Chapter 3
Algebraic Statistics, Tables, and Networks: The Fienberg Advantage

Elizabeth Gross, Vishesh Karwa, and Sonja Petrović

3.1 Introduction

Stephen Fienberg's early work on contingency tables [BFH74] relies on using intrinsic model geometry to understand the behavior of estimation algorithms, asymptotics, and model complexity. For example, in [Fie70], Fienberg gives a geometric proof of the convergence of the iterative proportional fitting algorithm for tables with positive entries. The result in [Fie70] relies on his study of the geometry of $r \times c$ tables in [Fie68] and his and Gilbert's geometric study of 2×2 tables [FG70]. This approach to understanding models would eventually fit within the field of algebraic statistics, a general research direction that would take hold in the 2000s, over 25 years after the publication of [Fie70] and the 1974 edition of Bishop, Fienberg, and Holland's book [BFH74], whose cover displayed the independence model for 2×2 tables as an algebraic surface.

The term "algebraic statistics" was coined in 2001 [PRW01] and generally refers to the use of broader algebraic—non-linear—and geometric—non-convex—tools in statistics. While the use of algebra and geometry had been long present in statistics, before the 2000s, linear algebra and convex geometry were the main tools used consistently. The field of algebraic statistics is now a branch of mathematical statistics that relies on insights from computational algebraic geometry, combinatorial

E. Gross
University of Hawai'i at Mānoa, Honolulu, HI, USA
e-mail: egross@hawaii.edu

V. Karwa
Temple University, Philadelphia, PA, USA
e-mail: vishesh@temple.edu

S. Petrović (✉)
Illinois Institute of Technology, Chicago, IL, USA
e-mail: sonja.petrovic@iit.edu

geometry, and commutative algebra to improve statistical inference. As algebraic statistics matured and caught the attention of many researchers, Fienberg and his students and collaborators reformulated several fundamental statistical problems, e.g. existence of maximum likelihood estimators and ensuring data privacy, into the language of polyhedral and algebraic geometry. Today Fienberg's intuition and influence remain central to one of the principal applications in algebraic statistics: testing goodness of fit of log-linear models for discrete data. Within the last decade or so, much of his work in this area focused on log-linear *network* models. In this regard, Fienberg defined new models, explained how to represent relational data as contingency tables in order to apply tools from categorical data analysis, and addressed the problems of estimation, model fit, and model selection. This paper presents a brief overview of this line of work heavily influenced by Fienberg's vision, which continues to inspire us.

3.2   Geometry and Algebra of Log-Linear Models

Let us recall the basics and fix notation. Let $\mathcal{I} = [d_1] \times \cdots \times [d_k]$ be a finite set that indexes *cells* in a contingency table $u \in \mathbb{Z}_{\geq 0}^{d_1 \times \cdots \times d_k}$. The $(i_1, \ldots, i_k)$-cell counts the number of occurrences of the event $\{X_1 = i_1, \ldots, X_k = i_k\}$ for k categorical random variables with X_i taking values on a finite set $[d_i] := \{1, \ldots, d_i\}$. Log-linear models are probability distributions on the discrete set $\mathcal{I}$ whose sufficient statistics are given by marginals, i.e. subtables of the table u obtained by summing u across a subset the index set $\mathcal{I}$; since marginalization is a linear map, it can be presented as matrix multiplication. Specifically, a log-linear model for $\mathcal{I}$ is a linear exponential family defined by an $m \times |\mathcal{I}|$ matrix A, called the *design matrix*, taking the following form:

$$P_\theta(U = u) = \exp\{\langle Au, \theta \rangle - \psi(\theta), \tag{3.1}$$

where $\theta \in \mathbb{R}^m$ is the vector of model parameters and $\psi(\theta)$ the normalizing constant. Note that specifying the matrix A completely specifies the contingency table model for $X_1, \ldots, X_k$, as it determines the vector of minimal sufficient statistics Au for the linear exponential family in (3.1). As is customary in algebraic statistics, we will denote the model (3.1) by $\mathcal{M}_A$.

Let us consider one of Fienberg's early favorite examples: the model of independence of two categorical random variables X_1 and X_2. Here, A is a $(d_1 + d_2) \times d_1 d_2$ matrix of the following form, where the first d_1 rows each have d_2 ones and the last d_2 rows contain d_1 copies of the $d_2 \times d_2$ identity matrix:

$$
A = \left[
\begin{array}{cccc|cccc|c|cccc}
1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 1 & 1 & \cdots & 1 \\
\hline
1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & \cdots & 1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 & 0 & 1 & \cdots & 0 & \cdots & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 1 & \cdots & 0 & 0 & \cdots & 1
\end{array}
\right].
$$

The sufficient statistic for $\mathcal{M}_A$ is the vector of marginal counts (that is, table row and column sums). For a contingency table u, these counts are computed as:

$$
A \begin{bmatrix} u_{11} \\ \vdots \\ u_{d_1 d_2} \end{bmatrix} = \begin{bmatrix} u_{1+} & \ldots & u_{+d_2} \end{bmatrix}. \tag{3.2}
$$

In [Fie68], Fienberg describes the geometry of $\mathcal{M}_A$ in detail, describing the model of independence as the intersection of the *manifold of independence* with the probability simplex. In algebraic geometry, the manifold of independence is a Segre variety, a categorical product, which Fienberg describes explicitly by detailing the linear spaces corresponding to the product fibers, or in other words, every preimage of the map described by Eq. (3.2). In addition, the defining equations of the Segre variety corresponding to the independence model are stated in [Fie68] in statistical terms (see Section 4 of [Fie68]). These equations, which are polynomial equations in indeterminates that represent joint cell probabilities, are a key ingredient to assessing model fit.

Indeed, assessing model fit for log-linear models, and consequently, log-linear network models, is possible due to a fundamental result in algebraic statistics that establishes a connection between model-defining polynomials and sampling from the conditional distributions of log-linear models. The model-defining polynomials of interest are generating sets of polynomial ideals called *toric ideals* [Stu96], §4 and §5. The essential component, which binds together the statistical and algebraic, is the vector of (minimal) sufficient statistics for the log-linear exponential family, the vector Au in the definition above.

One way to perform goodness-of-fit testing for log-linear models, especially in sparse settings such as networks, is to perform Fisher's exact test (see e.g. Section 2.6 in [Agr92]). In many cases, however, it is infeasible to compute the exact conditional p-value, thus it is estimated using a Markov chain Monte Carlo (MCMC) method. The *exact conditional p-value* of a contingency table u is the probability that the Pearson's χ^2 statistic of a random data table is larger than

the Pearson's χ^2 statistic of the table u, conditional on the observed values of the sufficient statistics. The set of all tables with the same sufficient statistics as u is called *the fiber of u under the model $\mathcal{M}_A$* and is defined as follows:

$$\mathcal{F}_A(u) := \{v \in \mathbb{Z}_{\geq 0}^{d_1 \times \ldots \times d_k} : Au = Av\}.$$

The naming of the reference set $\mathcal{F}_A(u)$ is derived from algebraic geometry: a fiber of a point of the linear map defined by A is its preimage under that map; in this case, we are considering the set of non-negative integer points in the preimage of the sufficient statistics Au. In order to perform the MCMC method to estimate the exact conditional p-value, a set of moves must be given, and this set of moves must connect all elements in the fiber $\mathcal{F}_A(u)$ so that the conditional distribution on the fiber can be sampled properly. Such a set of moves is called a Markov basis.

Definition 2.1 A *Markov basis* of the model $\mathcal{M}_A$ is a set of tables $\mathcal{B} := \{b_1, \ldots, b_n\} \subset \mathbb{Z}^{d_1 \times \ldots \times d_k}$ for which

$$Ab_i = 0$$

and such that for any contingency table $u \in \mathbb{Z}_{\geq 0}^{d_1 \times \ldots \times d_k}$ and for any $v \in \mathcal{F}_A(u)$, there exist $b_{i_1}, \ldots, b_{i_N} \in \mathcal{B}$ that can be used to reach v from u:

$$u + b_{i_1} + \ldots + b_{i_N} = v$$

while remaining in the fiber at each step:

$$u + \sum_{j=0}^{N} b_{i_j} \in \mathcal{F}_A(u), \text{ for } j = 1 \ldots N.$$

Note that the last requirement simply means that each move needs to preserve non-negativity of cells. As an example, let us consider the independence model with $N = 2$, $d_1 = 3$, and $d_2 = 3$. Then the fiber $F_A(u)$ for any u is a collection of 3×3 tables. Examples of three different Markov moves for the independence model in this setting are

1	−1	0
−1	1	0
0	0	0

−1	0	1
0	0	0
1	0	−1

0	0	0
0	1	−1
0	−1	1

It is hard to check *a priori* whether a given set of moves does in fact form a Markov basis for the model. However, the following foundational result from algebraic statistics allows one to compute a Markov basis by computing a generating set of a polynomial ideal.

Theorem 2.2 ([DS98]) *A set of vectors* $\mathcal{B} = \{b_1, \ldots, b_n\}$ *is a Markov basis of the log-linear model* $\mathcal{M}_A$ *if and only if* *the corresponding set of binomials* $\{x^{b_i^+} - x^{b_i^+}\}_{i=1,\ldots,n}$ *generates the toric ideal* $I_A := (x^u - x^v : u - v \in \ker_{\mathbb{Z}} A)$.

Considering again the independence model with $N = 2$, $d_1 = 3$, and $d_2 = 3$, the binomials associated with the three tables above are

$$x_{11}x_{22} - x_{12}x_{21}, \quad x_{13}x_{31} - x_{11}x_{33}, \quad x_{22}x_{33} - x_{23}x_{32}.$$

One can check that these three polynomials are not enough to generate the ideal I_A, and thus more moves are needed for a Markov basis.

Theorem 2.2 is a powerful result that connects categorical data analysis to algebra. By connecting network analysis to categorical data analysis, Fienberg was able to use the full force of this theorem for testing model fit of statistical network models.

3.3 Log-Linear ERGMs and Goodness-of-Fit Testing

As stated in the editorial piece [PSY19], Fienberg took joy in rediscovering old concepts from new points of view that gave them new interpretations and wider applicability; this was evident not only from his research articles and conference presentations, but various interviews, see, for example, [Vie15]. We follow his lead in the way we define *log-linear network models*.

Generally, a statistical network model is a collection of probability distributions over $\mathcal{G}_n$, the set of all (un)directed graphs on n vertices. The Fienberg approach to the analysis of statistical network models, dating back to the late '70s and early '80s, relies on explicitly making the connection to categorical data analysis by viewing graphs as contingency tables. For example, in [FW81a], Fienberg and Wasserman view a directed graph with n vertices as a $n \times n \times 2 \times 2$ table Y where $Y_{ij00} = 1$ if there is no edge between vertex i and j, $Y_{ij11} = 1$ if there is a reciprocated edge between i and j, $Y_{ij10} = 1$ if there is a non-reciprocated edge from i to j, and $Y_{ij01} = 1$ if there is a non-reciprocated edge from j to i, and all entries are 0 otherwise. Using this $n \times n \times 2 \times 2$ table, Fienberg and Wasserman then describe nine variants of a simple statistical network model, called the p_1 model [HL81], in terms of table marginals and show how these models can be fit using a version of iterative proportional scaling for multidimensional contingency tables. In addition, they also develop a variant of the p_1 model for K subgroups determined by nodal attributes, by collapsing the $n \times n \times 2 \times 2$ into a $K \times K \times 2 \times 2$ table; a precursor to the directed stochastic blockmodels.

The p_1 model and its variants described by Fienberg and Wasserman in [FW81b] are examples of log-linear ERGMs. Log-linear ERGMs are exponential family random graph models with a log-linear interpretation. Another example of log-linear models is stochastic blockmodels, which are given a contingency representation in

[FMW85]. Following the contingency table framework of the Fienberg approach, to define a log-linear ERGM, one chooses an embedding $\phi : \mathcal{G}_n \to \mathbb{R}^\ell$ such that for all $G = (V, E)$ we have $\phi(G) = \sum_{e \in E} \phi(e)$, and implicitly uses the embedding ϕ to represent G as a vector. For example, for directed graphs, a reasonable embedding would embed $\mathcal{G}_n$ into $\mathbb{R}^{n^2}$ and G would be represented by its vectorized adjacency matrix, while for undirected graphs $\mathbb{R}^{\binom{n}{2}}$ would work equally well. For directed graphs, a suitable embedding rooted in [FW81a] (see also [FW81b]) maps $\mathcal{G}_n$ into $\mathbb{R}^{n \times n \times 2 \times 2}$ by representing graphs by their vectorized $n \times n \times 2 \times 2$ Fienberg-Wasserman table as described above or a vectorized table of size $\binom{n}{2} \times 2 \times 2$ after removing redundant cells. These embeddings allow us to refer to graphs as vectors.

An exponential family random graph model, or an ERGM for short, is a collection of probability distributions on $\mathcal{G}_n$ that places the following probability on each graph $G \in \mathcal{G}_n$:

$$P_\theta(G) = Z(\theta)e^{\theta \cdot t(G)}, \tag{3.3}$$

where G is uniquely represented as a vector in $\mathbb{R}^\ell$, θ is a row vector of parameters of length q, the map $t : \mathbb{R}^\ell \to \mathbb{R}^q$ computes the sufficient statistics, and $Z(\theta)$ is a normalizing constant. The image of the sufficient statistic map t is a vector in which each entry is a network statistic used to specify the model, such as edge count, degree of a given vertex, number of edges in a given block of vertices, etc. When the sufficient statistic is a linear function on the entries of a natural contingency table representation of the graph, as in degree-based models or stochastic blockmodels, then the sufficient statistic map t can be described with a design matrix A and the model (3.3) takes the form of (3.1). When this happens, we call the model a log-linear ERGM.

Definition 3.1 We call an exponential family random graph model a *log-linear ERGM* if the sufficient statistic map t in the ERGM specification (3.3) is a *linear map* $t : \mathbb{R}^\ell \to \mathbb{R}^q$ from the space of graphs to the space of the minimal sufficient statistics of the model.

Log-linear ERGMs include degree-based models such as the β-model, models that include effects for reciprocity, such as p_1 models, and models for data with categorical nodal attributes, such as stochastic blockmodels. Since the sufficient statistic t is a linear map, dyadic independence is implied for a log-linear ERGM. Dyadic independence is another way to say that for each pair of vertices, i and j, the edge configuration (e.g. no edge between i and j, directed edge from i to j, directed edge from j to i, bidirected edge between i to j) is independent from the edge configuration between any other pair of vertices. Thus, we can fully specify a log-linear ERGM by specifying the distribution over each set of dyadic configurations.

Example 3.2 (Stochastic Blockmodels) Extremely popular in practice, this family of log-linear ERGMs models networks whose nodes are partitioned into groups–

blocks—according to some nodal attributes. For a directed network, each dyad can be in one of four states represented as follows: $(0, 0)$ represents no edge, $(1, 0)$ an edge from i to j, $(0, 1)$ an edge from j to i, and $(1, 1)$ a bidirected edge. Note that if the network is undirected, the model simply collapses to having only two dyadic states: $(0,0)$ and $(1,1)$. Denote by p_{ijkl} the probability of the dyad (i, j) to be in state (k, l).

Edge formation is governed by what Fienberg and Wasserman call choice parameters, denoted by δ^{rs}, and reciprocity effects ρ^{rs}. These parameters are defined on the level of blocks. In addition, Fienberg liked the use of an additional set of parameters λ_{ij} for normalization: ensuring that each dyad is observed in only one state at a time. Specifically, the model was defined in [FMW85] as follows:

$$\log p_{ij00} = \lambda_{ij} \tag{3.4}$$

$$\log p_{ij10} = \lambda_{ij} + \delta^{b(i)b(j)}$$

$$\log p_{ij01} = \lambda_{ij} + \delta^{b(j)b(i)}$$

$$\log p_{ij11} = \lambda_{ij} + \delta^{b(i)b(j)} + \delta^{b(j)b(i)} + \rho^{b(i)b(j)},$$

where each node in the graph belongs to one of K blocks, $B_1, \ldots, B_K$, and $b(i)$ denotes the (known) block assignment of vertex i.

There are various special cases of stochastic blockmodels. For example, we can choose $\delta^{rs} = \delta + \alpha^r + \beta^s$ and $\rho^{rs} = \rho$, as in ([FMW85], Equation (2.10)). Then the model is the following special case:

$$\log p_{ij00} = \lambda_{ij} \tag{3.5}$$

$$\log p_{ij10} = \lambda_{ij} + \delta + \alpha^{b(i)} + \beta^{b(j)}$$

$$\log p_{ij01} = \lambda_{ij} + \delta + \alpha^{b(j)} + \beta^{b(i)}$$

$$\log p_{ij11} = \lambda_{ij} + 2\delta + \alpha^{b(i)} + \alpha^{b(j)} + \beta^{b(j)} + \beta^{b(i)} + \rho.$$

In this setting, the sufficient statistics counted by the map t are the number of configurations for each dyad, the total number of edges, block in-degrees, block out-degrees, and the total number of reciprocated edges in the network. Here, the in-degree of block B_j (the number of edges that enter the block) is computed by adding in-degrees of all the nodes in the block, $d_{B_j}^{in} = \sum_{i \in B_j} d_i^{in}$. The out-degree is defined similarly.

Let us consider the space of directed graphs on $n = 3$ vertices $V = \{1, 2, 3\}$ with block structure $B_1 = \{1, 2\}$, $B_2 = \{3\}$, the design matrix A defining the linear map t would be as follows:

$$A = \begin{bmatrix} 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1 \\ 0\ 1\ 1\ 2\ 0\ 1\ 1\ 2\ 0\ 1\ 1\ 2 \\ 0\ 1\ 1\ 2\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1 \\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 1 \\ 0\ 1\ 1\ 2\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 1 \\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1 \\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1 \end{bmatrix}.$$

Let G be represented as a vector of length 12, where the first four entries correspond to the four possible dyadic configurations between vertices 1 and 2, the second four correspond to the four possible dyadic configurations between vertices 1 and 3, and the third four correspond to the four possible dyadic configurations between vertices 2 and 3. Then the first three rows of A count the number of configurations for each dyad (for simple graphs this count should always be one), the fourth row of A counts the total number of edges in G, the fifth and sixth rows count the block in-degrees, the seventh and eighth rows count the block out-degrees, and last row counts the total number of reciprocated edges in the network.

Example 3.3 (p_1 Models) The p_1-model for directed graphs was introduced by Holland and Leinhardt [HL81] and extended by Fienberg and Wasserman [FW81b]. It is a model that includes two nodal effects, one for popularity and another for expansiveness, and a reciprocation effect. Following Example 3.2, we denote p_{ijkl} the probability of the dyad (i, j) to be in state $(k, l) \in \{0, 1\}^2$. The dyadic probabilities for the p_1-model are specified as follows:

$$\log p_{ij00} = \lambda_{ij}, \tag{3.6}$$
$$\log p_{ij10} = \lambda_{ij} + \alpha_i + \beta_j + \delta,$$
$$\log p_{ij01} = \lambda_{ij} + \alpha_j + \beta_i + \delta,$$
$$\log p_{ij11} = \lambda_{ij} + \alpha_i + \alpha_j + \beta_j + \beta_j + 2\delta + \rho_{ij}.$$

The parameters α_i and β_i record the rates at which the node i sends and receives links, while ρ_{ij} controls reciprocation. Note that the model specification includes additional parameters. Namely, there is δ, a density parameter and $\binom{n}{2}$ dyadic effects, λ_{ij}, which are normalizing constants as described in Example 3.2.

The p_1 model has three main variants that capture different reciprocation effects: zero reciprocation, constant reciprocation, and dyad-specific reciprocation, also referred to as differential reciprocity. For example, in the constant reciprocation case, $\rho_{ij} = \rho$ for all i, j. The sufficient statistics for the p_1-model with constant reciprocation consists of the number of edges, the in-degree sequence, the out-degree sequence, and the number of reciprocated edges.

The design matrix A for several small examples can be found in [PRF10].

While Fienberg's work allows for a transfer of technology from the contingency table literature to networks, the interpretability of models and model equivalence was not always immediate and required additional insight. As noted in [Hab81] and reiterated by Fienberg and co-authors in [FPR10], even simple ERGMs, such as the p_1 model, pose fundamental challenges to the practitioner even within the contingency table setting, especially when testing model goodness of fit. For example, as pointed out by Fienberg and co-authors in [PRF10], many network models such as the p_1 model are theoretically problematic, since, in these models, the number of parameters depends on the number of vertices. This means that as the population size grows, the model complexity also increases, unlike traditional statistical models, where the complexity is often fixed and independent of the sample size. Another challenge to using existing traditional methods from categorical data analysis in goodness-of-fit testing and model selection is that the data are naturally sparse, making standard asymptotic methods unreliable. Under such conditions, exact conditional tests are preferred for model selection and goodness-of-fit testing. However, as mentioned in the previous section, exact conditional tests pose their own difficult problems for networks, mainly since the exact distribution is over a space that is combinatorially large, and in most cases, innumerable. Finally, the contingency tables described by Fienberg and Wasserman are highly redundant and are subject not only to symmetric constraints but also product multinomial constraints, e.g. since each dyad can only be in one of the four possible configurations $Y_{ij00} + Y_{ij10} + Y_{ij01} + Y_{ij00} = 1$ for all $i \neq j$.

Fienberg was able to provide a work-around to the difficulties posed by exact conditional tests by using Markov bases and algebraic statistics. In 1998, Sturmfels and Diaconis published Theorem 2.2 [DS98]. Afterwards, the idea of using toric ideals for goodness-of-fit testing for various log-linear models gained traction, and about 10 years later, Fienberg, Petrović, and Rinaldo applied Theorem 2.2 to three of the main variants of the p_1 model in [PRF10], essentially introducing algebraic statistics to the field of network analysis. In particular, they describe Markov moves for each variant and its corresponding simplified model (the model obtained after forgetting the normalizing parameters). The work not only provided a breakthrough in goodness-of-fit testing for log-linear ERGMs but also had an impact in combinatorial commutative algebra. The toric ideals corresponding the p_1 model are connected to toric ideals of graphs, defined in [SVV94] (see also [Vil95] and [OH00]) and more generally, toric ideals of hypergraphs. Indeed, the results of [PRF10] provided an applied motivation for the systematic study of toric ideals of hypergraphs in the field of combinatorial commutative algebra (see e.g. [GP13, HT08, PS14, PTV19]).

Before [PRF10], Markov bases were always used in the setting where the only constraints on the contingency tables were that every entry needed to be nonnegative. However, in the network setting, particularly in the case of a single sociometric relation, cells of the contingency tables are either 0 or 1 and there is only a single observation for each dyad. This was the first time in the Markov bases literature that sampling constraints of this form were directly incorporated in the study of Markov bases (note that related work [HT10], and relevant for the

problem here, on connecting tables with 0/1 entries appeared in the same volume). Fienberg and co-authors were able to effectively handle the network constraints by computing a minimal generating set of this ideal first and then by removing basis elements that violate the condition of one observation per dyad, which results in a product multinomial sampling scheme. Fienberg's idea of adding the normalizing parameters λ_{ij}s to the models directly enforced the 0/1 constraint in sampling. In particular, if a move produced by a Markov basis computation is applicable to the observed network, in that it does not attempt to remove edges that are not present, then it will follow the sampling constraint in that it will not add an edge where there is one already. Examples of applicable and inapplicable moves for the p_1 model and the Sampson data depicted in Fig. 3.1 are shown in Figs. 3.2, 3.3, and 3.4.

It should be noted that Fienberg's idea to prune non-applicable moves was novel and paved the way for practical implementation of a goodness-of-fit test for log-linear ERGMs [GPS16]. Indeed, in [DFR$^+$08], Fienberg and co-authors observed that Markov bases are *data independent*, meaning that they describe all the moves required to guarantee connectedness of any fiber; in other words, Markov bases do not depend on the observed network, only the model. This observation can help transform otherwise unwieldly sets of Markov moves into smaller and easier to manipulate sets of moves. For example, without pruning, the naive computation of a Markov basis for the p_1 model with constant reciprocation with 4 nodes has 80,610 moves, while the pruned Markov basis consisting of only elements applicable
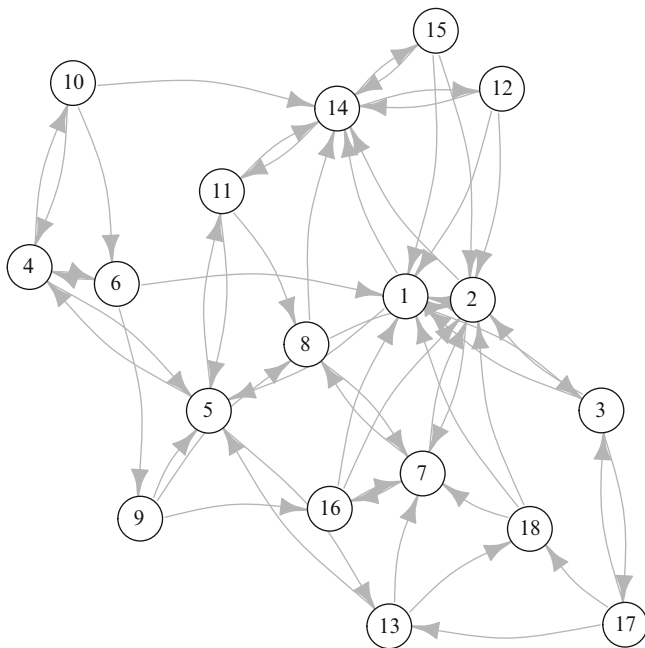

Fig. 3.1 The directed graph representation of Sampson's monastery dataset [Sam69]
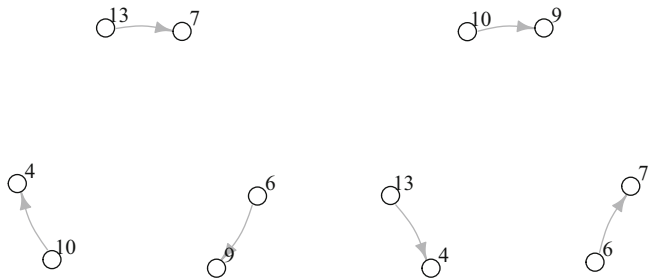
Fig. 3.2 A move from the Markov basis for the p_1 model with zero reciprocation. **Left:** Edges to remove. **Right:** Edges to add. This move can be applied to the network in Fig. 3.1 as it preserves node in-degrees and out-degrees. Note that edge $4 \leftarrow 10$ is reciprocated in the data, so after the move is applied, the total number of reciprocated edges is reduced by 1
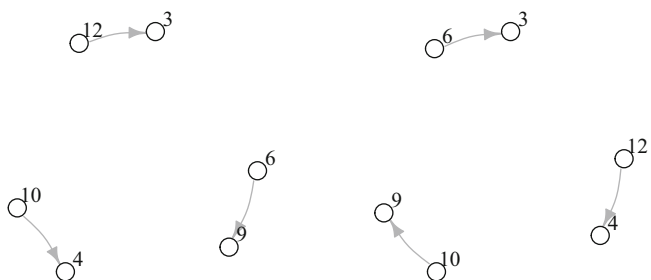
Fig. 3.3 A move from the Markov basis for the p_1 model with zero reciprocation. **Left:** Edges to remove. **Right:** Edges to add. However, this move cannot be applied to the network in Fig. 3.1 as the dyad $(3, 12)$ is observed in the state $(0, 0)$ rather than $(1, 0)$; that is, the edge $12 \rightarrow 3$ is not present, so it cannot be removed
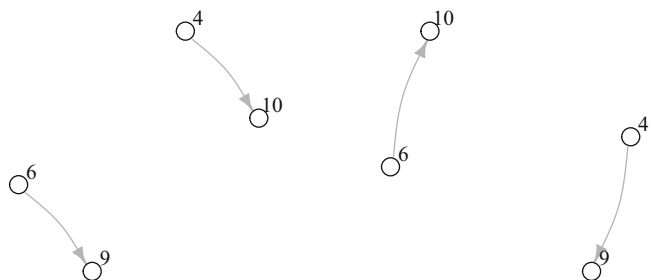
Fig. 3.4 A move from the Markov basis for the p_1 model with constant reciprocation. **Left:** Edges to remove. **Right:** Edges to add. This move can be applied to the network in Fig. 3.1. It preserves the number of reciprocated edges: the dyad $(4, 10)$ changes from reciprocated to directed edge, but the dyad $(6, 10)$ changes from directed to reciprocated

to simple networks and decomposed into essential building blocks, computed in [PRF10], has about 10 moves.

This idea was a starting point of departure from the algebraic status-quo approach, which is traditionally blind to data and as such leads to slow mixing times of the resulting Markov chains. After Fienberg's work in [PRF10], the main computational challenge remained open to make the theory useful for network data in practice. To this end, working within the *data dependent paradigm*, [GPS16] developed an algorithm to approximate the exact conditional p-value for log-linear ERGMs and implemented the algorithm for the p_1 model. The algorithm approximates the exact conditional p-value by using applicable Markov moves generated on an as-needed basis to move around the fiber. At each network in the chain, a goodness-of-fit statistic is computed and compared to the observed network. This adapted Metropolis-Hastings algorithm is described in detail in [GPS16].

For exposition and illustration of theoretical ideas, Fienberg saw great value in small data; for example, Sampson's monastery dataset [Sam69] (see Fig. 3.1) was the running example in [ABFX09] and also was an example dataset in Fienberg's survey of statistical network models with Goldenberg et al. [GZFA10]. Thus, the paper [GPS16] revisited the Sampson's monastery dataset and tested the fit of the p_1 model. The Sampson's monastery dataset, in Fienberg's words, was one of the reasons behind the construction of the Holland-Leinhardt p_1 model in the first place. However, this is not to say that Fienberg was not concerned with challenging big data problems, and the ideas described here do scale, e.g. [KP16] tests model fit for the β and p_1 models on co-authorship and citation networks of statisticians [JJ16] of about 3000 authors and 3000 papers. Finally, Fienberg was also an avid supporter of applications of statistics; it was he who suggested to the third author to study the Japanese corporate data set from *The New York Times* back in 2014 from the point of view of the p_1 model. As [Pet19] illustrates, the goodness-of-fit test confirms the Japanese Prime Minister's intuition.

3.4 Beyond Simple Graphs

The rapid increase of data-collecting mechanisms in recent decades has resulted in complex forms of network data, including multivariate and multi-agent networks. Still, in the growing field of network science, such data are still often represented in the form of a simple graph, mainly because simple random graph models are assumed to be easier to estimate and fit. However, such simplifications are not necessary with Fienberg's view of networks as contingency tables. This is because neither multiple observations on a single dyad, which increase cell counts in the table, nor multiway interactions, which increase table dimensions, present an additional layer of difficulty for estimation or testing model fit. On the contrary, the sampling algorithms based on Markov bases become easier, because the sampling constraint is relaxed.

One example of this simplification is when experiment data consisting of multiple observations is summarized as a simple graph by way of thresholding— preserving an edge between two nodes only if it was observed at least a fixed number of times. This happens very often in neuroscience and chemical reaction experiments. It is also often applied to social interactions data such as the co-authorship network in Fig. 3.5 below. In the co-authorship network, an edge (i, j) is present in the co-authorship graph if at least 4 joint papers were written by authors i and j. Why 4? This thresholding number of choices seems arbitrary at best (changing it may drastically change the structure of the graph), is done out of convenience, and in many applications results in significant information loss.

In [FMW80] and [FMW85], Fienberg, Meyer, and Wasserman set up the log-linear framework for multivariate directed graphs. We can think of a multivariate graph as a multi-layered network. For example, in the technical paper [FMW80], Fienberg, Meyer, and Wasserman consider a community of individuals and networks formed by three relations, information, money, and support; these relations are referred to as sociometric generators. In [FMW85], the authors develop extensions of [FMW80] to allow for covariates. Motivated by this, [RPF13] (see also [RPF10] for further details) study the *generalized β-model* for random graphs. They consider the log-linear model for undirected graphs whose sufficient statistics are node degrees, but they allow for the possibility that each dyad in the network be sampled a different number of times. Applying the geometric and combinatorial properties of log-linear models under product multinomial sampling schemes from [FR12], they derive necessary and sufficient conditions for MLE existence and discuss its asymptotics.

The second example of data simplification is also well illustrated using co-authorship data: it is common for multiway interactions to be collapsed to their
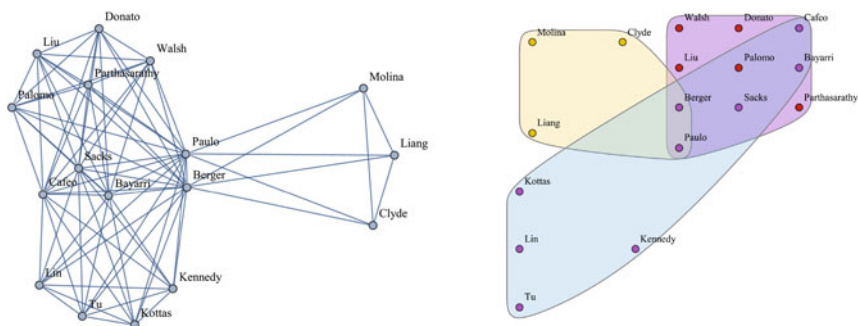
→ A hyperedge of size k corresponds to a paper with k authors

→ Hyper graph degree of i = number of hyperedges i participates in

Fig. 3.5 The graph and the hypergraph representing *the same* co-authorship data. In the graph on the left, it is not clear at all that the data corresponds to exactly 3 published papers, for example, which is clear in the hypergraph on the right. Graphs in the figure adapted from [KP16]
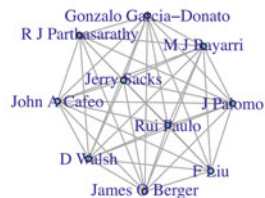
induced pairwise interactions. However, most of the time, capturing the multiway interaction is more realistic and informative. Figure 3.5 shows how the information from data that naturally comes in form of a hypergraph is obscured when represented by the underlying graph. Indeed, once the first co-authorship network data for statisticians was collected and released in [JJ16], the last two authors set out to explore the effects of these data summaries. In [KP16], it is shown what information is lost by reducing the data to a simple graph by presenting multi-observation table data summaries, core-decomposition summaries, and hypergraph data summaries, all of which suggest possibly different conclusions than those from the derived simple graphs. For example, the authors considered the inner-most clique, that is, the largest completely connected subgraph, of the co-authorship graph where there is an edge between two authors if they coauthored at least 4 joint papers. While these authors have many neighbors, i.e. their nodes have a high degree, we argue that degree-based modeling on the simple graph does not capture everything behind the data. Specifically, Fig. 3.6 shows that the secret behind these cliques is a single many-author paper in both cases.

With the issues illustrated in [KP16] in mind, Fienberg and co-authors introduce the β model for random hypergraphs in [SSR$^+$14], which builds upon and generalizes the well-studied β model for random graphs. Directly motivated by Fienberg's earlier foundational work, the authors provide two algorithms for fitting the model parameters, an iterative proportional scaling algorithm, and a fixed point algorithm. Furthermore, Fienberg and co-authors prove that both algorithms converge if the maximum likelihood estimator (MLE) exists, and they provide algorithmic and geometric ways of dealing the issue of MLE existence—one of Fienberg's favorite problems.

Fig. 3.6 The inner-most clique of each of the two co-authorship graphs studied in [JJ16]: each corresponds to a many-author paper. Graphs in the figure adapted from [KP16]

3.5   Closing Remarks

Fienberg always used to say how problems never go away, one just sees them under a new light. In this survey of Fienberg's work connecting categorical data analysis and algebraic statistics to network science, we hope we illustrated, in essence, this sentiment of continual discovery and rediscovery.

References

ABFX09   Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. In *Advances in Neural Information Processing Systems*, pages 33–40, 2009.

Agr92   Alan Agresti. A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153, 1992.

BFH74   Yvonne M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. Springer, 1974.

DFR$^+$08   Adrian Dobra, Stephen E. Fienberg, Alessandro Rinaldo, Aleksandra Slavković, and Yi Zhou. Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation and disclosure limitation. In *In IMA Volumes in Mathematics and its Applications: Emerging Applications of Algebraic Geometry*, pages 63–88. Springer Science+Business Media, Inc, 2008.

DS98   Persi Diaconis and Bernd Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, 26(1):363–397, 1998.

FG70   Stephen E. Fienberg and John P Gilbert. The geometry of a two by two contingency table. *Journal of the American Statistical Association*, 65(330):694–701, 1970.

Fie68   Stephen E. Fienberg. The geometry of an $r \times c$ contingency table. *The Annals of Mathematical Statistics*, 39(4):1186–1190, 1968.

Fie70   Stephen E. Fienberg. An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, 41(3):907–917, 1970.

FMW80   Stephen E Feinberg, Michael M Meyer, and Stanley Wasserman. Analyzing data from multivariate directed graphs: An application to social networks. Technical report, University of Minnesota, 1980.

FMW85   Stephen E. Fienberg, Michael M. Meyer, and Stanley S. Wasserman. Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80(389):51–67, 1985.

FPR10   Stephen E. Fienberg, Sonja Petrović, and Alessandro Rinaldo. *Algebraic statistics for p_1 random graph models: Markov bases and their uses*, volume Papers in Honor of Paul W. Holland, ETS. Springer, 2010.

FR12   Stephen E. Fienberg and Alessandro Rinaldo. Maximum likelihood estimation in log-linear models. *Annals of Statistics*, 40(2):996–1023, 2012.

FW81a   Stephen E. Fienberg and Stanley S. Wasserman. Categorical data analysis of single sociometric relations. *Sociological methodology*, 12:156–192, 1981.

FW81b   Stephen E. Fienberg and Stanley S. Wasserman. Discussion of Holland, P. W. and Leinhardt, S. "An exponential family of probability distributions for directed graphs". *Journal of the American Statistical Association*, 76:54–57, 1981.

GP13   Elizabeth Gross and Sonja Petrović. Combinatorial degree bound for toric ideals of hypergraphs. *International Journal of Algebra and Computation*, 23(6):1503–1520, 2013.

GPS16 Elizabeth Gross, Sonja Petrović, and Despina Stasi. Goodness-of-fit for log-linear network models: Dynamic Markov bases using hypergraphs. *Annals of the Institute of Statistical Mathematics*, 2016. https://doi.org/10.1007/s10463-016-0560-2.

GZFA10 Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

Hab81 Shelby J. Haberman. An exponential family of probability distributions for directed graphs: Comment. *Journal of the American Statistical Association*, 76(373):60–61, March 1981.

HL81   Paul W. Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association*, 76(373):33–65, 1981.

HT08   Huy Tài Hà and Adam Van Tuyl. Monomial ideals, edge ideals of hypergraphs, and their graded Betti numbers. *Journal of Algebraic Combinatorics*, 27(215–245), 2008.

HT10   Hisayuki Hara and Akimichi Takemura. Connecting tables with zero-one entries by a subset of a Markov basis. In Marlos Viana and Henry Wynn, editors, *Algebraic Methods in Statistics and Probability II*, volume 516 of *Contemporary Mathematics*. American Mathematical Society, 2010.

JJ16   Pengsheng Ji and Jiashun Jin. Coauthorship and citation networks for statisticians. *Annals of Applied Statistics*, 10(4):1779–1812, 2016.

KP16   Vishesh Karwa and Sonja Petrović. Coauthorship and citation networks for statisticians: Comment. invited comment on the paper by Jin and Ji. *Annals of Applied Statistics*, 10(4):1827–1834, 2016.

OH00   Hidefumi Ohsugi and Takayuki Hibi. Compressed polytopes, initial ideals and complete multipartite graphs. *Illinois J of Mathematics*, 44(2):391–406, 2000.

Pet19 Sonja Petrović. What is ... a Markov basis? *Notices of the American Mathematical Society*, 66(7):1088—1092, 2019.

PRF10 Sonja Petrović, Alessandro Rinaldo, and Stephen E. Fienberg. Algebraic statistics for a directed random graph model with reciprocation. In M. Viana and H. Wynn, editors, *Algebraic Methods in Statistics and Probability II*, volume 516 of *Contemporary Mathematics*, pages 261–283. American Mathematical Society, Providence RI, 2010.

PRW01 Giovanni Pistone, Eva Riccomagno, and Henry Wynn. Computational commutative algebra in discrete statistics. *Contemporary Mathematics*, 287:267–282, 2001.

PS14   Sonja Petrović and Despina Stasi. Toric algebra of hypergraphs. *Journal of Algebraic Combinatorics*, 39(1):187–208, 2014.

PSY19 Sonja Petrović, Aleksandra Slavkovic, and Ruriko Yoshida. " Old wine in new bottles," and some more new wine – Stephen Fienberg's influence on algebraic statistics. *Journal of Algebraic Statistics*, 10(1, Issue in honor of Stephen E. Fienberg):i–vi, 2019.

PTV19 Sonja Petrović, Apostolos Thoma, and Marius Vladoiu. Hypergraph encodings of arbitrary toric ideals. *Journal of Combinatorial Theory, Series A*, 166(11–41), 2019.

RPF10 Alessandro Rinaldo, Sonja Petrović, and Stephen E. Fienberg. On the existence of the MLE for a directed random graph network model with reciprocation. Technical report, 2010. http://arxiv.org/abs/1010.0745.

RPF13 Alessandro Rinaldo, Sonja Petrović, and Stephen E. Fienberg. Maximum likelihood estimation in the Beta model. *Annals of Statistics*, 41(3):1085–1110, 2013.

Sam69 S. Sampson. *Crisis in a cloister*. Unpublished doctoral dissertation, Cornell University, 1969.

SSR+14 Despina Stasi, Kayvan Sadeghi, Alessandro Rinaldo, Sonja Petrović, and Stephen E. Fienberg. Beta models for random hypergraphs with a given degree sequence. In *Proceedings of 21st International Conference on Computational Statistics*, 2014.

Stu96   Bernd Sturmfels. *Grobner bases and convex polytopes*, volume 8. American Mathematical Soc., 1996.

SVV94   Aaron Simis, Wolmer V. Vasconselos, and Rafael H. Villarreal. On the ideal theory of graphs. *Journal of Algebra*, 167(20):3890416, 1994.

Vie15   Statistics Views. "In most places, not only are statisticians not in control of Big Data efforts and data science, but sometimes they are totally excluded or at best, marginalised." An interview with Stephen Fienberg. http://www.statisticsviews.com/details/feature/8555981/In-most-places-not-only-are-statisticians-not-in-control-of-Big-Data-efforts-and.html, 10 Nov 2015.

Vil95   Rafael H. Villarreal. Rees algebras of edge ideals. *Communications in Algebra*, 23(9):3513–3524, 1995.