Why Does Sampling Matter?
Why Does Sampling Matter?
Why Does Sampling Matter?
Why Does Sampling Matter?
Why Does Sampling Matter?
Why Does Sampling Matter?
Sampling and Inference

- Two basic strategies for linking designs to inference
  - Model-based inference: incorporate the measurement process into the likelihood
  - Design-based inference: choose the design such that functions of data have known (usually asymptotic properties)

- Both have advantages, liabilities
  - Model-based inference more flexible, but design-based inference generally easier
  - Trade-off: assumptions about observation vs. assumptions about control
Probability Samples of Nodes

- Important practical case: probability samples of nodes
  - Individuals sampled independently from a well-defined (but possibly unframed) population
  - Individual sampling/inclusion probabilities known up to a constant

- Starting point for many applications
  - Estimation of attribute distributions
  - Estimation of local properties
    - Degree distribution
    - Ego-net properties
    - Attribute/Ego-net correlations
  - Seeds for $k$th order ego-net methods (Pattison, Robins)
Basic Design-based Estimators

- **Sampling with replacement: Hansen-Hurwitz**
  - $N$ elements, independent sample of size $n$
  - $i$th unit selected w/prob $p_i$ each draw; has value $y_i$
  - HH estimators of population mean, variance of mean estimator:
    \[
    \hat{\mu}_{HH} = \frac{1}{nN} \sum_{i=1}^{n} \frac{y_i}{p_i}
    \]
    \[
    \text{Var}(\hat{\mu}_{HH}) = \frac{1}{N^2 n(n-1)} \sum_{i=1}^{n} \left( \frac{y_i}{p_i} - N \hat{\mu}_{HH} \right)^2
    \]

- **Sampling without replacement: Horvitz-Thompson**
  - $N$ elements, sample of size $n$
  - Prob. of $i$th element appearing $p_i$, w/value $y_i$; joint $i,j$ inclusion probability $p_{ij}$
  - HT estimators of population mean, variance of mean estimator:
    \[
    \hat{\mu}_{HT} = \frac{1}{N} \sum_{i=1}^{n} \frac{y_i}{p_i}
    \]
    \[
    \text{Var}(\hat{\mu}_{HT}) = \sum_{i=1}^{n} \left( \frac{1-p_i}{p_i} \right) \frac{y_i^2}{N^2} + \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \left( \frac{p_{ij} - p_i p_j}{p_i p_j} \right) \frac{y_i y_j}{N^2}
    \]
Whence the Probability Sample?

• Wonderful, but where do we get these magical samples?

• Conventional methods
  – Where sampling frame exists and units are reachable, problem is well-studied and methods are direct
  – Typical cases: RDD and postal recruitment surveys, organizational samples
  – Nothing special about this setting, although what one administers to the respondents may be different...
Unconventional Cases: Population Sampling Unframed

- In some settings, no frame exists - or we can't use it
  - Hidden populations, online user communities, organizational populations, etc., often lack a sampling frame: population is indirectly defined, and cannot be enumerated
  - Alternately, may not be able to perform sampling from the frame - even if we know of a member, we have no way to recruit him or her

- Important workaround: network-based sampling methods
  - Exploit ties among actors to draw a probability sample
  - Today, we'll consider some basic techniques (some new!)
Extant Methods

• Primary family of methods: link-trace sampling
  – Exploits network structure for sampling purposes
  – Basic idea: find nodes by following links from an initial seed set
  – Many, many variants
    • Widely used for hidden population, online studies

• Some examples
  – Breadth-first search (BFS)
    • Visit all nodes at distances 1, 2, ... from seed
  – Random Walk sampling (RW)
    • Choose random neighbor of a node to visit
    • Repeat above step for a "long" time
BFS, Illustrated
BFS, Illustrated
BFS, Illustrated
BFS, Illustrated
Random Walk, Illustrated
Random Walk, Illustrated
Random Walk, Illustrated
Random Walk, Illustrated
Random Walk, Illustrated
Random Walk, Illustrated
Challenges to Effective Use

• Lack of a known equilibrium distribution
  – BFS, most ad hoc methods badly biased (unless whole network is captured)
  – RW biased, but converges to $1/d(v)$ in the undirected, connected case
    • Can observe $d(v)$, and thus adjust post-hoc
    • Directed case harder - can derive in theory, but not easily measure

• Unverified convergence
  – For methods with an equilibrium, need to verify convergence
    • These are really just MCMC methods; same issues apply
  – Methods do exist, but only recently applied to this problem
    • One area of progress: application of MCMC diagnostics to network sampling procedures
Avoiding Bias with MCMC Theory

• Why not derive a link-trace design that has a uniform (or other target) equilibrium distribution?

• Metropolis-Hastings Random Walk Sampling
  – Like simple RW, but rejects moves proportionally to ratio of old/new degrees
  – Equilibrium is uniform on sampled component (for version shown)
  – If converged, sample does not require reweighting for standard applications

• MHRW algorithm:
  – initialize: \( v^{(0)} \in V, G \)
  – Let CONVERGED := FALSE
  – Let \( i := 0 \)
  – while !CONVERGED do
    – Let \( i := i + 1 \)
    – Draw \( v^{(i)} \) from Unif(\( N(v^{(i-1)}) \))
    – if Unif(0,1) < \( d(v^{(i-1)})/d(v^{(i)}) \) then
      – Let \( v^{(i)} = v^{(i-1)} \)
    – endif
    – if \( v^{(0)},...,v^{(i)} \) has converged then
      – Let CONVERGED := TRUE
    – endif
  – endwhile
  – return \( v^{(0)},...v^{(i)} \)
MCMC Theory Part II: Knowing When to Quit

- When have we taken enough draws to attain convergence?
  - Markov chain must "forget" starting point ("burn-in")
  - Set of draws as a whole should approximate sample from chain equilibrium

- MCMC diagnostics: telling us when we've finished
  - Many heuristics (see, e.g., Gilks et al., 1996)
  - Simple example: Gewke's (1992) $z_G$

$$y_{a,b} = (y_a, \ldots, y_b), z_G = \frac{\bar{y}_{\alpha n, (\alpha + \delta) n} - \bar{y}_{(1-\delta) n, n}}{\sqrt{\text{var} \left( y_{\alpha n, (\alpha + \delta) n} \right) + \text{var} \left( y_{(1-\delta) n, n} \right)}}$$

- Converges to $\mathcal{N}(0,1)$ as $n \to \infty$
Application: Probability Sampling of Facebook Users

- Large online service (>2x10^8 users at time of study)
- Can no longer sample directly
  - (Could before, but few knew this!)
- Comparative study of sampling methods, using convergence diagnostics (M. Gjoka et al., 2010)
  - Goal: probability sample of non-isolate, publicly viewable users
  - Methods: BFS, RW, MHRW, Uniform (reference sample)
  - 28 seeds from uniform sample used to launch independent parallel traces
    - Each trace continued for exactly 81K steps (except Uniform, fixed at 982K)
    - Within (Geweke's $z_G$) and between (G+R's $R$) chain metrics used to extract final samples for RW, MHRW
Convergence for the MHRW Algorithm

Overall: acceptable convergence between 500 and 3000 iterations (depending on measure)

(M. Gjoka et al., 2010)
Comparative Estimation of Local Properties

(M. Gjoka et al., 2010)
Comparative Estimation of the Degree Distribution

(M. Gjoka et al., 2010)
Expansion: Multigraph Sampling

- Often, no **one** network on a given population supports sampling
  - May be fragmented, or clustered/heterogeneous (slowing convergence)
- **Solution:** multigraph sampling (Gjoka et al., 2011)
  - Walk on multiple graphs, or unions of graphs
  - Much better properties, esp if uncorrelated
  - Individual networks need not be well-connected to be useful
Multigraph Sampling Algorithm

- Want to sample from $\mathcal{G} = \{ G_1, \ldots, G_Q \}$ on vertex set $V$

- Trivial solution: just walk on the union graph
  - Fine, but requires taking unions of neighborhoods; becomes very expensive for graphs with large cliques (e.g., group co-membership)

- Alternate approach: mix across graphs
  - Add loops to each graph
  - Choose $G_j$ from $v_i$ w/prob $q_{ij} = d(v_i, G_j) / \sum_k d(v_i, G_k)$
  - Choose uniformly from $N_{G_j}(v_i)$

- Sampling Algorithm
  - initialize: $v^{(0)} \in V$, $\mathcal{G}$, $q$
  - Let CONVERGED := FALSE
  - Let $i := 0$
  - while !CONVERGED do
    - Let $i := i + 1$
    - Draw $G'$ from $\mathcal{G}$ with pmf $\text{Multinom}(q_{v^{(i-1)}}, \ldots, q_{v^{(i-1)}Q})$
    - Draw $v^{(i)}$ from $\text{Unif}(N_{G'}(v^{(i-1)}))$
    - if $v^{(0)}, \ldots, v^{(i)}$ has converged then
      - Let CONVERGED := TRUE
    - endwhile
  - return $v^{(0)}, \ldots, v^{(i)}$
Simulated Example

- **Vertex sampling probability converges to** $Pr(v(0)=v_j) \propto \sum_k d(v_j, G_k)$

- **Ex: 5 order 50 Bernoulli graphs, $\bar{d}=1.5$**
  - Each very fragmented, but union is well-connected

- **Sampling from a single vertex; sample of 1e5 thinned from Markov chain of length 5e6**
  - $\chi^2 p=0.52$ vs. theoretical

---

[Graphs and diagrams showing the fragmented and well-connected nature of the sampled graphs.]

---

**Observed vs. Theoretical Sampling Frequency, w/95% Confidence Intervals**
Network Sampling: Some Practical Guidelines

- Use principled sampling methods
- Check for convergence while sampling
  - (And be ready to use long chains!)
- If connectivity/mixing an issue, consider multigraph sampling
- If you use RW, remember to reweight!
Further Remarks

- Many ways in which MCMC theory can help with large-network sampling
  - Especially useful in the online case, since long chains are easy to obtain, sampling is cheap
  - Easy to apply "off the shelf" simulation/diagnostic strategies to network sampling problems
  - Minor extensions (e.g., multigraph sampling) can help patch weaknesses with current approaches

- Many applications
  - Link-trace sampling now vital in many important contexts (e.g., RDS for disease, drug use estimation)
  - Increasingly important in growing online communities