Sampling Graphs with a Prescribed Joint Degree Distribution Using Markov Chains

Isabelle Stanton and Ali Pinar

UC Berkeley and Sandia National Laboratories

January 22, 2011
Definition

The *degree distribution*, $P[k]$, is the fraction of nodes in a graph with degree $k$.

Definition

The *joint degree distribution*, $P[k, l]$ is the fraction of edges in a graph that connect nodes of degree $k$ and degree $l$. 
Definition

The degree distribution, \( P[k] \), is the fraction of nodes in a graph with degree \( k \).

Definition

The joint degree distribution, \( P[k, l] \) is the fraction of edges in a graph that connect nodes of degree \( k \) and degree \( l \).

The degree vector (or sequence), \( D \), is the count of the number of nodes of degree \( k \) in a specific graph.

The joint degree matrix, \( J \), is the count of the type of edges in a specific graph.

The degree vector can be obtained from the JDM via
\[
D_k = \frac{1}{k}(J_{k,k} + \sum_{l=1}^{n} J_{k,l}).
\]
Our Work

In this paper, we investigate three related problems:

- Given a $\mathcal{J}$, can one decide if it is graphically realizable?
- If so, can one construct a graph with that $\mathcal{J}$?
- Can one sample uniformly from the space of graphs with a given graphical $\mathcal{J}$?
Related Work

- Constructing graphs with a given degree sequence - [Havel-Hakimi]

Sampling Graphs with a Prescribed Joint Degree Distribution Using Markov Chains

Isabelle Stanton and Ali Pinar

The Joint Degree Distribution

Related Work

Constructing Graphs

Uniformly Sampling with Markov Chains

Autocorrelation

Results

Conclusions
Related Work

- Constructing graphs with a given degree sequence - [Havel-Hakimi]
- Sampling Graphs with a given degree sequence using importance sampling - [Steger, Wormald] and [Bayati, Kim and Saberi]
Related Work

- Constructing graphs with a given degree sequence - [Havel-Hakimi]
- Sampling Graphs with a given degree sequence using importance sampling - [Steger, Wormald] and [Bayati, Kim and Saberi]
- Sampling graphs with a given degree sequence using Markov Chains - [Kannan, Tetali, Vempala], [Jerrum, Sinclair], [Gkantsidis, Mihail, Zegura]
Related Work

- Constructing graphs with a given degree sequence - [Havel-Hakimi]
- Sampling Graphs with a given degree sequence using importance sampling - [Steger, Wormald] and [Bayati, Kim and Saberi]
- Sampling graphs with a given degree sequence using Markov Chains - [Kannan, Tetali, Vempala], [Jerrum, Sinclair], [Gkantsidis, Mihail, Zegura]
- Prior Work on the Joint Degree Distribution - [Amanatidis, Green, Mihail], [Newman], [Mahadevan et al.]
Necessary and Sufficient Conditions

**Theorem**

Let $J$ be given and $D$ be the associated degree distribution. $J$ can be realized as a simple graph if and only if

- $D_k$ is integer-valued for all $k$
- $\forall k, l$, if $k \neq l$ then $J_{k,l} \leq D_k D_l$. Otherwise, $\forall k$ $J_{k,k} \leq \binom{D_k}{2}$.

These simple conditions are the equivalent of the Erdős-Gallai Condition for a degree sequence to be graphical. Also due to Amanatidis, Green and Mihail.

In the paper, we give a simple $O(m)$ time algorithm to construct a graph iff these conditions are satisfied.
The JDM Configuration Model

This is a generalization of the configuration model. For each vertex of degree $k$, make $k$ mini-vertices. For each edge of type $(k, l)$ make 2 mini-vertices, 1 of type $k$ and 1 of type $l$. Connect all degree $s$ mini-vertices with type $s$ mini-vertices.

**Fact**

*Any perfect matching in the JDM configuration model corresponds with a pseudograph with the desired joint degree matrix.*

![Diagram showing the JDM Configuration Model with mini-vertices and endpoints connected by edges.]
The JDM Configuration Model gives us a simple way to construct a Markov Chain - select two edges in the same gadget and swap them.

Optional: If the swap creates a self-loop or multi-edge then reject the transition. The state space with this is reduced to only simple graphs.
Evaluating the Mixing Time

Current methods are not easily applied to the simple Markov Chain. Instead, we use autocorrelation to evaluate the mixing time of the chain.

**Definition**

Given a series of data $X$, the autocorrelation function is $R_X(t) = \frac{E[(X_i - \mu)(X_{i-t} - \mu)]}{\sigma^2}$.

**Definition**

The integrated autocorrelation time is

$\tau_{int,X} = \frac{1}{2} \sum_{t=-\infty}^{\infty} R_X(t) = \frac{1}{2} + \sum_{t=1}^{\infty} R_X(t)$ [Sokal].
Autocorrelation is only defined for a real-valued sequence, but we are sampling combinatorial objects. For each dataset, for each of the potential $\binom{n}{2}$ edges, we recorded the indicator variable if the edge was in the sampled graph.

- For each data set, we ran the chain 15 times for 50,000 iterations. For each edge and run, we calculated the estimated integrated autocorrelation time.
Autocorrelation is only defined for a real-valued sequence, but we are sampling combinatorial objects. For each dataset, for each of the potential $\binom{n}{2}$ edges, we recorded the indicator variable if the edge was in the sampled graph.

- For each data set, we ran the chain 15 times for 50,000 iterations. For each edge and run, we calculated the estimated integrated autocorrelation time.

- We used these same samples to estimate when the autocorrelation dropped below a threshold level.
Experimental Design

Autocorrelation is only defined for a real-valued sequence, but we are sampling combinatorial objects. For each dataset, for each of the potential $\binom{n}{2}$ edges, we recorded the indicator variable if the edge was in the sampled graph.

- For each data set, we ran the chain 15 times for 50,000 iterations. For each edge and run, we calculated the estimated integrated autocorrelation time.
- We used these same samples to estimate when the autocorrelation dropped below a threshold level.
- For each data set, we took samples every 200, 400, 800 ... steps until we had 10,000 or 20,000 samples. We calculated the deviation of sample mean for each edge from the true mean given by the joint degree distribution.
Datasets

We used several publicly available datasets. These were selected for size only.

<table>
<thead>
<tr>
<th></th>
<th>V</th>
<th>E</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdjNoun</td>
<td>112</td>
<td>425</td>
<td>159</td>
</tr>
<tr>
<td>Dolphins</td>
<td>62</td>
<td>159</td>
<td>61</td>
</tr>
<tr>
<td>Football</td>
<td>115</td>
<td>616</td>
<td>18</td>
</tr>
<tr>
<td>Karate</td>
<td>34</td>
<td>78</td>
<td>40</td>
</tr>
<tr>
<td>LesMis</td>
<td>77</td>
<td>254</td>
<td>99</td>
</tr>
</tbody>
</table>
The max, median and min values over the edges for the estimated integrated autocorrelation times. L to R: Karate, Dolphins, LesMis, Adjnoun and Football
Threshold Values

For each dataset and each run we calculated the unnormalized autocorrelation values for each edge for $t$ between 100 and 15,000 in multiples of 100.
The Sample Mean Versus the Real Mean

Design due to Raftery and Lewis.
Summary of Results

|       | $|E|$ | Max El | Mean Conv. | Thresh. |
|-------|-----|--------|------------|---------|
| AdjNoun | 425 | 1186   | 800-1600   | 700     |
| Dolphins | 159 | 528    | 400-600    | 600     |
| Football | 616 | 1546   | 800-1600   | 900     |
| Karate   | 78  | 382    | 200-400    | 400     |
| LesMis   | 254 | 894    | 800-1600   | 1000    |

Each experiment gives approximately the same answer as to the mixing time.
Further Experiments

We have repeated these experiments for larger graphs (2,000 to 110,000 edges) using observations discussed in the paper that allow us to focus on the edges that appear to mix the most slowly. These results continue the trend, allowing us to estimate that the mixing time for this chain is linear.
We have made several contributions to both the joint degree distribution problem and the design of experiments for Markov Chains

- **Theoretical**
  - We have given necessary and sufficient conditions for a JDM to be graphical.
  - We have introduced the JDM Configuration Model, and an algorithm for constructing graphs with a given JDM in $O(|E|)$ time using it.
  - We have proved that the state space of the Markov Chain is connected.
Conclusions

We have made several contributions to both the joint degree distribution problem and the design of experiments for Markov Chains

- **Theoretical**
  - We have given necessary and sufficient conditions for a JDM to be graphical.
  - We have introduced the JDM Configuration Model, and an algorithm for constructing graphs with a given JDM in $O(|E|)$ time using it.
  - We have proved that the state space of the Markov Chain is connected.

- **Experimental:** We have used autocorrelation as an efficient and effective heuristic for estimating mixing time of Markov Chains for combinatorial data.

**Contact:** isabelle@eecs.berkeley.edu