Inference methods for link-tracing sampling designs

Katie St. Clair

Fall/Winter 2019-20

Prerequisites: Math 245 and 275

Link Tracing Designs

Link-tracing sampling designs are ways of collecting data that exploit a network structure that exists within a population. These network structure often takes the form of social relationships, linking individuals in a population by friendship or familial links. For example, you could start with a small sample of individuals from the target population, then ask these people to recruit their friends or family who are also in the target population into the sample. These new recruits then recruit more individuals, and so on, until a stopping criteria has been met.

Such designs are often used in public health research to assist in studying marginalized or "hard to reach" populations that can be missed using conventional sampling designs due to undercoverage and nonresponse. Examples include studies of immigrant communities and "high risk" populations like intravenous drug users.

Estimation challenge

One challenge of link-tracing sampling methods is that they usually produce data that can't be considered a simple random sample from the population. It is often the case that measurements on linked units are more correlated than on unlinked units. Using standard inference methods that don't account for this correlation will result in biased estimates of population parameters and inaccurate SE estimates. Bias and incorrect SE estimates can also occur with standard methods because units that are highly connected are more likely to be selected than less connected units (e.g. there are unequal sampling probabilities).

For example, consider estimating the proportion of smokers who are denoted as filled-in symbols in the Figure 1 population. Lines between symbols represent friendships. There is a high density of smokers in the highly connected part of the graph, suggesting that smokers tend to have friendships with other smokers. Figure 2 shows a link-tracing sample; the circles represent the first round of sampling and the triangles represent units added after one round of recruitment. Because of the connectedness of smokers in the population, our sample in Figure 2 overrepresents smokers. The population is 27% smokers but the link-tracing sample is 38% smokers. Using the simple sample proportion of smokers will consistently overestimate the proportion of smokers in the population.

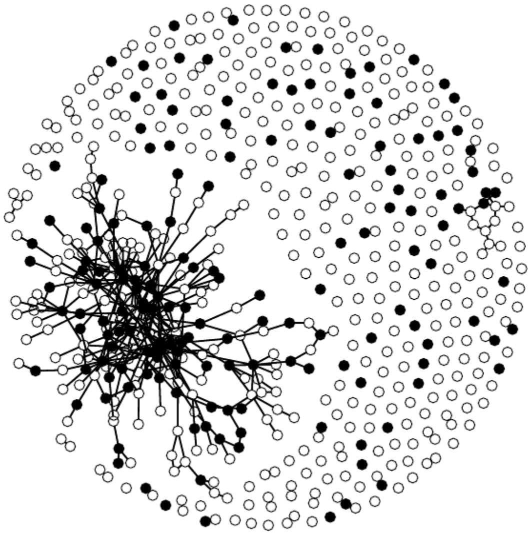

Figure 1: Population

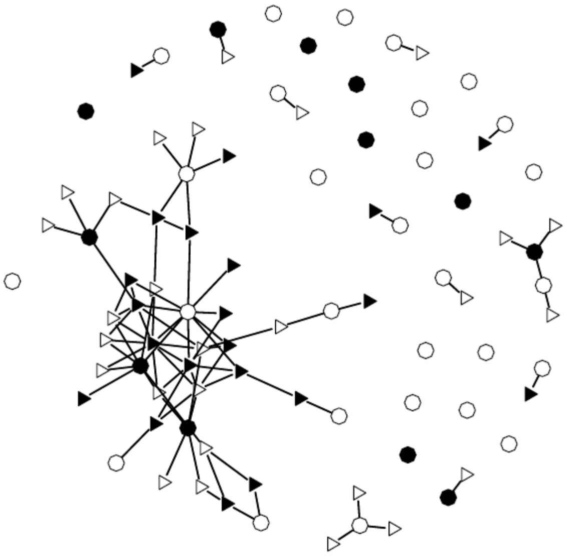

Figure 2: sample with one round of recruitment

The project

The primary idea for this comps project is to simulate network data with certain properties, then carry out the link-tracing sampling on the simulated networks. We can then explore what properties of the population network and what conditions of the sampling process have the greatest impact on biasing our inferences.

What type of inferences we focus on is an open question, though I am most inclined to explore how well standard regression models do at estimating the true population model when using link-tracing data. We could explore whether there are scenarios where a basic regression model can accurately estimate the population model parameters. For example, can a logistic regression model fit using the data collected in Figure 2 really tell us what factors are associated with smoking in the population shown in Figure 1?

We could also study more advanced modeling methods that may be able to account for the correlated structure in link-tracing data. Or we could develop our methodological framework for modeling associations using link-tracing data.

Expectations

I am looking for everyone in the group to have taken Math 245 and 275. The group would also benefit from anyone with background in sampling (Math 255), Bayesian inference (Math 315), or any experience working with network data. The "final product" for this comps will be submission of a 20 page paper to the Undergraduate Research Project Competition (https://www.causeweb.org/usproc/usresp). Since this is a ongoing area of research, there is also the possibility of a paper for publication in a research journal if the simulation results are interesting or a new method for modeling is derived!