Paintboxes and probability functions for edge-exchangeable graphs

Diana Cai
Department of Statistics
University of Chicago
Chicago, IL 60637
dcai@uchicago.edu

Trevor Campbell
CSAIL, MIT
Cambridge, MA 02139
tdjc@mit.edu

Tamara Broderick
CSAIL, MIT
Cambridge, MA 02139
tbroderick@csail.mit.edu

1 Introduction

Network datasets arise naturally in a number of applications, such as online social networks, biological networks, communication networks, and transportation networks [1]. As a result, there is growing interest in developing models for such data, and in particular—as individual networks continue to grow in size—statistical models that accurately reflect the real-life scaling properties of such graph sequences. One important real-life scaling property we want to capture in a network model is sparsity (e.g., [2–8]), where the number of edges grows sub-quadratically in the number of vertices [9–11].

Many popular network models make the assumption of exchangeability in the vertices (see Lloyd et al. [12, Sec. 7.1] for examples); i.e., the order that the vertices appear in the graph sequence does not affect the distribution of the graph. The Aldous-Hoover theorem [13, 14] characterizes the distribution of such vertex-exchangeable graphs, by guaranteeing the existence of a function, often called a *graphon*, from which we can sample vertex-exchangeable graphs. This representation theorem allows us to evaluate all vertex-exchangeable graphs at once; in particular, a corollary of the Aldous-Hoover theorem is that any vertex-exchangeable graph is almost surely dense or empty (see Orbanz and Roy [15]). This is an undesirable scaling property for modeling real-world graphs, as it implies that the number of edges grows quadratically in the number of vertices.

Recently, the notion of edge exchangeability for graphs was introduced; here the order that *edges* appear in the graph sequence—as opposed to the vertices—does not affect the distribution of the graph, and it was proven that there exist sparse, edge-exchangeable graphs [16, 17]. Thus, edge exchangeability provides a promising alternative to vertex exchangeability for developing hierarchical models for network and relational data, as, crucially, it admits properties of sparsity and power law behavior observed in many real-world networks. As in the vertex-exchangeable case with the Aldous-Hoover theorem, it is desirable to to characterize all edge-exchangeable graphs at once via a de-Finetti style representation.

In this paper, we provide a paintbox representation for edge-exchangeable graphs, which we call *the graph paintbox*, and show that it characterizes the distribution of edge-exchangeable graphs. The paintbox representation thus provides a way to analyze all edge-exchangeable graphs, and in particular, study desirable properties such as sparsity and power laws. In addition, we discuss a particular class of edge-exchangeable graphs examined by Cai, Campbell, and Broderick [16], called *graph frequency models*, and provide a characterization for an important class of graph frequency models via a function called the *exchangeable vertex probability function* (EVPF), which is reminiscent of the exchangeable partition probability function (EPPF) for random partitions [18]. In particular, the EVPF is convenient for practical inference in edge-exchangeable graph models, analogous to the role of the EPPF for performing practical inference Bayesian nonparametric models for clustering, such as with the Dirichlet process mixture [19] and more generally in other clustering models [20, 21].

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

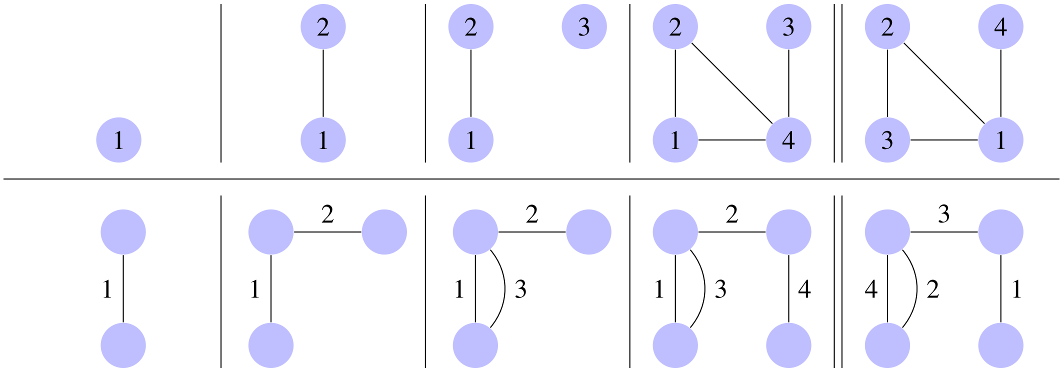

Figure 1: Vertex vs edge exchangeability. **Top row**: A vertex-exchangeable sequence, where a new vertex joins at each step, instantiating any edges with existing vertices. The two rightmost graphs have the same probability under vertex exchangeability. **Bottom row**: An edge-exchangeable sequence, where an edge joins at each step, instantiating vertices along the way. The two rightmost graphs have the same probability under edge exchangeability.

In what follows, we provide a high-level overview of our results. The detailed derivations and proofs for a generalized treatment (including results pertaining to hypergraphs, where edges connect to a finite number of vertices with finite multiplicity) can be found in Campbell, Cai, and Broderick [22].

2 Edge exchangeability

In order to discuss edge exchangeability in graphs, we need to first introduce the notion of a graph sequence. Let $(G_n)_{n\in\mathbb{N}} := G_1, G_2, \dots$ be a sequence of graphs where each graph $G_n = (V_n, E_n)$ consists of a (finite) set of vertices V_n and a (finite) multiset of edges E_n. Each edge connects two distinct vertices in V_n. We assume the sequence is both random and *projective*—or growing—so that $V_n \subseteq V_{n+1}$ and $E_n \subseteq E_{n+1}$ almost surely.

Figure 1 shows two different ways of representing such a graph sequence. The traditional approach is to label the vertices, and express each graph in the sequence as the collection of edges represented by pairs of vertices. For example, in the top row of Figure 1, the graph sequence is constructed by sequentially adding vertex 1 with no edge, then adding vertex 2 with the edge $\{1, 2\}$, then vertex 3 with no edge, and finally vertex 4 with edges $\{1, 4\}$, $\{2, 4\}$, and $\{3, 4\}$. We can represent the final graph as its collection of edges: $G_4 = \{\{1, 2\}, \{1, 4\}, \{2, 4\}, \{3, 4\}\}$. This representation allows us to specify *vertex exchangeability* as the invariance of the distribution of G_4 to permutations of the vertex labels. Hence, as shown in the upper right two graphs in Figure 1, the graphs G_4 and $\tilde{G}_4 = \{\{3, 2\}, \{1, 2\}, \{1, 3\}, \{1, 4\}\}$ have the same probability under a vertex-exchangeable distribution, as the vertex labels were reordered by the permutation $(134)(2)$ (in cycle notation).

An alternative approach, to which we adhere, is to label the *edges*, and express each graph in the sequence as the collection of vertices represented by the edge labels to which they connect. For example, in the bottom row of Figure 1, the graph sequence is constructed by the following sequential procedure. First, edge 1 is introduced between two new vertices, so $G_1 = \{\{1\}, \{1\}\}$. Next, edge 2 connects to one old and one new vertex, so $G_2 = \{\{1, 2\}, \{2\}, \{1\}\}$. Then edge 3 connects to the same vertices as edge 1, so $G_3 = \{\{1, 3, 2\}, \{2\}, \{1, 3\}\}$. Finally, edge 4 introduces one new vertex and connects to the vertex only connected to edge 2, so $G_4 = \{\{1, 3, 2\}, \{2, 4\}, \{1, 3\}, \{4\}\}$. This representation—which we call a *vertex allocation*, as it allocates labeled edges amongst the vertices—allows us to specify *edge exchangeability* as the invariance of the distribution of G_4 to permutations of the edge labels. Hence, as shown in the lower right two graphs in Figure 1, the graphs G_4 and $\tilde{G}_4 = \{\{2, 3, 4\}, \{3, 1\}, \{2, 4\}, \{1\}\}$ have the same probability under an edge-exchangeable distribution, as the edge labels were reordered by the permutation $(14)(23)$ (in cycle notation).

Definition 2.1. Consider the random graph sequence $(G_n)_n$ represented by vertex allocations. The sequence $(G_n)_n$ is infinitely *edge-exchangeable* if for every $n \in \mathbb{N}$ and for every permutation π of $\{1, \dots, n\}$, $G_n \overset{\mathrm{d}}{=} \tilde{G}_n$, where $\tilde{G}_n$ is the same vertex allocation as G_n with indices permuted by π.

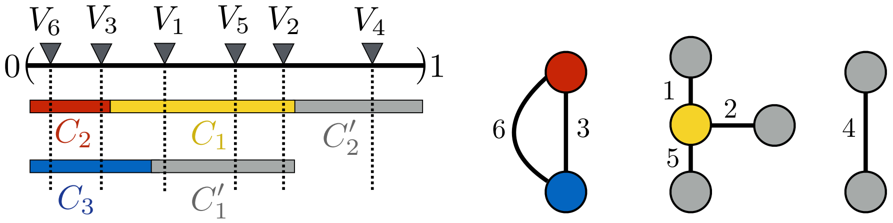

Figure 2: **Left**: An example graph paintbox and uniform random draws from it. **Right**: The resulting edge-exchangeable graph. Vertices are colored according to the paintbox subsets, and edges are labeled according to the indices of V_n. Gray vertices are dust vertices that only occur in a single edge.

3 The graph paintbox

We now present the paintbox representation that characterizes the distribution of edge-exchangeable graphs, referred to as the *graph paintbox*. In particular, *any* edge-exchangeable graph has essentially the same representation of its distribution. We first generate a random collection of open subsets of $(0, 1)$, $(C_k)_k$, C_1', and C_2', and independently sample an i.i.d. sequence $(V_n)_n$ of uniform random variables on $(0, 1)$. Each V_n corresponds to an edge in the graph sequence, which connects to exactly two vertices determined by the subsets in which it falls. The subsets $(C_k)_k$ represent *regular vertices*, and C_1', C_2' represent *dust vertices*. Regular vertices can connect to any number of edges, whereas dust vertices can connect only to a single edge in the entire infinite graph sequence. If $V_n \in C_k$ and $V_n \in C_j$ for $j \neq k$, then edge n is between two regular vertices. If $V_n \in C_k$ and $V_n \in C_1'$, then edge n is between a regular vertex and a dust vertex that only ever connects to edge n. Finally, if $V_n \in C_2'$, edge n connects to 2 dust vertices that both only ever connect to edge n. We call an edge-exchangeable graph *regular* if there are no dust vertices, or equivalently, if $C_1' = C_2' = \emptyset$.

Figure 2 illustrates an example graph paintbox, where the subsets C_1, C_2, C_3 are given by the yellow, red, and blue subsets of $(0, 1)$, respectively, and the gray subsets C_1', C_2' represent dust vertices. On the right is the resulting edge-exchangeable graph after making 6 i.i.d. uniform draws of V_n given the paintbox. The resulting vertex allocation is $G_6 = \{\{3, 6\}, \{1, 2, 5\}, \{3, 6\}, \{1\}, \{2\}, \{5\}, \{4\}, \{4\}\}$. Here the gray vertices are the dust vertices, which only ever connect to a single edge in the infinite graph sequence, and the dust vertices allow for structure such as isolated edges and stars in the graph. Note that Definition 3.1 could be modified to allow for both loops (edges that connect to only one vertex) and empty edges (those that connect to no vertices), but for clarity we only present the case where edges must connect to exactly two vertices.

Definition 3.1. An infinitely edge-exchangeable graph sequence $(G_n)_n$ has a *graph paintbox* if there exists a random sequence $(C_k)_k$ and random sets C_1', C_2' that are all Lebesgue-measurable open subsets of $(0, 1)$ such that

- the sets C_2' and $\bigcup_k C_k \cup C_1'$ are disjoint, and
- any $V \in (0, 1)$ is an element of either C_2' or exactly two sets from $(C_k)_k$ and C_1',

where the sequence $(G_n)_n$ is constructed by

- sampling $(V_n)_n \overset{i.i.d.}{\sim} \mathrm{Unif}(0, 1)$,
- setting $I_k = \{n : V_n \in C_k\}$ for each $k \in \mathbb{N}$,
- setting $I_1' = \{\{n\} : V_n \in C_1'\}$ and $I_2' = \{\{n\}, \{n\} : V_n \in C_2'\}$, and
- setting G_n for each $n \in \mathbb{N}$ to the collection $(I_k)_k, I_1', I_2'$ with indices above n removed.

The graph paintbox described in Definition 3.1 can also be viewed as a special case of the feature paintbox [23], in which the number of subsets a point $V_n \in (0, 1)$ can be an element of is constrained to be exactly 2; in this case, features correspond to vertices. Our treatment of the paintbox additionally includes the characterization of dust vertices.

Theorem 3.2. *A random graph $(G_n)_n$ is infinitely edge-exchangeable iff it has a graph paintbox.*

Note that the graph paintbox representation can be extended to more general *hypergraphs*, where edges connect a finite number of vertices (potentially with multiplicity) via recent work on the theory of trait allocations [22].

4 Frequency models and probability functions

In this section, we consider a particular construction of an edge-exchangeable graph, which is an important special case of the general class of graph frequency models considered by Cai, Campbell, and Broderick [16] (we refer to this special case with the same name for brevity). Graph frequency models are promising not only because they can produce sparse graphs [16], but also because they are particularly amenable to approximate posterior inference algorithms, such as variational inference, Hamiltonian Monte Carlo, and Gibbs sampling. We show the equivalence of random graph sequences having a frequency model with those having a marginal distribution that can be expressed as an *exchangeable vertex probability function (EVPF)*, the graph analog of the exchangeable partition probability function (EPPF) [18] and the exchangeable feature probability function (EFPF) [23]. Similar to how the EPPF has led to practical inference algorithms for Bayesian clustering, and how the EFPF has done the same for Bayesian feature learning, the equivalent characterization of the EVPF with graph frequency models provided in Theorem 4.2 implies a class of edge-exchangeable graph models that have practical posterior inference algorithms.

The particular graph frequency model we consider is as follows. Suppose we have a countably infinite collection of weights $(w_i)_{i \in \mathbb{N}}$, with $\sum_j w_j < \infty$ and $w_j \in (0, 1)$, where each index $i \in \mathbb{N}$ is a "popularity" associated with a potential vertex in the graph. To sample a sequence of graphs given these weights $(w_i)_i$, at each step n, edge n forms between vertex j and k, $j \neq k$, with probability proportional to $w_j w_k$.

Rather than specifying the distribution of a graph sequence conditioned on latent parameters (such as the weights $(w_i)_i$), we can specify its marginal distribution. Suppose at step n, we let $\bar{G}_n$ be the multiset of degrees of the vertices in G_n, i.e., a multiset containing the number of edges each vertex participates in. Using again the bottom row of Figure 1 as an example, since $G_4 = \{\{1, 3, 2\}, \{2, 4\}, \{1, 3\}, \{4\}\}$, we have that $\bar{G}_n = \{3, 2, 2, 1\}$. We also let $\kappa(G_n)$ be the number of unique orderings of the sets in G_n – in the previous example, $\kappa(G_4) = 4! = 24$ since there are 4 unique sets in the vertex allocation G_4. As another example, if $G_n = \{\{1, 2\}, \{1, 2\}, \{3\}, \{3\}\}$, there are $\kappa(G_n) = 4!/(2! \cdot 2!) = 6$ unique orderings. If the distribution of G_n depends only on n, $\bar{G}_n$, and $\kappa(G_n)$, as specified in Definition 4.1, we say G_n has an *exchangeable vertex probability function (EVPF)*.

Definition 4.1. A random graph G_n has a exchangeable vertex probability function if it admits the representation $\Pr(G_n) = \kappa(G_n) p(n, \bar{G}_n)$.

Intuitively, the EVPF says that the probability of the graph only depends on the degrees of the vertices. For an arbitrary edge-exchangeable graph, there may not exist an EVPF; however, any edge-exchangeable graph that has a graph frequency model also has an EVPF, and vice versa, as shown by Theorem 4.2.

Theorem 4.2. *A regular graph sequence $(G_n)_n$ has a frequency model iff it has an EVPF.*

5 Conclusions and future directions

We have presented a paintbox representation for all edge-exchangeable graphs, and the correspondence between exchangeable vertex probability functions and graph frequency models. The paintbox is a powerful representation that can be directly analyzed to characterize properties of all edge-exchangeable graphs. We are currently investigating the use of these results to characterize edge-exchangeable graphs with sparse power laws, i.e., graphs in which the number of edges grows with the number of vertices with an exponent strictly between 1 and 2. We are also investigating other types of power laws involving, for example, the number of triangles or the degree distribution of the graph. Finally, we are further exploring the connection between the EVPF and efficient algorithms for posterior inference in edge-exchangeable graph frequency models.

References

[1] Anna Goldenberg, Alice Zheng, Stephen Fienberg, and Edoardo Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2010.

[2] Francois Caron and Emily Fox. Sparse graphs using exchangeable random measures. *ArXiv e-print 1401.1137v3*, 2015.

[3] Victor Veitch and Daniel M. Roy. The class of random graphs arising from exchangeable random measures. *ArXiv e-print 1512.03099*, 2015.

[4] Christian Borgs, Jennifer Chayes, Henry Cohn, and Nina Holden. Sparse exchangeable graphs and their limits via graphon processes. *ArXiv e-print 1601.07134*, 2016.

[5] Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures Algorithms*, 31(1):3–122, 2007.

[6] Christian Borgs, Jennifer Chayes, Cohn Henry, and Yufei Zhao. An lp theory of sparse graph convergence i: limits, sparse random graph models, and power law distributions. *ArXiv e-print 1401.2906*, 2014.

[7] Patrick J. Wolfe and Sofia C. Olhede. Nonparametric graphon estimation. *ArXiv e-print 1309.5936*, September 2013.

[8] Christian Borgs, Jennifer Chayes, Henry Cohn, and Shirshendu Ganguly. Consistent nonparametric estimation for heavy-tailed sparse graphs. *ArXiv e-print 1401.1137*, 2015.

[9] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2003.

[10] Mark E.J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[11] Aaron Clauset, Cosma R. Shalizi, and Mark E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[12] James R. Lloyd, Peter Orbanz, Zoubin Ghahramani, and Daniel M. Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. In *Adv. Neural Inform. Process. Syst. (NIPS) 25*, 2012.

[13] David J. Aldous. Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.*, 11(4):581–598, 1981.

[14] Douglas N. Hoover. Relations on probability spaces and arrays of random variables. Preprint, Institute for Advanced Study, Princeton, NJ, 1979.

[15] Peter Orbanz and Daniel M. Roy. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2):437–461, Feb 2015.

[16] Diana Cai, Trevor Campbell, and Tamara Broderick. Edge-exchangeable graphs and sparsity. In *Advances in Neural Information Processing Systems*, 2016. To appear.

[17] Harry Crane and Walter Dempsey. Edge exchangeable models for network data. *ArXiv e-print 1603.04571*, 2016.

[18] Jim Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158, 1995.

[19] Michael Escobar. Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994.

[20] Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

[21] Jaeyong Lee, Fernando Quintana, Peter Müller, and Lorenzo Trippa. Defining predictive probability functions for species sampling models. *Statistical Science*, 28(2):209–222, 2013.

[22] Trevor Campbell, Diana Cai, and Tamara Broderick. Exchangeable trait allocations. *ArXiv e-print 1609.09147*, 2016.

[23] Tamara Broderick, Jim Pitman, and Michael I. Jordan. Feature allocations, probability functions, and paintboxes. *Bayesian Analysis*, 8(4):801–836, 2013.