STA6800 - Statistical Analysis of Network

Data Collection and Sampling

Ick Hoon Jin

Yonsei University, Department of Statistics and Data Science

1. Sampling Procedures
2. Sampling Designs
3. Coping Strategies
4. Big Data Solves Nothing

Difficulties in Network Data Collection

- The real foundation of any branch of statistics is data collection.
- For the sorts of statistics we have mostly seen before, where data are IID (or IID-ish), data comes from samples or from experiments.
- It is hard (though not impossible) to experiment with networks, so we mostly have to deal with samples, and even there, things become much more complicated than we are used to.
- Unfortunately, this complexity is all too often ignored when analyzing empirical networks.

Ideal Data: Network Census

- The ideal data would be a census or enumeration of the network.
- This would record every node, and every edge between nodes, with no spurious additional nodes or edges.
- If you are in the fortunate situation of having a complete network census, you can pretty much ignore the sampling process, and proceed to model network formation.

Imperfections in Network Censuses

- Unfortunately, even studies which try to get a complete census may fall short of perfection.
- The exact failure modes depend on the nature of the network and indeed on the details of the measurement process; for concreteness, I focus here on survey-based measurements of social networks.
- These surveys often work by approaching people and asking them questions like “Who are your friends?” or “From whom do you seek advice?” or “From whom have you borrowed money?”

Imperfections in Network Censuses

- There can be different results depending on whether are given suggestions, or a checklist of possibilities, or are asked to spontaneously recall names.
- Answers may be influenced by shame, boastfulness, or other emotions related to the “presentation of self in everyday life”.
- In older studies, it was common to frame the question as something like “name up to three colleagues you commonly go to for advice”; such censoring by degree necessarily prevented any recorded out-degree from being higher than three.

Sampling Designs

- If we cannot get hold of the true, “population” graph $G = (V, E)$, guided by the example of IID statistics, try to measure a “sampled” graph $G^* = (V^*, E^*)$, with $V^* \subset V$ and $E^* \subset E$.
- Different sampling designs amount to different ways of obtaining such sampled subgraphs.
- Our first step in understanding sampling is the concept of a simple random sample (SRS) of units from the population. In networks, even a simple random sample is complicated.

Induced and Incident Subgraph

- Start with a simple random sample of nodes, i.e., V^* is an SRS of V. Then take the induced subgraph, i.e., $(i,j) \in E^*$ if and only if $(i,j) \in E$, $i \in V^*$, and $j \in V^*$.
- This natural procedure, induced subgraph sampling, turns out be very biased for even very simple network statistics, though the biases can sometimes be calculated and compensated for.
- One the other hand, we start with a simple random sample of edges, i.e., E^* is an SRS oof E. Them, take the nodes which are incident on those edges, i.e., $i \in V^*$ if for some $j \in V$, $(i,j) \in E^*$.
- Experience with conventional surveys may make incident-subgraph sampling seem odd, but there are many situations where it's actually quite natural.

Example of a Bias

- The canonical example of how sampling can induce a bias, even when we're just doing a simple random sample of nodes, is the mean degree.
- Intuitively, we don't see any edges outside the induced subgraph, so the degree we record for each node is at most its real degree, and the mean degree in the sampled graph should be $\leq$ the true mean degree.

Example of a Bias

- Say k_i is the degree of node i, so $k_i = \sum_{j=1}^{n} A_{ij}$.
- The mean degree over the whole network is $\bar{k} = n^{-1} \sum_{i=1}^{n} k_i$.
- Take a simple random sample of m nodes, so the probability of seeing node i is the same for all nodes, $\pi = m/n$.
- $Z_i = 1$ if node i is in the sample, and $Z_i = 0$ otherwise, i.e., Z_i is the indicator for $i \in V^*$.
- The observed graph G^* has an observed adjacency matrix A^*, and $A^*_{ij} = 1$ if and only if $A_{ij} = 1$ and both i and j are om the sample.

Example of a Bias

- What is the expected value of the plug-in estimate $\bar{k}$ from G^*, say $\bar{k}^*$?

$$
\begin{aligned}
E\left(\bar{k}^*\right) &= E\left(\frac{1}{m}\sum_{i\in V*} k_i^*\right) = E\left(\frac{1}{m}\sum_{i\in V*}\sum_{j\in V*} A_{ij}^*\right) \\
&= E\left(\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}Z_iZ_j\right) = \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}E\left(Z_iZ_j\right) \\
&= \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}\pi^2 = \frac{1}{n\pi}\pi^2\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} \\
&= \frac{\pi}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} = \pi\bar{k}
\end{aligned}
$$

Exploratory Sampling Design

- For both induced- and incident-subgraph sampling, the sampling frame is in some sense separate from the actual, realized graph. The population from which we draw our SRS has to include all nodes, or all edges, but doesn't use the graph beyond that.
- In egocentric designs, we sample nodes and record information about their local neighborhoods, or ego networks.
- "ego": Other times we record edges and non-edges among the neighbors of the initial node; This is sometimes called a star design.
- When we deal with star designs, we collect multiple local graph neighborhoods, and an important question is whether those overlap; depending on the recording process, this information might be available

Snowball Sampling

- Start with a seed node, and record its immediate neighborhood.
- We then repeat the process for each of the neighbors, and then their neighbors, etc., until either no new nodes are found or we get tired, i.e., a pre-selected size is reached.
- There can be multiple seeds; there may then be an issue of determining when two snowballs which have formed around different seeds have over-lapped.

Snowball Sampling

- Snowball sampling leads to a different distribution over graphs than does either induced- or incident-subgraph sampling.
- Even if the seed is chosen by a simple random sample, the other nodes picked up by the snowball are not a random sample.
- Since they are nodes which can be reached by following paths from the seed, they must have degree at least 1, must be at least weakly connected to the seed, and in general tend to have higher-than-average degree.

Respondent-driven Sampling

- An important variant on snowball sampling, for social networks, is **respondent-driven sampling**.
- This originated as a way of studying members of hard-to-find ("hidden") sub-populations - often ones which were hidden because membership in them is stigmatized or illegal.
- The technique is to find some initial members of the group in question, and then persuade them to recruit other members whom they know as research subjects.

Respondent-driven Sampling

- Often, the respondents are given unique physical tokens to pass on those whom they recruit, so that links can be traced, and there may be some incentive for participation.
- Censoring by degree can result if, for instance, there is only a limited number of physical tokens per respondent.

Trace-route Sampling

- Trace-route sampling probes a network by tracing routes through it.
- The typical procedure goes as follows:
  1. Pick a set of source nodes.
  2. Pick a set of target nodes.
  3. For each source-target combination, find a path from the source to the target, and record all nodes and edges traversed along the path.

  Clearly, a lot will depend on how, precisely, paths are found, but this is an application-specific issue.

Trace-route Sampling

- Depending on exactly how route-tracing gets done, one may or may not get information from “failed” routes, i.e., ones which didn’t succeed in getting from source to target.
- Trace-route sampling systematically distorts the degree distribution, making all kinds of graphs look like they have heavy-tailed distributions whether they do or not.

Head in Sand

- That is, ignore distortions or biases due to sampling, and pretend that the graph we see is the whole graph. This is generally not a good idea.
- For induced-subgraph sampling, the mean degree is biased from the real degree by a calculable factor. Indeed, the sample values of motif counts for all motifs are also biased (again, in calculable ways). These would be pretty easy to compensate for.
- But degree distribution, for example, gets distorted in very complicated, hard-to-fix ways, even with induced-subgraph sampling.

Learn Sampling Theory

- Classical sampling theory is a theory of statistical inference in which probability assumptions are only made about the sampling process.
- The true population is regarded as unknown but fixed, and no stochastic assumptions are made about how it is generated. (One can always regard this as conditioning on the unknown population.)
- Because all the probability assumptions refer to the sampling design, and the validity of the inference depends only on whether the design has been accurately modeled, this is sometimes called design-based inference.

Learn Sampling Theory

- Try to estimate the mean μ of some quantity X_i over a finite population of size n, using a sample of units S.
- A simple, classic solution is the Horvitz-Thompson estimator

  $$\hat{\mu}_{HT} \equiv \frac{1}{n} \sum_{i \in S} \frac{X_i}{\pi_i}$$

  where π_i is the (assumed-known) inclusion probability of unit i, i.e., the probability of unit i being included in the sample.
- Notice that if all inclusion probabilities are equal, $\pi = |S|/n$, we get back the sample mean X.
- The intuition is that if we saw one unit with inclusion probability π_i, there are probably about $1/\pi_i$ others that we didn't see. More formally, we can show that this is an unbiased estimator.

Learn Sampling Theory

- The expectation of $\hat{\mu}_{HT}$, introducing indicator variables Z_i, $i \in 1:n$, which are 1 if $i \in S$ and 0 otherwise.

$$
\begin{aligned}
E\left(\hat{\mu}_{HT}\right) &= E\left(\frac{1}{n}\sum_{i \in S}\frac{X_i}{\pi_i}\right) = E\left(\frac{1}{n}\sum_{i \in 1:n}\frac{X_i}{\pi_i}Z_i\right) \\
&= \frac{1}{n}\sum_{i \in 1:n}\frac{X_i}{\pi_i}E\left(Z_i\right) = \frac{1}{n}\sum_{i \in 1:n}\frac{X_i}{\pi_i}P(Z_i = 1) \\
&= \frac{1}{n}\sum_{i \in 1:n}\frac{X_i}{\pi_i}\pi_i = \frac{1}{n}\sum_{i \in 1:n}X_i \\
&= \mu
\end{aligned}
$$

Learn Sampling Theory

- The variance of the estimator is

$$Var\left(\hat{\mu}_{HT}\right) = \frac{1}{n^2} \sum_{i \in 1:n} \sum_{j \in 1:n} X_i X_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right)$$

  with π_{ij} being the joint inclusion probability, i.2., the probability of including both i and j in the sample (with $\pi_{ii} = \pi_i$).

- Notice that if all the $\pi_i \to 1$, the variance goes to 0.
- We cannot actually calculate this truce variance, since we cannot sum over all the unknown units in the population, but there is a consistent empirical counter-part

$$\hat{Var}\left(\hat{\mu}_{HT}\right) = \frac{1}{n^2} \sum_{i \in 1:n} \sum_{j \in 1:n} X_i X_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - \frac{1}{\pi_{ij}} \right)$$

Strengths and Weaknesses

- The sampling-theory approach works well for stuff you can express as averages (or totals) of population quantities, and where you can work out inclusion probabilities from knowledge of the sampling design.
- Many network statistics can be expressed as averages (sometimes by defining the “unit” as, e.g., a dyad of nodes), but exact calculation of inclusion probabilities is harder.

Missing Data Tools

- Another approach is to treat the unobserved part of the network as missing data, and try to infer it.
- This can range from simple imputation strategies, to complex model-based strategies for inference, such as the EM algorithm.
- Successful imputation or EM is not design-based but model-based, and requires a model both of the network, and of the sampling process.
- It is very, very rare for anything to be “missing at random”, let alone “missing completely at random”.

Model the Effective Network

- A final strategy is to model the observed network. This means modeling both the observation/sampling process and the actual network, but combining them so that we get a family of probability distributions over the observed graph.
- That observed network is (or can be) still informative about the parameters of the underlying generative model.
- If that is all that's of interest, it may be possible to short-circuit the use of EM or imputation, which are more about recovering the full graph.

Big Data Solves Nothing

- Even when, as the promoters say, “n = all”, and the data are automatically recorded (voluntarily or involuntarily), almost all the network sampling issues we’ve gone over remain.
- After all, as the promoters do not say, you’re getting all of a biased convenience sample, not all of the truth.
- Three issues are particularly prominent for network
  1. Entity Resolution
  2. Diffusion
  3. Performativity

Entity Resolution

- **Entity resolution**, or **record linkage**, is a pervasive problem for data analysis.
- Generally speaking, it's the problem of determining when multiple data points all record information about the same thing (or records which are apparently co-referent really are about different things).
- In networks, this is usually about determining when two (or more) apparent nodes really refer to the same underlying entity.

Diffusion

- **Diffusion** refers to the way that many of the automatically-recorded networks which provide us with our big data have themselves spread over other, older social networks.
- What we see when we look at the network of (say) Facebook ties is a combination of the pre-Facebook social network and the results of the diffusion process.
- Comparatively little has been done to understand the results.
- Even if the diffusion process treats all nodes homogeneously, the network-as-diffused can differ radically in its properties from the underlying network.

Performativity

- **Performativity**, the way theories can become (partially) self-fulfilling prophecies. The companies which run online social networks are all very invested in getting very big, very dense networks of users. This is why they all offer link suggestion or link recommendation services.
- The algorithms behind these recommendations implement theories about how social networks form, and what sort of link patterns they should have.
- To the extent that people follow these recommendations, then, the recorded network will seem to conform to the theory.