Library & Information Science Research 38 (2016) 259–264

Contents lists available at ScienceDirect

Library & Information Science Research

State of the art review

An evaluation of exponential random graph modeling and its use in library and information science studies

Ana Dubnjakovic

College of Communications, Department of Library and Information Sciences, University of South Carolina, 800 Sumter Street, Columbia, SC 29208, USA

ARTICLE INFO

Article history:
Received 25 November 2014
Received in revised form 8 January 2016
Accepted 29 August 2016
Available online 6 September 2016

ABSTRACT

Social network analytical tools and theories have long been an accepted part of the research landscape in many social and physical sciences including: sociology, political science, psychology, communications, business, geography, biology, physics, and chemistry as well as library and information science (LIS). Given the level of activity in the social network analysis (SNA) area concerning LIS, it is important to review the latest trends in the SNA stochastic modeling, namely exponential random graph (ERG) models. Unlike previous SNA methods, ERG models offer insight into generative network properties through simultaneous inclusion of structural parameters and attributes in the analysis while accounting for the interdependent nature of network data. Additionally, when Monte Carlo Markov Chain Maximum Likelihood Estimator is used, ERG modeling results in parameter estimates superior to other methods (e.g., MRQAP). The current study will demonstrate the utility of ERG models in LIS through a brief overview of major concepts and techniques in SNA, followed by a detailed description of ERG modeling technique, a review of currently available software used in analysis and a brief examination of its current use in LIS studies.

1. Introduction

Social network analytical tools and theories have long been an accepted part of the research landscape in many social and physical sciences including: sociology, political science, psychology, communications, business, geography, biology, physics, chemistry as well as library and information sciences (LIS). At least two studies in major information sciences journals synthesize the social network analysis impact on the field of library and information sciences. The first examines, in some detail, techniques and theories of social network analysis (SNA) in the context of information exchange studies, including its unique qualities distinguishing it from other research approaches (Haythornthwaite, 1996). Schultz-Jones (2009) provides an update of this study in terms of recent research questions and a detailed literature review of seven disciplines, including LIS, using the SNA approaches to examine information behavior. This study also distinguishes between SNA theory and analytical tools and provides an overview of history and important developments in both. Additionally, recent LIS studies in the areas of bibliometrics, webometrics, knowledge management and user information behavior indicate interest the SNA stochastic modeling. Given the level of activity in the SNA area concerning LIS as demonstrated in the two studies as well as various content areas, it is important to review the latest trends in the SNA stochastic modeling, namely the exponential random graph (ERG) models.

Since ERG modeling requires substantial background in statistics and SNA, the aim of this study is to demystify the technique and promote its usage in LIS studies by demonstrating its unique value and advantages over other SNA descriptive and stochastic analytical tools. In keeping with this goal, the current study will demonstrate the utility of ERG models in LIS through a brief overview of major concepts and techniques in SNA, followed by a detailed description of ERG modeling technique, a review of currently available software used in analysis and a brief examination of its current use in LIS studies.

2. Literature review

2.1. SNA analytical procedures

SNA relies on relational data consisting of nodes, sometimes also referred to as actors, and connecting ties, also known as edges, which can be directed or non-directed (Wasserman & Faust, 1994). Resulting networks can be viewed from a single actor's perspective, termed egocentric, or a whole network perspective focusing on ties as reported by the entire set of actors. In information exchange studies, egocentric networks can provide information about who the actor goes to for information and where they receive it, while the whole network provides insight into information behavior of groups of actors (Haythornthwaite, 1996). Networks can also be of a one-

E-mail address: ana@mailbox.sc.edu.

http://dx.doi.org/10.1016/j.lisr.2016.08.005
0740-8188/

mode variety consisting of a single set of similar nodes (e.g., a set of authors) or two-mode variety consisting of two sets of nodes with ties connecting the two sets (e.g., a set of authors and a set of articles where the author is tied to an article if they are listed as an author). Data are commonly presented in graphs or adjacency matrices suitable for further analysis. An adjacency matrix is a square matrix with as many rows or columns as nodes in the network where the intersections between rows and columns indicate the presence or absence of ties (i.e., number one indicates a tie is present and a zero tie is absent). In undirected graphs (i.e., where all ties are reciprocated), the adjacency matrix is symmetrical, with zeros on the diagonal to avoid loops (i.e., ties to self). Consequently, directed graphs or digraphs are asymmetrical since ties are not necessarily reciprocated.

Before recent developments in ERG models were introduced, the majority of SNA techniques mostly focused on descriptions of network properties such as density, in-degree, out-degree, size, centralization/centrality and distance (Wasserman & Faust, 1994). Each of these provides valuable insight into how networks operate. Density provides an account of the level of connectivity of actors in a network, in-degree and out-degree detail numbers of connections attracted and radiating from an actor respectively, while distance provides a measure of actors' reachability. Centrality measures an individual actor's influence in the network while centralization measures the extent to which actors in a network are organized around a central node. Degree centrality posits higher centrality for the actors with the most ties. Closeness centrality assigns it to the actors who are most easily reached, while betweenness centrality considers those with the highest probability of occurring on the shortest path between two randomly chosen nodes to be the most central (Borgatti & Everett, 2006).

Additional advanced SNA analytical approaches concerning whole networks are grouped around notions of cohesion (i.e., the process of grouping actors according to common characteristics), brokerage (i.e., the information diffusion in connection with centralization and bridges), ranking as assessed through measures of prestige, dyadic and triadic analysis and positions as revealed through the blockmodeling procedure (i.e., grouping of structurally equivalent actors into clusters). With few exceptions (e.g., stochastic blockmodels and the quadratic assignment procedure), the advanced SNA analytical tools are best suited to describing properties of networks and fall short when it comes to describing their generative properties. Indeed, ample evidence suggests that most network studies conducted prior to 2003 focus on consequences rather than generative properties of networks (Schultz-Jones, 2009). Given the complexity of human information behaviors, it is reasonable to expect that behaviors giving rise to complex processes such as, for instance, information exchange networks are stochastic and that statistical analyses are needed.

Prior to recent advances in ERG modeling, the few statistical studies conducted mostly focused on inclusion of some of the aforementioned network concepts into the statistical models as independent variables along with other continuous attribute variables relying on techniques such as regression (e.g., Gest, Graham-Bermann, & Hartup, 2001; Oliver & Montgomery, 1996; Reich, 2007). Even from this brief discussion of social network data, it should be immediately apparent that, due to their relational nature, network data violate independence assumptions associated with such analytical tools thereby making the inferences questionable. The problem lies in standard error computation which relies on error variance. When actors are chosen in groups rather than as individuals, the possibility of correlated disturbances increases, making the coefficients unreliable (Allison, 1999). Permutation based regression based on Krackhardt's quadratic assignment procedure (QAP) (Krackhardt, 1987), where the rows and columns in the adjacency matrix are permuted simultaneously in such a way that the network structure is left intact (Snijders, 2011), emerges as the most popular solution. The sampling distribution is generated from the possible combinations of the sample space and the observed statistic is compared to a simulated distribution (Schaeffer, 2012).

Before QAP was introduced, very few options were available to researchers who wished to include attributes in their analysis along with network structural properties. Consequently, the temptation to include structural network properties as independent variables along with other independent continuous variables is understandable. However, even QAP as the earliest form of true stochastic network analysis suffers from notable faults. For instance, while multiple regression QAP (MRQAP) extends the QAP to include examination of more than two relations, both require data manipulation and make no attempt to model network dependencies (Snijders, 2011). In contrast, ERG models provide a statistical framework capable of directly manipulating network data and attributes associated with actors resulting in less statistical noise.

2.2. Exponential random graph (p^*) models

The basic assumption underlying the logic of ERG models is that the observed network is a result of some unknown stochastic process. The proposed model aims to explain this stochastic process by testing a set of hypotheses derived from theory or prior research and represented by the structural properties of the observed network (Robins, Pattison, Kalish, & Lusher, 2007). Specifically, ERG models test whether a generative process in a network occurs more frequently than expected by chance. Broadly speaking, these generative processes can be explained through network self-organization processes characterized by activity/popularity, reciprocity, closure and brokerage, actor attributes characterized by effects of the sender, effects of the receiver and their interaction, and exogenous contextual factors such as other networks or special factors (Lusher, Koskinen, & Robins, 2013). For instance, researchers could test if preferential attachment (i.e., new actors link to actors with high indegrees) can be modeled by including appropriate parameters into the model.

When exogenous factors and attributes are not included, structural configurations shape the form of the model. The simplest structural configuration in a directed network is an arc. Higher order parameters include star and triangle configurations. Stars generally reflect the degree distribution in a network and range from 2 stars to k stars. In directed networks, they can appear in the form of in-stars (i.e., all nodes are connected to the central node but not to each other), out-stars (i.e., the central node is connected to other nodes but they are not connected to each other) and mixed stars (i.e., some combination of in-stars and out-stars). Degree distribution reflects popularity and activity effects. For example, in the context of information exchange, degree distribution could provide valuable information about the existence of hubs (i.e., nodes receiving multiple ties) that play an important role in the way information is transferred through the network. Triangles reflect the process of closure, appearing in directed networks as transitive triangles where a node's connection to two other nodes increases the likelihood those nodes will be connected (e.g., a friend of my friend is a friend), and cyclic triangles where ties are unidirectional. In transitive triads, one node receives 2 ties, one node sends 2 ties and one receives and sends one tie. When modeling information exchange networks, the prevalence of such structures could indicate that a node receiving 2 ties has the most valuable information. In undirected networks, triadic relationships appear in 4 possible configurations (i.e., no ties, one tie, two ties, or all three ties). Dominance of any of these configurations indicates to what degree the nodes in that network are isolated, appear in couples, structural holes (i.e., when a node is connected to 2 other nodes but those nodes are not connected to each other) or clusters.

Attributes, in network parlance, represent individual characteristics of actors and can be dichotomous, categorical, and/or continuous. In ERG models, attributes are considered exogenous and network structural properties endogenous to the model. If the actors' attributes affect their involvement in the network in such a way that they become more active (e.g., similar actors might share more information) those effects are known as sender effects. If, in turn, they become more popular