Developing a methodology for social network sampling

Daniel W. Franks[1], Richard James[2], Jason Noble[3] and Graeme Ruxton[4]

[1]University of York
[2]University of Bath
[3]University of Southampton
[4]University of Glasgow
df525@york.ac.uk

Researchers are increasingly turning to network theory to describe and understand the social nature of animal populations. To make use of the statistical tools of network theory, ecologists need to gather relational data, typically by sampling the social relations of a population of animals over a given time-period. Due to effort constraints and the practical difficulty involved in tracking animals, these sampled relational data are almost always a subset of the actual network. Measurements of the sample – such as average path length, clustering, and assortativity – are assumed to be informative as to the structure of the real-world social network. However, this assumption is problematic. Due to artefacts of the sampling process, the various network measures taken on the sample may be biased estimators of the true values. For example, just as we would get a biased estimate of mean human height by selecting for a sample those people who stood out in a crowd, we will get a biased estimate of a measure like mean connectivity if we sample individuals who are socially prominent.

This problem can only be solved by developing a qualitative theory of network sampling, answering questions such as what proportion of the whole network needs to be sampled before a given level of accuracy is achieved, and what sampling procedures are least biased? To develop such a theory, we need to be able to generate networks from which to sample. Ideally, we need to perform a systematic study of sampling protocols on different known network structures. But currently available data on animal social networks are unsuitable as these networks were themselves sampled.

The simulation methods of artificial life provide the way forward. We have developed a computational tool for generating artificial social networks that have user-defined distributions for network properties (such as the number of nodes, and the density) and for key the measures of interest to ecologists (such as the average degree, average path length, clustering, betweenness, and assortativity). This tool allows us to perform the required systematic analyses of the biases inherent in different sampling regimes (e.g., snowball sampling) applied to different network structures. We will present details of this system, and show we are using it to develop robust sampling methods for social network data. We see the system as the first in a series of works that will allow us to develop a qualitative theory of social network sampling to aid ecologists, and eventually social scientists, in their social network data collection.