

Novel Algorithms for Sampling Abstract Simplicial Complexes

John Lombard

June 15, 2018

Abstract We provide dual algorithms for sampling the space of abstract simplicial complexes on a fixed number of vertices that aim to provide a balanced sampling over non-isomorphic complexes. Although sampling uniformly from geometrically distinct complexes is a difficult task with no known analytic algorithm, we develop a generative and descriptive sampler designed with heuristics to help balance the combinatorial multiplicities of the states and more widely sample across the space of inequivalent configurations. We provide a formula for the exact probabilities with which this algorithm will produce a requested labeled state, and compare the algorithm to Kahle’s multi-parameter model of exponential random simplicial complexes, demonstrating analytically that our algorithm performs better with respect to worst-case probability bounds on a given complex and providing numerical results illustrating the increased sampling efficiency over distinct classes. We also design a highly conductive local ergodic random walk with known transition probabilities. We characterize the autocorrelation of the walk, and numerically test it against our sampler to illustrate its efficacy.

Keywords Abstract Simplicial Complexes · Graph Sampling Algorithms · Geometric Probability · Numerical Combinatorics

Mathematics Subject Classification (2000) 05C81 · 05C85 · 68Q87 · 65C99

Acknowledgements This work was supported in part by the University of Washington. We would like to thank Hariharan Narayanan and Stephen Sharpe for their help in revising this manuscript.

1 Introduction to the Space and Use of Abstract Simplicial Complexes

Whether used to model information theoretic phenomena like social networks or to study the combinatorial properties of fundamental structures in understand-

John Lombard
Department of Physics, University of Washington, Seattle, Washington, 98195, USA
E-mail: jml448@uw.edu

ing emergent geometry, abstract simplicial complexes have a rich history of applications and are increasingly used in physics as powerful tools with extensive mathematical structures (Maletic and Rajkovic 2009). Unlike 1-dimensional graphs which only convey connectivity data between nodes, abstract simplicial complexes (ASCs) are generalizations that can allow representations of data through higher-dimensional geometric structures, such as surfaces and volumes in the form of combinatorial triangles and tetrahedra (and their higher dimensional equivalents). Informally, an ASC is the combinatorial abstraction of a geometric simplicial complex encoding the downward closure property. Unlike a geometric simplicial complex where the intersections of any two simplices in the complex must also be a simplex in the complex that is in the union of the boundaries of the intersecting simplices, ASCs only require that any boundary of a simplex is also a simplex in the complex. For example, the clique complex of a graph—the set of all complete subgraphs—is an abstract simplicial complex on the vertices. For a graphical picture of the differences of an ASC with a geometric simplicial complex when embedded into an ambient space, see Fig. 1.

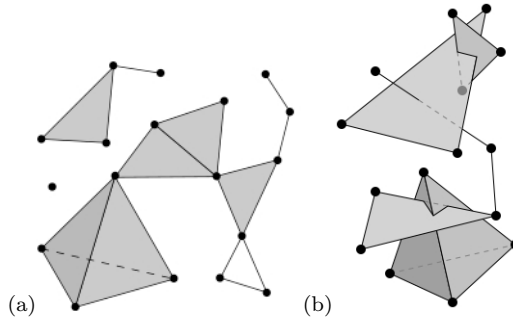


Fig. 1: A Simplicial Complex (a) and a Clique Complex (b), both Embedded in \mathbb{R}^3

This structure allows one to model more complex association data that may not be captured by the limited degrees of freedom in a traditional graph or directed graph. Many models that involve these structures are generative, which is to say that one has a well defined way of prescribing a constructive growth paradigm and studying the complex emergent properties of the resulting states (Wu et al 2015). However, statistical physics models on the space of simplicial complexes and ASCs with certain structures are becoming more popular (Bianconi et al 2015). Although work continues to formally understand the topological properties of this space, finding descriptive algorithms with known probability distributions still requires concentrated effort—especially for models that would be computational feasible (Costa and Farber 2016).

2 Challenges and Solutions in Sampling Abstract Complexes

Our goal is to introduce a new sampling algorithm that is both generative and descriptive on the ASC space C_n with a fixed number of nodes n that can then be used for sampling within algorithms that require random walks on this space, such as the oft used Metropolis Algorithms within Markov Chain Monte-Carlo methods employed throughout computational physics. Due to the combinatorial explosion, the cardinality of this space becomes very large very quickly with increasing n . Richard Dedekind in 1897 was the first to count the number of such configurations, as labeled ASCs are related to monotone boolean functions (Dedekind 1897). Dedekind numbers, which count the number of ASCs with m elements, are only known for $m \leq 8$; however, asymptotic formulas are also known for large m . For the purposes of sampling the unique (nonisomorphic) configurations in the space, we need to remove the labeling that introduces equivalence classes of states under label automorphisms. The inequivalent state cardinalities (and their asymptotic forms) are known only for $m \leq 7$, and grow to be on the order of 5×10^6 by $m = 7$ (Stephen and Yusun 2014). We note that these numbers provide an upper bound on $|C_n|$, as they also include nodal removal. Nevertheless, efficiently sampling such a high dimensional space, especially given the equivalence classes, is a challenge. Since there is not yet a general way to know the cardinalities of the isomorphism classes of simplicial complexes on n nodes, we can do little to tune our algorithm to accommodate this directly. Furthermore, designing either a reversible walk or a sampler with known transition probabilities on such a constrained space is an additional challenge that we face.

In the sections to follow, we introduce two new algorithms for sampling on C_n . We design some basic guiding principles that we show analytically yield a non-local uncorrelated fully ergodic sampler that exhibits extremely strong sampling properties. We numerically illustrate its fast and wide sampling capabilities in comparison to a benchmark model. We also design a local ergodic random walk with known transition probabilities that, at the cost of autocorrelation, samples even more efficiently. Lastly, we characterize the autocorrelation of the walk, and numerically test it against our sampler.

3 Notation and State Visualization

As there are a variety of ways to encode the data of a state $C \in C_n$, we take the opportunity to clarify for the reader the representation we will work with.

Definition 1 (Digraph Representation G)

A state $C \in C_n$ is expressed in a leveled digraph representation $G[C]$ if each node α_d in the digraph at level d represents a $(d - 1)$ -simplex in C , with α as a member of the indexing set on level d , $\alpha \in [1, |\{\alpha_d\}|]$. Defining the set $\{\alpha_1\}$ to be ‘roots’ of the graph with no incoming edges, the directed adjacency structure is constrained such that the following conditions are satisfied:

1. Directed edges exist only between levels $d \rightarrow (d + 1)$
2. The number of parents of node $\alpha_{d>1}$ must be d
3. The number of roots corresponding to the union of the heads of all dipaths leading to α_d must be d

The last condition guarantees simplicial closure, such that for each simplex, its boundary set are also nodes in the graph state with the proper completeness. There can be at most $\binom{n}{d}$ nodes in a level, corresponding to the ASC that is the complete clique complex on n nodes. Similarly, the maximum level is $d = n$.

This graph representation encodes an ASC uniquely up to α labeling. We denote the *geometric* state as one in which the labeling has been removed. For an example of a labeled state with a canonical ordering, we illustrate in Fig. 2 the complete state on 3 roots corresponding to a 2-simplex.

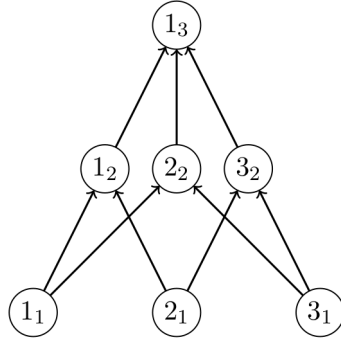


Fig. 2: A Representative Graph State Corresponding to a 2-simplex

The convenience of this representation allows us to repackage the boundary closure constraints into the adjacency structure of this digraph, with the directed nature proving useful for easily identifying branching subgraphs affected by said closure.

Let the complete state on n nodes be denoted C_n^* . Define $F : C_n^* \mapsto C$ as a boolean function that assigns $0 \vee 1$ to each $\alpha_d \in C_n^*$ such that the conditions 1 are satisfied, with $\alpha_d = 1$ indicating existence of node α_d and $\alpha_d = 0$ indicating removal of said node.

It is trivial to see that the space of all such functions $\mathbb{F} \ni F$ covers C_n . F provides an arbitrary labeled ASC in the graph representation $G[C]$. Isomorphic states are related by boolean functions equivalent up to subset permutations preserving the constraints.

On C_3 for example, using a representation of F as a vector with elements mapped to graph nodes in their level-canonical order starting at the roots, the masks $[1111100]$ and $[1111010]$ correspond to the same geometric state and can be shown to be equivalent through the allowed subset permutation on the elements corresponding to level $d = 2$.

4 Kahle's Inductive Construction

Kahle recently introduced a construction for generating random ASCs (Kahle 2014). We describe some of its properties here, using our above notation.

Definition 2 (Kahle’s Model) Kahle’s multi-parameter model $\Delta(n, p_2, \dots, p_n)$ builds an ASC inductively, starting at the edge set with $d = 2$. For every α_d , include the simplex with probability p_d provided it satisfies the boundary conditions in Def. 1.

The full state is built level by level, with constraints on the allowable set of nodes one can include at a given level due to the boundary existence requirements induced by the lower levels.

Let $|\alpha'_d|$ indicate the number of included simplices at level d and $|\alpha_d^*|$ indicate the number of possible simplices given the $(d - 1)$ structure:

$$|\alpha'_d| \leq |\alpha_d^*| \leq \binom{n}{d}. \quad (1)$$

A labeled state $G[C]$ is generated with probability P_Δ given by the following:

$$P_\Delta(C) = \prod_{d=2}^n p_d^{|\alpha'_d|} (1 - p_d)^{|\alpha_d^*| - |\alpha'_d|}. \quad (2)$$

As shown by Zuev *et al.*, Kahle’s model is an Exponential Random Simplicial Complex, implying that it generates a maximum entropy ensemble for an expected number of simplices in the skeletal structures (directly constrained by the probability parameters) (Zuev et al 2015).

We note that the probability of achieving a particular state decreases as a binomial power in the number of total nodes in $G[C]$. Even under a nonuniform probability weighting of the levels, it can be easily seen that the combinatorial multiplicities of nodes in each level create a sampling that is highly peaked around states with a given maximum level for large n —either one that terminates early at the lower levels leaving no higher structures, one that does the opposite, or one that samples toward the ‘half-graph’ state with $\approx \lceil n/2 \rceil$ levels in the case when we take the probabilities to be coin flips. Precise fine tuning would be needed to allow for sampling across a stretch of widely differing geometries, and the power behavior for finding a particular state will still not be mitigated. Additionally, the isomorphism classes of geometric states will be sampled from with additional probability based on their sizes. As the number density of labeled states concentrates toward those that terminate at the central level, we will take the model $\Delta(n, \frac{1}{2}, \dots, \frac{1}{2}) \equiv \Delta_{\frac{1}{2}}$ to benchmark against. Such an algorithm has a probability lower bound at the complete state as follows:

$$\tilde{P}_{\frac{1}{2}} \equiv \min_{C \in C_n} P_{\frac{1}{2}}(C) = P_{\frac{1}{2}}(C_n^*) = \left(\frac{1}{2}\right)^{\sum_{d=2}^n \binom{n}{d}} = \left(\frac{1}{2}\right)^{2^n - n - 1}. \quad (3)$$

We note Kahle’s construction was never claimed to be useful as a fast and broad sampler on C_n . However, from the class of both descriptive and generative algorithms, and as a producer of a maximum entropy ensemble, it is an incredibly simple and natural inductive construction that we feel would serve as a reasonable baseline to compare against our random sampler on this space with the goal of rounding small probability sets in mind.

5 The Balanced Algorithm

Our goal is to sample across geometrically inequivalent states with better mixing than the $\Delta_{\frac{1}{2}}$ model. To this end, we define three key properties that we wish our model to satisfy as heuristics that we intuitively suggest would promote more rapid and broad sampling.

1. Any isolated node such that $|\alpha_d^*| = 1$ should be given a probability of appearance of $p_d = \frac{1}{2}$. At this level in the induction, there are only two possible states that can be selected as the rest of the structure is already fixed. Each state should be given equal probability, as from the vantage of the current step in the algorithm, there is no differentiating property of either state that would induce a bias in the probability. For example, the highest dimension simplex should always have $p_n = \frac{1}{2}$.
2. The power law behavior of binomials in the probabilities should be avoided, which may also aid the associated issue in over-selecting multiple isomorphic states.
3. The completely disconnected state on n nodes, C_n^o , should have the same probability of occurrence as C_n^* . This heuristic aims to re-balance the combinatorial effects of the intimate coupling between nodes at different levels due to simplicial closure, since not including any nodes at $d = 2$ generates C_n^o , while in a construction like Δ , all nodes in $G[C]$ must be independently kept to generate C_n^* , regardless of what probabilities are assigned to each level or even each individual simplex.

To accomplish this, we first note that we will work inversely from Kahle's inductive constructive model and instead consider an equivalent inductive destructive model. Instead of starting from C_n^o , we start from C_n^* and remove nodes starting at $d = 2$ and work upwards in level. This is equivalent to sampling on the space \mathbb{F} , inductively building the boolean mask starting from the all-ones vector. This is computationally easier, as instead of checking the complicated closure conditions at each node we would like to place, we only have to solve for the complete graph state once (which involves finding all complete subgraphs on n nodes, the NP-complete clique problem), save this state to disk, and reference it at will. To retain the simpliciality, upon removing node α_d , one simply removes the unique directed tree associated with α_d as a starting node, which is a linear-time computation. In practice, this amounts to inductively applying a logical AND between the active masking function F and the logical vector NOT[IN TREE] for the removed head node.

Theorem 1 *Let $\mathbf{P}_d = [P_{d_0}, P_{d_1}, \dots, P_{d_{\binom{n}{d}-\hat{d}}}]$ be a probability vector such that $\|\mathbf{P}_d\|_1 = 1$ with P_{d_i} denoting the probability that i nodes are chosen uniformly at random and removed from level d , and \hat{d} indicating the number of nodes already removed from level d due to directed tree pruning from lower level removals.*

$$P_{d_i \neq 0} \equiv P_d = \frac{1}{1 + \sum_{k=d}^n \binom{n}{k} - \hat{k}} \quad (4)$$

$$P_{d_0} = 1 - \left(\binom{n}{d} - \hat{d} \right) P_d$$

satisfies all properties of conditions 1, 2 and 3.

Proof We first note that $0 < P_d \leq \frac{1}{2} \forall d$, as the total node set is positive, finite, and the maximum is achieved in condition 1 as proven below. Additionally, \hat{d} is defined such that $\binom{n}{d} - \hat{d} \in \mathbb{N}$. We need to show that $0 < P_{d_0} < 1$ to conclude that this is a valid probability vector element.

We only seek to show that $(\binom{n}{d} - \hat{d})P_d < 1$, as we already know this quantity is strictly positive due to above arguments. It should be clear that P_d is inversely proportional to the total number of nodes left in the state $G[C]$ at step d in the inductive construction (+1). The combinatorial prefactor is simply the total number of nodes remaining on level d , which must be less than or equal to the total number of nodes in the state. Hence, our claim is justified.

Lastly, we can safely conclude that $\|\mathbf{P}_d\|_1 = 1$ by our construction of $P_{d_0} = 1 - (\binom{n}{d} - \hat{d})P_d$.

To show that this distribution satisfies the condition 1, it can be seen from the definitions that

$$|\alpha_d^*| = 1 \Leftrightarrow \hat{k} = \begin{cases} \binom{n}{d} - 1 & k = d \\ \binom{n}{k} & k > d \end{cases}.$$

Hence,

$$\begin{aligned} P_d|_{|\alpha_d^*|=1} &= \frac{1}{1 + \binom{n}{d} - \binom{n}{d} + 1 + \sum_{k=d+1}^n (\binom{n}{k} - \binom{n}{k})} \\ &= \frac{1}{2}; \\ P_{d_0} &= 1 - P_d = \frac{1}{2}. \end{aligned} \tag{5}$$

Condition 2 is satisfied by algorithmic construction. In choosing groups of i nodes uniformly at random to remove from level d , we trade the power-binomial behavior in the probabilities that grow with the number of total nodes in $G[C]$ for a polynomial-binomial behavior that grows with the number of *levels* instead. Additionally, the Δ model will always pick out a specific labeled $G[C]$ insensitive to the number of isomorphic reachable graphs. In the balanced model, we select from a class of graphs with a certain number of simplicial elements. Although there can also be many such graphs that are not isomorphic but have the same number of elements of given dimensions, we sample the number of elements per level uniformly instead of with product probabilities, giving a key advantage in sets of small probability measure as will be seen exactly in the case of $n = 3$ shown in Section 6.

Satisfying condition 3 requires that the removal of all nodes at the edge level have the same probability as removing no nodes at any level:

$$P_2 = \prod_{d=2}^n P_{d_0}. \tag{6}$$

On the left-hand side,

$$\begin{aligned}
 P_2 &= \frac{1}{1 + \sum_{k=2}^n \binom{n}{k} - \hat{k}} \Big|_{\hat{k}=0} \\
 &= \frac{1}{1 + \sum_{k=2}^n \binom{n}{k}} \\
 &= \frac{1}{2^n - n}.
 \end{aligned} \tag{7}$$

On the right-hand side,

$$\begin{aligned}
 \prod_{d=2}^n P_{d_0} &= \prod_{d=2}^n (1 - (\binom{n}{d} - \hat{d})P_d) \Big|_{\hat{d}=0} \\
 &= \prod_{d=2}^n (1 - \frac{\binom{n}{d}}{1 + \sum_{k=d}^n \binom{n}{k}}) \\
 &= \prod_{d=2}^n \frac{1 + \sum_{k=d}^n \binom{n}{k} - \binom{n}{d}}{1 + \sum_{k=d}^n \binom{n}{k}} \\
 &= \prod_{d=2}^n \frac{1 + \sum_{k=d+1}^n \binom{n}{k}}{1 + \sum_{k=d}^n \binom{n}{k}} \\
 &= \frac{1}{1 + \sum_{k=2}^n \binom{n}{k}} \\
 &= \frac{1}{2^n - n}.
 \end{aligned} \tag{8}$$

Comparing Eq. 7 and 8 demonstrates equality.

We mention that the existence of such a solution to these constraints is very nontrivial. For example, the balancing condition 3 can be shown to have no solution for the $\Delta_{\frac{1}{2}}$ construction for $n > 2$ as equal probability of removal and acceptance would clearly require a solution to an equation of the form

$$\begin{aligned}
 x &= xy \\
 \text{s.t.} \quad &0 < \{x, y\} < 1.
 \end{aligned} \tag{9}$$

Since $n = 2$ doesn't admit more than one probability level (equivalently let $y = 1$), the conditions admit the trivial solution $x = \frac{1}{2}$.

For any constant probability model Δ_x on n roots enforcing the balancing condition 3 and condition 1 requires the probabilities to be roots of polynomials of the form

$$\begin{aligned}
 x^{\frac{n(n-1)}{2}} &= \frac{1}{2}(1-x)^{2^n-n-2} \\
 \text{s.t.} \quad &0 < x < 1.
 \end{aligned} \tag{10}$$

The computer algebra package Mathematica suggests that this equation does not have any rational solutions for x with $n > 2$, indicating that there is likely no natural combinatorial factor that can be attributed to the probability weighting for this model, and relaxing condition 1 does not help.

For any a generic $\Delta(n, p_2, \dots, p_n)$ model, our constraints require parameters that satisfy the following equation:

$$p_2^{\binom{n}{2}} = \frac{1}{2} \prod_{d=2}^{n-1} (1 - p_d)^{\binom{n}{d}} \quad (11)$$

s.t. $0 < p_d < 1$.

In the generic case with independent level probabilities, rational solutions only appear to exist if we remove condition 1; however, this may lead to an large imbalance in the state probabilities for states that are otherwise inductively identical—taking us further from our goal of uniformly sampling the geometric states. It is clear that although possible in theory to balance this algorithm, it requires finding numerical roots at each order and tuning the probabilities to best counteract the power behavior in the sampling, unlike the version we have presented that has closed-form analytic balancing and naturally handles the power structure.

We conclude this section with the probability of finding a given labeled state using this algorithm. As mentioned, this algorithm samples from classes of complexes with certain numbers of objects per skeletal level. In order to relate these probabilities to a specific geometric state, one must know how these classes decompose into nonisomorphic graphs, as well as the relative sizes of the equivalence classes, introducing an additional combinatorial factor.

Let the set of all graph isomorphisms between representations G of a geometric state C be denoted $\text{ISO}(G[C])$ such that the cardinality of this set gives the number of equivalent ways of representing C under Def. 1.

At each inductive step, let i_d nodes be removed from level d out of the total number of available nodes.

The fraction given by the number of labeled ways the selection can be made, weighted by the number of equivalent states at that level, yields the leveled combinatorial factor. Multiplying these factors over the full induction yields the resulting combinatorial factor $\xi(C)$ for achieving a particular geometric state:

$$\xi(C) = |\text{ISO}(G[C])| \prod_{d=2}^n \frac{1}{\binom{\binom{n}{d} - \hat{d}}{i_d}}. \quad (12)$$

However, since $|\text{ISO}(G[C])|$ is not known in advance, we can only compute probabilities analytically for labeled states as this breaks the symmetry factor. Thus, the combinatorial factor becomes

$$\xi_L(C) = \prod_{d=2}^n \frac{1}{\binom{\binom{n}{d} - \hat{d}}{i_d}}. \quad (13)$$

It is this quantity that we will use in our comparisons to the Δ model, as they both consider specific labeled states. In practice, the geometric probabilities are larger, with the labeled probabilities providing a lower bound.

Let $\{j\}$ be a boolean sequence representing whether any nodes were masked from C_n^* , with $j_d \equiv \{j\}_d = 0$ as an indicator that no nodes were removed from level d . In terms of our boolean function F , the elements correspond to a $\text{NOT}[\text{ALL}[F_d]]$ operation over the level subsets $F_d \subset F$. The probability of finding a labeled state is given by the following expression:

$$P(C) = \xi_L(C) \prod_{d=2}^n (P^d)^{\delta_{j_d}^1} (P_0^d)^{\delta_{j_d}^0}, \quad (14)$$

where δ_b^a is the Kronecker delta.

6 Properties of the Balanced Algorithm and Simulation Results

This algorithm samples across a weighted space of paths for inductively building a given state, as opposed to building a specific state itself. In the case where each such path yields a unique state up to relabeling, this algorithm will produce the uniform distribution on the space of complexes. Such a condition is only true for $n = \{2, 3\}$ where $\xi(C) = 1 \forall C \in C_{\{2,3\}}$, and is illustrated in a direct comparison with the $\Delta_{\frac{1}{2}}$ benchmark in Fig. 3. This graph bins the multiplicities for which each geometric state was sampled, subtracted by the mean multiplicity to give residuals, and normalized by the total number of samples. The bins themselves do not match to the same geometric state between the two algorithms, but map to the first encountered representative of a given state. One can clearly see the uniform sampling from the balanced algorithm, although given the number of total samples, both algorithms find all 5 geometric states. All simulations were performed using MATLAB.

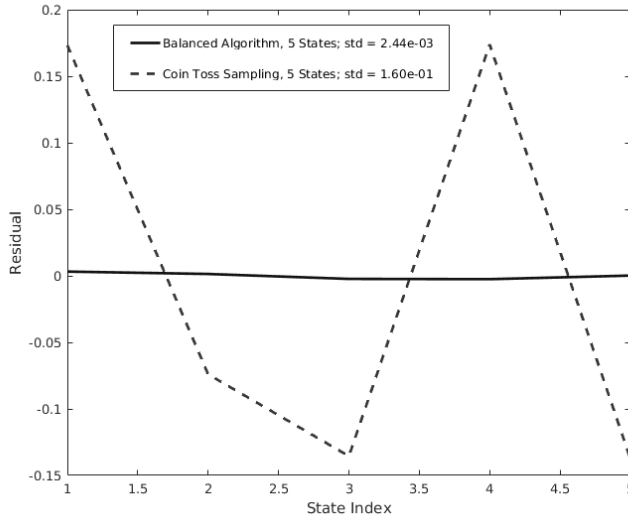


Fig. 3: Multiplicities Residuals of Unique Geometric States on 10000 Samples Drawn From C_3 , Linearly Interpolated

However, for $n = 4$ and higher, there exist nonisomorphic graphs with the same number of simplicial elements in each skeleton. This introduces a nonuniform

combinatorial factor that is not possible to account for at the time of writing due to the fact that there is no analytic algorithm for predicting the number of such inequivalent graphs and their combinatorial multiplicities. Of course, since we can explicitly compute the probabilities for generating a labeled state, we mention that this sampler can be equipped with a Metropolis filter to re-weight the probabilities to produce a uniform sampling on labeled states.

We now examine the raw probabilities for sampling a unique labeled state. Directly comparing the minimal probability in the $\Delta_{\frac{1}{2}}$ model with the equivalent complete state in the balanced model indicates that this state has a much greater probability of occurrence:

$$\frac{1}{2^n - n} > \left(\frac{1}{2}\right)^{2^n - n - 1} \forall n > 2. \quad (15)$$

To indicate whether the new algorithm has balanced the probabilities at large and removed sets of extremely suppressed measure would require looking at the minimal probability bound for this algorithm and comparing it to C_n^* as generated from $\Delta_{\frac{1}{2}}$. Here, we must use the labeled combinatorial factor ξ_L for adequate comparison. Due to the balancing, the probabilities are minimized toward the half-graph state, as this maximizes the binomial coefficients at each level with many combinatorial possibilities equivalent to the removal of certain numbers of nodes. As we would like a lower bound, we set $\hat{k} = 0 \forall k$. Even though we are removing approximately half of the nodes at each level, to maximize the binomial contribution, maintaining the full combinatorial degree of each level will further decrease the probabilities.

In total, this gives an estimate for a lower bound of the following form:

With

$$E(x) = \begin{cases} \frac{x}{2} & \text{mod } (x, 2) = 0 \\ \frac{x+1}{2} & \text{mod } (x, 2) = 1 \end{cases},$$

$$\tilde{P} = \min_{C \in C_n} P(C) \approx \prod_{d=2}^{E(n)} \frac{1}{\binom{n}{d} \frac{1}{E(\binom{n}{d}) + \sum_{k=d}^n \binom{n}{k}}} . \quad (16)$$

Numerical analysis confirms that $\tilde{P}_{\frac{1}{2}} \leq \tilde{P}$ for reasonable values of n before they become numerically unstable due to the combinatorial explosion, as illustrated in Fig. 4. It is immediately apparent that this algorithm has a much stronger probability behavior and actively works against the suppression found in a product model.

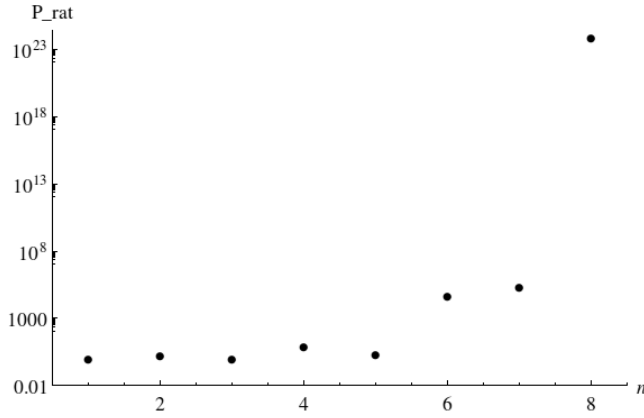


Fig. 4: A Log-Plot of the Ratio $\frac{\tilde{P}}{\tilde{P}_{\frac{1}{2}}}$ as a Function of the Number of Roots n

Lastly, we advertised that the combinatorial balancing would allow for a broader access of states. Below we provide some simulation results to illustrate this property. Fig. 5 shows the number of unique geometric states encountered while sampling C_6 for a variety of sampling lengths. We can see that the balanced algorithm samples states at a faster rate than the $\Delta_{\frac{1}{2}}$ benchmark test. This is again demonstrated in Fig. 6, where 50000 samples were drawn on C_5 . The balanced algorithm has appeared to converge, while the $\Delta_{\frac{1}{2}}$ benchmark has yet to find all of the inequivalent states. Naturally, the states with higher probability of being encountered were among the first to be sampled, explaining the correlation between the large initial fluctuations in the two algorithms given the first-representative binning process. However, the multiplicity fluctuations are much smaller for the balanced algorithm, indicating that the goal of heuristically rounding the space of state probabilities has been preliminarily accomplished by this algorithm.

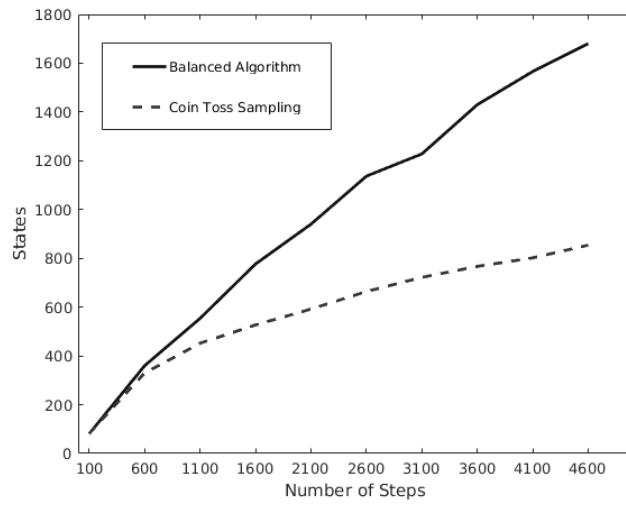


Fig. 5: Number of Unique Geometric States While Sampling C_6 as a Function of the Sample Size, Linearly Interpolated

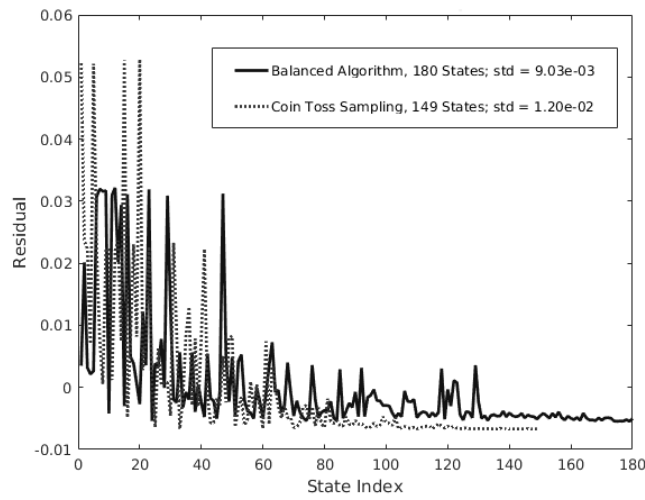


Fig. 6: Multiplicity Residuals of Unique Geometric States on 50000 Samples on C_5 , Linearly Interpolated

7 Local Random Walks

The algorithm introduced in Sec. 5 can be naturally used to perform an ergodic walk on the ASC space. We can jump from any state to another without a barrier as there is no dependence on the current state to restrict the space of next available states. This is a desirable feature from the perspective of sampling on the full space, as there are no regions of low conductance in the state space where our ‘walk’ can become trapped. However, when a Metropolis filter is utilized, the fact that this sampler can introduce transitions between arbitrary configurations may be a detriment to the acceptance rate if the filter is not naturally tuned to the intrinsic sampling probabilities of the walk. A local random walk between nearby configurations would be more likely to permit an acceptance with respect to a Metropolis filter, and as a result, may sample the full space faster. We would still like to be able to make this local random walk reversible, however, to ensure the Markov property. However, on the ASC space, the closure constraints make constructing a reversible local random walk very difficult. It is even still difficult to find *any* local random walk where one can compute forward and backward transition probabilities in order to force the walk to be reversible with respect to an additional Metropolis filter. In the following sections, we illustrate one example of a local random walk in the ASC space, and compare its properties to our global sampler.

8 A Local Random Walk on Abstract Simplicial Complexes

Our goal is to create a local random walk that has more favorable acceptance ratios for a Metropolis filter that is in some way sensitive to the topological structure of the current ASC. To that end, we define ‘local’ with respect to a new metric on the ASC space that is restricted to a unidirectional walk away from a given state.

Definition 3 (Unconstrained Nodes) Given an ASC in the graph representation $G[C]$, we define an ‘unconstrained node’ u as one that can be freely added or removed without requiring or destroying additional containment structure.

An unconstrained node is ‘removable’ if it has no children and is itself not a root (as the we hold the roots fixed in C_n).

An unconstrained node is ‘addable’ if it is a member of $C_n^* \setminus C$ (the complete graph with the current state excluded) that has all of its parents in C .

We work with unconstrained nodes for two reasons. Foremost, we would like to have a walk that admits a range of local movement as opposed to simply a one step nearest-neighbor walk on individual simplices. If we admit moves that can add or remove an arbitrary number of nodes within the state space, one needs to worry about the closure constraints. These constraints will make it very difficult to generate a walk that has computable probabilities for reversibility, as the number of admissible additions or removals would be dynamic with each sub-step within the same transition move, and there can be multiple paths with different probabilities that could lead to the same state. We want to restrict down this capability, but still admit larger jumps through the state space. Hence, we work with the space of unconstrained nodes as pure additions or removals within this space will prevent such issues from arising and admit a walk with computable probabilities. The

restriction that nodes are only added or removed in a single step additionally guarantees that we do not have any closed loops within our multi-step walk for a given transition.

Our notion of local distance is therefore the number of added or removed nodes in a given transition step, actioned by a binary flip on the boolean function representing C .

The algorithm mimics an exponential ball walk with respect to this distance measure. First, we compute the total number of nodes one could maximally flip on the state space. From this set, we establish a normalized probability function based on an exponential decrease in probability for larger numbers of binary flips. We decide to either add or remove nodes in a given transition. Once this choice is made, we compute all addable/removable nodes U for the current configuration. We select a distance δ to move based on the fixed probability measure. If that distance takes the state outside of the state space or beyond the number of admissible adds/removes, then the algorithm resets until an admissible move is found—this is our rejection sampling step similar to a ball walk on the edge of the state space. Further logic handles the complete state and the state with only roots to speed up the process in those cases. Once a good distance is accepted, a uniformly random selection of those unconstrained nodes $u \in U$ have their entries flipped in the boolean representation. The forward and backward probabilities are symmetric with respect to the exponential distance weighting, as this is not dependent on the state itself. Therefore, a Metropolis filter would only need to account for the uniform selection step, producing a combinatorial factor of

$$\frac{P_B}{P_F} = \frac{\binom{|U_B|}{\delta}}{\binom{|U_F|}{\delta}}. \quad (17)$$

9 Computational Results

Since our sampler now has local correlations, it becomes necessary to characterize more carefully the efficiency of the random walk and breadth of sampling. We present two extreme situations for the initial start of the walk: beginning at a corner of the state space, C_n^* , and beginning at a ‘central’ state consisting of roughly half of the available simplices being activated. We examine both the multiplicity residuals as before, as well as the autocorrelation length.

To characterize the autocorrelation length, we use an initial convex sequence method that involves the greatest common minorant Geyer (2011). First, we implement a Metropolis filter utilizing Eq. 17 such that our samples can be expected to be i.i.d. To measure autocorrelation, we compute the signed displacement of a transition between two states C and C' as the difference in the sums of their boolean representations:

$$\delta = \|C\|_1 - \|C'\|_1, \quad (18)$$

where $|\delta|$ still corresponds to the number of binary flips between the two, as discussed in the algorithm. We then look at the cumulative sum of the time sequence of δ_i values for each step i in the walk. This gives some sense of a 1-dimensional

projection of the random walk through the ASC space, making it a natural random variable to compute autocorrelations with.

Let an estimator for the sample mean on s samples be denoted

$$\hat{\mu}_s = \frac{1}{s} \sum_{i=1}^n \delta_i. \quad (19)$$

A natural estimator for the auto-covariance function at lag k is given by the following:

$$\hat{\gamma}_s = \frac{1}{s} \sum_{i=1}^{n-k} (\delta_i - \hat{\mu}_s)(\delta_{i+k} - \hat{\mu}_s). \quad (20)$$

The greatest convex minorant at lag k ,

$$\Gamma_k = \gamma_{2k} + \gamma_{2k+1}, \quad (21)$$

is a strictly positive, decreasing, and convex function for a reversible Markov chain. Therefore, examining our estimator $\hat{\Gamma}_k$ for the point at which it becomes nonpositive indicates the lag where we encounter autocorrelation. Due to the dependence on twice the lag, our autocorrelation is related to $2k'$ when $\hat{\Gamma}_{k'} \leq 0$.

We can see in Fig. 7 that the local random walk started from a central state, upon re-weighting with the Metropolis filter, produces autocorrelation out to about 16 steps, and the walk has a natural rejection rate on the order of 50%.

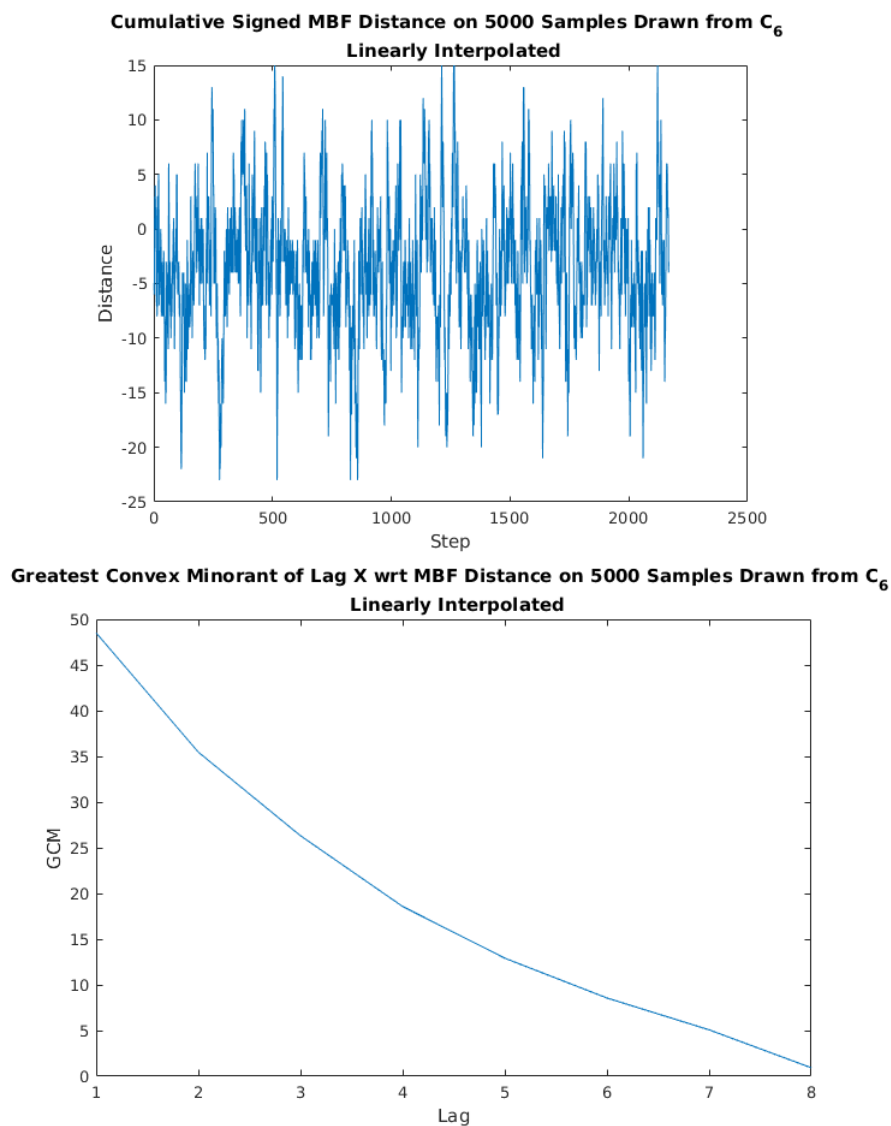


Fig. 7: Autocorrelation Statistics for a Random Walk on C_6 with 5000 Steps, Starting from a Central State

Fig. 8 which was produced starting from a corner state tells not much of a different story, which indicates that the edges in the state space are not incredibly narrow, and that this random walk is good at working its way out of those corners. We see less than double the autocorrelation, which is not unexpected due to the time spent in the region of small state density. A burn-in process would reduce this down to toward the autocorrelation lengths found in the central case.

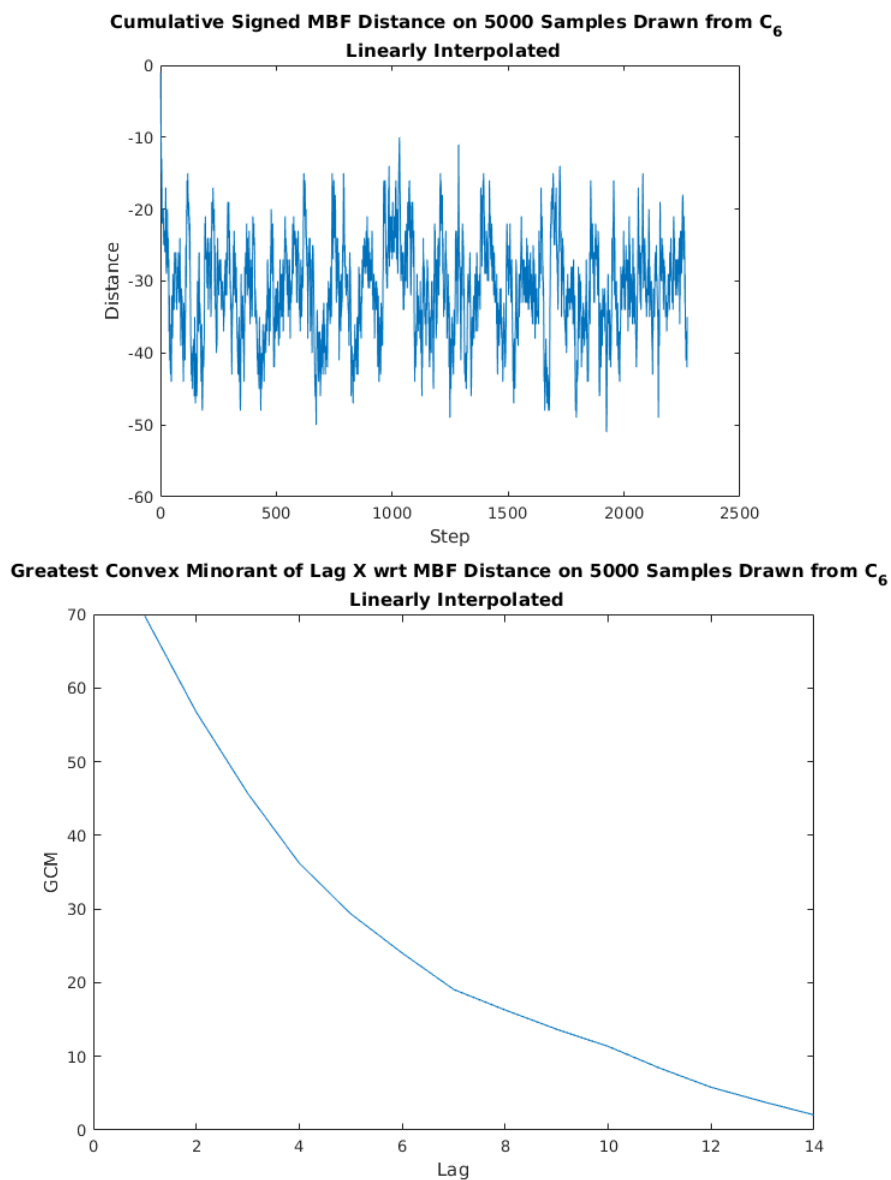


Fig. 8: Autocorrelation Statistics for a Random Walk on C_6 with 5000 Steps, Starting from a Corner State

We lastly compare the efficiency of all of these algorithms for sampling geometrically unique states. As seen in Figs. 9 and 10, the local random walk performs remarkably better, sampling more states with less accepted steps. This lends credence to the notion that the best sampler on this space would likely be a linear combination of the two Markov chains. Since such a construction still retains its

theoretical properties, we can achieve the best of both algorithms by choosing to perform a local walk with some large probability to reap the rewards of the rapid sampling, while occasionally using the balanced sampler to avoid regions of narrow conductance bands and to promote ergodicity and large nonlocal transitions.

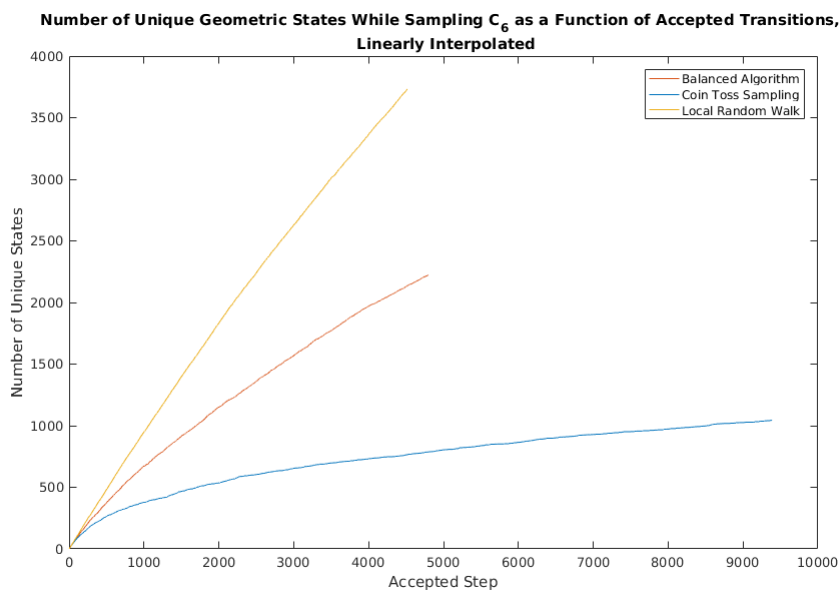


Fig. 9: A Comparison of the Unique Geometric States Sampled on C_6 as a Function of Accepted Transitions for All Three Samplers

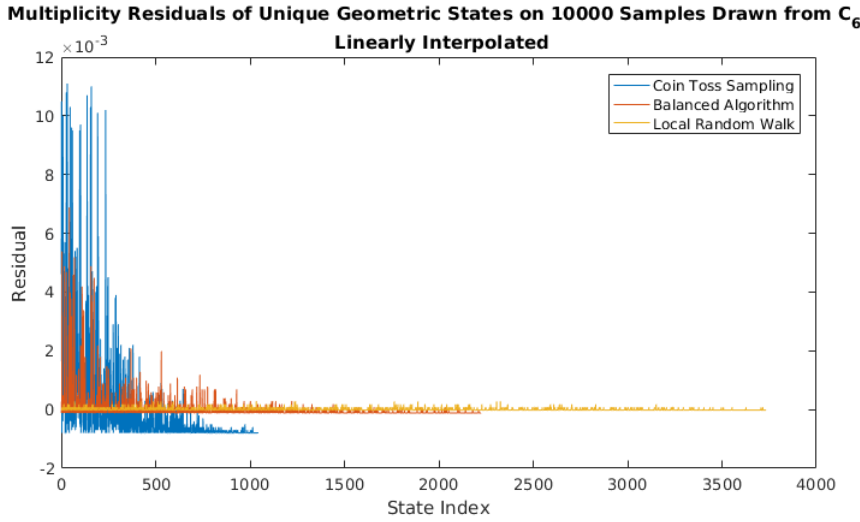


Fig. 10: A Comparison of the Multiplicity Residuals of Unique Geometric States Sampled on C_6 for All Three Samplers, Linearly Interpolated

10 Discussion on the New Samplers

As is the case with a wide variety of combinatorial spaces, it is often very difficult to develop a sampling procedure with transition probabilities that can *a priori* sample such that the uniform distribution is the stationary distribution without the use of a Metropolis filter. In the case of abstract simplicial complexes, the unknown isomorphism classes of configurations make this problem seemingly intractable. We have introduced an algorithm that uses three simple principles to attempt to re-balance the sampling such that the algorithm more readily samples inequivalent configurations with a wide breadth across the space. Our analytical results show that this algorithm has a worst case lower-bound on state probabilities that is larger than the equivalent sampling through a uniformly weighted Kahle process, which we used as an unoptimized benchmark. Our simulations confirm that a direct comparison between the two algorithms favors the balanced algorithm when attempting to sample across the geometric space of states.

We have also discussed a local random walk that can be made reversible. The advantage of this walk is to increase the acceptance rates for a Metropolis filter when sampling nearby states as opposed to large jumps in the state space, and we have illustrated through simulation its efficiency in also sampling from a wide range of states in the ASC space. However, in some applications with Metropolis filters, this walk may be sensitive to trapping regions, as it is not able to explore any possible configuration in a single transition step. Thus, a combination of our local walk and the balanced sampler can be used to promote ergodicity and rapid sampling.

Future work toward finding a better generative algorithm for sampling across equivalence classes of large random abstract simplicial complexes while maintaining analytical control is necessary in order to begin to probe the very large space

of states. With a variety of applications on the horizon, we anticipate this problem being approached from a broad range of perspectives, and we hope to have provided some insight through some practical, simple algorithms that accomplish the first steps toward this task.

References

- Bianconi G, Rahmede C, Wu Z (2015) Complex quantum network geometries: Evolution and phase transitions. *Phys Rev E* 92:022,815, DOI 10.1103/PhysRevE.92.022815, URL <http://link.aps.org/doi/10.1103/PhysRevE.92.022815>
- Costa A, Farber M (2016) Random simplicial complexes. In: Callegaro F, Cohen F, De Concini C, Feichtner EM, Gaiffi G, Salvetti M (eds) *Configuration Spaces: Geometry, Topology and Representation Theory*, Springer International Publishing, Cham, pp 129–153, DOI 10.1007/978-3-319-31580-5_6, URL http://dx.doi.org/10.1007/978-3-319-31580-5_6
- Dedekind R (1897) *Über zerlegungen von zahlen durch ihre grossten gemeinsamen teiler*. Festschrift der Technischen Hochschule zu Braunschweig bei Gelegenheit der 69 Versammlung Deutscher Naturforscher und Ärzte pp 1–40
- Geyer C (2011) Introduction to mcmc. In: Brooks S, Gelman A, Jones G, Meng XL (eds) *Handbook of Markov Chain Monte Carlo*, Chapman and Hall, chap Introduction to MCMC
- Kahle M (2014) Topology of random simplicial complexes: A survey. *AMS Contemp Math* 620:201–22
- Maletic S, Rajkovic M (2009) *Complex Networks (Studies in Computational Intelligence Series)*, vol 207, Springer-Verlag, Berlin, chap Simplicial Complex of Opinions on Scale-Free Networks, pp 127–34
- Stephen T, Yusun T (2014) Counting inequivalent monotone boolean functions. *Discrete Applied Mathematics* 167:15 – 24, DOI <http://dx.doi.org/10.1016/j.dam.2013.11.015>, URL <http://www.sciencedirect.com/science/article/pii/S0166218X13005611>
- Wu Z, Menichetti G, Rahmede C, Bianconi G (2015) Emergent complex network geometry. *Nature Scientific Reports* 5(10073):10,073, URL <http://arxiv.org/abs/1412.3405>
- Zuev K, Eisenberg O, Krioukov D (2015) Exponential random simplicial complexes. *Journal of Physics A: Mathematical and Theoretical* 48(46):465,002, URL <http://stacks.iop.org/1751-8121/48/i=46/a=465002>