Random walk models of sparse graphs and networks

Benjamin Reddy
Department of Statistics
Columbia University
reddy@stat.columbia.edu

Peter Orbanz
Department of Statistics
Columbia University
porbanz@stat.columbia.edu

Abstract

We consider the problem of modeling the structure of graphs representing networks. This is a qualitatively different task than fitting a statistic of a network, such as the edge density or degree distribution. We argue that, to obtain faithful models of network structure, the probability of edges being formed in the graph has to be stochastically dependent on the existing link structure. That is not the case in most models currently used in the literature, such as exchangeable models or the preferential attachment model, since such dependencies make statistical estimation hard. We point out a class of models which do permit such dependencies and can represent a rich class of graph structures. These models insert vertices into a connected graph in a manner similar to preferential attachment models, but can also connect existing vertices by linking the initial and terminal vertex of a random walk of random length on the graph. Such models are in general mathematically intractable, but we show that, if the random walk is simple and its length a Poisson variable, parameters can be estimated by maximum likelihood.

1 Introduction

The statistical approach to modeling a given type of data is to define families of distributions which, in some well-defined sense, match the properties of the data source. For graph-valued and network data, this problem remains largely unsolved, despite the considerable attention it has received in recent years. One reason is arguably that this type of data is still poorly understood. Another is tractability: Since edges in a random graph are in general highly correlated random variables, all but a few special cases are mathematically and statistically intractable, and the insights and technical tools to identify these special cases are emerging only slowly.

We consider problems where the data is a single, large, simple undirected graph (as opposed to multiple graphs of a fixed size); as more data is observed, the graph grows. There are two large classes of models for this type of data. One class consists of exchangeable models [e.g. 17] and sparsified exchangeable models [4]. These include stochastic blockmodels, and more generally models represented by graph limits [15], and have recently received considerable attention in statistics and machine learning [3, 14, 17, 1, 20]. Under exchangeable models and their derivatives, edges in the random graph are conditionally independent given suitable summary information on the graph (this information can be represented as a "graph limit"), and this conditional independence makes it impossible to represent the types of structure commonly encountered in network data.

The second large class of models, often called generative or sequential models, are those which sequentially insert edges and/or vertices into a growing, connected graph. The most widely used example is the preferential attachment model of Barabási and Albert [2]. This model and various modifications have been studied intensely in applied probability [10], but its applicability in statistics is limited by the fact that the model has no parameters aside from the number of edges per vertex. To fit a statistical model, a richer parametrization is required, which in turn raises tractability issues.

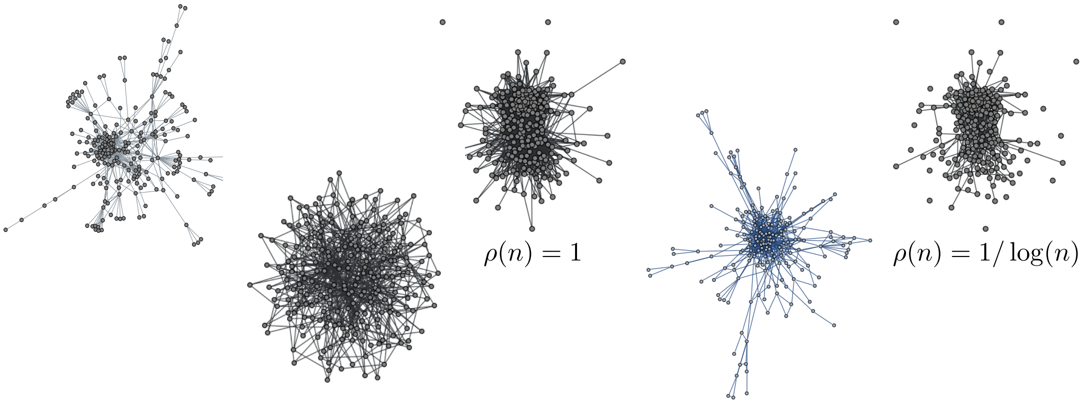

Figure 1: A protein-protein interaction network, and reconstructions generated by various models. *Top, left to right:* Input data, the exchangeable nonparametric model proposed in [14], chosen here because it performs competitively in comparison with other exchangeable models, and a sparsified version with $\rho(n) = 1/\log(n)$. *Bottom, left to right:* The preferential attachment model, and the uniform random walk model.

To obtain more expressive parametrizations, we consider models which permit the formation of links between vertices already contained in the graph—as opposed to the preferential attachment, which only describes how new vertices are inserted. If we assume that the insertion of an edge depends only on the current structure of the graph, not on its history (i.e. the order in which edges are inserted), the resulting model is a mixture of two processes: One which inserts vertices, and one which connects existing vertices.

This process can, for example, be interpreted as users in a social network meeting others users through their friends, or proteins in a protein-protein interaction network evolving to interact directly rather than through an intermediary protein. The technical motivation, however, is to obtain a tractable model of link formation in which the probability of connecting two vertices decays exponentially with distance in the current graph. Importantly, the rate of decay can be parameterized, and it controls the scale of the observed graph structure that is involved in edge formation. Although such random walk models are not tractable in general, we describe a specific case for which simple maximum likelihood estimators can be derived, based on the spectral properties of the graph.

2 Background and Motivation

In the following, we argue that random graph models which sample graphs vertex-wise—such as exchangeable models—are not suitable for the analysis of structure in most networks. It has already been recognized in the literature that exchangeable models necessarily generate graphs that are **dense**. In contrast, network data is typically **sparse**. There are several possible definitions of sparsity. We use the definition as in [17], where for a growing sequence of graphs, G_i, with edge sets $\mathbf{E}(G_i)$ and vertex sets $\mathbf{V}(G_i)$, the number of edges, $|\mathbf{E}(G_i)|$ is upper-bounded by $c \cdot |\mathbf{V}(G_i)|$, for some constant c; it is dense if $|\mathbf{E}(G_i)|$ lower-bounded by $c \cdot |\mathbf{V}(G_i)|^2$. Since exchangeable graphs are almost surely dense, they are misspecified as models of sparse data [17]. However, as we explain in this section, the misspecification stems not from lack of sparsity, but from the conditional independence between edges imposed by exchangeable models.

2.1 Exchangeable and Sparsified Exchangeable Models

Exchangeable models can be represented by a "graph limit" introduced by Lovász and Szegedy [16]. Denote by $\mathbf{W}$ the set of all measurable functions $[0,1]^2 \to [0,1]$. If a graph sequence converges, there is a function $w \in \mathbf{W}$ which defines the limiting distribution over graphs, $P =: P_w$. A finite random graph on N vertices, $G(w, N) \sim P_w$, with adjacency matrix X, can be generated as follows:

1. Sample $U_1, \dots, U_N \sim_{\text{iid}} \text{Uniform}[0,1]$.
2. For each pair $i, j \leq N$ of vertices, sample $X_{ij} \sim \text{Bernoulli}(w(U_i, U_j))$.

The function w is called a **graph limit** or **graphon**, and can be interpreted as a limiting representation of an adjacency matrix (see [15] for details). Since the variables U_i are i.i.d., the random graph $G(w, N)$ is **exchangeable**, i.e. its distribution is invariant under permutations of the vertex set. That is still true if we additionally randomize w, i.e. if W is a random function in $\mathbf{W}$ and we consider the random graph $G(W, \infty)$. The Aldous-Hoover theorem [13] shows that *any* exchangeable random graph G can be represented in this form, i.e. there is some random function W such that G is distributed as $G(W, \infty)$.

Unlike network graphs, a graph generated from an exchangeable graph model is almost surely dense [see 17, §7]. It can be sparsified by defining a decreasing **rate function** $\rho : \mathbb{N} \to (0, \infty)$, and substituting step 2. above by

$$X_{ij} \sim \text{Bernoulli}(\rho(N) w(U_i, U_j)) \ . \tag{2.1}$$

For suitable choices of ρ, such as $\rho(n) = \frac{1}{\log(n)}$ or $\rho(n) = \frac{1}{n}$, (2.1) generates sparse graphs [4].

Use in statistical modeling. Since the functions w parametrize random graph distributions, we can define a statistical model of random graphs by choosing a subset $\mathbf{W}_0 \subset \mathbf{W}$ as parameter space, resulting in the model $\{P_w | w \in \mathbf{W}\}$. If $\mathbf{W}_0$ is a finite-dimensional subspace, the model is parametric; if $\mathbf{W}_0$ is infinite-dimensional, it is non-parametric. Such models can be used in a Bayesian framework, by defining a prior distribution on $\mathbf{W}_0$ [14, 17], or in a non-Bayesian one by defining an estimator, either for w itself [1, 20] or for a functional of w [3].

Misspecification as models of network structure. Models based on graph limits, including the sparsified form (2.1), are inherently misspecified as models of network structure. The culprit is not the lack of sparsity, but rather conditional independence: In both the exchangeable model and in (2.1), the edges of the random graph are conditionally independent of each other given the function w and the uniform variables U_i. Consider the protein-protein interaction (PPI) network shown in Fig. 1 (top left). Under conditional independence, whether an edge is present in the graph depends only on its two end vertices, not on which other edges are or are not present. Hence, structures that appear in the PPI network—long chains, groups of vertices with degree one attached to a single center vertex, etc.—have negligible probability under such a model. Sparsification does not address this problem: Indeed, the sparsified model (2.1) can equivalently be sampled by first sampling a graph $G(w, N)$ from the exchangeable model, and then deleting each sampled edge independently with probability $(1 - \rho(N))$. To illustrate the misspecification problem, Fig. 1 compares the PPI network to its reconstructions from exchangeable and sparsified exchangeable models.

2.2 Sequential Models

Sequential sampling is the natural setting for models which grow a random graph by inserting edges and vertices, where the probability of insertion depends on the currently observed graph. The most prominent example is the **preferential attachment model** (PA), which generates a graph with N vertices as follows, starting from a single vertex: For $n = 2, \ldots, N$, sample m existing vertices biased by degree, then insert a new vertex connected to these m vertices. The model's only parameter is m, the number of connections per step. In the PA model, the distribution of edges at each step depends on the current graph, since it depends on the vertex degrees. The model is therefore not exchangeable. Unlike an exchangeable model, PA graphs are sparse, since $|\mathbf{E}(G)| = m(|\mathbf{V}(G)| - 1)$.

Although the PA model incorporates observed graph properties at each step, it is limited for two reasons. Firstly, edge formation between two existing vertices occurs with probability zero. Secondly, the *structure* of the graph at various scales is neglected; only the degree distribution, a global property, is considered. Variations of PA [6, chap. 3] that address the first limitation are still hampered by the second. A class of models called **duplication models** [8] attempts to incorporate local structure by copying some fraction of the edges of a randomly chosen vertex.

3 Sequential Models Based on Random Walks

The discussion in Section 2.1 shows that graphon-based models—although useful in a variety of contexts—are not adequate models of network structure. The same is true more generally for any model that renders edges conditionally independent, such as the sparsified model (2.1), since *to capture structures typically observed in network data, the probability of inserting edges in the sequential model must depend on the observed structure of the graph.* Our objective is hence to identify models which permit an at least slightly richer parametrization than preferential attachment models, but remain statistically tractable.

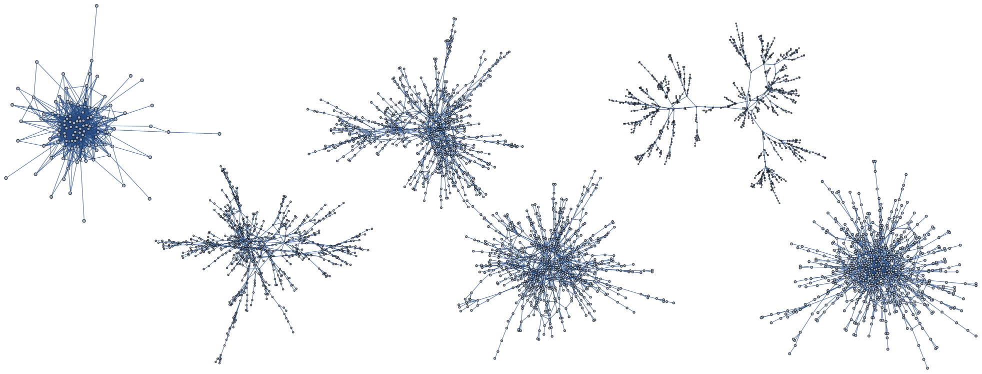

Figure 2: Graphs generated by the uniformly rooted random walk model. *Top:* $\lambda = 4$, with $\alpha \in \{0.1, 0.66, 0.9\}$ from left to right. *Bottom:* $\alpha = 0.66$, with $\lambda \in \{1, 10, 25\}$ from left to right.

3.1 Random Walk Models

We consider random graphs which grow edge-wise, i.e. at each step n of the sequential procedure, a graph G_{n+1} is obtained by inserting an edge (and vertices adjacent to this edge if necessary) into the previous graph G_n. A random graph G_{N} is hence generated as a sequence $G_{0:\mathrm{N}} = (G_0, \ldots, G_{\mathrm{N}})$, and $|\mathbf{E}(G_N)| = N$. We make two basic modeling assumptions: (1) Each graph G_n is connected, i.e. each vertex can be reached from every other vertex, and (2) the distribution of G_{n+1} depends only on G_n, not on the order in which the edges in G_n were generated. Hence, $G_{0:N}$ is a Markov chain on the set of connected graphs.

Connectedness implies that there are two ways to insert edges: Between two existing vertices, or between one existing and one new vertex. Since these two events are mutually exclusive, the probability of how the edge is inserted is a mixture; it is of the form

$$\alpha(G_n) \cdot \{\text{distribution on } \mathbf{V}(G_n)\} + (1 - \alpha(G_n)) \cdot \{\text{distribution on } \mathbf{V}(G_n) \times \mathbf{V}(G_n)\} \qquad (3.1)$$

for some $\alpha(G_n) \in [0, 1]$. Note that the Markov propery implies α is completely determined by G_n. Specification of a model requires choosing a distribution on vertices and one on pairs of vertices. We call the model a **random walk model** if pairs of vertices are chosen as the starting and terminal vertex of a finite-length random walk on G_n.

3.2 A Tractable Model

To obtain a statistically tractable model, our main design principle here is simplicity—as we will show below, even the simplest versions of this model generate a rich range of graphs. We define $\alpha(G_n)$ as a constant parameter $\alpha \in (0, 1]$, the random walk as simple random walk, and choose $\mu(v)$ to be one of the two basic distributions on the vertex set of a graph:

$$\mathbb{U}(v|G) = \frac{1}{|\mathbf{V}(g)|} \qquad \text{and} \qquad \mathbb{S}(v|G) = \frac{\deg(v)}{\sum_{u \in \mathbf{V}(G)} \deg(u)} \;, \qquad (3.2)$$

the uniform distribution $\mathbb{U}$ or the **size-biased** or **degree-biased distribution** $\mathbb{S}$, where $\deg(v)$ denotes the degree of the vertex v in G. Now generate a random graph G_{N} with N edges as follows:

1. For $n = 1, \ldots, N$, sample
   $$I_n \sim \text{Bernoulli}(\alpha) \qquad \text{and} \qquad K_n \sim \text{Poisson}(\lambda) \qquad \text{and} \qquad V_n \sim \mu_n \;. \qquad (3.3)$$
2. If $I_n = 1$, insert a new vertex into the graph and connect it to V_n.
3. Otherwise, start a simple random walk at V_n, terminate the walk after $K_n + 1$ steps, and connect V_n to the terminal vertex if the two are not currently connected.
4. If they are connected, restart the random walk. If V_n is connected to all current vertices, choose a new $V_n \sim \mu_n$ and proceed from step 3.

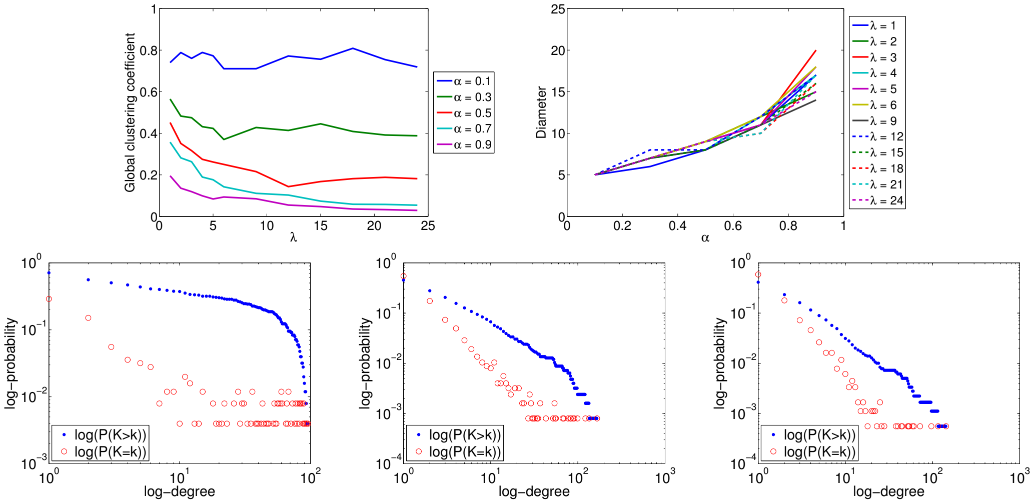

Figure 3: **Empirical results for the size-biased random walk model.** *Top left*: Plot of the global clustering coefficient as a function of λ. *Top right*: plots of the graph diameter as a function of α. *Bottom, left to right*: Empirical degree distribution for $\alpha \in \{0.1, 0.5, 0.7\}$, $\lambda = 5$. The power law is clearly expressed for $\alpha = 0.7$. Samples are all from graphs with $N = 2500$.

Step 3 means that the Bernoulli draw in step 1 is conditional on the graph not being complete (or $I_n = 1$ with probability 1), which occurs with negligible probability as n grows, and that V_n is chosen conditionally on not being connected to all other vertices.

We note that richer parameterizations, if implemented carefully, will still result in tractable estimation. In particular, one might parameterize the distribution μ, or use different distributions, μ_α and μ_λ, for sampling the attachment vertex.

3.3 Some Properties of the Model

This simple random walk model can generate a range of graph structures that is perhaps surprising, given that the model has only two parameters. Fig. 2 shows some examples. The model naturally exhibits a number of properties considered desirable in many network analysis problems:

- It is **transitive**, in the sense that vertices which share a common neighbor are more likely to be connected, especially when short random walks are likely. A measure of transitivity, the global clustering coefficient as defined in [18], is shown as a function of λ in Fig. 3.
- The **graph diameter**, i.e. maximum length of a shortest path, is primarily controlled by α. The smaller α, the more treelike the graph, resulting in larger diameters for a given number of edges.
- For suitable settings of the parameters, the size-biased random walk model shows clearly expressed **power law distribution** on the vertex degrees, see Fig. 3.
- Graphs generated by this model are **sparse**, since the number of edges added between the insertion of the i-th vertex and the $i+1$-st vertex is $D_i \sim_{\text{iid}} \text{Geometric}(\alpha)$, and the expected number of edges for n_v vertices is $(n_v - 1)\frac{1}{\alpha} \in \Theta(n_v)$.

4 Parameter Estimation

Parameter estimation in the model is possible due to the way in which a Poisson-length random walk interacts with the spectral properties of the graph. The two parameters of the model are the probability α of a new vertex, and the expected length λ of the random walks. Since these two parameter govern stochastically independent events, the joint parameter (α, λ) is identifiable.

The number of vertices is the outcome of N Bernoulli(α) experiments, thus we have $|\mathbf{E}(G_N)| = N$ and $|\mathbf{V}(G_N) - 1| \sim$ Binomial(α, N). The maximum likelihood estimator for α under the model is hence given simply by the binomial maximum likelihood estimator as $\hat{\alpha}_{\text{ML}} = N^{-1}(|\mathbf{V}(G_N)| - 1)$.

The distribution of a random walk depends on the graph structure, as does the estimate of λ. To estimate λ, we hence have to distinguish two cases:

(i) The case of "dynamic networks", where the order in which edges are generated is observed; that is, we observe a graph sequence $G_{0:\text{N}} := (G_0, \ldots, G_{\text{N}})$ in which G_n is obtained from G_{n-1} by inserting a single edge (and possibly a single vertex).

(ii) The case where only the final graph G_{N} is observed.

Conditional Probability of an Edge. Suppose G is a fixed graph, with adjacency matrix A and degree matrix $D := \text{diag}(\deg(v_1), \ldots, \deg(v_n))$. The probability that a simple random walk started at v_i reaches v_j on the k-th step is simply $(D^{-1}A)^k$. If we sample the length K of the walk at random from a distribution P, the probability of the random walk terminating at v_j is

$$\mathbb{P}(v_i \to v_j | G) = \sum_{k \in \mathbb{N}} \mathbb{P}(v_i \to v_j | G, K = k) P(k) = \Big(\sum_{k \in \mathbb{N}} (D^{-1}A)^k P(k) \Big)_{ij} . \tag{4.1}$$

The infinite series in general makes this expression intractable. In our model, however, P is specifically a Poisson distribution P_λ incremented by one, i.e. $P(k) = P_\lambda(k-1)$. In this case, a few lines of arithmetic reduce the infinite series (4.1) to a matrix exponential of the graph Laplacian Δ_{G},

$$\mathbb{P}(v_i \to v_j | G) = \big((\mathbf{I} - \Delta_{\text{G}}) e^{-\lambda \Delta_{\text{G}}} \big)_{ij} =: \mathbf{T}_{ij}(G, \lambda) . \tag{4.2}$$

Since the Laplacian is positive semi-definite and hence diagonalizable, the expression $e^{-\lambda \Delta_{\text{G}}}$, and thereby the infinite sum in (4.1), is tractable. We refer to [7] for details on graph Laplacians.

Step 3. in (3.3) requires that we only insert edges between two vertices that are not already connected. This alters the random walk probabilities in (4.2) as

$$\tilde{\mathbf{T}}_{ij}(G, \lambda) = \frac{\mathbf{T}_{ij}(G, \lambda) \mathbb{1}\{j \notin \mathcal{N}(i) \cup \{i\}\}}{1 - \sum_{k \in \mathcal{N}(i) \cup \{i\}} \mathbf{T}_{ik}(G, \lambda)} \mathbb{1}\{\mathcal{N}(i) \cup \{i\} \neq \mathbf{V}(G)\} , \tag{4.3}$$

where $\mathbb{1}$ is the indicator function, and $\mathcal{N}(i) \cup \{i\}$ is the union of vertex i and its neighbors.

Now suppose the graph sequence $G_{0:\text{N}}$ is observed. The two separate processes of edge insertion can be distinguished unambiguously: Suppose the edge distinguishing G_n from G_{n-1} connects vertices v_i and v_j in $\mathbf{V}(G_n)$. If only one vertex is contained in $\mathbf{V}(G_{n-1})$, a new vertex has been inserted, and the probability for this to happen depends only on α. Otherwise, the edge is the outcome of a random walk, in which case the probability of G_n conditioned on $I_n = 0$ is

$$\mathbb{P}(G_n | I_n = 0, G_{n-1}, \lambda) = \mu_n(v_i) \tilde{\mathbf{T}}_{ij}(G_{n-1}, \lambda) + \mu_n(v_j) \tilde{\mathbf{T}}_{ji}(G_{n-1}, \lambda) . \tag{4.4}$$

Case 1: Edge Sequence Observed Suppose the entire graph seqeunce $G_{0:\text{N}}$ is observed. Since the parameters are independent, the likelihood of λ is constant with respect to α, and hence

$$L(\lambda, G_{0:\text{N}}) = \prod_{n=1}^{N} \mathbb{P}(G_n | G_{n-1}, \alpha, \lambda) \propto \prod_{n | I_n = 0} \mathbb{P}(G_n | I_n = 0, G_{n-1}, \lambda) , \tag{4.5}$$

where, as in (3.3), $I_n = 0$ if the nth edge is inserted by random walk.

Case 2: Estimation From a Single Graph. If the order of edges is not observed, it can be imputed using a sampling algorithm: Given an input graph G, an ordering π of the edges is generated that is **valid**, meaning the edges of the graph could have been generated in that order without the graph being disconnected at any step. A graph G with N edges and a valid ordering π uniquely determine a graph sequence $G_{0:\text{N}}$, obtained by starting with a single vertex and adding edges in order π.

One way to generate a valid order given G is by means of importance sampling, where importance weights are computed that bias the generated order towards solutions plausible under the model. A similar sampler has been proposed for duplication-attachment models in [19]. A different and much more computationally efficient strategy, which one might call **extremely random sampling**[1], generates orderings purely at random, conditioned only on the ordering being valid, and an estimate of λ as follows:

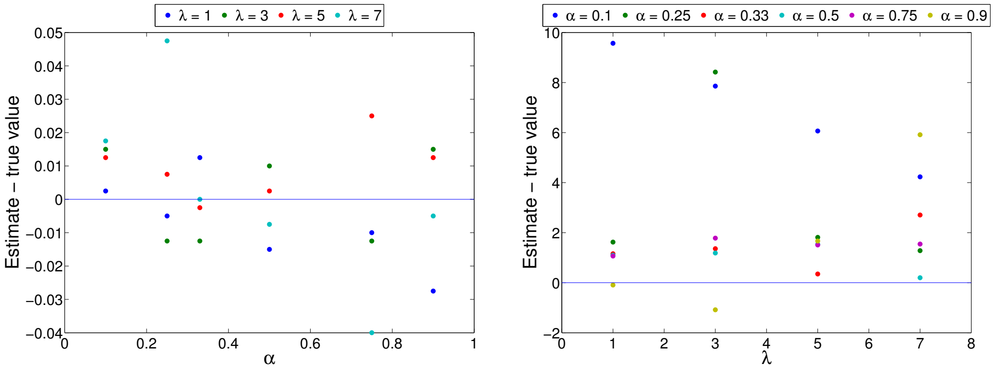

Figure 4: **Simulated accuracy of MLEs on a single graph.** Experiments on synthetic data generated using the size-biased distribution over vertices. The difference between the MLE and true parameter value (vertical) are plotted against the true values (horizontal); points on the blue line would be perfect estimates. In all simulations, $N = 400$, $M = 1000$. *Left*: α. *Middle*: λ.

1. For $m = 1, \ldots, M$, generate an edge ordering π^m as follows: Start with the entire graph G_{N}. Select a a vertex uniformly at random, then select one of its edges uniformly at random and delete it unless deletion disconnects the graph. Repeat until no edges are left.
2. For each $m = 1, \ldots, M$, compute the maximum likelihood estimator $\hat{\lambda}^m_{\text{ML}}$ from the graph sequence $G^m_{0:\text{N}}$ determined by the pair (G, π^m).
3. Return the averaged estimate $\hat{\lambda} := \frac{1}{M} \sum_m \hat{\lambda}^m_{\text{ML}}$.

This algorithm is obviously significantly faster than the importance sampler, which requires calculation of edge probabilities at each deletion step; intriguingly, the resulting parameter estimates are every bit as accurate. Even for moderate graph sizes, both importance sampling and extremely random sampling recover the model parameters with high accuracy. Results for the (more efficient) extremely random sampler are reported in Section 5.

5 Experimental Results

5.1 Parameter Estimation

Estimation on synthetic data. To test our estimation scheme, we sample graphs with $N = 400$ edges for various settings of (α, λ). Fig. 4 shows that, for a wide range of parameter settings, the estimation scheme recovers the true parameter. The value of α recovered with high accuracy, as one might predict for such a simple estimator. The estimates of λ have lower accuracy. One reason is that the effective sample size for estimating λ is $(1-\alpha)N$; as N increases, this source of inaccuracy goes away. Another reason is that when λ is large relative to the diameter of the graph (see Fig. 3), the random walk distributions approach the equilibrium distribution, i.e. the distribution of a random walk of infinite length. More precisely, $\mathbf{T}_{\cdot j}(G, \lambda) \to \text{deg}(j)/\text{vol}(G)$ for any starting vertex i ([5, chap. 1]). When this happens, information about λ is lost, and the variance of $\hat{\lambda}$ is high.

Estimation on a protein-protein interactome. With our model and estimation scheme, parameters of a single graph, such as the interactome ($N = 695$, $n_v = 230$) in Fig. 1, can be estimated by sampling. For the uniform random walk model, we estimate 95% confidence intervals of $(\hat{\alpha}, \hat{\lambda}) = (0.33, 4.06) \pm (0.035, 0.34)$. (Intervals for $\hat{\alpha}$ are estimated with the usual asymptotic properties of MLEs; for $\hat{\lambda}$ with the standard error of the samples $\hat{\lambda}^m$). With the size-biased random walk model, we estimate $(\hat{\alpha}, \hat{\lambda}) = (0.33, 3.99) \pm (0.035, 0.34)$. We can compare the reconstruction of a graph generated by the random walk model to that generated by other models—although these comparisons come with all the caveats of "visual" network comparisons. See Fig. 1 (bottom right).

[1] This strategy is loosely based on the notion of an **extremely randomized tree**, a type of random forest classifier in which the tree structure is generated purely at random and independent of the training data [11].

5.2 Comparison with Other Models: Link Prediction

As a measure of performance of our simple random walk models, we test them on the following link prediction task: Starting with the full PPI network, we select an edge uniformly at random (conditional on the resulting graph being connected), then estimate the model parameters on the remaining graph, and calculate the predicted edge probability of all $\frac{n_v(n_v-1)}{2} + n_v - (N-1)$ possible edges. False positives are identified as potential edges with greater predicted probability than the edge that was actually deleted. We then repeat for 10% of the total number of edges.

For comparison, we conducted the same experiment on the same set of deleted edges with different models for network data: the Barabási-Albert PA model; a more flexible version of the PA model analyzed by Chung [6], which is a limiting case of our model in which $\lambda \to \infty$, as all vertex and edge insertions are sampled from the size-biased distribution; and Hoff's eigenmodel [12], which was one of the first models to explicitly use the Aldous-Hoover theory of exchangeable random graphs.[2] Results are shown in the following table.

Model	BA-PA	Chung-PA	Eigenmodel, K=3	Eigenmodel, K=10,	U-RW	SB-RW
AUC	0.134	0.899	0.514	0.500	**0.981**	**0.981**

Clearly, the Barabási-Albert PA model is limited by its inability to insert edges between existing vertices; those potential edges have probability zero. The more flexible version of the PA model performs well. It is somewhat limited, though, by its inability to model structure on a smaller scale.

A word of caution when comparing our model to any exchangeable model, such as the eigenmodel, with link prediction: The way in which our model predicts edge probabilities is fundamentally different from that of an exchangeable model. The conditional independence between edges in exchangeable models means that an adjacency entry $X_{ij} = 0$ indicates *there is no edge*. In our model, $X_{ij} = 0$ means we have not observed an edge, but *an edge may be observed* later in the sequence. Distributions over exchangeable graphs of size N and $N+1$ are not projective, so adding an edge does not have the same meaning there as adding an edge in a sequential model. This difference is the cause of the eigenmodel performing slightly better than random guessing.

6 Conclusions

As we have argued in Section 2, faithful models of network structure need to express a certain amount of stochastic dependence between edges. There hence exists a trade-off between model accuracy and tractability: Exchangeable models are convenient and reasonably tractable, but the independence assumptions they impose are too strong to preserve network structure. On the other hand, models that introduce complicated dependencies may be relatively easy to define and to simulate from, but dependence between edges typically makes statistical estimation impossible. The random walk model presented here is an attempt to identify a "sweet spot" between these extremes.

The model has appealing properties—power law behavior, a wide range of expressible structures and strong predictive performance. We are not presently able to rigorously proof any of these properties, and to the best of our knowledge, doing so defies the mathematical toolkit currently available to study random graphs: The technical obstacle is the distribution of the random walk, which depends on the current graph, and hence changes at each step of the sequential process. In random walk jargon, the distribution of the walk is called its **environment**, and our model constitutes a **restarted random walk in a dynamic random environment**. A few results on such processes are available, but require strong simplifying assumptions which even a model as basic as ours does not satisfy (see [9] for a recent survey). It is hence perhaps surprising that statistical estimation under the model turns out to be feasible and well-behaved.

[2] We also conducted the experiment on the stochastic block model (SBM), which performed similarly to the eigenmodel. This is to be expected, as the eigenmodel is a generalization of the SBM.

References

[1] Airoldi, E. M., Costa, T. B., and Chan, S. H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation.

[2] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.

[3] Bickel, P. J., Chen, A., and Levina, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.*, **39**(5), 2280–2301.

[4] Bollobás, B., Janson, S., and Riordan, O. (2007). The phase transition in inhomogeneous random graphs. *Random Structures Algorithms*, **31**(1), 3–122.

[5] Chung, F. (1996). *Spectral Graph Theory*. Number 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Society.

[6] Chung, F. and Lu, L. (2006). *Complex Graphs and Networks*. Number 107 in CBMS Regional Conference Series in Mathematics. American Mathematical Society.

[7] Chung, F. and Yau, S.-T. (2000). Discrete Green's functions. *J. Combin. Theory Ser. A*, **91**(1-2), 191–214. In memory of Gian-Carlo Rota.

[8] Chung, F., Lu, L., Dewey, T. G., and Galas, D. J. (2003). Duplication models for biological networks. *Journal of Computational Biology*, **10**(5), 677–687.

[9] Drewitz, A. and Ramírez, A. F. (2013). Selected topics in random walk in random environment.

[10] Durrett, R. (2006). *Random Graph Dynamics*. Cambridge University Press.

[11] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, **63**(1), 3–42.

[12] Hoff, P. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Adv. Neural Inf. Process. Syst. 2007*.

[13] Kallenberg, O. (2005). *Probabilistic Symmetries and Invariance Principles*. Springer.

[14] Lloyd, J. R., Orbanz, P., Ghahramani, Z., and Roy, D. M. (2012). Random function priors for exchangeable arrays. In *Adv. in Neural Inform. Processing Syst. 25*, pages 1007–1015.

[15] Lovász, L. (2013). *Large Networks and Graph Limits*. American Mathematical Society.

[16] Lovász, L. and Szegedy, B. (2006). Limits of dense graph sequences. *J. Combin. Theory Ser. B*, **96**, 933–957.

[17] Orbanz, P. and Roy, D. M. (2013). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, to appear.

[18] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, **393**(6684), 440–442.

[19] Wiuf, C., Brameier, M., Hagberg, O., and Stumpf, M. (2006). A likelihood approach to analysis of network data. *Proceedings of the National Academy of Sciences*, **103**(20), 7566–7570.

[20] Wolfe, P. J. and Olhede, S. C. (2013). Nonparametric graphon estimation. Preprint. http://arxiv.org/abs/1309.5936.