A Plea for Clarity: What is a Network?

- graph vs. multigraph (are loops, multiple edges ok? What is a “simple” graph?)
- directed vs. undirected
- weighted vs. unweighted
- dynamics of vs. dynamics on
- labeled vs. unlabeled
- network as quantity of interest vs. quantities of interest on networks
Pairing Model (Configuration Model) for Random Graphs with Given Degrees

Create the desired number of stubs ("half-edges"), then randomly connect pairs.
Pairing Model (Configuration Model) for Random Graphs with Given Degrees

Create the desired number of stubs ("half-edges"), then randomly connect pairs.

Random Graph Dynamics (Durrett):

Theorem 2.1.1. Let $\mu = \sum k p_k$ and $\nu = \sum k(k - 1)p_k$. As $n \to \infty$, the number of self-loops $\chi_0$ and the number of parallel edges $\chi_1$ are asymptotically independent Poisson($\nu/2\mu$) and Poisson($((\nu/2\mu)^2)$. Thus the probability that the graph is simple has a positive limit.
Pairing Model (Configuration Model) for Random Graphs with Given Degrees

Create the desired number of stubs ("half-edges"), then randomly connect pairs.

Random Graph Dynamics (Durrett):

**Theorem 2.1.1.** Let \( \mu = \sum_k kp_k \) and \( \nu = \sum_k k(k - 1)p_k \). As \( n \to \infty \), the number of self-loops \( \chi_0 \) and the number of parallel edges \( \chi_1 \) are asymptotically independent Poisson(\( \nu/2\mu \)) and Poisson((\( \nu/2\mu \))^2). Thus the probability that the graph is simple has a positive limit.

**Bender-Canfield:** \[ P(\text{simple}) \sim e^{\frac{1-d^2}{4}} \text{ as } n \to \infty. \]
Markov Chain Monte Carlo

mixing times, burn-in, bottlenecks, autocorrelation,...

Switchings Chain

1 — 2

3 — 4
Markov Chain Monte Carlo

mixing times, burn-in, bottlenecks, autocorrelation,...

Switchings Chain

1 ---- 2

3 ---- 4

1 ---- 2

3 ---- 4
Importance Sampling

$\pi$: target distribution (desired to estimate with respect to)

$\sigma$: trial distribution (available to us)
Importance Sampling

$\pi$: target distribution (desired to estimate with respect to)
$\sigma$: trial distribution (available to us)

$$E_{\pi} f(X) = \sum_{x} f(x) \pi(x)$$
Importance Sampling

$\pi$: target distribution (desired to estimate with respect to)

$\sigma$: trial distribution (available to us)

$$E_\pi f(X) = \sum_x f(x)\pi(x) = \sum_x f(x)\frac{\pi(x)}{\sigma(x)}\sigma(x)$$
Importance Sampling

$\pi$: target distribution (desired to estimate with respect to)

$\sigma$: trial distribution (available to us)

$$E_\pi f(X) = \sum_x f(x)\pi(x) = \sum_x f(x)\frac{\pi(x)}{\sigma(x)}\sigma(x) = E_\sigma f(X)\frac{\pi(X)}{\sigma(X)}$$
Importance Sampling

\( \pi: \) target distribution (desired to estimate with respect to)

\( \sigma: \) trial distribution (available to us)

\[
E_\pi f(X) = \sum_x f(x)\pi(x) = \sum_x f(x)\frac{\pi(x)}{\sigma(x)}\sigma(x) = E_\sigma f(X)\frac{\pi(X)}{\sigma(X)}
\]

Unbiased estimator:
Importance Sampling

\( \pi \): target distribution (desired to estimate with respect to)

\( \sigma \): trial distribution (available to us)

\[
E_\pi f(X) = \sum_x f(x)\pi(x) = \sum_x f(x)\frac{\pi(x)}{\sigma(x)}\sigma(x) = E_\sigma f(X) \frac{\pi(X)}{\sigma(X)}
\]

Unbiased estimator:

\[
\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} f(X_i) \frac{\pi(X_i)}{\sigma(X_i)},
\]
Importance Sampling

$\pi$: target distribution (desired to estimate with respect to)

$\sigma$: trial distribution (available to us)

\[
E_\pi f(X) = \sum_x f(x)\pi(x) = \sum_x f(x)\frac{\pi(x)}{\sigma(x)}\sigma(x) = E_\sigma f(X)\frac{\pi(X)}{\sigma(X)}
\]

Unbiased estimator: 
\[
\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} f(X_i)\frac{\pi(X_i)}{\sigma(X_i)},
\]

The ratios $W_i = \pi(X_i)/\sigma(X_i)$ are called importance weights.
Importance Sampling

$\pi$: target distribution (desired to estimate with respect to)  
$\sigma$: trial distribution (available to us)

$$E_{\pi} f(X) = \sum_{x} f(x) \pi(x) = \sum_{x} f(x) \frac{\pi(x)}{\sigma(x)} \sigma(x) = E_{\sigma} f(X) \frac{\pi(X)}{\sigma(X)}$$

Unbiased estimator:  
$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} f(X_i) \frac{\pi(X_i)}{\sigma(X_i)},$$

The ratios $W_i = \pi(X_i)/\sigma(X_i)$ are called importance weights.

$$\tilde{\mu} = \frac{1}{\sum_i W_i} \sum_{i=1}^{N} f(X_i) W_i$$
**Importance Sampling**

\( \pi \): target distribution (desired to estimate with respect to)

\( \sigma \): trial distribution (available to us)

\[
E_\pi f(X) = \sum_x f(x)\pi(x) = \sum_x f(x)\frac{\pi(x)}{\sigma(x)}\sigma(x) = E_\sigma f(X)\frac{\pi(X)}{\sigma(X)}
\]

Unbiased estimator: 
\[
\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} f(X_i)\frac{\pi(X_i)}{\sigma(X_i)},
\]

The ratios \( W_i = \pi(X_i)/\sigma(X_i) \) are called *importance weights*.

Biased estimator: 
\[
\tilde{\mu} = \frac{1}{\sum_i W_i} \sum_{i=1}^{N} f(X_i)W_i
\]
Importance Sampling Approaches to Graphs with Given Degrees

Use combinatorial characterizations to avoid getting stuck! Make sure probabilities are known for the proposal distribution, up to a constant.

Blitzstein-Diaconis (to appear in Journal of Internet Mathematics)

Del Genio-Kim-Toroczkai-Bassler (PLOSonE 2010)
Volz-Heckathorn RDS Estimator

\[ E(\hat{Y}) = \frac{\sum_{j=1}^{n} Y_j / d_j}{\sum_{j=1}^{n} 1 / d_j} \]

This is a form of Horvitz-Thompson estimator, reweighting as in importance sampling.
Heckathorn Assumptions

- node degrees are known exactly
- recruitment is uniform
- ties are symmetric
- Markov chain converges rapidly to its stationary distribution (burn in? seediness?)
- sampling with vs. without replacement doesn’t matter

Handcock-Gile (Sociological Methodology, 2010) investigate many of these (mainly seed selection, recruitment preferences, and sampling with/without replacement)
Weighted vs. Unweighted

\[ \frac{\sum_i w_i y_i}{\sum_i w_i} - \bar{y} = \frac{S_y S_w}{\bar{w}} \text{Corr}(y, w) \]

So when do weights matter?
Can one afford to upweight/downweight so much in small samples?
To Model or Not To Model

• the underlying network? what about unknown nodes?
• the coupon assignment process?
• coupon refusal?
• the outcome variables (such as HIV status)?
Goel-Salganik (Stats in Medicine 2009, PNAS 2010):

RDS variances can be extremely large, especially if there are bottlenecks in the network from modularity/communities, and from multiple recruitment.

Work with Sergiy Nesterko (see poster presentation) explores the bias-variance tradeoff for random networks from a latent space model. What happens under different assumptions about:

(a) homophily?
(b) recruitment preferences?
(c) seed selection?
Recruitment Preferences

• What is a reasonable model for the recruitment process?

• Try various functional forms for probability of recruitment as a function of distance in a latent space, e.g., from Hoff et al model (2002)

\[ \eta_{i,j} = \log \text{odds}(y_{i,j} = 1 | z_i, z_j, x_{i,j}, \alpha, \beta) \]

\[ = \alpha + \beta' x_{i,j} - |z_i - z_j|. \]
Bias-variance tradeoff, homophily case
(simulation by Sergiy Nesterko)
Impasse!

Uniform recruitment is very implausible and makes a big difference, but what else can we do?
Impasse!

Uniform recruitment is very implausible and makes a big difference, but what else can we do?

Answer: get better data.
What would Fisher say?

To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.

-- R.A. Fisher
UC San Diego HIV/AIDS Data

- Program Project at the Antiviral Research Center at UCSD (PI Susan Little; working with Victor deGruttola at Harvard Biostat and others)

- aim to understand transmission and risk factors for HIV, the underlying social and sexual networks, and estimate prevalence in various “high risk” populations in San Diego

- participants recruited both via RDS and via other methods
First 16 responses to “Why do you think the person who gave you the coupon chose you?”

Cause i'm fly!
I'm positive and dating his friend
to collect his money for a referral friendship
He knew I would participate
He thought I would very likely participate
I bumped into him right after he completed his own participation
I'm accountable
He knew I would complete the study.
I am dependable.
I would follow through unknown recommendation
Help me with extra income 
we are friends
Model Ingredients

- overall mean parameter
- seed distribution
- strength of ties on each edge
- homophily, correlated distribution on $y$

Simple model: assume different recruitment trees are independent, and focus on modeling the correlations in each individual tree of the branching process.

$p_{seed}(y_1|\theta)$

\[ p(y_2|y_1, s_{12}, \theta)p(y_3|y_1, s_{13}, \theta)p(y_4|y_1, s_{14}, \theta) \]

\[ p(y_5|y_2, s_{25}, \theta)p(y_6|y_2, s_{26}, \theta)p(y_7|y_4, s_{47}, \theta) \]
MSEs averaged over networks and RDS runs

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volz-Heckathorn</td>
<td>0.15</td>
</tr>
<tr>
<td>vanilla mean</td>
<td>0.14</td>
</tr>
<tr>
<td>model-based MLE</td>
<td>0.13</td>
</tr>
</tbody>
</table>

(degrees treated as known exactly)
Likelihood functions give variance estimates naturally, and better convey uncertainty about parameters.
Some Open Problems

- adaptive coupon designs (number of coupons and eligibility on each)
- seedy problems
- better handle missing data in the y’s
- better models of the dependence structure
- combine with spatial models
- integration with other data sources