Generative models for social network data

Kevin S. Xu and James R. Foulds

Abstract

Due in part to the ubiquity of social network data today, interest in social network analysis has spread beyond its traditional home in the social sciences to many other disciplines including physics, computer science, statistics, and engineering. A topic of significant interest in social network analysis is the creation of statistical models for social network data. Many of these models are *generative*, which allows one to simulate random networks from a particular model. Additionally many models share a common structure: each vertex is assigned a set of latent or hidden attributes, and edges between vertices are generated with probability conditional on the hidden attributes of the vertices. These latent variables, which are automatically inferred from the network, can be valuable for understanding network structure with respect to sociological principles, and for making predictions about the network's current and future state.

In this tutorial, we cover four main classes of generative models for social network data, under which many of the commonly used statistical network models fall:

- *Latent space models*, which generally assume latent continuous attributes for vertices where the probability of an edge between two vertices is given by a distance function applied to the attributes of the vertices.
- *Block models*, which divide vertices into one of k latent classes where the probability of an edge between two vertices depends only on the classes of the vertices.
- *Latent feature models*, which allow vertices to have arbitrarily many unique (typically binary) features, where the probability of an edge between two vertices is given by a weighted sum of the elements of their feature vectors.
- *Mixed membership models,* in which each vertex has partial membership in their latent classes.

We discuss some of the challenges when it comes to applying these types of generative models on social network data, including

- Optimization methods to fit these generative models, which involve estimating the latent attributes of the vertices, in an optimal or near-optimal manner.
- Simulation approaches for Bayesian inference in these models using Markov chain Monte Carlo.
- Model selection and verification to validate a particular fit to a social network model.
- Interpretation of model parameters and their relationship to social network structure.

We conclude with an overview of recent developments on generative models for social network data, including models for a *collection* of networks, which can be used to represent relations at different times (dynamic networks) or different types of relations (multi-layer networks).

Expected audience

This tutorial should be applicable to attendees with interests in social network analysis from a statistical perspective and backgrounds in any of the topic areas covered by SBP-BRiMS, including behavioral and social sciences, public health, and computer and information sciences.

Biosketch

Kevin S. Xu received the B.A.Sc. degree in Electrical Engineering from the University of Waterloo in 2007 and the M.S.E. and Ph.D. degrees in Electrical Engineering: Systems from the University of Michigan in 2009 and 2012, respectively. He was a recipient of the Natural Sciences and Engineering Research Council of Canada (NSERC) Postgraduate Master’s and Doctorate Scholarships. He is currently an assistant professor in the EECS Department at the University of Toledo and has previously held industry research positions at Technicolor and 3M. His main research interests are in machine learning and statistical signal processing with applications to network science and human dynamics.

James (a.k.a. Jimmy) Foulds is a postdoctoral scholar at the University of California, San Diego. His research interests are in both applied and foundational machine learning, focusing on probabilistic latent variable models and the inference algorithms to learn them from social networks and text data. His work aims to promote the practice of latent variable modeling for applied research in disciplines such as computational social science and the digital humanities. He earned his PhD in computer science at the University of California, Irvine, and was a postdoctoral scholar at the University of California, Santa Cruz. His master's and bachelor's degrees were earned with first class honours at the University of Waikato, New Zealand, where he also contributed to the Weka data mining system.