Comparing Models for Extracting the Backbone of Bipartite Projections

Zachary P. Neal[1,*], **Rachel Domagalski**[2], **and Bruce Sagan**[2]

[1]Michigan State University, Psychology Department, East Lansing MI, USA
[2]Michigan State University, Mathematics Department, East Lansing MI, USA
[*]zpneal@msu.edu

ABSTRACT

Projections of bipartite or two-mode networks capture co-occurrences, and are used in diverse fields (e.g., ecology, economics, bibliometrics, politics) to represent unipartite networks that would otherwise be difficult or impossible to measure directly. A key challenge in analyzing such networks is determining whether an observed number of co-occurrences is significant. Several models now exist for doing so and thus for extracting the backbone of bipartite projections, but they have not been directly compared to each other. In this paper, we compare five such models – fixed fill model (FFM) fixed row model (FRM), fixed column model (FCM), fixed degree sequence model (FDSM), and stochastic degree sequence model (SDSM) – in terms of accuracy, speed, statistical power, similarity, and community detection. We find that the computationally-fast SDSM offers a statistically conservative but close approximation of the computationally-impractical FDSM under a wide range of conditions, and that it correctly recovers a known community structure even when the signal is weak. Therefore, although each backbone model may have particular applications, we recommend SDSM for extracting the backbone of most bipartite projections.

Introduction

Bipartite or two-mode networks are composed of two types of nodes, which we call *agents* and *artifacts*, and edges between nodes of one type and nodes of the other type. They can be used to represent a wide range of phenomena and therefore are studied in a diverse range of disciplines. For example, natural selection unfolds as species (the agents) compete over sites (the artifacts), commerce is possible as traders exchange resources, scientific advances are reported as scholars write papers, and laws are adopted as legislators sponsor bills. Although bipartite networks are useful in their own right, they can also be useful for inferring unipartite (i.e., one-mode) networks that would otherwise be difficult or impossible to measure directly. A bipartite projection transforms a bipartite network into a unipartite co-occurrence network in which agents are connected to the extent that they share artifacts. For example, competitive interaction networks can be inferred from species' co-occurrence in sites,[1] trade networks can be inferred from firm co-location[2] or product co-exchange,[3] scholarly collaboration networks can be inferred from paper co-authorship,[4] and political alliance networks can be inferred from bill co-sponsorship.[5] Throughout the paper we use these applications to offer concrete examples, however the models we discuss are perfectly general and can be applied to derive unipartite backbones in a range of contexts.[6–8] Indeed, in principle any unipartite network can be represented as the projection of some bipartite network.[9–11]

Despite their promise, bipartite projections (i.e., co-occurrence networks) are challenging to analyse because they are typically dense and weighted, and because the edge weights do not necessarily capture the strength of the relationship between nodes.[12] As a result, it is often useful to analyze the *backbone* of a bipartite projection, which is an unweighted and typically sparser network that retains only the most 'important' edges. Although well-known methods exist for extracting the backbone of weighted networks that are not bipartite projections,[13,14] methods designed specifically for bipartite projections have recently been developed.[12,15–17] However, relatively little is known about the similarities and differences of methods for extracting the backbone of bipartite projections. In this paper, we present five such methods – fixed fill model (FFM) fixed row model (FRM), fixed column model (FCM), fixed degree sequence model (FDSM), and stochastic degree sequence model (SDSM) – and conduct four related studies to better understand them.

The paper is organized in six sections. We begin by formally defining bipartite projections, backbones, and the five backbone models, presenting proofs of the probability mass functions for their respective edge weight distributions in the *Supplementary Information*. In study 1, we evaluate the accuracy and speed of different approaches for estimating cell-filling probabilities used by the SDSM. In study 2, we evaluate the statistical power of the SDSM relative to the FDSM. In study 3, we examine how degree distributions impact the similarity of backbones extracted using different models. In study 4, we examine the extent to which backbones extracted using different models accurately recover a known community structure. Finally, we conclude with recommendations for backbone model selection and opportunities for future model development.

Backbone extraction for bipartite projections

Preliminaries

A *bipartite network* captures connections between nodes of one type (*agents*) and nodes of a second type (*artifacts*). Throughout this section, we use the ecological case of Darwin's Finches to provide a concrete example.[18,19] On his voyage to the Galapagos Islands on the H.M.S. Beagle, Darwin observed that only some species of finches lived on each island. These patterns can be represented as a bipartite network in which finch species (the agent nodes) are connected to the islands (the artifact nodes) where they are found.[20] A bipartite network can be represented as a binary matrix in which the agents are arrayed as rows, and the artifacts are arrayed as columns. We use $\mathbf{B}$ to denote a bipartite network's representation as a matrix, where $B_{ik} = 1$ if agent i is connected to artifact k, and otherwise is 0. The sequence of row sums and the sequence of column sums of $\mathbf{B}$ are called the agent and artifact degrees sequences, respectively. These sequences are among the bipartite network's most significant features and are known to have implications for bipartite projections and backbones.[9,21,22] In the ecological case, the agent degree sequence captures the number of islands where each species is found, while the artifact degree sequence captures the number of species found on each island.

The *projection* of a bipartite network is a weighted unipartite co-occurrence network in which a pair of agents is connected by an edge with a weight equal to their number of shared artifacts. For example, the bipartite projection of Darwin's species location network is a species co-occurrence network in which a pair of species is connected by an edge with a weight equal to the number of islands where they are both found. We use $\mathbf{P}$ to denote the matrix representation of a bipartite projection, which is computed as $\mathbf{B}\mathbf{B}^T$, where $\mathbf{B}^T$ indicates the transpose of $\mathbf{B}$. In a projection $\mathbf{P}$, P_{ij} indicates the number of times both i and j were connected to the same artifact k in $\mathbf{B}$. The diagonal entries of $\mathbf{P}$, P_{ii}, are equal to the agent degrees, but in practice are ignored.

The *backbone* of a bipartite projection is a binary representation of $\mathbf{P}$ that contains only the most 'important' or 'significant' edges. For example, the backbone of a species co-occurrence network connects pairs of species if they are found on a significant number of the same islands, which might be interpreted as evidence that the two species do not compete for resources and perhaps are symbiotic. We use $\mathbf{P}'$ to denote the matrix representation of the backbone of $\mathbf{P}$. Because multiple methods exist for deciding when an edge is significant and thus should occur in the backbone, we use $\mathbf{P}'^{\mathrm{M}}$ denote a backbone extracted using method M.

Backbone extraction methods that were originally developed for non-projection weighted networks are often also applied to weighted bipartite projections. One simple method preserves an edge in the backbone if its weight in the projection exceeds some *universal threshold* T. However, when $T = 0$, which is common, the backbone is very dense and has a high clustering coefficient because each artifact of degree d induces $d(d-1)/2$ edges in the backbone.[23] Using $T > 0$ can yield a sparser and less clustered backbone,[24–26] but the choice of a particular threshold value is arbitrary, and applying the same threshold to all edges yields backbones that overlook agents with low degree in the projection.[13] More sophisticated methods, including the *disparity filter*[13] and *likelihood filter*,[14] aim to overcome these limitations of the universal threshold method by using a different threshold for each edge based on a null model. However, all methods that can be applied to non-projection weighted networks have the same shortcoming when applied to weighted bipartite projections: they ignore information about the artifacts.[12] In the ecological case, the universal threshold, disparity filter, and likelihood filter methods all decide whether two species should be connected in the backbone only by examining how many islands they are both found on, but do not consider the characteristics of those islands, including how many other species are found there, or even how many islands there are. Therefore, although these methods are promising for extracting the backbone from non-projection weighted networks, different methods are required for extracting the backbone from a bipartite projection.

Bipartite ensemble backbone models

Bipartite ensemble backbone models decide whether an edge's observed weight P_{ij} is significantly large, and thus whether a corresponding edge should be included in the backbone, in the following way. Let $\mathscr{B}$ be the set of all bipartite networks $\mathbf{B}^*$ having the same number of agents and artifacts as $\mathbf{B}$. In the ecological case, $\mathbf{B}^*$ might be viewed as representing a possible world containing the same species and islands, but in which locations of species on islands is different, and likewise $\mathscr{B}$ is the set of all such possible worlds. We will create our ensembles by taking a subset $\mathscr{B}^{\mathrm{M}}$ of $\mathscr{B}$ subject to certain constraints M and imposing a probability distribution on it. In all our models except the SDSM, we impose the uniform probability distribution on $\mathscr{B}^{\mathrm{M}}$, that is, each element of the ensemble is equally likely. We will then extract the backbone from the projection of $\mathbf{B}$ by using the distribution of edge weights arising from projections of members of the ensemble under consideration.

We use P^*_{ij} to denote a random variable equal to $(\mathbf{B}^*\mathbf{B}^{*T})_{ij}$ for $\mathbf{B}^* \in \mathscr{B}^{\mathrm{M}}$. That is, P^*_{ij} is the number of artifacts shared by i and j in a bipartite network randomly drawn from $\mathscr{B}^{\mathrm{M}}$. In the ecological case, P^*_{ij} represents the number of islands that are home to both species i and j in a possible world, while the distribution of P^*_{ij} is the distribution of the number of islands shared by species i and j in all possible worlds.

Decisions about which edges should appear in a backbone extracted at the two-tailed statistical significance level α are

made by comparing P_{ij} to P^*_{ij}

$$P'_{ij} = \begin{cases} 1 & \text{if } \Pr(P^*_{ij} \geq P_{ij}) < \frac{\alpha}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

This test preserves an edge in the backbone if its weight in the observed projection is uncommonly large compared to its weight in projections of members of the ensemble. A two-tailed significance test is used because, in principle, an edge's weight in the observed projection could be uncommonly *larger* or uncommonly *smaller* than its weight in projections of members of the ensemble. In the ecological case, two species are connected in the backbone if their number of shared islands in the observed world is uncommonly large compared to their number of shared islands in all possible worlds.

There are many ways that $\mathscr{B}$ can be constrained,[27] with each set of constraints describing a different ensemble $\mathscr{B}^{\mathrm{M}}$ and different ensemble backbone model; however, in this paper we focus on five possibilities. In the *fixed degree sequence model* (FDSM) $\mathbf{B}^* \in \mathscr{B}^{\mathrm{FDSM}}$ are constrained to have the same agent and artifact degree sequences as $\mathbf{B}$. Adopting the FDSM implies, for example, that in all possible worlds a given species is found on exactly the same number of islands, and a given island is home to exactly the same number of species. The distribution of P^*_{ij} arising from $\mathscr{B}^{\mathrm{FDSM}}$ is unknown, but can be approximated by uniformly sampling $\mathbf{B}^*$ from $\mathscr{B}^{\mathrm{FDSM}}$, constructing $\mathbf{P}^*$, and saving the values P^*_{ij}. In the studies below, we use 1000 samples of $\mathbf{B}^*$ generated using the 'curveball' algorithm, which is among the fastest methods to sample $\mathscr{B}^{\mathrm{FDSM}}$ uniformly at random.[28,29] The FDSM has been used to extract the backbone of bipartite projections of, for example, movies co-liked by viewers[15] and conference panel co-participation by scholars.[30,31] In this paper, we use the FDSM as the reference model to which other ensemble models are compared because it fully controls for both degree sequences.

The primary limitation of the FDSM is its computational cost. First, constructing each $\mathbf{P}^*$ requires matrix multiplication, which must be performed repeatedly and has complexity $\mathscr{O}(n^{2.37})$ for two $n \times n$ matrices using the fast Coppersmith-Winograd algorithm.[32] Second, computing $\Pr(P^*_{ij} \geq P_{ij})$ with sufficient precision to achieve a two-tailed familywise error rate of α requires at least $\frac{.5m^2 - .5m}{\alpha/2} + 1$ samples, where m is the number of rows (i.e., agents) in $\mathbf{B}$ and $\mathbf{P}$. Thus, for example, extracting the backbone of a bipartite projection with 1000 agents at a family-wise error rate of 0.05 would require performing at least 20 million matrix multiplications. Therefore, the tightly-constrained FDSM is frequently impractical for backbone extraction. However, models that rely on ensembles with more relaxed constraints offer computationally-feasible alternatives.

In the highly relaxed *fixed fill model* (FFM), $\mathbf{B}^* \in \mathscr{B}^{\mathrm{FFM}}$ are simply constrained to contain the same number of 1s as $\mathbf{B}$. Adopting the FFM implies, for example, that in all possible worlds only the total number of species-habitat pairs is fixed, but any given species may be found on a different number of islands and any given island may be home to a different number of species. The distribution of P^*_{ij} arising from B^{FFM} has not been described before, but is derived in *Supplementary Text S1.1*. We call it a *Jacobi distribution* because it is related to Jacobi polynomials.

In the more constrained *fixed row model* (FRM), $\mathbf{B}^* \in \mathscr{B}^{\mathrm{FRM}}$ are constrained to have the same agent degree sequence as $\mathbf{B}$, but have unconstrained artifact degree sequences. Adopting the FRM for backbone extraction implies, for example, that in all possible worlds a given species is found on the same number of islands, but a given island may be home to a different number of species. The distribution of P^*_{ij} arising from $\mathscr{B}^{\mathrm{FRM}}$ is hypergeometric (see *Supplementary Text S1.2*).[17,33] The FRM has been used to extract the backbone of bipartite projections of, for example, movies co-starring actors,[17] papers co-written by authors,[17] parties co-attended by women,[33] majority opinions joined by Supreme Court justices,[33], and microRNAs co-associated with diseases.[34]

In the closely related *fixed column model* (FCM), $\mathbf{B}^* \in \mathscr{B}^{\mathrm{FCM}}$ are constrained to have the same artifact degree sequence as $\mathbf{B}$, but have unconstrained agent degree sequences. Adopting the FCM for backbone extraction implies, for example, that in all possible worlds a given species may be found on a different number of islands, but a given island is home to the same number of species. The distribution of P^*_{ij} arising from $\mathscr{B}^{\mathrm{FCM}}$ has not been described before, but is derived in *Supplementary Text S1.3*, where we show it is Poisson-binomial.

Finally, the *stochastic degree sequence model* (SDSM) takes $\mathscr{B}^{\mathrm{SDSM}}$ to be all binary $m \times n$ matrices, but also gives a process for generating these matrices with different probabilities. Each $\mathbf{B}^*$ is generated by filling the cells B^*_{ik} with a 0 or 1 depending on the outcome of an independent Bernoulli trial with probability p^*_{ik}. The distribution of the random variable P^*_{ij} arising from $\mathscr{B}^{\mathrm{SDSM}}$ is Poisson-binomial with parameters which can be computed using the p^*_{ik} (see *Supplementary Text S1.4*).[21,35] There are many ways to choose p^*_{ik}, but in the studies below we choose p^*_{ik} so that it approximates $\Pr(B^*_{ik} = 1)$ for $\mathbf{B}^* \in \mathscr{B}^{\mathrm{FDSM}}$, with the goal of ensuring that the *expected* agent and artifact degree sequences of $\mathbf{B}^* \in \mathscr{B}^{\mathrm{SDSM}}$ match those of $\mathbf{B}$. Adopting such a version of SDSM implies, for example, that in each possible world a given species may be found on many or few islands and a given island may be home to many or few species, but the *average* number of islands on which a given species lives in all possible worlds and the *average* number of species that live on an given island in all possible worlds matches these values the observed world. The SDSM has been used to extract the backbone of bipartite projections of, for example, legislators

co-sponsoring bills,[5,12,36] zebrafish (*Danio rerio*) sharing operational taxonomic units,[37] countries sharing exports,[3] and genes expressed in genesets.[38]

Study 1: Choosing cell-filling probabilities for the SDSM

The SDSM requires choosing p^*_{ik}, which we want to approximate $\Pr(B^*_{ik} = 1)$ for $\mathbf{B}^* \in \mathscr{B}^{\text{FDSM}}$. There are three types of methods that might be used for doing so: arithmetic, general linear models, and entropy maximization. First, we can choose $p^*_{ik} = (r_i \times c_k)/f$, where r_i is the sum of entries in row i of $\mathbf{B}$, c_k is the sum of entries in column k of $\mathbf{B}$, and f is the sum of all entries in $\mathbf{B}$. When p^*_{ik} falls outside the $[0,1]$ range, it is truncated toward 0 or 1, respectively.[19] We call this method RCF because the value is chosen based on a row sum, a column sum, and the number of entries of $\mathbf{B}$ that are filled with a one. Second, an estimate can be obtained by fitting a general linear model of the form:

$$B_{ik} = \beta_0 + \beta_1 r_i + \beta_2 c_k + \varepsilon, \text{ or}$$
$$B_{ik} = \beta_0 + \beta_1 r_i + \beta_2 c_k + \beta_3 r_i c_k + \varepsilon,$$

where the β's are estimated coefficients and ε is an error term. If the model is treated as a linear regression and the coefficients are estimated using ordinary least squares, then the predicted value of B_{ik} is chosen for p^*_{ik}, either truncating values outside the required $[0,1]$ range (linear probability model; LPM) or transforming them into the required range using a linear discriminant model (LDM).[39] If the model is treated as a logistic regression and the coefficients are estimated using maximum likelihood, then the predicted probability that $B_{ik} = 1$ is chosen for p^*_{ik}. In prior work, the logistic regression approach has used a scobit or logit link function, with or without an interaction term (β_3).[5,12,36] Finally, an estimate can be obtained by entropy maximization methods, including the polytope method (Poly)[21,40] or bipartite configuration model (BiCM).[16] In this study, we evaluate the accuracy and speed of these methods for computing p^*_{ik}.

Methods

To evaluate accuracy, we begin by enumerating all the members of a small $\mathscr{B}^{\text{FDSM}}$. For example, given an agent degree sequence of $[1,1,2]$ and an artifact degree sequence of $[1,1,2]$, $\mathscr{B}^{\text{FDSM}}$ contains 5 members (see Table 1A). Second, from this complete enumeration, we compute the probabilities we wish p^*_{ik} to approximate (i.e., $\Pr(B^*_{ik} = 1)$ for $\mathbf{B}^* \in \mathscr{B}^{\text{FDSM}}$, see Table 1B). Third, we compute p^*_{ik} using each of nine methods (see Table 1C for the BiCM method). Finally, we quantify the accuracy with which p^*_{ik} approximates the desired probabilities using the absolute mean difference for all i,k. In the example shown in Table 1, BiCM's accuracy for these degree sequences is 0.028. That is, on average p^*_{ik} chosen using BiCM deviates from the desired probabilities by $\pm$ 0.028. Because evaluating accuracy in this way requires enumerating all members of $\mathscr{B}^{\text{FDSM}}$, it is possible only for short degree sequences that define $\mathscr{B}^{\text{FDSM}}$ with small cardinality. We focus on degree sequences ranging in length from 2 to 5, which define 384 unique $\mathscr{B}^{\text{FDSM}}$ ranging in cardinality from 4 to 2040.

(**A**) Members of $\mathscr{B}^{\text{FDSM}}$

1	0	0
0	0	1
0	1	1

0	0	1
1	0	0
0	1	1

0	0	1
0	0	1
1	1	0

0	0	1
0	1	0
1	0	1

0	1	0
0	0	1
1	0	1

(**B**) Desired probabilities

0.2	0.2	0.6
0.2	0.2	0.6
0.6	0.6	0.8

(**C**) p^*_{ik} computed using BiCM

0.216	0.216	0.568
0.216	0.216	0.568
0.568	0.568	0.863

Table 1. SDSM probabilities given agent and artifact degree sequences [1,1,2]

After identifying each method's accuracy, we evaluate the computational running time of the four most accurate methods by using them to choose p^*_{ik} for bipartite graphs defined by up to 3162 agents and up to 3162 artifacts, and thus requiring choosing up to 10,000,000 probabilities.

Results

Figure 1A shows the accuracy of each method's computation of p^*_{ik}. Each gray line plots the accuracy of each method for a single $\mathscr{B}^{\text{FDSM}}$, while the red line plots the mean accuracy of each method over all 384 $\mathscr{B}^{\text{FDSM}}$. We find that choosing p^*_{ik} using

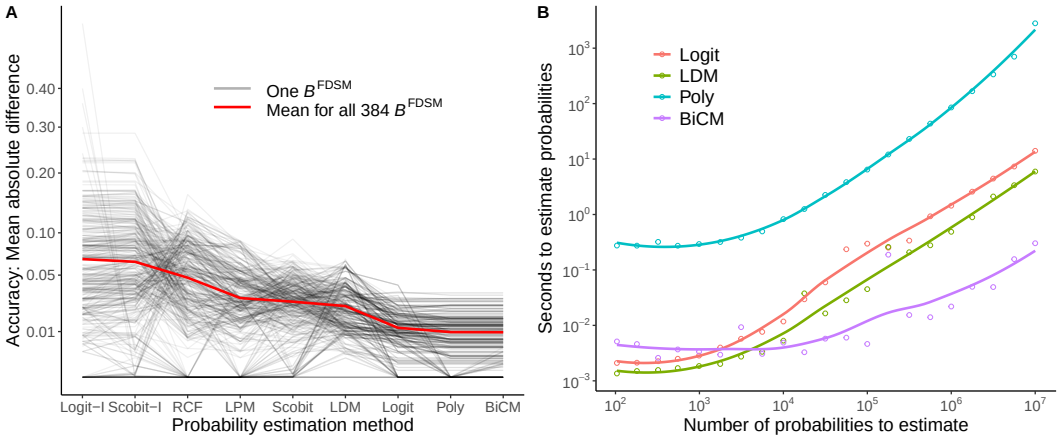

Figure 1. (A) Accuracy and (B) speed computing p^*_{ik} using different methods.

a logistic regression with an interaction term (i.e., (Scobit-I and Logit-I)) is on average least accurate,[5,12] while choosing p^*_{ik} using entropy maximization (i.e., BiCM and Poly) is on average most accurate.[3,21]

Figure 1B shows the number of seconds required to compute p^*_{ik} using a 2.3 GhZ Intel i7 processor. Among the two most accurate methods, BiCM is several orders of magnitude faster than Polytope. When computing more than 10^4 probabilities, BiCM is also faster than the two slightly less accurate Logit and LDM methods. In the largest case we evaluated, computing 10^7 probabilities, BiCM took only about 0.3 seconds. Therefore, we use BiCM for choosing p^*_{ik} when extracting SDSM backbones in the remaining studies because it is both the most accurate and fastest.

Study 2: Statistical power of SDSM

Ensemble backbone models require the specification of a statistical significance level α, which determines how uncommonly large an observed edge weight P_{ij} must be when compared to edge weights P^*_{ij} arising from an ensemble in order for a corresponding edge to be included in the backbone. For a given model, smaller values of α represent more stringent criteria for retaining edges, and therefore yield sparser backbones. Although FDSM and SDSM define their respective ensembles by constraining both agent and artifact degree sequences, and thus aim to yield similar backbones, a given α does not necessarily represent the same level of stringency in these two models. Because the SDSM allows variation in the degree sequences of $\mathbf{B}^* \in \mathscr{B}^{\text{SDSM}}$, the distribution of P^*_{ij} is wider. These wider distributions mean that the SDSM provides a more conservative test of edge weight significance than FDSM, or alternatively the SDSM has less statistical power to detect significant edges than FDSM.

A concrete example serves to illustrate this difference. In economic geography, it is common to study the world city network using a bipartite projection where two cities are linked to the extent that firms maintain locations in both of them. The Globalization and World Cities (GaWC) dataset has been widely-used in this context, and takes the form of a bipartite network recording the presence or absence of 100 firms (artifacts) in 196 cities (agents) in the year 2000.[2,22] In this bipartite network, the agent degrees are right-tailed because most cities contain only a few firms, while a few cities (e.g., New York) contain many. Likewise, the artifact degrees are also right tailed because most firms maintain locations in only a few cities, while a few firms (e.g., KPMG) maintain locations in many.

Figure 2A illustrates the distribution of the Milan-Paris edge weight in projections arising from $\mathscr{B}^{\text{FDSM}}$ and $\mathscr{B}^{\text{SDSM}}$ of which the observed bipartite network is a member (i.e., the random variable P^*_{ij}). These distributions allow a researcher to decide whether Milan and Paris's observed number of co-located firms is significantly large, and therefore whether Milan and Paris should be connected in a world city network backbone. The SDSM distribution is wider than the FDSM distribution, which has implications for whether the Milan-Paris edge will be included in a backbone extracted at a given significance level using each model. In the observed data, there are 26 firms co-located in Milan and Paris (i.e., $P_{ij} = 26$). The probability of observing the same or larger edge weight from the FDSM ensemble is 0.0033, which is less than $\frac{0.05}{2}$, and therefore a Milan-Paris edge is deemed significant by the FDSM model and is included in the FDSM backbone extracted at $\alpha = 0.05$. In contrast, the probability of observing the same or larger edge weight from the SDSM ensemble is 0.0275, which is *not* less than $\frac{0.05}{2}$, and therefore a Milan-Paris edge is *not* deemed significant by the SDSM model and is *not* included in the SDSM backbone extracted at $\alpha = 0.05$. This difference in statistical power leads the SDSM backbone to be sparser than the FDSM backbone (density = 0.004 vs. 0.012), and means that these two backbones are dissimilar (Jaccard = 0.36).

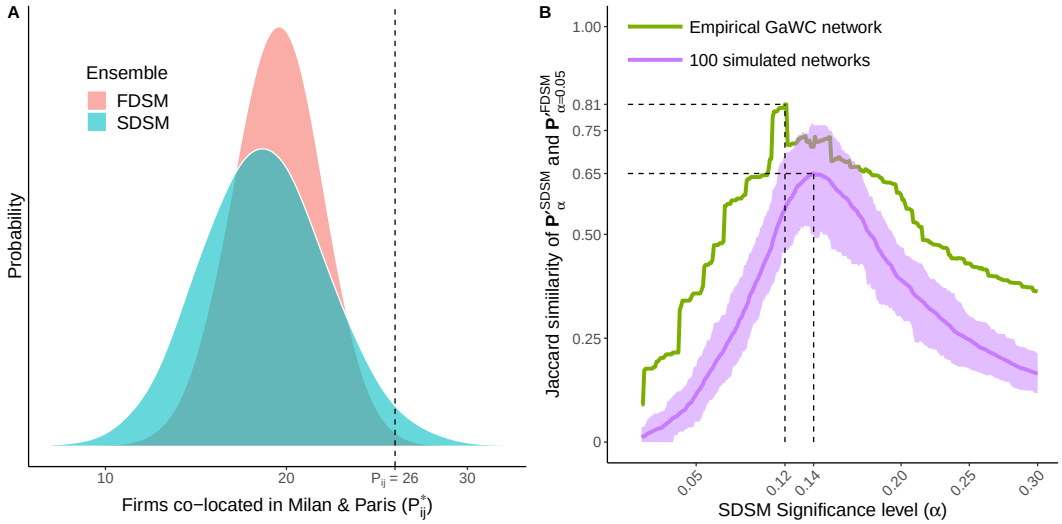

Figure 2. Statistical power of SDSM. (A) Distribution of weights for the Paris-Milan edge in projections derived from FDSM and SDSM ensembles. (B) Similarity of an FDSM backbone extracted at $\alpha = 0.05$ to SDSM backbones extracted at various α from an empirical bipartite network (green line) and from 100 synthetic bipartite networks (purple line = mean, purple region = 10^{th}–90^{th} percentile).

In this study, we investigate SDSM's statistical power relative to FDSM, and specifically whether extracting an SDSM backbone using a more liberal (i.e., larger) α makes it more similar to an FDSM backbone extracted at $\alpha = 0.05$.

Methods

To evaluate SDSM's statistical power and the effect of significance levels on the similarity of SDSM and FDSM backbones, we first extracted the FDSM backbone from the GaWC bipartite network at $\alpha = 0.05$. We then extracted several SDSM backbones from the GaWC bipartite network at $0.01 \leq \alpha \leq 0.3$ in 0.001 increments, each time computing the Jaccard index (J) to measure the similarity between the SDSM and FDSM backbones. After comparing SDSM and FDSM backbones extracted from the empirical GaWC bipartite network, we repeat this process using 100 synthetic bipartite networks with the same dimensions (196×100), density (0.08) and right-tailed agent and artifact degree distributions.

Results

The green line in Figure 2B shows the Jaccard similarity between an FDSM backbone extracted from the empirical GaWC network at $\alpha = 0.05$ and SDSM backbones extracted at the significance levels shown on the x-axis. We find that an SDSM backbone achieves its maximum similarity to the FDSM backbone ($J = 0.81$) when it is extracted using the more liberal significance level of $\alpha = 0.12$. Returning to the example in Figure 2A, using this more liberal significance level would result in the Milan-Paris edge being deemed significant and included in the SDSM backbone because its SDSM p-value $0.0275 < \frac{0.12}{2}$. Because this more liberal significance level results in the inclusion of additional edges, the new SDSM backbone extracted at $\alpha = 0.12$ has a density (0.01) close to that of the FDSM backbone extracted at $\alpha = 0.05$ (0.012).

The purple line in Figure 2B shows the mean Jaccard similarity between an FDSM backbone extracted using $\alpha = 0.05$ and SDSM backbones extracted using $0.01 \leq \alpha \leq 0.3$ from 100 bipartite networks generated to resemble the empirical GaWC network. The shaded purple region shows the 10^{th} and 90^{th} percentile of Jaccard similarities of these backbones. We find that these synthetic networks behave similarly to the empirical network. Specifically, SDSM and FDSM backbones extracted from a low-density 196×100 bipartite network with right-tailed degree distributions achieve a maximum similarity of $0.49 < J < 0.76$ when the FDSM backbone is extracted using $\alpha = 0.05$ and the SDSM backbone is extracted using $\alpha = 0.14$. This is promising because it suggests that, given the characteristics of an empirical bipartite network, it may be possible to select a significance level for extracting a computationally-efficient SDSM backbone that closely resembles a computationally-infeasible FDSM backbone.

Study 3: Backbone equivalence under varying degree distributions

Agent and artifact degree distributions are a key feature of a bipartite network, and are known to have implications for bipartite projections.[9,21,22] The FDSM is particularly appealing because it allows decisions about the significance of edges in a projection

Degree Distribution	Authors (agents)	Papers (artifacts)
Right-tailed $\sim \beta(1,10)$	Most write some papers, but a few are prolific (most departments).	Most papers are sole-authored, but some are written by large teams (e.g., sociology).
Left-tailed $\sim \beta(10,1)$	Most are prolific, but some are inactive (elite departments).	Most papers are written by large teams, but some are sole-authored (e.g., physics).
Uniform $\sim \beta(1,1)$	There is substantial diversity in scholarly output (e.g., interdisciplinary departments).	There is substantial diversity in the size of authorship teams (e.g., an entire university).
Constant $\sim \beta(10000,10000)$	There are strong norms about how many papers an author should have (e.g., for performance evaluations).	There are strong norms about how many authors a paper should have (e.g., senior/lead & junior)
Normal $\sim \beta(10,10)$	Scholarly output varies around some typical level.	Authorship teams vary around some typical size.

Table 2. Bipartite degree distributions, with examples in the context of a scholarly authorship bipartite network

to be conditioned on both bipartite degree sequences, thereby taking into account these important features. However, because the computational requirements of the FDSM make it impractical for extracting the backbone from most bipartite projections, it is often necessary to use a different backbone model. In this study, we evaluate the equivalence of an FDSM backbone and backbones extracted using more computationally efficient models. We perform this comparison for backbones extracted from bipartite networks characterized by five types of degree distributions: right-tailed, left-tailed, normal, constant, and uniform.

For the sake of concreteness, in this section we use the example of a bipartite network in which authors (agents) are linked to the papers they have written (artifacts). The projection of such a network yields a co-authorship network in which the edge weight between a pair of authors indicates their number of co-authored papers.[4] These edge weight values will depend heavily on the distribution of papers written by authors (i.e., the agent degree sequence), and on the distribution of authors on each paper (i.e., the artifact degree sequence). Different distributions describe different kinds of scholarly environments as shown in Table 2. The choice of a backbone model affects how and whether these distributions are considered, and in this example affects the extent to which decisions about the significance of two authors' number of co-authored papers consider the scholarly environment. The FDSM compares their observed number of co-authored papers to the number that might be observed in alternative realizations *of the same environment*, while other backbone models relax the extent to which the environment is held constant.

Methods

We evaluate similarities among the backbones extracted using different models by comparing backbones extracted from synthetic 100×100 bipartite networks with a density of 0.1. Following our example, these synthetic bipartite networks might represent a college of 100 faculty who collectively wrote 100 papers, where each individual had a 10% chance of being an author on each paper. First, we generate separate agent and artifact degree sequences that sum to 1000 and that approximately follow one of the beta distributions in Table 2. Second, we generate a bipartite network with these agent and artifact degree sequences using the Gale-Ryser algorithm.[41] Third, we extract five different backbones from the generated bipartite network, using the fixed fill model, fixed row model, fixed column model, stochastic degree sequence model, and fixed degree sequence model; in all cases we use $\alpha = 0.05$. Finally, we compute the similarity of the first four backbones to the FDSM backbone using a Jaccard index. We repeat this process 100 times for each of the 25 possible combinations of agent and artifact degree distributions.

Results

The heatmaps in Figure 3 illustrate the similarity between an FDSM backbone and a backbone extracted using an alternative model. The rows of each heat map correspond to different agent degree distributions, and the columns correspond to different artifact degree distributions, in the synthetic bipartite networks from which the backbones were extracted. The lightest patches identify conditions under which a given backbone model yields a backbone that is similar to what would be obtained using the computationally costly FDSM, while darker patches identify conditions under which these two backbones differ. We find that when agent degrees are constant (i.e., every agent has the same degree) and artifact degrees are constant or left-tailed, all backbone models yield essentially the same backbone as FDSM (Mean $J = 0.999$). However, beyond this special case, which is likely to be rare in empirical data, similarity to FDSM-extracted backbones varies.

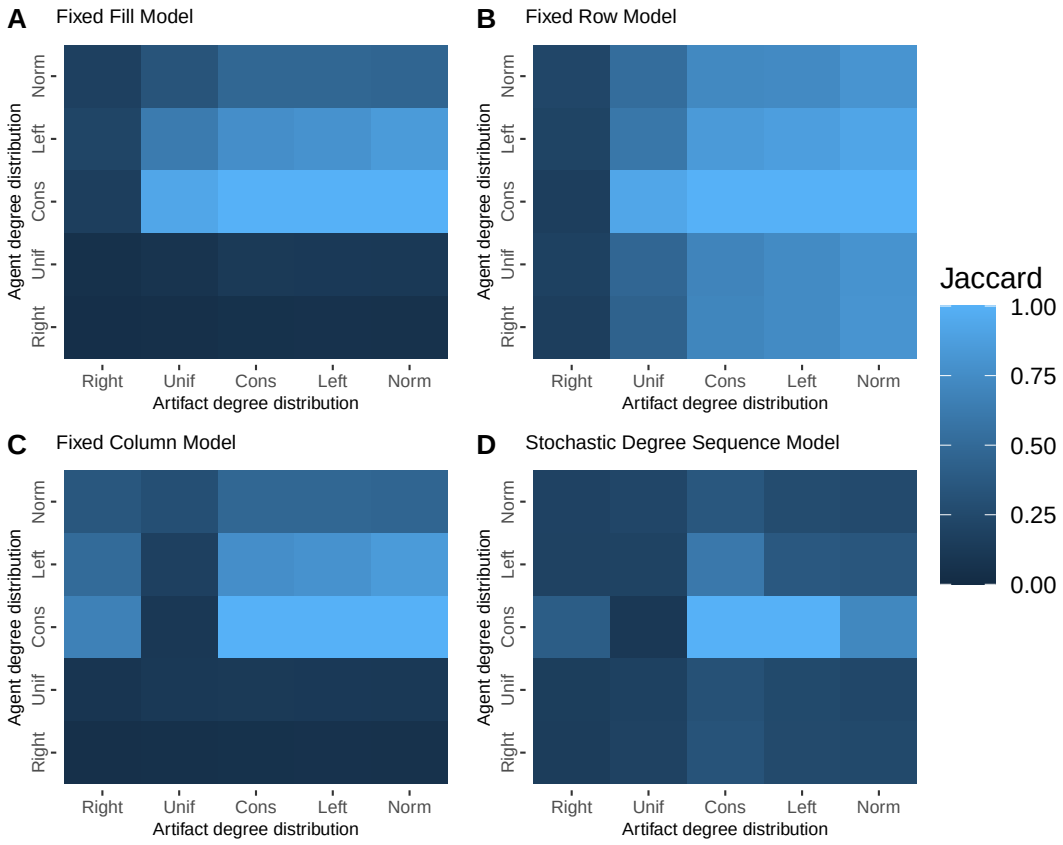

Figure 3. Jaccard similarity of a backbone extracted at $\alpha = 0.05$ using the Fixed Degree Sequence Model and a backbone extracted using (A) the Fixed Fill Model, (B) Fixed Row Model, (C) Fixed Column Model, (D) Stochastic Degree Sequence Model. Each cell represents the mean over 100 instances of a 100×100 bipartite network with given agent and artifact degree distributions.

As expected, the similarity of backbones extracted using FRM and FDSM depends primarily on the distribution of artifact degrees, not agent degrees (see Figure 3B). For example, for any agent degree distribution, these two models yield very different backbones when artifact degrees follow a right-tailed distribution (Mean $J = 0.186$), but very similar backbones when artifact degrees follow a normal distribution (Mean $J = 0.863$). This occurs because both models exactly control for agent degrees, however FDSM also controls for artifact degrees, while FRM does not.

A similar but rotated pattern emerges when considering the FCM: the similarity of backbones extracted using FCM and FDSM depends primarily on the distribution of agent degrees, not artifact degrees (see Figure 3C). For any artifact degree distribution, these two models yield very different backbones when agent degrees follow a right-tailed or uniform (Mean $J = 0.084$) distribution , but more similar backbones when agent degrees follow a left-tailed distribution or are constant (Mean $J = 0.617$). This occurs because both models exactly control for artifact degrees, however FDSM also controls for agent degrees, while FRM does not. However, there is a notable exception to this general pattern: when artifact degrees follow a uniform distribution, FCM and FDSM always yield different backbones (Mean $J = 0.151$).

The conditions under which the FFM yields FDSM-similar backbones occur at the intersection of the conditions under which the FRM and FCM both yield FDSM-equivalent backbones (see Figure 3A). When artifact degrees follow a right-tailed distribution and/or the agent degrees follow a right-tailed or uniform distribution, then FFM and FDSM backbones differ (Mean $J = 0.1$). In contrast, for other combinations of degree distributions, FFM and FDSM backbones are more similar (Mean $J = 0.724$).

Finally, as expected based on the findings from study 2, we observe that the SDSM generally yields different backbones than FDSM when both are extracted at $\alpha = 0.05$ (see Figure 3D). Specifically, except in the narrow case where agent degrees are constant and artifact degrees are constant or left-tailed (Mean $J = 1$), SDSM and FDSM backbones exhibit only modest similarity (Mean $J = 0.314$). This lack of similarity or equivalence occurs because SDSM offers a less statistically powerful (or more conservative) test of edges statistical significance than FDSM, and therefore retains fewer edges in the backbone. However, findings from study 2 also suggested that careful selection of the significance level used for extracting an SDSM backbone can yield results more similar to FDSM.

To explore this possibility, we repeated the analysis reported in figure 3D, finding that when a suitably more liberal (i.e., larger) significance level α is used to extract an SDSM backbone, the resulting SDSM backbone is very similar to an FDSM backbone extracted at $\alpha = 0.05$ (see Figure 4A). Specifically, for backbones extracted from bipartite networks with *any* agent or artifact degree distributions, these two backbones tend to be nearly equivalent (Mean $J = 0.865$). This suggests that in principle the fast SDSM can be used to obtain a close approximation of a computationally-infeasible FDSM backbone from any bipartite network.

In practice, using SDSM to obtain an FDSM-like backbone requires selecting an α value for the SDSM that corresponds to $\alpha = 0.05$ in the FDSM. We observe that there are three distinct values of such an 'optimal' α that depend on agent and artifact degree distributions (see Figure 4B). First, when agent degrees are constant, a value only slightly higher than 0.05 (Mean $= 0.062$, SD $= 0.021$) achieves the best approximation of an FDSM backbone. Second, when artifact degrees are constant, a value roughly double (Mean $= 0.09$, SD $= 0.022$) achieves the best approximation of an FDSM backbone. Finally, when neither agent nor artifact degrees are constant, which is likely in most empirical bipartite networks, a value roughly 2.5 times larger (Mean $= 0.13$, SD $= 0.014$) achieves the best approximation of an FDSM backbone. Although further work is needed to facilitate the *a priori* selection of an α that allows an SDSM backbone to closely approximate an FDSM backbone, these results suggest that under the most common circumstances (i.e., when there is variation in degrees) $\alpha \approx 0.13$ may be appropriate.

Study 4: Recovery of community structure

Studies 1-3 examine the backbones extracted from synthetic random bipartite networks; however, empirical bipartite networks are generally not random. For backbones of bipartite projections to be useful, they must be able to capture this non-random structure. Community structure – the clustering of nodes into groups such that within-group edges are more common than between-group edges – is among the most widely studied patterns in unipartite networks.[42] In this study, we evaluate the extent to which backbones extracted using different models reflect a known community structure that is encoded in the bipartite data from which they are extracted. Prior work has shown that SDSM and FDSM backbones extracted from a bipartite network representing bill co-sponsorship in the 114th session of the US Senate more clearly captured the known partisan community structure than an FRM backbone.[21] For the sake of concreteness, we use this context as an example in this section, but we extend this prior work by considering a broader range of backbone models, and by examining their ability to recover community structures from bipartite data containing varying levels of evidence for this structure.

Methods

We investigate the ability for backbones to recover a known community structure in three steps. First, we simulate a 200×1000 bipartite network with a density of 0.1 and right-tailed agent and artifact degree distributions. We focus on a bipartite

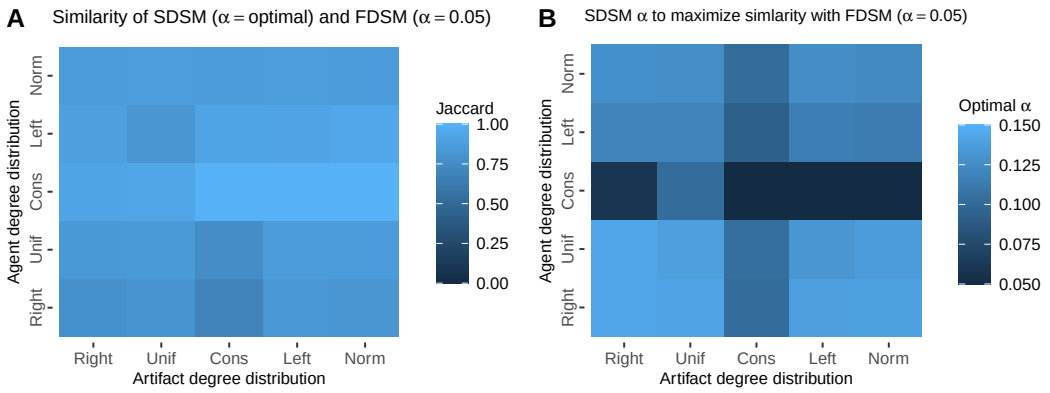

Figure 4. (A) Given agent and artifact degree distributions, there exists a statistical significance level α that maximizes the similarity between an SDSM backbone extracted at this level and an FDSM backbone extracted at $\alpha = 0.05$, and (B) when used yields an SDSM backbone that is very similar to the corresponding FDSM backbone.

network with more artifacts than agents to ensure that these data contain sufficient information to encode potential community memberships. We focus on a bipartite network with right-tailed degree distributions because they are common in many empirical unipartite[43] and bipartite networks.[5,6,22] This synthetic bipartite network could represent a legislative body composed of 200 legislators casting votes on 1000 bills, where any given legislator had a 10% chance of voting in favor of any given bill. The right-tailed degree distributions capture the fact that most legislators vote in favor of only a few bills, and that most bills receive the support of only a few legislators, which is typical of legislative bodies. The backbone of a projection of such a bipartite network would represent a network of collaboration or ideological alignment among legislators.[5]

Second, we incorporate evidence of communities in this bipartite network by randomly assigning each agent and each artifact to one of two groups. We then shuffle the edges, preserving the degree distributions, such that a given fraction of edges W are within-group, connecting an agent and artifact from the same group. Figure 5A provides graphical depictions of the matrices describing synthetic bipartite networks at two values of W. In each plot, the rows represent agents assigned to group A or B, the columns represent artifacts assigned to group A or B, and a cell is shaded black if the row agent is connected to the column artifact. When $W = 0.5$, agents in a given group are equally likely to associate with artifacts in either group, placing ≈ 0.5 of the edges (i.e., shaded cells) in the diagonal blocks and ≈ 0.5 of the edges in the off-diagonal blocks. In contrast, when $W = 0.8$, agents in a given group are much more likely to associate with artifacts from their own group than artifacts in the other group, placing ≈ 0.8 of the edges in the diagonal blocks and ≈ 0.2 of the edges in the off-diagonal blocks. Returning to our example, the groups could represent political parties: each legislator belongs to one of two parties (i.e., there are conservative and liberal legislators), and each bill advances the agenda of one of these parties (i.e., there are conservative and liberal bills). When $W = 0.5$, a conservative legislator is equally likely to vote for conservative and liberal bills, while when $W = 0.8$, a conservative legislator is four-times more likely to vote for a conservative bill than a liberal bill.

Finally, we extract a backbone from the bipartite network using a given model and compute the backbone's modularity Q with respect to the agents' group assignments. If a backbone model is able to recover the community structure from evidence in the bipartite network, then we expect a positive association between W and Q, and more specifically we expect that as W approaches 1 (i.e., agents are associated only with artifacts from their own group) Q will approach 0.5 (i.e., a backbone with distinct communities).[44] In the legislative example, if legislators are bipartisan in their voting patterns (i.e., $W = 0.5$), then legislators should not be clustered by party in the backbone (i.e., $Q \approx 0$). In contrast, if legislators are strongly partisan in their voting patterns (i.e., $W = 0.8$), then legislators should be clustered by party in the backbone (i.e., $Q \approx 0.5$).

We repeat these three steps 10 times for $0.5 \leq W \leq 0.8$ in 0.05 increments. When evaluating the SDSM backbone, we consider both a backbone extracted using the conventional significance level of $\alpha = 0.05$ and one extracted at the more liberal $\alpha = 0.13$, which study 3 suggests yields a backbone similar to FDSM.

Results

Figure 5B shows the modularity (y-axis; with respect to known community memberships) of backbones extracted using different models from bipartite networks with different fractions of within-community edges (x-axis). All six lines increase monotonically, confirming that all backbone models yield backbones that can recover a known community structure. However, there is notable variation among the models. As evidence of community structure grows stronger in the bipartite network, the modularity of backbones extracted using the FFM and FCM slowly increase, but reach average values of only $Q = 0.15$ and

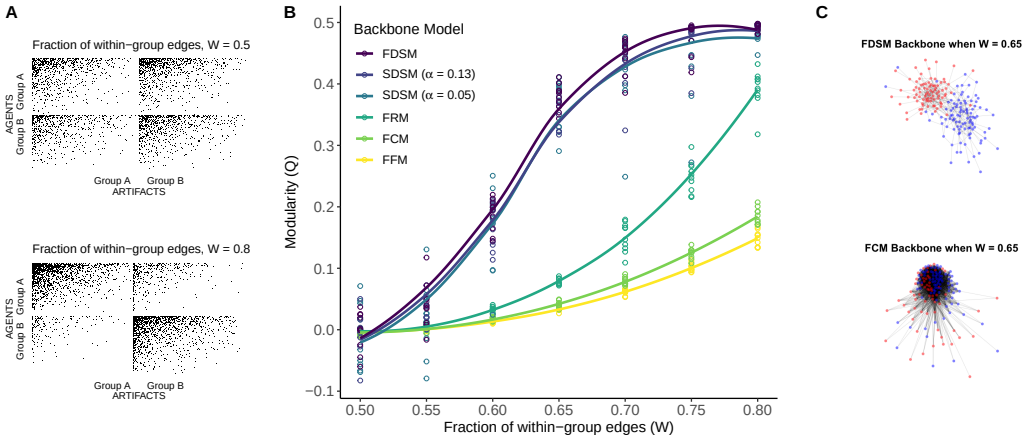

Figure 5. (A) Synthetic bipartite networks with varying levels of block structure, from which (B) backbones extracted using different models exhibit varying modularity. (C) When 65% of bipartite edges are within-block, a backbone extracted using FDSM shows a clear group structure (top) while a backbone extracted using FCM does not (bottom).

0.18 (respectively), even when the evidence of such a structure is quite strong (i.e., when $W = 0.8$). Backbones extracted using the FRM display a similar pattern, but achieve a higher average value when W is large ($Q = 0.39$).

In contrast, backbones extracted using FDSM and SDSM are virtually indistinguishable in their ability to recover the known community structure, and do so very well. As evidence of a community structure grows stronger in the bipartite network, the modularity of backbones extracted using these models rapidly increases. When the evidence of community structure is strong (i.e., when $W = 0.8$), these backbones have very high modularity (mean $Q = 0.49$). However, even when there is only modest evidence of community structure in the bipartite network (e.g., when $W = 0.65$), these backbones are still able to identify the community structure and have a distinctively high modularity (mean $Q = 0.37$).

Figure 5C illustrates the difference between two backbone models' abilities to recover a known community structure, when evidence of that structure is modest in the bipartite data from which the backbone is extracted ($W = 0.65$). In both plots, agents from group A (e.g., conservatives, in the legislative example) are colored red, while agents from group B (e.g., liberals, in the legislative example) are colored blue. The FDSM-extracted backbone clearly places agents from different groups in separate clusters. In contrast, the FCM-extracted backbone is unable to distinguish this group structure and fails to cluster agents according to their known group memberships. These findings suggest that although all backbone models can yield backbones that recover a known community structure, SDSM and FDSM backbones are able to detect this structure more clearly and from a weaker signal.

Discussion

Bipartite networks can be used to represent a wide range of phenomena in the social and natural worlds including interspecies competition, global trade, scientific advances, and legislative deliberation. Likewise, projections of bipartite networks, which take the form of co-occurrence networks, can be useful for inferring unipartite networks that would otherwise be difficult to measure directly. Several models have been proposed for extracting the backbone of bipartite projections, and thus for making such inferences, including the fixed fill model (FFM) fixed row model (FRM), fixed column model (FCM), fixed degree sequence model (FDSM), and stochastic degree sequence model (SDSM). In this paper we have systematically compared these models in terms of their relative accuracy, speed, statistical power, similarity, and ability to recover a known community structure.

In study 1, we examined several methods for choosing the probabilities necessary for applying the stochastic degree sequence model (SDSM), finding that the bipartite configuration model (BiCM) is both the fastest and most accurate. In study 2, we examined the statistical power of the SDSM relative to the fixed degree sequence model (FDSM), finding that the SDSM can be viewed as a statistically less powerful (or more conservative) variant of the FDSM. In study 3, we examined the similarity of an FDSM-extracted backbone to backbones extracted using other models, finding that the SDSM and FDSM extract very similar backbones from bipartite networks with a wide range of possible degree distributions when an appropriate significance level α is chosen. Finally, in study 4, we examined the ability for backbones extracted using different models to recover a known community structure, finding that although all models can recover the structure, SDSM and FDSM can detect a community structure more clearly and from a weaker signal.

Based on these findings, and with the goal of offering researchers some guidance in extracting the backbones of bipartite projections, we offer three recommendations. First, *we recommend the stochastic degree sequence model (SDSM) for extracting the backbones of bipartite projections* because it is fast, controls for both agent and artifact degree sequences, and yields modular backbones when the bipartite data contains even modest evidence of within-community clustering. Second, when the SDSM is used, *we recommend that the cell-filling probabilities p^*_{ik} be chosen using the Bipartite Configuration Model (BiCM)* because it is faster and more accurate than any other currently available method. Third, when an FDSM backbone extracted at the $\alpha = 0.05$ significance level is desired but computationally infeasible, *we recommend extracting an SDSM backbone at the $\alpha = 0.13$ significance level*, which we observe is very similar when there is variation in the agent and artifact degree sequences. The models and options necessary to adopt these recommendations are implemented in the `backbone` package for **R**.[21]

These findings and recommendations must be viewed in light of the fact that, due to the computational requirements of the FDSM and of extracting a large number of backbones across the four studies, these studies have relied on small synthetic bipartite networks ranging in size from 3×3 (study 1) to 200×1000 (study 4). However, in practice bipartite networks may be several orders of magnitude larger. For example, a bipartite network used to infer collaborations in the US House of Representatives includes 435 agents (representatives) and over 6000 artifacts (bills),[5,40] while a bipartite network used to infer movie recommendations includes 17,770 agents (films) and nearly 500,000 artifacts (viewers).[15] Future research should explore whether these findings extend to backbones extracted from such large bipartite networks. Limitations of existing backbone models also point to directions for future research. First, using the FDSM will generally be computationally infeasible in practice because the distribution of P^*_{ij} arising from $\mathscr{B}^{\text{FDSM}}$ must be estimated via numerical simulation. Identifying this distribution's probability mass function, which is known for the other ensembles (see *Supplementary Text S1*), would facilitate the use of this otherwise attractive model; however, this is a well-studied problem and so is probably very hard to solve. Second, all the ensemble models we have considered impose constraints on the degree sequences, but other types of constraints may also be useful. For example, in some contexts it may be necessary to constrain all members of an ensemble to contain a 0 in a particular cell (e.g., to represent that an author was not alive to co-author a paper, or a legislator was not present to co-sponsor a bill). These limitations and future directions notwithstanding, the results presented above provide a starting point for further development of backbone models, and provide applied researchers with some practical guidance and a preliminary rationale for adopting the stochastic degree sequence model.

References

1. Diamond, J. M. Assembly of species communities. In Cody, M. L. & Diamond, J. M. (eds.) *Ecology and evolution of communities*, 342–444 (Harvard University Press, 1975).
2. Taylor, P. J., Catalano, G. & Walker, D. R. Measurement of the world city network. *Urban Stud.* **39**, 2367–2376 (2002).
3. Saracco, F., Di Clemente, R., Gabrielli, A. & Squartini, T. Randomizing bipartite networks: the case of the world trade web. *Sci. Reports* **5**, 1–18 (2015).
4. Newman, M. E. Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E* **64**, 016131 (2001).
5. Neal, Z. P. A sign of the times? Weak and strong polarization in the us congress, 1973–2016. *Soc. Networks* **60**, 103–112 (2020).
6. Ahn, Y.-Y., Ahnert, S. E., Bagrow, J. P. & Barabási, A.-L. Flavor network and the principles of food pairing. *Sci. Reports* **1**, 1–7 (2011).
7. Tollefson, J. Tracking QAnon: How Trump turned conspiracy-theory research upside down. *Nature* (2021).
8. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4** (2005).
9. Vasques Filho, D. & O'Neale, D. R. J. Transitivity and degree assortativity explained: The bipartite structure of social networks. *Phys. Rev. E* **101**, 052305, DOI: 10.1103/PhysRevE.101.052305 (2020).
10. Guillaume, J.-L. & Latapy, M. Bipartite structure of all complex networks. *Inf. Process. Lett.* **90**, 215–221 (2004).
11. Newman, M. E. & Park, J. Why social networks are different from other types of networks. *Phys. review E* **68**, 036122 (2003).
12. Neal, Z. P. The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. *Soc. Networks* **39**, 84–97, DOI: 10.1016/j.socnet.2014.06.001 (2014).
13. Serrano, M. Á., Boguná, M. & Vespignani, A. Extracting the multiscale backbone of complex weighted networks. *Proc. Natl. Acad. Sci.* **106**, 6483–6488 (2009).

14. Dianati, N. Unwinding the hairball graph: Pruning algorithms for weighted complex networks. *Phys. Rev. E* **93**, 012304 (2016).

15. Zweig, K. A. & Kaufmann, M. A systematic approach to the one-mode projection of bipartite graphs. *Soc. Netw. Analysis Min.* **1**, 187–218, DOI: 10.1007/s13278-011-0021-0 (2011).

16. Saracco, F. *et al.* Inferring monopartite projections of bipartite networks: An entropy-based approach. *New J. Phys.* **19**, 053022 (2017).

17. Tumminello, M., Micciché, S., Lillo, F., Piilo, J. & Mantegna, R. N. Statistically validated networks in bipartite complex systems. *PLoS One* **6**, e17994, DOI: 10.1371/journal.pone.0017994 (2011).

18. Sanderson, J. G. Testing ecological patterns. *Am. Sci.* **88**, 332 (2000).

19. Gotelli, N. J. Null model analysis of species co-occurrence patterns. *Ecology* **81**, 2606–2621 (2000).

20. Neal, Z. P. & Neal, J. W. Out of bounds? The boundary specification problem for centrality in psychological networks, DOI: 10.31234/osf.io/nz6k3 (2020).

21. Domagalski, R., Neal, Z. P. & Sagan, B. Backbone: An R package for extracting the backbone of bipartite projections. *PloS One* **16**, e0244363 (2021).

22. Neal, Z. P., Domagalski, R. & Sagan, B. Analysis of spatial networks from bipartite projections using the R backbone package. *Geogr. Analysis* (2021).

23. Latapy, M., Magnien, C. & Del Vecchio, N. Basic notions for the analysis of large two-mode networks. *Soc. Networks* **30**, 31–48 (2008).

24. Derudder, B. & Taylor, P. The cliquishness of world cities. *Glob. Networks* **5**, 71–91 (2005).

25. Fong, C. Expertise, networks, and interpersonal influence in congress. *The J. Polit.* **82**, 269–284 (2020).

26. Bratton, K. A. & Rouse, S. M. Networks in the legislative arena: How group dynamics affect cosponsorship. *Legislative Stud. Q.* **36**, 423–460 (2011).

27. Strona, G., Ulrich, W. & Gotelli, N. J. Bi-dimensional null model analysis of presence-absence binary matrices. *Ecology* **99**, 103–115, DOI: 10.1002/ecy.2043 (2018).

28. Strona, G., Nappo, D., Boccacci, F., Fattorini, S. & San-Miguel-Ayanz, J. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nat. Commun.* **5**, 4114, DOI: 10.1038/ncomms5114 (2014).

29. Carstens, C. J. Proof of uniform sampling of binary matrices with fixed row sums and column sums for the fast curveball algorithm. *Phys. Rev. E* **91**, DOI: 10.1103/PhysRevE.91.042812 (2015).

30. Stegbauer, C. & Rausch, A. How international are international congresses? *Connections* **32**, 1–11 (2012).

31. Derudder, B. & Liu, X. How international is the annual meeting of the association of american geographers? A social network analysis perspective. *Environ. Plan. A* **48**, 309–329 (2016).

32. Coppersmith, D. & Winograd, S. Matrix multiplication via arithmetic progressions. *J. Symb. Comput.* **9**, 251–280 (1990).

33. Neal, Z. P. Identifying statistically significant edges in one-mode projections. *Soc. Netw. Analysis Min.* **3**, 915–924, DOI: 10.1007/s13278-013-0107-y (2013).

34. Chen, X. *et al.* BNPMDA: Bipartite network projection for mirna–disease association prediction. *Bioinformatics* **34**, 3178–3186 (2018).

35. Liebig, J. & Rao, A. Fast extraction of the backbone of projected bipartite networks to aid community detection. *Europhys. Lett.* **113**, 28003, DOI: 10.1209/0295-5075/113/28003 (2016).

36. Schoch, D. & Brandes, U. Legislators' roll-call voting behavior increasingly corresponds to intervals in the political spectrum. *Sci. Reports* **10**, 1–9 (2020).

37. Buerger, A. N. *et al.* Gastrointestinal dysbiosis following diethylhexyl phthalate exposure in zebrafish (danio rerio): Altered microbial diversity, functionality, and network connectivity. *Environ. Pollut.* **265**, 114496 (2020).

38. Marini, F., Ludt, A., Linke, J. & Strauch, K. Genetonic: an r/bioconductor package for streamlining the interpretation of rna-seq data. *bioRxiv* (2021).

39. Allison, P., Williams, R. A. & von Hippel, P. Better predicted probabilities from linear probability models with applications to multiple imputation. In *2020 Stata Conference*, 1 (Stata Users Group, 2020).

40. Neal, Z. P., Domagalski, R. & Yan, X. Homophily in collaborations among us house representatives, 1981–2018. *Soc. Networks* **68**, 97–106 (2022).

41. Brualdi, R. A. & Sagan, B. Dihedral transportation and (0, 1)-matrix classes. *Linear Multilinear Algebr.* **66**, 2557–2568, DOI: 10.1080/03081087.2017.1406892 (2017).

42. Fortunato, S. Community detection in graphs. *Phys. Reports* **486**, 75–174 (2010).

43. Broido, A. D. & Clauset, A. Scale-free networks are rare. *Nat. communications* **10**, 1–10 (2019).

44. Newman, M. E. & Girvan, M. Finding and evaluating community structure in networks. *Phys. review E* **69**, 026113 (2004).

Acknowledgements

This work was supported by the National Science Foundation (#1851625 & #2016320) and the Michigan State University Center for Business and Social Analytics.

Author contributions statement

Z.P.N. conceived the research questions, designed and conducted the analysis, and wrote the first draft. R.D. and Z.N. wrote the backbone package. B.S. wrote the proofs. All authors analysed the results and revised the manuscript.

Additional information

Competing interests

The authors declare no competing interests.

Supplementary information

S1 Probability Mass Functions of projection edge weights under ensemble backbone models

In the subsections below, we derive the probability mass functions of P^*_{ij} used by ensemble backbone models to evaluate the statistical significance of the weight of edge P_{ij} in a bipartite projection. We use the following notation:

- Let $\mathbf{B}$ be an $m \times n$ bipartite matrix, with a vector of row sums $R = (r_1, \ldots, r_m)$, a vector of column sums $C = (c_1, \ldots, c_n)$, and f cells containing a 1. So
$$f = \sum_{i=1}^{m} r_i = \sum_{j=1}^{n} c_j.$$
- Let $\mathscr{B}^{\mathrm{M}}$ be the ensemble of all $m \times n$ matrices $\mathbf{B}^* = (B^*_{ij})$ that obey the constraints of the respective model. In all models, the probability distribution on $\mathscr{B}^{\mathrm{M}}$ is uniform except in the stochastic case.
- Let P^*_{ij} be a random variable equal to $(\mathbf{B}^*\mathbf{B}^{*T})_{ij}$ for all $\mathbf{B}^* \in \mathscr{B}^{\mathrm{M}}$. Note that we have

$$P^*_{ij} = B^*_{i1}B^*_{j1} + B^*_{i2}B^*_{j2} + \cdots + B^*_{in}B^*_{jn}. \tag{1}$$

S1.1 Fixed Fill Model (FFM)

Let the *fixed fill model* constrain all $\mathbf{B}^* \in \mathscr{B}^{\mathrm{FFM}}$ to contain the same number of 1s (i.e. fill) as $\mathbf{B}$.

Theorem S1.1. *Under the fixed fill model, the distribution of P^*_{ij} for $i \neq j$ satisfies*

$$\Pr(P^*_{ij} = k) = \frac{\binom{n}{k}\sum_r 2^{n-k-r}\binom{n-k}{r}\binom{(m-2)n}{f-n-k+r}}{\binom{mn}{f}}. \tag{2}$$

Proof. For the denominator we need to compute the cardinality $\#\mathscr{B}^{\mathrm{FFM}}$. If $\mathbf{B}^* \in \mathscr{B}^{\mathrm{FFM}}$ then $\mathbf{B}^*$ has mn entries of which f must be chosen to be ones. So

$$\#\mathscr{B}^{\mathrm{FFM}} = \binom{mn}{f}.$$

For the numerator, suppose $P^*_{ij} = k$. We see from equation (1) that there are exactly k columns c where $B^*_{ic} = B^*_{jc} = 1$. There are $\binom{n}{k}$ ways to choose these columns. Now define the following parametrs:

$$\begin{aligned} p &= \text{number of coumns } c \text{ where } B^*_{ic} = 1 \text{ and } B^*_{jc} = 0, \\ q &= \text{number of coumns } c \text{ where } B^*_{ic} = 0 \text{ and } B^*_{jc} = 1, \\ r &= \text{number of coumns } c \text{ where } B^*_{ic} = 0 \text{ and } B^*_{jc} = 0. \end{aligned}$$

The nunber of ways to pick the columns counted by these parameters from the $n-k$ columns which do not contains ones in both rows is the trinomial coefficients $\binom{n-k}{p,q,r}$. Now we have used $2k+p+q$ ones in rows i and j. So there are $f-2k-p-q$ left to distribute to the remaining $m-2$ rows. And these rows have $(m-2)n$ entries. So the number of possibilities for these remaining ones is $\binom{(m-2)n}{f-2k-p-q}$. Thus the total number of choices from this and the previous paragraph is

$$\begin{aligned} \binom{n}{k} \sum_{p+q+r=n-k} \binom{n-k}{p,q,r} \binom{(m-2)n}{f-2k-p-q} &= \binom{n}{k} \sum_{p+q+r=n-k} \binom{n-k}{r} \binom{n-k-r}{p} \binom{(m-2)n}{f-n-k+r} \\ &= \binom{n}{k} \sum_r \binom{n-k}{r} \binom{(m-2)n}{f-n-k+r} \sum_p \binom{n-k-r}{p} \\ &= \binom{n}{k} \sum_r 2^{n-k-r} \binom{n-k}{r} \binom{(m-2)n}{f-n-k+r} \end{aligned}$$

as desired. □

For even modestly large **B**, computing equation (2) involves values larger than can be handled by some programs. In practice, we use logs to make these computations practical.

We now show that the sum in the numerator of this probability is related to the famous Jacobi orthogonal polynomials. This sum is a terminating hypergeometric series. Given a real number a and a nonnegative integer r the corresponding *Pochhammer symbol* or *rising factorial* is

$$(a)_r = a(a+1)(a+2)\cdots(a+r-1).$$

Note that if a is an integer with $-r < a \le 0$ then $(a)_r = 0$ because the product contains 0 as a factor. Given real numbers $a_1, a_2, \ldots, a_p$ and $b_1, b_2, \ldots, b_q$ as well as a variable z, the corresponding *hypergeometric series* is

$${}_pF_q\left[\begin{array}{cccc} a_1 & a_2 & \ldots & a_p \\ b_1 & b_2 & \ldots & b_q \end{array}; z\right] = \sum_{r\ge 0} \frac{(a_1)_r (a_2)_r \cdots (a_p)_r}{(b_1)_r (b_2)_r \cdots (b_q)_r} \frac{z^r}{r!}.$$

Note that if any of the a_i are negative integers then, because of the remark above, this series will terminate and become a polynomial in z.

To convert a binomial coefficient into Pochhammer symbols, we write

$$\begin{aligned} \binom{n}{r} &= \frac{(n)(n-1)\cdots(n-r+1)}{r!} \\ &= \frac{(-1)^r(-n)(-n+1)\cdots(-n+r-1)}{(1)_r} \\ &= \frac{(-1)^r(-n)_r}{(1)_r}. \end{aligned}$$

The following identity will also be useful

$$\begin{aligned} (a)_{b+r} &= (a)(a+1)\cdots(a+b-1)\times(a+b)(a+b+1)\cdots(a+b+r-1) \\ &= (a)_b(a+b)_r. \end{aligned}$$

We now return to the sum in the numerator of equation (2). We will ignore the factor of 2^{n-k} since it is constant with respect to the sum and so can be pulled outside. For simplicity of calculation we will also use the substitutions

$$s = (m-2)n, \qquad t = f-n-k.$$

Thus we have

$$\begin{aligned}
\sum_r 2^{-r}\binom{n-k}{r}\binom{(m-2)n}{f-n-k+r} &= \sum_r \binom{n-k}{r}\binom{s}{t+r}(1/2)^r \\
&= \sum_r \frac{(-1)^r(k-n)_r}{(1)_r}\cdot\frac{(-1)^{t+r}(-s)_{t+r}}{(1)_{t+r}}(1/2)^r \\
&= (-1)^t\sum_r \frac{(k-n)_r(-s)_t(-s+t)_r}{(1)_t(t+1)_r}\frac{(1/2)^r}{(1)_r} \\
&= \frac{(-1)^t(-s)_t}{(1)_t}\sum_r \frac{(k-n)_r(-s+t)_r}{(t+1)_r}\frac{(1/2)^r}{r!} \\
&= \binom{s}{t}\,{}_2F_1\left[\begin{matrix} k-n & -s+t \\ & t+1 \end{matrix}; \frac{1}{2}\right]
\end{aligned}$$

We are indebted to Marko Petkovšek [personal communication] for pointing out that this ${}_2F_1$ is, up to a factor, a specialization of a Jacobi polynomial. Given a nonnegative integer ℓ and real numbers α,β the associated *Jacobi polynomial* is

$$P_\ell^{(\alpha,\beta)}(z) = \binom{\alpha+\ell}{\ell}\,{}_2F_1\left[\begin{matrix} -\ell & \ell+\alpha+\beta+1 \\ & \alpha+1 \end{matrix}; \frac{1-z}{2}\right]$$

To make these ${}_2F_1$ polynomials agree we can let $\ell = n-k$, $\alpha = t = f-n-k$,

$$\beta = -s+t-(\ell+\alpha+1) = k-(m-1)n-1$$

and $z=0$. With these substitutions we get

$$\sum_r 2^{-r}\binom{n-k}{r}\binom{(m-2)n}{f-n-k+r} = \frac{\binom{(m-2)n}{f-n-k}}{\binom{f-2k}{n-k}}\,P_{n-k}^{(f-n-k,\;k-(m-1)n-1)}(0).$$

S1.2 Fixed Row Model (FRM)

Let the *fixed row model* constrain all $\mathbf{B}^* \in \mathscr{B}^{\mathrm{FRM}}$ to have the same row sums as $\mathbf{B}$.

Theorem S1.2. *Under the fixed row model, the distribution of P_{ij}^* for $i \neq j$ is hypergeometric and satisfies*

$$\Pr(P_{ij}^* = k) = \frac{\binom{r_j}{k}\binom{n-r_j}{r_i-k}}{\binom{n}{r_i}}.$$

Proof. The total number of ways to pick r_i of the n columns for ones in the ith row and r_j of the n columns for ones in the jth row is

$$\binom{n}{r_i}\binom{n}{r_j} = \binom{n}{r_i}\frac{n!}{r_j!(n-r_j)!}. \tag{3}$$

So that will go in the denominator of the desired probability.

For the numerator we follow the same line of reasoning as in the previous proof, where the parameters therein can be expressed as

$$\begin{aligned}
p &= r_i - k, \\
q &= r_j - k, \\
r &= n - r_i - r_j + k.
\end{aligned}$$

So we have a total of

$$\binom{n}{k}\binom{n-k}{p,q,r} = \frac{n!}{k!(r_i-k)!(r_j-k)!(n-r_i-r_j+k)!} \tag{4}$$

choices.

Dividing equation (4) by (3) and cancelling $n!$ gives

$$\Pr(P^*_{ij}=k) = \frac{\frac{r_j!}{k!(r_j-k)!}\cdot\frac{(n-r_j)!}{(r_i-k)!(n-r_i-r_j+k)!}}{\binom{n}{r_i}} = \frac{\binom{r_j}{k}\binom{n-r_j}{r_i-k}}{\binom{n}{r_i}}.$$

as desired. □

S1.3 Distribution of projection edge weights under the Fixed Column Model (FCM)

Let the *fixed column model* constrain all $\mathbf{B}^* \in \mathscr{B}^{\text{FCM}}$ to have the same column sums as $\mathbf{B}$.

Let $X_1,\ldots,X_n$ be independent Bernoulli random variables. Let the probability of success for X_i be

$$\Pr(X_i=1) = p_i.$$

The random variable

$$X = X_1+\cdots+X_n \tag{5}$$

is said to have the *Poisson binomial distribution* with parameters $p_1,\ldots,p_n$.

Theorem S1.3. *Under the fixed column model, the distribution of P^*_{ij} for $i\neq j$ is Poisson binomial with parameters*

$$p_1 = \frac{c_1(c_1-1)}{m(m-1)},\ p_2 = \frac{c_2(c_2-1)}{m(m-1)},\ \ldots,\ p_n = \frac{c_n(c_n-1)}{m(m-1)}.$$

Proof. The B^*_{ik} are all either zero or one and are independent in different columns when only the column sums are fixed. So as k varies, the products $B^*_{ik}B^*_{jk}$ are independent Bernoulli random variables. Comparing equations (1) and (5), we see that the distribution of P^*_{ij} is Poisson binomial.

If column k has column sum $c=c_k$ then all zero-one vectors with sum c are equally likely for that column of $\mathbf{B}^*$. So there are $\binom{m}{c}$ possible kth columns. The number of ways to have a success is the number of possible columns which have ones in both positions i and j where $i\neq j$. So the number of choices is the number of ways to choose the remaining $c-2$ ones in that column from the other $m-2$ positions, that is, $\binom{m-2}{c-2}$. Thus

$$p_k = \Pr(B^*_{ik}B^*_{jk}=1) = \frac{\binom{m-2}{c-2}}{\binom{m}{c}} = \frac{c(c-1)}{m(m-1)}$$

which finishes the demonstration. □

S1.4 Stochastic Degree Sequence Model (SDSM)

In the *stochastic degree sequence model*, $\mathscr{B}^{\text{SDSM}}$ consists of all binary $m\times n$ matrices. A method is then chosen to generate probabilities p^*_{ik}. Finally, matrices $\mathbf{B}^*\in\mathscr{B}^{\text{SDSM}}$ are generated using these probabilities for independent Bernoulli trials, where B^*_{ik} is filled with a one with probability p^*_{ik} and zero otherwise.

Theorem S1.4. *Under the stochastic degree sequence model, the distribution of P^*_{ij} for $i\neq j$ is Poisson binomial with parameters*

$$p_1 = p^*_{i1}p^*_{j1},\ \ldots,\ p_n = p^*_{in}p^*_{jn}.$$

Proof. The fact that the distribution is Poisson binomial follows immediately from the independence assumption on the $\Pr(B^*_{ik})$ and equation (1). Furthermore, the probability that the kth variable is one is

$$p_k = \Pr(B^*_{ik}B^*_{jk}=1) = \Pr(B^*_{ik}=1)\Pr(B^*_{jk}=1) = p^*_{ik}p^*_{jk}.$$

So we are done. □