

Sampling Networks and the Inference of Network Characteristics

Eric D. Kolaczyk
Department of Mathematics and Statistics
Boston University
111 Cummington Street
Boston, MA 02215
kolaczyk@math.bu.edu

ABSTRACT

In the study of complex networks, a substantial amount of work has focused on the summary description of characteristics of measured network graphs (e.g., summaries of size, diameter, degree, clustering, betweenness, etc.). Findings based on such work have in turn been used for a variety of purposes, including the motivation of a wide variety of models aimed at reproducing certain of the observed characteristics. However, as in any measurement process, network measurement is open to sampling-based biases and errors, and these may have the effect of rendering summary characteristics of measured networks unrepresentative of the same characteristics in the underlying network that was sampled. As a result, recent efforts have begun exploring in some depth the implications of sampling networks and, to a lesser extent, ways in which the adverse effects of bias and error may be alleviated through modification of the sampling and/or the statistical summaries used.

In this lecture I aim to provide an overview of the relevant principles and issues, as well as the tools for addressing them. I will focus primarily on the effects of bias in sampling. We will review elements of the basic theory and methods of statistical sampling and construction of estimators. We will then look at the problem of sampling networks from this perspective, starting with results from an older body of work in the social network literature and ending with results of the past few years in the complex network literature. I will finish by presenting recent work with colleagues for constructing estimators of network characteristics that seek to correct for certain sampling biases, primarily in the context of Traceroute sampling in the Internet.

Keywords

Complex network, estimation, sampling bias, species.