DEGREE DISTRIBUTIONS AND RANDOM GRAPHS

EDDIE MALDONADO AND ROY OURSLER

DR. GLENCORA BORRADAILE

ABSTRACT. We analyze existing random graph models and the structure of the graphs that they produce. Using a graph decomposition we call a box decomposition, we compare random graphs to datasets representing actual networks. We demonstrate that existing random graph models fail to capture properties found by the box decomposition in real data sets. We propose a new graph model, the "density column" model, and show that the new model generates random graphs with properties more similar to those of real-world networks.

1. INTRODUCTION

In this work, we use the box decomposition (section 1.2), originally studied in [BIM+12], to analyze the structure of simple undirected graphs: the Barabási-Albert random graph model and actual data sets (section 3). We propose and analyze a random graph model which better resembles real-world data in section 2.2.

1.1. **Notation.**

Let G be a directed graph where $V(G)$ denotes the set of vertices of G and $E(G)$ denotes the set of edges. If $v \in V(G)$, then $\delta(v)$ denotes the total degree of v, $\delta^+(v)$ denotes the indegree of v, and $\delta^-(v)$ is the outdegree of v.

1.2. **Box Decomposition for a Directed Graph.**

Let G be a directed graph, and let k be the maximum indegree of G.

Definition 1.1. *Let B_k, also called the* top box*, be the set of all vertices of indegree k, along with all vertices which have a directed path to a vertex of indegree k. For all $\ell < k$, we define:*

$$B_\ell = \left\{ v \in V(G) \setminus \bigcup_{i=\ell+1}^{k} B_i \mid v \text{ has indegree } \ell \text{ or has a directed path to a vertex of indegree } \ell \right\}$$

Definition 1.2. *The* rank *of a graph G, written* $\mathrm{Rank}(G)$*, is* $\max_{v \in V(G)}\{\delta^+(v)\}$.

For an arbitrary directed graph, this decomposition does not yield much information, but for directed graphs with certain properties it does. One such property is used to define the box decomposition on an undirected graph.

Date: 16 August 2012.

This work was done during the Summer 2012 REU program in Mathematics at Oregon State University.

1.3. **Box Decomposition for an Undirected Graph.** The box decomposition for an undirected graph G is the box decomposition of a directed version of G whose indegrees have minimum lexicographic order.

Definition 1.3. *A directed graph H has minimum lexicographic order if $\delta^+(v)$ for all $v \in V(H)$, sorted from highest to lowest, is lexicographically less than or equal to $\delta^+(h)$ of all vertices $h \in V(H')$, sorted from highest to lowest, where $H' = H$ after a set of edges is reversed.*

To find such an orientation, we used an algorithm called PATH-REVERSAL [BIM+12]. The process goes as follows: suppose there is a procedure called ORIENT that takes an undirected graph and orients the edges arbitrarily. A directed path from u to v in G is called *reversible* if $\delta^+(u) < \delta^+(v) - 1$. FIND-REVERSIBLE-PATH returns a reversible path. The pseudocode of the algorithm itself follows:

PATH-REVERSAL(G)

```
1  ORIENT(G) ▷ Assume G starts undirected
2  while HAS-REVERSIBLE-PATH(G)
3      do p ← FIND-REVERSIBLE-PATH(G)
4         REVERSE-PATH(p)
```

When a directed graph has minimum lexicographic order, the box decomposition has been shown to have the following properties [BIM+12]:

(1) The top box contains the densest subgraph.
(2) All edges between vertices in different boxes will be directed from a box B_j to B_i where $j > i$.
(3) The box decomposition is the same for any orientation with minimum lexicographic order.

Since the box decomposition is the same for any graph orientation of minimum lexicographic order, the box decomposition for an undirected graph is unique. This decomposition can then be used to obtain information about the structure of graphs. We used this decomposition to analyze the structure of some random graph models.

1.4. **Random Graph Models.** Generating simple graphs randomly is a problem that has appeared in numerous situations. One of the first models of random graphs that was studied is the random $G(n,p)$ graph model, also known as the Erdős-Rényi model, after its creators, Paul Erdős and Alfréd Rényi [ER59]. In a graph on n vertices, there are $\binom{n}{2}$ possible edges. The $G(n,p)$ model simply adds each of these possible edges to the graph with probability p. This model has significant limitations for mimicking real world graphs because many graph properties and structures which arise in real life do not commonly appear in $G(n,p)$ graphs, such as

Definition 1.4. *The degree distribution of a graph G is a probability mass function $f(x)$ where $f(x) = \dfrac{\delta(x)}{\sum_{v \in G} \delta(v)}$ for $x \in G$.*

The expected degree distribution of a graph generated with the $G(n,p)$ graph model is binomial, however many real-world graphs do not exhibit that degree distribution. For this reason, the random $G(n,p)$ is not a useful model of real-world networks.

To overcome this limitation, researchers have proposed other random graph models. One these is the Barabási-Albert model, which generates graphs which have an expected power law distribution

— similar to many graphs that arise in the real world [AB02]. A random graph is generated in this model by starting with a seed graph on m_0 vertices. Further vertices are added iteratively with each new vertex attached to m existing vertices chosen with a probability that is proportional to the degree of the existing vertices. This is also known as a preferential attachment model since vertices with higher degrees will be attached to more often than vertices with lower degrees.

The Barabási-Albert model has been used to model many types of networks, such as social networks [BDML06]. The reason the Barabási-Albert model is used in social networks is that the more people someone knows, the more likely it is that they will meet someone new. There are many networks with this property making this model useful for modeling real-world data.

Between the $G(n,p)$ graph model and the Barabási-Albert model, there are only two degree distributions, the binomial and power law distributions, which can be modeled by randomly generated graphs. This points out a major limitation in those models: arbitrary degree distribution. Some researchers have suggested models which will generate graphs with an arbitrary degree distribution. A survey of models used to generate graphs with arbitrary degree distributions by Britton, Deijfen, and Martin-Löf found four basic types of random graph models to generate a given degree distribution [BDML06]. One of these is a generalization of the random $G(n,p)$ model. We propose a modification of the generalized $G(n,p)$ model, the iterated $G(n,p)$ discussed in section 2.1, which converges to the given distribution faster than the model give by Britton et al.

Another model for generating random graphs is the probabilistic planted model, another generalization of the $G(n,p)$ [Rou10, p. 6-8]. The planted model partitions vertices in a graph into separate sets S_i. There is then a set of probabilities P where each element p_{ij} represents the probability of an edge existing between a vertex in S_i and a vertex in S_j. By doing this, each set S_i induces a $G(n,p)$ model subgraph and these graphs are then connected to generate a graph with a more complicated degree distribution. We analyze a specific instance of this model, the density column model (section 2.2), as a method to generate random graphs with a given box decomposition.

We analyze the random graph models mentioned above for the box decompositions found in those graphs. Along with this, we analyze actual data sets and compare their box decompositions to those of the random graphs to see if the box decomposition shows a structural difference between the random graphs and the actual data sets. We find that the density column model that is analyzed best reproduces the structures found by the box decomposition.

2. Results

2.1. **Iterated $G(n,p)$ Graph.** Let P be a matrix used to calculate the iterated $G(n,p)$ graph defined below in 2.1.1. Each element p_{ij} of P is the probability the edge between vertices i and j exists.

2.1.1. *Defining P.* Let δ_0 be a vector with n elements representing the desired degree distribution of vertices in the random graph with the distribution satisfying $\frac{\delta_{0_j}^2}{\sum_{i=0}^{k}\delta_{0_i}} \leq \frac{1}{2}$. Let S_i be a sequence of matrices defined below. Let δ_i be a vector defined by the diagonal of S_i where the element

$\delta_{i_k} = S_{i_{kk}}$. Now the sequence S_i is inductively defined by

$$\begin{aligned} S_1 &= \frac{\delta_0 \delta_0^T}{\mathbf{1} \cdot \delta} \\ S_2 &= \frac{\delta_1 \delta_1^T}{\mathbf{1} \cdot \delta_1} = \frac{\delta_1 \delta_1^T}{\mathrm{tr}(S_1)} \\ &\vdots \\ S_k &= \frac{\delta_{k-1} \delta_{k-1}^T}{\mathbf{1} \cdot \delta_{k-1}} = \frac{\delta_{k-1} \delta_{k-1}^T}{\mathrm{tr}(S_{k-1})} \\ &\vdots \end{aligned}$$

Then for every p_{ij} in P

$$p_{ij} = \begin{cases} \sum_{k=1}^{\infty} S_{n_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

2.1.2. *Properties of P.* The matrix P has properties that allow it to produce a distribution that is the original distribution: the probabilities should be symmetric ($p_{ij} = p_{ji}$), $p_{ij} < 1$ for all i and j, and should give the desired degree distribution ($\delta_{0_j} = \sum_{i=0}^{k} p_{ij}$).
To show that $p_{ij} = p_{ji}$, we note that S_i is a product of a vector with its transpose and so is semidefinite and therefore symmetric. Since P is just a sum of these matrices, P is symmetric.

To show the last two properties note $S_{i_{jj}} = \sum_{k=0}^{n} S_{i+1_{jk}}$:

$$S_{i_{jj}} = \delta_{i_j} = \delta_{i_j} \frac{\sum_{k=0}^{n} \delta_{i_k}}{\sum_{m=0}^{n} \delta_{i_m}} = \sum_{k=0}^{n} \frac{\delta_{i_j} \delta_{i_k}}{\sum_{m=0}^{n} \delta_{i_m}} = \sum_{k=0}^{n} S_{i+1_{kj}}$$

To show that $p_{ij} \leq 1$, $S_{i_{jj}} = \sum_{k=0}^{n} S_{i+1_{jk}}$ implies that $p_{ij} \leq S_{k_{ij}} + S_{k_{jj}}$. For $k = 0$, $S_{0_{ij}} < \frac{1}{2}$ and $S_{0_{jj}} < \frac{1}{2}$ because of the limitations on δ_0. So $p_{ij} \leq \frac{1}{2} + \frac{1}{2} = 1$.
To show that $\sum_{i=0}^{n} p_{ij} = \delta_{0_j}$:

$$\begin{aligned} \delta_{0_j} &= \sum_{i=0}^{n} S_{1_{ij}} = \sum_{i \neq j}^{n} S_{1_{ij}} + S_{1_{ii}} = \sum_{i \neq j}^{n} S_{1_{ij}} + \sum_{i \neq j}^{n} S_{2_{ij}} \\ &= \sum_{i \neq j}^{n} (S_{1_{ij}} + S_{2_{ij}}) + S_{2_{ii}} = \sum_{i \neq j}^{n} \sum_{k=1}^{\infty} S_{k_{ij}} = \sum_{i=0}^{n} p_{ij} = \end{aligned}$$

2.1.3. *Convergence of P.*

Theorem 2.1. *P converges at least as fast as* $\frac{1}{2^{2^k}}$.

Proof. We start by showing

$$\frac{S_{k_{jj}}}{\sum_{i=0}^{n} \delta_{k_i}} < \frac{1}{2^{2^k}} \tag{1}$$

Base Case:

$$\frac{S_{1_{jj}}}{\sum_{i=0}^{n}\delta_{k_i}} = \frac{S_{1_{jj}}}{\sum_{i=0}^{n} S_{1_{ii}}} = \frac{\frac{\delta_{0_j}^2}{\sum_{k=0}^{n}\delta_{0_k}^2}}{\sum_{i=0}^{n}\frac{\delta_{0_i}^2}{\sum_{k=0}^{n}\delta_{0_k}^2}} = \frac{\delta_{0_j}^2}{\sum_{i=0}^{n}\delta_{0_i}^2} < \frac{1}{2}$$

because this distribution is achievable by a simple graph

Inductive Step:

$$\frac{1}{2^{2^k}} > \frac{S_{k_{jj}}}{\sum_{i=0}^{n}\delta_{k_i}} \implies \frac{1}{2^{2^{k+1}}} > \frac{S_{k_{jj}}^2}{(\sum_{i=0}^{n}\delta_{k_i})^2} = \frac{S_{k_{jj}}^2}{\sum_{i=0}^{n}\delta_{k_i}^2}\frac{\sum_{i=0}^{n}\delta_{k_i}^2}{(\sum_{i=0}^{n}\delta_{k_i})^2}$$

$$\geq \frac{S_{k_{jj}}^2}{(\sum_{i=0}^{n}\delta_{k_i})^2} = \frac{S_{k+1_{jj}}^2}{(\sum_{i=0}^{n}\delta_{k+1_i})^2}$$

This proves equation 1. Consider the sum of all the elements of a row.

$$\delta_{0_j} = \sum_{i=0}^{n} p_{ij} = \sum_{i\neq j}^{n}\sum_{m=0}^{k} S_{k_{ij}} + S_{k_{jj}}$$

Therefore $S_{k_{jj}}$ is the error between δ_{0_j} and $\sum_{i\neq j}^{n}\sum_{m=0}^{k} S_{k_{ij}}$. So:

$$S_{k_{jj}} = S_{k-1_{jj}}\frac{S_{k-1_{jj}}}{\sum_{i=0}^{n} S_{k-1_{ii}}} < \frac{1}{2^{2^{k-1}}}$$

□

2.2. **The Density Column Model.** The probabilistic planted model suggests many variations on the basic $G(n,p)$ idea. Here we will present a special case of the planted model that can produce graphs with a variety of degree distributions and box decompositions by choosing the input parameters appropriately. Suppose that we are given a sequence of m ordered pairs of the form (n_i,k_i). Each of these indicates that we want n_i vertices of expected degree k_i in our final graph. The idea is that, if we can generate such a graph, then appropriate choices of n and k values can be used to obtain varying degree distributions with the same model. We wish to find probabilities $p_1, p_2, \ldots, p_m$ such that we can perform the following procedure and get a graph with the desired properties:

(1) Create a graph on $N = \sum_{i=1}^{m} n_i$ vertices with no edges.
(2) Begin on layer $i = 1$.
(3) For each vertex in layer i , join it with an edge to each vertex of layer i or higher with probability p_i.
(4) Increment i and repeat the process until the top layer is reached.

A little bit of background is in order at this point. This model began as an attempt to generate a graph with a desired box decomposition, motivated by the following conjecture:

Conjecture 2.2. *Let G be a $G(n,p)$ random graph that has been oriented by path reversal. Then $\mathbb{E}(\text{Rank}(G)) = np/2 + o(1)$ when $p = o(1)$.*

If this conjecture holds, we can let the desired lower boxes be the lower layers in this model. We can use the expected rank, plus the fact that edges between boxes must always point to the lower one, to get:

$$\ell = \frac{n_\ell p_\ell}{2} + p_\ell \sum_{i=\ell+1}^{k} n_i$$

and then solve for p_ℓ to get

$$p_\ell = \frac{2\ell}{n_\ell + 2\sum_{i=\ell+1}^{k} n_i}$$

in order to give the desired box decomposition. This model has promise, yet the conjecture remains unproven, however, given that we know how to find the degree distributions of $G(n,p)$ graphs and generalizations of them, this method can be used to generate graphs of different degree distributions.

We can now return to the newer model and solve for p_ℓ. Note that the existence of each possible edge in the graph is simply a Bernoulli trial, with the probability of success defined by the p_ℓ values and our algorithm. Keeping the Bernoulli trial nature of the events in mind and thinking in terms of how the algorithm builds the graph, we can work backwards and compute the expected degree of the vertices of a particular layer:

$$k_\ell = p_\ell \left(\sum_{i=\ell}^{m} n_i - 1 \right) + \sum_{i=1}^{\ell-1} p_i n_i$$

Now, we just solve for p_ℓ to get:

$$p_\ell = \left(k_\ell - \sum_{i=1}^{\ell-1} p_i n_i \right) \Bigg/ \left(\sum_{i=\ell}^{m} n_i - 1 \right) \tag{2}$$

Note that p_ℓ depends on $p_1, p_2, \ldots, p_{\ell-1}$, so this method allows us to find the needed probabilities recursively.

Now we wish to find the overall distribution of the resulting random graph. To do so, we can employ the law of total probability.

$$\mathbb{P}(\delta(v) = x) = \sum_{i=1}^{m} \mathbb{P}(v \text{ in layer } i)\mathbb{P}(\delta(v) = x \mid v \text{ in layer } i)$$

If we're wondering about an arbitrary vertex v chosen uniformly, then we have:

$$\mathbb{P}(v \text{ in layer } i) = \frac{n_i}{N}$$

Now we just need to find the conditional probability of $\delta(v)$ taking a particular value given that v is in layer i. Let $\mathbf{P}(i) \in [0,1]^{N-1}$ be a vector whose components are the probabilities of attachment between a vertex in layer i and all other vertices in the graph, not necessarily in any particular order. Let $\mathbf{P}(i)_j$ denote the j^{th} component of $\mathbf{P}(i)$. Furthermore, let:

$$S_n = \{A \in \mathcal{P}(\{0,1,\ldots,N-1\}) \mid |A| = n\}$$

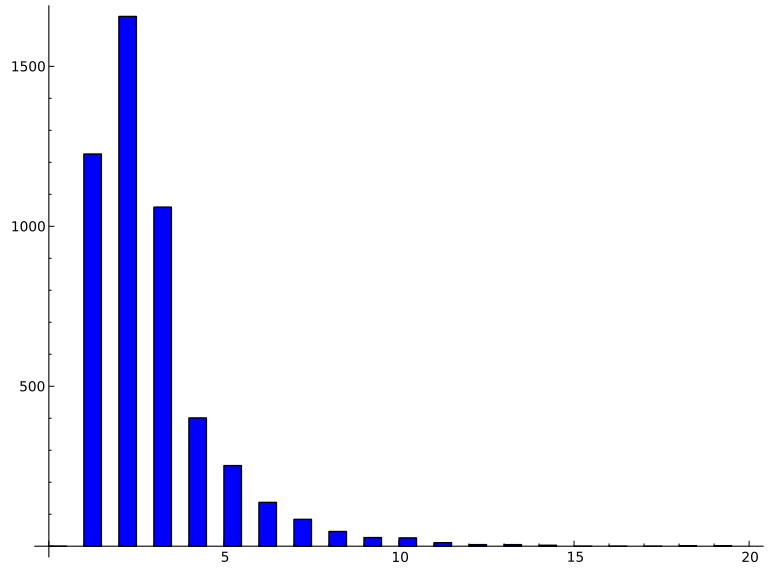

FIGURE 1. Degree Distribution of Power Grid Graph.

Then we have:

$$\mathbb{P}(\delta(v) = x \mid v \text{ in layer } i) = \sum_{A \in S_x} \prod_{a \in A} \mathbf{P}(i)_a \prod_{a \in A^C} (1 - \mathbf{P}(i)_a)$$

So, finally, this gives us:

$$\mathbb{P}(\delta(v) = x) = \sum_{i=1}^{m} \frac{n_i}{N} \sum_{A \in S_x} \prod_{a \in A} \mathbf{P}(i)_a \prod_{a \in A^C} (1 - \mathbf{P}(i)_a)$$

As necessitated by the nature of the distribution, this is not the easiest expression to work with. However, it does give us some idea of the structure of the resulting overall degree distribution.

3. ADVANTAGES OF THE DENSITY COLUMN MODEL

Our new model has practical value in terms of generating random graphs that are useful for studying real-life networks. For example, let's consider a graph representing the power grid of the western US assembled and analyzed by Watts and Strogatz in [WS98]. The graph's degree distribution (without normalization, as in all of our figures) is shown in figure 1. As it turns out, this graph has an interesting box decomposition:

i	$\lvert B_i \rvert$
1	1571
2	3313
3	57

Here we have a concrete example of a graph representing a real-world network that would be an interesting object of study. How can we generate random graphs with similar properties in order to analyze these networks further? The most popular idea currently would be to use some sort of preferential attachment scheme. The Barabási-Albert model, as previously discussed, generates

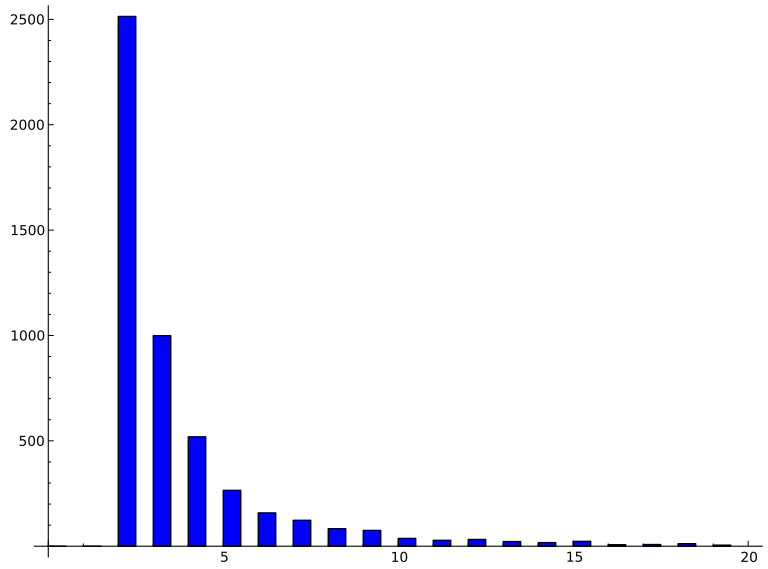

FIGURE 2. Degree Distribution of a Barabási-Albert Graph ($n = 5000$ and $m = 2$). The tail has been truncated to make the graph more legible.

graphs with a power law distribution (as seen in many datasets, including this one) and comes along with a nice story about how the model mimics the growth of real-life networks [AB02]. So, how well does this model actually simulate the data? We can get a similar size and degree distribution by creating a Barabási-Albert graph with $n = 5000$ and $m = 2$. The degree distribution of the resulting graph is shown in figure 2.

The degree distribution does, in fact, appear similar. The one notable difference is the distribution of low degree vertices — note that the actual power grid data doesn't quite obey a power law in this respect. Now, in order to gain further insight to the structure of the graph we have generated, we can look at its box decomposition:

i	$\lvert B_i \rvert$
1	0
2	5000

Clearly, we are missing something. The box decomposition captures certain elements of the graph's overall structure, so this dissimilarity in box decompositions indicates that there is a corresponding dissimilarity between the empirically-obtained graph and the random one.

Next, we can compare this to a density column graph generated with appropriate parameters. We can approximate a power law distribution with the following n and k values:

n	k
1000	2
250	4
111	6
63	8

Using equation 2, we can compute the necessary probability values and generate a corresponding random graph. The degree distribution of this graph is shown in figure 3. Note that the distribution

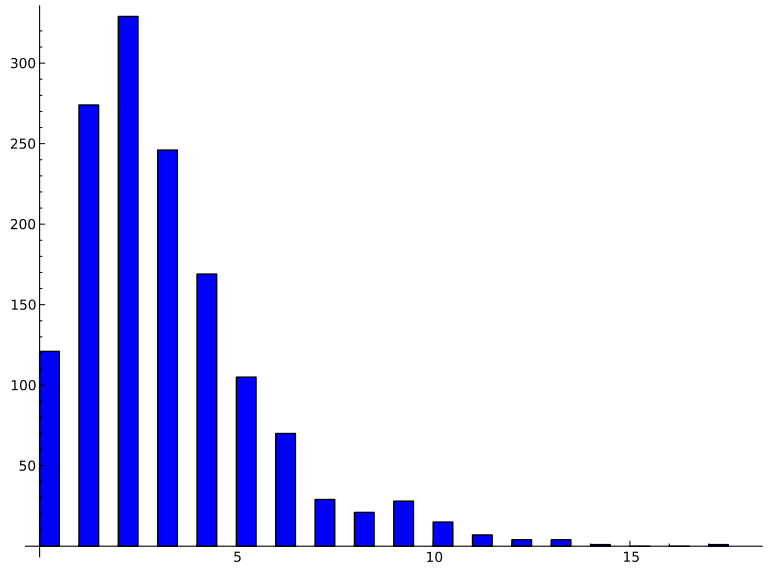

FIGURE 3. Degree Distribution of our Density Column Graph.

still falls off in a power law-like fashion, but through our choice of parameters, we were able to more accurately model the distribution of low-degree vertices.

Now, we can perform the same test as earlier with the box decomposition. For this graph, we have:

i	$\lvert B_i \rvert$
0	121
1	312
2	991

So, just like in the power grid data, we have a nontrivial box decomposition. This indicates that we have made progress in terms of replicating the structure of real-world networks.

4. CONCLUSION

We have made progress in developing interesting new random graph models that are motivated by a desire to replicate the structure of real-world networks. These models generalize the existing family of Erdős - Rényi models in interesting ways and produce properties that the basic $G(n,p)$ model typically does not produce. We were ultimately able to justify the one model that did come to fruition through demonstrating its application to modeling real-world networks.

4.1. **Further Work.**

There are numerous directions we can pursue with this problem. For one thing, a better understanding of the box decomposition is needed — it clearly says something about the structure of a graph, but a clearer understanding of what sort of structure is needed. In particular, a non-algorithmic definition of the box decomposition for an undirected graph would be useful.

Furthermore, we could develop a method which takes an arbitrary "target" degree distribution and finds appropriate parameters for the density column model to approximate that distribution.

Finally, we might be able to take insights from this work to reassess the path reversal algorithm and the original problem of finding dense subgraphs through the use of the box decomposition. While work in this particular area was not terribly fruitful in the few weeks that we had, the work that we did do together with some of the other work suggested in this section might open the door to further progress.

REFERENCES

[AB02] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002.

[BDML06] Tom Britton, Maria Deijfen, and Anders Martin-Löf. Generating simple random graphs with prescribed degree distribution. *Journal of Statistical Physics*, 124:1377–1397, 2006. 10.1007/s10955-006-9168-x.

[BIM$^+$12] Glencora Borradaile, Jennifer Iglesias, Theresa Migler, Antonio Ochoa, Gordon Wilfong, and Lisa Zhang. Egalitarian graph orientations. Manuscript, 2012.

[CL02] Fan Chung and Linyuan Lu. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, (6):125–145, 2002.

[ER59] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[NSW01] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64:026118, Jul 2001.

[NWS02] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(Suppl 1):2566–2572, 2002.

[Rou10] Tim Roughgarden. Lecture #4: Probabilistic and semirandom models for clustering and graph partitioning. Course notes for CS369N: Beyond Worst-Case Analysis, 2010.

[WS98] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, June 1998.

REED COLLEGE AND UNIVERSITY OF WYOMING

E-mail address: lmaldona@reed.edu and Roursler@uwyo.edu