Spectral Estimation of Conditional Random Graph Models for Large-Scale Network Data

Antonino Freno  Mikaela Keller∗  Gemma C. Garriga  Marc Tommasi∗
INRIA Lille – Nord Europe, Villeneuve d’Ascq (France)
{antonino.freno, mikaela.keller, gemma.garriga, marc.tommasi}@inria.fr

Abstract

Generative models for graphs have been typically committed to strong prior assumptions concerning the form of the modeled distributions. Moreover, the vast majority of currently available models are either only suitable for characterizing some particular network properties (such as degree distribution or clustering coefficient), or they are aimed at estimating joint probability distributions, which is often intractable in large-scale networks. In this paper, we first propose a novel network statistic, based on the Laplacian spectrum of graphs, which allows to dispense with any parametric assumption concerning the modeled network properties. Second, we use the defined statistic to develop the Fiedler random graph model, switching the focus from the estimation of joint probability distributions to a more tractable conditional estimation setting. After analyzing the dependence structure characterizing Fiedler random graphs, we evaluate them experimentally in edge prediction over several real-world networks, showing that they allow to reach a much higher prediction accuracy than various alternative statistical models.

1 INTRODUCTION

Arising from domains as diverse as bioinformatics and web mining, large-scale data exhibiting network structure are becoming increasingly available. Network models are commonly used to represent the relations among data units and their structural interactions. Recent studies, especially targeted at social network modeling, have focused on random graph models of those networks. In the simplest form, a social network is a configuration of binary random variables $X_{uv}$ such that the value of $X_{uv}$ stands for the presence or absence of a link between nodes $u$ and $v$ in the network. The general idea underlying random graph modeling is that network configurations are generated by a stochastic process governed by specific probability laws, so that different models correspond to different families of distributions over graphs.

The simplest random graph model is the Erdős-Rényi (ER) model [Erdős and Rényi, 1959], which assumes that the probability of observing a link between two nodes in a given graph is constant for any pair of nodes in that graph, and it is independent of which other edges are being observed. In preferential attachment models [Barabási and Albert, 1999], the probability of linking to any specified node is proportional to the degree of the node in the graph, leading to “rich get richer” effects. Small-world models [Watts and Strogatz, 1998] try to capture instead some phenomena often observed in real networks such as high clustering coefficients and small diameters [Newman, 2010]. A sophisticated attempt to model complex dependences between edges in the form of Gibbs-Boltzmann distributions is made by exponential random graph (ERG) models [Snijders et al., 2006], which subsume the ER model as a special case. Finally, a recent attempt at modeling real networks through a stochastic generative process is made by Kronecker graphs [Leskovec et al., 2010], which try to capture phenomena such as heavy-tailed degree distributions and shrinking diameter properties while paying attention to the temporal dynamics of network growth.

While some of these models behave better than others in terms of computational tractability, one basic limitation affecting all of them is what we may call a parametric assumption concerning the probability laws underlying the observed network properties. In other words, currently available models of network structure assume that the probability distribution gener-