EXPONENTIAL RANDOM SIMPLICIAL COMPLEXES

KONSTANTIN ZUEV, OR EISENBERG AND DMITRI KRIOUKOV

Abstract. Exponential random graph models have attracted significant research attention over the past decades. These models are maximum-entropy ensembles under 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 cast as maximum-entropy ensembles under certain constraints. We introduce and analyze the most general random simplicial complex ensemble $\Delta$ with statistically independent simplices. Our analysis is simplified by the observation that any distribution $P(O)$ on any collection of objects $O = \{O\}$, including graphs and simplicial complexes, is maximum-entropy under the constraint that the expected value of $-\ln P(O)$ is equal to the entropy of the distribution. With the help of this observation, we prove that ensemble $\Delta$ is maximum-entropy under two types of constraints that fix the expected numbers of simplices and their boundaries.

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 by graphs has proved to be useful for understanding 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 [6,8,28].

A host of developed network models (e.g. see [15] for a survey) can be roughly divided into two classes: “generative” models and “descriptive” models [1]. A generative model is an algorithm that describes how to generate a network using some probabilistic rules for connecting nodes. These models mainly 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 $G(n,p)$ [11,12,34]: 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 [3,7,24] and the small-world model [29,36,37] that explain the power-law degree distributions and small distances between most nodes, two universal properties empirically observed in many real networks. Any generative model gives rise to an ensemble $(G,P)$, where $G$ is a set of all graphs the model can possibly generate and $P$ is the probability distribution on $G$, where $P(G)$ is the probability that the model generates $G \in G$. One can always readily sample from $P$ (using the network generating algorithm), but often cannot obtain a closed-form expression for $P(G)$, or even implicitly describe $P$ 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 $(G,P_\theta)$, where $G$ is a set of graphs and $P_\theta$ is the joint probability distribution on $G$ parameterized by a vector of parameters $\theta$, which are to be inferred from the observed network data. For any graph $G \in G$, a descriptive model gives a closed-form expression for $P_\theta(G)$ 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 $\mathbb{P}_\theta$, and often it is 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, and its solution could be very useful for applications.

Exponential random graphs (ERGs) [13,17,23,30], often called $p^*$ models in the social network research community [2,32,35], are among the most popular and best studied descriptive models that provide a conceptual framework for statistical modeling of network data. Let $\mathcal{G}_n$ be the set of all simple graphs (no self-loops or multiedges) with $n$ nodes, $x_1, \ldots, x_r$ be functions on $\mathcal{G}_n$, referred to as the graph observables, and let $\bar{x}_1, \ldots, \bar{x}_r$ be the values of these observables $x_1(\bar{G}), \ldots, x_r(\bar{G})$ 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 $\bar{x}_1, \ldots, \bar{x}_r$ 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 $\theta = (\theta_1, \ldots, \theta_r)$ is a vector of model parameters that satisfy

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

Whereas originally (1.1) was simply postulated and used in empirical studies [2], it was later recognized [30] 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

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

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

Despite some known problems with ERGs whose edges are not statistically independent [4,18,33], 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 by a simplicial complex — a high-dimensional analog of a graph — is conceptually more sound than the network representation, and provide 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 2-simplex on the three nodes. This is illustrated in Fig 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 be merely pairwise connected.

In general, compared to graphs, simplicial complexes allow to 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 [22], opinion formation and dynamics [26, 27], coverage and hole-detection in sensor networks [14], broadcasting in wireless networks [31], to name just a few. We remark that prior to using simplicial complexes for studying complex interactions, they were used in a rich variety of geometric problems, ranging from grid generation in finite element analysis to modeling configuration spaces of dynamical systems [9]. Further details and applications can be found in [10].

In this paper, we introduce exponential random simplicial complexes (ERSCs) that are higher dimensional generalizations of exponential random graphs, develop the formalism for ERSCs, and show that several popular generative models of random simplicial complexes — random flag complexes [19], Linial-Meshulam complexes [25], and Kahle’s multi-parameter model [21] — can all be explicitly represented as ERSCs. We also introduce the most general ensemble of random simplicial complexes $\Delta$ 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 [16].

A simplicial complex $C$ on $n$ vertices $V = \{1, \ldots, n\}$ 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 $\tau$ is called a face of simplex $\sigma$, and $\sigma$ is a coface of $\tau$. A simplex $\sigma$ 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 $\mathbb{R}^K$, $K \geq k$ [16]. 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 dimension 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

$$\mathcal{C}_n = \otimes_{d=1}^{n} a_d,$$  \hspace{1cm} (2.1)
where $a_d = \{a_{i_1, \ldots, i_d}\}$, $i_j = 1, \ldots, n$, $j = 1, \ldots, d$, is a boolean symmetric tensor of order $d$ with zeros on all its diagonals. These conditions require precisely that $a_{i_1, \ldots, i_d} = a_{i_{\kappa(1)}, \ldots, i_{\kappa(d)}}$ for any permutation $\kappa$ of subsubindices $1, \ldots, d$, and $a_{i_1, \ldots, i_d} = 0$ if $i_j = i_k$ for any pair of $j$ and $k$. The non-redundant elements of tensor $a_d$ are thus $a_{i_d}$, where by multi-index $i_d$ we denote a $d$-tuple of indices with increasing values:

$$i_d = i_1, \ldots, i_d,$$

$$1 \leq i_1 < \ldots < i_d \leq n. \quad (2.2)$$

The only requirement for $\otimes_{d=1}^n a_d$ to be bijective with $C_n$ is then the following compatibility condition:

$$a_{i_d} = 1 \implies b_{i_d} \eqdef \prod_{k=1}^d a_{i_k} = 1, \quad (2.4)$$

$$i_d^k = i_1, \ldots, \widehat{i_k}, \ldots, i_d \quad (2.5)$$

is the $(d - 1)$-long multi-index with $i_k$ omitted. Condition (2.4) simply formalizes the requirement that if the complex contains simplex $\{i_d\}$, then it also contains all its faces.

For a simplicial complex $C \in C_n$, $a_d = \{a_{i_d}\}$ is thus its “adjacency” tensor that encodes the presence of $(d-1)$-simplices: $a_{i_d} = 1$ if simplex $\{i_d\} \in C$, and zero otherwise. Since we assume that $C$ has $n$ vertices, then we trivially have $a_1 = 1_n = (1, \ldots, 1)$. 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 1-skeleton of a complex, for example, is a graph.

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

$$C^{[d+1]} = C^{(d)} \cup \{\{i_{d+2}\} : b_{i_{d+2}} = 1\}. \quad (2.6)$$
In other words, $C^{[d+1]}$ is obtained from $C^{(d)}$ by adding $(d + 1)$-simplices as follows. For every $(d + 1)$-simplex $\{i_{d+2}\}$, if $C^{(d)}$ contains all $(d + 2)$ $d$-simplices $\{i^k_{d+2}\}$, $k = 1, \ldots, d + 2$, we add $\{i_{d+2}\}$ to $C^{(d)}$. Intuitively, we add $\{i_{d+2}\}$ if its $d$-dimensional boundary is already in $C^{(d)}$. Note that in this case we add $\{i_{d+2}\}$ even if $\{i_{d+2}\} \notin C$, 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 1-skeleton of $C$ with all its triangular subgraphs filled by 2-simplices. Figure 3 illustrates the construction of a filled skeleton.

Thus, we have the following hierarchy of “empty” and “filled” skeletons:

\[ V = C^{(0)} \subset C^{(1)} \subset \ldots \subset C^{(d-1)} \subset C^{(d)} \subset \ldots \subset C^{(m-1)} \subset C^{(m)} = C, \]

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

\[ f_d = \sum_{i_{d+1}} a_{i_{d+1}} \quad \text{and} \quad \phi_d = \sum_{i_{d+1}} b_{i_{d+1}}. \quad (2.7) \]

Figure 6 shows all simplicial complexes $C \in C_3$ and the values of $f_1$, $f_2$, and $\phi_2$ for each $C$.

### 3. Exponential Random Simplicial Complexes

Let $\mathcal{S}$ be any subset of $\mathcal{C}_n$, $\{x_1, \ldots, x_r\}$ be a set of functions on $\mathcal{S}$, $x_i : \mathcal{S} \to \mathbb{R}$, and $\{\bar{x}_1, \ldots, \bar{x}_r\}$ be a set of numbers, $\bar{x}_i \in \mathbb{R}$. We define the exponential random simplicial complex (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.** 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
\begin{align}
\mathbb{E}_P[x_i] &= \sum_{C \in S} x_i(C)P(C) = \bar{x}_i, \quad (3.2) \\
\sum_{C \in S} P(C) &= 1. \quad (3.3)
\end{align}

We can define ERSC for any set of simplicial complexes, but, for the most of the paper, we restrict ourselves to \( \mathcal{C}_n \) and its subsets. If we use \( S = G_n \subset \mathcal{C}_n \), 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 \( P \) defined by \((3.1)-(3.3)\) can be written as follows
\begin{align}
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 S} e^{-H(C)}, \quad (3.4)
\end{align}
where \( H(C) \) is the Hamiltonian of simplicial complex \( C \in S \), \( Z(\theta) \) is the normalizing constant, called the partition function, and \( \theta = (\theta_1, \ldots, \theta_r) \) are the parameters satisfying the following system of \( r \) equations
\begin{align}
-\frac{\partial \ln Z}{\partial \theta_i} &= \bar{x}_i. \quad (3.5)
\end{align}

The proof is nearly identical to the proof for ERGs \([30]\), 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 \( \theta_1, \ldots, \theta_r \) and \( \alpha \) be the Lagrange multipliers for the constraints in \((3.2)\) and \((3.3)\). The Lagrangian is then
\begin{equation}
\mathcal{L} = -\sum_{C \in S} P(C) \ln P(C) + \sum_{i=1}^r \theta_i \left( \bar{x}_i - \sum_{C \in S} x_i(C)P(C) \right) + \alpha \left( 1 - \sum_{C \in S} P(C) \right). \quad (3.6)
\end{equation}
The maximum entropy is achieved if the distribution \( P \) satisfies \( \frac{\partial \mathcal{L}}{\partial \theta_i} = 0 \) for any \( C \in S \). This gives
\begin{equation}
-\ln P(C) - 1 - \sum_{i=1}^r \theta_i x_i(C) - \alpha = 0, \quad (3.7)
\end{equation}
or,
\begin{equation}
P(C) \propto \exp \left( -\sum_{i=1}^r \theta_i x_i(C) \right), \quad (3.8)
\end{equation}
which is equivalent to \((3.4)\) since \( \sum_{C \in S} P(C) = 1 \). It remains to check that \((3.5)\) indeed holds:
\begin{equation}
-\frac{\partial \ln Z}{\partial \theta_i} = -\frac{1}{Z} \frac{\partial}{\partial \theta_i} \sum_{C \in S} e^{-H(C)} = \frac{1}{Z} \sum_{C \in S} \frac{\partial H(C)}{\partial \theta_i} e^{-H(C)}
= \frac{1}{Z} \sum_{C \in S} x_i(C) e^{-H(C)} = \sum_{C \in S} x_i(C)P(C) = \bar{x}_i, \quad (3.9)
\end{equation}
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 1-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 $\binom{n}{2}p$ \cite{er2}:

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

4.2. Random Flag Complexes. The flag complex $X(G)$ of a graph $G \in G_n$, also called the clique complex or the Vietoris-Rips complex, is a (deterministic) simplicial complex in $C_n$ whose 1-skeleton is $G$ and $k$-simplices correspond to complete subgraphs of $G$, called cliques, of size $k + 1$. Since any simplicial complex is homeomorphic to a flag complex, they arise in different applications, and are often used for topological data analysis \cite{tda}.

Kahle \cite{k1,k2} 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

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

Before giving the proof, we comment on what exactly Proposition 4 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 4 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 1-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:

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

Therefore, indeed $\mathcal{S}_{X(n,p)} = \mathcal{F}_n$. To prove (4.2), we need to show that $\mathbb{P}_{X(n,p)}$ is in fact the ERSC probability distribution (4.4), (4.5). Since every flag complex $C \in \mathcal{F}_n$ is completely defined by the adjacency matrix of its 1-skeleton, $\mathcal{F}_n = \otimes_{d=1}^2 a_d = \mathcal{I}_n \otimes a_2$. The partition function $Z$ can then be computed as follows:

$$Z(\theta_1) = \sum_{C \in \mathcal{F}_n} e^{-H(C)} = \sum_{C \in \mathcal{F}_n} e^{-\theta_1 f_1(C)} = \sum_{a_2} e^{-\theta_1 \sum_{i_2} a_{i_2}} = \sum_{a_2} \prod_{i_2} e^{-\theta_1 a_{i_2}} = \prod_{i_2} \sum_{a_{i_2}=0}^{1} e^{-\theta_1 a_{i_2}} = \left( 1 + e^{\theta_1} \right)^{\binom{n}{2}}.$$ \hfill (4.4)

We can now solve (3.5) with $\bar{x}_1 = \left( \binom{n}{2} \right)p$ for the parameter $\theta_1$.

$$\theta_1 = - \ln \frac{p}{1 - p},$$ \hfill (4.5)

and check that indeed

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

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 1-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 2-complexes. To generate $Y \sim Y(n, p)$, we start with a complete graph on $n$ vertices, the 1-skeleton of a future simplicial complex, and then add each of the \( \binom{n}{2} \) possible triangle faces independently at random with probability $p$. Linial and Meshulam introduced this model in [25] 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 [11]. The model $Y(n, p)$ can be readily generalized for higher dimensions: start with a full $d$-complex on $n$ vertices, $1 \leq d \leq n - 2$, and then 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 $C^{(d+1)}_n \subset C_n$ be a set of all simplicial complexes of dimension $(d+1)$ or less, and $\mathcal{Y}_d \subset C^{(d+1)}_n$ be a subset of complexes with full $d$-skeleton. In other words,

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

Since for any $C \in \mathcal{Y}_d$, the first $(d+1)$ adjacency tensors $a_1, \ldots, a_{d+1}$ are unit tensors with zero diagonals, $\mathcal{Y}_d = a_{d+2}$.

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

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

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

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

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

$$Z(\theta_1) = \sum_{C \in \mathcal{Y}_d} e^{-H(C)} = \sum_{C \in \mathcal{Y}_d} e^{-\theta_1 f_{d+1}(C)} = \sum_{a_{d+2}} \sum_{\mathcal{Y}_d} e^{-\theta_1 \sum_{l_{d+2}} a_{l_{d+2}}} \quad (4.10)$$

$$= \sum_{a_{d+2}} \prod_{l_{d+2}} e^{-\theta_1 a_{l_{d+2}}} = \prod_{l_{d+2}} \sum_{a_{l_{d+2}}=0}^{1} e^{-\theta_1 a_{l_{d+2}}} = (1 + e^{\theta_1})^{\binom{n}{d+2}}.$$ 

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

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

as claimed. \qed

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 that now we are creating $(d+1)$-simplices instead of 1-simplices (edges).

5. ANY DISTRIBUTION IS MAXIMUM-ENTROPY

To deal with more complicated ERSCs, we will rely on the fact that any given distribution $\mathbb{P}$ is maximum-entropy under the constraint that the expected value of $-\ln \mathbb{P}$ is equal to the entropy of the distribution.

Let us consider a simplicial complex ensemble $(S, \mathbb{P}^*)$, where $\mathbb{P}^*$ is some fixed probability distribution on $S \subset C_n$. Define the Hamiltonian of $C \in S$ as

$$H^*(C) = -\ln \mathbb{P}^*(C), \quad (5.1)$$
and let $\bar{H}^*$ denote the expectation of the observable $H^* : S \to \mathbb{R}$ with respect to $\mathbb{P}^*$, which is exactly the entropy of $\mathbb{P}^*$,

$$\bar{H}^* = \mathbb{E}_{\mathbb{P}^*}[H^*] = \sum_{C \in S} H^*(C) \mathbb{P}^*(C) = S(\mathbb{P}^*).$$  \hfill (5.2)

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

$$S(\mathbb{P}) \to \max, \quad \sum_{C \in S} \mathbb{P}(C) = 1, \quad \mathbb{E}_{\mathbb{P}}[H^*] = \bar{H}^*.$$  \hfill (5.3)

Or, equivalently,

$$(S, \mathbb{P}^*) = \text{ERSC} \left( S, H^*, \bar{H}^* \right).$$  \hfill (5.4)

**Lemma 1** is by no means specific to simplicial complexes. It can be formulated for any ensemble of “objects.” Even more generally, it applies to any discrete or continuous distribution. It states that any probability distribution is a maximum-entropy distribution under the constraint that the expected value of its Hamiltonian is the entropy of that distribution. We will see that this fact simplifies dramatically the proofs for ERSCs in the subsequent section. Yet we will need a more general version of **Lemma 1**.

Suppose that the Hamiltonian $H^*$ can be written as a linear combination of $r$ simplicial complex observables $h_i^* : S \to \mathbb{R}$, and a constant $\xi \in \mathbb{R}$:

$$H^*(C) = -\ln \mathbb{P}^*(C) = \sum_{i=1}^{r} \lambda_i h_i^*(C) + \xi,$$  \hfill (5.5)

and let $\bar{h}_i^*$ denote the expectation of $h_i^*$ with respect to $\mathbb{P}^*$,

$$\bar{h}_i^* = \mathbb{E}_{\mathbb{P}^*}[h_i^*] = \sum_{C \in S} h_i^*(C) \mathbb{P}^*(C).$$  \hfill (5.6)

**Lemma 2.** *Probability distribution $\mathbb{P}^*$ is the solution of the following optimization problem:*

$$S(\mathbb{P}) \to \max, \quad \sum_{C \in S} \mathbb{P}(C) = 1, \quad \mathbb{E}_{\mathbb{P}}[h_i^*] = \bar{h}_i^*, \quad i = 1, \ldots, r.$$  \hfill (5.7)

Or, equivalently,

$$(S, \mathbb{P}^*) = \text{ERSC} \left( S, \{h_i^*\}, \{\bar{h}_i^*\} \right).$$  \hfill (5.8)

We remark that **Lemma 1** is a special case of **Lemma 2** with $\xi = 0, r = 1$, and $\lambda_1 = 1$. The proof is similar to Theorem 11.1.1 in [5].

**Proof.** Let $\mathbb{P}$ be any distribution that satisfies the constraints in (5.7). Then its entropy

$$S(\mathbb{P}) = -\sum_{C \in S} \mathbb{P}(C) \ln \mathbb{P}(C) = -\sum_{C \in S} \mathbb{P}(C) \ln \frac{\mathbb{P}(C) \mathbb{P}^*(C)}{\mathbb{P}^*(C)}$$

$$= -D_{KL}(\mathbb{P} \parallel \mathbb{P}^*) + \sum_{C \in S} \mathbb{P}(C) H^*(C),$$  \hfill (5.9)

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

$$S(\mathbb{P}) \geq \sum_{C \in S} \mathbb{P}(C) H^*(C) = \sum_{C \in S} \mathbb{P}(C) \left( \sum_{i=1}^{r} \lambda_i h_i^*(C) + \xi \right)$$

$$= \sum_{i=1}^{r} \lambda_i \mathbb{E}_{\mathbb{P}}[h_i^*] + \xi = \sum_{i=1}^{r} \lambda_i \bar{h}_i^* + \xi = S(\mathbb{P}^*).$$  \hfill (5.10)
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 $\mathbb{P} = \mathbb{P}^*$. □

In many cases, Lemmas 1&2 are not useful, since for many generative models $(\mathcal{S}, \mathbb{P}^*)$ the distribution $\mathbb{P}^*$ is not known, and so neither are Hamiltonian $H^*$ and observables $\{h^*_i\}$. In generative models, the complexity of the generating algorithm makes it often impossible to explicitly write the probability $\mathbb{P}^*(C)$ for a given $C$.

In preferential attachment [3], where the algorithm that grows 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 $\mathbb{P}^*$, 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 $G \in \mathcal{G}_n$ to be generated by the model is

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

The corresponding Hamiltonian is then

$$
H_G(n,p)(G) = - f_1(G) \ln p - \left( \frac{n}{2} - f_1(G) \right) \ln(1 - p)
$$

$$
= \ln \left( \frac{1 - p}{\lambda_1} \right) f_1(G) - \left( \frac{n}{2} \right) \ln(1 - p),
$$

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 $E_{\mathbb{P}_G(n,p)}[f_1] = \binom{n}{2}p$. Similarly, for $X(n,p)$ and $Y_d(n,p)$,

$$
H_{X(n,p)}(C) = \ln \frac{1 - p}{p} f_1(C) - \left( \frac{n}{2} \right) \ln(1 - p),
$$

$$
\bar{f}_1 = E_{X(n,p)}[f_1] = \binom{n}{2}p,
$$

$$
H_{Y_d(n,p)}(C) = \ln \frac{1 - p}{p} f_1(C) - \left( \frac{n}{d + 1} \right) \ln(1 - p),
$$

$$
\bar{f}_{d+1} = E_{Y_d(n,p)}[f_{d+1}] = \binom{n}{d + 2}p,
$$

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 known, the computation of the partition function, which tends to be a nontrivial task in general, is not necessary.

6. Kahle’s $\Delta$-Ensembles

We now turn to a more general model that has the Erdős-Rényi random graphs, the random flag complexes, and the Linial-Meshulam complexes as special cases. In a recent survey [21], Kahle introduced the following multi-parameter model $\Delta(n; p_1, \ldots, p_{n-1})$ that generates random simplicial complexes inductively by dimension. First, built a 1-skeleton by putting an edge between any two vertices with probability $p_1$. Then, for $d = 2, \ldots, n-1$, 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. $\Delta(n; p_1, \ldots, p_{n-1})$ is a random simplicial complex model that generates $C \in \mathcal{C}_n$ as follows: for $d = 1, \ldots, n-1$, for every $i_{d+1}$,
\[
\begin{align*}
&\text{if } b_{d+1} = 0 \Rightarrow \text{set } a_{i_{d+1}} = 0, \\
&\text{if } b_{d+1} = 1 \Rightarrow \text{set } a_{i_{d+1}} = \begin{cases} 1 & \text{with probability } p_d, \\ 0 & \text{with probability } 1 - p_d. \end{cases}
\end{align*}
\]

In Appendix A.1 we prove the following proposition.

Proposition 3. Let $C \sim \Delta(n; p_1, \ldots, p_{n-1})$. The expected numbers of $d$-simplices in $C^{(d)}$ and $C^{[d]}$ are
\[
\tilde{f}_d = \left( \frac{n}{d+1} \right) \prod_{k=1}^{d+1} p_{k-1} \quad \text{and} \quad \bar{f}_d = \left( \frac{n}{d+1} \right) \prod_{k=1}^{d} p_{k-1}.
\]

The Kahle’s model unifies all random simplicial complexes we have considered so far:
\[
\begin{align*}
&G(n, p) = \Delta(n; p, 0, \ldots, 0), \\
&X(n, p) = \Delta(n; p, 1, \ldots, 1), \\
&Y(n, p) = \Delta(n; 1, p, 0, \ldots, 0), \\
&Y_d(n, p) = \Delta(n; 1, \ldots, 1, p, 0, \ldots, 0).
\end{align*}
\]

Since all these special cases are ERSCs, it is natural to expect that so is $\Delta(n; p_1, \ldots, p_{n-1})$. We cannot prove this in general 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 computing the partition function for $\Delta(n; p_1, \ldots, p_{n-1})$ with $n = 3$. However, with the help of Lemmas 1 & 2 in Section 5 there exists a simpler alternative proof. The fact that $\Delta(n; p_1, \ldots, p_{n-1})$ is an ERSC is a direct corollary of those lemmas.

Theorem 2. The Kahle’s $\Delta$-ensemble is the ERSC ensemble:
\[
\Delta(n; p_1, \ldots, p_{n-1}) = \text{ERSC } \left( \mathcal{C}_n, \{ \{ f_d \}_{d=1}^{n-1}, \{ \bar{f}_d \}_{d=2}^{n-1} \}, \{ \{ \tilde{f}_d \}_{d=1}^{n-1}, \{ \tilde{f}_d \}_{d=2}^{n-1} \} \right),
\]
where $\tilde{f}_d$ and $\bar{f}_d$ are the expected numbers of $d$-simplices in $C^{(d)}$ and $C^{[d]}$.

Proof. For any $C \in \mathcal{C}_n$, the probability $P_\Delta(C)$ that $\Delta(n; p_1, \ldots, p_{n-1})$ generates $C$, $C \sim \Delta(n; p_1, \ldots, p_{n-1})$, follows by an induction argument:
\[
P_\Delta(C) = \prod_{d=1}^{n-1} P_\Delta \left( C^{(d)} \bigg| C^{(d-1)} \right) = \prod_{d=1}^{n-1} p_d^{f_d(C)} \left( 1 - p_d \right)^{\phi_d(C) - f_d(C)}.
\]

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 $P_\Delta \left( C^{(d)} \big| C^{(d-1)} \right) = p_d^{f_d(C)} \left( 1 - p_d \right)^{\phi_d(C) - f_d(C)}$, where $f_d(C)$ is the actual number of $d$-simplices in $C^{(d)}$.

The Hamiltonian of $C$ is therefore
\[
H_\Delta(C) = \sum_{d=1}^{n-1} \left( f_d(C) \log \frac{1 - p_d}{p_d} + \phi_d(C) \log \frac{1}{1 - p_d} \right)
\]
\[
= \sum_{d=1}^{n-1} f_d(C) \log \frac{1 - p_d}{p_d} + \sum_{d=2}^{n-1} \phi_d(C) \log \frac{1}{1 - p_d} + \left( \frac{n}{2} \right) \log \frac{1}{1 - p},
\]

\[
\]
adjacency tensors

a complexes with

\( \Delta \) which are not necessarily equal to 1. We denote this new model by

\( \text{general as possible, we must allow for even the 0-simplices (vertices) to be present with any probabilities, independent simplices. In this case each simplex has its own individual probability of appearance. To stay completes the proof.}

\[ \begin{align*}
(6.7) \\
\xi = n \ln \frac{1}{1 - p_1},
\end{align*} \]

completes the proof. \( \square \)

7. General Random Simplicial Complexes with Independent Simplices

We finally 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 0-simplices (vertices) to be present with any probabilities, which are not necessarily equal to 1. We denote this new model by \( \Delta(n; \mathbf{p}_1, \ldots, \mathbf{p}_n) \), or \( \Delta \) for brevity, where \( \mathbf{p}_d = \{ p_{d,i} \} \) is a collection of \( \binom{n}{d} \) appearance probabilities for each \( (d - 1) \)-simplex. Whereas in \( \Delta \) in the previous section, the subindex \( d \) in \( \mathbf{p}_d \) refers to the simplex dimension, in \( \Delta \) the sub-multi-index \( \mathbf{i}_d \) in \( p_{d,i} \) refers to the specific \( (d - 1) \)-simplex \( \{ \mathbf{i}_d \} \). To generate \( C \sim \Delta \), we first create its 0-skeleton by having vertices \( \{ \mathbf{i}_1 \} \in C \) with probabilities \( p_{1,i} \), \( i_1 = 1, \ldots, n \). That is, vertex 1 exists with probability \( p_1 \), vertex 2 has probability \( p_2 \), and so on. Then, for \( d = 1, \ldots, n - 1 \), we add every \( d \)-simplex \( \{ \mathbf{i}_{d+1} \} \) with probability \( p_{d+1,i} \), but only if the entire \( (d - 1) \)-dimensional boundary of that simplex is already in place. Figure 4 illustrates the generation of a 2-complex from \( \Delta \) with \( n = 10 \).

More formally, we have the following definition. Let \( C_{\leq n} = \cup_{k=0}^{n} C_k \) denote the set of all simplicial complexes with \( n \) vertices or less. As in Section 2, any \( C \in C_{\leq n} \) is uniquely determined by a collection of its adjacency tensors \( \mathbf{a}_d = \{ a_{d,i} \}, d = 1, \ldots, n \), 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.** $\Delta(n; p_1, \ldots, p_n)$ is a random simplicial complex model that generates $C \in \mathcal{C}_{\leq n}$ as follows: for $d = 0, \ldots, n - 1$, for every $i_d$, 

\[
\begin{align*}
\text{if } b_{i_d+1} = 0 & \Rightarrow \text{ set } a_{i_d+1} = 0, \\
\text{if } b_{i_d+1} = 1 & \Rightarrow \text{ set } a_{i_d+1} = \begin{cases} 1 & \text{with probability } p_{i_d+1}, \\ 0 & \text{with probability } 1 - p_{i_d+1}. \end{cases}
\end{align*}
\] (7.1)

Here, for convenience, we use a natural convention that $b_i = 1$ for all $i = 1, \ldots, n$. If $p_{i_d+1} = p_d$ for $d = 0, \ldots, n - 1$ and $p_0 = 1$, then we recover the original Kahle’s model $\Delta(n; p_1, \ldots, p_{n-1})$ from the previous section.

To give the expressions for the expected values of the observables $a_{i_d}$ and $b_{i_d}$ in $\Delta(n; p_1, \ldots, p_n)$, we need a bit of new notation. Let multi-index $k_m = k_1, \ldots, k_m$ denote an $m$-tuple with increasing values $1 \leq k_1 < \ldots < k_m \leq d$, and $i_k = i_1, \ldots, k_1, \ldots, k_{m-1}, \ldots, i_d$ be the $(d-m)$-long multi-index with $i_k, \ldots, i_m$ omitted, with a convention that $k_0 = \emptyset$ and $i_k = i_d$. In Appendix A.2 we prove the following proposition.

**Proposition 4.** Let $C \sim \Delta(n; p_1, \ldots, p_n)$. The expected values of the observables $a_{i_d}$ and $b_{i_d}$ are

\[
\bar{a}_{i_d} = \prod_{m=0}^{d-1} p_{i_d} \quad \text{and} \quad \bar{b}_{i_d} = \prod_{m=1}^{d-1} p_{i_d},
\] (7.2)

**Lemma 2** helps again to prove that the general model $\Delta$ is also an ERSC.

**Theorem 3.** The $\Delta$-ensemble is the ERSC ensemble:

\[
\Delta(n; p_1, \ldots, p_n) = \text{ERSC} \left( \mathcal{C}_{\leq n}, \{ \{ a_{i_d} \}_{d=1}^n, \{ b_{i_d} \}_{d=2}^n \}, \{ \{ \bar{a}_{i_d} \}_{d=1}^n, \{ \bar{b}_{i_d} \}_{d=2}^n \} \right),
\] (7.3)

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

**Proof.** The probability that $\Delta(n; p_1, \ldots, p_n)$ generates $C \in \mathcal{C}_{\leq n}$ is

\[
\mathbb{P}_\Delta(C) = \mathbb{P}_\Delta \left( C^{(0)} \right) \prod_{d=1}^{n-1} \mathbb{P}_\Delta \left( C^{(d)} | C^{(d-1)} \right)
\]

\[
= \prod_{i_1} a_{i_1} (1 - p_1)^{1 - a_{i_1}} \times \prod_{d=1}^{n-1} \prod_{i_{d+1}} b_{i_{d+1}} (1 - p_{i_{d+1}})^{b_{i_{d+1}} - a_{i_{d+1}}} 
\] (7.4)

\[
= \prod_{d=1}^{n-1} \prod_{i_d} a_{i_d} (1 - p_{i_d})^{b_{i_d} - a_{i_d}}
\]

The Hamiltonian of $C$ is then

\[
H_\Delta(C) = \sum_{d=1}^{n} \sum_{i_d} a_{i_d} \ln \frac{1 - p_{i_d}}{p_{i_d}} + b_{i_d} \ln \frac{1}{1 - p_{i_d}},
\] (7.5)

\[
= \sum_{d=1}^{n} \sum_{i_d} \alpha_{i_d} a_{i_d} + \sum_{d=2}^{n} \sum_{i_d} \beta_{d} b_{i_d} + \sum_{i_1} \ln \frac{1}{1 - p_{i_1}},
\]

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

\[
\alpha_{i_d} = \ln \frac{1 - p_{i_d}}{p_{i_d}} \quad \text{and} \quad \beta_{i_d} = \ln \frac{1}{1 - p_{i_d}}.
\] (7.6)
Using Lemma 2 with
\[ \{h_i^*\} = \{\{a_{i1}\}, \ldots, \{a_{in}\}, \{b_{i2}\}, \ldots, \{b_{in}\}\}, \]
\[ \{\lambda_i\} = \{\{a_{i1}\}, \ldots, \{a_{in}\}, \{\beta_{i2}\}, \ldots, \{\beta_{in}\}\}, \]
\[ \xi = \sum_{i} \ln \frac{1}{1 - \overline{p_i}}, \]
completes the proof. \(\square\)

8. Discussion

In summary, exponential random simplicial complexes (ERSCs) are a natural higher dimensional analog of exponential random graphs used extensively 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 \(\Delta\) with statistically independent simplices. This model contains as special cases several popular models studied in the literature: Erdős-Rényi random graphs, random flag complexes, Linial-Meshulam complexes, and Kahle’s \(\Delta\)-ensembles. As all these models, \(\Delta\) is an ERSC ensemble. The constraints in this ensemble are expected number of simplices and their boundaries.

This result a direct corollary of the general observation that any probability distribution \(P\) is maximum-entropy under the constraint that the expected value of \(-\ln P\) is equal to the entropy of \(P\). 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 exponential random graphs 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 exponential random graphs with statistically independent edges \([30]\). However the analogous calculation for the general case of random simplicial complexes \(\Delta\) with statistically independent simplices appears to be intractable.

The multi-parameter model \(\Delta(n; p_1, \ldots, p_n)\) is the ERSC ensemble with two types of constrained observables: \(\{a_{id}\}\) and \(\{b_{id}\}\). The first type observables are simplices themselves: \(a_{id}(C) = 1\) if the \((d - 1)\)-simplex \(\{i_d\}\) belongs to \(C\), and zero otherwise. The second type observables are their boundaries: \(b_{id}(C) = 1\) if the entire \((d - 2)\)-dimensional boundary of simplex \(\{i_d\}\) belongs to \(C\), and zero otherwise. Theorem 3 states that \(\Delta(n; p_1, \ldots, p_n)\) is a solution of the following optimization problem:

\[ S(P) \rightarrow \max, \quad \sum_{C \in \mathcal{C}_{\leq n}} P(C) = 1, \quad (8.1) \]
\[ \mathbb{E}_P[a_{id}] = \overline{a}_{id}, \quad d = 1, \ldots, n, \quad \mathbb{E}_P[b_{id}] = \overline{b}_{id}, \quad d = 2, \ldots, n. \quad (8.2) \]

If we drop the second type observables in this optimization problem, we alter the maximum-entropy distribution as illustrated in Figure 5. Since the distribution has changed, ensemble \(\Lambda(n; p_1, \ldots, p_n) = \text{ERSC} (\mathcal{C}_{\leq n}; \{a_{id}\}_{d=1}^{n}; \{\overline{a}_{id}\}_{d=1}^{n})\) defined by this distribution is now also different from \(\Delta(n; p_1, \ldots, p_n)\).

The fact that the second type of boundary-presence observables are also constrained in \(\Delta(n; p_1, \ldots, p_n)\) may appear quite unexpected at the first glance. The reason for the presence of these constraints is that simplex existence probabilities are actually conditional, and the conditions are the presence of simplex boundaries. If we go from conditional to unconditional probabilities, we change \(\Delta\) to \(\Lambda\). Indeed, in \(\Delta\), \(p_{id}\) is the conditional probability of the \((d - 1)\)-simplex \(\{i_d\}\) to appear in \(C\), given that its \((d - 2)\)-dimensional boundary is already in place,

\[ p_{id} = \frac{P_{\Delta}(a_{id} = 1, b_{id} = 1)}{P_{\Delta}(b_{id} = 1)} = \frac{P_{\Delta}(a_{id} = 1)}{P_{\Delta}(b_{id} = 1)}, \quad (8.3) \]
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 $C_{\leq n}$. The global maximum corresponds to the uniform distribution $U$, which is the maximum-entropy distribution among all distributions supported on $C_{\leq n}$. Theorem 3 shows that if we have two sets of constraints, $E_P[a_{id}] = \bar{a}_{id}$ and $E_P[b_{id}] = \bar{b}_{id}$, then the resulting maximum-entropy distribution is $P_\Delta$. If we drop the second set of constraints, then we get some other maximum-entropy distribution $P_\Lambda \neq P_\Delta$ for ensemble $\Lambda \neq \Delta$.

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

$$P_\Delta(a_{id} = 1) = p_{id} P_\Delta(b_{id} = 1),$$

(8.4)

and, therefore, the expected values of observables $a_{id}$ and $b_{id}$ satisfy

$$\bar{a}_{id} = E_\Delta[a_{id}] = P_\Delta(a_{id} = 1) = p_{id} P_\Delta(b_{id} = 1) = p_{id} \bar{b}_{id}. $$

(8.5)

Thus, if we want to represent $\Delta$ as an ERSC and we fixed the expected values of the first type observables $a_{id}$, we must also fix the expected values of the second type observables $b_{id}$. Moreover, these expected values are not independent and must satisfy $a_{id} = p_{id} \bar{b}_{id}$, which is consistent with Proposition 4. In Appendix A.4 we consider a special case with $n = 3$, $p_1 = p_2 = p_1$, and $p_3 = p_2$, 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, ensemble $\Lambda$, with only the first type observables constrained, appears more natural than $\Delta$. Yet $\Delta$ is more natural than $\Lambda$ 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 $P_\Lambda(C)$ in ensemble $\Lambda$, and to efficiently sample from it, i.e., to easily generate simplicial complexes $C$ with this probability.

Acknowledgments

We thank Alexandru Suciu, 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.
A.1. Proof of Proposition 3. Let us first compute the expected number of $d$-simplices in $C^{[d]}$, where $C \sim \Delta(n; p_1, \ldots, p_{n-1})$.

\[
\tilde{\phi}_d = \mathbb{E}[\phi_d] = \mathbb{E} \left[ \sum_{i_{d+1}} b_{i_{d+1}} \right] = \mathbb{E} \left[ \sum_{i_{d+1}} \prod_{k=1}^{d+1} a_{i_{d+1}^k} \right] = \mathbb{E} \left[ \sum_{i_{d+1}} \prod_{k=1}^{d+1} \mathbb{E} \left[ a_{i_{d+1}^k} \mid a_{d-1} \right] \right]. \tag{A.6}
\]

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

\[
\mathbb{E} \left[ a_{i_{d+1}^k} \mid a_{d-1} \right] = p_{d-1} b_{i_{d+1}^k}.
\tag{A.7}
\]

So,

\[
\tilde{\phi}_d = \mathbb{E} \left[ \sum_{i_{d+1}} \prod_{k=1}^{d+1} p_{d-1} b_{i_{d+1}^k} \right] = p_{d-1} \mathbb{E} \left[ \sum_{i_{d+1}} \prod_{k=1}^{d+1} a_{i_{d+1}^k} \right], \tag{A.8}
\]

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

\[
\tilde{\phi}_d = p_{d-1}^{d+1} \mathbb{E} \left[ \sum_{i_{d+1}} \prod_{k_2} a_{i_{d+1}^k} \mid a_{d-2} \right] = p_{d-1}^{d+1} \mathbb{E} \left[ \sum_{i_{d+1}} \prod_{k_3} a_{i_{d+1}^k} \mid a_{d-2} \right] = \ldots \tag{A.9}
\]

The last equation holds because $a_{i_{d+1}^k} = 1$ for any $i_{d+1}$ and $k$, since we all simplicial complexes $C \sim \Delta(n; p_1, \ldots, p_{n-1})$ have exactly $n$ vertices. The expected number of $d$-simplices in $C^{(d)}$ is now:

\[
\tilde{f}_d = \mathbb{E}[f_d] = \mathbb{E}[\mathbb{E}[f_d|\phi_d]] = \mathbb{E}[p_d \phi_d] = p_d \tilde{\phi}_d = \left( \frac{n}{d+1} \right) \prod_{k=1}^{d} p_k^{(d+1)}.
\tag{A.10}
\]
A.2. Proof of Proposition 4

Computations are similar to those in the previous section.

\[ \bar{b}_{k_d} = \mathbb{E}[b_{k_d}] = \mathbb{E}\left[ \prod_{k=1}^{d} a_{i_{kd}}^{-k_d} \right] = \mathbb{E}\left[ \prod_{k=1}^{d} \mathbb{E}\left[ a_{i_{kd}}^{-k_d} | a_{d-2} \right] \right] \]

\[ = \mathbb{E}\left[ \prod_{k=1}^{d} \mathbb{E}\left[ a_{i_{kd}} | a_{d-2} \right] \right] = \mathbb{E}\left[ \prod_{k=1}^{d} p_{i_{kd}} b_{k_d} \right] = \prod_{k=1}^{d} p_{i_{kd}} \mathbb{E}\left[ \prod_{k=1}^{d} b_{k_d} \right] = \ldots \]

\[ = \prod_{k=1}^{d} p_{i_{kd}} \cdots \prod_{k_{d-1}}^{d} \mathbb{E}\left[ \prod_{k=1}^{d} b_{k_{d-1}} \right] = \prod_{m=1}^{d-1} \prod_{k_{d-1}}^{d} p_{k_{d-1}} \]

since \( b_{k_{d-1}} = 1 \) for any \( i_d \) and \( k_{d-1} \). Finally,

\[ \tilde{a}_{i_d} = \mathbb{E}[a_{i_d}] = \mathbb{E}[\mathbb{E}[a_{i_d} | h_{i_d}]] = \mathbb{E}[p_{i_d} b_{i_d}] = \prod_{m=0}^{d-1} \prod_{k_m}^{d-1} p_{k_m}. \]

(A.12)

A.3. Special case: \( \Delta(3; p_1, p_2) \). Theorem 2 in Section 6 explicitly represents the Kahle’s multi-parameter model of random simplicial complexes \( \Delta(n; p_1, \ldots, p_{n-1}) \) as an ERSC for any values of the parameters. This theorem is a direct corollary of Lemmas 4, 6 in Section 5 that 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 \( P_\Delta \) using Theorem 1. We successfully used this method, which is based on computing the partition function, in Section 6 for the Erdős-Rényi random graphs and the random flag and Linial-Meshulam complexes. For Kahle’s \( \Delta \)-ensemble, however, the partition function becomes intractable.

Consider a special case of the 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 3 vertices with three constraints:

\[ \mathbb{E}[f_1] = 3p_1, \quad \mathbb{E}[f_2] = p_1^3 p_2, \quad \mathbb{E}[\phi_2] = p_1^3. \]

(A.13)

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

\[ Z(\theta_1, \theta_2, \theta_3) = \sum_{C \in C_3} e^{-H(C)} = \sum_{C \in C_3} e^{-\theta_1 f_1(C) - \theta_2 f_2(C) - \theta_3 \phi_2(C)} \]

\[ = 1 + 3e^{-\theta_1} + 3e^{-2\theta_1} + e^{-3\theta_1} + e^{-3\theta_1 - \theta_2 - \theta_3}, \]

where the last equality follows from Figure 6 where we list all complexes in \( C_3 \) along with the corresponding values of observables \( f_1, f_2, \) and \( \phi_2 \). To find parameters \( \theta_1, \theta_2, \) and \( \theta_3 \), which are the Lagrange multipliers coupled to observables \( f_1, f_2, \) and \( \phi_2 \), we need to solve the system of three equations (3.5), where \( \tilde{x}_i \) are replaced by the expected values in (A.13):

\[ \frac{3e^{-\theta_1} + 6e^{-2\theta_1} + 3e^{-3\theta_1} e^{-\theta_2} e^{-\theta_3}}{1 + 3e^{-\theta_1} + 3e^{-2\theta_1} + e^{-3\theta_1} e^{-\theta_2} e^{-\theta_3}} = 3p_1, \]

\[ e^{-3\theta_1} e^{-\theta_2} e^{-\theta_3} = p_1^3 p_2, \]

(A.15)

\[ \frac{3e^{-\theta_1} + 6e^{-2\theta_1} + 3e^{-3\theta_1} e^{-\theta_2} e^{-\theta_3}}{1 + 3e^{-\theta_1} + 3e^{-2\theta_1} + e^{-3\theta_1} e^{-\theta_2} e^{-\theta_3}} = p_1^3. \]
After some tedious algebra, one can show that the solution is
\[ e^{-\theta_1} = \frac{p_1}{1-p_1}, \quad e^{-\theta_2} = \frac{p_2}{1-p_2}, \quad e^{-\theta_3} = 1 - p_2. \] (A.16)

The partition function simplifies then to
\[ Z = \frac{1}{(1-p_1)^3}. \] (A.17)

Therefore, the maximum-entropy distribution is
\[ P_{\Delta(3;p_1,p_2)}(C) = \frac{e^{-H(C)}}{Z} = \frac{e^{-\theta_1 f_1(C)-\theta_2 f_2(C)-\theta_3 \phi_2(C)}}{Z} = (1-p_1)^3 \left( \frac{p_1}{1-p_1} \right)^{f_1(C)} \left( \frac{p_2}{1-p_2} \right)^{f_2(C)} (1-p_2)^{\phi_2(C)}. \] (A.18)

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 \( P_{\Delta} \) cannot be extended to the general case \( \Delta(n; p_1, \ldots, p_{n-1}) \): when \( n > 3 \) the partition function \( Z \) and the corresponding analog of system (A.15) become analytically intractable. This makes Lemmas 1&2 an essential tool for proving Theorem 2 and a more general Theorem 3.

A.4. \( \text{ERSC}(C_3, \{f_1, f_2\}, \{\bar{f}_1, \bar{f}_2\}) \). Here we derive the maximum-entropy distribution on \( C_3 \) only under the constraints of the first type, \( E[f_1] = \bar{f}_1 \) and \( E[f_2] = \bar{f}_2 \), and show that it is different from \( P_{\Delta(3;p_1,p_2)} \). This explicitly demonstrates that the constraint of the second type, \( E[\phi_2] = \bar{\phi}_2 \), is not redundant, and, if dropped, the resulting maximum-entropy ensemble will no longer be \( \Delta \).
Let \((C_3, \bar{P})\) be the maximum-entropy ensemble \(\text{ERSC}(C_3, \{f_1, f_2, \{\tilde{f}_1, \tilde{f}_2\}\})\). In other words, \(\bar{P}\) is the maximum-entropy distribution on \(C_3\) under the constraints
\[
\mathbb{E}[f_1] = \tilde{f}_1 \quad \text{and} \quad \mathbb{E}[f_2] = \tilde{f}_2.
\]
(A.19)

We can find \(\bar{P}\) using Theorem 1 as in the previous section. The partition function
\[
\bar{Z}(\theta_1, \theta_2) = \sum_{C \in C_3} e^{-\tilde{H}(C)} = \sum_{C \in C_3} e^{-\theta_1 f_1(C) - \theta_2 f_2(C)}
\]
\[
= (1 + e^{-\theta_1})^3 + e^{-3\theta_1} e^{-\theta_2},
\]
(A.20)

where the last equality is obtained with the help of Figure 6. The system of equations (A.5) for \(\theta_1\) and \(\theta_2\) is then
\[
3 e^{-\theta_1} \frac{(1 + e^{-\theta_1})^2 + e^{-3\theta_1} e^{-\theta_2}}{(1 + e^{-\theta_1})^3 + e^{-3\theta_1} e^{-\theta_2}} = \tilde{f}_1,
\]
\[
e^{-3\theta_1} e^{-\theta_2} = \tilde{f}_2,
\]
(A.21)

and one can check that the solution is given by
\[
e^{-\theta_1} = \frac{\tilde{f}_1}{3} - \tilde{f}_2 \quad \text{and} \quad e^{-\theta_2} = \frac{\tilde{f}_2(1 - \tilde{f}_2)^2}{\left(\frac{\tilde{f}_1}{3} - \tilde{f}_2\right)^3}.
\]
(A.22)

The partition function, as a function of \(\tilde{f}_1\) and \(\tilde{f}_2\), is then
\[
\bar{Z} = \frac{(1 - \tilde{f}_2)^2}{\left(1 - \frac{\tilde{f}_1}{3}\right)^3}.
\]
(A.23)

Therefore, the maximum-entropy distribution is
\[
\bar{P} = \frac{e^{-\tilde{H}(C)}}{Z} = \frac{e^{-\theta_1 f_1(C) - \theta_2 f_2(C)}}{Z}
\]
\[
= \left(1 - \tilde{f}_2\right)^2 \left(\frac{\tilde{f}_1}{3} - \tilde{f}_2\right) f_1(C) \left(\frac{\tilde{f}_2(1 - \tilde{f}_2)^2}{\left(\frac{\tilde{f}_1}{3} - \tilde{f}_2\right)^3}\right)
\]
\[
= \left(\frac{\tilde{f}_1}{3} - \tilde{f}_2\right)^{f_1(C) - 3 f_2(C)} \left(1 - \frac{\tilde{f}_1}{3}\right)^3 f_2(C) \left(1 - \tilde{f}_2\right)^{2 f_2(C) - 2}.
\]
(A.24)

This is a general expression for \(\bar{P}\) for any expected values \(\tilde{f}_1\) and \(\tilde{f}_2\). In the special case, when \(\tilde{f}_1\) and \(\tilde{f}_2\) coincide with the corresponding values for \(\Delta(3; p_1, p_2)\) in (A.13), that is \(\tilde{f}_1 = 3p_1\) and \(\tilde{f}_2 = p_1^2 p_2\), the distribution \(\bar{P}\) reduces to
\[
\bar{P} = p_1 f_1(C) (1 - p_1)^3 f_2(C) p_2 f_2(C) (1 - p_1^2 p_2) f_1(C) - 3 f_2(C) (1 - p_1^2 p_2)^2 f_2(C) - 2.
\]
(A.25)

We see that \(\bar{P} \neq P_{\Delta(3; p_1, p_2)}\). This means that the two maximum-entropy ensembles \(\Delta(3; p_1, p_2)\) and \((C_3, \bar{P})\) are different,
\[
\text{ERSC}(C_3, \{f_1, f_2, \phi_2\}, \{\tilde{f}_1, \tilde{f}_2, \tilde{\phi}_2\}) \neq \text{ERSC}(C_3, \{f_1, f_2\}, \{\tilde{f}_1, \tilde{f}_2\}),
\]
(A.26)

and, more generally,
\[
\text{ERSC} \left(\{f_1\}_{d=1}^{n-1}, \{\phi_d\}_{d=2}^{n-1}, \{f_1\}_{d=1}^{n-1}, \{\phi_d\}_{d=2}^{n-1}\right) \neq \text{ERSC} \left(\{f_1\}_{d=1}^{n-1}, \{f_1\}_{d=1}^{n-1}, \{f_1\}_{d=1}^{n-1}\right).
\]
(A.27)
References

[1] E. Airoldi, D. M. Blei, S. E. Fienberg, A. Goldenberg, E. P. Xing, and A. X. Zheng (eds.), Statistical network analysis: Models, issues, and new directions, Springer-Verlag, Berlin Heidelberg, 2007.

[2] C. J. Anderson, S. Wasserman, and B. Crouch, A \( p^* \) primer: Logit models for social networks, Social Networks 21 (1999), 37–66.

[3] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, Science 286 (1999), 509–512.

[4] S. Chatterjee and P. Diaconis, Estimating and understanding exponential random graph models, Ann. Statist. 41 (2013), 2428–2461.

[5] T. M. Cover and J. A. Thomas, Elements of information theory, Wiley-Interscience, New York, 1991.

[6] S. N. Dorogovtsev, Lectures on complex networks, Oxford University Press, Oxford, 2010.

[7] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Structure of growing networks with preferential linking, Phys Rev Lett 85 (2000), no. 21, 4633–4636.

[8] D. Easley and J. Kleinberg, Networks, crowds, and markets: Reasoning about a highly connected world, Cambridge University Press, Cambridge, 2010.

[9] H. Edelsbrunner, Modeling with simplicial complexes: topology, geometry, and algorithms, Proc. 6th canad. conf. comput. geom., 1994, pp. 36–44.

[10] H. Edelsbrunner and J. L. Harer, Computational topology: An introduction, Amer. Math. Soc., Providence, Rhode Island, 2010.

[11] P. Erdös and A. Rényi, On random graphs, Publ. Math. 6 (1959), 290–297.

[12] ____, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci. 5 (1960), 17–61.

[13] O. Frank and D. Strauss, Markov graphs, J Am Stat Assoc 81 (1986), no. 395, 832–842.

[14] R. Ghrist and A. Muhammad, Coverage and hole-detection in sensor networks via homology, Proc. 4th int. symp. on information processing in sensor networks, 2005.

[15] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi, A survey of statistical network models, Found. Trends Mach. Learn. 2 (2010), no. 2, 129–233.

[16] A. Hatcher, Algebraic topology, Cambridge University Press, New York, NY, 2002.

[17] P. W. Holland and S. Leinhardt, An exponential family of probability distributions for directed graphs, J. Am. Stat. Assoc. 76 (1981), no. 373, 33–50.

[18] Sz. Horvát, É. Czabarka, and Z. Toroczkai, Reducing degeneracy in maximum entropy models of networks, arXiv:1407.0991 [cond-mat.stat-mech] (2014).

[19] M. Kahle, Topology of random clique complexes, Discrete Mathematics 309 (2009), no. 6, 1658–1671.

[20] ____, Sharp vanishing thresholds for cohomology of random flag complexes, Annals of Mathematics 179 (2014), 1085–1107.

[21] ____, Topology of random simplicial complexes: a survey, AMS Contemp. Math. 620 (2014), 201–222.

[22] K. F. Kee, L. Sparks, D. C. Struppa, and M. Mannucci, Social groups, social media, and higher dimensional social structures: A simplicial model of social aggregation for computational communication research, Communication Quarterly 61 (2013), no. 1, 35–58.

[23] E. Kolaczyk, Statistical analysis of network data, Springer, New York, 2009.

[24] P. L. Krapivsky, S. Redner, and F. Leyvraz, Connectivity of growing random networks, Phys Rev Lett 85 (2000), no. 21, 4629–4632.

[25] N. Linial and R. Meshulam, Homological connectivity of random 2-complexes, Combinatorica 26 (2006), 475–487.

[26] S. Maletić and M. Rajković, Simplicial complex of opinions on scale-free networks, Complex networks, 2009, pp. 127–134.

[27] ____, Consensus formation on a simplicial complex of opinions, Physica A 397 (2014), 111–120.

[28] M. E. J. Newman, Networks: An introduction, Oxford University Press, Oxford, 2010.

[29] M. E. J. Newman and D. J. Watts, Renormalization group analysis of the small-world network model, Physics Letters A 263, 341–346.

[30] J. Park and M. E. J. Newman, Statistical mechanics of networks, Phys Rev E 70 (2004), 066117.

[31] W. Ren, Q. Zhao, R. Ramanathan, J. Gao, A. Swami, A. Bar-Noy, M. Johnson, and P. Basu, Broadcasting in multi-radio multi-channel wireless networks using simplicial complexes, Wireless Networks 19 (2013), no. 6, 1121–1133.

[32] G. Robins, P. Pattison, Y. Kalish, and D. Lusher, An introduction to exponential random graph (\( p^* \)) models for social networks, Social Networks 29 (2007), 173–191.

[33] C. R. Shalizi and A. Rinaldo, Consistency under sampling of exponential random graph models, Ann. Statist. 41 (2013), no. 2, 508–535.

[34] R. Solomonoff and A. Rapoport, Connectivity of random nets, Bull. Math. Biophys. 13 (1951), 107–117.

[35] S. Wasserman and P. E. Pattison, Logit models and logistic regressions for social networks: I. an introduction to markov graphs and \( p^* \), Psychometrika 61 (1996), 401–425.
[36] D. J. Watts, Small worlds: The dynamics of networks between order and randomness, Princeton University Press, Princeton, NJ, 1999.

[37] D. J. Watts and S. H. Strogatz, Collective dynamics of “small-world” networks, Nature 393 (1998), 440–442.

[38] A. Zomorodian, Topological data analysis, Advances in applied and computational topology, 2012.