Lecture 3: Simulating social networks

Short Course on Social Interactions and Networks,

CEMFI, May 26-28th, 2014

Bryan S. Graham

28 May 2014

Consider a network with adjacency matrix $\mathbf{D} = \mathbf{d}$ and corresponding degree sequence $\mathbf{D}_+ \overset{def}{=} (D_{1+}, \ldots, D_{N+}) = (d_{1+}, \ldots, d_{N+}) \overset{def}{=} \mathbf{d}_+$. Let $\mathbb{D}_{N,\mathbf{d}_+}$ denote the set of all undirected $N \times N$ adjacency matrices with degree counts equal to $\mathbf{d}_+$. This note describes an algorithm for sampling uniformly from the set $\mathbb{D}_{N,\mathbf{d}_+}$.

Sampling uniformly from $\mathbb{D}_{N,\mathbf{d}_+}$ has a number of uses. First, as we will see later, it facilitates conditional inference on, and conditional estimation of, certain models of network formation. Second, in some settings the researcher may only observe agents' degree and not their actual links. For example a researcher may know how many friends an individual has, but not who they are. In such cases, properties of the *class* of networks consistent with the available degree information may be of interest.

A concrete example is helpful. Let $f(\mathbf{D})$ be some function of the adjacency matrix, say its transitivity index. Among all undirected networks with degree sequences coinciding with $\mathbf{D}$'s what fraction have a transitivity index less than the one observed in the network in hand? Let $\left|\mathbb{D}_{N,\mathbf{d}_+}\right|$ denote the size, or cardinality, of $\mathbb{D}_{N,\mathbf{d}_+}$. We seek to evaluate

$$\Pr\left(f(\mathbf{D}) \leq c\right) = \frac{\sum_{\mathbf{v} \in \mathbb{D}_{\mathbf{N},\mathbf{d}_+}} \mathbf{1}\left(f(\mathbf{v}) \leq c\right)}{\left|\mathbb{D}_{N,\mathbf{d}_+}\right|}. \tag{1}$$

Direction enumeration of all the elements of $\mathbb{D}_{N,\mathbf{d}_+}$ is generally not feasible. Even for networks that includes as few as 10 agents, this set may have millions of elements. We therefore require a method of sampling from $\mathbb{D}_{N,\mathbf{d}_+}$ uniformly and also estimating its size.

Two complications arise. First, it is not straightforward to construct a random draw from $\mathbb{D}_{N,\mathbf{d}_+}$. Second, we must draw uniformly from this set. Fortunately the first challenge is

solvable using ideas from the discrete math literature. To ensure our draws are uniform we use importance sampling (e.g., Owen, 2013).

Researchers in graph theory and discrete math have studied the construction of graphs with fixed degrees and, in particular, provided conditions for checking whether a particular degree sequence is graphical (e.g., Sierksma and Hoogeveen, 1991). We say that $\mathbf{D}_+$ is **graphical** if there is feasible undirected network with degree sequence $\mathbf{D}_+$. Not all integer sequences are graphical. For example, there is no feasible undirected network of three agents with degree sequence $\mathbf{D}_+ = (3, 2, 1)$.

Recently, Del Genio, Kim, Toroczkai and Bassler (2010), Blitzstein and Diaconis (2011), Zhang and Chen (2012) and others have constructed (reasonably) efficient procedures for sampling uniformly from the set $\mathbb{D}_{N,\mathbf{d}_+}$. In this note I outline the importance sampling algorithm of Blitzstein and Diaconis (2011).[1]

Determining whether a candidate degree sequence is graphical

A sequential network construction algorithm begins with a matrix of zeros and sequentially adds links to it until its rows and columns sum to the desired degree sequence. Unfortunately, unless the links are added appropriately, it is easy to get "stuck" (in the sense that a certain point in the process it becomes impossible to reach a graph with the desired degree and the researcher must restart the process (e.g., Snijders, 1991)).

Blitzstein and Diaconis (2011) propose an algorithm that is *guaranteed* to produce a matrix from the set $\mathbb{D}_{\mathbf{d}_+,N}$. A key feature of this algorithm is cleverly using checks for whether an integer sequence is graphic when adding links.

Let $\mathbf{D}_+ = (D_{1+}, \ldots, D_{N+})$ be a sequence of candidate degrees for each of N agents in a network. Without loss of generality assume that the elements of $\mathbf{D}_+$ are arranged in descending order so that $D_{1+} \geq D_{2+} \geq \cdots \geq D_{N+}$. In a paper published in Hungarian, Erdos and Gallai (1961) showed $\mathbf{D}_+$ is graphical if and only if $\sum_{i=1}^{N} D_{i+}$ is even and

$$\sum_{i=1}^{k} D_{i+} \leq k\,(k-1) + \sum_{i=k+1}^{N} \min\left(k, D_{i+}\right) \text{ for each } k \in \{1, \ldots, N\}\,.$$

To show necessity of the condition observe that for any set S of k agents in the network there can be at most $\binom{k}{2} = \frac{1}{2}k\,(k-1)$ links between them. For the remaining $N-k$ agents with $i \notin S$ there can be at most $\min\left(k, D_{i+}\right)$ links from i to agents in S.

[1]The work of Del Genio, Kim, Toroczkai and Bassler (2010) and Blitzstein and Diaconis (2011) was evidently undertaken independently. The method of Zhang and Chen (2012) improves upon their algorithms.

The study of graphic integer sequences has a long history in discrete math. Sierksma and Hoogeveen (1991) summarize several criteria that can be used to check whether $\mathbf{D}_+$ is is graphical. Blitzstein and Diaconis (2011) base their sampling algorithm on a simple recursive test for whether D_+ is graphical due to Havel (1955) and Hakimi (1962).

Theorem 1. *(Havel-Hakimi) Let $D_{i+} > 0$, if $\mathbf{D}_+$ does not have at least D_{i+} positive entries other than i it is not graphical. Assume this condition holds. Let $\tilde{\mathbf{D}}_+$ be a degree sequence of length $N-1$ obtained by*
[i] deleting the i^{th} entry of $\mathbf{D}_+$ and
[ii] subtracting 1 from each of the D_{i+} highest elements in $\mathbf{D}_+$ (aside from the i^{th} one).
$\mathbf{D}_+$ is graphical if and only if $\tilde{\mathbf{D}}_+$ is graphical. If $\mathbf{D}_+$ is graphical, then it has a realization where agent i is connected to any of the D_{i+} highest degree agents (other than i).

Proof. See Blitzstein and Diaconis (2011). □

Theorem 1 is suggestive of a sequential approach to building an undirected network with degree sequence $\mathbf{D}_+$. The procedure begins with a target degree sequence $\mathbf{D}_+$. It starts by choosing a link partner for the lowest degree agent (with at least one link). It chooses a partner for this agent from among those with higher degree. A one is then subtracted from the lowest degree agent and her chosen partner's degrees. This procedure continues until the **residual degree sequence** (the sequence of links that remain to be chosen for each agent) is zero.

To describe the method proposed Blitzstein and Diaconis (2011) we require some additional notation. Let $(\oplus_{i_1,\ldots,i_k}\mathbf{D}_+)$ be the vector obtained by adding a one to the $i_1,\ldots,i_k$ elements of $\mathbf{D}_+$:

$$(\oplus_{i_1,\ldots,i_k}\mathbf{D}_+)_j = \begin{cases} D_{j+}+1 & \text{for } j \in \{i_1,\ldots,i_k\} \\ D_{j+} & \text{otherwise} \end{cases}$$

Let $(\ominus_{i_1,\ldots,i_k}\mathbf{D}_+)$ be the vector obtained by subtracting one from the $i_1,\ldots,i_k$ elements of $\mathbf{D}_+$:

$$(\ominus_{i_1,\ldots,i_k}\mathbf{D}_+)_j = \begin{cases} D_{j+}-1 & \text{for } j \in \{i_1,\ldots,i_k\} \\ D_{j+} & \text{otherwise} \end{cases}$$

Algorithm 1. *(Blitzstein and Diaconis) A sequential algorithm for constructing a random graph with degree sequence $\mathbf{D}_+ = (D_{1+},\ldots,D_{N+})'$ is*

1. Let $\mathbf{G}$ be an empty adjacency matrix.
2. If $\mathbf{D}_+ = \mathbf{0}$ terminate with output $\mathbf{G}$

3. Choose the agent i with minimal positive degree D_{i+}.

4. Construct a list of candidate partners $J = \{j \neq i : \mathbf{G}_{ij} = \mathbf{G}_{ji} = 0 \text{ and } \ominus_{i,j} \mathbf{D}_+ \text{ graphical}\}$.

5. Pick a partner $j \in J$ with probability proportional to its degree in $\mathbf{D}_+$.

6. Set $\mathbf{G}_{ij} = \mathbf{G}_{ji} = 1$ and update $\mathbf{D}_+$ to $\ominus_{i,j}\mathbf{D}_+$.

7. Repeat steps 4 to 6 until the degree of agent i is zero.

8. Return to step 2.

The input for Algorithm 1 is the target degree sequence $\mathbf{D}_+$ and the output is an undirected adjacency matrix $\mathbf{G}$ with $\mathbf{G}'\iota = \mathbf{D}_+$.
An example is:

$$(3,2,2,2,1) \to (3,1,2,2,0) \to (2,0,2,2,0) \to (1,0,2,1,0) \to (0,0,1,1,0) \to (0,0,0,0,0)$$

Importance sampling

Let $\mathcal{Y}_{N,\mathbf{d}_+}$ denote the set of all possible sequences of links outputted by Algorithm 1 given input $\mathbf{D}_+ = \mathbf{d}_+$. Let $\mathcal{G}(Y)$ be the adjacency matrix induced by link sequence Y. Let Y and Y' be two different sequences produced by the algorithm. These sequences are equivalent if their "end point" adjacency matrices coincide (i.e., if $\mathcal{G}(Y) = \mathcal{G}(Y')$). We can partition $\mathcal{Y}_{N,\mathbf{d}_+}$ into a set of equivalence classes, the number of such classes coincides with the number of feasible networks with degree distribution $\mathbf{D}_+$ (i.e., with the cardinality of $\mathbb{D}_{N,\mathbf{d}_+}$). Let $c(Y)$ denote the number of possible link sequences produced by Algorithm 1 that produce Y's end point adjacency matrix (i.e., the number of different ways in which Algorithm 1 can generate a given adjacency matrix).
Let $i_1, i_2, \ldots, i_M$ be the sequence of agents chosen in step 3 of Algorithm 1 in which Y is the output. Let $a_1, \ldots, a_m$ be the degree sequences of $i_1, \ldots, i_M$ at the time when each agent was first selected in step 3, then

$$c(Y) = \prod_{k=1}^{M} a_k!$$

Let $\sigma(Y)$ be the probability that Algorithm 1 produces link sequence Y. Note that $\sigma(Y)$ is easy to compute. Each time the algorithm choose a link in step 5 record the probability with which it was chosen (i.e., the residual degree of the chosen agent divided by the sum of the residual degrees of all agents in the choice set). The product of all these probabilities equals $\sigma(Y)$.

Let $f(\mathbf{G})$ be some function of the adjacency matrix and consider the expected value

$$\begin{aligned}
\mathbb{E}\left[\frac{\pi(\mathcal{G}(Y))}{c(Y)\,\sigma(Y)} f(\mathcal{G}(Y))\right] &= \sum_{y\in\mathbb{Y}_{N,\mathbf{d}}} \frac{\pi(\mathcal{G}(y))}{c(y)\,\sigma(y)} f(\mathcal{G}(y))\,\sigma(y) \\
&= \sum_{y\in Y_{N,\mathbf{d}}} \frac{\pi(\mathcal{G}(y))}{c(y)} f(\mathcal{G}(y)) \\
&= \sum_{\mathbf{g}\in\mathbb{D}_{N,\mathbf{d}}} \sum_{\{y:G(y)=\mathbf{g}\}} \frac{\pi(\mathbf{g})}{c(y)} f(\mathbf{g}) \\
&= \sum_{\mathbf{g}\in\mathbb{D}_{N,\mathbf{d}}} \pi(\mathbf{g})\, f(\mathbf{g}) \\
&= \mathbb{E}_{\pi}[f(\mathbf{G})].
\end{aligned}$$

The ratio $\pi(\mathbf{G}(Y_t))/c(Y_t)\,\sigma(Y_t)$ is called the likelihood ratio or the **importance weight.** If we set $f(\mathbf{G})$ to the constant function we see that the expected value of this weight is one. This suggests the analog estimator

$$\hat{\mu}_{f(\mathbf{G})} = \left[\sum_{t=1}^{T} \frac{\pi(\mathbf{G}(Y_t))}{c(Y_t)\,\sigma(Y_t)}\right]^{-1} \times \sum_{t=1}^{T} \frac{\pi(\mathbf{G}(Y_t))}{c(Y_t)\,\sigma(Y_t)} f(\mathbf{G}(Y_t)).$$

Setting $\pi(\mathbf{G}) = 1$ we get a procedure for estimating the expectation of $f(\mathbf{G})$ when $\mathbf{G}$ is drawn uniformly from $\mathbb{D}_{N,\mathbf{d}_+}$.

References

[1] Blitzstein, Joseph and Persi Diaconis. (2011). "A sequential importance sampling algorithm for generating random graphs with prescribed degrees," *Internet Mathematics* 6 (4): 489 - 522.

[2] Del Genio, Charo I., Hyunju Kim, Zoltan Toroczkai and Kevin Bassler. (2010). "Efficient and exact sampling of simple graphs with given arbitrary degree sequence," *Plos One* 5 (4): e100012.

[3] Erdös, Paul and Tibor Gallai. (1961). "Graphen mit punkten vorgeschriebenen grades," *Matematikai Lapok* 11: 264 - 274.

[4] Hakimi, S. L. (1962). "On realizability of a set of integers as degrees of the vertices of a linear graph. I," *Journal of the Society for Industrial and Applied Mathematics* 10 (3): 496 – 506.

[5] Havel, Václav J. (1955). "A remark on the existence of finite graph," *Časopis Pro Pěstování Matematiky* 80: 477 - 480.

[6] Owen, Art. B. (2013). *Monte Carlo Theory, Methods and Examples* available online at `http://statweb.stanford.edu/owen/mc/`.

[7] Sierksma, Gerard and Han Hoogeveen. (1991). "Seven criteria for integer sequences being graphic," *Journal of Graph Theory* 15 (2): 223 – 231.

[8] Snijders, Tom A. B. (1991). "Enumeration and simulation methods for 0-1 matrices with given marginals," *Psychometrika* 56 (3): 397 - 417.

[9] Zhang, Jingfei and Yuguo Chen. (2013). "Sampling for conditional inference on network data," *Journal of the American Statistical Association* 108 (504): 1295 - 1307.