Social Networks 39 (2014) 84–97

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors

Zachary Neal*

Psychology Department, 316 Physics Road, Michigan State University, East Lansing, MI 48824, United States

ARTICLE INFO

Keywords:
Bill co-sponsorship
Binarize
Dichotomize
Political network
Projection
Two-mode

ABSTRACT

The analysis and visualization of weighted networks pose many challenges, which have led to the development of techniques for extracting the network's backbone, a subgraph composed of only the most significant edges. Weighted edges are particularly common in bipartite projections (e.g. networks of co-authorship, co-attendance, co-sponsorship), which are often used as proxies for one-mode networks where direct measurement is impractical or impossible (e.g. networks of collaboration, friendship, alliance). However, extracting the backbone of bipartite projections requires special care. This paper reviews existing methods for extracting the backbone from bipartite projections, and proposes a new method that aims to overcome their limitations. The stochastic degree sequence model (SDSM) involves the construction of empirical edge weight distributions from random bipartite networks with stochastic marginals, and is demonstrated using data on bill sponsorship in the 108th U.S. Senate. The extracted backbone's validity as a network reflecting political alliances and antagonisms is established through comparisons with data on political party affiliations and political ideologies, which offer an empirical ground-truth. The projection and backbone extraction methods discussed in this paper can be performed using the -onemode- command in Stata.

1. Introduction

Analyzing and visualizing weighted networks presents a number of challenges, which has lead to the development of methods for extracting the 'backbone' of these networks. Such backbone extraction methods aim to reduce the original, weighted network into a simpler, binary network that preserves only those edges whose weights are sufficiently large to suggest they are significant. The challenge lies in determining how strong an edge's weight must be before deeming it significant. Several techniques for assessing a edge's significance and thus achieving this reduction have been proposed, ranging from relatively simple methods like an unconditional threshold that retains the strongest edges to more sophisticated methods that compare observed edge weights to expectations from a null model (Serrano et al., 2009) or to empirical distributions (Foti et al., 2011).

Weighted networks arise in many different contexts, but are particularly common in the case of bipartite network projections, including networks where authors are linked by the number of papers they have co-authored (e.g. de Stefano et al., 2013) or actors are linked by the number of movies they have both appeared in (e.g. Watts and Strogatz, 1998). However, the backbone extraction methods developed for natively one-mode networks are not well suited for one-mode networks that have been obtained from bipartite data via projection. Such methods fail to incorporate information present in the original bipartite data into their decisions about whether a given edge should be preserved in the backbone network. Several alternative backbone extraction methods have been developed specifically for bipartite projections (e.g. Zweig and Kaufmann, 2011; Neal, 2013), but they are computationally complex and risk imposing too many or too few assumptions. The purpose of this paper is to review the existing methods for extracting the backbone from bipartite projections, then to propose and demonstrate a new method that aims to overcome some of the existing methods' limitations.

After defining bipartite networks and some key features of their projections, I briefly review the most commonly used methods for extracting the backbone of bipartite projections, noting their strengths and weaknesses. I then describe a new method – the stochastic degree sequence model (SDSM) – that involves building empirical probability distributions using a sample of random bipartite networks with stochastic row and column degree sequences. In Section 5, I provide a step-by-step demonstration of this method using data on bill sponsorship activities in the 108th U.S. Senate

* Tel.: +1 517 432 1811.
E-mail address: zpneal@msu.edu

http://dx.doi.org/10.1016/j.socnet.2014.06.001
0378-8733/

to infer political alliances and antagonisms among senators. In this context, the SDSM involves asking whether two senators co-sponsored significantly more bills (suggesting an alliance) or significantly fewer bills (suggesting an antagonism) than they might have co-sponsored in plausible alternate worlds in which the senators randomly sponsored roughly the same number of bills and the bills were randomly sponsored by roughly the same number of senators. It yields a backbone network of political alliances and antagonisms that exhibits a high level of criterion validity when compared to expectations based on political party and ideology data, which offer an empirical ground-truth. The paper concludes with a discussion of the proposed method's limitations and directions for future research on the analysis of bipartite projections.

2. Bipartite networks and projections

A bipartite network is composed of two mutually exclusive sets of nodes; edges may exist between nodes in different sets, but not between nodes in the same set. Also known as two-mode or affiliation networks, bipartite networks have been discussed in many different contexts including southern women attending social events (Davis et al., 1941), individuals sitting on corporate boards (Mizruchi, 1996), actors appearing in movies (Watts and Strogatz, 1998), world cities hosting branches of multinational firms (Taylor, 2001), supreme court justices joining majority opinions (Doreian et al., 2004), legislators sponsoring bills (Fowler, 2006a), and ingredients possessing flavor compounds (Ahn et al., 2011). Formally defined, an m-by-n bipartite network, $\mathbf{B}$, in which $B_{ik} = 1$ if there is an edge between i and k and otherwise is zero, can be projected onto an m-by-m unipartite or one-mode network, $\mathbf{P}$, as $\mathbf{BB}'$ (Breiger, 1974). Using this approach, for example, a bipartite network that describes legislators' sponsorship of bills is transformed into a unipartite or one-mode network of legislators linked to one another by their co-sponsorship of bills.

To facilitate a discussion of bipartite projections, some generic terminology is useful. Throughout this paper, I use the terms *agent* and *artifact* to describe the two sets of nodes in a bipartite network. Agents are represented as rows in $\mathbf{B}$ and are the primary nodes of interest. An agent's degree is the row marginal of $\mathbf{B}$, and indicates an agent's total number of artifacts, for example, how many social events (the artifacts) a given person (an agent) attended or how many bills (the artifacts) a given legislator (an agent) sponsored. Artifacts are represented as columns in $\mathbf{B}$ and are instrumental in forging the linkages between agents in $\mathbf{P}$, but are not of direct interest in the bipartite projection. An artifact's degree is the column marginal of $\mathbf{B}$, and indicates an artifact's total number of agents, for example, how many people (the agents) attended a given social event (an artifact) or how many legislators (the agents) sponsored a given bill (an artifact).

Much has already been written about the mathematical properties of bipartite projections (see Latapy et al., 2008), however the nature of edge weights in bipartite projections is of particular concern in the methods discussed below, and thus warrants brief consideration. The weight of an edge in the projection, P_{ij}, reflects the number of artifacts that agents i and j have in common (e.g. the number of bills two legislators both sponsored). Some have argued that bipartite projections are easier to analyze than the original bipartite network because they are one-mode networks, noting that "there is no need to develop any new techniques to analyze [bipartite projections]. . .for which the full range of network analytic methods are available" (Borgatti and Everett, 1997, p. 246). However, because bipartite projections are nearly always weighted networks, their analysis is not as straightforward as this claim implies. Projecting a bipartite network into a one-mode network merely "transforms the problem of analysing a bipartite structure into the problem of analysing a weighted one, which is not easier" (Latapy et al., 2008, pp. 34–35). One important but rarely noted feature of these edge weights is their constrained range of possible values. The range of values an edge between agents i and j may take in a bipartite projection can be expressed as a function of these agents' degrees (i.e. D_i and D_j) and the total number of artifacts (A):

$$\min(D_i, D_j) - (A - \max(D_i, D_j)) \leq P_{ij} \leq \min(D_i, D_j) \quad (1)$$

A simple example serves to illustrate. Suppose Tom attends 5 of 10 parties, and Jerry attends 7 of the same 10 parties. From Eq. (1), we know that Tom and Jerry must have co-attended at least 2 parties, and could not have co-attended more than 5 parties. A critical implication of this identity is that, *ceteris paribus*, higher-degree agents will necessarily have stronger edges than lower-degree agents.

Before turning to methods for dealing with these edge weights, it is also useful to consider why one would examine a bipartite projection at all. Indeed, the projection transformation involves the loss of information including the specific identity of the artifacts responsible for forging linkages between agents (Latapy et al., 2008), and methods are emerging for analyzing bipartite networks without requiring their projection (Borgatti and Everett, 1997; Agneessens and Everett, 2013). Nonetheless, bipartite projections remain an important methodological tool in research where the interest is in a natively one-mode network, but where measurement of this network is impossible or impractical. In developmental psychology research on peer relationships among youth, high non-response rates and challenges associated with obtaining parental permission required by Institutional Review Boards make the direct collection of one-mode social network data is difficult. As a solution, a method known in this literature as Social Cognitive Mapping uses bipartite projections in which children are linked by their co-participation in social groups to infer the unobserved social network of interest (e.g. Cairns and Cairns, 1994; Gest et al., 2007; Neal and Neal, 2013). Similarly, in political science research on relationships of political alliance and influence, politicians' compelling strategic reasons for wanting to conceal their alliances makes direct collection of such data impossible. As a solution, some have turned to bipartite projections reflecting bill co-sponsorship or committee co-membership to infer the unobserved social network of interest (e.g. Porter et al., 2005; Fowler, 2006a). Finally, in geography research on global economic relations between cities, although some types of data exist on trade and foreign direct investment between countries, no such data exists at the city level. As a solution, a method known in this literature as the Interlocking World City Network Model uses bipartite projections in which cities are linked by the co-location of branches of advanced producer service firms (e.g. institutional banks, law firms, accounting agencies, etc.) to infer the unobserved economic network of interest (e.g. Taylor, 2001; Neal, 2008). In each case, a bipartite projection is used as a proxy to infer an unobserved, natively one-mode network of interest. When used as a proxy measurement tool, the methods for handling edge weights in bipartite projections must permit such inferences to be made in a principled way.

3. Existing methods for backbone extraction

All methods of network backbone extraction, whether applied to natively one-mode networks or to bipartite projections, involve the use of a threshold. Edges whose weights exceed the threshold value are retained in the backbone, while those whose weights are below the threshold value are omitted from the backbone. Backbone extraction methods vary, however, in how threshold values are identified. In this section, I review three broad approaches that can be distinguished by the information on which the selection of threshold values is conditioned. Table 1 summarizes examples of