Statistics > Methodology

Estimating Network Degree Distributions Under Sampling: An Inverse Problem, with Applications to Monitoring Social Media Networks

Yaonan Zhang, Eric D. Kolaczyk, Bruce D. Spencer

(Submitted on 21 May 2013)

Networks are a popular tool for representing elements in a system and their interconnectedness. Many observed networks can be viewed as only samples of some true underlying network. Such is frequently the case, for example, in the monitoring and study of massive, online social networks. We study the problem of how to estimate the degree distribution -- an object of fundamental interest -- of a true underlying network from its sampled network. In particular, we show that this problem can be formulated as an inverse problem. Playing a key role in this formulation is a matrix relating the expectation of our sampled degree distribution to the true underlying degree distribution. Under many network sampling designs, this matrix can be defined entirely in terms of the design and is found to be ill-conditioned. As a result, our inverse problem frequently is ill-posed. Accordingly, we offer a constrained, penalized weighted least-squares approach to solving this problem. A Monte Carlo variant of Stein's unbiased risk estimation (SURE) is used to select the penalization parameter. We explore the behavior of our resulting estimator of network degree distribution in simulation, using a variety of combinations of network models and sampling regimes. In addition, we demonstrate the ability of our method to accurately reconstruct the degree distributions of various sub-communities within online social networks corresponding to Friendster, Orkut, and LiveJournal. Overall,

our results show that the true degree distributions from both homogeneous and inhomogeneous networks can be recovered with substantially greater accuracy than reflected in the empirical degree distribution resulting from the original sampling.