Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Incorporating Assortativity and Degree Dependence into Scalable Network Models

Stephen Mussmann[1,2]* and **John Moore**[1]* and **Joseph J. Pfeiffer III**[1] and **Jennifer Neville**[1]
Departments of Computer Science and Statistics[1], and Mathematics[2]
Purdue University, West Lafayette, IN
Email: {somussma, moore269, jpfeiffer, neville}@purdue.edu
*Authors contributed equally

Abstract

Due to the recent availability of large complex networks, considerable analysis has focused on understanding and characterizing the properties of these networks. Scalable generative graph models focus on modeling distributions of graphs that match real world network properties and scale to large datasets. Much work has focused on modeling networks with a power law degree distribution, clustering, and small diameter. In network analysis, the *assortativity* statistic is defined as the correlation between the degrees of linked nodes in the network. The assortativity measure can distinguish between types of networks—social networks commonly exhibit positive assortativity, in contrast to biological or technological networks that are typically disassortative. Despite this, little work has focused on scalable graph models that capture assortativity in networks.

The contributions of our work are twofold. First, we prove that an unbounded number of pairs of networks exist with the same degree distribution and assortativity, yet different joint degree distributions. Thus, assortativity as a network measure cannot distinguish between graphs with complex (nonlinear) dependence in their joint degree distributions. Motivated by this finding, we introduce a generative graph model that explicitly estimates and models the joint degree distribution. Our *Binned Chung Lu* method accurately captures both the joint degree distribution and assortativity, while still matching characteristics such as the degree distribution and clustering coefficients. Further, our method has subquadratic learning and sampling methods that enable scaling to large, real world networks. We evaluate performance compared to other scalable graph models on six real world networks, including a citation network with over 14 million edges.

1 Introduction

Networks are commonly used as a representation for understanding complex systems. For example, interactions of physical systems (e.g., internet topology), biology (e.g., protein interactions), the Web (e.g., hyperlink structure) and social networks (e.g., Facebook, email) can be represented as networks. This has led to considerable interest in understanding these networks and constructing generative network models that can capture a wide range of network behavior. Generally speaking, generative graph models are used to reason about observed graph structures by constructing graphs that are, ideally, structurally similar to an input graph. The methods use an input (observed) graph to estimate parameters of the model, then the learned parameters are used to generate a new network by drawing a sample from the model. A variety of scalable generative network models exist that capture common network structure such as a power law degree distribution, small diameter and clustering coefficients (Chung and Lu 2002; Kolda et al. 2013; Leskovec et al. 2010; Pfeiffer III et al. 2012).

However, the *assortativity* of a network, or the tendency for vertices to form connections with other vertices of similar degree, is largely overlooked in these modeling efforts. This is surprising in the sense that assortativity typically reflects the network type: social networks generally have positive assortativity, while biological and technological networks are generally disassortative (i.e., negative assortativity) (Stanton and Pinar 2011). Thus, the assortativity in a network can provide significant insight into the type of complex network observed, making it a popular statistic within social network analysis (Newman 2002; Zhou et al. 2012).

Although assortativity is rarely directly modeled, it may be produced as a byproduct of the modeling process. For example, the Block Two-Level Erdos Renyi (BTER) model (Kolda et al. 2013) groups clusters of vertices with similar degree together in such a way that the vertices *within* a cluster have higher probability of linking to each other; this representation produces networks with high clustering and assortativity. However, this approach cannot produce disassortative networks and moreover, the assortativity and clustering cannot be varied independently, since they are tied together in the process. Stanton and Pinar (2011) explicitly model the joint degree distribution of a network, which generalizes the assortativity summary statistic. However, their approach that directly models the entire joint degree distribution is quadratic in the number of unique degrees in the network, and their model cannot easily model other network statistics (e.g., clustering).

In this work, we explore two aspects of this problem. First, we formally discuss the relationship between the assortativity of a network and its joint degree distribution. Namely, we prove that assortativity alone cannot summarize the joint degree distribution of a network: even networks

with identical degree distributions and assortativity can have provably different joint degree distributions. Second, we develop a *Binned Chung Lu* (BCL) generative graph model that can *learn* assortativity patterns from an input graph and reflect them accurately in the sampled networks. Specifically, we enhance the Chung Lu (CL) family of models (Chung and Lu 2002; Pfeiffer III et al. 2012) to learn and incorporate a model of the joint degree distribution (JDD)—using a coarse *binning* technique to summarize the JDD with a small number of parameters and a statistical accept-reject sampling process that augments the original CL model to sample networks to match the the observed patterns in the JDD. Unlike previous approaches, our BCL method can accurately *learn* and model a wide range of joint degree distributions (i.e., with both positive and negative assortativity). Moreover, since our binning approach provably matches the coarse JDD of the original network, it will also coarsely preserve the assortativity. We also prove that our BCL approach maintains the same expected degree distributions as the baseline CL model and remains scalable (i.e., subquadratic). Our contributions can be summarized as follows:

- Introduction of a coarse binning technique to capture the joint degree distribution and assortativity found in a range of real world networks.
- Proof that assortativity cannot distinguish joint degree distributions, including a constructive proof for an infinite class of graphs and an illustrative real-world example.
- Development of an accept-reject sampling process to efficiently sample networks with a prescribed JDD and assortativity, and a method to efficiently learn the corresponding model parameters.
- Proofs that our BCL approach only incurs an extra subquadratic learning and constant sampling cost.
- Proofs that our model and sampling algorithm preserves the degree distribution and bin frequencies (i.e., approximate JDD).

Our BCL approach is general enough to use with any CL model, but for this paper we implement and experiment with two different baseline models: Fast Chung Lu with Binning (FCLB), which extends the Fast Chung Lu model of (Kolda et al. 2013), and Transitive Chung Lu with Binning (TCLB), which extends the Transitive Chung Lu model of (Pfeiffer III et al. 2012). We apply our methods to six network datasets and show they are able to accurately capture assortativity (significantly better than baseline methods) while maintaining other graph properties of the original models. Additionally, to empirically demonstrate the scalability of the models, we apply them to learn and sample from a Patents dataset with 14 million edges.

In the following section, we outline the notation used, introduce the Chung Lu model with its variants, and discuss related work. Section 3 discusses assortativity in detail. Section 4 outlines our BCL method and analyzes its characteristics. In section 5, we compare BCL experimental to competing models. We conclude in section 6.

2 Notation and Background

Let a graph $G = \langle \mathbf{V}, \mathbf{E} \rangle$ define a set of *vertices* or nodes $\mathbf{V}$, with a corresponding set of edges $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$. We denote the individual edges $e_{ij} \in \mathbf{E}$ where $e_{ij} = (v_i, v_j)$. As the graph is undirected, $e_{ij} \in \mathbf{E}$ if and only if $e_{ji} \in \mathbf{E}$. Let D_i be the degree of node v_i in the graph. Let M be the total number of edges in the graph. Note that $|\mathbf{E}| = 2M$ since each undirected edge corresponds to two elements in $\mathbf{E}$. We refer to the number of nodes of each degree as the degree distribution of the graph.

Assortativity is a graph measure defined in (Newman 2002); formally, it is the correlation of degrees across edges observed in the network. Define $T_\beta(e_{ij}) = D_i$ and $T_\epsilon(e_{ij}) = D_j$. Intuitively, $\{T_\beta(e)|e \in \mathbf{E}\}$ is the distribution of degrees of the startpoints of edges in graph G, while $\{T_\epsilon(e)|e \in \mathbf{E}\}$ is the distribution of degrees of the endpoints in graph G. The assortativity is the Pearson correlation coefficient for these variables:

$$\mathcal{A} = \frac{cov(\{T_\beta(e), T_\epsilon(e)|e \in \mathbf{E}\})}{\sqrt{var(\{T_\beta(e)|e \in \mathbf{E}\})} \cdot \sqrt{var(\{T_\epsilon(e)|e \in \mathbf{E}\})}} \quad (1)$$

Note that as the graph is undirected, the variances of the random variables are equal; however, the *covariance* can widely vary depending on the network structure (as we will explore in detail in Section 3). A more in depth discussion of assortativity can be found in (Newman 2002).

Chung-Lu Models

The Chung Lu models are a family of models such that the expected degree of a vertex in a sampled network is equal to the degree of the node in the original network (Chung and Lu 2002). Chung Lu models define the marginal probability of an edge to be proportional to the product of the endpoint degrees normalized by a constant; i.e., $(D_i D_j)/2M$.

The structural assumptions of the model enable a fast sampling method for sparse networks; although the model defines a set of edge probabilities quadratic in the number of vertices, efficient sampling algorithms exist that are linear in the number of edges. This Fast Chung Lu (FCL) algorithm (Pinar, Seshadhri, and Kolda 2011) generates a graph by repeatedly sampling proportional to their defined probabilities and adds them to the network. Let G be an observed graph and π be a normalized degree distribution ($\pi(i) = D_i/2M$). For M iterations, the FCL algorithm samples twice from π to select an i, j pair and then adds the edge e_{ij} to the output graph G'. FCL provably samples a given network's degree distribution in expectation. However, for most graphs, the observed clustering coefficient is much higher than what is generated by FCL.

In contrast, the Transitive Chung Lu (TCL) graph model extends the simple FCL method to incorporate *transitivity* into the sampling process (Pfeiffer III et al. 2012). In particular, TCL is a mixture model: with probability $1 - \rho$ it samples an edge using FCL, but with probability ρ performs a two hop random walk across the current network sample. Starting at a vertex v_i, once it takes two random steps it places an edge between v_i and the endpoint v_j. Thus, TCL

incorporates transitivity into the network distribution by inserting triangles at a higher probability. This adjustment provably maintains the degree distribution, but allows TCL to accurately model the clustering of real world networks. TCL has a simple and scalable learning algorithm, meaning that it already can accurately learn p from large scale networks. Further, TCL provably samples twice from the corresponding marginal degree distributions, which will be an important criteria for this work. Our method will be able to be used with TCL, allowing us to model large scale networks' degree distributions, clustering coefficients and joint degree distributions.

Attributed Graph Models

Recently, the *Attributed Graph Model* (Pfeiffer III et al. 2014) was proposed as an extension to generative graph models to allow for jointly modeling the dependencies between vertex attributes and network structure. This method augments generative graph models to incorporate correlated attributes while accurately maintaining patterns in the graph structure. In that work, *accept-reject* sampling was used to keep proposed FCL and TCL edges from being inserted into a final network if they did not match the desired correlations (with some probability). This representation provably samples from the set of edges conditioned on the vertex attributes, and has scalable learning and sampling algorithms for determining the acceptance probabilities given an attributed network. In this work, we expand on this to prove that higher order graph statistics (namely, the joint degree distribution) can be efficiently learned and sampled while maintaining lower order statistics (degree, clustering).

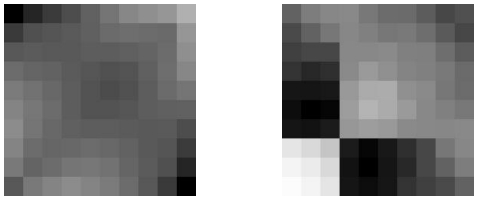
Accept-reject sampling is a framework used in statistics (Liang, Liu, and Carrol 2010) that we will employ in our BCL method. In particular, given a proposal distribution Q' that is easy to sample from, we wish to use Q' to sample from a more complex target distribution Q. Using accept reject sampling, we can repeatedly draw from Q' but only *accept* the sample some of the time. In particular, for each value x that is drawn from Q', we flip a Bernoulli coin with probability proportional to $\frac{Q(x)}{Q'(x)}$ and only accept the sample x if the Bernoulli trial is a success. In particular the probability that x is sampled is $A(x) = \frac{Q(x)}{Q'(x)Z}$, where Z is a constant such that $\forall\ x,\ A(x) \leq 1$. Note that as samples are drawn from $Q'(x)$, the resulting distribution (after the rejection process) is simply $Q(x)$.

3 Assortativity in Graph Models

Assortativity is a measure of the correlation among the degrees of linked nodes in a network (Newman 2002). It is a popular measure used to categorize and understand the structure of a network, since different types of networks have varying amounts of assortativity. For example, social networks tend to have positive assortativity while biological and technological networks tend to have negative assortativity (Stanton and Pinar 2011). See Table 1 for the assortativity ($\mathcal{A}$) we observed across six networks of varying sizes that we considered in this work, ranging from 0.18 to -0.29 (the details on each is discussed in Section 5).

Graph	Nodes	Edges	$\mathcal{A}$	$\hat{\mathcal{A}}_{TCL}$
Facebook Wall	444,829	1,014,542	-0.297	-0.0021
Purdue Email	54,076	880,693	-0.1161	-0.0092
Gnutella	36,682	88,328	-0.1034	0.0006
Epinions	75,865	385,418	0.0226	-0.0363
Rovira Email	1,133	5,451	0.0782	-0.0200
Patents	2,745,762	13,965,410	0.1813	0.0004

Figure 1: Network Statistics



(a) Purdue Email (b) Gnutella

Figure 2: Joint degree distribution representations $\mathcal{B}$; $k = 10$.

However, despite its popularity as a graph measure, assortativity can fail to capture important dependencies in the joint degree distribution. This is because it measures degree correlation and thus focuses on linear relationships. Consider the real-world examples depicted in Figures 2.a-b, which illustrate the joint degree distributions of the Purdue email and Gnutella networks respectively. To (coarsely) visualize the joint degree distribution, we divide the degrees of a graph G into k quantiles that we refer to as $\mathbf{B}_k = [B_1, B_2, ..., B_k]$. In Figure 2, we use $k = 10$. Let $b(D)$ be a function that returns the set membership in $\mathbf{B}_k$ for a given degree D. We can now construct a $k \times k$ matrix $\mathcal{B}$ to represent the joint degree distribution, where each cell i, j counts the number of edges between nodes with degrees in B_i and those with degree B_j. In other words, an edge e_{ij} would be counted in cell $[b(D_i), b(D_j)]$. When visualizing $\mathcal{B}$, we use a gray scale intensity plot to indicate the number of edges in each cell (i.e., a cell without any edges will be colored white and a cell with the largest amount of edges will be close to black).

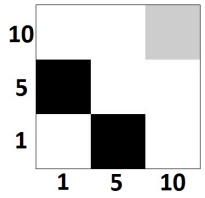
Assuming the axes of $\mathcal{B}$ are ordered by increasing degree, as we do above, these plots show information related to assortativity. If the darker boxes form a line with a positive slope, the graph will have positive assortativity. In contrast, if the dark boxes form a line with a negative slope, the graph will have negative assortativity. More precisely, the *binned* plots are a histogram approximation to the full joint degree distribution; they graphically represent the dependencies between the various degrees in the network.

The Purdue Email and Gnutella datasets have similar assortativity values of -0.1161 and -0.1034, respectively. However, their joint degree distributions are quite different as can be seen in Figures 2.a-b. These examples illustrate evidence in support our claim that the single dimensional measure of assortativity does not fully capture the dependencies we observe in joint degree distributions in real networks.

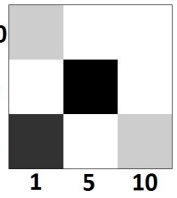
To expand on this individual example, we prove in Theo-

D	A	B
1	8910	8910
5	1782	1782
10	176	176

(a) Degrees



(b) Graph A



(c) Graph B

Figure 3: (a) Degree distribution; (b-c) Joint degree distribution representations $\mathcal{B}$; $k=3$.

rem 1 that there are infinitely many pairs of graphs with the same assortativity and degree distribution, but maximally different joint degree distributions (Proof in Appendix).

Theorem 1. *Let A_w, B_w be two networks comprised of graphlets, where w parameterizes the size and counts of the graphlets. There exist an infinite set of pairs of graphs $\{A_w, B_w\}$ such that for any $w \geq 2$, A_w and B_w have the same degree distribution, the same assortativity, but infinite KL-Divergence between their joint degree distributions.*

As an example of a pair of graphs that are covered by Thm. 1, we construct two Graphs A and B with $w = 5$. The graphs are composed of disconnected subgraphs that are either stars or cliques. Graph A consists of 1782 stars of size 5 and 16 cliques of size 11. Graph B consists of 176 stars of size 10, 297 cliques of size 6, and 3575 cliques of size 2. Both of these graphs have the exact same degree distribution (see Table 3.a) and the same assortativity: $\mathcal{A} = \frac{9}{187}$. However, Figures 3.b-c show that the two graphs have very different joint degree distributions, despite their identical assortativity and degree distributions. Further, as the two joint degree distributions have disjoint support, the KL-Divergence between them is infinite. Next, we propose a novel approach to modeling assortativity in networks that considers dependencies in the full joint distribution.

4 Binning Chung Lu

Most generative graph models are not able to reproduce assortativity, and even fewer model negative assortativity. For example, although Chung Lu (CL) models preserve other network statistics, the processes produce no correlation between the degrees of edge endpoints. This results in network samples with near zero assortativity (see Figure 1). We now propose the Binning Chung Lu (BCL) method to model assortativity in networks.

Generally speaking, our BCL methods use an existing edge-by-edge generative graph model (such as FCL or TCL) to *propose* possible edges from their respective distributions. BCL then filters, or conditionally accepts, a subset of the proposed edges into a final network sample. By estimating the correct acceptance probabilities, the final accepted edge samples are provably from the desired joint degree distribution. Our approach is a form of accept reject sampling and general enough to augment any Chung Lu model. In this paper, we implement our approach using both FCL and TCL—referring to them as FCLB and TCLB respectively.

More specifically, BCL expands on the binning visualization introduced in the previous section. First, BCL creates a "degree vector", where the pair (D_i, v_i) is inserted in the vector D_i times. We then sort this vector by the degrees of the vertices, followed by dividing the vector into k equal segments, or *quantiles*. These quantiles correspond to the bins discussed in the visualization from the previous section: every vertex v_i is assigned to its corresponding bin, or quantile, as determined by the vertex's degree D_i. Similarly, the vertices contained within a bin B_p are the vertices with degrees in the p^{th} quantile of the degree vector. Then the $k \times k$ matrix $\mathcal{B}$ represents the count of the number of observed edges e_{ij} that fall in a particular cell indexed by $[b(D_i), b(D_j)]$.

Given $\mathcal{B}$, which is our coarse, *binned* representation of the joint degree distribution, we can outline an accept-reject sampling method to extend a given FCL or TCL model. Formally, the original CL model is a proposal distribution Q': FCL and TCL are chosen due to their scalable sampling processes. We define the target distribution Q to be a network distribution with the same binned joint degree distribution as the input graph.

To implement BCL, we first compute the bin representation $\mathbf{B}_k$ and associated $\mathcal{B}$ from an input graph G. We then generate a graph G' from the CL model, using G as input, and compute the resulting joint degree distribution $\mathcal{B}'$ from G' (using $\mathbf{B}_k$). With $\mathcal{B}$ and $\mathcal{B}'$, we estimate the acceptance probabilities A:

$$R(m,n) = \frac{\#\{e_{ij} \in \mathbf{E} \mid b(D(i)) = m \wedge b(D(j)) = n\}}{\#\{e_{ij} \in \mathbf{E}' \mid b(D(i)) = m \wedge b(D(j)) = n\}} = \frac{\mathcal{B}_{mn}}{\mathcal{B}'_{mn}}$$

$$A(m,n) = \frac{R(m,n)}{\max_{mn} R(m,n)} = \frac{R(m,n)}{R_{max}}$$

We refer to the sampling process of the CL model as $\pi(\Theta)$. Thus, we sample two nodes i, j from π and after each sample, determine whether to accept or reject the edge e_{ij} depending on the corresponding bin's acceptance probability. We repeat this process until we have sampled as many edges as the original graph.

This process is described in Algorithm 1. G is the original graph, π is the CL model, and Θ is the parameters for the CL model. Note that we must first generate a preliminary graph using the CL model so that we can compute the acceptance probabilities. For a proposed edge e_{ij} (line 7), $A(b(D_i), b(D_j))$ returns the acceptance probability from the appropriate cell of $\mathcal{B}$ (Line 8). Lastly, lines 9-12 determine whether to accept the proposed edge into the final set.

In the next subsection, we prove that our BCL approach maintains the original degree distributions as guaranteed by FCL and TCL, while also modeling the true joint degree distribution of the original network and (by extension) the assortativity.

Analysis of Binning Chung Lu Models

In this section, we prove the following key theorems:

- **Theorem 2:** In expectation, the Binning Chung Lu models provably sample from the original coarse JDD.

Algorithm 1 Binning Chung Lu Method(G, π, Θ, k)

Compute $k \times k$ bin frequencies $\mathcal{B}$ from G
Generate G' from π, using G and Θ
Compute $k \times k$ bin frequencies $\mathcal{B}'$ from G'
Compute $A(m, n)$ from $\mathcal{B}$ and $\mathcal{B}'$
Create empty graph G^{BCL}
while $|E^{BCL}| \leq |E|$ **do**
  $< i, j >=$ edge sampled using $\pi(\Theta)$
  $a = A(b(D_i), b(D_j))$
  $r = bernoulli\_sample(a)$
  if $r = 1$ **then**
    $E^{BCL} = E^{BCL} \cup e_{ij}$
  end if
end while
$return(G^{BCL})$

- **Theorem 3:** The expected degree of a node sampled using Binning Chung Lu models equals the original degree.

These proofs exploit a key relationship between the Chung Lu graph models and the defined quantiles; namely, the Chung Lu graph models sample uniformly from the coarse JDD representation. In the following Lemma, we use this to derive the probability a given sample from a Chung Lu model is accepted by Binning Chung Lu. This is subsequently used to prove both Theorem 2 and 3.

Lemma 1. *In BCL, if the edges are sampled from a Chung Lu model, the marginal probability that an edge sample is accepted is* $\frac{1}{R_{max}}$.

Proof. First, the acceptance probability is equivalent to marginalizing over all the acceptance probabilities for each e_{ij}. Let k_1, k_2 indicate a particular bin index. Then:

$$\begin{aligned}
P(\text{accepted}) &= \sum_{e_{ij}} P(e_{ij}) P(\text{accepted}|e_{ij}) \\
&= \sum_{e_{ij}} \frac{D_i D_j}{(2M)(2M)} \sum_{k_1,k_2} \frac{\mathbb{I}[e_{ij} \in R(k_1,k_2)] \cdot R(k_1,k_2)}{R_{max}} \\
&= \frac{1}{R_{max}} \sum_{k_1,k_2} \sum_{e_{ij} \in Bin(k_1,k_2)} \frac{D_i D_j}{(2M)(2M)} R(k_1,k_2) \\
&= \frac{1}{R_{max}} \sum_{k_1,k_2} \sum_{e_{ij} \in Bin(k_1,k_2)} \frac{D_i D_j}{(2M)(2M)} \frac{\mathcal{B}_{k_1 k_2}}{\mathcal{B}'_{k_1 k_2}}
\end{aligned}$$

The edge probabilities can be further split by their respective quantiles $Qu(k_1), Qu(k_2)$, i.e., $Qu(k) = \{v_i | b(D_i) = B_k\}$:

$$\begin{aligned}
P(\text{accepted}) &= \frac{1}{R_{max}} \sum_{k_1} \sum_{k_2} \left(\sum_{i \in Qu(k_1)} \frac{D_i}{2M} \right) \left(\sum_{j \in Qu(k_2)} \frac{D_j}{2M} \right) \frac{\mathcal{B}_{k_1 k_2}}{\mathcal{B}'_{k_1 k_2}} \\
&= \frac{1}{R_{max}} \frac{1}{k} \frac{1}{k} \sum_{k_1} \sum_{k_2} \frac{\mathcal{B}_{k_1 k_2}}{\mathcal{B}'_{k_1 k_2}}
\end{aligned}$$

As Chung Lu models have uniform bin distribution $\forall\ k_1, k_2$, we have $\mathcal{B}'_{k_1 k_2} = \frac{2M}{k^2}$. Thus:

$$P(\text{accepted}) = \frac{1}{R_{max}} \frac{1}{k} \frac{1}{k} \frac{k^2}{2M} \sum_{k_1} \sum_{k_2} \mathcal{B}_{k_1,k_2}$$

As the sum over all the bin frequencies is $2M$, we simplify the above to recover $P(\text{accepted}) = \frac{1}{R_{max}}$. □

Using the above lemma, we show that although the proposal distribution (FCL or TCL) proposes edges from a uniform joint degree distribution, the edges *accepted* into the network sample provably model the joint degree distribution of the original network.

Theorem 2. *For a graph G^{BCL} generated by BCL, the expected edge frequency in bin $\mathcal{B}^{BCL}_{k_1 k_2}$ is equal to the frequency in bin $\mathcal{B}_{k_1 k_2}$, from the original input graph G.*

Proof. The probability of accepting an edge in a bin is:

$$\begin{aligned}
P(\text{edge in } \mathcal{B}^{BCL}_{k_1 k_2}) &= \sum_{e_{ij} \in \mathcal{B}^{BCL}_{k_1 k_2}} P(e_{ij} \text{ sampled}) A(k_1, k_2) \\
&= \sum_{e_{ij} \in \mathcal{B}^{BCL}_{k_1 k_2}} 2 \frac{D_i D_j}{(2M)(2M)} \frac{R(k_1,k_2)}{R_{max}}
\end{aligned}$$

Summing over quantiles:

$$\begin{aligned}
&P(\text{edge in } \mathcal{B}^{BCL}_{k_1 k_2}) \\
&= \frac{2}{R_{max}} \left(\sum_{i \in Qu(k_1)} \frac{D_i}{2M} \right) \left(\sum_{j \in Qu(k_2)} \frac{D_j}{2M} \right) \frac{\mathcal{B}_{k_1 k_2}}{\mathcal{B}'_{k_1 k_2}}
\end{aligned}$$

Chung Lu models have uniform bin distribution so $\mathcal{B}'_{k_1 k_2} = \frac{2M}{k^2}$ and further by definition, the sum of degrees are uniformly distribution among the k quantiles:

$$\begin{aligned}
P(\text{edge in } \mathcal{B}^{BCL}_{k_1 k_2}) &= \frac{2\mathcal{B}_{k_1,k_2}}{R_{max}} \frac{1}{k} \frac{1}{k} \frac{k^2}{2M} \\
&= \frac{\mathcal{B}_{k_1,k_2}}{R_{max} M}
\end{aligned} \tag{2}$$

The overall acceptance probability is $\frac{1}{R_{max}}$ from Lemma 1, meaning the expected number of draws to insert a single edge is R_{max}. Therefore, the expected number of total draws from the underlying CL model is MR_{max}. Thus, combined with Equation 2, the expected number of edges in $\mathcal{B}^{BCL}_{k_1,k_2}$ is equal to $\mathcal{B}_{k_1,k_2}$.

□

Thus, the resulting graph samples will have edges drawn from the coarse joint degree distribution representation that parameterizes the original network G. By extension, the assortativity (which is simply a statistic of the joint degree distribution) of the generated networks G^{BCL} will approximately match that of G. Additionally, our BCL method preserves the modeled expected degree distribution.

Theorem 3. *For a graph G^{BCL} generated by BCL, the expected degree of a node i is D_i, the degree of the node in the original graph G.*

Proof. First, we derive the probability of accepting a sampled edge incident to node i (within bin $b(D_i)$) for a given node i:

$$
\begin{aligned}
&P(\text{edge incident to } i \text{ accepted}) \\
&\quad = \sum_j P(e_{ij}) P(\text{accepted} | e_{ij}) \\
&\quad = \sum_{k' \in [1..k]} \sum_{j \in Qu(k')} P(e_{ij}) A(b(D_i), k') \\
&\quad = \sum_{k' \in [1..k]} \sum_{j \in Qu(k')} 2 \frac{D_i D_j}{(2M)(2M)} \frac{R(b(D_i), k')}{R_{max}}
\end{aligned}
$$

Summing over the quantile $B_{k'}$ of node j:

$$
\begin{aligned}
&P(\text{edge incident to } i \text{ accepted}) \\
&\quad = \frac{D_i}{R_{max} M} \sum_{k' \in [1..k]} \left(\sum_{j \in Qu(k')} \frac{D_j}{2M} \right) \frac{\mathcal{B}_{b(D_i),k'}}{\mathcal{B}'_{b(D_i),k'}} \\
&\quad = \frac{D_i}{R_{max} M} \frac{1}{k} \frac{k^2}{2M} \sum_{k' \in [1..k]} \mathcal{B}_{b(D_i),k'}
\end{aligned}
$$

Because of the definition of quantiles, the marginal bin frequencies are uniform: $\frac{2M}{k}$

$$
\begin{aligned}
P(\text{edge incident to } i \text{ accepted}) &= \frac{D_i}{R_{max} M} \frac{1}{k} \frac{k^2}{2M} \frac{2M}{k} \\
&= \frac{D_i}{R_{max} M}
\end{aligned}
$$

As in the proof of Theorem 2, the expected number of total draws from the underlying CL model is is MR_{max}. Therefore, the expected number of edges incident to node i, and thus the degree of node i, is D_i. □

Runtime Complexity

For a given CL algorithm, let the sampling complexity be $O(M \cdot \lambda)$, where λ is strictly sublinear. For a given k number of bins, the learning step must first (a) construct the degree vector and (b) sort the degree vector. Next, it must (c) construct the k^2 bins, (d) sample a network from the model, and (e) insert all the edges into the appropriate bin. Thus, the total cost is $O(M \log M + k^2 + M \cdot \lambda)$. If the chosen model is scalable and k is a constant, the total learning time is subquadratic. To sample a network, each accept-reject lookup costs $O(1)$, and we have a geometric acceptance distribution. Since the mean of a geometric distribution is a constant, the expected total sampling cost remains $O(M \cdot \lambda)$.

5 Experiments

Experiments were performed to assess the algorithms and to assess how varying the amount of binning effects error rates. To empirically evaluate the models, we learned model parameters from real-world graphs and then generated new graphs using those parameters. We then compared the network statistics of the generated graphs with those of the original networks.

Datasets

We used six different datasets to evaluate our experimental results. Their node and edge counts can be found in Figure 1, with all networks being cast into an undirected and unweighted representation.

First, we study two email datasets: a small publicly available dataset from an email network of students at University Rovira i Virgili in Tarragona (RoviraEmail) (Guimer et al. 2003), and a large email network from Purdue University (Email). Each dataset is a collection of SMTP logs representing when users send an email to one another, with every email sent representing a link between instances.

The next two networks we study are examples of social networks, with a collection of Facebook wall postings (FacebookWall) and the publicly available Epinions trust network (Epinions) (Richardson, Agrawal, and Domingos 2003).

Another dataset is Gnutella, a publicly available Peer2Peer network where users are attempting to find seeds for file sharing (Ripeanu, Iamnitchi, and Foster 2002). In a Peer2Peer network, a user queries its peers to determine if they can seed a file. If not, the peer refers them to other users who might have a file. This repeats until a seed is found.

Lastly, we study a publicly available citation network of US Patents (Leskovec, Kleinberg, and Faloutsos 2005). Nodes in this network are published patents, while edges indicate where one patent cited the other. This is a large network, with over 10 million citations between 2 million edges and demonstrates the scalability of our proposed methods.

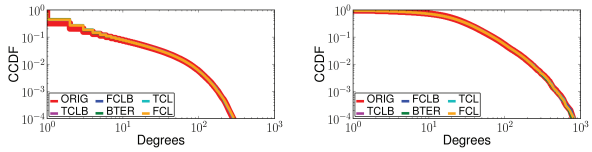
Models Compared

We compare our proposed methods against three baselines: FCL, TCL and the Block Two-Level Erdos-Renyi (BTER) model[1] (Kolda et al. 2013). The BTER model groups vertices with similar degrees into blocks with high probability, resulting in networks with a high amount of clustering and positive assortativity. As a result, BTER cannot model networks where the assortativity is independent of the clustering, meaning augmenting BTER with our Binning Chung Lu method would interfere with the clustering that BTER models. In contrast, the degree and clustering statistics of FCL and TCL are independent from the assortativity. Thus, we implement our augmentation to both the FCL and TCL models, creating the FCLB and TCLB methods. We demonstrate how, in particular, TCLB can jointly capture the degree distribution, clustering, joint degree distributions and assortativity, in contrast to any of the baseline methods.

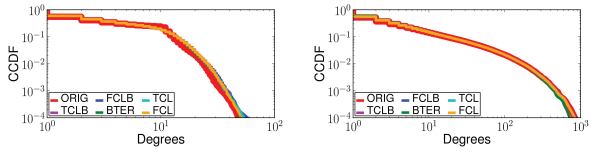
Methodology

We ran experiments on six real world datasets using five different algorithms. For evaluation, we compared the graphs generated by the algorithms using the complementary cumulative distribution function for both the degree distribution and the distribution of local clustering coefficients. We also compared the assortativity coefficient and the joint degree distributions visually. To compare the distributions, we choose a binning number of 10 bins and plot the original and generated graphs' joint degree distribution.

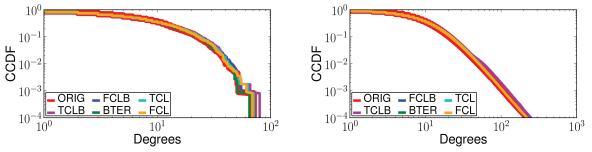
[1] Downloaded from $www.sandia.gov/\ tgkolda/feastpack$



(a) Facebook Wall (b) Purdue Email

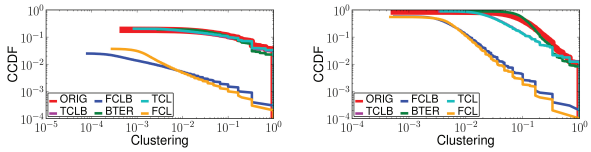


(c) Gnutella (d) Epinions

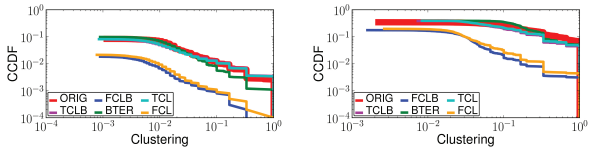


(e) Rovira (f) Patents

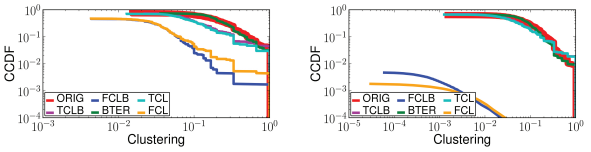
Figure 4: Degree Distributions



(a) Facebook Wall (b) Purdue Email



(c) Gnutella (d) Epinions



(e) Rovira (f) Patents

Figure 5: Clustering Coefficients

Results

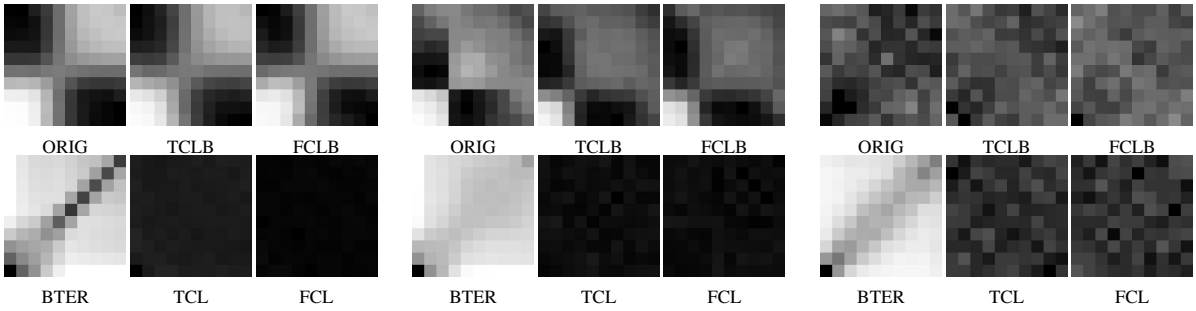
To begin, Figure 4 demonstrates that all the compared methods closely model the degree distributions of the datasets. In the next figure, Figure 5, we demonstrate that only TCL, TCLB and BTER preserve the local clustering coefficients found in the original network. As expected, the FCL method fails to model the clustering found in the datasets; this is reflected in FCLB, which only captures the assortativity of the FCL model provided to it. Thus, the binning models (FCLB and TCLB) reflect the underlying proposal distributions (FCL and TCL) and preserve the corresponding statistics that each model .

In Figure 6, we plot the joint degree distributions for all six of our datasets. FCLB and TCLB are able to capture not only the correct assortativity coefficient, but also accurately model the joint degree distribution. In contrast, BTER creates a joint degree distribution with a linear degree correlation.

Correspondingly, we present the assortativity of each of the models for each of the datasets in Figure 7.a. In particular, the non-binning Chung Lu models have assortativity very close to zero. However, FCLB and TCLB closely match the assortativity for all datasets. Although BTER exhibits positive assortativity, it doesn't model the assortativity found in the corresponding real world networks.

In order to test the impact of the 10 bin selection, we compare *error* rates for FCLB and TCLB on the Gnutella dataset as we vary the bin size (Figure 7.b-c) The error we use is the *skew*-divergence to measure the difference between the model distribution and the original data distribution (Lee 2001). The skew-divergence is used as it has a slight mixture between the two distributions, meaning that there is always support between the measures (unlike KL Divergence). We generate 25 different graphs and take the mean and standard deviations (bars) of error rates at various points, measuring the error for the degree distribution, joint degree distribution and clustering coefficients. In addition to the binning methods (solid lines, plotted by means), we give the original models' values for each statistic (dashed lines). First, we see that the degree distribution for the binning models are not significantly different from the original models until a relatively large number of bins are used (50 or more). However, the joint degree distribution quickly improves over the baseline models, with considerably less error when only 5 bins are used. Additionally, the clustering coefficient distribution is not significantly different from the original model. Hence, our models are largely stable over a variety of binning choices, maintaining the original degree distribution and clustering of a given model and additionally modeling joint degree distribution.

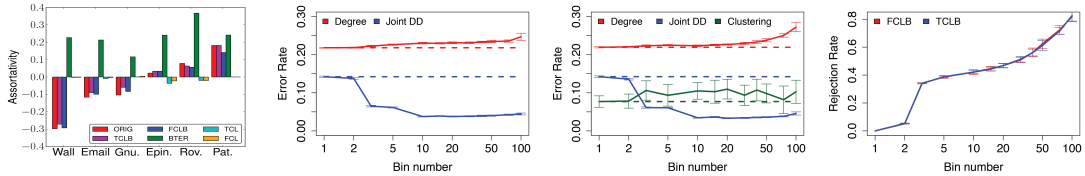
Lastly, we plot the rejection rates for the binning models, as we vary the number of bins used, in Figure 7.d. Note that the original Chung Lu models sample from the edge distribution M times. Let $\alpha_{rejection}$ be the rejection rate. Since the acceptance of an edge has a geometric distribution, we can conclude that the binning version samples from the edge distribution $\frac{M}{1-\alpha_{rejection}}$ on average. Thus, even with large amounts of rejection for the large bins (100), the binned version is only 5x longer than the original model. For smaller bin sizes, which also accurately model the joint degree distribution, the rejection rate is considerably less (around 2x rejection rate). Thus, with a constant factor of extra samples we are able to accurately model the joint degree distribution, while maintaining the degree distribution and clustering. Empirically, the Patents dataset containing 14 million edges ran in under 20 minutes with only 10 bins.



(a) Facebook Wall (c) Gnutella (e) Rovira

(b) Purdue Email (d) Epinions (f) Patents

Figure 6: Visualization of joint degree distribution representations $\mathcal{B}$; $k = 10$.



(a) Assortativity (b) FCLB Error Rates (c) TCLB Error Rates (d) Rejection Rates

Figure 7: (a) shows assortativity for the datasets and algorithms. The effect of varying bin sizes for (b) FCLB and (c) TCLB. (d) illustrates the rejection rates for each.

6 Conclusions

In this work, we presented the BCL extension to the Chung Lu family of models. We began with a discussion of the assortativity statistic commonly used in social network analysis, and demonstrated that existing random network models are unable to model it. We next discussed that the assortativity is only a linear summary statistic of the joint degree distribution, demonstrating that real world networks with similar assortativity coefficients but largely different joint degree distributions exist. Further, we proved that there are an infinite number of graphs with the same degree distribution and assortativity, yet different joint degree distributions. This motivated our solution: the coarse binning technique to capture the joint degree distribution of a network. Using accept reject sampling, this binning method can pair with any existing Chung Lu model to preserve the assortativity, degree distribution and clustering, without increasing the runtime complexity.

We implemented our method with both the FCL and TCL models, with the corresponding TCLB capturing the degree distribution, clustering, assortativity and joint degree distribution better than any baselines. We tested this implementation on six large, real world networks, including a Patents dataset with nearly 14 million edges. Our BCL representation empirically preserves an underlying model's statistics while incorporating the true joint degree distribution. Future work includes expanding our methods to scalably model additional graph statistics, all while provably maintaining guarantees of current network models.

Acknowledgements

This research is supported by NSF under contract numbers IIS-1149789, CCF-0939370 and IIS-1219015, and the undergraduate research *Channels Scholars Program*, from the Center for the Science of Information (CCF-0939370).

Appendix

This section provides the detailed proof for Theorem 1.The proof states there are infinitely many pairs of graphs (name A and B) with the following three conditions:

- **Lemma 2**: Graphs A, B have equal degree distributions.

- **Lemma 3**: Graphs A and B have the same assortativity.
- **Proposition 1**: The joint degree distributions of graphs A and B have infinite KL divergence.

Graphs A and B are defined in terms of a positive integer w, where $w \geq 2$. In particular, A and B are defined solely in terms of stars and cliques that relate to w:

Graph A	Graph B
N_S w-stars	N_{2s} 2w-stars
N_{2c} (2w+1)-cliques	N_C (w+1)-cliques
	N_p Pairs

In particular, given the same $w \geq 2$, the graphs A and B have different joint degree distributions.

Proposition 1. *For any $w \geq 2$, the joint degree distributions of Graphs A and B have infinite KL-Divergence.*

Proof. Note that Graph A consists solely of links between nodes of degree 1 to w, and $2w$ to $2w$. In contrast, Graph B consists solely of links between nodes of degree $2w$-1, w to w, and 1 to 1. These sets are disjoint, meaning neither graph has full (or any) support of the opposite graph. Thus, the joint degree distributions have infinite KL-Divergence. □

Next, we define values for the parameters of A and B that result in the same degree distributions between the two networks (computation omitted for space).

Lemma 2. *For any $w \geq 2$, if $N_S = 2(w+1)(2w+1)(2w-1)^2$ and $N_{2s} = (w+1)(2w+1)(w-1)^2$ and $N_C = 2(2w+1)(2w-1)^2$ and $N_{2c} = (w+1)(w-1)^2$ and $N_p = w^2(w+1)(2w+1)(3w-2)$, then graphs A and B have identical degree distributions.*

For an undirected graph, the first moments of these two distributions are identical.

Definition 1. *For a bidirectional graph G, let:*

$$\mu_G = \frac{\sum_{e \in \mathbf{E}} T_*(e)}{|\mathbf{E}|} = \frac{\sum_{e_{ij} \in \mathbf{E}} D_i}{|\mathbf{E}|}$$

$$\sigma_G^2 = \frac{\sum_{e \in \mathbf{E}} (T_*(e) - \mu_G)^2}{|\mathbf{E}|} = \frac{\sum_{e_{ij} \in \mathbf{E}} (D_i - \mu_G)^2}{|\mathbf{E}|}$$

where $* \in \{\beta, \epsilon\}$ This occurs due to the symmetry of a bidirectional network: for every $e_{ij} \in \mathbf{E}$ there is a corresponding $e_{ji} \in \mathbf{E}$.

Lastly, although the variances of $T_\beta(e)$ and $T_\epsilon(e)$ are equal, the covariance between them is not equal to the variance. We use this to define the assortativity.

Definition 2. *The assortativity of a network is defined as the covariance of two variables divided by the standard deviation of each. Thus,*

$$\mathcal{A}_G = \frac{cov(\{T_\beta(e), T_\epsilon(e) | e \in \mathbf{E}\})}{\sigma_G \cdot \sigma_G} = \frac{\sum_{e_{ij} \in \mathbf{E}} (D_i - \mu_G)(D_j - \mu_G)}{\sigma_G^2}$$

With these defined, the next lemma proves values for graphs A and B are equal (computation omitted for space).

Lemma 3. *For any $w \geq 2$, if N_S, N_{2s}, N_C, N_{2c}, and N_p are defined as in Lemma 2, then graphs A and B have identical assortativity values.*

We now combine Lemmas 1 and 2 (matching Degree Distribution and Assorativity) with Proposition 1 (infinite KL divergence in the joint degree distribution).

Theorem 1. *Let A_w, B_w be two networks comprised of graphlets, where w parameterizes the size and counts of the graphlets. There exist an infinite set of pairs of graphs $\{A_w, B_w\}$ such that for any $w \geq 2$, A_w and B_w have the same degree distribution, the same assortativity, but infinite KL-Divergence between their joint degree distributions.*

Proof. We construct $\{A_w, B_w\}$ by using the equations from Lemma 2 and w to find N_S, N_{2s}, N_C, N_{2c}, and N_p. The value of these five constants can be used to construct two graphs, A_w and B_w, as in Proposition 1.

By Lemma 2, A_w and B_w have the same degree distribution. By Lemma 3, A_w and B_w have the same assortativity. Finally, by Proposition 1, A_w and B_w have infinite KL-divergence. □

References

Chung, F., and Lu, L. 2002. The average distances in random graphs with given expected degrees. *Internet Mathematics* 1.

Guimer, R.; Danon, L.; Diaz-Guilera, A.; Giralt, F.; and Arenas, A. 2003. Self-similar community structure in a network of human interactions. *Phys. Rev. E* 68(6):065103.

Kolda, T. G.; Pinar, A.; Plantenga, T.; and Seshadhri, C. 2013. A scalable generative graph model with community structure. arXiv:1302.6636. revised March 2013.

Lee, L. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, 65–72.

Leskovec, J.; Chakrabarti, D.; Kleinberg, J.; Faloutsos, C.; and Ghahramani, Z. 2010. Kronecker graphs: An approach to modeling networks. *J. Mach. Learn. Res.*

Leskovec, J.; Kleinberg, J.; and Faloutsos, C. 2005. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *In KDD*, 177–187. ACM Press.

Liang, F.; Liu, C.; and Carrol, R. J. 2010. *Advanced Markov chain Monte Carlo methods: learning from past samples*. Wiley Series in Computational Statistics. New York, NY: Wiley.

Newman, M. E. J. 2002. Assortative mixing in networks. *Physical Review Letters* 89(20):208701.

Pfeiffer III, J. J.; La Fond, T.; Moreno, S.; and Neville, J. 2012. Fast generation of large scale social networks while incorporating transitive closures. In *Fourth ASE/IEEE International Conference on Social Computing (SocialCom)*.

Pfeiffer III, J. J.; Moreno, S.; La Fond, T.; Neville, J.; and Gallagher, B. 2014. Attributed graph models: Modeling network structure with correlated attributes. In *Proceedings of the 23rd International World Wide Web Conference (WWW 2014)*.

Pinar, A.; Seshadhri, C.; and Kolda, T. G. 2011. The similarity between stochastic kronecker and chung-lu graph models. *CoRR* abs/1110.4925.

Richardson, M.; Agrawal, R.; and Domingos, P. 2003. Trust management for the semantic web. In *Proceedings of the Second International Semantic Web Conference*, 351–368.

Ripeanu, M.; Iamnitchi, A.; and Foster, I. 2002. Mapping the gnutella network. *IEEE Internet Computing* 6(1):50–57.

Stanton, I., and Pinar, A. 2011. Constructing and sampling graphs with a prescribed joint degree distribution. *CoRR* abs/1103.4875.

Zhou, D.; Stanley, H. E.; D'Agostino, G.; and Scala, A. 2012. Assortativity decreases the robustness of interdependent networks. *Phys. Rev. E* 86:066103.