

Exponential random simplicial complexes

Konstantin Zuev¹, Or Eisenberg and Dmitri Krioukov

Department of Physics, Department of Mathematics, Department of Electrical & Computer Engineering, Northeastern University, Boston, MA 02115, USA

E-mail: k.zuev@neu.edu, eisenberg.o@husky.neu.edu and dima@neu.edu

Received 18 June 2015, revised 24 August 2015

Accepted for publication 5 October 2015

Published 26 October 2015



CrossMark

Abstract

Exponential random graph models have attracted significant research attention over the past decades. These models are maximum-entropy ensembles subject to the constraints that the expected values of a set of graph observables are equal to given values. Here we extend these maximum-entropy ensembles to random simplicial complexes, which are more adequate and versatile constructions to model complex systems in many applications. We show that many random simplicial complex models considered in the literature can be casted as maximum-entropy ensembles under certain constraints. We introduce and analyze the most general random simplicial complex ensemble Δ with statistically independent simplices. Our analysis is simplified by the observation that any distribution $\mathbb{P}(\mathcal{O})$ on any collection of objects $\mathcal{O} = \{O\}$, including graphs and simplicial complexes, is maximum-entropy subject to the constraint that the expected value of $-\ln \mathbb{P}(\mathcal{O})$ is equal to the entropy of the distribution. With the help of this observation, we prove that ensemble Δ is maximum-entropy subject to the two types of constraints which fix the expected numbers of simplices and their boundaries.

Keywords: random simplicial complexes, random graphs, maximum-entropy distributions, exponential random graphs model, network models

(Some figures may appear in colour only in the online journal)

1. Introduction

When studying complex systems consisting of many interconnected, interacting components, it is rather natural to represent the system as a graph or, more generally, as a simplicial complex. Modeling complex systems with graphs has proved to be useful for understanding

¹ Author to whom any correspondence should be addressed.

systems as intricate as the Internet, the human brain, and interwoven social groups, and has led to a new area of research, called network science [12, 14, 35].

A host of developed network models (e.g. see [21] for a survey) can be roughly divided into two classes: ‘generative’ models and ‘descriptive’ models [1]. Generative models are algorithms which describe how to generate a network using some probabilistic rules for connecting nodes. These models primarily aim to uncover the hidden evolution mechanisms responsible for certain properties observed in real networks. A classical, and perhaps the simplest and best studied, example of a generative model is the Erdős–Rényi random graph $G(n, p)$ [17, 18, 43]: given n nodes, place a link between every two nodes independently at random with probability p . Among other prominent examples are the preferential attachment model [4, 13, 30] and the small-world model [36, 47, 48] which explain the power-law degree distributions and small distances between most nodes, two universal properties observed in many real networks. Any generative model gives rise to an ensemble $(\mathcal{G}, \mathbb{P})$, where \mathcal{G} is a set of all graphs the model can possibly generate and \mathbb{P} is the probability distribution on \mathcal{G} , where $\mathbb{P}(G)$ is the probability that the model generates $G \in \mathcal{G}$. One can always readily sample from \mathbb{P} (using the network generating algorithm), but often cannot obtain a closed-form expression for $\mathbb{P}(G)$, or even implicitly describe \mathbb{P} as a solution of some optimization problem equation.

Generative models can help to understand the fundamental organizing principles behind real networks and explain their qualitative behavior, but they are not specifically designed for network data analysis. Descriptive models attempt to fill this gap. A descriptive model is explicitly defined as an ensemble $(\mathcal{G}, \mathbb{P}_\theta)$, where \mathcal{G} is a set of graphs and \mathbb{P}_θ is the joint probability distribution on \mathcal{G} parameterized by a vector of parameters θ , which are to be inferred from the observed network data. For any graph $G \in \mathcal{G}$, a descriptive model gives a closed-form expression for $\mathbb{P}_\theta(G)$ which can be used for further statistical inference, e.g. for estimating ensemble averages $\sum_{G \in \mathcal{G}} x(G) \mathbb{P}_\theta(G)$, where x is a network property of interest. In contrast to generative models, however, a descriptive model does not specify how to sample networks from \mathbb{P}_θ , which is often a challenging task. In simple cases, a network model can be represented as both generative and descriptive model. For example, the Erdős–Rényi random graph $G(n, p)$ can be defined, as above, by a generative algorithm, or by the formula for the probability distribution $\mathbb{P}(G) = p^{f_1(G)} (1-p)^{\binom{n}{2} - f_1(G)}$, where $f_1(G)$ is the number of edges in G . In general, however, representing a generative model as descriptive (and vice versa) is a very difficult problem whose solution could be very useful for applications.

Exponential random graphs (ERGs) [19, 23, 29, 37, 42, 45], often called p^* models in the social network research community [3, 40, 46], are among the most popular and best studied descriptive models which provide a conceptual framework for statistical modeling of network data. Let \mathcal{G}_n be the set of all simple graphs (without self-loops or multi-edges) with n nodes, x_1, \dots, x_r be functions on \mathcal{G}_n , henceforth referred to as the graph observables, and let $\bar{x}_1, \dots, \bar{x}_r$ be the values of these observables $x_1(\bar{G}), \dots, x_r(\bar{G})$ for a network of interest $\bar{G} \in \mathcal{G}_n$ computed from available network data. The ERG model defined by \bar{G} and its observables $\bar{x}_1, \dots, \bar{x}_r$ is the exact analog of the Boltzmann distribution in statistical mechanics:

$$\mathbb{P}_\theta(G) = \frac{e^{-H_\theta(G)}}{Z(\theta)}, \quad H_\theta(G) = \sum_{i=1}^r \theta_i x_i(G), \quad (1.1)$$

where $H_\theta(G)$ is called the graph Hamiltonian, $Z(\theta)$ the partition function (the normalization constant), and $\theta = (\theta_1, \dots, \theta_r)$ is a vector of model parameters which satisfy

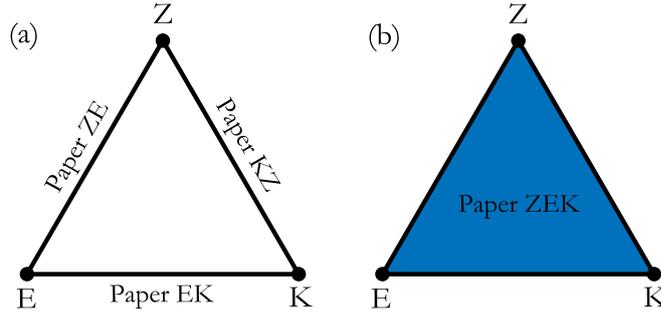


Figure 1. Networks versus simplicial complexes. The outcome of collaboration between scientists Z , E , and K could be three different papers co-authored by Z and E , E and K , and K and Z (panel (a)), or a single paper co-authored by all three scientists (panel (b)). While the network representation does not distinguish between these two cases and results in the graph in panel (a), the simplicial complex representation does by adding in the latter case the triangle $\{Z, E, K\}$ in panel (b).

$$-\frac{\partial \ln Z}{\partial \theta_i} = \bar{x}_i. \quad (1.2)$$

Whereas originally (1.1) was simply postulated and used in empirical studies [23], it was later recognized [19, 37, 45] that ERGs are maximum-entropy ensembles. Namely, the distribution defined by (1.1) and (1.2) maximizes the Gibbs entropy

$$S(\mathbb{P}) = -\sum_{G \in \mathcal{G}_n} \mathbb{P}(G) \ln \mathbb{P}(G), \quad (1.3)$$

subject to the r ‘soft’ constraints and the normalization condition

$$\mathbb{E}_{\mathbb{P}}[x_i] = \sum_{G \in \mathcal{G}_n} x_i(G) \mathbb{P}(G) = \bar{x}_i, \quad (1.4)$$

$$\sum_{G \in \mathcal{G}_n} \mathbb{P}(G) = 1. \quad (1.5)$$

The general principle of maximum entropy is thoroughly reviewed in [38]. In the context of complex networks, the principle of maximum entropy and different entropy measures are discussed in [2]. Despite some known problems with ERGs with nonlinearly correlated constraints [7, 24, 41], ERGs remain one of the most popular descriptive models for network data analysis, especially in social science.

In many cases, however, representing a complex system with a simplicial complex—a higher-dimensional analog of a graph—is conceptually more sound than the basic network representation, and provides a ‘higher order approximation’ of the system. Consider for example a social system of scientific collaboration. Three researchers may co-author a single article or they may have three different papers with two authors each. The network representation, where nodes are connected if the corresponding scientists co-authored a paper, will not distinguish between these two cases. But we can do this by placing (in the former case), or not (in the latter case), a two-simplex on the three nodes. This is illustrated in figure 1. Other examples, where the simplicial complex representation is more accurate include biological protein-interaction systems, where proteins form protein complexes often consisting of more than two proteins, economic systems of financial transactions often involving several parties,

and social systems, where groups of people are united by a common motive, interest, or goal, as opposed to merely being pairwise connected.

In general, compared to graphs, simplicial complexes encode more relevant information about a complex system, and make possible modeling beyond dyadic interactions. They have been used in many applications, including modeling social aggregation [28], agent interaction [44], opinion formation and dynamics [32, 33], coverage and hole-detection in sensor networks [20], and broadcasting in wireless networks [39], to name just a few. We remark that prior to their being used for studying complex interactions, simplicial complexes were used in a rich variety of geometric problems, ranging from grid generation in finite element analysis to modeling configuration spaces of dynamical systems [15]. Further details and applications can be found in [16].

In this paper, we introduce *exponential random simplicial complexes* (ERSCs) which are higher dimensional generalizations of ERGs, develop the formalism for ERSCs, and show that several popular generative models of random simplicial complexes—random flag complexes [25], Linial–Meshulam complexes [31], and Kahle’s multi-parameter model [27]—can all be explicitly represented as ERSCs. We also introduce the most general ensemble of random simplicial complexes Δ with statistically independent simplices, and show that this ensemble is an ERSC ensemble as well.

2. Basic definitions and notations

Here we recall a few basic definitions and introduce notation that we use throughout the paper. For a comprehensive reference on simplicial complexes the reader is referred to [34].

A *simplicial complex* C on n vertices $V = \{1, \dots, n\}$ is a collection of non-empty subsets of V , called simplices. Complex C contains all vertices, $\{i\} \in C$, and is closed under the subset relation: if $\sigma \in C$ and $\tau \subset \sigma$, then $\tau \in C$, where τ is called a *face* of simplex σ , and σ is a *coface* of τ . A simplex σ is called a *k-simplex* of dimension k if its cardinality is $|\sigma| = k + 1$. It is useful to think of a k -simplex as the convex hull of $(k + 1)$ points in general position in \mathbb{R}^K , $K \geq k$ [22]. For instance, 0-, 1-, 2-, and 3-simplices are, respectively, vertices, edges, triangles, and tetrahedra. A simplicial complex is then a collection of simplices of different dimensions properly glued together. We say that C has dimension m if it has at least one m -simplex, but does not have simplices of higher dimension. Clearly, $m \leq n - 1$.

Let \mathcal{C}_n be the set of all simplicial complexes on n vertices. By analogy with graphs, where there exists a one-to-one correspondence between \mathcal{G}_n and the set all boolean symmetric n -by- n matrices with zeros on the diagonal, known as adjacency matrices, we can represent \mathcal{C}_n by a tensor product

$$\mathcal{C}_n = \bigotimes_{d=1}^n \mathbf{a}_d, \quad (2.1)$$

where $\mathbf{a}_d = \{a_{i_1, \dots, i_d}\}$, $i_j = 1, \dots, n$, $j = 1, \dots, d$, is a boolean symmetric tensor of order d with zeros on all its diagonals. These conditions require precisely that $a_{i_1, \dots, i_d} = a_{i_{\kappa(1)}, \dots, i_{\kappa(d)}}$ for any permutation κ of subindices $1, \dots, d$, and $a_{i_1, \dots, i_d} = 0$ if $i_j = i_k$ for any pair of j and k . The non-redundant elements of tensor \mathbf{a}_d are thus $a_{\mathbf{i}_d}$, where multi-index \mathbf{i}_d denotes a d -tuple of indices with increasing values:

$$\mathbf{i}_d = i_1, \dots, i_d, \quad (2.2)$$

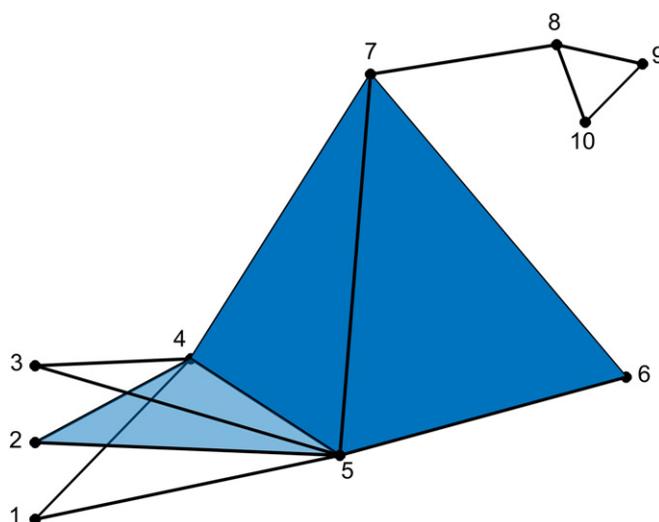


Figure 2. Simplicial complex and its adjacency tensors. In this example, $C \in \mathcal{C}_{10}$, $\dim C = 3$, and the non-zero elements a_{i_d} of adjacency tensor \mathbf{a}_d , $d = 1, 2, 3, 4$, are: $a_{i_1} = 1$ for all $i_1 = 1, \dots, 10$; $a_{i_2} = 1$ for $i_2 = (1, 4), (1, 5), (2, 4), (2, 5), (3, 4), (3, 5), (4, 5), (4, 6), (4, 7), (5, 6), (5, 7), (6, 7), (7, 8), (8, 9), (8, 10),$ and $(9, 10)$; $a_{i_3} = 1$ for $i_3 = (2, 4, 5), (4, 5, 6), (4, 5, 7), (4, 6, 7),$ and $(5, 6, 7)$; $a_{i_4} = 1$ only for $i_4 = (4, 5, 6, 7)$. The edge $\{4, 6\}$ is not visible because of the three-simplex $\{4, 5, 6, 7\}$.

$$1 \leq i_1 < \dots < i_d \leq n. \tag{2.3}$$

The only requirement for $\otimes_{d=1}^n \mathbf{a}_d$ to be in bijection with \mathcal{C}_n is then the following compatibility condition:

$$a_{i_d} = 1 \Rightarrow b_{i_d} \stackrel{\text{def}}{=} \prod_{k=1}^d a_{i_d^k} = 1, \quad \text{where} \tag{2.4}$$

$$\mathbf{i}_d^{\hat{k}} = i_1, \dots, \hat{i}_k, \dots, i_d \tag{2.5}$$

is the $(d - 1)$ -long multi-index obtained from multi-index \mathbf{i}_d by omitting index i_k . It is useful to think of $\mathbf{i}_d^{\hat{k}}$ as the result of operation $(\cdot)^{\hat{k}}$, which is the deletion of the k th index, applied to multi-index \mathbf{i}_d . Condition (2.4) simply formalizes the requirement that if the complex contains simplex $\{\mathbf{i}_d\}$, then it also contains all its faces.

For a simplicial complex $C \in \mathcal{C}_n$, $\mathbf{a}_d = \{a_{i_d}\}$ is thus its ‘adjacency’ tensor that encodes the presence of $(d - 1)$ -simplices: $a_{i_d} = 1$ if $\{\mathbf{i}_d\} \in C$, and zero otherwise. Since we assume that C has n vertices, we trivially have $\mathbf{a}_1 = \mathbf{1}_n = (1, \dots, 1)$. Figure 2 illustrates the correspondence between simplicial complexes and their adjacency tensors.

A subcomplex of C is a subset $C' \subset C$ that is also a simplicial complex. The d -skeleton of C , denoted $C^{(d)}$, is a subcomplex consisting of all k -simplices of C with $k \leq d$. The one-skeleton of a simplicial complex, for example, is a graph.

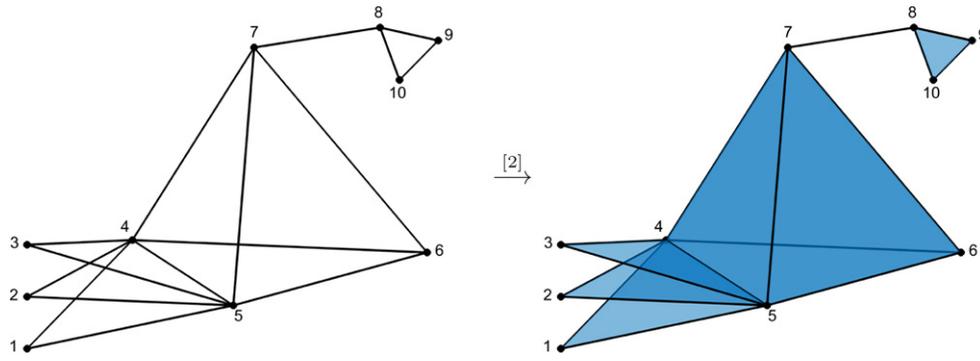


Figure 3. Empty and filled skeletons. The left simplicial complex is the one-skeleton $C^{(1)}$ of the complex C in figure 2. The filled one-skeleton $C^{[2]}$ on the right is obtained by adding all two-simplices based on triangular subgraphs of $C^{(1)}$. The three-simplex $\{4, 5, 6, 7\} \in C$ does not belong to $C^{[2]}$. The filled skeleton $C^{[2]}$ is not a subcomplex of C since, for example, $\{8, 9, 10\} \notin C$. The filling operation is denoted by $\xrightarrow{[2]}$.

Definition 1. The filled d -skeleton, denoted $C^{[d+1]}$, is a simplicial complex

$$C^{[d+1]} = C^{(d)} \cup \{ \{ \mathbf{i}_{d+2} \} : b_{\mathbf{i}_{d+2}} = 1 \}. \tag{2.6}$$

In other words, $C^{[d+1]}$ is obtained from $C^{(d)}$ by adding $(d + 1)$ -simplices as follows. For every $(d + 1)$ -simplex $\{ \mathbf{i}_{d+2} \}$, if $C^{(d)}$ contains all $(d + 2)$ d -simplices $\{ \mathbf{i}_{d+2}^k \}$, $k = 1, \dots, d + 2$, we add $\{ \mathbf{i}_{d+2} \}$ to $C^{(d)}$. Intuitively, we add $\{ \mathbf{i}_{d+2} \}$ if its d -dimensional boundary is already in $C^{(d)}$. Note that in this case we add $\{ \mathbf{i}_{d+2} \}$ even if $\{ \mathbf{i}_{d+2} \} \notin C$, and, therefore, $C^{[d+1]}$ is not necessarily a subcomplex of C . For example, $C^{[1]}$ is a complete graph on n vertices, and $C^{[2]}$ is the one-skeleton of C with all its triangular subgraphs filled by two-simplices. We denote the filled d -skeleton by $C^{[d+1]}$ (instead of $C^{[d]}$), to emphasize that generally it has dimension $(d + 1)$. Figure 3 illustrates the construction of a filled skeleton.

Thus, we have the following hierarchy of ‘empty’ and ‘filled’ skeletons:

$$\begin{array}{ccccccc}
 & & C^{[1]} & & C^{[d]} & & C^{[m]} \\
 & \nearrow [1] & \cup \! \! \! \cup & & \cup \! \! \! \cup & & \cup \! \! \! \cup \\
 V = C^{(0)} & \subset & C^{(1)} \subset \dots \subset C^{(d-1)} & \subset & C^{(d)} \subset \dots \subset C^{(m-1)} & \subset & C^{(m)} = C,
 \end{array}$$

where $\xrightarrow{[d]}$ denotes the filling operation. Let f_d denote the number of d -simplices in $C^{(d)}$ (and therefore in C), and ϕ_d be the number of d -simplices in $C^{[d]}$. By construction, $\phi_d \geq f_d$, and

$$f_d = \sum_{\mathbf{i}_{d+1}} a_{\mathbf{i}_{d+1}} \text{ and } \phi_d = \sum_{\mathbf{i}_{d+1}} b_{\mathbf{i}_{d+1}}. \tag{2.7}$$

Figure A1 shows all simplicial complexes $C \in \mathcal{C}_3$ and the values of f_1, f_2 , and ϕ_2 for each C .

3. Exponential random simplicial complexes

Let \mathcal{S} be any subset of $\mathcal{C}_n, \{x_1, \dots, x_r\}$ be a set of functions on $\mathcal{S}, x_i: \mathcal{S} \rightarrow \mathbb{R}$, and $\{\bar{x}_1, \dots, \bar{x}_r\}$ be a set of numbers, $\bar{x}_i \in \mathbb{R}$. We define the ERSC as a maximum-entropy ensemble of complexes with ‘soft’ constraints that require the observables x_i to have the expected values \bar{x}_i in the ensemble.

Definition 2. An ERSC $(\mathcal{S}, \{x_i\}, \{\bar{x}_i\})$ is a pair $(\mathcal{S}, \mathbb{P})$, where \mathbb{P} is a probability distribution on \mathcal{S} that maximizes the entropy

$$S(\mathbb{P}) = - \sum_{C \in \mathcal{S}} \mathbb{P}(C) \ln \mathbb{P}(C) \rightarrow \max, \quad (3.1)$$

subject to the following constraints

$$\mathbb{E}_{\mathbb{P}}[x_i] = \sum_{C \in \mathcal{S}} x_i(C) \mathbb{P}(C) = \bar{x}_i, \quad (3.2)$$

$$\sum_{C \in \mathcal{S}} \mathbb{P}(C) = 1. \quad (3.3)$$

An ERSC is thus a descriptive model for random simplicial complexes. Generative models have been recently introduced and analyzed in [5, 6, 49].

We can define ERSC for any set of simplicial complexes, but, for most of the paper, we restrict ourselves to \mathcal{C}_n and its subsets. If we use $\mathcal{S} = \mathcal{G}_n \subset \mathcal{C}_n$, then we recover the definition of ERGs. As with ERGs, the solution of the constrained optimization problem (3.1)–(3.3) belongs to the exponential family, hence the name of the ensemble.

Theorem 1. *The maximum-entropy distribution \mathbb{P} defined by (3.1)–(3.3) can be written as follows*

$$\mathbb{P}(C) = \frac{e^{-H(C)}}{Z(\theta)}, \quad H(C) = \sum_{i=1}^r \theta_i x_i(C), \quad Z(\theta) = \sum_{C \in \mathcal{S}} e^{-H(C)}, \quad (3.4)$$

where $H(C)$ is the Hamiltonian of simplicial complex $C \in \mathcal{S}$, $Z(\theta)$ is the normalizing constant, called the partition function, and $\theta = (\theta_1, \dots, \theta_r)$ are the parameters satisfying the following system of r equations

$$-\frac{\partial \ln Z}{\partial \theta_i} = \bar{x}_i. \quad (3.5)$$

The proof is nearly identical to the proof for ERGs [37], but we give it here for completeness.

Proof. We use the standard method of Lagrange multipliers to solve the optimization problem (3.1)–(3.3). Let $\theta_1, \dots, \theta_r$ and α be the Lagrange multipliers for the constraints in (3.2) and (3.3). The Lagrangian is then

$$\mathcal{L} = - \sum_{C \in \mathcal{S}} \mathbb{P}(C) \ln \mathbb{P}(C) + \sum_{i=1}^r \theta_i \left(\bar{x}_i - \sum_{C \in \mathcal{S}} x_i(C) \mathbb{P}(C) \right) + \alpha \left(1 - \sum_{C \in \mathcal{S}} \mathbb{P}(C) \right). \quad (3.6)$$

The maximum entropy is achieved if the distribution \mathbb{P} satisfies $\frac{\partial \mathcal{L}}{\partial \mathbb{P}(C)} = 0$ for any $C \in \mathcal{S}$. This gives

$$-\ln \mathbb{P}(C) - 1 - \sum_{i=1}^r \theta_i x_i(C) - \alpha = 0, \quad (3.7)$$

or

$$\mathbb{P}(C) \propto \exp \left(- \sum_{i=1}^r \theta_i x_i(C) \right), \quad (3.8)$$

which is equivalent to (3.4), since $\sum_{C \in \mathcal{S}} \mathbb{P}(C) = 1$. It remains to check that (3.5) indeed holds:

$$\begin{aligned} - \frac{\partial \ln Z}{\partial \theta_i} &= - \frac{1}{Z} \frac{\partial}{\partial \theta_i} \sum_{C \in \mathcal{S}} e^{-H(C)} = \frac{1}{Z} \sum_{C \in \mathcal{S}} \frac{\partial H(C)}{\partial \theta_i} e^{-H(C)} \\ &= \frac{1}{Z} \sum_{C \in \mathcal{S}} x_i(C) e^{-H(C)} = \sum_{C \in \mathcal{S}} x_i(C) \mathbb{P}(C) = \bar{x}_i, \end{aligned} \quad (3.9)$$

since the expected value of the observable x_i in the ensemble is \bar{x}_i . □

4. Simple examples of ERSCs

Here we illustrate ERSCs with three simple examples: Erdős–Rényi random graphs $G(n, p)$, random flag complexes $X(n, p)$ and Linial–Meshulam random complexes $Y(n, p)$.

4.1. Erdős–Rényi random graphs

Perhaps the simplest nontrivial example of an ERSC is the Erdős–Rényi random graph ensemble $G(n, p)$, which can be viewed as a generative model for one-dimensional simplicial complexes. $G(n, p)$ is a maximum-entropy ensemble with only one constraint that the expected number of edges f_1 in the ensemble is $\binom{n}{2} p$ [37]:

$$G(n, p) = \text{ERSC} \left(\mathcal{G}_n, f_1, \binom{n}{2} p \right). \quad (4.1)$$

4.2. Random flag complexes

The flag complex $X(G)$ of a graph $G \in \mathcal{G}_n$, also called the clique complex or the Vietoris–Rips complex, is a (deterministic) simplicial complex in \mathcal{C}_n whose one-skeleton is G and whose k -simplices correspond to complete subgraphs of G , called cliques, of size $k + 1$. Since any simplicial complex is homeomorphic to a flag complex, simplicial complexes arise in different applications and are often used for topological data analysis [50].

Kahle [25, 26] defines the random flag complex $X(n, p)$ as the flag complex of the Erdős–Rényi random graph, $X(n, p) = X(G(n, p))$, and studies phase transitions of its homology groups. Here we show that $X(n, p)$ is, in fact, an ERSC.

Proposition 1. *Let $\mathcal{F}_n \subset \mathcal{C}_n$ be the set of all flag complexes on n vertices, then*

$$X(n, p) = \text{ERSC}\left(\mathcal{F}_n, f_1, \binom{n}{2}p\right). \tag{4.2}$$

Before giving the proof, we comment on what exactly proposition 1 states. $X(n, p)$ is a generative model of simplicial complexes: to generate $C \sim X(n, p)$, one first generates $G \sim G(n, p)$, and then sets $C = X(G)$. Let $\mathcal{S}_{X(n,p)} \subset \mathcal{F}_n$ denote the sample space of this random generative process, and $\mathbb{P}_{X(n,p)}$ be the resulting probability distribution on $\mathcal{S}_{X(n,p)}$. The random flag complex $X(n, p)$ can therefore be viewed as ensemble $(\mathcal{S}_{X(n,p)}, \mathbb{P}_{X(n,p)})$. Proposition 1 claims that $(\mathcal{S}_{X(n,p)}, \mathbb{P}_{X(n,p)})$ is a maximum-entropy ensemble with $\mathcal{S}_{X(n,p)} = \mathcal{F}_n$ and a single constraint that the expected number of one-simplices is $\binom{n}{2}p$. The proof is the same as for (4.1), but we give it here for illustrative purposes.

Proof. First, note that any flag complex $C \in \mathcal{F}_n$ can be generated by $X(n, p)$ with a non-zero probability:

$$\mathbb{P}_{X(n,p)}(C) = \mathbb{P}_{G(n,p)}(C^{(1)}) = p^{f_1(C)}(1-p)^{\binom{n}{2}-f_1(C)}. \tag{4.3}$$

Therefore, $\mathcal{S}_{X(n,p)}$ is indeed equal to \mathcal{F}_n . To prove (4.2), we need to show that $\mathbb{P}_{X(n,p)}$ is in fact the ERSC probability distribution (3.4), (3.5). Since every flag complex $C \in \mathcal{F}_n$ is completely defined by the adjacency matrix of its one-skeleton, $\mathcal{F}_n = \otimes_{d=1}^2 \mathbf{a}_d = \mathbf{1}_n \otimes \mathbf{a}_2$. The partition function Z can then be computed as follows:

$$\begin{aligned} Z(\theta_1) &= \sum_{C \in \mathcal{F}_n} e^{-H(C)} = \sum_{C \in \mathcal{F}_n} e^{-\theta_1 f_1(C)} = \sum_{\mathbf{a}_2} e^{-\theta_1 \sum_{i_2} a_{i_2}} \\ &= \sum_{\mathbf{a}_2} \prod_{i_2} e^{-\theta_1 a_{i_2}} = \prod_{i_2} \sum_{a_{i_2}=0}^1 e^{-\theta_1 a_{i_2}} = (1 + e^{-\theta_1})^{\binom{n}{2}}. \end{aligned} \tag{4.4}$$

We can now solve (3.5) with $\bar{x}_1 = \binom{n}{2}p$ for the parameter θ_1 ,

$$\theta_1 = -\ln \frac{p}{1-p}, \tag{4.5}$$

and check that indeed

$$\mathbb{P}_{X(n,p)}(C) = \frac{e^{-\theta_1 f_1(C)}}{Z(\theta_1)}. \tag{4.6}$$

This completes the proof. □

Given (4.1), the result in (4.2) is intuitively expected since the random part of generating $C \sim X(n, p)$ is sampling the one-skeleton $C^{(1)} \sim G(n, p)$. The rest of the construction, $C = X(C^{(1)})$, is fully deterministic.

4.3. Linial–Meshulam random complexes

Another example of ERSC is a generative model $Y(n, p)$ for random two-complexes. To generate $Y \sim Y(n, p)$, we start with a complete graph on n vertices, the one-skeleton of a

future simplicial complex, and add each of the $\binom{n}{3}$ possible triangle faces independently at random with probability p . Linial and Meshulam introduced this model in [31] and studied its topological properties. In particular, they proved for $Y(n, p)$ a cohomological analog of the celebrated Erdős–Rényi theorem on connectivity of the Erdős–Rényi random graphs [17]. The model $Y(n, p)$ can be readily generalized to higher dimensions: start with a full d -complex on n vertices, $1 \leq d \leq n - 2$, and add each of the $\binom{n}{d+2}$ possible $(d + 1)$ -simplices independently at random with probability p . We denote this model by $Y_d(n, p)$. The original Linial–Meshulam random complex $Y(n, p)$ is then $Y_1(n, p)$.

Let $\mathcal{C}_n^{(d+1)} \subset \mathcal{C}_n$ be a set of all simplicial complexes of dimension $(d + 1)$ or less, and $\mathcal{Y}_d \subset \mathcal{C}_n^{(d+1)}$ be a subset of complexes with full d -skeleton. In other words

$$\mathcal{Y}_d = \left\{ C \in \mathcal{C}_n^{(d+1)} : C^{(k)} = C^{[k]}, k = 1, \dots, d \right\}. \tag{4.7}$$

Since for any $C \in \mathcal{Y}_d$, the first $(d + 1)$ adjacency tensors $\mathbf{a}_1, \dots, \mathbf{a}_{d+1}$ are unit tensors with zero diagonals, $\mathcal{Y}_d = \mathbf{a}_{d+2}$.

Proposition 2. *The Linial–Meshulam random complex $Y_d(n, p)$ is the ERSC ensemble:*

$$Y_d(n, p) = \text{ERSC} \left(\mathcal{Y}_d, f_{d+1}, \binom{n}{d+2} p \right). \tag{4.8}$$

Proof. The proof is similar to that for random flag complexes. Given $C \in \mathcal{Y}_d$, the probability that the complex has been generated by $Y_d(n, p)$ is

$$\mathbb{P}_{Y_d(n,p)}(C) = p^{f_{d+1}(C)} (1 - p)^{\binom{n}{d+2} - f_{d+1}(C)}. \tag{4.9}$$

We need to show that this is in fact the maximum-entropy distribution under the constraint $\mathbb{E}[f_{d+1}] = \binom{n}{d+2} p$. The partition function:

$$\begin{aligned} Z(\theta_1) &= \sum_{C \in \mathcal{Y}_d} e^{-H(C)} = \sum_{C \in \mathcal{Y}_d} e^{-\theta_1 f_{d+1}(C)} = \sum_{\mathbf{a}_{d+2}} e^{-\theta_1 \sum_{i_{d+2}} a_{i_{d+2}}} \\ &= \sum_{\mathbf{a}_{d+2}} \prod_{i_{d+2}} e^{-\theta_1 a_{i_{d+2}}} = \prod_{i_{d+2}} \sum_{a_{i_{d+2}}=0}^1 e^{-\theta_1 a_{i_{d+2}}} = (1 + e^{-\theta_1})^{\binom{n}{d+2}}. \end{aligned} \tag{4.10}$$

The Lagrange multiplier is then $\theta_1 = -\ln \frac{p}{1-p}$, and

$$\mathbb{P}_{Y_d(n,p)}(C) = \frac{e^{-\theta_1 f_{d+1}(C)}}{Z(\theta_1)}, \tag{4.11}$$

as claimed. □

This result is also expected, since the Linial–Meshulam random complex is a higher dimensional analog of the Erdős–Rényi random graph: sampling from $Y_d(n, p)$ is the same Bernoulli trials process as in $G(n, p)$, with the only difference being that now we are creating $(d + 1)$ -simplices instead of one-simplices (edges).

5. Any distribution is maximum-entropy

It is a well-known fact in statistics and information theory (e.g. [11]) that *any* discrete distribution \mathbb{P}^* is maximum-entropy under properly specified constraints. Specifically, if one can write $-\ln \mathbb{P}^*$ as a linear combination $\sum \lambda_i h_i^* + \xi$ of some functions $\{h_i^*\}$, then distribution \mathbb{P}^* uniquely maximizes entropy $S(\mathbb{P})$ across all distributions \mathbb{P} that satisfy constraints $\mathbb{E}_{\mathbb{P}}[h_i^*] = \mathbb{E}_{\mathbb{P}^*}[h_i^*]$. In this section, we briefly review this general result, and show how it applies to the already considered models $G(n, p)$, $X(n, p)$, and $Y_d(n, p)$, where $-\ln \mathbb{P}^*$ can be written as a linear combination. We will see in the next section that this result simplifies dramatically the proofs for more complicated ERSCs.

Let us consider a discrete probability space (Ω, \mathbb{P}^*) , where Ω is a finite sample space and \mathbb{P}^* is some fixed probability distribution on Ω . Let us represent the distribution \mathbb{P}^* in the ‘Gibbs form’ as follows:

$$\mathbb{P}^*(\omega) = e^{-(\ln \mathbb{P}^*(\omega))} = e^{-H^*(\omega)}, \quad (5.1)$$

where

$$H^*(\omega) = -\ln \mathbb{P}^*(\omega). \quad (5.2)$$

Let \bar{H}^* denote the expectation of the function $H^*: \Omega \rightarrow \mathbb{R}$ with respect to \mathbb{P}^* , which is exactly the entropy of \mathbb{P}^* ,

$$\bar{H}^* = \mathbb{E}_{\mathbb{P}^*}[H^*] = \sum_{\omega \in \Omega} H^*(\omega) \mathbb{P}^*(\omega) = S(\mathbb{P}^*). \quad (5.3)$$

Lemma 1. *The probability distribution \mathbb{P}^* is the solution of the following optimization problem:*

$$S(\mathbb{P}) = -\sum_{\omega \in \Omega} \mathbb{P}(\omega) \ln \mathbb{P}(\omega) \rightarrow \max, \quad (5.4)$$

subject to the constraints

$$\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1 \text{ and } \mathbb{E}_{\mathbb{P}}[H^*] \equiv -\sum_{\omega \in \Omega} \mathbb{P}(\omega) \ln \mathbb{P}^*(\omega) = \bar{H}^*. \quad (5.5)$$

In other words, lemma 1 states that any discrete probability distribution is a maximum-entropy distribution. The entropy maximization is across all possible distributions \mathbb{P} satisfying the constraint that the expected value of H^* in distribution \mathbb{P} is equal to H^* ’s expected value in distribution \mathbb{P}^* , which is \mathbb{P}^* ’s entropy. In what follows, we will need a more general version of lemma 1.

Suppose that function H^* in (5.2) can be written as a linear combination of r other functions $h_i^*: \Omega \rightarrow \mathbb{R}$:

$$H^*(\omega) = -\ln \mathbb{P}^*(\omega) = \sum_{i=1}^r \lambda_i h_i^*(\omega) + \xi, \quad (5.6)$$

where $\lambda_i, \xi \in \mathbb{R}$ are constants. Let \bar{h}_i^* denote the expectation of h_i^* with respect to \mathbb{P}^* ,

$$\bar{h}_i^* = \mathbb{E}_{\mathbb{P}^*}[h_i^*] = \sum_{\omega \in \Omega} h_i^*(\omega) \mathbb{P}^*(\omega). \quad (5.7)$$

Lemma 2. (Theorem 11.1.1 [11]) The probability distribution \mathbb{P}^* is the solution of the following optimization problem:

$$S(\mathbb{P}) = - \sum_{\omega \in \Omega} \mathbb{P}(\omega) \ln \mathbb{P}(\omega) \rightarrow \max, \quad (5.8)$$

subject to the constraints

$$\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1 \text{ and } \mathbb{E}_{\mathbb{P}}[h_i^*] \equiv \sum_{\omega \in \Omega} h_i^*(\omega) \mathbb{P}(\omega) = \bar{h}_i^*, i = 1, \dots, r. \quad (5.9)$$

We note that the main utility of lemma 2 is not in observing that $\mathbb{P}^*(\omega) \propto e^{-\sum_{i=1}^r \lambda_i h_i^*(\omega)}$ is a maximum-entropy distribution (in fact, lemma 1 states that any distribution is), but in specifying more general constraints (5.9) under which distribution \mathbb{P}^* is maximum-entropy. Lemma 1 is a special case of more general lemma 2 with $\xi = 0$, $r = 1$, and $\lambda_1 = 1$. Indeed, in this case $H^* = h_1^*$, and the constraints in (5.5) and (5.9) become manifestly identical. Lemma 2 is identical to theorem 11.1.1 in [11], but we provide the proof here for completeness.

Proof. Let \mathbb{P} be any distribution that satisfies the constraints in (5.9). Then its entropy

$$\begin{aligned} S(\mathbb{P}) &= - \sum_{\omega \in \Omega} \mathbb{P}(\omega) \ln \mathbb{P}(\omega) = - \sum_{\omega \in \Omega} \mathbb{P}(\omega) \ln \frac{\mathbb{P}(\omega) \mathbb{P}^*(\omega)}{\mathbb{P}^*(\omega)} \\ &= - D_{\text{KL}}(\mathbb{P} \parallel \mathbb{P}^*) + \sum_{\omega \in \Omega} \mathbb{P}(\omega) H^*(\omega), \end{aligned} \quad (5.10)$$

where $D_{\text{KL}}(\mathbb{P} \parallel \mathbb{P}^*)$ is the Kullback–Leibler divergence of \mathbb{P} from \mathbb{P}^* . Since the KL divergence is always non-negative

$$\begin{aligned} S(\mathbb{P}) &\leq \sum_{\omega \in \Omega} \mathbb{P}(\omega) H^*(\omega) = \sum_{\omega \in \Omega} \mathbb{P}(\omega) \left(\sum_{i=1}^r \lambda_i h_i^*(\omega) + \xi \right) \\ &= \sum_{i=1}^r \lambda_i \mathbb{E}_{\mathbb{P}}[h_i^*] + \xi = \sum_{i=1}^r \lambda_i \bar{h}_i^* + \xi = S(\mathbb{P}^*). \end{aligned} \quad (5.11)$$

This shows that \mathbb{P}^* indeed maximizes the entropy. The uniqueness follows from the fact that $D_{\text{KL}}(\mathbb{P} \parallel \mathbb{P}^*) = 0$ if and only if $\mathbb{P} = \mathbb{P}^*$. \square

Lemmas 1 and 2 can be formulated for any ensemble of discrete ‘objects,’ including sets of graphs and simplicial complexes. Using the notation introduced in definition 2 and applying the lemmas to $\Omega = \mathcal{S} \subset \mathcal{C}_n$, we can concisely write lemma 1 as

$$(\mathcal{S}, \mathbb{P}^*) = \text{ERSC}(\mathcal{S}, H^*, \bar{H}^*), \quad (5.12)$$

and lemma 2 as

$$(\mathcal{S}, \mathbb{P}^*) = \text{ERSC}(\mathcal{S}, \{h_i^*\}, \{\bar{h}_i^*\}). \quad (5.13)$$

In many cases, lemmas 1 and 2 are not useful, since for many generative models $(\mathcal{S}, \mathbb{P}^*)$ the distribution \mathbb{P}^* cannot be explicitly written in the Gibbs form with linear Hamiltonian (5.6). Moreover, in generative models, the complexity of the generating algorithm often makes it impossible to even explicitly compute the probability $\mathbb{P}^*(C)$ for a given C . Even in the preferential attachment model [4], where the algorithm which generates a network appears

to be fairly simple—a new node connects to existing node i with probability proportional to its degree $p_i \propto k_i$ —the resulting distribution is unknown. However, if we do know \mathbb{P}^* as a function of observables, $\mathbb{P}^*(C) \propto e^{-\sum_{i=1}^r \lambda_i h_i^*(C)}$, lemma 2 is very helpful in representing $(\mathcal{S}, \mathbb{P}^*)$ as an ERSC.

Indeed, let us briefly see how lemma 2 applies to the already considered generative models. For $G(n, p)$, the probability of a graph $G \in \mathcal{G}_n$ in the model is

$$\mathbb{P}_{G(n,p)} = p^{f_1(G)} (1 - p)^{\binom{n}{2} - f_1(G)}. \tag{5.14}$$

The corresponding Hamiltonian is then

$$\begin{aligned} H_{G(n,p)}(G) &= -f_1(G) \ln p - \left(\binom{n}{2} - f_1(G) \right) \ln(1 - p) \\ &= \underbrace{\ln \frac{1-p}{p}}_{\lambda_1} \underbrace{f_1(G)}_{h_1^*(G)} - \underbrace{\binom{n}{2} \ln(1-p)}_{\xi}, \end{aligned} \tag{5.15}$$

where the bottom notations refer to the notations in lemma 2. The observation that $G(n, p)$ is an ERG (4.1) then follows from lemma 2, since $\mathbb{E}_{\mathbb{P}_{G(n,p)}}[f_1] = \binom{n}{2} p$. Similarly, for $X(n, p)$ and $Y_d(n, p)$,

$$\begin{aligned} H_{X(n,p)}(C) &= \ln \frac{1-p}{p} f_1(C) - \binom{n}{2} \ln(1-p), \\ \bar{f}_1 &= \mathbb{E}_{X(n,p)}[f_1] = \binom{n}{2} p, \\ H_{Y_d(n,p)}(C) &= \ln \frac{1-p}{p} f_1(C) - \binom{n}{d+1} \ln(1-p), \\ \bar{f}_{d+1} &= \mathbb{E}_{Y_d(n,p)}[f_{d+1}] = \binom{n}{d+2} p, \end{aligned} \tag{5.16}$$

and the observations (4.2) and (4.8) that these ensembles are ERSCs are direct corollaries of lemma 2.

The main point of this section is that in case the probability distribution is a known exponential function of a linear combination of structural observables, the computation of the partition function, which tends to be a nontrivial task in general, is not necessary to show that the distribution is the unique maximizer of entropy across all the distributions that satisfy the constraints that the expected values of these observables are equal to their expected values in this distribution.

6. Kahle’s Δ -ensembles

We now turn to a more general model which contains the Erdős–Rényi random graphs, the random flag complexes, and the Linial–Meshulam complexes as special cases. In a recent survey [27], Kahle introduced the following multi-parameter model $\Delta(n; p_1, \dots, p_{n-1})$ which generates random simplicial complexes inductively by dimension. First, build a one-skeleton by putting an edge between any two vertices with probability p_1 . Then, for $d = 2, \dots, n - 1$, add every d -simplex with probability p_d , but only if the entire $(d - 1)$ -dimensional boundary of that simplex is already in place. More formally, we have the following definition.

Definition 3. The Kahle model $\Delta(n; p_1, \dots, p_{n-1})$ is a random simplicial complex model that generates $C \in \mathcal{C}_n$ as follows: for $d = 1, \dots, n - 1$, for every \mathbf{i}_{d+1} ,

$$\begin{aligned} &\text{if } b_{\mathbf{i}_{d+1}} = 0 \Rightarrow \text{set } a_{\mathbf{i}_{d+1}} = 0. \\ &\text{if } b_{\mathbf{i}_{d+1}} = 1 \Rightarrow \text{set } a_{\mathbf{i}_{d+1}} = \begin{cases} 1 & \text{with probability } p_d, \\ 0 & \text{with probability } 1 - p_d. \end{cases} \end{aligned} \tag{6.1}$$

Topological properties of the Khale model are studies in [8–10]. Here we study its maximum-entropy properties. In appendix A.1 we prove the following proposition.

Proposition 3. Let $C \sim \Delta(n; p_1, \dots, p_{n-1})$. The expected numbers of d -simplices in $C^{(d)}$ and $C^{[d]}$ are

$$\bar{f}_d = \binom{n}{d+1} \prod_{k=1}^d p_k^{\binom{d+1}{d-k}} \text{ and } \bar{\phi}_d = \binom{n}{d+1} \prod_{k=1}^{d-1} p_k^{\binom{d+1}{d-k}}. \tag{6.2}$$

The Kahle model unifies all the random simplicial complexes we have considered so far:

$$\begin{aligned} G(n, p) &= \Delta(n; p, 0, \dots, 0), \\ X(n, p) &= \Delta(n; p, 1, \dots, 1), \\ Y(n, p) &= \Delta(n; 1, p, 0, \dots, 0), \\ Y_d(n, p) &= \Delta\left(n; \underbrace{1, \dots, 1}_d, p, 0, \dots, 0\right). \end{aligned} \tag{6.3}$$

Since all these special cases are ERSCs, it is natural to expect that so is $\Delta(n; p_1, \dots, p_{n-1})$. We cannot prove this using the same method as for the Erdős–Rényi random graphs and the random flag and Linial–Meshulam complexes in section 4. As with ERGs, analytical computation of the partition function $Z(\theta)$ for ERSCs is rarely possible, and $G(n, p)$, $X(n, p)$, and $Y_d(n, p)$ are lucky exceptions. In appendix A.3 we illustrate difficulties one has to be prepared to experience when attempting to compute the partition function for $\Delta(n; p_1, \dots, p_{n-1})$ with $n = 3$. However, with the help of lemmas 1 and 2 in section 5 there exists a simpler alternative proof. The fact that $\Delta(n; p_1, \dots, p_{n-1})$ is an ERSC is a direct corollary of those lemmas.

Theorem 2. The Kahle Δ -ensemble is the ERSC ensemble:

$$\Delta(n; p_1, \dots, p_{n-1}) = \text{ERSC}\left(\mathcal{C}_n, \left\{ \left\{ f_d \right\}_{d=1}^{n-1}, \left\{ \phi_d \right\}_{d=2}^{n-1} \right\}, \left\{ \left\{ \bar{f}_d \right\}_{d=1}^{n-1}, \left\{ \bar{\phi}_d \right\}_{d=2}^{n-1} \right\} \right), \tag{6.4}$$

where \bar{f}_d and $\bar{\phi}_d$ are the expected numbers of d -simplices in $C^{(d)}$ and $C^{[d]}$.

Proof. For any $C \in \mathcal{C}_n$, the probability $\mathbb{P}_\Delta(C)$ that $\Delta(n, p_1, \dots, p_{n-1})$ generates C can be computed by induction:

$$\mathbb{P}_\Delta(C) = \prod_{d=1}^{n-1} \mathbb{P}_\Delta(C^{(d)} | C^{(d-1)}) = \prod_{d=1}^{n-1} p_d^{f_d(C)} (1 - p_d)^{\phi_d(C) - f_d(C)}. \tag{6.5}$$

Indeed, given the $(d - 1)$ -skeleton $C^{(d-1)}$, the maximum possible number of d -simplices in $C^{(d)}$ is exactly $\phi_d(C)$, the number of d -simplices in the filled skeleton $C^{[d]}$. Since each of these d -simplices appears independently with probability p_d , the conditional probability

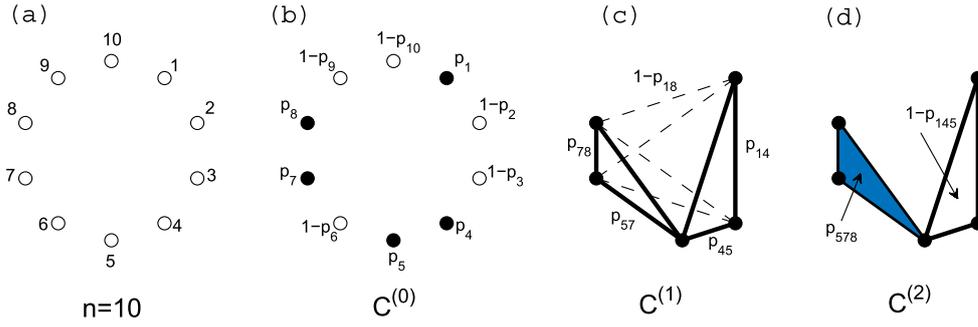


Figure 4. Sampling from Δ with $n = 10$. Panel (a) shows the starting point: $n = 10$ potential zero-simplices (vertices) represented by empty dots. At the first step, shown in panel (b), we create the zero-skeleton of a future complex C by including (excluding) every vertex i in (from) $C^{(0)}$ with probability p_i (with probability $1 - p_i$). In this example, five vertices represented by filled dots belong to $C^{(0)}$. Next, in panel (c), we generate the one-skeleton by creating one-simplices (edges). Each of the $\binom{5}{2}$ possible edges $\{i_2\} = \{1, 4\}, \{1, 5\}, \dots, \{7, 8\}$ appears in $C^{(1)}$ with probability p_{i_2} . We represent the accepted (rejected) edges by solid (dashed) lines. Finally, in panel (d), we create the two-skeleton by adding two-simplices (triangles). There are only two possible triangles in $C^{(2)}$: $\{1, 4, 5\}$ and $\{5, 7, 8\}$. Here, the former (empty) was rejected with probability $1 - p_{145}$, and the latter (filled) was accepted with probability p_{578} .

$\mathbb{P}_\Delta(C^{(d)} | C^{(d-1)}) = p_d^{f_d(C)} (1 - p_d)^{\phi_d(C) - f_d(C)}$, where $f_d(C)$ is the actual number of d -simplices in $C^{(d)}$.

The Hamiltonian of C is therefore

$$\begin{aligned}
 H_\Delta(C) &= \sum_{d=1}^{n-1} \left(f_d(C) \ln \frac{1 - p_d}{p_d} + \phi_d(C) \ln \frac{1}{1 - p_d} \right) \\
 &= \sum_{d=1}^{n-1} f_d(C) \ln \frac{1 - p_d}{p_d} + \sum_{d=2}^{n-1} \left(\phi_d(C) \ln \frac{1}{1 - p_d} \right) + \binom{n}{2} \ln \frac{1}{1 - p_1},
 \end{aligned} \tag{6.6}$$

since $\phi_1(C) = \binom{n}{2}$ for any C . Using lemma 2 with

$$\begin{aligned}
 \{h_i^*\} &= \{f_1, \dots, f_{n-1}, \phi_2, \dots, \phi_{n-1}\}, \\
 \{\lambda_i\} &= \left\{ \ln \frac{1 - p_1}{p_1}, \dots, \ln \frac{1 - p_{n-1}}{p_{n-1}}, \ln \frac{1}{1 - p_2}, \dots, \ln \frac{1}{1 - p_{n-1}} \right\}, \\
 \xi &= \binom{n}{2} \ln \frac{1}{1 - p_1},
 \end{aligned} \tag{6.7}$$

completes the proof. □

7. General random simplicial complexes with independent simplices

Finally, we introduce and consider the most general case of random simplicial complexes with statistically independent simplices. In this case each simplex has its own individual probability of appearance. To stay as general as possible, we must allow for even the zero-

simplices (vertices) to be present with any probabilities, which are not necessarily equal to 1. We denote this new model by $\Delta(n; \mathbf{p}_1, \dots, \mathbf{p}_n)$, or Δ for brevity, where $\mathbf{p}_d = \{p_{\mathbf{i}_d}\}$ is a collection of $\binom{n}{d}$ appearance probabilities for each $(d - 1)$ -simplex. Whereas in Δ in the previous section, the subindex d in p_d refers to the simplex dimension, in Δ the sub-multi-index \mathbf{i}_d in $p_{\mathbf{i}_d}$ refers to the specific $(d - 1)$ -simplex $\{\mathbf{i}_d\}$. To generate $C \sim \Delta$, we first create its zero-skeleton by having vertices $\{\mathbf{i}_1\} \in C$ with probabilities $p_{\mathbf{i}_1}$, $\mathbf{i}_1 = 1, \dots, n$. Then, for $d = 1, \dots, n - 1$, we add every d -simplex $\{\mathbf{i}_{d+1}\}$ with probability $p_{\mathbf{i}_{d+1}}$, but only if the entire $(d - 1)$ -dimensional boundary of that simplex is already in place. Figure 4 illustrates the generation of a two-complex from Δ with $n = 10$.

More formally, we have the following definition. Let $\mathcal{C}_{\leq n} = \cup_{k=1}^n \mathcal{C}_k$ denote the set of all simplicial complexes with n vertices or less. As in section 2, any $C \in \mathcal{C}_{\leq n}$ is uniquely determined by a collection of its adjacency tensors $\mathbf{a}_d = \{a_{\mathbf{i}_d}\}$, $d = 1, \dots, n$, except that now \mathbf{a}_1 is not necessarily equal to the all-ones vector $\mathbf{1}_n$, since C may have less than n vertices.

Definition 4. The model $\Delta(n; \mathbf{p}_1, \dots, \mathbf{p}_n)$ is a random simplicial complex model that generates $C \in \mathcal{C}_{\leq n}$ as follows: for $d = 0, \dots, n - 1$, for every \mathbf{i}_{d+1} ,

$$\begin{aligned} & \text{if } b_{\mathbf{i}_{d+1}} = 0 \Rightarrow \text{set } a_{\mathbf{i}_{d+1}} = 0, \\ & \text{if } b_{\mathbf{i}_{d+1}} = 1 \Rightarrow \text{set } a_{\mathbf{i}_{d+1}} = \begin{cases} 1 & \text{with probability } p_{\mathbf{i}_{d+1}}, \\ 0 & \text{with probability } 1 - p_{\mathbf{i}_{d+1}}. \end{cases} \end{aligned} \quad (7.1)$$

Here, for convenience, we use the convention that $b_{\mathbf{i}_1} = 1$ for all $\mathbf{i}_1 = 1, \dots, n$. If $p_{\mathbf{i}_{d+1}} = p_d$ for $d = 0, \dots, n - 1$ and $p_0 = 1$, then we recover the original Kahle's model $\Delta(n; p_1, \dots, p_{n-1})$ from the previous section.

To give the expressions for the expected values of the observables $a_{\mathbf{i}_d}$ and $b_{\mathbf{i}_d}$ in $\Delta(n; \mathbf{p}_1, \dots, \mathbf{p}_n)$, we need a bit of new notation. Let multi-index $\mathbf{k}_m = k_1, \dots, k_m$ denote an m -tuple with increasing values $1 \leq k_1 < \dots < k_m \leq d$, and $\mathbf{i}_d^{\widehat{\mathbf{k}}_m} = i_1, \dots, \widehat{i_{k_1}}, \dots, \widehat{i_{k_m}}, \dots, i_d$ be the $(d - m)$ -long multi-index with i_{k_1}, \dots, i_{k_m} omitted, with a convention that $\mathbf{k}_0 = \emptyset$ and $\mathbf{i}_d^{\widehat{\mathbf{k}}_0} = \mathbf{i}_d$. In appendix A.2 we prove the following proposition.

Proposition 4. Let $C \sim \Delta(n; \mathbf{p}_1, \dots, \mathbf{p}_n)$. The expected values of the observables $a_{\mathbf{i}_d}$ and $b_{\mathbf{i}_d}$ are

$$\bar{a}_{\mathbf{i}_d} = \prod_{m=0}^{d-1} \prod_{\mathbf{k}_m} p_{\mathbf{i}_d^{\widehat{\mathbf{k}}_m}} \text{ and } \bar{b}_{\mathbf{i}_d} = \prod_{m=1}^{d-1} \prod_{\mathbf{k}_m} p_{\mathbf{i}_d^{\widehat{\mathbf{k}}_m}}. \quad (7.2)$$

Lemma 2 helps again to prove that the general model Δ is also an ERSC.

Theorem 3. The Δ -ensemble is the ERSC ensemble:

$$\Delta(n; \mathbf{p}_1, \dots, \mathbf{p}_n) = \text{ERSC}\left(\mathcal{C}_{\leq n}, \left\{ \{a_{\mathbf{i}_d}\}_{d=1}^n, \{b_{\mathbf{i}_d}\}_{d=2}^n \right\}, \left\{ \{\bar{a}_{\mathbf{i}_d}\}_{d=1}^n, \{\bar{b}_{\mathbf{i}_d}\}_{d=2}^n \right\} \right), \quad (7.3)$$

where $\bar{a}_{\mathbf{i}_d}$ and $\bar{b}_{\mathbf{i}_d}$ are the expected values of the observables $a_{\mathbf{i}_d}$ and $b_{\mathbf{i}_d}$.

Proof. The probability that $\Delta(n; \mathbf{p}_1, \dots, \mathbf{p}_n)$ generates $C \in \mathcal{C}_{\leq n}$ is

$$\begin{aligned} \mathbb{P}_{\Delta}(C) &= \mathbb{P}_{\Delta}(C^{(0)}) \prod_{d=1}^{n-1} \mathbb{P}_{\Delta}(C^{(d)} | C^{(d-1)}) \\ &= \prod_{\mathbf{i}_1} p_{\mathbf{i}_1}^{a_{\mathbf{i}_1}} (1 - p_{\mathbf{i}_1})^{1 - a_{\mathbf{i}_1}} \times \prod_{d=1}^{n-1} \prod_{\mathbf{i}_{d+1}} p_{\mathbf{i}_{d+1}}^{a_{\mathbf{i}_{d+1}}} (1 - p_{\mathbf{i}_{d+1}})^{b_{\mathbf{i}_{d+1}} - a_{\mathbf{i}_{d+1}}} \\ &= \prod_{d=1}^n \prod_{\mathbf{i}_d} p_{\mathbf{i}_d}^{a_{\mathbf{i}_d}} (1 - p_{\mathbf{i}_d})^{b_{\mathbf{i}_d} - a_{\mathbf{i}_d}}. \end{aligned} \quad (7.4)$$

The Hamiltonian of C is then

$$\begin{aligned} H_{\Delta}(C) &= \sum_{d=1}^n \sum_{\mathbf{i}_d} \left(a_{\mathbf{i}_d} \ln \frac{1 - p_{\mathbf{i}_d}}{p_{\mathbf{i}_d}} + b_{\mathbf{i}_d} \ln \frac{1}{1 - p_{\mathbf{i}_d}} \right) \\ &= \sum_{d=1}^n \sum_{\mathbf{i}_d} \alpha_{\mathbf{i}_d} a_{\mathbf{i}_d} + \sum_{d=2}^n \sum_{\mathbf{i}_d} \beta_{\mathbf{i}_d} b_{\mathbf{i}_d} + \sum_{\mathbf{i}_1} \ln \frac{1}{1 - p_{\mathbf{i}_1}}, \end{aligned} \quad (7.5)$$

where $\alpha_{\mathbf{i}_d}$ and $\beta_{\mathbf{i}_d}$ are the Lagrange multipliers coupled to observables $a_{\mathbf{i}_d}$ and $b_{\mathbf{i}_d}$,

$$\alpha_{\mathbf{i}_d} = \ln \frac{1 - p_{\mathbf{i}_d}}{p_{\mathbf{i}_d}} \text{ and } \beta_{\mathbf{i}_d} = \ln \frac{1}{1 - p_{\mathbf{i}_d}}. \quad (7.6)$$

Using lemma 2 with

$$\begin{aligned} \{h_i^*\} &= \{ \{a_{\mathbf{i}_1}\}, \dots, \{a_{\mathbf{i}_n}\}, \{b_{\mathbf{i}_2}\}, \dots, \{b_{\mathbf{i}_n}\} \}, \\ \{\lambda_i\} &= \{ \{\alpha_{\mathbf{i}_1}\}, \dots, \{\alpha_{\mathbf{i}_n}\}, \{\beta_{\mathbf{i}_2}\}, \dots, \{\beta_{\mathbf{i}_n}\} \}, \\ \xi &= \sum_{\mathbf{i}_1} \ln \frac{1}{1 - p_{\mathbf{i}_1}}, \end{aligned} \quad (7.7)$$

completes the proof. \square

8. Discussion

In summary, ERSCs are a natural higher dimensional analog of ERGs which are extensively used for modeling network data and statistical inference. An ERSC ensemble is a maximum-entropy ensemble of simplicial complexes under ‘soft’ constraints that fix expected values of some observables or properties of simplicial complexes. We have developed the formalism for ERSCs, and introduced the most general generative model of random simplicial complexes Δ with statistically independent simplices. This model has as special cases several popular models studied in the literature: Erdős–Rényi random graphs, random flag complexes, Linial–Meshulam complexes, and Kahle’s Δ -ensembles. As all these models, Δ is an ERSC ensemble. The constraints in this ensemble are expected number of simplices and their boundaries.

This result is a direct corollary of the general observation that any probability distribution \mathbb{P} is maximum-entropy under the constraint that the expected value of $-\ln \mathbb{P}$ is equal to the entropy of \mathbb{P} . This observation dramatically simplifies the representation of many ensembles of random simplicial complexes as ERSCs since the calculation of the partition function is no longer needed. For example, to show that the Erdős–Rényi random graphs $G(n, p)$ are ERGs with a given expected number of edges, one does not really have to calculate the partition function. This calculation is trivial in the Erdős–Rényi case or in the general case of ERGs with statistically

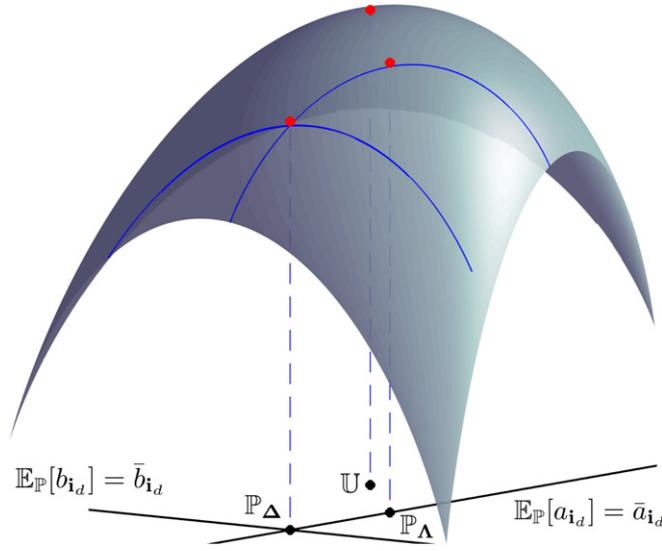


Figure 5. Constrained entropy maximization. The surface represents the Gibbs entropy S , which, in this schematic example, is a function on the set of all probability distributions on $\mathcal{C}_{\leq n}$. The global maximum corresponds to the uniform distribution \mathbb{U} , which is the maximum-entropy distribution among all distributions supported on $\mathcal{C}_{\leq n}$. theorem 3 shows that if we have two sets of constraints, $\mathbb{E}_{\mathbb{P}}[a_{i_d}] = \bar{a}_{i_d}$ and $\mathbb{E}_{\mathbb{P}}[b_{i_d}] = \bar{b}_{i_d}$, then the resulting maximum-entropy distribution is \mathbb{P}_{Δ} . If we drop the second set of constraints, then we get some other maximum-entropy distribution $\mathbb{P}_{\Lambda} \neq \mathbb{P}_{\Delta}$ for ensemble $\Lambda \neq \Delta$.

independent edges [37]. However, the analogous calculation for the general case of random simplicial complexes Δ with statistically independent simplices appears to be intractable.

The multi-parameter model $\Delta(n; \mathbf{p}_1, \dots, \mathbf{p}_n)$ is the ERSC ensemble with two types of constrained observables: $\{a_{i_d}\}$ and $\{b_{i_d}\}$. The observables of the first type are simplices themselves: $a_{i_d}(C) = 1$ if the $(d - 1)$ -simplex $\{i_d\}$ belongs to C , and zero otherwise. The observables of the second type are their boundaries: $b_{i_d}(C) = 1$ if the entire $(d - 2)$ -dimensional boundary of simplex $\{i_d\}$ belongs to C , and zero otherwise. Theorem 3 states that $\Delta(n; \mathbf{p}_1, \dots, \mathbf{p}_n)$ is a solution of the following optimization problem:

$$S(\mathbb{P}) \rightarrow \max, \sum_{C \in \mathcal{C}_{\leq n}} \mathbb{P}(C) = 1, \tag{8.1}$$

$$\mathbb{E}_{\mathbb{P}}[a_{i_d}] = \bar{a}_{i_d}, d = 1, \dots, n, \mathbb{E}_{\mathbb{P}}[b_{i_d}] = \bar{b}_{i_d}, d = 2, \dots, n. \tag{8.2}$$

If we drop the observables of the second type in this optimization problem, we alter the maximum-entropy distribution as illustrated in figure 5. Since the distribution has changed, ensemble $\Lambda(n; \mathbf{p}_1, \dots, \mathbf{p}_n) = \text{ERSC}(\mathcal{C}_{\leq n}, \{a_{i_d}\}_{d=1}^n, \{\bar{a}_{i_d}\}_{d=1}^n)$ defined by this distribution is now also different from $\Delta(n; \mathbf{p}_1, \dots, \mathbf{p}_n)$.

The fact that the second type of boundary-presence observables are also constrained in $\Delta(n; \mathbf{p}_1, \dots, \mathbf{p}_n)$ may appear quite unexpected at first glance. The reason for the presence of these constraints is that simplex existence probabilities are actually conditional, where the conditions are the presence of simplex boundaries. If we go from conditional to unconditional probabilities, we change Δ to Λ . Indeed, in Δ , p_{i_d} is the conditional probability of the

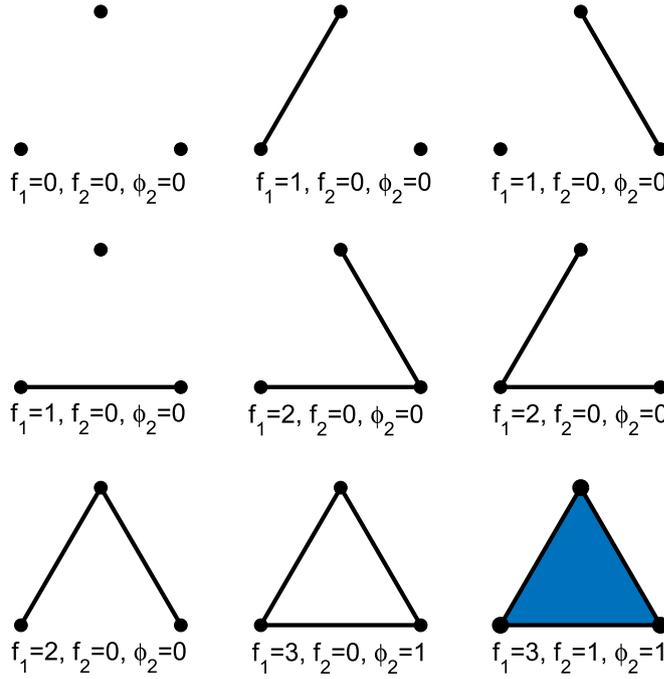


Figure A1. Simplicial complexes on three vertices. Here we show all $C \in \mathcal{C}_3$ and the values of f_1 (number of one-simplices), f_2 (number of two-simplices), and ϕ_2 (number of two-simplices in $C^{[2]}$) for each C .

$(d - 1)$ -simplex $\{\mathbf{i}_d\}$ to appear in C , given that its $(d - 2)$ -dimensional boundary is already in place

$$p_{\mathbf{i}_d} = \mathbb{P}_{\Delta}(a_{\mathbf{i}_d} = 1 | b_{\mathbf{i}_d} = 1) = \frac{\mathbb{P}_{\Delta}(a_{\mathbf{i}_d} = 1, b_{\mathbf{i}_d} = 1)}{\mathbb{P}_{\Delta}(b_{\mathbf{i}_d} = 1)} = \frac{\mathbb{P}_{\Delta}(a_{\mathbf{i}_d} = 1)}{\mathbb{P}_{\Delta}(b_{\mathbf{i}_d} = 1)}, \quad (8.3)$$

where the last equation follows from the compatibility condition $a_{\mathbf{i}_d} = 1 \Rightarrow b_{\mathbf{i}_d} = 1$. This means that the unconditional probability of having $\{\mathbf{i}_d\} \in C$ is

$$\mathbb{P}_{\Delta}(a_{\mathbf{i}_d} = 1) = p_{\mathbf{i}_d} \mathbb{P}_{\Delta}(b_{\mathbf{i}_d} = 1), \quad (8.4)$$

and, therefore, the expected values of observables $a_{\mathbf{i}_d}$ and $b_{\mathbf{i}_d}$ satisfy

$$\bar{a}_{\mathbf{i}_d} = \mathbb{E}_{\Delta}[a_{\mathbf{i}_d}] = \mathbb{P}_{\Delta}(a_{\mathbf{i}_d} = 1) = p_{\mathbf{i}_d} \mathbb{P}_{\Delta}(b_{\mathbf{i}_d} = 1) = p_{\mathbf{i}_d} \bar{b}_{\mathbf{i}_d}. \quad (8.5)$$

Thus, if we want to represent Δ as an ERSC and we fixed the expected values of the observables of the first type $a_{\mathbf{i}_d}$, we must also fix the expected values of the observables of the second type $b_{\mathbf{i}_d}$. Moreover, these expected values are not independent and must satisfy $\bar{a}_{\mathbf{i}_d} = p_{\mathbf{i}_d} \bar{b}_{\mathbf{i}_d}$, which is consistent with proposition 4. In appendix A.4 we consider a special case with $n = 3$, $p_{\mathbf{i}_1} = 1$, $p_{\mathbf{i}_2} = p_1$, and $p_{\mathbf{i}_3} = p_2$, and explicitly show that the maximum-entropy distributions with and without the second type constraints are different.

To conclude, $\Delta \neq \Lambda$. From the maximum-entropy point of view, the ensemble Λ , with only the observables of the first type constrained, appears more natural than Δ . Yet Δ is more natural than Λ in terms of simplicity of its constructive definition 4 that allows for efficient sampling of simplicial complexes. We leave open the questions of whether there exist ways to

calculate the probability distribution $\mathbb{P}_\Lambda(C)$ in ensemble Λ , and to efficiently sample from it, i.e., to easily generate simplicial complexes C with this probability.

Acknowledgments

We thank Alexandru Suciuc, Gabor Lippner, and Christopher King for very useful discussions, comments, and suggestions. This work was supported by DARPA grant No. HR0011-12-1-0012; NSF grants No. CNS-1344289, CNS-1442999, CNS-0964236, CNS-1441828, CNS-1039646, and CNS-1345286; and by Cisco Systems.

Appendix

A.1. Proof of proposition 3

Let us first compute the expected number of d -simplices in $C^{[d]}$, where $C \sim \Delta(n; p_1, \dots, p_{n-1})$.

$$\begin{aligned} \bar{\phi}_d &= \mathbb{E}[\phi_d] = \mathbb{E}\left[\sum_{\mathbf{i}_{d+1}} b_{\mathbf{i}_{d+1}}\right] = \mathbb{E}\left[\sum_{\mathbf{i}_{d+1}} \prod_{k=1}^{d+1} a_{\mathbf{i}_{d+1}^k}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\sum_{\mathbf{i}_{d+1}} \prod_{k=1}^{d+1} a_{\mathbf{i}_{d+1}^k} \mid \mathbf{a}_{d-1}\right]\right] = \mathbb{E}\left[\sum_{\mathbf{i}_{d+1}} \prod_{k=1}^{d+1} \mathbb{E}\left[a_{\mathbf{i}_{d+1}^k} \mid \mathbf{a}_{d-1}\right]\right]. \end{aligned} \quad (\text{A.1})$$

If the boundary of $(d-1)$ -simplex $\{\mathbf{i}_{d+1}^k\}$ belongs to C , i.e. $b_{\mathbf{i}_{d+1}^k} = 1$, then $\{\mathbf{i}_{d+1}^k\} \in C$, i.e. $a_{\mathbf{i}_{d+1}^k} = 1$, with probability p_{d-1} . Otherwise, if $b_{\mathbf{i}_{d+1}^k} = 0$, then automatically $a_{\mathbf{i}_{d+1}^k} = 0$. Therefore, the inner expected value:

$$\mathbb{E}\left[a_{\mathbf{i}_{d+1}^k} \mid \mathbf{a}_{d-1}\right] = p_{d-1} b_{\mathbf{i}_{d+1}^k}. \quad (\text{A.2})$$

So

$$\bar{\phi}_d = \mathbb{E}\left[\sum_{\mathbf{i}_{d+1}} \prod_{k=1}^{d+1} p_{d-1} b_{\mathbf{i}_{d+1}^k}\right] = p_{d-1}^{d+1} \mathbb{E}\left[\sum_{\mathbf{i}_{d+1}^{\mathbf{k}_2}} \prod_{k=1}^{d+1} a_{\mathbf{i}_{d+1}^k}\right], \quad (\text{A.3})$$

where $\mathbf{k}_2 = k_1, k_2$ is a pair of indices $1 \leq k_1 < k_2 \leq d+1$, and $\mathbf{i}_{d+1}^{\mathbf{k}_2} = i_1, \dots, \widehat{i_{k_1}}, \dots, \widehat{i_{k_2}}, \dots, i_{d+1}$ is the $(d-1)$ -long multi-index with i_{k_1} and i_{k_2} omitted. Proceeding in this manner, we have:

$$\begin{aligned} \bar{\phi}_d &= p_{d-1}^{d+1} \mathbb{E}\left[\mathbb{E}\left[\sum_{\mathbf{i}_{d+1}^{\mathbf{k}_2}} \prod_{k=1}^{d+1} a_{\mathbf{i}_{d+1}^k} \mid \mathbf{a}_{d-2}\right]\right] = p_{d-1}^{d+1} \mathbb{E}\left[\sum_{\mathbf{i}_{d+1}^{\mathbf{k}_2}} \mathbb{E}\left[a_{\mathbf{i}_{d+1}^{\mathbf{k}_2}} \mid \mathbf{a}_{d-2}\right]\right] \\ &= p_{d-1}^{d+1} \mathbb{E}\left[\sum_{\mathbf{i}_{d+1}^{\mathbf{k}_2}} p_{d-2} b_{\mathbf{i}_{d+1}^{\mathbf{k}_2}}\right] = p_{d-1}^{\binom{d+1}{1}} p_{d-2}^{\binom{d+1}{2}} \mathbb{E}\left[\sum_{\mathbf{i}_{d+1}^{\mathbf{k}_3}} \prod_{k=1}^{d+1} a_{\mathbf{i}_{d+1}^k}\right] = \dots \\ &= p_{d-1}^{\binom{d+1}{1}} \dots p_1^{\binom{d+1}{d-1}} \mathbb{E}\left[\sum_{\mathbf{i}_{d+1}^{\mathbf{k}_d}} \prod_{k=1}^{d+1} a_{\mathbf{i}_{d+1}^k}\right] = \binom{n}{d+1} \prod_{k=1}^{d-1} p_k^{\binom{d+1}{d-k}}. \end{aligned} \quad (\text{A.4})$$

The last equation holds because $a_{\mathbf{i}_{d+1}^{\mathbf{k}_d}} = 1$ for any \mathbf{i}_{d+1} and \mathbf{k}_d , since all simplicial complexes $C \sim \Delta(n; p_1, \dots, p_{n-1})$ have exactly n vertices. The expected number of d -simplices in $C^{(d)}$

is now:

$$\bar{f}_d = \mathbb{E}[f_d] = \mathbb{E}[\mathbb{E}[f_d | \phi_d]] = \mathbb{E}[p_d \phi_d] = p_d \bar{\phi}_d = \binom{n}{d+1} \prod_{k=1}^d p_k \binom{d+1}{d-k}. \quad (\text{A.5})$$

A.2. Proof of proposition 4

Computations are similar to those in the previous section

$$\begin{aligned} \bar{b}_{\mathbf{i}_d} &= \mathbb{E}[b_{\mathbf{i}_d}] = \mathbb{E}\left[\prod_{k=1}^d a_{\mathbf{i}_d^k}\right] = \mathbb{E}\left[\mathbb{E}\left[\prod_{k=1}^d a_{\mathbf{i}_d^k} \mid \mathbf{a}_{d-2}\right]\right] \\ &= \mathbb{E}\left[\prod_{k=1}^d \mathbb{E}\left[a_{\mathbf{i}_d^k} \mid \mathbf{a}_{d-2}\right]\right] = \mathbb{E}\left[\prod_{k=1}^d p_{\mathbf{i}_d^k} b_{\mathbf{i}_d^k}\right] = \prod_{k=1}^d p_{\mathbf{i}_d^k} \mathbb{E}\left[\prod_{k=1}^d b_{\mathbf{i}_d^k}\right] \\ &= \prod_{k=1}^d p_{\mathbf{i}_d^k} \mathbb{E}\left[\prod_{\mathbf{k}_2} a_{\mathbf{i}_d^k} \right] = \prod_{k=1}^d p_{\mathbf{i}_d^k} \prod_{\mathbf{k}_2} p_{\mathbf{i}_d^k} \mathbb{E}\left[\prod_{\mathbf{k}_2} b_{\mathbf{i}_d^k}\right] = \dots \\ &= \prod_{k=1}^d p_{\mathbf{i}_d^k} \dots \prod_{\mathbf{k}_{d-1}} p_{\mathbf{i}_d^k} \mathbb{E}\left[\prod_{\mathbf{k}_{d-1}} b_{\mathbf{i}_d^k}\right] = \prod_{m=1}^{d-1} \prod_{\mathbf{k}_m} p_{\mathbf{i}_d^k}, \end{aligned} \quad (\text{A.6})$$

since $b_{\mathbf{i}_d^k} = 1$ for any \mathbf{i}_d and \mathbf{k}_{d-1} . Finally

$$\bar{a}_{\mathbf{i}_d} = \mathbb{E}[a_{\mathbf{i}_d}] = \mathbb{E}[\mathbb{E}[a_{\mathbf{i}_d} | b_{\mathbf{i}_d}]] = \mathbb{E}[p_{\mathbf{i}_d} b_{\mathbf{i}_d}] = \prod_{m=0}^{d-1} \prod_{\mathbf{k}_m} p_{\mathbf{i}_d^k}. \quad (\text{A.7})$$

A.3. Special case: $\Delta(3; p_1, p_2)$.

Theorem 2 in section 6 explicitly represents Kahle’s multi-parameter model of random simplicial complexes $\Delta(n; p_1, \dots, p_{n-1})$ as an ERSC for any values of the parameters. This theorem is a direct corollary of lemmas 1 and 2 in section 5 which assert that any distribution is, in fact, the maximum-entropy distribution under certain constraints. Here we illustrate the difficulties that arise when one tries to compute the maximum-entropy distribution \mathbb{P}_Δ using theorem 1. We successfully used this method, which is based on computing the partition function, in section 4 for the Erdős–Rényi random graphs and the random flag and Linial–Meshulam complexes. For Kahle’s Δ -ensemble, however, the partition function becomes intractable.

Consider a special case of Kahle’s model with $n = 3$. According to theorem 2 and proposition 3, $\Delta(3; p_1, p_2)$ is the maximum-entropy ensemble of simplicial complexes on three vertices with three constraints:

$$\mathbb{E}[f_1] = 3p_1, \quad \mathbb{E}[f_2] = p_1^3 p_2, \quad \mathbb{E}[\phi_2] = p_1^3. \quad (\text{A.8})$$

Let us compute the corresponding maximum-entropy distribution $\mathbb{P}_{\Delta(3; p_1, p_2)}$ using theorem 1. The partition function Z in (3.4) is

$$\begin{aligned} Z(\theta_1, \theta_2, \theta_3) &= \sum_{C \in \mathcal{C}_3} e^{-H(C)} = \sum_{C \in \mathcal{C}_3} e^{-\theta_1 f_1(C) - \theta_2 f_2(C) - \theta_3 \phi_2(C)} \\ &= 1 + 3e^{-\theta_1} + 3e^{-2\theta_1} + e^{-3\theta_1 - \theta_3} + e^{-3\theta_1 - \theta_2 - \theta_3}, \end{aligned} \quad (\text{A.9})$$

where the last equality follows from figure A1, where we list all complexes in \mathcal{C}_3 along with the corresponding values of observables f_1, f_2 , and ϕ_2 . To find parameters θ_1, θ_2 , and θ_3 ,

which are the Lagrange multipliers coupled to observables f_1, f_2 , and ϕ_2 , we need to solve the system of three equations (3.5), where \bar{x}_i are replaced by the expected values in (A.8):

$$\begin{aligned} \frac{3e^{-\theta_1} + 6e^{-2\theta_1} + 3e^{-3\theta_1}e^{-\theta_3} + 3e^{-3\theta_1}e^{-\theta_2}e^{-\theta_3}}{1 + 3e^{-\theta_1} + 3e^{-2\theta_1} + e^{-3\theta_1}e^{-\theta_3} + e^{-3\theta_1}e^{-\theta_2}e^{-\theta_3}} &= 3p_1, \\ \frac{e^{-3\theta_1}e^{-\theta_2}e^{-\theta_3}}{1 + 3e^{-\theta_1} + 3e^{-2\theta_1} + e^{-3\theta_1}e^{-\theta_3} + e^{-3\theta_1}e^{-\theta_2}e^{-\theta_3}} &= p_1^3 p_2, \\ \frac{e^{-3\theta_1}e^{-\theta_3} + e^{-3\theta_1}e^{-\theta_2}e^{-\theta_3}}{1 + 3e^{-\theta_1} + 3e^{-2\theta_1} + e^{-3\theta_1}e^{-\theta_3} + e^{-3\theta_1}e^{-\theta_2}e^{-\theta_3}} &= p_1^3. \end{aligned} \quad (\text{A.10})$$

After some tedious algebra, one can show that the solution is

$$e^{-\theta_1} = \frac{p_1}{1 - p_1}, \quad e^{-\theta_2} = \frac{p_2}{1 - p_2}, \quad e^{-\theta_3} = 1 - p_2. \quad (\text{A.11})$$

The partition function simplifies then to

$$Z = \frac{1}{(1 - p_1)^3}. \quad (\text{A.12})$$

Therefore, the maximum-entropy distribution is

$$\begin{aligned} \mathbb{P}_{\Delta(3;p_1,p_2)}(C) &= \frac{e^{-H(C)}}{Z} = \frac{e^{-\theta_1 f_1(C) - \theta_2 f_2(C) - \theta_3 \phi_2(C)}}{Z} \\ &= (1 - p_1)^3 \left(\frac{p_1}{1 - p_1} \right)^{f_1(C)} \left(\frac{p_2}{1 - p_2} \right)^{f_2(C)} (1 - p_2)^{\phi_2(C)} \\ &= p_1^{f_1(C)} (1 - p_1)^{3 - f_1(C)} p_2^{f_2(C)} (1 - p_2)^{\phi_2(C) - f_1(C)}. \end{aligned} \quad (\text{A.13})$$

As expected, the obtained distribution coincides with the distribution in (6.5), where $n = 3$ and $\phi_1(C) = 3$. Unfortunately, this method of computing \mathbb{P}_{Δ} cannot be extended to the general case $\Delta(n; p_1, \dots, p_{n-1})$: when $n > 3$ the partition function Z and the corresponding analog of system (A.10) become analytically intractable. This makes lemmas 1 and 2 an essential tool for proving theorem 2 and a more general theorem 3.

A.4. ERSC $(\mathcal{C}_3, \{f_1, f_2\}, \{\bar{f}_1, \bar{f}_2\})$

Here we derive the maximum-entropy distribution on \mathcal{C}_3 only under the constraints of the first type, $\mathbb{E}[f_1] = \bar{f}_1$ and $\mathbb{E}[f_2] = \bar{f}_2$, and show that it is different from $\mathbb{P}_{\Delta(3;p_1,p_2)}$. This explicitly demonstrates that the constraint of the second type, $\mathbb{E}[\phi_2] = \bar{\phi}_2$, is not redundant, and, if dropped, the resulting maximum-entropy ensemble will no longer be Δ .

Let $(\mathcal{C}_3, \tilde{\mathbb{P}})$ be the maximum-entropy ensemble ERSC($\mathcal{C}_3, \{f_1, f_2\}, \{\bar{f}_1, \bar{f}_2\}$). In other words, $\tilde{\mathbb{P}}$ is the maximum-entropy distribution on \mathcal{C}_3 under the constraints

$$\mathbb{E}[f_1] = \bar{f}_1 \quad \text{and} \quad \mathbb{E}[f_2] = \bar{f}_2. \quad (\text{A.14})$$

We can find $\tilde{\mathbb{P}}$ using theorem 1 as in the previous section. The partition function

$$\begin{aligned} \tilde{Z}(\theta_1, \theta_2) &= \sum_{C \in \mathcal{C}_3} e^{-\tilde{H}(C)} = \sum_{C \in \mathcal{C}_3} e^{-\theta_1 f_1(C) - \theta_2 f_2(C)} \\ &= (1 + e^{-\theta_1})^3 + e^{-3\theta_1} e^{-\theta_2}, \end{aligned} \quad (\text{A.15})$$

where the last equality is obtained with the help of figure A1. The system of equations (3.5) for θ_1 and θ_2 is then

$$\begin{aligned} 3 \frac{e^{-\theta_1}(1 + e^{-\theta_1})^2 + e^{-3\theta_1}e^{-\theta_2}}{(1 + e^{-\theta_1})^3 + e^{-3\theta_1}e^{-\theta_2}} &= \bar{f}_1, \\ \frac{e^{-3\theta_1}e^{-\theta_2}}{(1 + e^{-\theta_1})^3 + e^{-3\theta_1}e^{-\theta_2}} &= \bar{f}_2, \end{aligned} \quad (\text{A.16})$$

and one can check that the solution is given by

$$e^{-\theta_1} = \frac{\frac{\bar{f}_1}{3} - \bar{f}_2}{1 - \frac{\bar{f}_1}{3}} \quad \text{and} \quad e^{-\theta_2} = \frac{\bar{f}_2(1 - \bar{f}_2)^2}{\left(\frac{\bar{f}_1}{3} - \bar{f}_2\right)^3}. \quad (\text{A.17})$$

The partition function, as a function of \bar{f}_1 and \bar{f}_2 , is then

$$\tilde{Z} = \frac{(1 - \bar{f}_2)^2}{\left(1 - \frac{\bar{f}_1}{3}\right)^3}. \quad (\text{A.18})$$

Therefore, the maximum-entropy distribution is

$$\begin{aligned} \tilde{\mathbb{P}} &= \frac{e^{-\tilde{H}(C)}}{\tilde{Z}} = \frac{e^{-\theta_1 f_1(C) - \theta_2 f_2(C)}}{\tilde{Z}} \\ &= \frac{\left(1 - \frac{\bar{f}_1}{3}\right)^3}{(1 - \bar{f}_2)^2} \left(\frac{\frac{\bar{f}_1}{3} - \bar{f}_2}{1 - \frac{\bar{f}_1}{3}}\right)^{f_1(C)} \left(\frac{\bar{f}_2(1 - \bar{f}_2)^2}{\left(\frac{\bar{f}_1}{3} - \bar{f}_2\right)^3}\right)^{f_2(C)} \\ &= \left(\frac{\bar{f}_1}{3} - \bar{f}_2\right)^{f_1(C) - 3f_2(C)} \left(1 - \frac{\bar{f}_1}{3}\right)^{3 - f_1(C)} \bar{f}_2^{f_2(C)} (1 - \bar{f}_2)^{2f_2(C) - 2}. \end{aligned} \quad (\text{A.19})$$

This is a general expression for $\tilde{\mathbb{P}}$ for any expected values \bar{f}_1 and \bar{f}_2 . In the special case, when \bar{f}_1 and \bar{f}_2 coincide with the corresponding values for $\Delta(3; p_1, p_2)$ in (A.8), that is $\bar{f}_1 = 3p_1$ and $\bar{f}_2 = p_1^3 p_2$, the distribution $\tilde{\mathbb{P}}$ reduces to

$$\tilde{\mathbb{P}} = p_1^{f_1(C)} (1 - p_1)^{3 - f_1(C)} p_2^{f_2(C)} (1 - p_1^2 p_2)^{f_1(C) - 3f_2(C)} (1 - p_1^3 p_2)^{2f_2(C) - 2}. \quad (\text{A.20})$$

We see that $\tilde{\mathbb{P}} \neq \mathbb{P}_{\Delta(3; p_1, p_2)}$. This means that the two maximum-entropy ensembles $\Delta(3; p_1, p_2)$ and $(\mathcal{C}_3, \tilde{\mathbb{P}})$ are different

$$\text{ERSC}\left(\mathcal{C}_3, \{f_1, f_2, \phi_2\}, \{\bar{f}_1, \bar{f}_2, \bar{\phi}_2\}\right) \neq \text{ERSC}\left(\mathcal{C}_3, \{f_1, f_2\}, \{\bar{f}_1, \bar{f}_2\}\right), \quad (\text{A.21})$$

and, more generally

$$\begin{aligned} &\text{ERSC}\left(\mathcal{C}_n, \left\{ \left\{ f_d \right\}_{d=1}^{n-1}, \left\{ \phi_d \right\}_{d=2}^{n-1} \right\}, \left\{ \left\{ \bar{f}_d \right\}_{d=1}^{n-1}, \left\{ \bar{\phi}_d \right\}_{d=2}^{n-1} \right\}\right) \\ &\neq \text{ERSC}\left(\mathcal{C}_n, \left\{ f_d \right\}_{d=1}^{n-1}, \left\{ \bar{f}_d \right\}_{d=1}^{n-1}\right). \end{aligned} \quad (\text{A.22})$$

References

- [1] Airoidi E *et al* (ed) 2007 *Statistical Network Analysis: Models, Issues, and New Directions* (Berlin: Springer)
- [2] Anand K and Bianconi G 2009 Entropy measures for networks: toward an information theory of complex topologies *Phys. Rev. E* **80** 045102(R)
- [3] Anderson C J, Wasserman S and Crouch B 1999 A p^* primer: logit models for social networks *Soc. Netw.* **21** 37–66
- [4] Barabási A-L and Albert R 1999 Emergence of scaling in random networks *Science* **286** 509–12
- [5] Bianconi G and Rahmede C 2015 Complex quantum network manifolds in dimension $d > 2$ are scale-free *Sci. Rep.* **5** 13979
- [6] Bianconi G, Rahmede C and Wu Z 2015 Complex quantum network geometries: evolution and phase transitions *Phys. Rev. E* **92** 022815
- [7] Chatterjee S and Diaconis P 2013 Estimating and understanding exponential random graph models *Ann. Stat.* **41** 2428–61
- [8] Costa A and Farber M 2015 Homological domination in large random simplicial complexes arXiv:1503.03253
- [9] Costa A and Farber M 2015 Large random simplicial complexes I *J. Topology Anal.* at press
- [10] Costa A and Farber M 2015 Random simplicial complexes *Configuration Spaces – Geometry, Topology and Representation Theory (Springer INdAM Series)* at press
- [11] Cover T M and Thomas J A 1991 *Elements of Information Theory* (New York: Wiley-Interscience)
- [12] Dorogovtsev S N 2010 *Lectures on Complex Networks* (Oxford: Oxford University Press)
- [13] Dorogovtsev S N, Mendes J F F and Samukhin A N 2000 Structure of growing networks with preferential linking *Phys. Rev. Lett.* **85** 4633–6
- [14] Easley D and Kleinberg J 2010 *Networks, Crowds, and Markets: Reasoning about a Highly Connected World* (Cambridge: Cambridge University Press)
- [15] Edelsbrunner H 1994 Modeling with simplicial complexes: topology, geometry, and algorithms *Proc. 6th Canadian Conf. Computational Geometry* pp 36–44
- [16] Edelsbrunner H and Harer J L 2010 *Computational Topology: An Introduction* (Providence, RI: American Mathematical Society)
- [17] Erdős P and Rényi A 1959 On random graphs *Publ. Math.* **6** 290–7
- [18] Erdős P and Rényi A 1960 On the evolution of random graphs *Publ. Math. Inst. Hung. Acad. Sci.* **5** 17–61
- [19] Frank O and Strauss D 1986 Markov graphs *J. Am. Stat. Assoc.* **81** 832–42
- [20] Ghrist R and Muhammad A 2005 Coverage and hole-detection in sensor networks via homology *Proc. 4th Int. Symp. on Information Processing in Sensor Networks*
- [21] Goldenberg A, Zheng A X, Fienberg S E and Airoidi E M 2010 A survey of statistical network models *Found. Trends Mach. Learn.* **2** 129–233
- [22] Hatcher A 2002 *Algebraic Topology* (New York: Cambridge University Press)
- [23] Holland P W and Leinhardt S 1981 An exponential family of probability distributions for directed graphs *J. Am. Stat. Assoc.* **76** 33–50
- [24] Horvát S, Czabarka É and Toroczkai Z 2015 Reducing degeneracy in maximum entropy models of networks *Phys. Rev. Lett.* **114** 158701
- [25] Kahle M 2009 Topology of random clique complexes *Discrete Math.* **309** 1658–71
- [26] Kahle M 2014 Sharp vanishing thresholds for cohomology of random flag complexes *Ann. Math.* **179** 1085–107
- [27] Kahle M 2014 Topology of random simplicial complexes: a survey *AMS Contemp. Math.* **620** 201–22
- [28] Kee K F, Sparks L, Struppa D C and Mannucci M 2013 Social groups, social media, and higher dimensional social structures: a simplicial model of social aggregation for computational communication research *Commun. Q.* **61** 35–58
- [29] Kolaczyk E 2009 *Statistical Analysis of Network Data* (New York: Springer)
- [30] Krapivsky P L, Redner S and Leyvraz F 2000 Connectivity of growing random networks *Phys. Rev. Lett.* **85** 4629–32
- [31] Linial N and Meshulam R 2006 Homological connectivity of random 2-complexes *Combinatorica* **26** 475–87

- [32] Maletić S and Rajković M 2009 Simplicial complex of opinions on scale-free networks *Complex Networks (Studies in Computational Intelligence Series vol 207)* (Berlin: Springer-Verlag) pp 127–34
- [33] Maletić S and Rajković M 2014 Consensus formation on a simplicial complex of opinions *Physica A* **397** 111–20
- [34] Munkres J R 1993 *Elements of Algebraic Topology* (New York: Perseus Books)
- [35] Newman M E J 2010 *Networks: An Introduction* (Oxford: Oxford University Press)
- [36] Newman M E J and Watts D J 1999 Renormalization group analysis of the small-world network model *Phys. Lett. A* **263** 341–6
- [37] Park J and Newman M E J 2004 Statistical mechanics of networks *Phys. Rev. E* **70** 066117
- [38] Pressé S, Ghosh K, Lee J and Dill K A 2013 Principles of maximum entropy and maximum caliber in statistical physics *Rev. Mod. Phys.* **85** 1115–41
- [39] Ren W, Zhao Q, Ramanathan R, Gao J, Swami A, Bar-Noy A, Johnson M and Basu P 2013 Broadcasting in multi-radio multi-channel wireless networks using simplicial complexes *Wirel. Netw.* **19** 1121–33
- [40] Robins G, Pattison P, Kalish Y and Lusher D 2007 An introduction to exponential random graph (p^*) models for social networks *Soc. Netw.* **29** 173–91
- [41] Shalizi C R and Rinaldo A 2013 Consistency under sampling of exponential random graph models *Ann. Stat.* **41** 508–35
- [42] Snijders T A B 2002 Markov chain Monte Carlo estimation of exponential random graph models *J. Soc. Struct.* **3** 1–40
- [43] Solomonoff R and Rapoport A 1951 Connectivity of random nets *Bull. Math. Biophys.* **13** 107–17
- [44] Stasi D, Sadeghi K, Rinaldo A, Petrovic S and Fienberg S E 2014 Beta models for random hypergraphs with a given degree sequence *Proc. 21st Int. Conf. on Comput. Stat.* (arXiv:1407.1004)
- [45] Strauss D 1986 On a general class of models for interaction *SIAM Rev.* **28** 513–27
- [46] Wasserman S and Pattison P E 1996 Logit models and logistic regressions for social networks: I. An introduction to markov graphs and p^* *Psychometrika* **61** 401–25
- [47] Watts D J 1999 *Small Worlds: The Dynamics of Networks Between Order and Randomness* (Princeton, NJ: Princeton University Press)
- [48] Watts D J and Strogatz S H 1998 Collective dynamics of ‘small-world’ networks *Nature* **393** 440–2
- [49] Wu Z, Menichetti G, Rahmede C and Bianconi G 2015 Emergent complex network geometry *Sci. Rep.* **5** 10073
- [50] Zomorodian A 2012 *Topological data analysis Advances in Applied and Computational Topology* (Providence, RI: American Mathematical Society)