Modeling and analysis of affiliation networks with preferential attachment and subsumption

Alexey Nikolaev[1] and Saad Mneimneh[1,2]
[1]*Department of Computer Science, The Graduate Center of CUNY, 365 5th Avenue, New York, New York 10016, USA*
[2]*Department of Computer Science, Hunter College of CUNY, 695 Park Avenue, New York, New York 10065, USA*

(Received 13 April 2022; revised 5 April 2023; accepted 19 May 2023; published 31 July 2023)

Preferential attachment describes a variety of graph-based models in which a network grows incrementally via the sequential addition of new nodes and edges, and where existing nodes acquire new neighbors at a rate proportional to their degree. Some networks, however, are better described as groups of nodes rather than a set of pairwise connections. These groups are called affiliations, and the corresponding networks affiliation networks. When viewed as graphs, affiliation networks do not necessarily exhibit the power law distribution of node degrees that is typically associated with preferential attachment. We propose a preferential attachment mechanism for affiliation networks that highlights the power law characteristic of these networks when presented as hypergraphs and simplicial complexes. The two representations capture affiliations in similar ways, but the latter offers an intrinsic feature of the model called subsumption, where an affiliation cannot be a subset of another. Our model of preferential attachment has interesting features, both algorithmic and analytic, including implicit preferential attachment (node sampling does not require knowledge of node degrees), a locality property where the neighbors of a newly added node are also neighbors, the emergence of a power law distribution of degrees (defined in hypergraphs and simplicial complexes rather than at a graph level), implicit deletion of affiliations (through subsumption in the case of simplicial complexes), and to some extent a control over the affiliation size distribution. By varying the parameters of the model, the generated affiliation networks can resemble different types of real-world examples, so the framework also serves as a synthetic generation algorithm for simulation and experimental studies.

DOI: 10.1103/PhysRevE.108.014310

I. INTRODUCTION

The term *preferential attachment* was introduced in [1] and is now considered an umbrella mechanism to grow graph-based networks by sequentially adding new nodes and edges in such a way that existing nodes acquire their new neighbors at a rate proportional to their degree. Examples of such networks include social, biological, and information networks, where preferential attachment leads to a power law distribution in node degrees, i.e., the probability of a node with degree k satisfies $P(k) \propto k^{-\gamma}$. Networks with such degree distributions are often called scale free because $P(ak) \propto k^{-\gamma}$ persists for any scaling factor a.

While the notion of scale-free networks became very influential after it was popularized by Barabási and Albert, the idea was not entirely new: The preferential attachment mechanism (under different names) and the power law in the tail of the degree distribution were studied earlier in several settings [2–4], including networks, leading to what is now known as the Yule-Simon distribution $P(k) = \rho\Gamma(\rho+1)\Gamma(k)/\Gamma(k+\rho+1)$, where $P(k) \propto k^{-(\rho+1)}$ for large k.

Despite the success of preferential attachment models in capturing power laws and other characteristics of real-world networks [5–9], as opposed to, say, the random graph model proposed by Erdös and Rényi [10,11] and Gilbert [12], not every network can be explained by a straightforward preferential attachment growth. For instance, friendship networks [13] and coauthorship networks [14] are two such examples. However, as shown below, viewing these networks as groups of nodes rather then a set of pairwise connections offers better insight into the mechanisms underlying their growth.

Figure 1(a) shows the degree distribution of a PNAS authorship network [13] modeled as a graph, where two authors are connected by an edge if they have some paper in common. Figure 1(b) shows the degree distribution of the same network modeled as a hypergraph, where each paper defines a hyperedge on the set of its authors (and degree is the number of hyperedges a node belongs to). The degree distribution in Fig. 1(a) is described in [13] as one that exhibits three regimes (generalized Poisson, a crossover region, and a power law). On the other hand, the distribution in Fig. 1(b) can be completely explained by a power law.

Figure 2 shows similar distributions for the degrees in a Facebook friendship ego network [15], modeled as a graph in Fig. 2(a) with two people connected by an edge if they are friends, and as a clique simplicial complex in Fig. 2(b), where a clique simplicial complex consists of all the maximal cliques of the graph as hyperedges, now called facets in simplicial complexes, and degree becomes the number of facets a node belongs to. As in the previous example, the degrees in the simplicial complex exhibit the power law. Figure 3 replicates Fig. 2 with logarithmic binning.

Networks where nodes are joined not by pairwise connections but as groups of multiple nodes are called *affiliation* networks. Affiliation networks can be modeled in several ways, including the use of bipartite graphs, hypergraphs, and simplicial complexes. In both examples shown above, the degree distribution in the affiliation network is easily cap-

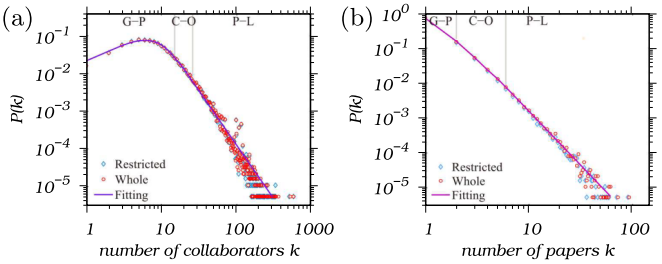

FIG. 1. The degree distribution in a coauthorship network (copied from [13]). (a) $P(k)$ is the proportion of the authors with k collaborators (i.e.. graph model). (b) $P(k)$ is the proportion of the authors with k papers (i.e., hypergraph model).

tured by a power law, suggesting first the existence of an underlying preferential attachment mechanism, and second that this mechanism is operating at the affiliation network level and not at the graph level. We present and analyze a preferential attachment model for affiliation networks when seen as hypergraphs and simplicial complexes, and hence provide an understanding of how underlying mechanisms contribute to the incremental growth of networks that seemingly do not comply (when modeled as graphs) with preferential attachment and power law distributions.

II. AFFILIATION NETWORKS AS HYPERGRAPHS AND SIMPLICIAL COMPLEXES

An affiliation network (also known as two-mode network, membership network, or hypernetwork) is a network that describes the dual relation between nodes and their *affiliations*. In the most general terms, it is given by a set of nodes V, a set of affiliations S, and a relation $R \subseteq V \times S$, that describes which nodes belong to which affiliations. For instance, a collaborative network of overlapping teams can be represented by a set of persons V and a set of teams S, where a pair $(v, s) \in R$ means that person v works in team s. We assume that every node must belong to at least one affiliation, and every affiliation must contain at least one node. Given an affiliation network, the corresponding skeleton graph is the simple graph such that two nodes are connected by an edge if they belong to some common affiliation.

Affiliation networks are used in the modeling of several phenomena, including interlocking directorates [16–18], where a member of one company's board of directors also serves within another company's management; citation networks [13], where papers serve as affiliations with overlapping authors; and co-occurrence studies of biological species [19,20], where each species is a node and each geographic site is an affiliation.

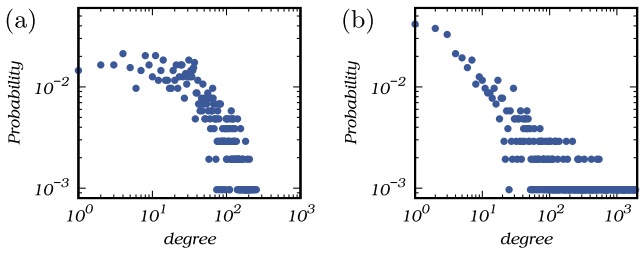

FIG. 2. The degree distribution in a Facebook friendship ego network [14] modeled as (a) a graph and (b) as a clique simplicial complex.

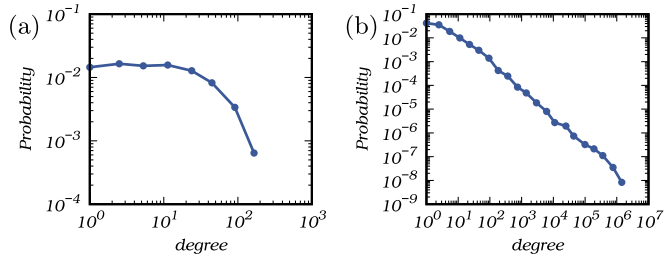

FIG. 3. The same degree distributions of Fig. 2 binned with a constant bin size on the logarithmic scale.

We consider two growth models of affiliation networks: hypergraphs and simplicial complexes. In both models, the network grows by adding new affiliations that contain existing nodes and newly created nodes.

A. Hypergraphs

Any affiliation network, when not requiring special structural properties, can be modeled as a hypergraph (V, E), where V is a set of nodes and E is a collection of nonempty subsets of V (the hyperedges). Each affiliation becomes a hyperedge. Our network growth algorithm (described in Sec. IV) guarantees that hyperedges are unique by making each added affiliation contain at least one newly created node. The degree of a node is defined as the number of hyperedges it belongs to, so it is nondecreasing as the network grows (this does not hold in the case of simplicial complexes described below).

B. Simplicial complex: Affiliation networks with subsumption

An abstract simplicial complex is a collection Δ of nonempty sets such that if $f \in \Delta$, then every nonempty subset of f is in Δ as well. The sets of Δ are called the faces of the simplicial complex. The dimension of a face f is $\dim(f) = |f| - 1$. The elements of the faces are the nodes, so the set of all nodes of a simplicial complex Δ is $V = \cup_{f\in\Delta}$. The maximal faces (those that are not subsets of other faces) are called facets, and we use F to denote the set of facets.

Simplicial complexes have been used in many applications, ranging from biology [21–23], to communication networks [24–29], to scientific collaboration [30–35], to epidemic spreading and opinion propagation [36–38].

Any simplicial complex can be fully described by the set of its facets. While the faces of the complex can be generated if needed, their number grows exponentially in the dimensions of the facets. Therefore, when modeling an affiliation network as a simplicial complex [such as the example in Fig. 2(b)], it is reasonable to view the facets as the affiliations of the network, and define the *degree* of a node as the number of facets it belongs to.

This approach in defining affiliations leads to a property that distinguishes simplicial complexes from hypergraphs, which we call *subsumption*, following a terminology used in [39–43] for instance. To illustrate subsumption, observe that

in a simplicial complex network, no affiliation (i.e., facet) can be a subset of another affiliation. Therefore, as a network grows, an affiliation must be removed from the network if it ever becomes a subset of another (existing or newly added) affiliation. Consider a simplicial complex network that consists of two facets: $\{A, B\}$ and $\{B, C, D\}$. Adding $\{A, B, E\}$ to the network would lead to the subsumption of $\{A, B\}$. On the other hand, adding $\{B, C\}$ would not change it because $\{B, C\}$ would be subsumed by the existing affiliation $\{B, C, D\}$. To simplify analysis, our network growth algorithm (described in Sec. IV) avoids this latter scenario of subsumption by making each added affiliation contain at least one newly created node. We also use the term *absorption* to describe the former scenario of subsumption. Unlike the case of a hypergraph, the degree of an existing node can decrease upon the addition of a facet when absorption occurs: for instance, when two or more of the node's facets are absorbed. However, this scenario is practically eliminated given the no overlap assumption explained in Sec. IV; therefore, a node may either gain a degree or maintain its current degree.

III. PRIOR WORK AND OUR CONTRIBUTION

Some efforts of studying preferential attachment in affiliation networks exist. The work in [44,45] grows simplicial complexes with interesting structural properties by repeatedly adding facets of size three (facets are restricted to only triangles). The work in [46] generates hypergraphs that exhibit a symmetric property, where the same power law distribution governs both degrees and hyperedge sizes. The work in [39] on simplicial complexes produces facets with a target size distribution, but relies on an explicit selection of nodes based on degrees (which is a computational bottleneck) and eliminates the chance of subsumption when networks become large enough (large simplicial complexes behave as hypergraphs).

Our model for generating affiliation networks exhibits, by design, a set of desirable features on both the algorithmic and the analytic levels:

(i) implicit preferential attachment, where nodes are not explicitly chosen based on degree;

(ii) locality parameter (the neighbors of a newly added node are also neighbors);

(iii) power law distribution of degrees;

(iv) absorption: implicit deletion of existing affiliations by virtue of the subsumption property in simplicial complexes (see Sec. II); this feature does not die out in large networks;

(v) controlled affiliation size by specifying a distribution for the number of newly added nodes.

The above framework offers a convenient tool for efficiently generating synthetic networks that resemble real-world networks, which is useful in simulation and experimental studies. For instance, by controlling several parameters of the generation algorithm, different properties of networks can be achieved including average distances, the clustering coefficients [47,48], and assortativity [49].

IV. OUR MODEL AND GENERATION ALGORITHM

Below we present the algorithm for growing an affiliation network described for both hypergraphs and simplicial complexes, where affiliation stands for either hyperedge or facet accordingly. Given parameters $0 \leqslant \alpha < 1$, $\ell \in \mathbb{N} = \{1, 2, 3, \ldots\}$, and a fixed probability distribution P_c for $c \in \mathbb{N}$ (with a finite mean $E[c]$), the algorithm is defined as follows:

Grow_Network(α, ℓ, P_c)
  start with one affiliation of some size
  while some stopping condition is not met
    $f \leftarrow$New_affiliation(α, ℓ, P_c)
    add f to the network
    (simplicial complex) delete facets absorbed by f

New_Affiliation(α, ℓ, P_c)
  $f \leftarrow$ empty affiliation
  sample ℓ existing affiliations $f_1, \ldots, f_\ell$ uniformly at random with replacement
  (variant 1) $A \leftarrow f_1 \cup \ldots \cup f_\ell$ (**union**)
  (variant 2) $A \leftarrow f_1 \cup_+ \ldots \cup_+ f_\ell$ (**multiset**)
  for each node x in A
    $f \leftarrow f \cup \{x\}$ with probability α/ℓ
  sample $c \in \mathbb{N}$ with probability P_c
  add c new nodes to f
  return f

The parameter ℓ controls "locality;" for instance, when $\ell = 1$, all nodes are sampled from a single affiliation to make a new one. Absorption is part of the model, and for fixed α and ℓ, an affiliation of size s is absorbed with probability $(\alpha/\ell)^s$. Therefore, the probability of absorption is $\sum_s (\alpha/\ell)^s P(s)$, where $P(s)$ is the probability of size s affiliations, and hence absorption is not vanishing with a growing network size, making the simplicial complex model distinct from that of a hypergraph. Finally, the number of newly created nodes in each step is controlled by the random parameter c, which may also be a constant.

When designing the algorithm, we decided not to allow c to be zero, and not to elaborate on the case of $\alpha = 1$. This design choice of $c \in \mathbb{N} = \{1, 2, 3, \ldots\}$ guarantees that no empty affiliations are created, i.e., $P(s = 0) = 0$. The smallest affiliation size is the smallest c such that $P_c > 0$. In addition, this choice avoids the possibility of adding the same hyperedge multiple times, as well as the possibility of subsuming a newly created facet, which will simplify the analysis of simplicial complexes. Similarly, the design choice of $\alpha \neq 1$ avoids a boundary case; for instance, when $\alpha = \ell = 1$, the newly created facet in each iteration will absorb the existing facet, leading to a degenerate network in simplicial complexes with all nodes belonging to one large facet.

In terms of implementation, both nodes and affiliations are stored as hash tables, which are efficiently updated. Moreover, each node maintains a list of the affiliations it belongs to, and each affiliation maintains a list of the nodes it contains. This means all updates to the network are local and, therefore, efficient. For instance, to check whether some facets are absorbed in simplicial complexes, we first obtain a list of facets for each sampled node. These facets can then be checked for inclusion in the newly created one. The time complexity of this entire operation is proportional to the sum of the sizes of all facets containing any sampled node. This depends on the average node degree and the average facet size, thus on the parameters

of the algorithm and not the size of the network. In addition, sampling ℓ affiliations followed by a sampling of nodes with probability α/ℓ implies that the algorithm does not rely on an explicit sampling of nodes based on their degrees, which could otherwise depend on the number of nodes. Indeed, sampling an affiliation uniformly at random requires only a constant time, and sampling its nodes requires a time proportional to the affiliation size.

A. Probability of gaining a degree

Since the degree of a node is the number of affiliations it belongs to (in both hypergraphs and simplicial complexes), in a network with m affiliations, any given node with degree k is selected to be part of a new affiliation with probability

$$P_+^{\text{union}}(k) = \left[1 - \left(1 - \frac{k}{m}\right)^{\ell}\right]\frac{\alpha}{\ell},$$

$$P_+^{\text{multiset}}(k) = 1 - \left(1 - \frac{k}{m}\frac{\alpha}{\ell}\right)^{\ell}.$$

When $\ell = 1$, both variants give $\alpha k/m$ (preferential attachment). However, for $\ell > 1$, they differ. In small networks where k and m are comparable, a feature of the union variant is that the above probability is approximately α/ℓ, especially for large ℓ since $(1 - k/m)^{\ell} \approx 0$, giving a uniform attachment model for small networks, and explaining an important transitional feature for the average distance in the skeleton graph when the network grows [50] (see Sec. VIII). The multiset variant remains preferential for small networks, albeit not linear in the degree, with an approximate probability of $1 - e^{-\alpha k/m}$.

In large networks where k/m goes to zero as m goes to infinity, both variants will have the same expression for the probability of selecting a given node with degree k, providing the preferential attachment feature:

$$P_+(k) \approx \frac{\alpha k}{m}. \tag{1}$$

The assumption that $\lim_{m\to\infty} k/m = 0$ is not unrealistic given the no overlap assumption stated below.

B. Two algorithm assumptions

Define P_{overlap} as the probability that among the $\ell > 1$ sampled affiliations, at least one pair overlaps, then the existence of a node with degree k such that $k/m \geqslant \epsilon$ implies $P_{\text{overlap}} \geqslant 2\binom{k}{2}/m^2 \geqslant \epsilon^2 - k/m^2 \geqslant \epsilon^2 - 1/m \approx \epsilon^2$. Our algorithm makes the following two assumptions:

No overlap: P_{overlap} goes to zero as m goes to infinity [hence k/m goes to zero and $P_+(k) \approx \alpha k/m$ is valid].

No accidental absorption: When a new facet is added to the network, the probability of absorbing a facet that is not among the ℓ sampled ones goes to zero as m goes to infinity (this will be useful for the analysis of simplicial complexes).

Both assumptions are justified experimentally in Fig. 4; for instance, $P_{\text{overlap}}(m) \propto m^b$ in hypergraphs, where b is always negative, even as $\alpha \to 1$. Similarly, we have a negative exponent b' in the average number of accidental absorptions per iteration $N_{\text{acc}}(n) \propto n^{b'}$ (n is the number of nodes in the network). Observe that the probability of accidental absorption is bounded from above by $N_{\text{acc}}(n)$ when n is large, as the latter should converge to the expected number of accidental absorptions. The two assumptions are proven true theoretically under reasonable conditions [14].

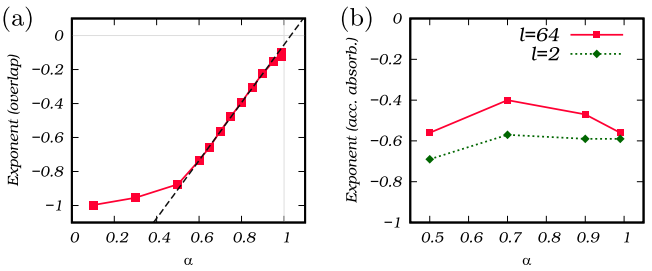

FIG. 4. (a) Empirically found (best-fit) exponent b in the probability of overlap for $c = 1$ and $\ell = 2$, assuming the probability has the form $P_{\text{overlap}}(m) \propto m^b$, and (b) the exponent b' of $N_{\text{acc}}(n) \propto n^{b'}$.

V. EMERGENCE OF POWER LAW

Let n_t and m_t denote the number of nodes and affiliations at iteration t, respectively, and let $n_{t,k}$ be the number of nodes of degree k at iteration t. To establish the power law, we track how $E[n_{t,k}]$ changes from one iteration to another in a classical master equation approach. We use $P(k)$ and $P(s)$ to describe the probability distributions for node degree and affiliation size, respectively.

A. Degree distribution in hypergraphs

For hypergraphs, m_t is the number of hyperedges at iteration t and satisfies $m_t = t$. Recall that each node of degree k gains a degree with probability $P_+(k) = \alpha k/m_t$; therefore, we have the following master equation:

$$E[n_{t+1,k}|n_t] = E[n_{t,k}|n_t] + E[c]\mathbb{1}_{k=1} - E[n_{t,k}|n_t]\frac{\alpha k}{t} + E[n_{t,k-1}|n_t]\frac{\alpha(k-1)}{t}.$$

If $P_t(k|n_t)$ is the probability that a node has degree k at iteration t given n_t nodes, then in the limit as $t \to \infty$, when a limiting distribution exists, this is just $P(k)$ for all "typical" values of n_t. Therefore, $E[n_{t,k}|n_t] = n_t P_t(k|n_t) = n_t P(k)$. Similarly, $E[n_{t+1,k}|n_t] = E[E[n_{t+1,k}|n_t, n_{t+1}]|n_t] = E[n_{t+1} P_{t+1}(k|n_t, n_{t+1})|n_t] = (n_t + E[c])P_{t+1}(k|n_t, n_{t+1}) = (n_t + E[c])P(k)$. The master equation becomes

$$E[c]P(k) = E[c]\mathbb{1}_{k=1} - \lim_{t\to\infty} n_t P(k)\frac{\alpha k}{t} + \lim_{t\to\infty} n_t P(k-1)\frac{\alpha(k-1)}{t}.$$

Replacing $\lim_{t\to\infty} n_t/t$ by $E[c]$ (and since $E[c] \neq 0$), we obtain the recurrence

$$P(k) = \mathbb{1}_{k=1} - \alpha k P(k) + \alpha(k-1)P(k-1). \tag{2}$$

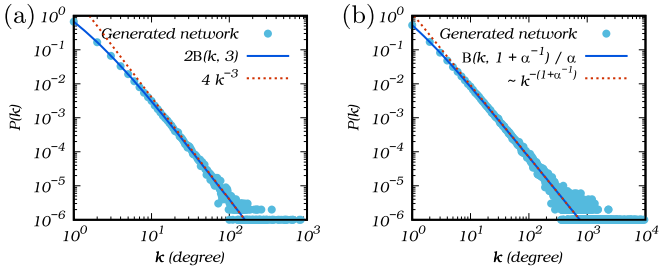

FIG. 5. Degree distribution for (a) $\alpha = 0.5$, $c = 1$, $\ell = 1$, and (b) $\alpha = 0.9$, $c = 1$, $\ell = 4$. Each panel shows an empirical distribution from a generated network with $n = 10^6$ nodes, the corresponding exact solution, and its Stirling's power law approximation.

Iterating the above recurrence gives

$$P(k) = \frac{(k-1)!}{\alpha(1/\alpha+1)^{(k)}} = \frac{\Gamma(k)\Gamma(1/\alpha+1)}{\alpha\Gamma(1/\alpha+1+k)} = \frac{1}{\alpha}B\left(k, \frac{1}{\alpha}+1\right), \tag{3}$$

which is the Yule-Simon distribution [2–4] (see Fig. 5 for some examples). The Yule-Simon distribution is asymptotically a power law satisfying $P(k) \propto k^{-(1/\alpha+1)} = k^{-\gamma}$ for large k, with γ given below for hypergraphs:

$$\gamma_{hg} = \frac{1}{\alpha} + 1. \tag{4}$$

Therefore, the power of the distribution has tail exponent $\gamma = \frac{1}{\alpha} + 1 > 2$ when $0 \leqslant \alpha < 1$. Observe that such distribution with infinite domain is valid when $\gamma > 1$, and admits a finite mean when $\gamma > 2$ ($\alpha < 1$) and a finite variance when $\gamma > 3$ ($\alpha < 0.5$). Our algorithm (for both hypergraphs and simplicial complexes) does not asymptotically generate networks with $\gamma < 2$. In these networks, degrees (and thus the number of affiliations) grow at a much faster rate than nodes. An example of such dense networks is the Facebook friendship network described in Figs. 2 and 3, with $1 < \gamma < 2$.

B. Degree distribution in simplicial complexes

Using a similar master equation technique (see Appendix A for detail), we derive $P(k)$ for a simplicial complex by relying on the notion of a "meganode," which is a node with a large degree k. To that end, define the conditional probability $P(s|k)$ as the probability that a facet has size s given it contains a node of degree k, and assume $Q(s) = \lim_{k\to\infty} P(s|k)$ exists, describing the facet size distribution of a meganode. Then for large k,

$$P(k) = [-P(k)k + P(k-1)(k-1)]\alpha^*,$$

where

$$\alpha^* = \alpha\frac{1 - \frac{\ell}{\alpha}h\left(\frac{\alpha}{\ell}\right)}{1 - \ell f\left(\frac{\alpha}{\ell}\right)}, \tag{5}$$

and $f(z) = \sum_s z^s P(s)$ and $h(z) = \sum_s z^s Q(s)$ are the generating functions for the facet size distribution in the network and the facet size distribution of a meganode, respectively.

This is the same as Eq. (2) for large k ($\mathbb{1}_{k=1}$ is dropped) with α replaced by α^*. Therefore, $P(k) \propto k^{-\gamma}$ for large k, with γ given below for simplicial complexes:

$$\gamma_{sc} = \frac{1}{\alpha^*} + 1. \tag{6}$$

Compared to hypergraphs, one can think of α^* as the *effective* α for simplicial complexes. Section VII provides numerical and approximation techniques for computing α^* using the generating functions $f(z)$ and $h(z)$.

VI. HYPEREDGE AND FACET SIZE DISTRIBUTIONS

We study the size distributions $P(s)$ and $Q(s)$ (see above) by exploring their generating functions $f(z) = \sum_s z^s P(s)$ and $h(z) = \sum_s z^s Q(s)$, respectively. Refer to Appendix B for the detailed mathematical derivations.

A. Hypergraph network

As shown in in Appendix B, the generating functions in a hypergraph satisfy

$$f(z) = g(z)f^\ell\left(1 - \frac{\alpha}{\ell} + \frac{\alpha}{\ell}z\right), \tag{7}$$

$$h(z) = \frac{zg(z)}{1 - \frac{\alpha}{\ell} + \frac{\alpha}{\ell}z}h\left(1 - \frac{\alpha}{\ell} + \frac{\alpha}{\ell}z\right)f^{\ell-1}\left(1 - \frac{\alpha}{\ell} + \frac{\alpha}{\ell}z\right), \tag{8}$$

where $g(z) = \sum_c z^c P_c$ is the generating function for the fixed distribution given by P_c, the probability of adding c new nodes in one iteration of the algorithm described in Sec. IV. For the special case of $\ell = 1$ and a constant c where $g(z) = z^c$, $f(z)$ can be iterated to obtain $f(z) = (1-z;\alpha)_\infty^c$ and $P(s=c) = \sum_{s\geqslant c} P(s)(1-\alpha)^s = f(1-\alpha) = (\alpha;\alpha)_\infty^c$, where $(x;q)_\infty$ is the q-Pochhammer symbol defined as $\prod_{i=0}^{\infty}(1-xq^i)$.

Given $f(z)$, one can compute several properties of the size distribution; for instance, the expected hyperedge size is given by

$$E[s] = f'(1) = \frac{E[c]}{1-\alpha}. \tag{9}$$

Since the sum of degrees is equal to the sum of affiliation sizes,

$$\frac{\sum_i d_i}{n_t} = \frac{\sum_f |f|}{m_t}\frac{m_t/t}{n_t/t}.$$

This implies at the limit that the expected node degree is given by

$$E[k] = \frac{E[s]}{E[c]} = \frac{1}{1-\alpha}. \tag{10}$$

B. Simplicial complex network

Similarly, we obtain the following for simplicial complexes:

$$f(z) = \frac{g(z)f^\ell\left(1 - \frac{\alpha}{\ell} + \frac{\alpha}{\ell}z\right) - \ell f\left(\frac{\alpha}{\ell}z\right)}{1 - \ell f\left(\frac{\alpha}{\ell}\right)}, \tag{11}$$

with

$$E[s] = \frac{E[c] - \alpha f'(\alpha/\ell)}{1 - \alpha - \ell f(\alpha/\ell)},\ E[k] = \frac{E[s]}{E[c]}[1 - \ell f(\alpha/\ell)]$$

and

$$\left[1 - \frac{\ell}{\alpha}h\left(\frac{\alpha}{\ell}\right)\right]h(z) + \frac{\ell}{\alpha}h\left(\frac{\alpha}{\ell}z\right) = \frac{zg(z)}{1 - \frac{\alpha}{\ell} + \frac{\alpha}{\ell}z}h\left(1 - \frac{\alpha}{\ell} + \frac{\alpha}{\ell}z\right)f^{\ell-1}\left(1 - \frac{\alpha}{\ell} + \frac{\alpha}{\ell}z\right). \tag{12}$$

It is easy to verify that for hypergraphs, and for simplicial complexes with $\ell = 1$, meganetworks exhibit a shift in the size distribution, where $h(z) = zf(z)$ and hence $Q(s) = P(s-1)$. This also implies that $\alpha^* = \alpha$ when $\ell = 1$ [see Eq. (5)].

VII. NUMERICAL SOLUTIONS AND POISSON APPROXIMATIONS

In this section, we describe numerical solutions for $f(z)$ and $h(z)$ given by Eqs. (7), (8), (11), and (12), as well as their approximate solutions when ℓ is large. Both approaches are useful, in particular for simplicial complexes to obtain α^*, which in turn determines the exponent of the degree distribution given by $\gamma = 1/\alpha^* + 1$ [Eq. (6)]. Solutions for $f(z)$ and $h(z)$ can also be used to find $E[s]$ and $E[k]$ (see previous section).

A. Numerical solutions for constant c

To solve $f(z)$ numerically for given ℓ and c, we start with a small α and $f_0(z) = z^c$ as an initial solution, and repeatedly compute $f(z)$ from previous values of $f(z)$, using Eq. (11) for instance, until convergence:

$$f_{i+1}(z) = \frac{z^c f_i^\ell\left(1 - \frac{\alpha}{\ell} + \frac{\alpha}{\ell}z\right) - \ell f_i\left(\frac{\alpha}{\ell}z\right)}{1 - \ell f_i\left(\frac{\alpha}{\ell}\right)},$$

where $f_i(z)$ is evaluated at $z_j = j/(N-1)$ for $j \in \{0, \ldots, N-1\}$ for some large integer N. When evaluating $f(z)$ for a general $z_j \leqslant z < z_{j+1}$, we interpolate between $f(z_j)$ and $f(z_{j+1})$. We then increase α and use the solution just computed as initial solution. The process is repeated until $f(z)$ is numerically computed for all values of α. To solve for $h(z)$ numerically, we do the same except that we use the corresponding $f(z)$ computed above as initial solution in $h_0(z) = f(z)$ for each α.

The iterative approach described above fails to converge in simplicial complexes for large α and $\ell = 1$ (Appendix C sheds some light on this difficulty when $\ell = 1$). Fortunately, when $\ell = 1$, $\alpha^* = \alpha$ (see the last paragraph of the previous section). Nevertheless, $f(\alpha/\ell) = \sum_s (\frac{\alpha}{\ell})^s P(s)$ determines the probability of absorption and $f'(1) = E[s]$, so it is useful to compute $f(z)$ even when $\ell = 1$. We present an alternative method to compute $f(z)$ [and $h(z) = zf(z)$] based on $P(s)$. Equation (B1) for $\ell = 1$ in Appendix B gives

$$P(s)[1 - f(\alpha) + \alpha^s] = \sum_{s_1 \geqslant c} P(s_1)\binom{s_1}{s-c}\alpha^{s-c}(1-\alpha)^{s_1-s+c}.$$

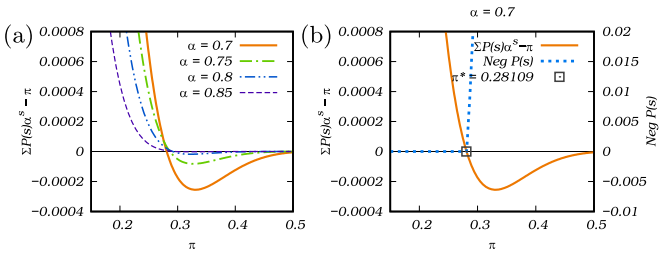

FIG. 6. (a) The difference between $\sum_s \alpha^s P(s)$ and π plotted as a function of π for $c = 1$ and $\alpha \in \{0.7, 0.75, 0.8, 0.85\}$. (b) $\sum_s \alpha^s P(s) - \pi$ and Neg$P$ for $\alpha = 0.7$.

By treating $f(\alpha)$ as a constant parameter π, we transform the above into a linear system with $s_{\max} - c + 1$ equations for some large integer $s_{\max}$, where $f(z)$ is approximated as $f(z) \approx \sum_{s=c}^{s_{\max}} z^s P(s)$:

$$P(s)(1 - \pi + \alpha^s) = \sum_{s_1=c}^{s_{\max}} P(s_1)\binom{s_1}{s-c}\alpha^{s-c}(1-\alpha)^{s_1-s+c}$$

$$\text{for } s \in \{c, \ldots, s_{\max} - 1\},$$

$$\sum_{s=c}^{s_{\max}} P(s) = 1.$$

We then search for a π that yields a solution in which π is approximately $f(\alpha) = \sum_{s=c}^{s_{\max}} \alpha^s P(s)$. Therefore, π is chosen by searching for $\pi^* = \arg\min |\pi - \sum_{s=c}^{s_{\max}} \alpha^s P(s)|$. The solution is validated by making sure that $\text{Neg}P = \sum_s |P(s)| \mathbb{1}_{P(s)<0} = 0$. Figure 6 shows how π^* is found.

Table I provides a comparison for the power γ between the numerical computation and the experimental fit from simulation. The iterative approach and the linear system can also be generalized for nonconstant c. For more elaborate tables showing α^*, $E[s]$, and $E[k]$ in simplicial complexes for various values of α, c, and ℓ, refer to [14].

B. Approximations for large ℓ

Given $0 \leqslant \alpha < 1$, when ℓ is large and $0 \leqslant z \leqslant 1$,

$$f\left(1 - \frac{\alpha}{\ell} + \frac{\alpha}{\ell}z\right) \approx f(1) + f'(1)\frac{\alpha(z-1)}{\ell} = 1 + E[s]\frac{\alpha(z-1)}{\ell}.$$

TABLE I. The exponent γ of the degree distribution in simplicial complex networks for $\alpha = 0.5$ obtained in two ways: (Left) computed from the formula $\frac{1}{\alpha^*} + 1$ using $f(z)$ and $h(z)$ of the numerical approach. (Right) as the linear fit of the $1 - CMF(k)$ (complement of the cumulative distribution) from simulation data.

	Computation				Simulation		
	$\ell = 1$	$\ell = 2$	$\ell = 4$	vs	$\ell = 1$	$\ell = 2$	$\ell = 4$
$c = 1$	3	2.835	2.799	$c = 1$	3.006	2.830	2.792
$c = 2$	3	2.976	2.984	$c = 2$	2.984	2.980	2.990
$c = 4$	3	3	3	$c = 4$	3.010	2.992	3.003

Since $\lim_{\ell\to\infty}(1+x/\ell)^\ell = e^x$, $f^\ell(1-\frac{\alpha}{\ell}+\frac{\alpha}{\ell}z) \approx e^{\alpha E[s](z-1)}$, and we have for hypergraphs

$$h_{hg}(z) = zf_{hg}(z) \approx zg(z)e^{\alpha E[s](z-1)}.$$

When c is constant, $f(z) \approx z^c e^{\alpha E[s](z-1)}$ and, therefore,

$$P(s) \approx \frac{(\alpha E[s])^{s-c}e^{-\alpha E[s]}}{(s-c)!},$$

which is a "shifted" Poisson distribution, so we refer to the approximations in this section as Poisson approximations. Intuitively, the size of the affiliation before the addition of c nodes is a binomial random variable with probability $p = \alpha/\ell$ and a number of trials $n = s_1 + \cdots + s_\ell$ (the sum of sizes of ℓ sampled affiliations), giving an approximate Poisson with parameter $\lambda = \lim_{\ell\to\infty} pn = \lim_{\ell\to\infty} \frac{\alpha}{\ell}(s_1 + \cdots + s_\ell) = \alpha E[s]$.

For simplicial complexes, the situation is similar. When ℓ is large, the arguments for $f(\alpha/\ell)$ and $f(\alpha z/\ell)$ are close to 0. For small z,

$$f(z) \approx f(0) + f'(0)z = zf'(0) = zP(s=1).$$

Therefore, $\ell f(\alpha/\ell) \approx \alpha P(s=1)$ and $\ell f(\frac{\alpha}{\ell}z) \approx \alpha z P(s=1)$. Using Eq. (11) we get

$$f_{sc}(z) \approx \frac{g(z)e^{\alpha E[s](z-1)} - \alpha z P(s=1)}{1-\alpha P(s=1)}. \tag{13}$$

This is interesting because it shows that as long as $P(s=1) = 0$, i.e., there are no facets with singleton nodes, simplicial complexes behave like hypergraphs at the limit when ℓ goes to infinity [and the probability of absorption $f(\alpha/\ell) = \sum_s (a/\ell)^s P(s)$ vanishes].

On the other hand, $Q(s=1) = 0$ because facet sizes in a meganode's network must be greater than 1 (they all contain that node). Therefore, when ℓ is large, and using Eq. (12), we can similarly derive

$$h_{sc}(z) \approx h_{hg}(z) = zf_{hg}(z).$$

We can now approximate α^* and γ for large ℓ:

$$\alpha^* = \alpha\frac{1-\frac{\ell}{\alpha}h_{sc}\left(\frac{\alpha}{\ell}\right)}{1-\ell f_{sc}\left(\frac{\alpha}{\ell}\right)} \approx \alpha\frac{1-0}{1-\alpha P(s=1)},$$

$$\gamma = \frac{1}{\alpha^*} + 1 \approx \frac{1}{\alpha} + 1 - P(s=1).$$

It remains to approximate $P(s=1)$. Equation (13) for $f_{sc}(z)$ above can be expanded as

$$f_{sc}(z) = \sum_{s\geqslant 1} z^s P(s) \approx \frac{\sum_{c\geqslant 1} z^c P_c e^{\alpha E[s](z-1)} - \alpha z P(s=1)}{1-\alpha P(s=1)}.$$

Dividing by z on both sides and taking the limit as z goes to zero, we get

$$P(s=1) \approx \frac{P_1 e^{-\alpha E[s]} - \alpha P(s=1)}{1-\alpha P(s=1)}.$$

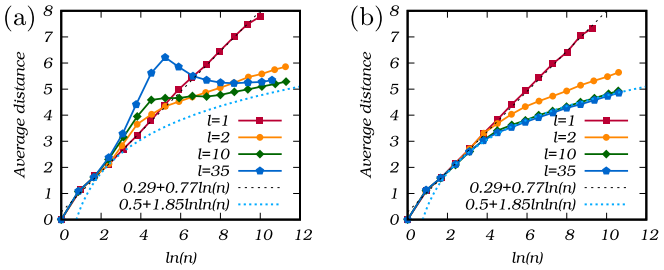

FIG. 7. Average distance in a hypergraph network for (a) the union and (b) the multiset variants of the node sampling process plotted for $\alpha = 0.5$, $c = 1$ and $\ell = 1\ldots35$. Dashed lines show logarithm and double-logarithm fits for the multiset variant. There is an "anomaly" in the union variant at $n \approx 50$–500 (or equivalently, $\ln n \approx 4$–6) when ℓ is large. This effect fades away with the growth of the network when the number of nodes reaches $n \approx 10^5$–10^6.

This quadratic equation for $P(s=1)$ has one solution greater than 1, and another given by

$$P(s=1) \approx \frac{1+\alpha-\sqrt{(1+\alpha)^2 - 4\alpha e^{-\alpha E[s]}P_1}}{2\alpha},$$

where $E[s]$ can be approximated by its hypergraph counterpart $E[c]/(1-\alpha)$. For $\alpha = 0.5$ and a constant c, this approximation (for large ℓ) gives $\gamma = 2.7305$ when $c = 1$, and $\gamma = 3$ when $c > 1$, which represent the convergence of γ in Table I if ℓ continues to increase.

VIII. SOME EMPIRICAL PROPERTIES OF THE NETWORKS

In this section, we discuss the average distance, the clustering coefficient, and the assortativity of our generated networks, thus highlighting some of their features in resembling real-world networks.

A. Average distance

We use the standard definition of distance between two nodes in the skeleton graph of the network. To compute all pairwise distances, an exact solution can be found by performing "breadth first search" from each node, which requires n single-source shortest path computations. Instead, we used the approximation algorithm described in [51], which requires $O(\epsilon^{-2}\ln n)$ such computations. The estimate of the average distance is guaranteed to be within an ϵ relative error from the real value with high probability. Practically, in a network with $n = 10^6$ nodes, allowing an error $\epsilon = 0.1$ (the actual error is better in practice), the algorithm would require only in the order of 1000 single-source shortest path computations.

Figure 7 shows the average distance in a hypergraph for the union and multiset variants described in Sec. IV. The union variant exhibits an "anomaly" with two regimes when ℓ is large: Small networks ($n < 200$) have uniform attachment (which leads to a geometric degree distribution [5,14]), and large networks ($n > 200$) have preferential attachment with a power law in the tail of the degree distribution. A strikingly similar feature has been observed in some networks, e.g., word-adjacency networks [50].

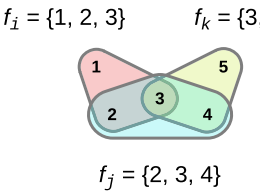

FIG. 8. An example of a "cat" structure.

B. Clustering coefficient

Many real-world graph networks are characterized by a nonzero global clustering coefficient gcc. For instance, in a social context, a nonzero clustering coefficient implies that if Alice knows Bob and Carol, then there is a chance Bob and Carol know each other as well (informally, there is a "triangle"). However, nonzero gcc generally does not occur in most synthetic graph networks unless the network growth algorithm includes an explicit triangle closure operation to create triangles from 2-paths. Affiliation networks naturally join nodes as groups, and so have a built-in mechanism for forming these triangles in the skeleton graph.

To avoid this trivial behavior, we extend the definition of gcc to specifically handle affiliation networks. We adopt Opsahl's definition [47]

$$\mathrm{gcc}_{\mathrm{Opsahl}} = \frac{\text{number of closed 4-paths}}{\text{number of 4-paths}}, \tag{14}$$

where a 4-path is defined as (u, f, v, g, w) with $u \neq v \neq w \in V$, $f \neq g \in F$, $u, v \in f$, and $v, w \in g$. A 4-path is closed if its end points u and w are included in a third affiliation h that is distinct from f and g. Therefore, in a social network setting, this extension does not simply compute the proportion of a person's acquaintances who know each other. Instead, it requires that each pair in a closed 4-path be connected through a different social group. For example, Alice and Bob are siblings, Alice and Carol are classmates, and Bob and Carol go to the same club.

Figure 9(a) illustrates $\mathrm{gcc}_{\mathrm{Opsahl}}$ in hypergraphs and simplicial complexes for $c = 1$, $\ell = 1$, and $\alpha \leqslant 0.5$, showing that it reaches approximately 0.05 when $\alpha = 0.4$. The value of $\mathrm{gcc}_{\mathrm{Opsahl}}$ does not converge as quickly in networks with

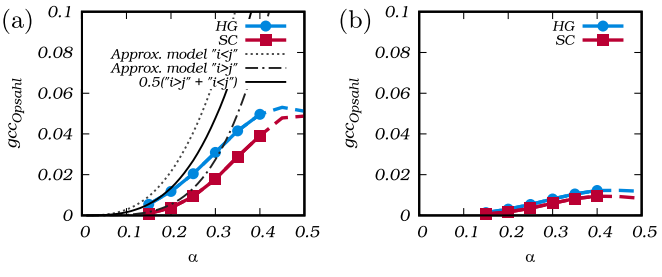

FIG. 9. (a) Experimentally computed clustering coefficient and two approximate models of cat structure formation for $c = 1$ and $\ell = 1$. The relation $i < j$ ($i > j$) indicates that f_i was created before (after) f_j in the formation of a cat structure (Fig. 8). When α is small, the two models provide a good approximation of the clustering coefficient in hypergraphs and simplicial complexes, respectively. (b) Experimental clustering coefficient for $c = 1$ and $\ell = 2$.

larger values of α, making it hard to experimentally study the clustering coefficient in these networks.

We believe that the positive clustering coefficient is due to a spatial closed 4-path formed by the structure in Fig. 8, which we call a "cat structure." A cat structure consists of three affiliations that form a closed 4-path, where one of the affiliations contains all three nodes of that path. Therefore, a cat structure contains three affiliations f_i, f_j, and f_k, and three nodes u, v, and w, such that $u, v \in f_i$, $u, v, w \in f_j$, and $v, w \in f_k$, making (u, f_i, v, f_k, w), (v, f_k, w, f_j, u), and (w, f_j, u, f_i, v) all closed 4-paths (closed by f_j, f_i, and f_k, respectively). Such a structure can be formed by adding the affiliations in one of two orders f_i, f_j, f_k (thus $i < j$) or f_j, f_i, f_k (thus $i > j$). Cat structures can occur even when $\ell = 1$, which does not conform to the intuitive notion of a triangle formed by sampling nodes u and w from two different affiliations, and joining them in one newly created affiliation h, which closes some 4-path (u, f, v, g, w) (f and g may or may not be among the ℓ sampled affiliations). This becomes rare as the network grows in size: the no overlap assumption precludes f and g to be both among the sampled affiliations (because they overlap on v), and in general, the distance between u and w is expected to be large (while the existence of v makes that distance at most 2). Cat structures do no rely on such rare events and, therefore, they become the primary way of creating closed 4-paths. However, the formation of a cat structure requires affiliations to overlap on more than one node which is easier to achieve when α/ℓ is large.

In Fig. 9, two approximate models for $\mathrm{gcc}_{\mathrm{Opsahl}}$ match their experimental counterpart for hypergraphs and simplicial complexes when $c = 1$, $\ell = 1$, and α is small.

C. Assortativity

Assortativity is the tendency of a network to form links between *similar* nodes. The most direct and immediately available property of the nodes is their degree. We consider three definitions of assortativity, all motivated by the definition in a graph introduced by [49], which is based on the *Pearson correlation coefficient* of the degrees of adjacent nodes.

Let p_{ij} be a joint probability over a subset of the ordered pairs $\{(i, j) \mid i, j \in \mathbb{N}\}$, where $\mathbb{N} = \{1, 2, 3 \ldots\}$. The Pearson correlation coefficient is defined as

$$r = \frac{E[ij] - E[i]E[j]}{\sqrt{(E[i^2] - E[i]^2)(E[j^2] - E[j]^2)}}.$$

We consider the case when p_{ij} is symmetric, i.e., $p_{ij} = p_{ji}$. The above becomes

$$r = \frac{E[ij] - E[i]^2}{E[i^2] - E[i]^2} = \frac{\sum_{(i,j)} ijp_{ij} - \left[\sum_{(i,j)} ip_{ij}\right]^2}{\sum_{(i,j)} i^2 p_{ij} - \left[\sum_{(i,j)} ip_{ij}\right]^2}. \tag{15}$$

If we then define $q_{ij} = p_{ij} + p_{ji} = 2p_{ij}$ when $i \neq j$, and $q_{ij} = p_{ij}$ when $i = j$, then the above equation becomes

$$r = \frac{\sum_{\{i,j\}} ijq_{ij} - \left[\sum_{\{i,j\}} \frac{i+j}{2} q_{ij}\right]^2}{\sum_{\{i,j\}} \frac{i^2+j^2}{2} q_{ij} - \left[\sum_{\{i,j\}} \frac{i+j}{2} q_{ij}\right]^2}, \tag{16}$$

where the sums are over multisets (for instance, $\{i, i\}$ is among them). This has the same form as the assortativity defined

in [49]. Therefore, to define an assortativity coefficient, we describe how we map some random process to the set of ordered pairs of degrees (i, j), as long as the mapping produces a symmetric joint probability p_{ij}. For instance, to compute the assortativity in a graph $G = (V, E)$, we can sample uniformly (with probability $\frac{1}{2|E|}$) from the set

$$S_{\text{graph}} = \bigcup_{\substack{u \neq v \\ (u,v) \in E}} \{(u, v)\},$$

and map (u, v) to the ordered pair of degrees $(i, j) = (d(u), d(v))$. This corresponds to an experiment in which we sample an edge (u, v) uniformly at random, then make the pair of degrees be either $(d(u), d(v))$ or $(d(v), d(u))$ with probability $\frac{1}{2}$. Equation (16) [and equivalently Eq. (15)] then becomes identical to the definition in [49].

While we describe the assortativity coefficient in terms of a sampling and mapping process, we do not actually perform any sampling. Instead, we explicitly compute p_{ij} by enumerating all outcomes that map to (i, j) and use Eq. (15).

Simple neighbor (SN): Assortativity is defined by Eq. (15) when sampling (u, v) uniformly at random from the set

$$S_{\text{SN}} = \bigcup_{\substack{u \neq v \\ \exists f \in F : (u,v) \in F}} \{(u, v)\},$$

where (u, v) is then mapped to $(i, j) = (d(u), d(v))$. This corresponds to an experiment in which each connected (ordered) pair of nodes has the same probability of being chosen; therefore, we essentially get the assortativity of the skeleton graph. For a simple graph, SN is equivalent to [49], but the two will differ on multigraphs.

Multineighbor (MN): Assortativity is defined by Eq. (15) when sampling (u, v, f) uniformly from the set

$$S_{\text{MN}} = \bigcup_{f \in F} \bigcup_{\substack{u \neq v \\ (u,v) \in f}} \{(u, v, f)\},$$

so each (u, v, f) is sampled with probability

$$\frac{1}{\sum_{f \in F} |f|(|f| - 1)}$$

and we map (u, v, f) to a pair of degrees $(i, j) = (d(u), d(v))$. This corresponds to an experiment in which we first sample (u, v, f) uniformly at random, where $(u, v) \in f$, $f \in F$, and $u \neq v$, and then obtain $(d(u), d(v))$. Therefore, an (ordered) pair of nodes is chosen with probability proportional to the number of affiliations they both belong to. When α is small and ℓ is large, affiliations are unlikely to overlap on more than one node; therefore, MN becomes effectively the same as SN (see Fig. 10).

Weighted multineighbor (WgtMN): Assortativity is defined by Eq. (15) when sampling from the set

$$S_{\text{WgtMN}} = \bigcup_{f \in F} \bigcup_{\substack{u \neq v \\ (u,v) \in f}} \{(u, v, f)\},$$

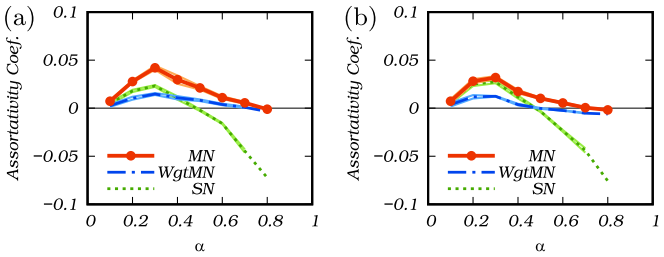

FIG. 10. Multineighbor (MN), weighted multineighbor (WgtMN), and simple neighbor (SN) assortativity coefficients in simplicial complexes plotted as the average of multiple runs for (a) $c = 1$, $\ell = 2$ and (b) $c = 1$, $\ell = 8$. The thin lines show the standard error of the mean.

where (u, v, f) has probability

$$\frac{1}{|F_{\geqslant 2}|} \cdot \frac{1}{|f|(|f| - 1)},$$

and $F_{\geqslant 2}$ is the set of all affiliations of size at least two. As before, we map (u, v, f) to $(i, j) = (d(u), d(v))$. This corresponds to an experiment in which we sample an affiliation f ($|f| \geqslant 2$) uniformly at random, then choose an ordered pair of nodes $(u, v) \in f$ with probability $\frac{1}{|f|(|f|-1)}$.

The main difference between MN and WgtMN is that the former assigns equal weight to each (u, v, f), while the latter assigns equal weight to each affiliation f. As a result, when using WgtMN, small affiliations, such as edges and triangles, contribute equally to the assortativity coefficient as the large affiliations.

Figure 10 shows the assortativity coefficient when $c = 1$, with a transition from positive to negative depending on which definition of assortativity we adopt. In general, a larger c reduces the difference between MN and WgtMN because it makes all affiliations relatively large. Both MN and WgtMN increase with c until they stabilize since a larger c means more nodes that are similar in degree are added in each iteration. For instance, when $c \geqslant 4$ and $\alpha \approx 0.3$, MN and WgtMN coefficients are around 0.12 (see Fig. 12). A larger ℓ, on the other hand, leads to lower MN and WgtMN coefficients since the sampled nodes that make up new affiliations will come from ℓ (many) existing ones and, therefore, are more likely to be different in degree.

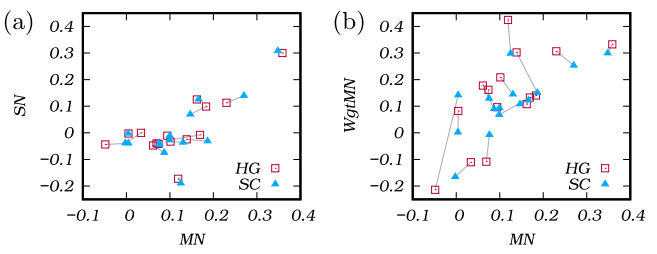

FIG. 11. Assortativity coefficients of real networks from Table II plotted as (a) SN vs MN and (b) WgtMN vs MN. Hypergraphs (red squares) and the corresponding simplicial complexes (blue solid triangles) are connected by gray lines.

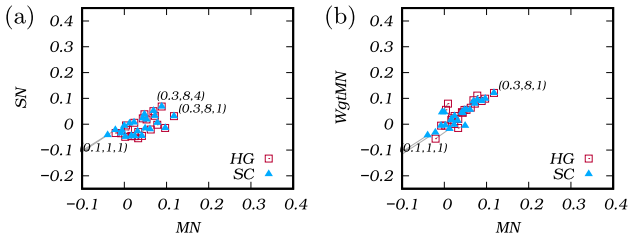

FIG. 12. Assortativity coefficients of generated networks for $\alpha \in \{0.1, 0.3, 0.5, 0.7\}$, $c \in \{1, 2, 8\}$, $\ell \in \{1, 8\}$, and number of nodes $n = 2 \times 10^5$, plotted similarly to Fig. 11 as (a) SN vs MN and (b) WgtMN vs MN. Each simplicial complex is obtained from the corresponding hypergraph by subsuming hyperedges that are subsets of others (instead of generating a completely new network). Some data points are labeled with the triplet (α, c, ℓ). Note that hypergraph networks have SN $\approx$ MN $\approx$ WgtMM ≈ -0.4 when $\alpha = 0.1$ and $c = 1$ and are outside the plotted area, while the corresponding simplicial complexes have assortativity close to 0. On the other hand, maximum assortativity is achieved when $\alpha \approx 0.3$ and $c \gg 1$.

D. Assortativity and clustering coefficient in real networks

Table II shows the assortativity and the clustering coefficient for several networks from the KONECT collection [52] obtained in the following way: Bipartite graphs are converted to their affiliation network equivalent, where edge (u, v) in the bipartite graph means that node u belongs affiliation v. Directed and undirected graphs are converted to their neighborhood affiliation network, where all neighbors of node u in the graph (not including u itself) make one affiliation. Each of the obtained affiliation networks is then interpreted as a hypergraph and as a simplicial complex. For simplicity, we did not check for duplicate hyperedges in the resulting hypergraphs.

In addition to Table II, Figs. 11 and 12 provide a comparison among the assortativity of the KONECT networks in Table II and that of our generated networks for several values of α, c, and ℓ. Finally, Table III summarizes the general effect of α, c, and ℓ on different properties.

IX. GUIDELINES FOR GENERATING NETWORKS WITH GIVEN PROPERTIES

Although our model was not designed for network replication, its parameters α, ℓ, and P_c are flexible enough to generate networks with some desired properties. In general, α controls the exponent γ of the power law in $P(k) \propto k^{-\gamma}$, P_c determines the affiliation size distribution $P(s)$, and ℓ can fine tune $P(k)$ and $P(s)$ [since it affects $P(s)$ in hypergraphs and both $P(k)$ and $P(s)$ in simplicial complexes].

TABLE II. Assortativity MN, WgtMN, and SN, and clustering coefficient $\text{gcc}_{\text{Opsahl}}$ computed for several real networks from the KONECT collection [52]. Missing $\text{gcc}_{\text{Opsahl}}$ values are due to avoiding the time complexity in larger networks.

Network	Connection nature	Transformation	SN	MN	WgtMN	$\text{gcc}_{\text{Opsahl}}$
Corporate leadership [53,54]	Company	Bip G → HG	−0.0437	−0.0483	−0.2142	0.6720
		Bip G → SC	−0.0392	0.0046	0.1416	0.5473
Amazon products ratings [55,56]	Product rating	Bip G → HG	−0.0027	0.0048	0.0823	0.0011
		Bip G → SC	−0.0024	0.0041	0.0018	0.0010
MovieLens movie ratings [57,58]	Movie rating	Bip G → HG	−0.0108	0.0940	0.0963	
		Bip G → SC	−0.0114	0.0995	0.0933	
FilmTrust movie ratings [59,60]	Movie rating	Bip G → HG	−0.0475	0.0613	0.1775	
		Bip G → SC	−0.0745	0.0870	0.0901	
Crime [61]	Committed crime	Bip G → HG	0.0000	0.0336	−0.1103	0.0318
		Bip G → SC	−0.0386	−0.0024	−0.1658	0.0270
arXiv cond-mat [62,63]	Coauth/Paper	Bip G → HG	0.0988	0.1831	0.1406	0.2769
		Bip G → SC	0.0695	0.1468	0.1080	0.1661
DBpedia writers [64,65]	Coauth/Work	Bip G → HG	0.3002	0.3585	0.3325	0.1310
		Bip G → SC	0.3078	0.3482	0.2994	0.0728
DBpedia producers [65,66]	Coauth/Work	Bip G → HG	0.1129	0.2300	0.3064	0.3037
		Bip G → SC	0.1397	0.2702	0.2529	0.2888
US airports [67,68]	Flight	Dir G → Nbhd HG	−0.1724	0.1189	0.4244	
		Dir G → Nbhd SC	−0.1898	0.1249	0.2981	
Enron emails [69,70]	Email sent	Dir G → Nbhd HG	−0.0329	0.1014	0.2085	
		Dir G → Nbhd SC	−0.0365	0.1303	0.1446	
arXiv high energy physics [71,72]	Citation	Dir G → Nbhd HG	−0.0413	0.0739	0.1620	
		Dir G → Nbhd SC	−0.0396	0.0754	0.1287	
Zachary karate club [73,74]	Personal ties	Un G → Nbhd HG	−0.0076	0.1693	0.1327	0.4534
		Un G → Nbhd SC	−0.0274	0.0988	0.0680	0.4054
Internet topology [75,76]	Connection	Un G → Nbhd HG	−0.0243	0.1385	0.3024	
		Un G → Nbhd SC	−0.0313	0.1865	0.1505	
C. elegans metabolism [77,78]	Protein interaction	Un G → Nbhd HG	−0.0399	0.0693	−0.1085	0.5879
		Un G → Nbhd SC	−0.0439	0.0762	−0.0075	0.5584
Yeast metabolism [79,80]	Protein interaction	Un G → Nbhd HG	0.1256	0.1618	0.1079	0.0839
		Un G → Nbhd SC	0.1261	0.1662	0.1223	0.0659

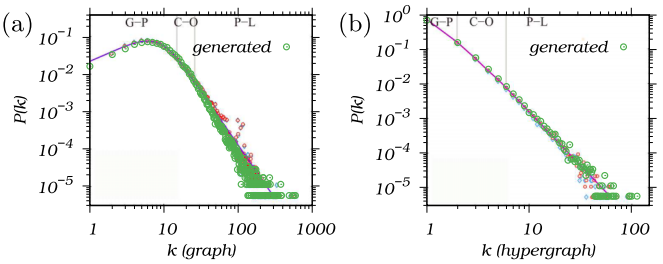

FIG. 13. Replication of the PNAS coauthorship network from Fig. 1. The green circles show the distributions of a generated hypergraph network with $\alpha = 0.4$, $c \sim \text{Geom}(0.26)$ (with support $\mathbb{N} = \{1, 2, 3, \ldots\}$), $\ell = 3$, and a number of nodes $n = 180\,000$. (The replicated distributions are superimposed on the original ones.)

A. Tail of the power law distribution $P(k)$

Given a degree distribution $P^*(k)$, the power exponent of the tail γ^* can be estimated, for instance, as described in [81–83]. Alternatively, γ^* itself could be available in the first place. To get the same power law tail exponent in a generated hypergraph network, one could then set $\alpha = \frac{1}{\gamma^*-1}$ [see Eq. (4)]. For simplicial complexes, this formula is only a good approximation [see Eq. (6)], but still provides a starting point to search for a better α, especially if ℓ is large and P_c favors large c, so that absorption is not very likely and the simplicial complex behaves like a hypergraph.

Whether the network is hypergraph or a simplicial complex, matching the exponent of the tail does not guarantee that $P(k)$ replicates $P^*(k)$ entirely; nevertheless, the focus here is on the tail. Once this is done, ℓ can be chosen to fine tune the result, although its effect might be small. Figure 13 shows that we can produce the degree distribution in the hypergraph network of Fig. 1(b) with $\alpha = 0.4$ and $\ell = 3$.

B. Affiliation size distribution $P(s)$

Given an affiliation size distribution $P^*(s)$, one can attempt to mimic it by a careful choice of P_c. For instance, if $P^*(s)$ has a mean μ, one can use $c = \max(1, [s^* - \alpha\mu])$ where $s^* \sim P^*(s)$ and $[\ldots]$ denotes the closest integer function. This means that $P_c = \sum_{s|\max(1,[s-\alpha\mu])=c} P^*(s)$. Appendix D considers a theoretical treatment of this choice with some examples, but to explain it intuitively, given s^*, c is chosen such that $c \in \mathbb{N} = \{1, 2, 3, \ldots\}$ and when combined with the mean of sampled nodes in the newly created affiliation (approximately $\alpha\mu$ as described in Appendix D), we get an expected total of $s = c + \alpha\mu \approx (s^* - \alpha\mu) + \alpha\mu = s^*$ new nodes. Again, this approach can be tried for few small values of ℓ to fine tune the result.

When $P^*(s)$ is not available, one could possibly determine P_c by relying on other properties of the network; for instance, the degree distribution in the skeleton graph [e.g., Fig. 1(a)], which indirectly captures some information about affiliation sizes (larger affiliation sizes lead to larger degrees in the skeleton graph). An initial manual search for P_c to obtain Fig. 1(a) revealed that $\frac{P_1}{P_2} \approx \frac{P_2}{P_3} \approx \cdots$. We then searched among geometric distributions and found that $c \sim \text{Geom}(0.26)$ will reasonably match the degree distribution in the skeleton graph. Figure 13 shows that we can produce the network given in Fig. 1 with $\alpha = 0.4$, $c \sim \text{Geom}(0.26)$, and $\ell = 3$.

TABLE III. The general trend for the effects of α, c, and ℓ on different properties of the generated network. This is approximate and is not intended to provide an exact behavior; for instance, properties may stop changing while some parameters continue to increase, and $\text{gcc}_{\text{Opsahl}}$ was only computed for $\alpha \leqslant 0.5$.

	α	c	ℓ
Average distance	↘		↘
Clustering coefficient $\text{gcc}_{\text{Opsahl}}$	↗		↘
Assortativity coefficients MN and WgtMN		↗	↘

C. Other network properties

As described in Sec. VIII, α, c, and ℓ can produce different results in terms of average distances, the clustering coefficient, and assortativity. Choices of these parameters can be made to obtain the desired properties (although not necessarily simultaneously), guided by Figs. 7, 9, and 10. Larger values of α and ℓ lead to better connectivity and, therefore, shorter distances. The clustering coefficient is higher for a larger α (but only tested for $\alpha \leqslant 0.5$) and a small ℓ. Larger values of c (and smaller values of ℓ in case of MN and WgtMN) lead to higher assortativity as mentioned in Sec. VIII and illustrated in Fig. 12.

X. CONCLUSION

Real-world networks that do not appear to be scale free when given as graphs can still exhibit scale-free properties if seen as affiliation networks based on hypergraphs or simplicial complexes. This suggests that the inherent pairwise relations among the nodes in a graph may be insufficient in capturing the growth of these networks, which prompts the search for new models of network growth targeted at affiliation networks. To that end, we propose an extension to the preferential attachment mechanism to generate affiliation networks with a power law in the tail of the distribution of node degrees. Our model uses implicit preferential attachment and introduces the concept of locality, where the neighbors of newly added nodes are also neighbors. In addition, and by considering simplicial complexes as an underlying structure for the affiliation network, our model allows the subsumption of existing affiliations when a newly created affiliation becomes their superset. We present a theoretical analysis of the node degree and affiliation size distributions $P(k)$ and $P(s)$ respectively. While the replication of networks is not the main goal of our generation algorithm, we provide experimental evidence that the algorithm produces a distribution tail of the form $P(k) \propto k^{-\gamma}$, where $\gamma > 2$ can be chosen, and that it has enough flexibility to make $P(s)$ close enough to a given distribution. We also include an experimental study of average distance, clustering coefficient, and assortativity. Overall, we show that the model is rich enough to provide new ways of controlling network growth in practice, while still providing interesting theory. Therefore, our model can be used as a basis for explaining the formation of real-world networks, as

well as for creating new synthetic networks with some desired properties.

The source code and documentation for the algorithm described in Sec. IV can be found at the GitHub repository [84].

APPENDIX A: POWER LAW IN SIMPLICIAL COMPLEXES

The case of simplicial complexes requires two adjustments for the master equation. First, the number of facets m_t is no longer deterministically equal to t since absorption may occur. We therefore condition on m_t. Second, each node of degree k now gains a degree with probability

$$P_+(k)\left[1-\sum_s \left(\frac{\alpha}{\ell}\right)^{s-1} P(s|k)\right],$$

where $P_+(k) \approx \alpha k/m_t$ [from Eq. (1) in Sec. IV]. The additional factor is the probability that the sampled facet of size s is not absorbed, so aside from the given node itself, it is the probability that not all of the remaining $s-1$ nodes get sampled. The probability $P(s|k)$ is that of a facet of size s given it contains a node of degree k since one cannot assume independence between the size of a facet and the degrees of its nodes; for instance, when c is constant, a facet of size c can only have nodes of degree one.

The no overlap assumption means that a node cannot lose a degree since at most one sampled facet is absorbed. In addition, with the no accidental absorption assumption, we need only consider absorption of the sampled facets. The master equation for *large* k becomes (no $E[c]\mathbb{1}_{k=1}$ term)

$$\begin{aligned}E[n_{t+1,k}|n_t,m_t] = {} & E[n_{t,k}|n_t,m_t] \\ & - E[n_{t,k}|n_t,m_t]\frac{\alpha k}{m_t}\left[1-\frac{\ell}{\alpha}h\left(\frac{\alpha}{\ell}\right)\right] \\ & + E[n_{t,k-1}|n_t,m_t]\frac{\alpha(k-1)}{m_t}\left[1-\frac{\ell}{\alpha}h\left(\frac{\alpha}{\ell}\right)\right],\end{aligned}$$

where $h(z)=\sum_s Q(s)z^s$ is the generating function for $Q(s)=\lim_{k\to\infty}P(s|k)$, which represents the facet size distribution of a "meganode" (a node with large degree k).

Using similar ideas to the case of hypergraphs, we have at the limit $E[n_{t,k}|n_t,m_t]=n_tP(k)$ and $E[n_{t+1,k}|n_t,m_t]=(n_t+E[c])P(k)$. We then rewrite the above as

$$\begin{aligned}E[c]P(k) = {} & -\lim_{t\to\infty} n_tP(k)\frac{\alpha k}{m_t}\left[1-\frac{\ell}{\alpha}h\left(\frac{\alpha}{\ell}\right)\right] \\ & + \lim_{t\to\infty} n_tP(k-1)\frac{\alpha(k-1)}{m_t}\left[1-\frac{\ell}{\alpha}h\left(\frac{\alpha}{\ell}\right)\right].\end{aligned}$$

Since in every iteration we add one facet, and each of the ℓ sampled facets can be absorbed,

$$\begin{aligned}E[m_t] &= (t-\tau)\left[1-\ell\sum_s\left(\frac{\alpha}{\ell}\right)^s P(s)\right]+O(\tau) \\ &= (t-\tau)\left[1-\ell f\left(\frac{\alpha}{\ell}\right)\right]+O(\tau),\end{aligned}$$

where τ is large enough for $P(s)$ to converge, and $f(z)=\sum_s z^sP(s)$ is the generating function for $P(s)$. We can therefore replace $\lim_{t\to\infty} n_t/m_t$ by $E[c]/[1-\ell f(\alpha/\ell)]$ to obtain

$$\begin{aligned}P(k) &= [-P(k)k+P(k-1)(k-1)]\alpha\frac{1-\frac{\ell}{\alpha}h\left(\frac{\alpha}{\ell}\right)}{1-\ell f\left(\frac{\alpha}{\ell}\right)} \\ &= [-P(k)k+P(k-1)(k-1)]\alpha^*,\end{aligned}$$

which satisfies $P(k)\propto k^{-(1/\alpha^*+1)}$ for large k.

APPENDIX B: HYPEREDGE AND FACET SIZE DISTRIBUTIONS

With m_t representing the number of affiliations at iteration t, let $m_{t,s}$ denote, among those, the number of affiliations of size s. We study the size distribution $P(s)$ in hypergraphs and simplicial complexes by deriving the generating function for $P(s)$ from the expression of $E[m_{t,s}]$. As before, we use the idea that $P_t(s|m_t)$ in the limit for large t is just $P(s)$. Then $E[m_{t,s}]=E[E[m_{t,s}|m_t]=E[m_tP_t(s|m_t)]=E[m_tP(s)]=E[m_t]P(s)$.

1. Hypergraph network

Consider a constant c, and assume that the hyperedge size distribution converged to a stationary $P(s)$ after τ iterations. For $t\gg\tau$,

$$\begin{aligned}E[m_{t,s}] = {} & O(\tau)+(t-\tau)\sum_{s_1,\ldots,s_\ell} P(s_1)\ldots P(s_\ell) \\ & \times\left(\frac{\alpha}{\ell}\right)^{s-c}\left(1-\frac{\alpha}{\ell}\right)^{\sum s_i-s+c}\binom{\sum s_i}{s-c},\end{aligned}$$

where $s_1,\ldots,s_\ell$ represent the sizes of the sampled hyperedges, and the probability of creating a hyperedge of size s is given by the binomial probability of selecting $s-c$ nodes from a total of $\sum s_i$ nodes (given the no overlap condition).

Using $E[m_{t,s}]=E[m_t]P(s)$, $m_t=t$, and finally dividing by t as $t\to\infty$, we obtain

$$\begin{aligned}P(s) = {} & \sum_{s_1,\ldots,s_\ell} P(s_1)\ldots P(s_\ell) \\ & \times\left(\frac{\alpha}{\ell}\right)^{s-c}\left(1-\frac{\alpha}{\ell}\right)^{\sum s_i-s+c}\binom{\sum s_i}{s-c}.\end{aligned}$$

Multiplying both sides of the equation by z^s and summing over s, we find the generating function $f(z)=\sum_s z^sP(s)$:

$$\begin{aligned}f(z) = {} & z^c\sum_{s_1,\ldots,s_\ell} P(s_1)\ldots P(s_\ell) \\ & \times\sum_s\left(\frac{\alpha}{\ell}z\right)^{s-c}\left(1-\frac{\alpha}{\ell}\right)^{\sum s_i-s+c}\binom{\sum s_i}{s-c}.\end{aligned}$$

By the binomial theorem,

$$f(z)=z^c\sum_{s_1,\ldots,s_\ell} P(s_1)\ldots P(s_\ell)\left(1-\frac{\alpha}{\ell}+\frac{\alpha}{\ell}z\right)^{\sum s_i}.$$

The sum on the right hand side can be expressed as the product $\prod_{i=1}^{\ell}\sum_{s_i}(1-\alpha/\ell+\alpha z/\ell)^{s_i}P(s_i)$; thus,

$$f(z)=z^cf^\ell\left(1-\frac{\alpha}{\ell}+\frac{\alpha}{\ell}z\right).$$

To generalize for a nonconstant c, and since c is sampled from a fixed distribution P_c in each iteration, $P(s)$ will change to

$$P(s) = \sum_c P_c \sum_{s_1,\ldots,s_\ell} P(s_1)\ldots P(s_\ell) \times \left(\frac{\alpha}{\ell}\right)^{s-c}\left(1-\frac{\alpha}{\ell}\right)^{\sum s_i - s + c}\binom{\sum s_i}{s-c},$$

which, using the same strategy, will yield

$$f(z) = \sum_c P_c z^c f^\ell\left(1-\frac{\alpha}{\ell}+\frac{\alpha}{\ell}z\right) = g(z) f^\ell\left(1-\frac{\alpha}{\ell}+\frac{\alpha}{\ell}z\right),$$

where $g(z)$ is the generating function $\sum_c z^c P_c$.

2. Simplicial complex network

For simplicial complexes, we have to account in $E[m_{t,s}]$ for the expected number of facets of size s that are absorbed in each iteration, which is given by $\ell P(s)(\alpha/\ell)^s$ since only sampled facets can be absorbed given the no accidental absorption condition

$$E[m_{t,s}] = O(\tau) + (t-\tau)\sum_{s_1,\ldots,s_\ell} P(s_1)\ldots P(s_\ell) \times \left(\frac{\alpha}{\ell}\right)^{s-c}\left(1-\frac{\alpha}{\ell}\right)^{\sum s_i - s + c}\binom{\sum s_i}{s-c} - (t-\tau)\ell P(s)\left(\frac{\alpha}{\ell}\right)^s.$$

Using the exact same strategy above, and the fact that $E[m_t] = (t-\tau)[1-\ell f(\alpha/\ell)] + O(\tau)$, we obtain for a constant c

$$P(s)\left[1-\ell f\left(\frac{\alpha}{\ell}\right)+\ell\left(\frac{\alpha}{\ell}\right)^s\right] = \sum_{s_1,\ldots,s_\ell} P(s_1)\ldots P(s_\ell)\left(\frac{\alpha}{\ell}\right)^{s-c}\left(1-\frac{\alpha}{\ell}\right)^{\sum s_i - s + c}\binom{\sum s_i}{s-c}, \tag{B1}$$

which then yields

$$f(z) = \frac{g(z) f^\ell\left(1-\frac{\alpha}{\ell}+\frac{\alpha}{\ell}z\right) - \ell f\left(\frac{\alpha}{\ell}z\right)}{1-\ell f\left(\frac{\alpha}{\ell}\right)}.$$

3. Meganetwork (hypergraphs and simplicial complexes)

A meganode v is a node with a large degree, and the meganode network of v is the set of affiliations containing v (the notion of ego network). As explained in Sec. V, we assume that the affiliation size distribution in a meganetwork is $Q(s) = \lim_{k\to\infty} P(s|k)$. We will find the generating function $h(z) = \sum_s z^s Q(s)$.

For a given meganode v, we now let t count only iterations in which at least one of v's affiliations is among the ℓ sampled ones. In addition, and by the no overlap condition, exactly one affiliation is sampled from v's meganetwork in each of these "effective" iterations. Let M_t and $M_{t,s}$ be the expected number of affiliations, and those of size s, respectively, in the meganode network at iteration t.

In computing $E[M_{t,s}]$, the changes from the previous section are as follows [ignoring $O(\tau)$ terms]:

(i) The expression

$$\sum_{s_1,\ldots,s_\ell} P(s_1)\ldots P(s_\ell)\left(\frac{\alpha}{\ell}\right)^{s-c}\left(1-\frac{\alpha}{\ell}\right)^{\sum s_i - s + c}\binom{\sum s_i}{s-c}$$

is replaced by

$$\sum_{s_1} Q(s_1)\sum_{s_2,\ldots,s_\ell} P(s_2)\ldots P(s_\ell) \times \frac{\alpha}{\ell}\left(\frac{\alpha}{\ell}\right)^{s-c-1}\left(1-\frac{\alpha}{\ell}\right)^{\sum s_i - s + c}\binom{\sum s_i - 1}{s-c-1}.$$

(ii) For hypergraphs,

$$E[M_t] = \frac{\alpha}{\ell}t.$$

(iii) For simplicial complexes,

$$E[M_t] = t\left[\frac{\alpha}{\ell} - \sum_s \left(\frac{\alpha}{\ell}\right)^s Q(s)\right] = \frac{\alpha}{\ell}t\left[1-\frac{\ell}{\alpha}h\left(\frac{\alpha}{\ell}\right)\right].$$

(iv) For simplicial complexes, the expected number of facets of size s absorbed in the meganetwork in each iteration is (the probability that the one sampled facet has size s and is absorbed)

$$Q(s)\left(\frac{\alpha}{\ell}\right)^s = \frac{\alpha}{\ell}Q(s)\left(\frac{\alpha}{\ell}\right)^{s-1}.$$

Repeating the same work yields for the hypergraph meganetworks

$$h(z) = \frac{z g(z)}{1-\frac{\alpha}{\ell}+\frac{\alpha}{\ell}z} h\left(1-\frac{\alpha}{\ell}+\frac{\alpha}{\ell}z\right) f^{\ell-1}\left(1-\frac{\alpha}{\ell}+\frac{\alpha}{\ell}z\right),$$

and for the simplicial complex meganetworks

$$\left[1-\frac{\ell}{\alpha}h\left(\frac{\alpha}{\ell}\right)\right]h(z) + \frac{\ell}{\alpha}h\left(\frac{\alpha}{\ell}z\right) = \frac{z g(z)}{1-\frac{\alpha}{\ell}+\frac{\alpha}{\ell}z} h\left(1-\frac{\alpha}{\ell}+\frac{\alpha}{\ell}z\right) f^{\ell-1}\left(1-\frac{\alpha}{\ell}+\frac{\alpha}{\ell}z\right).$$

APPENDIX C: THE ITERATIVE DIFFICULTY FOR $f(z)$ IN SIMPLICIAL COMPLEXES WHEN $\ell = 1$

Recall the generating function for the facet size distribution in simplicial complexes given by Eq. (11) when c is constant [$g(z) = z^c$]:

$$f(z) = \frac{z^c f^\ell\left(1-\frac{\alpha}{\ell}+\frac{\alpha}{\ell}z\right) - \ell f\left(\frac{\alpha}{\ell}z\right)}{1-\ell f\left(\frac{\alpha}{\ell}\right)}.$$

If we start with $f(z) = z^q$ for $q \geqslant c \geqslant 1$, then iterating once gives

$$\hat{f}(z) = \frac{z^c\left(1-\frac{\alpha}{\ell}+\frac{\alpha}{\ell}z\right)^{q\ell} - \ell\left(\frac{\alpha}{\ell}z\right)^q}{1-\ell\left(\frac{\alpha}{\ell}\right)^q}.$$

We need $\hat{f}(z) \geqslant 0$. One can prove that when $\ell > 1$, there exists $q > c$ such that for every $0 \leqslant \alpha < 1$, $\hat{f}(z) \geqslant 0$ for $0 \leqslant z \leqslant 1$ (the proof is not included here). However, when $\ell = 1$,

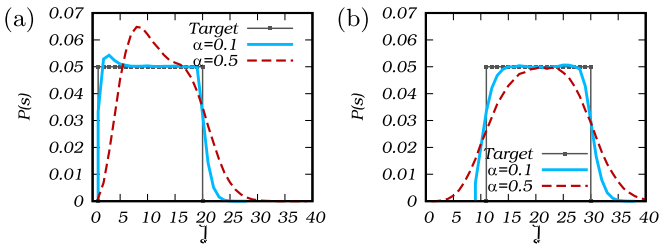

FIG. 14. Mimicking of uniform distributions (a) $P^*(s) = U_{[1,20]}$, (b) $P^*(s) = U_{[11,30]}$ in a simplicial complex with $\ell = 2$ and number of nodes $n = 10^6$.

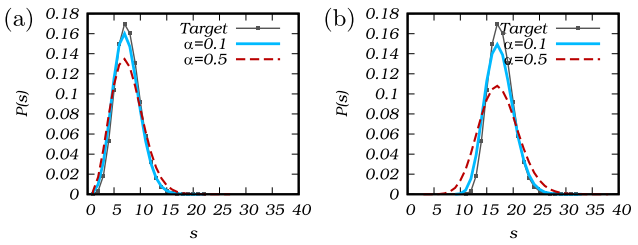

FIG. 15. Mimicking of distributions generated by $\alpha = 0.75$, $c = 2$, and $\ell = 8$ in a simplicial complex with $\ell = 2$ and number of nodes $n = 10^6$: (a) the generated distribution, (b) the same generated distribution shifted by 10.

no such q exists (a large α becomes problematic as shown below). This reflects one aspect of the inherent difficulty of the iterative approach when $\ell = 1$.

Since the denominator is always positive when $0 \leqslant \alpha < 1$, $\ell \geqslant 1$, and $q \geqslant 1$, the requirement for $\hat{f}(z) \geqslant 0$ is equivalent to the numerator being non-negative. This is expressed below in logarithm space for $\ell = 1$:

$$q[\ln(1-\alpha+\alpha z) - \ln(\alpha z)] \geqslant \ln \frac{1}{z^c}.$$

For every fixed q and z, there is a large enough $\alpha < 1$ that makes the left hand side of the inequality arbitrarily close to 0. Since $\ln \frac{1}{z^c}$ is not affected by α, no fixed value for q will satisfy the inequality for every α.

APPENDIX D: TARGETED AFFILIATION SIZE DISTRIBUTION

Given a target affiliation size distribution $P^*(s)$ with mean μ, we describe how to make a best effort to approximately mimic it through the choice of c (but not necessarily achieve it theoretically): We first sample s^* from $P^*(s)$, and then set $c = \max(1, [s^* - \alpha\mu])$, where $[\dots]$ denotes the closest integer function.

Assume that $P^*(s^* \leqslant \alpha\mu)$ is small; for instance, if s^* is subexponential, then $P^*(s^* \leqslant \alpha\mu) = P^*(s^* - \mu \leqslant -(1-\alpha)\mu) \leqslant e^{-\Theta((1-\alpha)\mu)}$ when $(1-\alpha)\mu \gg 1$ and is, therefore, small. Then,

$$\begin{aligned} E[c] &\approx \sum_{s^* > \alpha\mu} (s^* - \alpha\mu) P^*(s^*) + \sum_{s^* \leqslant \alpha\mu} 1 \cdot P^*(s^*) \\ &= \sum_{s^*} (s^* - \alpha\mu) P^*(s^*) + \sum_{s^* \leqslant \alpha\mu} (\alpha\mu - s^* + 1) P^*(s^*) \\ &\approx (1-\alpha)\mu, \end{aligned}$$

which follows from that the last sum is upper bounded by $\alpha\mu P^*(s^* \leqslant \alpha\mu) \leqslant \alpha\mu e^{-\Theta((1-\alpha)\mu)}$. Therefore, Eq. (9) for a hypergraph gives $E[s] = E[c]/(1-\alpha) \approx \mu$.

But what about $P(s)$? For simplicity, we consider the case of large ℓ.

Since ℓ is large, the size of an affiliation prior to the addition of c nodes is a Poisson random variable with parameter $\alpha E[s] \approx \alpha\mu$ (Sec. VII). Therefore, the size of a newly created affiliation is given by the random variable

$$s = c + \text{Pois}(\alpha\mu) \approx s^* - \alpha\mu + \text{Pois}(\alpha\mu).$$

If s^* has high concentration around μ and $|-\alpha\mu + \text{Pois}(\alpha\mu)| \ll \mu$, then s is close to s^*. Indeed, for any small ϵ, a subexponential s^* is highly concentrated and has a small probability $P^*(|s^* - \mu| \geqslant \epsilon\mu) \leqslant 2e^{-\Theta(\epsilon\mu)}$ when $\epsilon\mu \gg 1$.

If $X \sim \text{Pois}(\alpha\mu)$, then a standard Chernoff bound gives for $0 \leqslant \epsilon \leqslant \alpha$:

$$P(|X - \alpha\mu| \leqslant \epsilon\mu) = P(|X - \alpha\mu| \leqslant \delta\alpha\mu) \geqslant 1 - 2e^{\Theta(\delta^2\alpha\mu)},$$

where $\delta = \epsilon/\alpha \leqslant 1$. This makes $|-\alpha\mu + \text{Pois}(\alpha\mu)| \ll \mu$ with high probability when ϵ is small and $\delta^2\alpha\mu = \epsilon^2 \mu/\alpha \gg 1$.

The above analysis suggests that one could reasonably mimic $P^*(s)$ using $c = \max(1, [s^* - \alpha\mu])$ where $s^* \sim P^*(s)$ has high concentration around $\mu = E[s^*]$ if μ is large or if α is small. Experimentally, Fig. 14 shows the mimicking of uniform distributions, and Fig. 15 shows the mimicking of distributions generated by the algorithm itself for different parameters.

[1] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, Science **286**, 509 (1999).

[2] G. U. Yule, Ii.-a mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S, Philos. Trans. R. Soc. London B **213**, 21 (1925).

[3] H. A. Simon, On a class of skew distribution functions, Biometrika **42**, 425 (1955).

[4] D. d. S. Price, A general theory of bibliometric and other cumulative advantage processes, J. Am. Soc. Inf. Sci. **27**, 292 (1976).

[5] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády, The degree sequence of a scale-free random graph process, Random Struct. Algorithms **18**, 279 (2001).

[6] B. Bollobás and O. M. Riordan, Mathematical results on scale-free random graphs, *Handbook of Graphs and Networks: From the Genome to the Internet* (Wiley-VCH, Weinhein, 2003), pp. 1–34.

[7] B. Bollobás and O. Riordan, The diameter of a scale-free random graph, Combinatorica **24**, 5 (2004).

[8] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Structure of Growing Networks with Preferential Linking, Phys. Rev. Lett. **85**, 4633 (2000).

[9] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, Resilience of the Internet to Random Breakdowns, Phys. Rev. Lett. **85**, 4626 (2000).

[10] P. Erdős and A. Rényi, On random graphs, i, Publ. Math. Debrecen **6**, 290 (1959).

[11] P. Erdős and A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci **5**, 17 (1960).

[12] E. N. Gilbert, Random graphs, Ann. Math. Stat. **30**, 1141 (1959).

[13] Z. Xie, M. Li, J. Li, X. Duan, and Z. Ouyang, Feature analysis of multidisciplinary scientific collaboration patterns based on PNAS, EPJ Data Sci. **7**, 5 (2018).

[14] A. Nikolaev, Modeling and analysis of affiliation networks with subsumption, Ph.D. thesis, the City University of New York, 2021.

[15] Stanford network analysis project (snap), https://snap.stanford.edu/data/ego-Facebook.html.

[16] M. P. Allen, The identification of interlock groups in large corporate networks: Convergent validation using divergent techniques, Social Networks **4**, 349 (1982).

[17] M. S. Mizruchi, What do interlocks do? an analysis, critique, and assessment of research on interlocking directorates, Annu. Rev. Sociol. **22**, 271 (1996).

[18] G. F. Davis and H. R. Greve, Corporate elite networks and governance changes in the 1980s, Am. J. Sociol. **103**, 1 (1997).

[19] D. M. Griffith, J. A. Veech, C. J. Marsh *et al.*, Cooccur: probabilistic species co-occurrence analysis in R, J. Stat. Software **69**, 1 (2016).

[20] J. G. Sanderson, Testing ecological patterns, Am. Sci. **88**, 332 (2000).

[21] E. Estrada and G. J. Ross, Centralities in simplicial complexes. applications to protein interaction networks, J. Theor. Biol. **438**, 46 (2018).

[22] N. Malod-Dognin and N. Przulj, Functional geometry of protein-protein interaction networks, arXiv:1804.04428.

[23] M. W. Reimann, M. Nolte, M. Scolamiero, K. Turner, R. Perin, G. Chindemi, P. Dłotko, R. Levi, K. Hess, and H. Markram, Cliques of neurons bound into cavities provide a missing link between structure and function, Front. Comput. Neurosci. **11**, 48 (2017).

[24] W. Ren, Q. Zhao, R. Ramanathan, J. Gao, A. Swami, A. Bar-Noy, M. P. Johnson, and P. Basu, Broadcasting in multi-radio multi-channel wireless networks using simplicial complexes, Wireless Networks **19**, 1121 (2013).

[25] R. Ghrist and A. Muhammad, Coverage and hole-detection in sensor networks via homology, in *Proceedings of the 4th International Symposium on Information Processing in Sensor Networks* (IEEE, Piscataway, NJ, 2005), p. 34.

[26] V. de Silva and R. Ghrist, Coordinate-free coverage in sensor networks with controlled boundaries via homology, Int. J. Robotics Res. **25**, 1205 (2006).

[27] V. De Silva and R. Ghrist, Homological sensor networks, Not. Am. Math. Soc. **54**, 10 (2007).

[28] J. Kanno, J. G. Buchart, R. R. Selmic, and V. Phoha, Detecting coverage holes in wireless sensor networks, in *2009 17th Mediterranean Conference on Control and Automation* (IEEE, Piscataway, NJ, 2009), pp. 452–457.

[29] H. Chintakunta and H. Krim, Divide and conquer: Localizing coverage holes in sensor networks, in *2010 7th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)* (IEEE, Piscataway, NJ, 2010), pp. 1–8.

[30] J. Johnson, Complexity science in collaborative design, CoDesign **1**, 223 (2005).

[31] T. J. Moore, R. J. Drost, P. Basu, R. Ramanathan, and A. Swami, Analyzing collaboration networks using simplicial complexes: A case study, in *2012 Proceedings IEEE INFOCOM Workshops* (IEEE, Piscataway, NJ, 2012), pp. 238–243.

[32] C. J. Carstens and K. J. Horadam, Persistent homology of collaboration networks, Math. Problems Eng. **2013**, 815035 (2013).

[33] M. X. Hoang, R. Ramanathan, and A. K. Singh, Structure and evolution of missed collaborations in large networks, in *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (IEEE, Piscataway, NJ, 2014), pp. 849–854.

[34] M. Bampasidou and T. Gentimis, Modeling collaborations with persistent homology, arXiv:1403.5346.

[35] S. Pal, T. J. Moore, R. Ramanathan, and A. Swami, Comparative topological signatures of growing collaboration networks, in *International Workshop on Complex Networks* (Springer, Berlin, 2017), pp. 201–209.

[36] B. Jhun, M. Jo, and B. Kahng, Simplicial sis model in scale-free uniform hypergraph, J. Stat. Mech. (2019) 123207.

[37] I. Iacopini, G. Petri, A. Barrat, and V. Latora, Simplicial models of social contagion, Nat. Commun. **10**, 2485 (2019).

[38] S. Maletić and M. Rajković, Consensus formation on a simplicial complex of opinions, Phys. A (Amsterdam) **397**, 111 (2014).

[39] E. N. Ciftcioglu, R. Ramanathan, and P. Basu, Generative models for global collaboration relationships, Sci. Rep. **7**, 11160 (2017).

[40] B. Guo, H. He, Z. Yu, D. Zhang, and X. Zhou, Groupme: Supporting group formation with mobile sensing and social graph mining, in *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services* (Springer, Berlin, 2012), pp. 200–211.

[41] D. MacLean, S. Hangal, S. K. Teh, M. S. Lam, and J. Heer, Groups without tears: mining social topologies from email, in *Proceedings of the 16th International Conference on Intelligent User Interfaces* (ACM, New York, 2011), pp. 83–92.

[42] A. Carzaniga, D. S. Rosenblum, and A. L. Wolf, Achieving scalability and expressiveness in an internet-scale event notification service, in *Proceedings of the nineteenth annual ACM symposium on Principles of Distributed Computing* (ACM, New York, 2000), pp. 219–227.

[43] P. Triantafillou and A. Economides, Subscription summarization: A new paradigm for efficient publish/subscribe systems, in *Proceedings of 24th International Conference on Distributed Computing Systems, 2004* (IEEE, Piscataway, NJ, 2004), pp. 562–571.

[44] Z. Wu, G. Menichetti, C. Rahmede, and G. Bianconi, Emergent complex network geometry, Sci. Rep. **5**, 10073 (2015).

[45] O. T. Courtney and G. Bianconi, Dense power-law networks and simplicial complexes, Phys. Rev. E **97**, 052303 (2018).

[46] L. Hébert-Dufresne, A. Allard, V. Marceau, P.-A. Noël, and L. J. Dubé, Structural preferential attachment: Stochastic process for

the growth of scale-free, modular, and self-similar systems, Phys. Rev. E **85**, 026108 (2012).
[47] T. Opsahl, Triadic closure in two-mode networks: Redefining the global and local clustering coefficients, Social Networks **35**, 159 (2013).
[48] E. Estrada and J. A. Rodríguez-Velázquez, Subgraph centrality and clustering in complex hyper-networks, Phys. A (Amsterdam) **364**, 581 (2006).
[49] M. E. J. Newman, Assortative Mixing in Networks, Phys. Rev. Lett. **89**, 208701 (2002).
[50] A. Kulig, S. Drożdż, J. Kwapień, and P. Oświecimka, Modeling the average shortest-path length in growth of word-adjacency networks, Phys. Rev. E **91**, 032810 (2015).
[51] S. Chechik, E. Cohen, and H. Kaplan, Average distance queries through weighted samples in graphs and metric spaces: High scalability with tight statistical guarantees, arXiv:1503.08528.
[52] J. Kunegis, Konect: the koblenz network collection, in *Proceedings of the 22nd International Conference on World Wide Web* (ACM, New York, 2013), pp. 1343–1350.
[53] Corporate leaderships, http://www.konect.cc/networks/brunson_corporate-leadership/.
[54] R. Barnes and T. Burkett, Structural redundancy and multiplicity in corporate networks, Connections **30**, 4 (2010).
[55] Konect: Amazon wang, http://www.konect.cc/networks/wang-amazon/.
[56] H. Wang, Y. Lu, and C. Zhai, Latent aspect rating analysis on review text data: a rating regression approach, in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2010), pp. 783–792.
[57] Konect: Movielens 100k, http://www.konect.cc/networks/movielens-100k_rating/.
[58] Konect: Grouplens movielens, https://grouplens.org/datasets/movielens/.
[59] Konect: Filmtrust ratings, http://www.konect.cc/networks/librec-filmtrust-ratings/.
[60] G. Guo, J. Zhang, and N. Yorke-Smith, A novel evidence-based bayesian similarity measure for recommender systems, ACM Trans. Web (TWEB) **10**, 1 (2016).
[61] Konect: Crime, http://www.konect.cc/networks/moreno_crime/.
[62] Konect: arxiv cond-mat, http://www.konect.cc/networks/opsahl-collaboration/.
[63] M. E. Newman, The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. USA **98**, 404 (2001).
[64] Konect: Dbpedia writers, http://www.konect.cc/networks/dbpedia-writer/.
[65] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, Dbpedia: A nucleus for a web of open data, in *Proceedings of The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007* (Springer, Berlin, 2007), pp. 722–735.
[66] Konect: Dbpedia producers, http://www.konect.cc/networks/dbpedia-producer/.
[67] Konect: Us airports, http://www.konect.cc/networks/opsahl-usairport/.
[68] T. Opsahl, Why anchorage is not (that) important: Binary ties and sample selection, https://toreopsahl.com/2011/08/12/.
[69] Konect: Enron, http://www.konect.cc/networks/enron/.
[70] B. Klimt and Y. Yang, The enron corpus: A new dataset for email classification research, in *Machine Learning: ECML 2004: Proceedings of the 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004* (Springer, Berlin, 2004), pp. 217–226.
[71] Konect:http://www.konect.cc/networks/cit-HepPh/.
[72] J. Leskovec, J. Kleinberg, and C. Faloutsos, Graph evolution: Densification and shrinking diameters, ACM Trans. Knowl. Discov. Data **1**, 2 (2007).
[73] Konect: Zachary karate club, http://www.konect.cc/networks/ucidata-zachary/.
[74] W. W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. **33**, 452 (1977).
[75] Konect: Internet topology, http://www.konect.cc/networks/topology/.
[76] B. Zhang, R. Liu, D. Massey, and L. Zhang, Collecting the internet as-level topology, ACM SIGCOMM Comput. Commun. Rev. **35**, 53 (2005).
[77] Konect: Caenorhabditis elegans, http://www.konect.cc/networks/arenas-meta/.
[78] J. Duch and A. Arenas, Community detection in complex networks using extremal optimization, Phys. Rev. E **72**, 027104 (2005).
[79] Konect: Yeast, http://www.konect.cc/networks/moreno_propro/.
[80] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, Lethality and centrality in protein networks, Nature (London) **411**, 41 (2001).
[81] M. Newman, Power laws, pareto distributions and zipf's law, Contemp. Phys. **46**, 323 (2005).
[82] A. Clauset, C. Shalizi, and M. Newman, Power-law distributions in empirical data, SIAM Rev. **51**, 661 (2009).
[83] plfit: Fitting power-law distributions to empirical data, according to the method of clauset, shalizi and newman, https://github.com/ntamas/plfit.
[84] https://github.com/a-nikolaev/gen-networks.