $f(\mathbf{M}, n)$ can then be used to synthesize networks containing $n$ nodes using the learned action-based model $\mathbf{M}$, leading to action-based network generators (ABNG). The fundamental idea behind action-based networks is to define a unifying network generative process, which follows from observations by [33] who note that there must exist an assembling algorithm to combine local mechanisms for emergence of different complex network structures. For an observed network $G^*$, the action-based model $\mathbf{M}$ is determined by solving a multi-objective optimization problem. In our empirical analysis, we used degree distribution, local assortativity [25], and local transitivity of the observed network as the set of network properties in the objective function.

## 3 Experimental Setup

Evaluation of the distributional properties of a generative model requires a well-defined methodology that correctly represents the distribution over networks. Although a model-based technique for hypothesis testing of networks has been proposed in the literature [18], it heavily relies on the choice of a baseline model. Alternatively, one could build on the concept of a network morphospace [3], which provides a coarse-grained approach for classifying and mapping network architectures according to a set of network-level structural characteristics. The network morphospace can be transformed to a network dissimilarity space ($\mathfrak{D}_G \subset \mathbb{R}^d$), where networks are placed based on their dissimilarity to the observed network $G^* \in G$ with respect to a variety of node-level structural characteristics. The true process and network models also have counterpart distributions $\mathbb{P}_{\mathfrak{D}_G}(A^*)$ and $\mathbb{P}_{\mathfrak{D}_G}(M)$ in the network dissimilarity space. In an appropriately defined dissimilarity space, if $\mathbb{P}_{\mathfrak{D}_G}(M)$ sufficiently approximates $\mathbb{P}_{\mathfrak{D}_G}(A^*)$, we might be able to conclude that model $M$ can synthesize networks that belong to the same population as the observed network $G^*$.

The utility of such a network dissimilarity space relies heavily on the choice of node-level metrics used for network comparison. Network science provides numerous quantitative tools to measure and classify different patterns of local and global network architectures across disparate types of systems. A set of node-level measures that could prove particularly useful for the network dissimilarity space is provided by the $dk$-series [21], which is a systematic series of properties $(Y_0, Y_1, \dots)$ of network structure defined in a way such that each $Y_i$ provides more detailed information about the network structure and $Y_n$ fully characterizes a network with $n$ nodes. [21] have shown that the first three terms in the $dk$-series ($Y$ = degrees + correlations + clustering/transitivity) are capable of almost fully defining local and global organization of most real-world networks that do not exhibit community structure.

## 4 Experimental Results

In our experiments to evaluate the distributional properties of generative models, we propose to use the Kolmogorov-Smirnov statistic for evaluating the dissimilarity between networks based on node-level properties of degree, correlations and clustering. To examine the ability of existing generative models to approximate the ground truth process using a single network observation (assuming it is representative of the true process with respect to the measures of interest), we propose two different experiments: (i) a controlled experiment where the true process is known, and (ii) set of real-world networks that have most likely evolved from a common generative process (for example, social interaction networks of different villages).

The following datasets were used for real-world network populations: (i) Contact Networks: 69 daily cumulated networks where nodes represent visitors of a Science Gallery while the edges represent close-range face-to-face proximity between the concerned persons [14]; (ii) Social Networks in Indian Villages: Data from a survey of social networks in 75 villages in rural southern Karnataka, a state in India [4]; and (iii) Autonomous Systems: The graph of routers comprising the Internet organized into sub-graphs called Autonomous Systems (AS). The dataset [30] contains 733 daily instances which span an interval of 785 days from November 8 1997 to January 2 2000. The first 100 networks were used in this study. The networks obtained from each of these datasets have most likely evolved from similar underlying social mechanisms, hence belonging to the same network population.

Figure 2 shows the results for the first set of experiments when the Barabási-Albert [5] and Forest Fire models [15] are used as the true processes $A^*$. For the second experiment, we consider the three real-world network populations described above, with results presented in Figs. 3 and 5. Results presented in Figs. 2, 3, 4 and 5 are composed of three different plots:

1. Scatter plots below the diagonal show each synthesized/real network as a point in the network dissimilarity space, where the coordinates are computed using the Kolmogorov-Smirnov distance of the associated properties when the network is compared to the observed network $G^*$ (the observed network itself is at the (0,0) position). Network models (colored triangles) showing higher overlap with networks originating from the true process (black dots) are better.
2. In the blocks above the diagonal, we evaluate the amount of overlap between $\mathbb{P}_{\mathfrak{D}_G}(A^*)$ and $\mathbb{P}_{\mathfrak{D}_G}(M)$ using the 2-D KS distance [22] (lower the better). This quantifies the extent to which a given generative model is able to reproduce the distributional properties of the population representing the true process.
3. Plots along the diagonal show the density distributions of the Kolmogorov-Smirnov distance of the associated properties when the network is compared to the observed network $G^*$.

Based on Figs. 2 and 3 we can easily conclude that ABNG consistently outperforms the other models considered here by replicating the natural variability

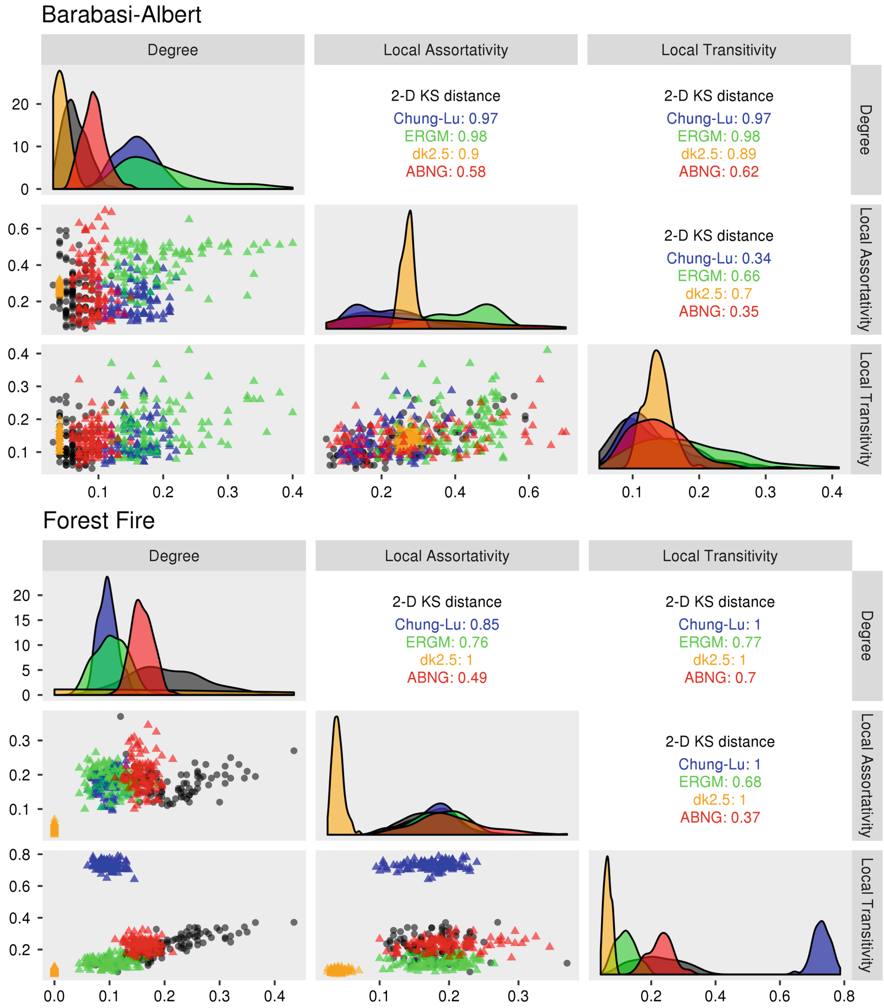


**Fig. 2.** Empirical evaluation of the ability of network models to approximate the ground truth system based on observation of a single network. The Barabási-Albert and Forest Fire models are used as the true generators.

in the network population of the true process in both the experimental settings. The plots also show that *dk*-random graphs, which are considered to be the state-of-art, fail to capture the variability of the true generative process and potentially over-fit the observed network. This leads us to question the fundamental idea behind *dk*-random graphs, i.e. whether exactly preserving the distribution of differently sized subgraphs of a given network leads to a good model for real-world networks. In fact, in most cases we see that the Chung-Lu model, by matching the degree distribution in expectation, outperforms *dk*-random graphs. These results highlight the need for evaluating the ability of a generative model to

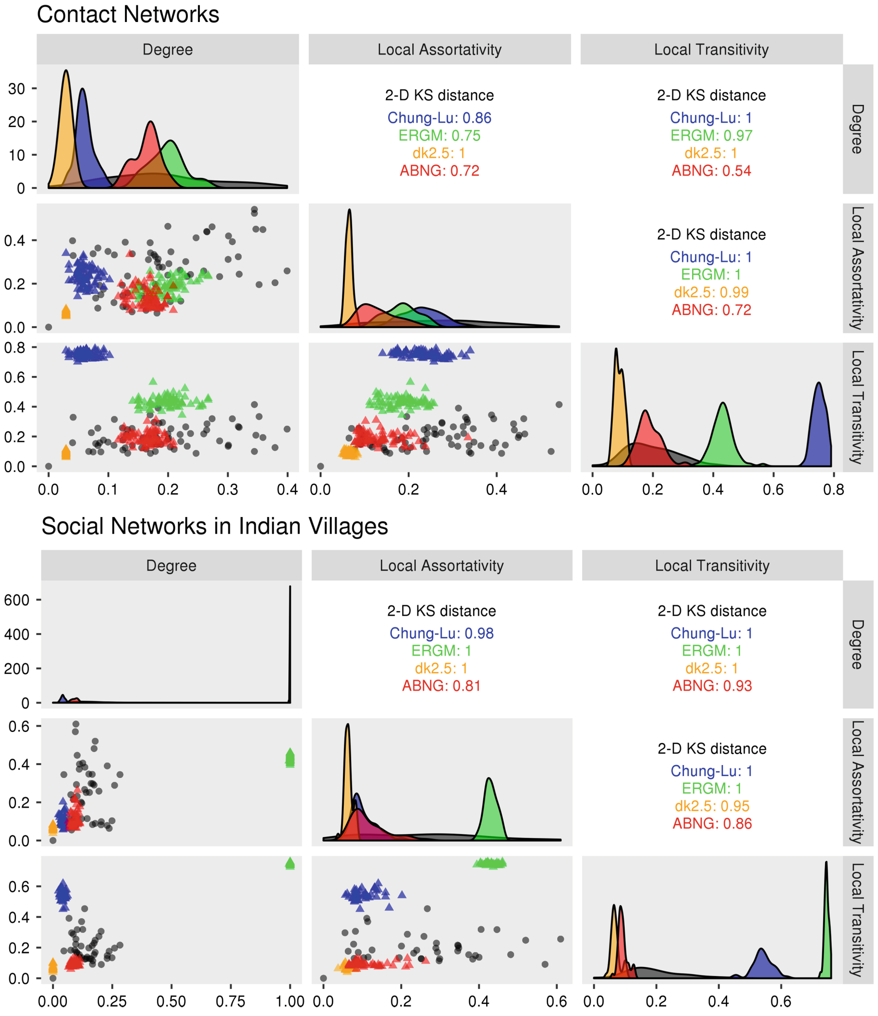


**Fig. 3.** Empirical evaluation of the ability of network models to approximate the ground truth system based on observation of a single network. Two real-world datasets were considered: contact networks, and social networks in Indian villages.

capture the distributional properties of a network population as comparing only with the observed network might produce misleading results.

### 4.1 Networks with Community Structure

While the network dissimilarity space defined in Sect. 3 works well for networks without communities, it will prove ineffective for networks with community structures, which is property seen in most real-world networks [10]. In this section, we extend the network dissimilarity space by adding a fourth dimension to compare the community structures of two networks. To compare community structures, we first use a community detection algorithm[1] to decide the membership for each node, followed by evaluation of the normalized mutual information measure [8]. We also add the microcanonical stochastic block model (referred to as SBM-fit in the plots) [23] to our set of generative models and evaluate its ability to replicate the community structure of these networks.

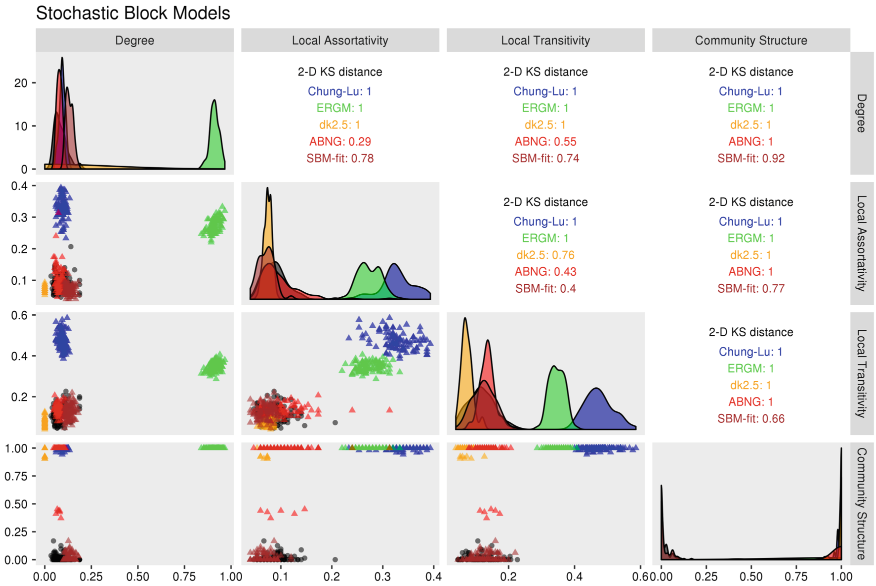


**Fig. 4.** Empirical evaluation of the ability of network models to approximate the ground truth system based on observation of a single network. The stochastic block model is used as the true generator, and the ability of different models to replicate the community structure is tested.

Again, we performed two different experiments to test the validity of our extended network dissimilarity space: (i) a controlled experiment where the true process is known, and (ii) set of real-world networks (with communities) that have most likely evolved from a common generative process. For the first case, we used the standard version of the stochastic block model [9,31] with 3 communities of different sizes, and the results can be seen in Fig. 4. As expected, ABNG performs well on the original measures, but fails to reproduce the community structure, while the fitted SBM is the most likely candidate capable of replicating the true process. This is an expected result as the four original models are not designed to create networks with communities. Fig. 5 shows the results for the network of Autonomous Systems, where only the fitted SBM was able to capture some of the features of the true process. Results presented in Fig. 5 show the

---

[1] Infomap community detection algorithm [26] was used in our experiments.

inability of the microchanonical block model to reproduce the local transitivity of the true generative process, thus creating an exciting direction for future research.

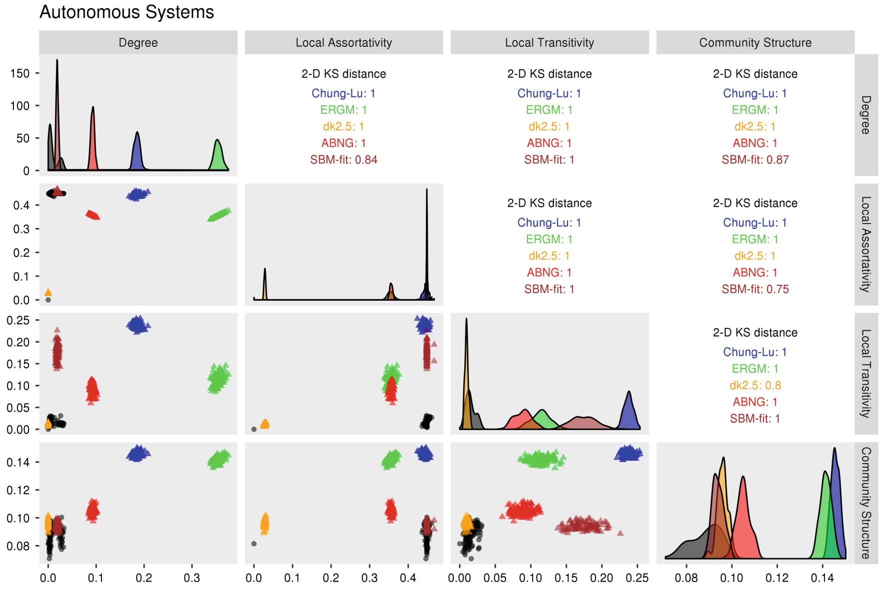


**Fig. 5.** Empirical evaluation of the ability of network models to approximate the ground truth system based on observation of a single network. The ability of different models to replicate the community structure of networks of autonomous systems is tested.

In summary, our empirical analysis has highlighted the importance of considering distributional properties of network populations for evaluating generative models of complex networks. This shows that there is an urgent need to rethink the network modeling problem and create new models that can reproduce the variability in the structural properties of network populations.

## 5 Conclusions

Traditional approaches for evaluating the ability of a network model to synthesize networks exhibiting real-world characteristics have compared the similarity of the synthesized networks with the observed network. While this approach assumes that the particular observation is representative of the underlying process that created the observation, it does not account for the natural variability of the population from which it is sampled. Our experiments have highlighted the importance of considering network populations for evaluating generative models. Although it is difficult to obtain data corresponding to network populations, we have shown that it is possible to establish a baseline test set to evaluate the ability of a network model to capture the distribution of network populations. This test set can then be used for preliminary validation of a network model before it is used for drawing conclusions about real-world networks.

**Acknowledgements.** This material is based upon work supported by the National Science Foundation under Grant No. 1762633.

## References

1. Alderson, D.L.: Catching the "network science" bug: insight and opportunity for the operations researcher. Oper. Res. **56**(5), 1047–1065 (2008). https://doi.org/10.1287/opre.1080.0606
2. Arora, V., Ventresca, M.: Action-based modeling of complex networks. Sci. Rep. **7**(1), 6673 (2017). https://doi.org/10.1038/s41598-017-05444-4
3. Avena-Koenigsberger, A., Goni, J., Sole, R., Sporns, O.: Network morphospace. J. R. Soc. Interf. **12**(103), 20140,881–20140,881 (2014). https://doi.org/10.1098/rsif.2014.0881
4. Banerjee, A., Chandrasekhar, A.G., Duflo, E., Jackson, M.O.: The diffusion of microfinance. Science **341**(6144), 1236,498 (2013)
5. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(October), 509–512 (1999). https://doi.org/10.1126/science.286.5439.509
6. Chung, F., Lu, L.: Connected components in random graphs with given expected degree sequences. Annal. Comb. **6**(2), 125–145 (2002)
7. Chung, F., Lu, L.: The average distances in random graphs with given expected degrees. Proc. Nat. Acad. Sci. **99**(25), 15879–15882 (2002)
8. Danon, L., Díaz-Guilera, A., Duch, J., Arenas, A.: Comparing community structure identification. J. Stat. Mech. Theory Exp. (09), P09,008–P09,008 (2005). https://doi.org/10.1088/1742-5468/2005/09/P09008
9. Faust, K., Wasserman, S.: Blockmodels: interpretation and evaluation. Soc. Netw. **14**(1–2), 5–61 (1992). https://doi.org/10.1016/0378-8733(92)90013-W
10. Fortunato, S., Hric, D.: Community detection in networks: a user guide. Phys. Rep. **659**, 1–44 (2016). https://doi.org/10.1016/j.physrep.2016.09.002
11. Goldenberg, A.: A survey of statistical network models. Foundations and trends® *R*. Mach. Learn. **2**(3), 235–274 (2009). https://doi.org/10.1561/2200000008
12. Gutfraind, A., Meyers, L.A., Safro, : I.: Multiscale Network Generation. arXiv:1207.4266, 28 (2012)
13. Hunter, D.R.: Curved exponential family models for social networks. Soc. Netw. **29**(2), 216–230 (2007). https://doi.org/10.1016/j.socnet.2006.08.005
14. Isella, L., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J.F., den Broeck, W.: What's in a crowd? analysis of face-to-face behavioral networks. J. Theoret. Biol. **271**(1), 166–180 (2011)
15. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution. ACM Trans. Knowl. Discov. Data **1**(1), 2-es (2007). https://doi.org/10.1145/1217299.1217301
16. Mahadevan, P., Krioukov, D., Fall, K., Vahdat, A.: Systematic topology analysis and generation using degree correlations. ACM SIGCOMM Comput. Commun. Rev. **36**(4), 135 (2006). https://doi.org/10.1145/1151659.1159930
17. Moreno, S., Neville, J.: an investigation of the distributional characteristics of generative graph models. In: Proceedings of the 1st Workshop on Information in Networks (2009)

18. Moreno, S., Neville, J.: Network hypothesis testing using Mixed Kronecker product graph models. In: Xiong, H., Karypis, G., Thuraisingham, B.M., Cook, D.J., Wu, X. (eds.), pp. 1163–1168 (2013)
19. Moreno, S., Neville, J., Kirshner, S.: Tied Kronecker product graph models to capture variance in network populations. ACM Trans. Knowl. Discov. Data **20**(3), 1–40 (2018). https://doi.org/10.1145/3161885
20. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**(2), 167–256 (2003)
21. Orsini, C., Dankulov, M.M., Colomer-de Simón, P., Jamakovic, A., Mahadevan, P., Vahdat, A., Bassler, K.E., Toroczkai, Z., Boguñá, M., Caldarelli, G., Fortunato, S., Krioukov, D.: Quantifying randomness in real networks. Nat. Commun. **6**(May), 8627 (2015). https://doi.org/10.1038/ncomms9627
22. Peacock, J.A.: Two-dimensional goodness-of-fit testing in astronomy. Month. Notices R. Astron. Soc. **202**(3), 615–627 (1983)
23. Peixoto, T.P.: Nonparametric Bayesian inference of the microcanonical stochastic block model. Phys. Rev. E **95**(1), 1–21 (2017). https://doi.org/10.1103/PhysRevE.95.012317
24. Pinar, A., Seshadhri, C., Kolda, T.G.: The similarity between stochastic Kronecker and Chung-Lu graph models. In: Zaki, M., Obradovic, Z., Tan, P.N., Banerjee, A., Kamath, C., Parthasarathy, S. (eds.), Proceedings of the 2012 SIAM International Conference on Data Mining, pp. 1071–1082. Society for Industrial and Applied Mathematics, Philadelphia, PA (2012). https://doi.org/10.1137/1.9781611972825.92
25. Piraveenan, M., Prokopenko, M., Zomaya, : A.Y.: Local assortativeness in scale-free networks. EPL (Europhysics Letters) **84**(2), 28,002 (2008). https://doi.org/10.1209/0295-5075/84/28002
26. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proc. Nat. Acad. Sci. **105**(4), 1118–1123 (2008). https://doi.org/10.1073/pnas.0706851105
27. Schieber, T.A., Carpi, L., Díaz-Guilera, A., Pardalos, P.M., Masoller, C., Ravetti, M.G.: Quantification of network structural dissimilarities. Nat. Commun. **8**(May 2016), 13,928 (2017). https://doi.org/10.1038/ncomms13928
28. Snijders, T.A.B., Pattison, P.E., Robins, G.L.: Specifications for exponential random graph models. In: Sociological Methodology, p. 44., Handcock, M.S.: New (2004). https://doi.org/10.1111/j.1467-9531.2006.00176.x
29. Strauss, D.: On a general class of models for interaction. SIAM Rev. **28**(4), 513–527 (1986)
30. Views, R.: University of Oregon route views project (2000)
31. Wasserman, S., Anderson, C.: Stochastic a posteriori blockmodels: construction and assessment. Soc. Netw. **9**(1), 1–36 (1987). https://doi.org/10.1016/0378-8733(87)90015-3
32. Wasserman, S., Pattison, P.: Logit models and logistic regressions for social networks. Psychometrika **60**, 401–425 (1996)
33. Zheng, B., Wu, H., Kuang, L., Qin, J., Du, W., Wang, J., Li, D.: A simple model clarifies the complicated relationships of complex networks. Sci. Rep. **4**, 6197 (2014). https://doi.org/10.1038/srep06197