Estimation of global network statistics from incomplete data

Catherine A. Bliss, Christopher M. Danforth, Peter Sheridan Dodds

Computational Story Lab, Department of Mathematics and Statistics, Vermont Complex Systems Center
& the Vermont Advanced Computing Core, University of Vermont, Burlington, VT, 05405

Abstract

Complex networks underlie a variety of social, biological, physical, and virtual systems. In many settings, it is impossible to observe all nodes and all network interactions. Previous work addressing the impacts of partial network data, which is surprisingly limited and focuses primarily on missing nodes, suggests that network statistics derived from subsampled data are not suitable plug in estimators for network statistics describing the overall network topology. Our aim is to generate scaling methods to predict true network parameters from only partial knowledge of nodes, links, or weights. We validate analytical results on four simulated network classes (Erdös-Rényi, Scale-free, Small World, and Range dependent networks) each with $N = 2 \times 10^5$ and $k_{\text{avg}} = 10$ and empirical data sets of various sizes. We perform 100 subsampling experiments by varying proportions of sampled data and demonstrate that our scaling methods provide very good estimates of the true network parameters. Lastly, we apply our techniques to a set of rich and evolving large-scale social networks, Twitter reply networks. From over 100 million tweets, we use our scaling techniques to propose a statistical characterization of the Twitter interactome from September 2008-February 2009.

February 25, 2014