Impact on the topology of power-law networks from anisotropic and localized access to information

Zhenfeng Cao,[1,*] Zhou He,[2] and Neil F. Johnson[3]
[1]*Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong*
[2]*School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China*
[3]*Physics Department, George Washington University, Washington, DC 20052, USA*

(Received 17 December 2017; published 9 October 2018)

Preferential attachment is a popular candidate mechanism for generating power-law networks. However, incoming nodes require global information about existing nodes' connectivities before connecting, whereas such information access within real-world networks may be only anisotropic and localized. Here we investigate how anisotropic and localized information access affect the resulting network topology. We find that anisotropy impacts the power-law exponent significantly but has only a weak influence on the clustering coefficient. By contrast, we find that locality influences the clustering coefficient significantly but has only weak influence on the power-law exponent. We show that this generalized network-generation mechanism is capable of generating networks with a broad range of power-law exponents and clustering coefficients. Our findings contribute to the debate about why so many real-world networks have degree distributions that crudely resemble power laws, even if this resemblance doesn't survive strict statistical testing procedures.

DOI: 10.1103/PhysRevE.98.042307

I. INTRODUCTION

Understanding the possible mechanisms that generate networks with approximate power-law degree distributions has been a major topic in recent years [1–3]. Although real-world networks are never perfect power laws, the fact that so many resemble a crude approximation to a power law—coupled with physicists' intrinsic interest in scale-free behavior—provides motivation for continuing to explore minimal mechanisms that generate power-law networks.

Among such minimal mechanisms, the so-called preferential attachment (PA) mechanism has garnered significant interest in connection with the growth of network systems as diverse as scientific collaboration networks [3–8], the World Wide Web [4,9–12], actor collaboration networks [11,13,14], social networks [15–22], and chemical and biological networks [23–26]. Existing mechanisms proposed to produce power-law network degree distributions can crudely be divided into two categories, depending on whether the network generation mechanism assumes a newly added node has access to some global information about the existing nodes as in Refs. [3,6,27–30] when selecting which node to connect to, or it does not as in the copying models of Refs. [24,31–33]. Real-world scenarios would seem to favor the latter category of models, since it is more likely in reality that a new node has only local access to information about existing nodes [32]. Although the topic of how a localized PA mechanism can generate a power-law distribution has been well studied, the question of how anisotropic and localized access to information influence the network topology—not only the power-law exponent, but also other topological measures such as the clustering coefficient—has been studied less.

In this paper, we address this question by presenting and analyzing a network generation mechanism that features anisotropic and localized access to information. Although our model joins a large category of existing network copying models, it distinguishes itself from existing works by allowing both anisotropic and localized access to information. In Sec. II we present our model mechanism and derive the out-degree distribution for a simple version of it. We show explicitly how the resulting power-law exponent is related to the accessibility that nodes have to upstream and downstream information within the network. In Sec. III we study numerically how the various parameters in our model influence the network clustering coefficient. In Sec. IV we investigate the extent to which the predictions of our model are consistent with three typical networks from different real-world domains. The conclusion is given in Sec. V, where we summarize our main contributions, discuss the limitations of the present work, and comment on open questions.

II. MECHANISTIC MODEL

We now introduce our model and derive the analytical expression for the out-degree distribution for a simple version of it. We will use the concrete setting of a citation network in order to explain our mechanistic model and help make it intuitively easier to understand; however, we stress that this same mechanistic model can in principle be applied to any network that has similar properties. Imagine an author is exploring which citations to include in his or her upcoming paper and directly accesses (e.g., using a search engine) a highly relevant article that has previously been published (i.e., the article that is listed as most relevant given their search keywords) and then uses this to start a local exploration of articles that are related to it. We define this highly relevant published article (i.e., this network node) as a *center*, since it forms a center for the author's subsequent exploration; we define the published articles (nodes) that are cited by this center as its *parents*; and we define the published articles

*Corresponding author: zhenfeng.cao@polyu.edu.hk

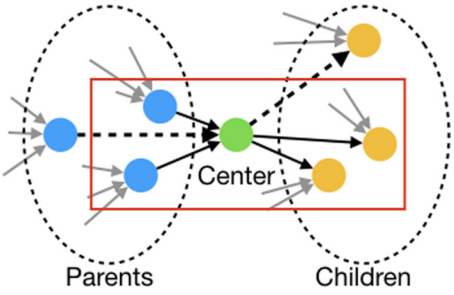

FIG. 1. The family of a center node consists of the nodes within the red solid rectangle, i.e., the family comprises the center (green), its two accessible parents (blue), and its two accessible children (yellow). The solid black arrows (edges) connect nodes whose information is accessible, while the dashed black arrows connect nodes whose information is inaccessible.

(nodes) that subsequently cite this center as its *children*. We use the convention that if node A (i.e., published paper A) is cited by node B (published paper B), A is the source node and B is the target node of a directed edge from node A to node B in the network. We now assume that the author does not (or cannot) access, and hence does not cite, the complete set of all parents and children of this center. We will refer to this situation as the author having limited *access* to the parents and children. We stress that it is irrelevant to the present work whether this limited access is because the author has finite time, finite resources, or both. We define a and b as the probabilities that a given parent and a child can be accessed, respectively. A center's *family* can hence be defined as a set of nodes comprising the center, its accessible parents, and its accessible children. This set is illustrated in Fig. 1, and it is obviously a small subset of the complete network of published papers.

Our mechanistic model's network generation mechanism then proceeds as follows:

(1) In the first time step, we initialize the network with a single published article (i.e., a single node). This initial condition could be made more realistic, for example, by initializing the network with a small number of nodes having random citations among them, but in the long time limit the results are insensitive to the precise choice of initial condition.

(2) At every subsequent time step, we assume a new article is published, i.e., a new node is added to the network, and that it cites a subset of the existing published articles (and hence forms links to existing nodes) in the following way:

(a) With a probability ρ, the new published article (i.e., new node) randomly selects a published article to be a center, from the set of previously published articles. By contrast with a probability $1-\rho$, it randomly selects a published article to be a center from the subset of previously accessed articles. This subset of previously accessed articles comprises the union of all families from the previous citing processes by the new article. It mimics the idea of an author drawing a highly relevant article from a list of published papers that he or she has accessed in the past and using this as a starting point to explore other related articles. For the case of the first citation by the new article, there are no previously accessed articles, and hence we choose the center randomly from the previously published articles (i.e., $\rho = 1$).

(b) The new published article (i.e., new node) then randomly selects a published article to cite from the family of this selected center. We generate this family by a random sampling process: each parent of the selected center has a probability a of being accessed (i.e., becoming a member of the family), while each child of the selected center has a probability b of being accessed (i.e., becoming a member of the family). This mimics the idea of an author drawing citations from all papers directly related to a center paper (i.e., the center paper itself, its citations, and the papers citing it).

We let this citing process be repeated on average m times by the newly published article, i.e., the in-degree of the new node and hence the number of references of the newly published article is m on average. For technical convenience when implementing the random selection of nodes, we use random sampling with replacement (i.e., allowing repeated citations). However, we have checked that our results are insensitive to whether we use sampling with or without replacement, as long as we run the simulation for a sufficiently long period.

We now derive the out-degree distribution of such a network when $\rho = 1$, i.e., a center is always selected randomly from the entire set of previously published articles. The $\rho = 1$ case is easier to study analytically than the $\rho < 1$ case, but it is sufficient for indicating how asymmetric access to information impacts the power-law exponent. In addition, we will show in a later section that the out-degree distribution is insensitive to ρ. Hence we focus our analytical study here on this simpler case $\rho = 1$. Let $p(q,s,t)$ denote the probability of an article published at time step s having q children when observed at step t. The probability density function (PDF) of the number of children (i.e., the out-degree, meaning the number of citations that an article receives) observed at time t can then be expressed as

$$P(q,t) = \frac{1}{t}\sum_{s=0}^{t} p(q,s,t). \tag{1}$$

Consider an article published at time s and observed at time step t to have degree q. At every instance of the on-average m independent citations, this article will be cited by an article published at $t+1$ as either (1) a center, (2) an accessible parent, or (3) an accessible child of a family whose center is randomly selected. For case 1, the probability that the article will be selected as the center will be $1/t$, and the probability that it will be further selected as the one to be cited among the $am + bq + 1$ family members is $1/(am + bq + 1)$. For case 2, the probability that it will be an accessible parent in a randomly selected family (or equivalently, that a randomly selected center will be its child and have access to it) is given by aq/t, and the probability that it will be further selected as the one to be cited among the $am + b\bar{q} + 1$ family members is $1/(am + b\bar{q} + 1)$. Similarly for case 3, the probability that it will be an accessible child in a randomly selected family (or equivalently, that a randomly selected center will be its parent and have access to it) is given by bm/t, and the probability that it will be further selected as the one to be cited among the $am + b\bar{q} + 1$ family members is $1/(am + b\bar{q} + 1)$.

This citing process is repeated on average m times. Hence the probability that this article (i.e., the one published at s and observed at time step t to have degree q) will be cited by an article published at $t+1$ is given by

$$A(q,t) = \frac{m}{t}\frac{1}{am+bq+1} + \frac{m}{t}\frac{aq+bm}{am+b\bar{q}+1}, \quad (2)$$

where the first term represents the probability that the article will be cited as a center, and the second term represents the probability that it will be cited as an accessible parent or child. $\bar{q}$ is the average q which equals m. If we assume $aq+bm$ is much greater than 1, the first term is then usually much smaller than the second one when q is $\gtrsim\bar{q}$. This is very likely to be true in real-world data sets. For example, the average number of citations given by $m \equiv \bar{q}$ for an article in a citation network, for hyperlinks of a web page on the Internet, and follows for a user in the online social network studied in this work are found to be around 10.5, 8.2, and 29.8, respectively. In addition, we note that a and b cannot be both very small: otherwise the network would approach a random one, which is inconsistent with the observation that the out-degrees of these real-world networks exhibit a fat-tail distribution. For simplicity, we make the approximation $q \equiv \bar{q}$ in the first term, which then yields

$$A(q,t) \approx \frac{m}{t}\frac{aq+bm+1}{(a+b)m+1}. \quad (3)$$

The evolution of $p(q,s,t)$ is given by

$$\begin{aligned} &p(q,s,t+1) - p(q,s,t) \\ &\quad = -A(q,t)p(q,s,t) + A(q-1,t)p(q-1,s,t), \end{aligned} \quad (4)$$

together with the initial condition $p(q,s,s) = \delta_{q,0}$. Summing over s from 0 to t, and considering the long-time limit in which we can transition to a continuous-time approximation, we obtain

$$\begin{aligned} \frac{\partial P(q,t)}{\partial t} = &-\frac{P(q,t)}{t} - A(q,t)P(q,t) \\ &+ A(q-1,t)P(q-1,t) + \delta_{q,s}, \end{aligned} \quad (5)$$

where $P(q,t)$ is given by Eq. (1). Consequently, the solution for the stable state [i.e., $\partial P(q,t)/\partial t = 0$ when $t \to \infty$] is given by

$$P(q) = C^{-1}B\left(2+\frac{b}{a}+\frac{1}{am}, \frac{bm+1}{a}+q\right). \quad (6)$$

Here B is the beta function, and C is the normalization constant given by

$$\begin{aligned} C = {}&{}_2\tilde{F}_1\left(1, \frac{bm+1}{a}; \frac{2a+b+bm+1}{a}+\frac{1}{am}; 1\right) \\ &\times B\left(2+\frac{b}{a}+\frac{1}{am}, \frac{bm+1}{a}\right), \end{aligned} \quad (7)$$

where ${}_2\tilde{F}_1$ is the hypergeometric function. For large q, this produces a power-law distribution [i.e., $P(q) \sim q^{-\alpha}$] with the power-law exponent

$$\alpha = 2+\frac{b}{a}+\frac{1}{am}. \quad (8)$$

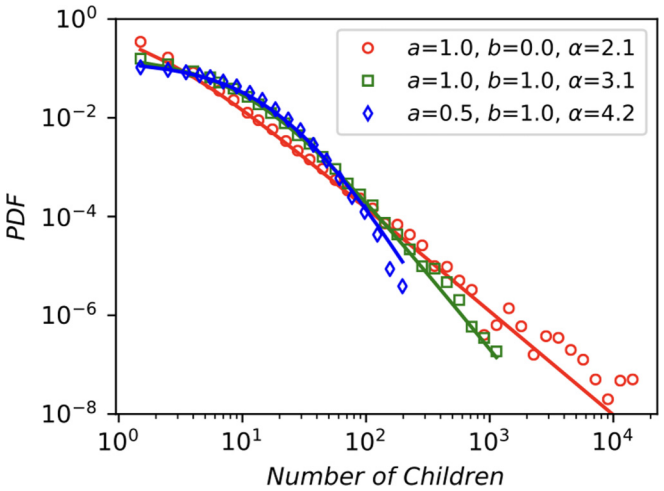

FIG. 2. Comparison between the analytical results (solid lines) derived from Eq. (6) for our mechanistic model as described in Sec. II, and the simulation results (symbols are same color as respective line) for different parameter values. The quantity shown is the out-degree (i.e., number of children) distribution of the resulting network. The agreement is good. As explained in this paper, a (b) is the probability that a parent (child) is accessible, and m is the average member of parents that a center has. For all cases here, $m = 10$ and $\rho = 1$.

This tells us that increasing b will make α increase linearly when a is constant. By decreasing a with b constant, α will increase rapidly and can reach any value above 2. We note that in all the above, $0 < a \leqslant 1$ and $0 \leqslant b \leqslant 1$.

Figure 2 shows the good agreement obtained between our analytical result in Eq. (6) and the result of numerical simulations for our mechanistic network-generating model. We explicitly show the out-degree (i.e., number of children) distributions of the resulting networks. We ran each simulation until it generated a large number of nodes (10^5). When plotting the analytical results, we used the same values of the parameters a, b, and m as the ones used in the corresponding simulations. The simulations shown in Fig. 2 have a constant in-degree of m for all nodes; however, we have checked that our findings are unchanged when we allow the in-degree distributions to vary. This makes sense since our model indicates that the out-degree distribution is driven by the average number of in-degrees but is insensitive to the exact form of the in-degree distribution. This is confirmed in Fig. 3, which compares the out-degree distributions of the simulations for a variety of in-degree distributions: all the simulations result in very similar out-degree distributions. All other parameter values in the analytical result and simulations are kept to be the same, apart from the in-degree distribution.

As further verification of our model's predictions [i.e., Eq. (8)], we ran simulations for different combinations of parameters. The simulation results are shown in Table I and demonstrate good agreement with our analytical results. By looking at the errors, we find that the cases with $a < b$ show the largest errors. This is understandable since the power-law exponents (α) in those cases will be large, and hence there will be fewer data points in the tails of the distributions

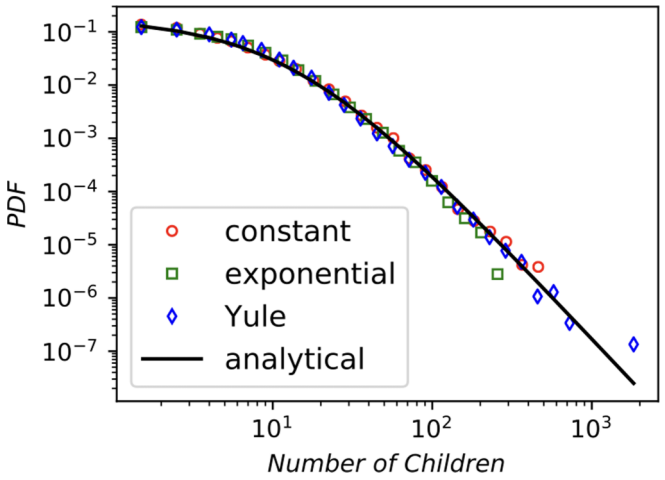

FIG. 3. Simulations showing that the out-degree distribution is insensitive to the in-degree distribution. The three cases shown as examples are where m is a constant, or it follows an exponential distribution, or it follows a Yule-Simon distribution. All the other parameter values are held at the same value. The cases when $a = b = 0.5$, $m = 10$, and $\rho = 1$ are shown, but our further investigations have shown that this main result is valid for other combinations of parameters. In addition, the analytical result (the black solid line) is shown for comparison, with the same set of parameter values.

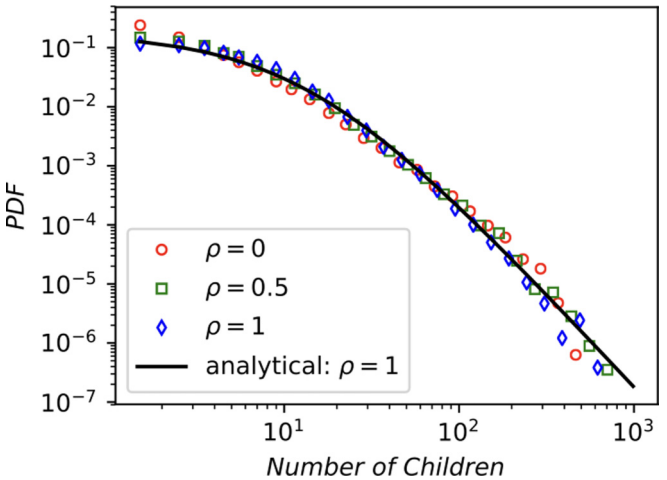

FIG. 4. Simulations showing that the out-degree distribution is insensitive to ρ. In all cases, we ensured that all other parameters are kept the same ($a = b = 0.5$) and that the in-degree follows a Yule-Simon distribution with an average degree number $m = 10$. Other distributions could be used, but we find that the main conclusions are unchanged. In addition, the analytical result (black solid line) for $\rho = 1$ is shown for comparison.

as compared to other cases; thus this will result in poorer statistics as observed.

Figure 4 shows the results from simulations in which we vary the value of ρ, which is a measure of the nonlocality of a new node's access to information. These show that the out-degree distribution is similarly insensitive to ρ. This also makes sense because although a new node's access to information is localized when $\rho < 1$, the out-degree distribution is a measure that is averaged over all nodes, and we have assumed that there is no correlation among the nodes (except for their temporal order). Hence the impact of this localization effect on the out-degree distribution has been averaged out in the long run.

III. CLUSTERING COEFFICIENTS

Our finding that the network out-degree is insensitive to both the in-degree distribution and ρ confirms the well-known fact that although the out-degree distribution is a relevant measure of a network, it is insufficient for inferring the network's detailed architecture or the process that generated it. Hence to gain a better understanding of the network, we need to look at some other topological measure such as the clustering coefficient (CC). In this section, we study numerically how the parameters in our model influence the global CC of the network. It seems clear that the form of the in-degree distribution should have some influence on the CC; however, in this paper we are more interested in exploring how anisotropy and locality influence the CC. Hence we will set the in-degree distribution of all our simulations to be the Yule-Simon distribution with average degree number 5. Though other choices of distribution and of the average degree number can of course be made, all the cases that we checked have given the same main results. In each run, we allow a sufficient amount of time to pass so that the network grows to be large ($>10^4$ nodes).

We find that the anisotropy, as described by the parameters a and b, has only a weak influence on the CC

TABLE I. Power-law exponents for different combinations of a, b, and m: analytical (α) vs simulation (α^*) results. Here α is calculated according to Eq. (8). α^* is estimated using the MLE scheme [see Eq. (9)] and is the mean value since we repeated the simulation 20 times for each case. The percentage error $\%err = |\alpha - \alpha^*|/\alpha^* \times 100\%$. The standard deviations are shown in parentheses following their corresponding mean values. When doing the MLE, we fit Eq. (6) to the data from simulations to estimate a, b but with m treated as a known parameter to reduce overfitting, and then we used the estimated a and b to calculate α^* according to Eq. (8). For all simulations, the total number of steps in each run is 10^6, and ρ is set to be 1.

a	1.0	1.0	1.0	0.75	1.0	1.0	1.0	0.75
b	0.0	0.5	1.0	1.0	0.0	0.5	1.0	1.0
m	8	8	8	8	8	16	16	16
α	2.125	2.625	3.125	3.500	2.063	2.563	3.063	3.417
α^*	2.128(2)	2.568(10)	3.060(29)	3.612(11)	2.073(1)	2.447(8)	2.962(16)	3.573(4)
$\%err$	0.14(9)%	2.26(40)%	2.12(97)%	3.10(30)%	0.48(5)%	4.74(34)%	3.41(56)%	4.37(11)%

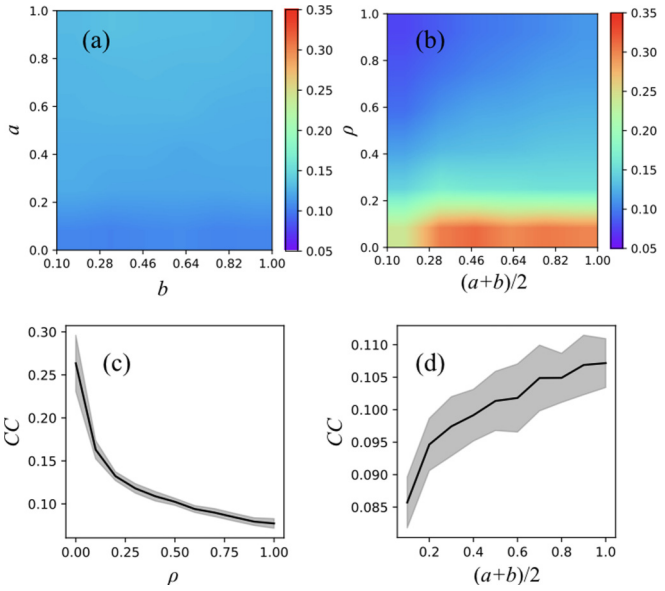

FIG. 5. (a) A heat map showing how the clustering coefficient (CC) generated by our mechanistic model described in Sec. II is influenced by a and b when $\rho = 0.5$ and $m = 5$. (b) A heat map showing how CC is influenced by ρ and $(a+b)/2$ when $a = b$ and $m = 5$. The colors of the heat maps show the values of CC. (c) CC vs ρ when $a = b = 0.5$ and $m = 5$. (d) CC vs $(a+b)/2$ when $\rho = 0.5$ and $m = 5$. The gray bands represent a one-sigma (i.e., one standard deviation) error spread.

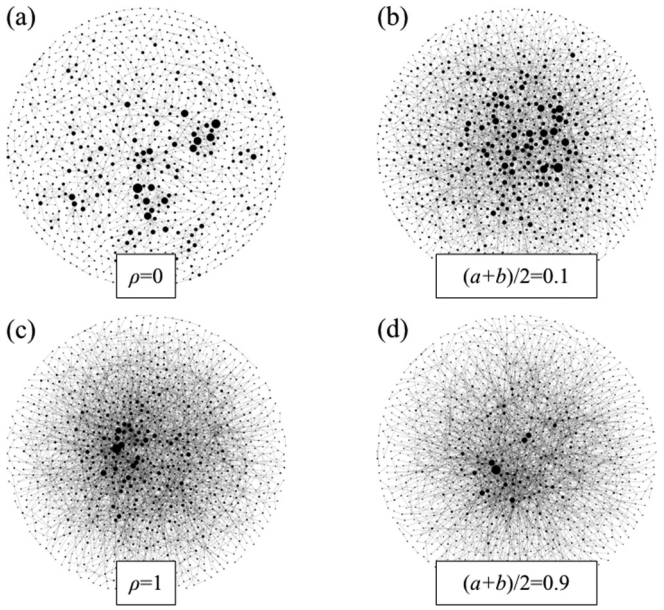

FIG. 6. Network graphs from simulations for different ρ and $(a+b)/2$ values (see the labels on the graphs). In each simulation, the number of nodes is 1500. Only nodes with at least one edge have been shown. In (a) and (c), $a = b = 0.5$ and $m = 2$. In (b) and (d), $a = b$, $m = 2$, and $\rho = 0.5$. The size of a node is proportional to its out-degree.

[see Fig. 5(a)] but that the locality has significant impact on the CC [see Figs. 5(b) and 5(c)]. The CC increases rapidly with the decrease of ρ, indicating that localized access to information plays a pivotal role in the formation of clusters. This is also consistent with the case studies in Figs. 6(a) and 6(c). The CCs are calculated using the function *Graph.transitivity_undirected()* in the open-source Python package *igraph* [34], which follows the definition of transitivity given in Ref. [35].

We also studied the influence of the average (and total) accessibility to upstream and downstream information by looking at the relationship between $(a+b)/2$ and the CC. We found that the average accessibility is positively correlated to the CC, though the influence of it is much weaker than ρ. This is because reduced accessibility would enhance randomness in access to information when the in-degree (i.e., number of parents) of a new node is given. To gain a better understanding of what is happening, Figs. 6(b) and 6(d) show particular cases from the simulation. We can see that the case of Fig. 6(d) looks slightly more structured (i.e., less random) than Fig. 6(b).

IV. APPLICATION TO REAL-WORLD NETWORKS

A. Anisotropy and power-law exponent

In previous sections, we found our model can produce a broad spectrum of power-law exponents and clustering coefficients. It is known that anisotropic and localized access to information can be relatively common scenarios in real-world networks. Since the main focus of this work is not to fully reproduce the topology of real-world networks but instead to show theoretically how anisotropic and localized access to information influence the topology, we now examine the extent to which our model is able to reproduce features of the topology of real-world networks.

To realize this comparison, which we stress has the modest goal of exploring model consistency with real-world networks, not proof of their generating mechanisms, we compare the predictions of our model to three real-world networks from different domains: specifically, the citation network of the American Physical Society, which is available at https://www.journals.aps.org/datasets and will be referred to as "APS"; the hyperlink network of https://www.stanford.edu [36], which we will refer to as "Stanford"; and the relationship network of Twitter [37], which we will refer to as "Twitter." These networks have some common features: (1) nodes and edges are gradually added to the network (i.e., they are formed through some growing process), (2) the mean value of their average in-degree is finite, partly due to finite time and energy available to a new node for adding edges, and (3) their out-degrees exhibit a fat-tail distribution, though we note that rigorous power-law tests fail for all of them. Specifically, applying the goodness-of-fit test to the tail of them [2] yields p values of < 0.01 for all three cases. This is not a surprise since it is known that only a few real-world networks pass a rigorous power-law test in which the tail is sufficiently perfect [38,39].

We now investigate the extent to which the analytical result [Eq. (6)] from our mechanistic model, can reproduce the observed out-degree distributions. To achieve this, we applied maximum likelihood estimation (MLE) to estimate the parameters in Eq. (6). To minimize overfitting, we determined m directly from the data by measuring the average number of parents of all the nodes. There are then two parameters, a

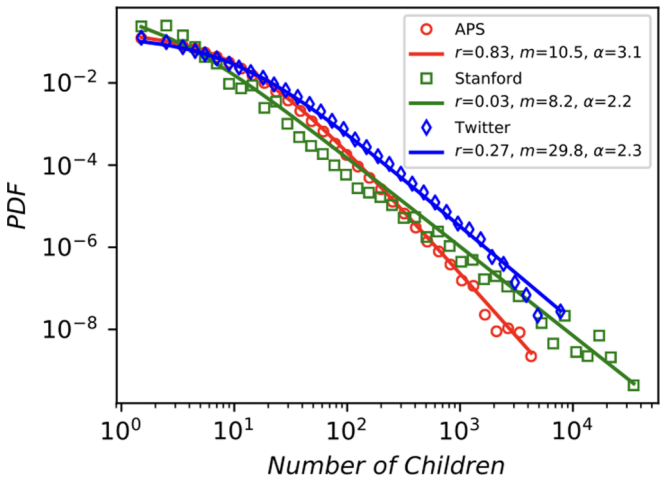

FIG. 7. Best fit (solid lines) for the three empirical distributions (symbols of the same color as lines) using the analytical expression from our mechanistic model [Eq. (6)]. The legend shows the values of m directly measured from the data set, the optimal r, and the corresponding α value [see Eq. (8)] of the best fits.

and b, to be determined. Suppose there are N nodes in total and x_i is the number of children of node i in the observation. The average log likelihood is then given by

$$\hat{\ell}(a,b;x) = \frac{1}{N}\sum_{i=1}^{N} \log(P(x_i;a,b)). \tag{9}$$

where $P(x_i;a,b) = P(q = x_i)$ is given in Eq. (6) but with a and b adjustable and m fixed to its empirical values. The problem then becomes one of finding the globally optimal a and b values that maximize $\hat{\ell}(a,b;x)$. This optimization can be done using conventional multivariable global optimization algorithms such as *basinhopping* in the Python package *scipy*. We find that although overfitting still exists (the errors for the estimated a and b are ~35%), the ratio of b to a with $r \equiv b/a$ converges well to a globally optimal value. Specifically, we find that $r = 0.83 \pm 0.06$, 0.03 ± 0.03, and 0.27 ± 0.05 for the APS, Stanford, and Twitter data sets respectively. Even though the fluctuations and hence variance of a and b could both be very large, their ratio remains largely unchanged, and hence the variance of the ratio would be small. This is possible if a and b are positively correlated, and hence a and b essentially increase or decrease together. The standard deviation of r for each data set is obtained by repeating this optimization procedure 100 times. Figure 7 shows how these fits look of the model compared to the real-world networks, with all repeated runs resulting in very similar results. It is understandable that these best fits also fail the rigorous Kolmogorov-Smirnov (K-S) test [p value < 0.01 are found for all three cases for our proposed model, i.e., Eq. (6)] since real-world networks are not in general strict power laws [39].

Despite the limitation that the errors in the estimated parameters are not small, the robustness and simplicity of the analytical result [Eq. (6)] suggest that it could be useful for evaluating the levels of asymmetric access to upstream and downstream information in these networks. In the APS network, the best-fit value $r = 0.83$ suggests that information about the citing and cited articles of a center is almost equally accessible to an author. This is consistent with everyday experience; for example, search engines such as Google Scholar make it convenient to access both the citing and the cited articles. With respect to the Stanford network, which is a hyperlink network, the best-fit value $r = 0.03$ is consistent with the notion that a user knows only what page a selected center page cites and has little information about which pages have cited this page when he or she is deciding which pages to cite as part of new page creation. With respect to the Twitter network, $r = 0.27$ has the reasonable implication that when adding a new follow, users favor the follows of the current follows over their followers. In future work, we will try to elucidate this further and compare across different types of social media platform.

B. Locality and clustering coefficient

We showed that our analytical result [Eq. (6)] explains well the out-degree distributions from the perspective of anisotropic access to information. However, we also need to check other topological details such as the CC, since these may be quite different for a given out-degree distribution. We will perform this check in an albeit limited way in the present study by adopting the in-degree distributions directly from the data sets. Since the out-degree distribution is insensitive to ρ and $(a+b)/2$ but very sensitive to the anisotropy measure r, by fixing r we ensure that our simulations reproduce well the out-degree distributions. Then, by adjusting ρ and $(a+b)/2$ with r fixed, we can expect to reproduce well the CCs and the out-degree distributions. Since $(a+b)/2$ has much weaker influence on CC than ρ, we fix it to be 1 for all three networks and adjust only ρ. We could set $(a+b)/2$ to other values, keeping it away from zero to avoid a pure random process, but its impact is minor compared to that of changing ρ [recall Fig. 5(b)].

The average CCs of the APS, Stanford, and Twitter networks are found to be 0.08 ± 0.02, 0.01 ± 0.01, and 0.18 ± 0.02, respectively. Since the data sets are very large, we sampled them 50 times when calculating the average CCs and their one-sigma errors, with the size of each sample set to 10^4 nodes. We found that by setting ρ near 0, 1, and 0 for the APS, Stanford, and Twitter networks, respectively, we are able to reproduce the observed CCs within their one-sigma errors. For the purposes of this paper, we simply choose these specific ρ values so that they match the results from the respective data sets reasonably well, without carrying out any rigorous optimal fitting, or deriving them or inferring them from some first-principles approach. This is because our point in this paper is not to explain the mechanisms that generate these data sets in some unique microscopic way, but rather to show how these two data sets can be interpreted within the same single model of growth. The value $\rho = 0$ for APS paper citations suggests that the mechanism of citing APS papers tends toward selecting an article from all previously accessed articles (i.e., locality plays an important role). Similarly, $\rho = 0$ for the Twitter network indicates a user tends to add new friends or follows by exploring the neighbors of the already accessed ones. By contrast, the value of 1 for the Stanford network suggests that the locality is less relevant for a

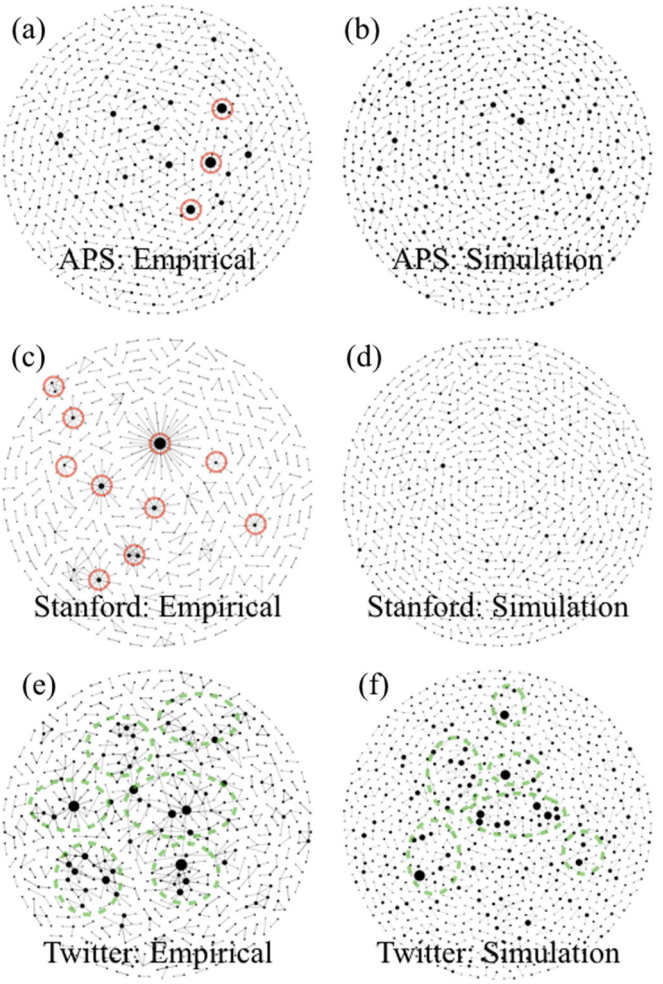

FIG. 8. Empirical results (left) for subsets of the three real-world networks versus simulations of our mechanistic model simulation (right). The size of a node is proportional to its out-degree. When making the graphs, we sampled 5000 nodes for the APS and Stanford networks and 2000 nodes for the Twitter network. Only nodes with at least one edge are included. The nodes at the center of the red circles in (a) and (c) are outliers that cannot be reproduced by our mechanistic model. The green dashed circles in (e) and (f) show the approximate positions of the emerging clusters.

web-page creator's decision with regards to which web pages to cite.

Figure 8 compares sample subnetworks from these simulations to those from the empirical data sets. From Fig. 8, it can be seen that our model successfully reproduces a crude version of the real network structures. For example, clustering structures similar to the observational ones emerge from the simulation of the Twitter network. However, there are outliers in the APS and Stanford data sets that cannot be reproduced by our current model, specifically, the large nodes centered at the red circles in Figs. 8(a) and 8(c). Those nodes are rare but have unusually high out-degrees. This indicates that there could be some mechanisms or external coordinating factors that are missing from our current model. It also suggests that global measures such as the out-degree distribution and the global CC tend to overlook such rare, but probably influential, individuals with unusually high out-degrees. The existence of these outliers is also possibly one reason why the tails of the out-degree distributions do not pass a rigorous power-law test.

V. CONCLUSION

We have proposed a network generation mechanism with a focus on investigating the influence of anisotropic and localized access to information on the topology of a power-law network. Our model shows that anisotropy impacts the power-law exponent significantly, but has only a weak influence on the clustering coefficient. By contrast, locality influences the clustering coefficient effectively but has only weak influence on the power-law exponent. Our proposed model is capable of generating networks spanning a broad spectrum of power-law exponents and clustering coefficients, and hence could feed into the debate about the nature of real-world networks.

As expected in any comparison between minimal models and complex real-world systems, there are several limitations of this work. First, we have been unable to uniquely determine a, b, and ρ by matching the out-degree distributions and the clustering coefficient. This suggests that additional network quantities should be compared. Second, there are outliers (nodes with unusually high out-degrees) that cannot be reproduced by our current model, which in turn suggests that there may be additional mechanisms or real-world external factors that need to be included in an improved version of our mechanistic model. For example, we assumed that all nodes are equally attractive to a new node; however, in real-world networks, such attractiveness of the nodes could be highly inhomogeneous. In particular, a high-quality article could be much more influential than a low-quality one in the APS network. Likewise, the home page of a department or a school is more likely to cited by a newly created page related to research or teaching; and some people through their profession (e.g., politicians) naturally have more followers than others in the Twitter network. We hope that this work will stimulate future discussions to address these open issues.

ACKNOWLEDGMENTS

N.F.J. acknowledges funding from the National Science Foundation Grant CNS 1522693 and Air Force (AFOSR) Grant FA9550-16-1-0247. The views and conclusions contained herein are solely those of the authors and do not represent official policies or endorsements by any of the entities named in this paper.

[1] M. E. J. Newman, Contemp. Phys. **46**, 323 (2005).
[2] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Rev. **51**, 661 (2009).
[3] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[4] H. Jeong, Z. Néda, and A.-L. Barabási, Europhys. Lett. **61**, 567 (2003).

[5] M. E. J. Newman, Phys. Rev. E **64**, 025102(R) (2001).
[6] S. Lehmann, A. Jackson, and B. Lautrup, Europhys. Lett. **69**, 298 (2005).
[7] S. Milojević, J. Am. Soc. Inf. Sci. Tech. **61**, 1410 (2010).
[8] T. Kuhn, M. Perc, and D. Helbing, Phys. Rev. X **4**, 041036 (2014).
[9] L. A. Adamic and B. A. Huberman, Science **287**, 2115 (2000).
[10] M. Faloutsos, P. Faloutsos, and C. Faloutsos, in *ACM SIGCOMM Computer Communication Review*, Vol. 29 (ACM, New York, 1999), pp. 251–262.
[11] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[12] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, Phys. Rev. E **74**, 036116 (2006).
[13] P.-P. Zhang, K. Chen, Y. He, T. Zhou, B.-B. Su, Y. Jin, H. Chang, Y.-P. Zhou, L.-C. Sun, B.-H. Wang *et al.*, Physica A (Amsterdam) **360**, 599 (2006).
[14] A.-L. Barabási *et al.*, Science **325**, 412 (2009).
[15] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement* (ACM, New York, 2007), pp. 29–42.
[16] F. Fu, L. Liu, and L. Wang, Physica A (Amsterdam) **387**, 675 (2008).
[17] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Åberg, Nature (London) **411**, 907 (2001).
[18] J. H. Jones and M. S. Handcock, Proc. R. Soc. London B **270**, 1123 (2003).
[19] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, in *Proceedings of the First Workshop on Online Social Networks* (ACM, New York, 2008), pp. 25–30.
[20] B. Ribeiro, W. Gauvin, B. Liu, and D. Towsley, in *INFOCOM IEEE Conference Computer Communications Workshops, 2010* (IEEE, San Diego, 2010), pp. 1–6.
[21] L. A. Adamic and E. Adar, Social Networks **25**, 211 (2003).
[22] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao, in *Proceedings of the 4th ACM European Conference on Computer Systems* (ACM, New York, 2009), pp. 205–218.
[23] M. Kaiser and C. C. Hilgetag, Biol. Cybern. **90**, 311 (2004).
[24] R. V. Solé, R. Pastor-Satorras, E. Smith, and T. B. Kepler, Adv. Complex Syst. **5**, 43 (2002).
[25] A.-L. Barabasi and Z. N. Oltvai, Nat. Rev. Genet. **5**, 101 (2004).
[26] E. Eisenberg and E. Y. Levanon, Phys. Rev. Lett. **91**, 138701 (2003).
[27] K. Choromański, M. Matuszak, and J. Miękisz, J. Stat. Phys. **151**, 1175 (2013).
[28] F. Pammolli, D. Fu, S. V. Buldyrev, M. Riccaboni, K. Matia, K. Yamasaki, and H. E. Stanley, Eur. Phys. J. B **57**, 127 (2007).
[29] C. Roth, in *ISWC 4th International Semantic Web Conference, Workshop on Semantic Network Analysis* Vol. 171 (CEUR-WS, Galway, 2005), pp. 1613–0073.
[30] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, arXiv:cond-mat/0004434.
[31] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani, Complexus **1**, 38 (2003).
[32] A. Vázquez, arXiv:cond-mat/0006132.
[33] P. L. Krapivsky and S. Redner, Phys. Rev. E **71**, 036118 (2005).
[34] G. Csardi and T. Nepusz, InterJournal Complex Systems **1695**, 1 (2006).
[35] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Phys. Rep. **424**, 175 (2006).
[36] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, Internet Math. **6**, 29 (2009).
[37] J. Leskovec and J. J. Mcauley, in *Advances in Neural Information Processing Systems* (NIPS, Lake Tahoe, 2012), pp. 539–547.
[38] A. D. Broido and A. Clauset, arXiv:1801.03400.
[39] B. Albert-László, Love is all you need (Barabási Lab, 2018).