ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Algebraic Statistics in Practice: Applications to Networks

Marta Casanellas,[1] Sonja Petrović,[2] and Caroline Uhler[3,4]

[1]Department of Mathematics and BGSMath, Universitat Politècnica de Catalunya, 14 08028 Barcelona, Spain; email: marta.casanellas@upc.edu

[2]Department of Applied Mathematics, Illinois Institute of Technology, Chicago, Illinois 60616, USA; email: sonja.petrovic@iit.edu

[3]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

[4]Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; email: cuhler@mit.edu

Annu. Rev. Stat. Appl. 2020. 7:227–50

First published as a Review in Advance on October 21, 2019

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-031017-100053

Keywords

algebraic statistics, network models, graphical models, causal structure discovery, phylogenetics

Abstract

Algebraic statistics uses tools from algebra (especially from multilinear algebra, commutative algebra, and computational algebra), geometry, and combinatorics to provide insight into knotty problems in mathematical statistics. In this review, we illustrate this on three problems related to networks: network models for relational data, causal structure discovery, and phylogenetics. For each problem, we give an overview of recent results in algebraic statistics, with emphasis on the statistical achievements made possible by these tools and their practical relevance for applications to other scientific disciplines.

1. INTRODUCTION

Algebraic statistics is a branch of mathematical statistics that focuses on the use of algebraic, geometric and combinatorial methods in statistics. The term "algebraic statistics" itself was coined as the title of a book on the use of techniques from commutative algebra in experimental design (Pistone et al. 2001). An early influential paper (Diaconis & Sturmfels 1998) connected the problem of sampling from conditional distributions for the analysis of categorical data to commutative algebra, thereby showcasing the power of the interplay between these areas. In the two decades that followed, increasing interest in applying new algebraic tools to key problems in statistics has generated a growing literature.

The use of algebra, geometry, and combinatorics in statistics did not start a mere two decades ago. Combinatorics and probability theory have gone hand in hand since their beginnings. The first standard mathematical method in statistics may be the method of least squares, which has been used extensively since shortly after 1800 and relies heavily on systems of linear equations. Nonlinear algebra has played a major role in statistics since the 1940s; readers are directed to, for example, Wilks (1946), Votaw (1948), James (1954), Andersson (1975), Bailey (1981), and Jensen (1988). In addition, the development of the theory of exponential families relied heavily on convex geometry (Barndorff-Nielsen 1978). However, Diaconis & Sturmfels (1998) and Pistone et al. (2001) introduced new algebraic disciplines in statistics, including modern computational algebraic geometry and commutative algebra. In this review, we concentrate on the developments in algebraic statistics in the past two decades and, in particular, on applications to networks.

The analysis of networks as relational data and to represent probabilistic interactions between variables is becoming increasingly popular, with applications in fields including the social sciences, genomics, neuroscience, economics, linguistics, and medicine. Theoretical and algorithmic developments for exploring such data sets are found at the intersection of statistics, applied mathematics, and machine learning. In this review we focus on some of the key statistical problems and their solutions using algebraic techniques in three application areas: network models (based on relational data encoded as observations on the edges of a random network), causal structure discovery (based on multivariate data encoded as observations on the nodes of an unknown underlying causal network), and phylogenetics (a particular network structure discovery problem where the underlying network is a tree with latent variables).

Section 2 focuses on statistical models for relational data, typical uses of which arise in the social and biological sciences. In these applications, nodes in the network may represent individuals, organizations, proteins, neurons, or brain regions, while links represent observed relationships between the nodes, such as personal or organizational affinities, social or financial relationships, binding between proteins, or physical links between brain regions. A key problem in this area is to test whether a proposed statistical model fits the data at hand; such a test typically involves generating a sufficiently large sample of networks from the model and comparing it to the observed network. Perhaps somewhat surprisingly, algorithms for sampling networks with given network statistics for goodness-of-fit testing are often efficiently encoded by algebraic constraints. In Section 2, we outline how techniques from commutative algebra and combinatorics are applied to this problem for several families of network models for which a formal test is otherwise unavailable.

In Section 3, we turn to applications where the network structure cannot directly be observed and we only have access to observations on the nodes of the network. Such applications range from modeling data on consumer behavior to click statistics for ads or websites, DNA sequences of related species, gene expression data, etc. The use of such data to gain insight into complex phenomena requires characterizing the relationships among the observed variables. Probabilistic graphical models explicitly capture the statistical relationships between the variables as a network.

A good representation of a complex system should enable prediction of not only the state of one component given the others, but also the effect that local operations have on the global system. This requires causal modeling and use of interventional data. In Section 3, we discuss the role that algebraic and discrete geometry play in analyzing prominent algorithms for causal structure discovery and in developing the first provably consistent algorithms for causal inference from a mix of observational and interventional data.

In Section 4, we discuss a particular directed network model, phylogenetic trees, for evolutionary reconstruction. Algebra and related areas have always been present in the study of evolutionary processes but have played minor roles relative to combinatorics or optimization. However, since the beginning of this century, the developments in algebraic statistics have given rise to techniques with a major impact on three different problems in phylogenetics: model selection, model identifiability, and phylogenetic reconstruction. Models that best fit the data should only be selected from among those whose parameters are identifiable, so understanding model identifiability is crucial. The final step, given an evolutionary model and data, is to reconstruct the phylogenetic tree and infer the evolutionary parameters. In Section 4, we explain how algebraic techniques can be used to address these problems and discuss their applications to complex evolutionary models and phylogenetic networks.

2. NETWORK MODELS FOR RELATIONAL DATA

Network models for relational data, that is, various types of interactions between a fixed set of entities, such as neurons, proteins, people, or corporations, have grown in popularity in recent decades. The interactions can be directed (e.g., affinity or one-way influence) or undirected (e.g., mutual affiliation) and may be counted with multiplicity or weight. Consider two recently collected data sets on statisticians who publish in five top-rated journals (Ji & Jin 2016). The data can be represented as a bipartite graph of authors and papers, in which a link exists between nodes i and j if author i wrote paper j; or as a citation network, in which a directed edge from i to j denotes that paper i cites paper j; or as collapsed coauthorship or citation networks among authors (**Figure 1**).

Taking a model-based approach, we study the effects of various types of author interactions on network analysis and inference by concentrating on goodness of fit of a network model. This is central for estimating network features, appropriately simulating data, and correctly interpreting the results. In addition, goodness-of-fit testing is considered to be very challenging in the network science community due to the size of the networks in many applications as well as their sparsity and particular structure. Methods rooted in algebraic statistics help answer such questions efficiently and reliably for a variety of network models.

Relational data can be modeled using random graphs, in which the interactions are modeled as random variables. This results in a statistical model with random directed or undirected edges on a fixed set of nodes. There is a rich literature on different random graph models, starting from the classical Erdös–Rényi graphs (Erdös & Rényi 1961), exponential random graph models (ERGMs) (Holland & Leinhardt 1981), and Markov graphs (Frank & Strauss 1986), to models that capture more intricate relational behavior, such as stochastic block models (SBMs) (Holland et al. 1983), latent space models (Hoff et al. 2002), and mixed membership SBMs (Airoldi et al. 2009; see also Goldenberg et al. 2010). The question of whether any of these models provides an adequate fit to data has received relatively little attention.

Here we consider the broad and flexible class of exponential family models for random graphs, ERGMs. To specify an ERGM, one first selects a vector of network characteristics $T(g) \subset \mathbb{R}^p$ that represent an interpretable and meaningful summary of the network, such as the number of neighbors of each node or block membership. The resulting model is the collection of probability

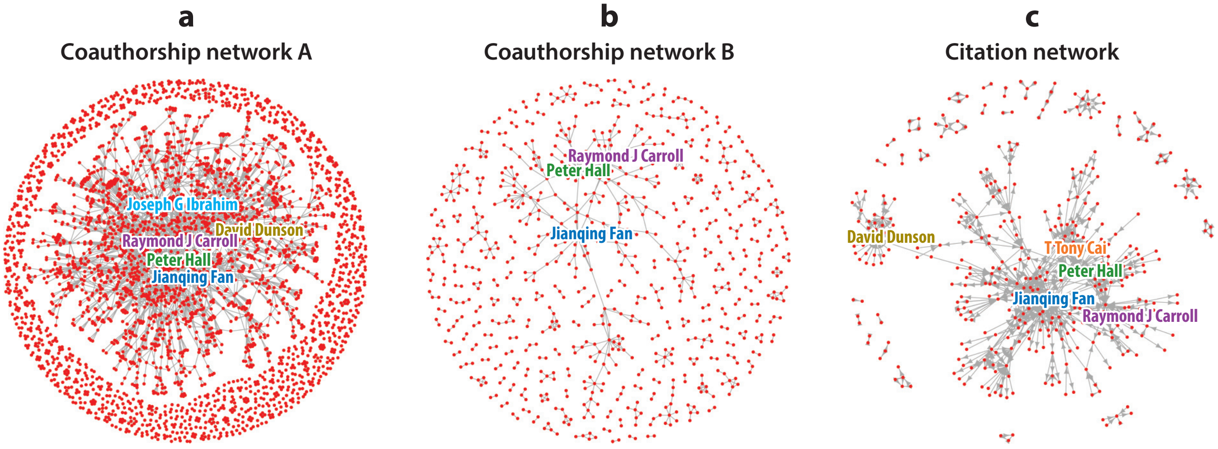

Figure 1

Author networks constructed from the data collected by Ji & Jin (2016). (*a*) In coauthorship network A, there is an undirected edge between nodes i and j if authors i and j coauthored at least two papers. (*b*) In coauthorship network B, there is an undirected edge between nodes i and j if authors i and j coauthored at least one paper. (*c*) In the citation network, there is a directed edge from author i to author j if i cited at least one paper by j.

measures $\mathcal{M} = \{p_\theta : \theta \in \Theta\}$ indexed by points in $\Theta \subset \mathbb{R}^p$ such that for any $\theta \in \Theta$, the probability of observing a given network $G = g$ takes the exponential form

$$p_\theta(g) = \exp\{\langle T(g), \theta\rangle - \psi(\theta)\},$$

where $\psi(\theta) = \log \sum_g \exp\{\langle T(g), \theta\rangle\}$ is the normalizing function (also known as the log-partition function) and $T(g)$ is the vector of minimal sufficient statistics for $\mathcal{M}$. Tools from commutative algebra can be applied to construct a finite-sample test for goodness of fit of such a model to the observed network, while graph-theoretic and combinatorial considerations can render the resulting algorithms scalable and applicable in practice for large networks.

2.1. Testing Model Fit: State of the Art

Studies devoted to goodness-of-fit tests for network models fall into two categories.

Heuristic tests are based on graphical comparisons between observed statistics and the corresponding statistics obtained from the fitted model (see Handcock 2003, Hunter et al. 2008, Carnegie et al. 2015). Given an observed graph g_{obs}, the goal is to evaluate how well a model $\mathbb{P}_\theta(G)$ fits g_{obs}. Let $s(g)$ be a vector of network statistics. The most popular include the number of edges, triangles, or two-stars in g; the vector of counts of neighbors of every node (the degree sequence); or other summaries of a node's connectedness or centrality in g. The graphical method proceeds by computing a maximum likelihood estimator (MLE) $\hat{\theta}$ of θ and simulating several graphs $g^1, \ldots, g^B$ from $\mathbb{P}_{\hat{\theta}}$. Departures from the model are detected by comparing the sample distribution of $s(g^1), \ldots, s(g^B)$ with the observed value $s(g_{\text{obs}})$. Central to the graphical method is the choice of complementary statistics $s(g)$ used for evaluating the fit. Although widely used, graphical tests have two limitations: First, they are not based on any formal discrepancy measure between the model and observed network since the choice of $s(g)$ is arbitrary. Second, the distribution of the complementary statistics is unknown under the null hypothesis, so calibration and formal type I error rates are difficult to obtain.

Asymptotic tests are a natural alternative and rely on formal testing criteria for evaluating model fit. However, classical test criteria such as the log-likelihood ratio, Akaike information criterion (AIC), or Bayesian information criterion (BIC) cannot be directly applied to general network models, mainly because the usual asymptotics do not apply to models other than very simplistic ones. This is due to the fact that the independent and identically distributed (i.i.d.) assumption on the random edges does not hold, which undermines the use of results on asymptotic distributions of various test statistics. In addition, in many network models, the number of parameters increases with the number of nodes. This issue was first pointed out by Fienberg & Wasserman (1981b) and noted also in several later works (Holland & Leinhardt 1981, Hunter et al. 2008, Chatterjee et al. 2011, Carnegie et al. 2015, Krivitsky & Kolaczyk 2015, Yan et al. 2016). In addition, many commonly used ERGMs suffer from the lack of a natural notion of projectability (Shalizi & Rinaldo 2013), which relates the marginal distribution of a network on p nodes to the same model on $p+1$ nodes, essentially ruling out consistency of MLEs.

To remedy these issues, one can derive modified asymptotic distributions, when they exist, of various test statistics for special cases. For example, Yan et al. (2014) consider testing the degree-corrected block model with the usual SBM; Wang & Bickel (2017) derive an asymptotic Gaussian distribution of the likelihood ratio test statistic for selecting between two SBMs with different numbers of communities; Gao & Lafferty (2017) consider testing an Erdös–Rényi model against an SBM, construct a chi-square-like test statistic using a combination of edge, 2-star, and triangle counts, and show that its limiting distribution is a chi-square distribution; Lei (2016) constructs a goodness-of-fit test for the SBM by using the extreme eigenvalues of a certain residual matrix as a test statistic and deriving its asymptotic distribution; and similarly, Banerjee & Ma (2017) derive a central limit theorem for linear spectral statistics for testing an Erdös–Rényi model against a two-block block model. A common limitation of these studies is that the asymptotic distributions are derived in specialized asymptotic regimes that may not hold in practice and are difficult to verify, given a single network. For instance, Lei (2016) shows that the asymptotic null distribution of the test statistic requires that the entries of the estimated edge probabilities be uniformly bounded away from 0 and 1, which rules out certain types of sparse network.

2.2. From Networks to Contingency Tables: Log-Linear Models

Network data on p nodes can be naturally summarized by a contingency table of format $p \times p \times i_1 \times \cdots \times i_k$, classifying the type of a relationship (directed, undirected, block-dependent, etc.) that holds for each dyad in the graph. This representation means that certain ERGMs can be represented by equivalent models for contingency tables that have a long history in the statistics literature. The models amenable to such a representation are called log-linear ERGMs, and their vector of sufficient statistics is a linear function of the network. For such models there exists a matrix A such that $T(g) = Ag$, where the network g has been flattened to vector format. Log-linear ERGMs encompass many of the popular models in use today, including all undirected and directed degree-based models [e.g., the β model (Chatterjee et al. 2011, Rinaldo et al. 2013)], SBMs with or without mixed membership but with known block assignment (Holland & Leinhardt 1981, Fienberg et al. 1985, Airoldi et al. 2009), combinations of these such as the degree-corrected SBM (Karrer & Newman 2011), and extensions of any of these models using covariates (Yan et al. 2018).

The connection to contingency tables dates back to Fienberg & Wasserman (1981a), Fienberg et al. (1985), and Fienberg & Wasserman (1981b), who consider some models for relational data that were very novel at the time and remain very popular. By viewing the network representation of the data as a union of independent dyads that can appear in various configurations, they express in table format a set of models that are now considered canonical under the ERGM framework. The first advantage of this viewpoint, also pointed out in these early works, is that the MLE can efficiently and accurately be computed using iterative proportional fitting, thus avoiding the usual convergence issues that are the main drawback of Markov chain Monte Carlo (MCMC) approaches typically used for ERGMs (see, e.g., Hunter et al. 2008). The second advantage became apparent in the 2000s with the development of tools from algebraic statistics for contingency tables: A generating set of a polynomial ideal can be translated to a set of networks that preserve an ERGM's sufficient statistics, then used as input to a sampling algorithm that provides a reference set for testing model fit. Coupled with a valid discrepancy measure for model fit also ported from the contingency table literature into networks, this approach solves the issues outlined in Section 2.1.

2.3. Goodness-of-Fit Testing for Log-Linear Exponential Random Graph Models

Let $\mathcal{M}_T$ be a log-linear ERGM, where T denotes the vector of sufficient statistics. A canonical way to test model fit is to compute the exact p-value conditional on the sufficient statistics for the null hypothesis that $p_{\hat{\theta}}(g)$ lies in the model $\mathcal{M}_T$, where $\hat{\theta}$ is the MLE, against the general alternative [see Fienberg & Wasserman (1981b) for further motivation]. The p-value is computed by comparing the observed network g against all other networks whose sufficient statistics are the same; this set,

$$\mathcal{F}_T(g) := \{g' : T(g') = T(g)\},$$

is called the fiber of g under the model $\mathcal{M}_T$. In virtually all instances of interest for applications, the fiber is too large to enumerate, so one resorts to sampling from it.

To sample from this conditional distribution for any log-linear model, Diaconis & Sturmfels (1998) introduce a notion of a basis that can be used as input to the Metropolis–Hastings algorithm. In the context of networks, a Markov basis of the log-linear ERGM $\mathcal{M}_T$ is any set of networks $\mathcal{B} = \{b_1, \ldots, b_n\}$ for which $T(b_i) = 0$ and such that for any given network g and any $h \in \mathcal{F}_T(g)$, there exist $b_{i_1}, \ldots, b_{i_N} \in \mathcal{B}$ that can be used to reach h from g, i.e.,

$$g + b_{i_1} + \cdots + b_{i_N} = h,$$

while walking through elements of the fiber, meaning that each partial sum $u + \sum_{j=0}^{N} b_{i_j}$, for any $j = 1, \ldots, N$, represents a valid network (see **Figure 2**). Note that $T(u) = T(u + b_i)$ means that adding a move b_i to any network does not change the values of the sufficient statistics, so to remain in the fiber, we need only ensure that adding a move did not produce negative entries in the vector, as the count of edges in a graph cannot be negative. The resulting Markov chain is irreducible, symmetric, and aperiodic; Drton et al. (2009, algorithm 1.13) outline a vanilla implementation.

The connection to commutative algebra translates each move b_i into a binomial: a difference of products of indeterminates, each corresponding to a cell in the contingency table. In the example from **Figure 2**, the depicted move can be written as $e_{17}e_{24}e_{34}e_{58} - e_{14}e_{23}e_{45}e_{78}$, where e_{ij} is the indeterminate representing the dyad $\{i, j\}$. The move is thus a polynomial in the random dyads. This translation is straightforward but leads to a fundamental and surprising result: A set of moves is a Markov basis if and only if the corresponding binomials generate the toric ideal defined by T (Diaconis & Sturmfels 1998). Consequently, each log-linear model does have a finite Markov basis by the Hilbert basis theorem, and all the basis elements can be obtained using tools for computing bases of toric ideals.

Markov bases are a popular theoretical construct in algebraic statistics, but in practice they pose serious challenges in large networks, and in general for large sparse contingency tables. One is that

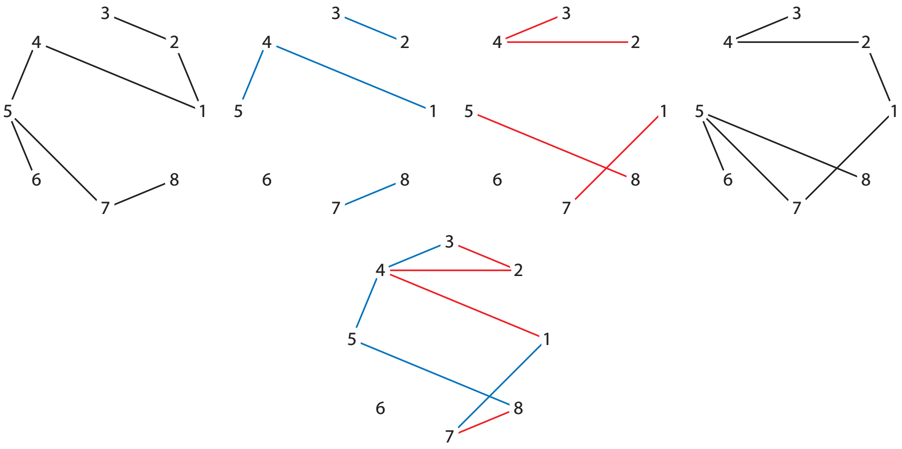

Figure 2

Example of a move for the exponential random graph models with $T(g) = (d_1, \ldots, d_8) = (2, 2, 1, 2, 3, 1, 2, 1)$, where d_i is the number of neighbors of node i. (*top*) From left, a starting graph g (*black*), set of edges to be removed from g (*blue*), set of edges to be added to g (*red*), and the resulting graph h (*black*). (*bottom*) The move b represented as a bicolored graph, with blue edges carrying weight -1 and red $+1$. Since blue edges contribute negative neighbors, $T(b) = 0$ and thus $T(h) = T(g)$.

they are complicated to compute a priori and that algebra produces many moves inapplicable to the observed data. To circumvent this difficulty, one may construct only those moves that can be applied to the observed network, doing so on the fly (without precomputing the entire basis) while ensuring that the fiber is connected. Gross et al. (2017) and Karwa et al. (2016) implement such a dynamic algorithm for generating Markov moves for the β and p_1 models, some of the basic variants of the SBMs, and combinations of these. Another difficulty concerns the mixing time of the Markov chain constructed using Markov bases, as any chain that is slow to mix will not be scalable to large networks in practice; to this end, we will only mention that there is a large body of literature in discrete mathematics that implies rapid mixing of this chain for almost all fibers; details are provided by Dillon (2016).

> **Example.** Considering the largest connected component of the citation network of authors, Ji & Jin (2016) count the neighbors in this directed graph and propose using them for author rankings. We perform an exact test of model fit for the ERGM whose sufficient statistic is this vector of neighbor counts, namely, the p_1 model with dyad-specific reciprocation. As a measure of discrepancy between the observed graph and the MLE, we use the chi-square statistic. The p-value reported is the proportion of the sampled networks in the fiber whose chi-square value is at least as large as that of g_{obs}. The test is done by running the Markov chain described by Gross et al. (2017). After $N = 100{,}000$ steps, the estimated p-value is 0.0072.

This result indicates that the p_1 model does not fit the citation network of authors, so the network may possess transitive effects and the dyads may not be independent. Similarly, we perform an exact test of fit of the β model in the largest connected component of the coauthorship network A. The p-value from the goodness-of-fit test obtained by running the Markov chain for $N = 100{,}000$ steps is 0.997. This suggests that node degrees are a surprisingly good summary of the graph and that the degree of an author could be used to determine an author ranking. However, this graph was obtained by thresholding the original data (a popular technique in

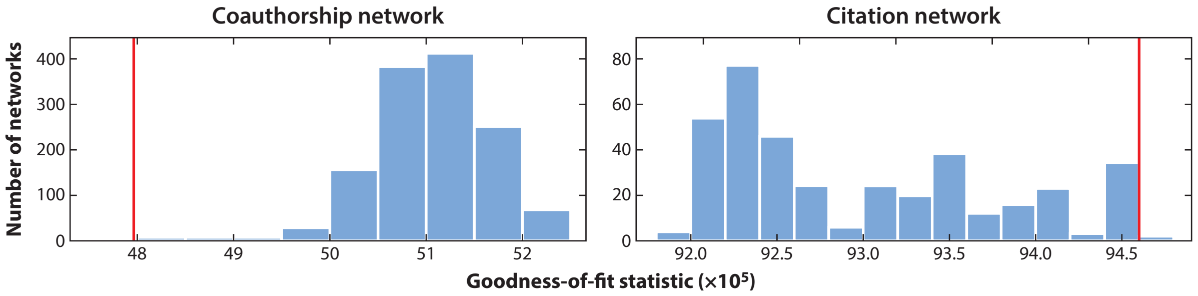

Figure 3

Testing model fit for the exponential random graph models with node degrees and sufficient statistics. (*left*) Histogram of the chi-square statistics, as a measure of discrepancy, for 100,000 networks in the fiber of the coauthorship network from **Figure 1**. (*right*) Same information for the citation network. The vertical red line indicates the observed value of the chi-square statistic.

network analysis used to avoid multiple edges), as well as by reducing multi-author papers to pairs of authors. These two tests are summarized in **Figure 3**, which depicts the sampling distributions of the chi-square statistic. We used the usual MCMC diagnostics in R to check that the Markov chains converged fairly well.

2.4. Generalizations to Weighted Graphs

The previous example opens up several interesting questions: How can we preserve the underlying data structure and still use an interesting network model with scalable estimation and goodness-of-fit methods? Karwa & Petrović (2016) argue that thresholding and reducing to a graph are not necessary; one can instead work with a hypergraph representation of the data, which preserves more of the coauthorship structure than the network representation. For I authors, J research areas, and K journals, consider an $I \times I \times J \times K$ contingency table whose (i, i', j, k) entry counts the number of times author i cites author i' in research area j and journal k. A similar representation can be obtained for the coauthorship network, where we count the number of times authors i and j wrote a joint paper. These representations preserve the citation and coauthorship count data. We can then collapse the table to an $I \times I$ author-by-author table and fit log-linear models to the citation counts. In essence, we seek to avoid thresholding, as in the generalized β model discussed by Rinaldo et al. (2013) for weighted networks represented in table form. Generalizing to weighted or multiple graphs is straightforward in the contingency table setting, with MLE algorithms unaffected and Markov basis algorithms becoming—perhaps surprisingly—more efficient and easier to implement. This opens up several lines of research on generalizing these models and enriches the network science literature with goodness-of-fit tests for many popular ERGMs.

By definition, log-linearity means that the sufficient statistics are a linear function of the graph, which in turn implies dyadic independence. The assumption of dyadic independence may seem restrictive, but Yan et al. (2018) show that it includes many popular models and avoids the degeneracy that plagues other ERGMs. In addition, Karwa et al. (2016) develop goodness-of-fit testing methods combining the Bayesian and algebraic approaches for mixture models of log-linear ERGMs, which do not assume dyadic independence.

3. CAUSAL STRUCTURE DISCOVERY

From random graph models, where each edge of the network is associated with a random variable, we now turn to graphical models, where each node of the network is associated with a random variable. In many applications, the underlying network must be learned from data on the nodes. We here consider the problem of learning the causal relationships among the nodes.

Causal inference is the basis of scientific discovery, because it asks "why?" The gold standard for inferring a causal relationship is a randomized controlled trial. However, in many applications, running such trials to test for a causal effect is impractical, unethical, or prohibitively expensive, so there have been large efforts to develop a theory of causal inference based purely on observational data. This began with two crucial advances made independently in the 1920s. Jerzy Neyman established a formal distinction between random variables under randomization and ordinary random variables via the potential outcome notation (Neyman 1923). Sewall Wright independently pioneered the use of graphs to represent cause-effect relationships using structural equation models (Wright 1921, 1934). However, skepticism among statisticians resulted in the causal interpretation of structural equation models being overlooked and almost forgotten [see Pearl (2012) for a historical account]. The reemergence of causal inference from observational data in statistics began in the 1970s and led to major contributions by Pearl (2000), Robins (1999), Rubin (1974, 2005), and Spirtes et al. (2001).

While it has in general been unethical, too expensive or even impossible to perform large-scale interventional studies, the development of genome editing technologies in biological studies (Cong et al. 2013) and the explosion of interventional data in online advertisement and education represent a unique opportunity for the development of new causal inference methodologies. It is now possible to obtain large-scale interventional data sets relatively easily. This calls for a theoretical and algorithmic framework for learning causal networks from a mix of observational and interventional data.

In this section, we show how methods from algebraic geometry, combinatorics, graph theory, and discrete geometry have been brought to bear in the analysis and development of causal structure discovery algorithms. In Section 3.1, we introduce the framework of structural equation models for causal modeling and then discuss open problems in combinatorics and graph theory related to the degree of identifiability of causal effects. In Section 3.2, we review a prominent causal structure discovery algorithm. It relies on the so-called faithfulness assumption, and using algebraic geometry, we show that this assumption is very restrictive and hard to satisfy in practice. In Section 3.3, we will discuss an alternative algorithm that makes critical use of discrete geometry to overcome the limitations of the faithfulness assumption and leads to the first provably consistent algorithm for causal inference from a mix of observational and interventional data. Finally, in Section 3.4, we discuss various open problems and related literature in algebraic statistics.

3.1. Structural Equation Models and Markov Equivalence

We represent a causal network by a directed graph $G = (V, E)$ consisting of vertices $V = \{1, \dots, p\}$ and directed edges E representing direct causal relationships. We make the common assumption that G is a directed acyclic graph (DAG), meaning there are no directed cycles $i_0 \to i_1 \to \cdots \to i_m \to i_0$, since causal effects only act forward in time. In a structural equation model (Wright 1921, 1934), each node $i \in V$ is associated with a random variable X_i and is a deterministic function of its parents, denoted by $\mathrm{pa}(i)$, and independent noise, denoted by ϵ_i. For example, a structural equation model on the 4-node DAG $1 \to 2$, $2 \to 3$, $3 \to 4$, $1 \to 4$ is given by

$$X_1 = f_1(\epsilon_1), \quad X_2 = f_2(X_1, \epsilon_2), \quad X_3 = f_3(X_2, \epsilon_3), \quad X_4 = f_4(X_1, X_3, \epsilon_4). \qquad 1.$$

Gaussian linear structural equation models are special instances of this model class, where $X_j = \sum_{i \in \mathrm{pa}(j)} a_{ij} X_i + \epsilon_j$ and the noise $\epsilon = (\epsilon_1, \dots, \epsilon_p)$ follows a Gaussian distribution $\mathcal{N}(0, D)$ with

Table 1 The number of MECs, along with the ratios of the numbers of MECs to DAGs and the ratios of the counts of MECs of size 1 (MEC_1) to the total number of MECs up to 10 nodes

p (nodes)	1	2	3	4	5	6	7	8	9	10
# MEC	1	2	11	185	8,782	1,067,825	312,510,571	212,133,402,500	326,266,056,291,213	1,118,902,054,495,975,141
(# MEC)/(# DAG)	1.00000	0.66667	0.44000	0.34070	0.29992	0.28238	0.27443	0.27068	0.26888	0.26799
(# MEC_1)/(# MEC)	1.00000	0.50000	0.36364	0.31892	0.29788	0.28667	0.28068	0.27754	0.27590	0.27507

Data are from Gillispie & Perlman (2001). Abbreviations: DAG, directed acyclic graph; MEC, Markov equivalence class.

diagonal covariance matrix D. In this case, the joint distribution of $X = (X_1, \ldots, X_p)$ is a Gaussian $\mathcal{N}(0, \Sigma)$, where $\Sigma^{-1} = (I - A)D^{-1}(I - A)^T$ and A is the weighted adjacency matrix of G containing the causal effects a_{ij}. While this model is of interest for its mathematical simplicity, in many applications including genomics the linear and Gaussian assumptions are often violated and it is preferable to work with the general nonparametric model in Equation 1.

A structural equation model encodes not only the observational distribution, i.e., the distribution of X, but also the interventional distributions. For instance, in Equation 1, an intervention on node X_3 that sets its value to 0 would change the distribution of the nodes X_3 and X_4, but not the others, since they are not downstream of X_2. Such an intervention could, for example, be used to model a gene knockout experiment, where the expression of certain genes is set to zero (Cong et al. 2013).

A structural equation model provides a factorization of the joint distribution, which implies certain conditional independence (CI) relations through the Markov property, namely $X_i \perp\!\!\!\perp X_{\mathrm{nd}(i)} \mid X_{\mathrm{pa}(i)}$, where $\mathrm{nd}(i)$ denotes the nondescendents of node i (see, e.g., Lauritzen 1996 for an introduction to graphical models). A standard approach for causal structure discovery is to infer CI relations from the sample distribution and then infer the DAG from these relations. However, in general a DAG is not identifiable, since multiple DAGs can encode the same set of CI relations; such DAGs are called Markov equivalent. Verma & Pearl (1990) provided a graphical characterization of when two DAGs are Markov equivalent, namely, when they have the same skeleton (i.e., undirected edges) and immoralities (i.e., induced subgraphs of the form $i \to j \leftarrow k$).

Since from observational data it is only possible to identify a DAG up to its Markov equivalence class (MEC), it is important to study the sizes of MECs and their distributions. However, while a recurrence relation for the number of DAGs on p nodes is known (Robinson 1973), no such formula is known for MECs. Gillispie & Perlman (2001) enumerated all MECs up to 10 nodes (see **Table 1**). The first row shows that the number of MECs grows very quickly in the number of nodes p. The second row shows the ratio of the number of MECs to the number of DAGs, suggesting that this sequence converges to $\approx 1/4$. This combinatorial conjecture would have important consequences for causal inference, since it would imply that on average an MEC consists of about 4 DAGs, meaning that in general only very few interventional experiments would be required to identify the true causal DAG. Finally, the last row suggests that the ratio of the number of MECs of size 1 to the total number of MECs also converges to $\approx 1/4$. This would imply that $\approx 1/4$ of all causal DAGs can be uniquely identified without any interventional data. While a combinatorial analysis of the number of MECs for particular families of DAGs was initiated by Radhakrishnan et al. (2017, 2018), these problems in general are wide open.

3.2. Causal Structure Discovery Algorithms and Faithfulness

Since the overwhelming majority of available data have been observational, most causal inference algorithms have been developed in this setting. A standard approach to causal structure discovery is constraint based, i.e., to treat causal inference as a constraint satisfaction problem, with the constraints being the CI relations inferred from the data. A prominent example is the PC algorithm (named for its authors, Peter and Clark), which starts in the complete undirected graph and iteratively removes edges (i, j) if there exists $S \subset V \setminus \{i, j\}$ such that $X_i \perp\!\!\!\perp X_j \mid X_S$. This results in the skeleton of the DAG; the immoralities are determined in a second step using the identified CI relations.

For such an algorithm to output the correct MEC, it is necessary that the inferred CI relations are faithful to the true DAG. In particular, it has to hold that

$$X_i \not\!\perp\!\!\!\perp X_j \mid X_S \quad \text{for all } (i, j) \in E \text{ and all } S \subset V \setminus \{i, j\}, \qquad 2.$$

which is known as the adjacency faithfulness assumption (Ramsey et al. 2006). Faithfulness violations can occur through cancellation of causal effects in the graph. Assumption 2 seems harmless at first, since it is highly unlikely that causal effects in a DAG cancel each other out exactly. However, CI relations are inferred from data via hypothesis testing, so Assumption 2 must be strengthened in the finite sample regime. In the Gaussian setting, where CI relations can be tested using partial correlations $\rho_{ij|S}$, Assumption 2 leads to the definition of strong faithfulness (Zhang & Spirtes 2003):

$$\rho_{ij|S} \geq \lambda \quad \text{for all } (i, j) \in E \text{ and all } S \subset V \setminus \{i, j\},$$

where $\lambda \asymp \sqrt{\log(p)/n}$ to guarantee uniform consistency of the PC algorithm (Kalisch & Bühlmann 2007).

Since the strong faithfulness assumption is critical for the consistency of various prominent causal inference algorithms, it is important to understand how many samples are needed in general to satisfy it. Algebraic geometry has played a major role in answering this question (Uhler et al. 2013, Lin et al. 2014). To see why, consider a Gaussian linear structural equation model on the fully connected DAG on three nodes with edges $1 \to 2, 1 \to 3, 2 \to 3$. For simplicity, we assume that all error variances are equal to unity and hence $(X_1, X_2, X_3) \sim \mathcal{N}(0, \Sigma)$, where $\Sigma^{-1} = (I - A)(I - A)^T$ and A is strictly upper triangular containing the causal effects a_{12}, a_{13} and a_{23}. Since no edge is missing, any CI relation is unfaithful to the DAG. On three nodes, there are six possible CI relations. For Gaussian distributions, any CI relation corresponds to the vanishing of an almost principal minor, as shown in **Figure 4**. Hence, faithfulness violations correspond to a collection of real algebraic hypersurfaces and understanding how restrictive the strong faithfulness assumption is requires the computation of the volume of tubes around these hypersurfaces. This was achieved using tools from real algebraic geometry, namely, Crofton's formula and the Lojasiewicz inequality in Uhler et al. (2013) and real log-canonical thresholds in Lin et al. (2014). These results were then used to compute the scaling of number of samples to number of variables that lead to the tubes filling up the whole space, meaning that no faithful distribution exists. In the high-dimensional setting, this scaling was shown to be as bad as $p_n = O(\log n)$, a real limitation for the application of algorithms that rely on faithfulness, including the PC algorithm. These results also provide an example of how methods from algebraic geometry can be applied in high-dimensional statistics.

- $X_1 \perp\!\!\!\perp X_2 \quad \Longleftrightarrow \quad \det((\Sigma^{-1})_{13,23}) = a_{12} = 0$
- $X_1 \perp\!\!\!\perp X_3 \quad \Longleftrightarrow \quad \det((\Sigma^{-1})_{12,23}) = a_{13} + a_{12}a_{23} = 0$
- $X_1 \perp\!\!\!\perp X_3 \quad \Longleftrightarrow \quad \det((\Sigma^{-1})_{12,13}) = a_{12}^2 a_{23} + a_{12}a_{13} + a_{23} = 0$
- $X_1 \perp\!\!\!\perp X_2 \mid X_3 \quad \Longleftrightarrow \quad \det((\Sigma^{-1})_{1,2}) = a_{13}a_{23} - a_{12} = 0$
- $X_1 \perp\!\!\!\perp X_3 \mid X_2 \quad \Longleftrightarrow \quad \det((\Sigma^{-1})_{1,3}) - a_{13} = 0$
- $X_2 \perp\!\!\!\perp X_3 \mid X_1 \quad \Longleftrightarrow \quad \det((\Sigma^{-1})_{2,3}) - a_{23} = 0$

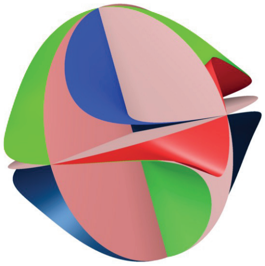

Figure 4

The unfaithful distributions for a three-node, fully connected DAG correspond to a collection of six hypersurfaces, each of which is defined by the vanishing of an almost principal minor; the three linear hypersurfaces are shown in pink, and the three nonlinear hypersurfaces are in red, blue, and green. Adapted with permission from Uhler et al. (2013). Abbreviation: DAG, directed acyclic graph.

3.3. Directed Acyclic Graph Associahedra for Causal Inference from Interventional Data

With this understanding of unfaithful distributions as a collection of hypersurfaces, it is clear that obtaining algorithms with better consistency guarantees requires removing some hypersurfaces, i.e., testing fewer CI relations. Given an ordering of the nodes π that is consistent with the true DAG G (i.e., if $i \to j$ in G, then $i < j$ in the ordering π), then by the Markov property, G can be recovered by testing only one CI relation per edge, namely, using the conditioning set consisting of all ancestors of i and j with respect to π, i.e.,

$$X_i \perp\!\!\!\perp X_j \mid X_S, \quad \text{where } S = \{k \in V : k \leq i \text{ or } k \leq j \text{ with respect to } \pi\} \setminus \{i, j\}. \qquad 3.$$

The true ordering, however, must be inferred from data. A natural approach following Occam's razor is to associate to each permutation π a DAG G_π using the conditional dependence relations in 3. and then return the sparsest permutation, i.e., the sparsest DAG among all permutations. This approach is uniformly consistent under strictly weaker conditions than strong faithfulness, namely, provided the sparsest DAG is in the true MEC (Raskutti & Uhler 2018). However, these improved consistency guarantees were achieved at a high computational price, since determining the sparsest permutation requires searching over all $p!$ permutations.

This raises the question of whether replacing the exhaustive permutation search by a greedy search could be used for causal inference. Greedy search algorithms are commonly applied for causal inference, most notably greedy equivalence search, a greedy search over the space of MECs (Chickering 2002). The convex hull of all permutations of length p gives rise to a $(p-1)$-dimensional polytope, known as the permutohedron, whose vertices are the permutations. Two permutations are connected by an edge in the permutohedron if and only if they differ by a neighboring transposition. The three-dimensional permutohedron of all permutations of length 4 is shown in **Figure 5**. Solus et al. (2017) show that a greedy search in the permutohedron is consistent, i.e., it outputs the correct MEC when the sample size goes to infinity, under strictly weaker conditions than faithfulness. The sequences in **Table 1** suggest that the number of MECs grows much faster than the number of permutations. Hence, it is remarkable that a greedy search on the space of permutations has similar consistency guarantees as greedy search on the space of MECs, despite a large reduction in the search space.

In fact, the search space can be reduced further by identifying permutations whose DAGs G_π and $G_{\pi'}$ are the same, since the number of edges in such graphs is necessarily the same. Such permutations are connected by edges in the permutohedron. Contracting these edges gives rise to a polytope (Mohammadi et al. 2018) known as the DAG associahedron, which can also be obtained

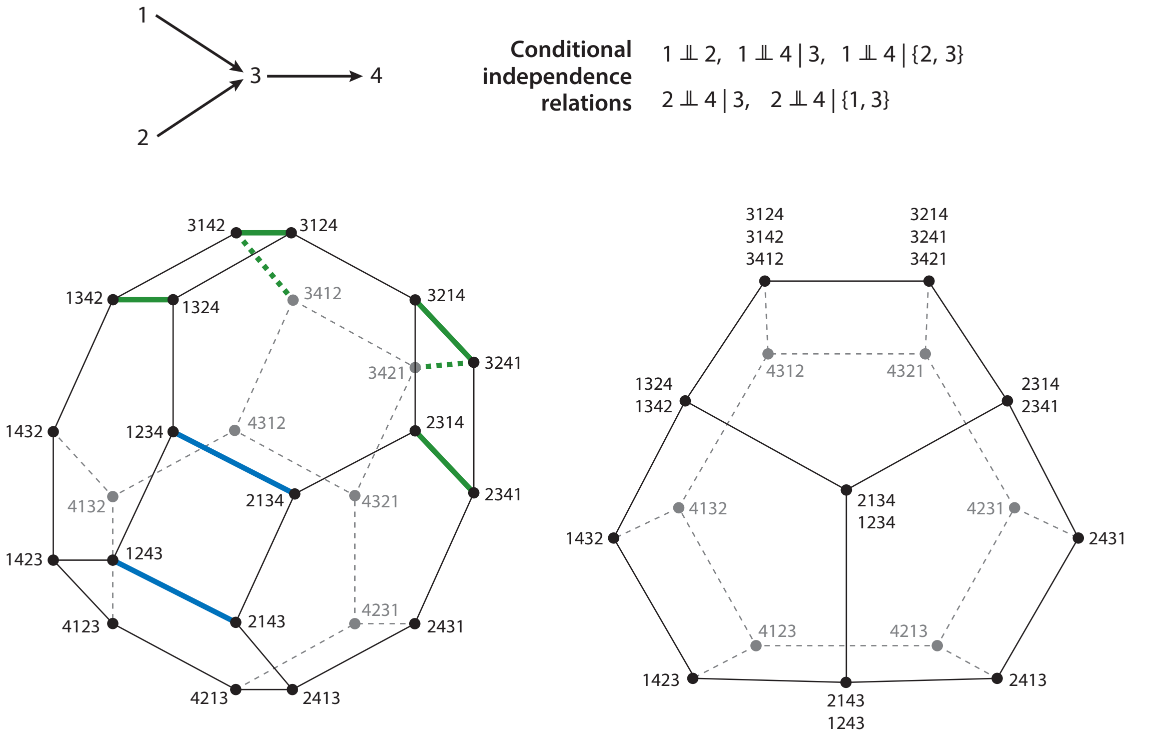

Figure 5

Three-dimensional permutohedron consisting of all permutations of length 4 and the DAG associahedron for a particular 4-node DAG model; blue and green edges correspond to the conditional independence relations in this model. Adapted from Mohammadi et al. (2018). Abbreviation: DAG, directed acyclic graph.

by a different construction, namely, by associating to each edge in the permutohedron a CI relation as described by Morton et al. (2009) and contracting all edges corresponding to CI relations in the underlying DAG (Mohammadi et al. 2018) (see **Figure 5**). Thus, DAG associahedra are a generalization of the prominent (undirected) graph associahedra (Carr & Devadoss 2006) that are obtained from the permutohedron by contracting all edges corresponding to separations or CI statements in an undirected graphical model. Hence, the quest for a causal inference algorithm that is consistent under strictly weaker conditions than faithfulness and as a consequence achieves higher accuracy than previous algorithms in the high-dimensional setting led to the development of DAG associahedra and new results in convex geometry that are of independent interest.

Recent years have witnessed a paradigm shift in the kinds of data collected. In genomics, but also various other application areas, large-scale interventional data sets are being produced by deliberately altering components such as genes. By further reducing the search directions, the greedy sparsest permutation search algorithm described above was extended to the first provably consistent algorithm for causal inference from a mix of observational and interventional data (Wang et al. 2017, Yang et al. 2018). For applications of these algorithms to learning gene regulatory networks, readers are directed to, e.g., Wang et al. (2017, 2018) and Yang et al. (2018).

3.4. Open Problems and Related Literature

While faithfulness is well-understood from a geometric perspective, it is an open problem in algebraic geometry/combinatorics to understand the assumptions needed for consistency of the sparsest permutation algorithm. This is of great interest, since it is conjectured that these are the weakest assumptions that guarantee consistency of any algorithm for learning the true MEC (Raskutti & Uhler 2018). Other polyhedral approaches for causal inference have been described (Cussens et al. 2017, Jaakkola et al. 2010), and it would be interesting to better understand how they relate to each other. So far, we have only considered causal inference when all variables are observed. However, for applications in the social sciences, latent variables are ubiquitous. The FCI (fast causal inference) algorithm and its variants generalize the PC algorithm to the latent setting (Spirtes et al. 2001). It is an open problem to generalize greedy permutation search algorithms to the setting with latent variables. In addition, while CI relations are the only constraints that act on structural equation models in the fully observed setting, in the latent setting, there are additional constraints such as the Verma constraints (Richardson et al. 2017). While a full algebraic description of these constraints is not known, for linear Gaussian structural equation models, a large subset has recently been characterized as nested determinants (Drton et al. 2018). In addition to these equality constraints, there are inequality constraints. Describing these is very challenging, as demonstrated by the ongoing search for the semialgebraic description of the set of matrices of fixed nonnegative rank (Allman et al. 2015, Kubjas et al. 2015), which correspond to simple latent tree models in the discrete setting. Explicit knowledge of the defining equations and inequalities is crucial to answer questions of identifiability (e.g., Allman et al. 2009) or model selection (e.g., Drton et al. 2017, Evans 2018). Finally, we return to the beginnings of algebraic statistics on experimental design (Pistone et al. 2001) to point out a critical problem in the era of interventional data: deciding which interventions to perform in order to gain the most information about the underlying causal system (e.g., Agrawal et al. 2019).

4. PHYLOGENETICS

This section treats a particular class of directed graphical models with latent variables, called phylogenetic trees. Algebraic tools have been used since the end of the twentieth century to address problems in phylogenetics (Felsenstein 1978, Hendy & Penny 1989, Evans & Speed 1993, Hendy et al. 1994). In particular, Lake (1987) and Cavender & Felsenstein (1987) opened the door to the development of phylogenetic reconstruction methods based on the polynomial equations implied by a particular evolutionary model and tree structure. Below, we survey this approach and then describe the major impact that algebraic statistics has had on phylogenetic reconstruction, model selection, and identifiability.

More detailed introductions to algebraic phylogenetics can be found in publications by Allman & Rhodes (2007), Pachter & Sturmfels (2005), Sullivant (2018), Zwiernik (2015), and Steel (2016, sections 7 and 8).

4.1. Phylogenetic Reconstruction

A phylogenetic tree is a tree graph whose leaves correspond to molecular sequences (for example, the whole genome of a species or a single gene) of different living species and represents the speciation process that led to them: the interior nodes represent ancestral sequences, the edges represent evolutionary processes, and the leaves are labeled with the names of the living species. The topology of a phylogenetic tree is the topology of the labeled graph; for example, **Figure 6** shows the three different (unrooted) tree topologies for the species $\{1, 2, 3, 4\}$, which are denoted as $12|34$, $13|24$, and $14|23$. While phylogenetic trees can be reconstructed using a variety of data including DNA or protein molecules, we here assume that the available data are a sequence of characters A, C, G, and T (corresponding to the four nucleotides) of length N for each living species in the tree.

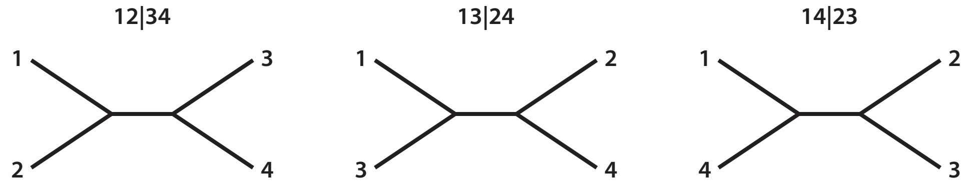

Figure 6

The three different unrooted phylogenetic tree topologies on leaves {1, 2, 3, 4} are denoted as 12|34 (*left*), 13|24 (*middle*), and 14|23 (*right*).

In order to model the evolution of nucleotide data, it is convenient to assume that the substitution of nucleotides occurs randomly and follows a Markov process on the phylogenetic tree T, where the internal nodes are latent variables. The state space of the random variables at the nodes of T is $\{A, C, G, T\}$ and the parameters of the model are a distribution π at a fixed interior node (which plays the role of the root) and the entries of the 4 × 4 transition matrices M^e associated to the edges e of T. According to this hidden Markov process on T, where two nodes are independent given their least common ancestor, the probability of observing a character pattern at the leaves of the tree can be expressed as a polynomial in the model parameters. Assuming that the characters in each sequence have all evolved following the same evolutionary process and are independent of each other, the data are N independent samples from a multinomial distribution.

Different restrictions imposed on the transition matrices give different evolutionary models, from the simplest Jukes & Cantor (1969) model (JC69), where π is uniform and there is only one free parameter per edge (that is, on each edge, all conditional probabilities $P(x|y)$ are equal if $x \neq y$), to the general Markov model (GMM) with no restrictions on the transition matrices or on π.

As an example, we consider the JC69 model on the tree 12|34 shown in **Figure 6** relating the set of species $\{1, 2, 3, 4\}$. Denoting by $p_{x_1x_2x_3x_4}$ the joint probability of observing nucleotide x_i at species i, the fact that the JC69 model is invariant under permutations of the four states implies that

$$\begin{aligned} p_{\mathtt{AAAA}} &= p_{\mathtt{CCCC}} = p_{\mathtt{GGGG}} = p_{\mathtt{TTTT}}, \\ p_{\mathtt{AAAC}} &= p_{\mathtt{AAAG}} = p_{\mathtt{AAAT}} = \cdots = p_{\mathtt{TTTG}}, \end{aligned} \qquad 4.$$

for any parameters of the model. These are the first natural algebraic equations that arise from this type of evolutionary model. Although they are linear equations, they can be useful in model selection (see Section 4.3). They cannot be used to estimate the tree topology because they hold for any joint distribution arising from a JC69 model on any of the trees in **Figure 6**; they are therefore called model invariants.

Consider now

$$p_{\mathrm{ACAC}} + p_{\mathrm{ACGT}} = p_{\mathrm{ACGC}} + p_{\mathrm{ACAT}}, \qquad 5.$$

$$p_{\mathrm{ACCA}} + p_{\mathrm{ACTG}} = p_{\mathrm{ACCG}} + p_{\mathrm{ACTA}}. \qquad 6.$$

Both Equations 5 and 6 hold for any set of JC69 parameters on the tree 12|34, but Equation 5 does not hold for all distributions on the tree 13|24 and Equation 6 does not hold for 14|23 (Lake 1987). Such equations, which are satisfied for all joint distributions on a particular phylogenetic tree but not for all distributions on another tree, are called topology invariants (Steel 2016, section 8.3).

These equations were used by Lake (1987) to design a statistical test based on the χ^2-statistic to infer the tree topology. These first attempts were not very successful (Huelsenbeck 1995), were only valid for simple models, and only used two of the relevant algebraic equations (Casanellas & Fernández-Sánchez 2010). The use of topology invariants for phylogenetic reconstruction was thus halted until Allman & Rhodes (2008). We explain their main contribution in what follows.

Let $p \in \mathbb{R}^{256}$ be a distribution of character patterns on species $\{1, 2, 3, 4\}$ as above, and consider its flattening matrix $\mathit{flatt}_{12|34}(p)$ according to the split $12|34$, namely:

$$\begin{array}{cc} & \text{states at leaves } 3,4 \\ \begin{array}{r} \text{states} \\ \text{at} \\ \mathit{flatt}_{12|34}(p) = \text{leaves} \\ 1,2 \end{array} & \begin{pmatrix} p_{\mathtt{AAAA}} & p_{\mathtt{AAAC}} & p_{\mathtt{AAAG}} & \cdots & p_{\mathtt{AATT}} \\ p_{\mathtt{ACAA}} & p_{\mathtt{ACAC}} & p_{\mathtt{ACAG}} & \cdots & p_{\mathtt{ACTT}} \\ p_{\mathtt{AGAA}} & p_{\mathtt{AGAC}} & p_{\mathtt{AGAG}} & \cdots & p_{\mathtt{AGTT}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{\mathtt{TTAA}} & p_{\mathtt{TTAC}} & p_{\mathtt{TTAG}} & \cdots & p_{\mathtt{TTTT}} \end{pmatrix}. \end{array}$$

Allman & Rhodes (2008) prove that if p is a distribution from a Markov process on the tree $12|34$ (in the GMM), then $\mathit{flatt}_{12|34}(p)$ has rank ≤ 4; moreover, it has rank 16 if p is a joint distribution on any of the two other trees $13|24$, $14|23$ (arising from a Markov process with generic parameters).

This result has allowed the development of new topology invariants under the GMM but also, most importantly, the use of techniques such as rank approximation to propose methods to select the tree that best represents the data. This approach has been exploited in work that has attracted the attention of biologists (Chifman & Kubatko 2014, Fernández-Sánchez & Casanellas 2016) and has allowed a generalization to the multispecies coalescent model and the use of these methods to estimate both gene trees and species trees (Chifman & Kubatko 2015). Some of these methods have been implemented in PAUP* (Swofford 2003), one of the most widely used software packages in phylogenetics, which has opened the use of these tools to the biological community at large and has allowed the application in areas such as biodiversity preservation (Devitt et al. 2019).

When dealing with real data, one only has access to a finite number of samples from the corresponding multinomial distribution. Since many methods are based on asymptotic tests, they may not be suitable for small samples. Current research attempts to solve this issue using statistical tests based on algebraic tools are starting to be developed for the finite-sample regime (Gaither & Kubatko 2016, Sumner et al. 2017).

These first approaches to algebraic phylogenetics have been restricted to trees on four species but can be used in quartet-based methods to infer large phylogenetic trees using only quartet data as input (that is, topologies of four species, in addition to an assessment score of the reliability of each quartet topology) (Strimmer & von Haeseler 1996, Ranwez & Gascuel 2001, Snir & Rao 2010, Davidson et al. 2018). This approach has been used by Fernández-Sánchez & Casanellas (2016) to provide new support for the phylogenetic tree of eight species of yeast that was suggested by biological evidence and only obtained by certain reconstruction methods restricted to certain models. It would be of great interest to develop algebraic methods that can directly infer large trees; a first result in this direction is Sumner (2017).

The evolutionary models used in these algebraic approaches are more general than those commonly used by biologists. Indeed, the usual approach in phylogenetics is to use a continuous-time Markov process. In this case, the transition matrix M corresponding to an edge is of type $M = e^{tQ}$, where Q is an instantaneous mutation rate matrix that operates for the duration $t \geq 0$. Not all transition matrices are of this type (i.e., not all Markov matrices are embeddable in a continuous-time process); indeed, the logarithm of a transition matrix may not be real, and if it is real, it may not be a rate matrix. Roca-Lacostena & Fernández-Sánchez (2018) proved that, for the Kimura three-parameter model of nucleotide substitution, the set of embeddable matrices represents only 9.4% of all transition matrices. Moreover, it is commonly assumed that Q is the same for all edges of the tree (i.e., the Markov process is homogeneous in time) and that the process is stationary and time reversible. This leads to one of the most common models in phylogenetics, the so-called general time-reversible (GTR) model. While restricting to this model is quite controversial (Sumner et al. 2012) and it might be too restrictive, as we just insinuated, using a GTR model might be convenient because it considers fewer parameters than the GMM, so parameter estimation is more feasible. Algebraic approaches to phylogenetics avoid parameter inference altogether and make phylogenetic inference feasible for the most general Markov model, the GMM.

So far we have mainly focused on the recovery of the tree. As the number of trees grows superexponentially in the number of leaves, accurately recovering the tree topology is a basic first step toward parameter recovery using methods such as maximum likelihood. Nevertheless, algebraic statistics can also lead to important results in obtaining estimates for continuous parameters of small phylogenetic trees. For example, using computational algebra, one can compute the number of critical points of the log-likelihood (Catanese et al. 2006) and then use tools from numerical algebraic geometry to obtain the global optimum (Kosta & Kubjas 2019). Moreover, tools from computational algebra have provided major insights into the existence of a unique global optimum and provided analytical expressions to obtain it (Chor et al. 2006, Dinh et al. 2017). In addition to equality constraints given by the model, the continuous parameters must also satisfy biological constraints and stochastic conditions, which are encoded as inequality constraints. While understanding these semialgebraic constraints is difficult, Zwiernik & Smith (2011), Matsen (2009), Allman et al. (2014), and Steel & Faller (2009) discuss which semialgebraic constraints suffice to describe the model together with the algebraic constraints.

These algebraic tools are also being used in phylogenetic networks; for instance, Chifman & Kubatko (2019) introduce a new technique based on algebraic statistics to detect events in hybridization networks. These new tools for topology reconstruction are opening a new direction for phylogenetic reconstruction, with many interesting challenges from both statistical and algebraic points of view.

4.2. Identifiability

Although the use of algebraic statistics for proving identifiability of parameters of statistical models in phylogenetics is primarily of theoretical interest, it has been critical for proving the consistency of many phylogenetic reconstruction methods, including those based on likelihood. In the following, we provide a short overview.

Chang (1996) used algebraic tools to prove that the GMM is generically identifiable, that is, generic parameters are identifiable from the joint distributions of triplets of species (up to label swapping). The same holds for simpler models. Unfortunately, identifiability becomes much more involved for more complex models. Of particular interest are extensions that allow different sites to evolve at different rates, either by considering a gamma distribution of rates across sites (but assuming that all sites evolve according to the same tree topology) or by considering a mixture model (i.e., the joint distribution p is a mixture of a certain number of distributions p^i that have arisen from trees T_i under a certain model $\mathcal{M}$, with unknown trees, continuous parameters, and mixing parameters).

One well-known model that allows for different rates is the GTR+Γ, where all sites evolve based on the same tree topology and with the same instantaneous mutation rate matrix, but the rate at which each site evolves follows a gamma distribution (of certain fixed parameters). Although maximum likelihood estimation for this model has been widely used by the biological community, identifiability of its parameters was established using algebraic statistics only by Allman et al. (2008).

As far as mixture models are concerned, the first problem is to prove identifiability of the tree parameters. Without any constraints on the number of distributions, overfitting occurs and the continuous parameters are not identifiable. However, with appropriate constraints, the trees can be identified. Consider, for instance, mixtures on a single tree of four leaves evolving under the JC69 model. Equations 5 and 6 are satisfied for all distributions on the tree 12|34, and as they are linear, they are also satisfied for any mixture of distributions on this tree. Therefore, as mixtures of distributions on either of the other two trees in **Figure 6** do not satisfy one of these equations, these topological invariants are able to identify the tree for this type of mixture. In a similar way, the rank conditions mentioned in Section 4.1 allow a generalization that proves identifiability of trees in the case of mixtures on a single tree (Rhodes & Sullivant 2012). When mixtures on two different trees are considered, few positive results have been obtained (Allman et al. 2011). For example, it is unknown whether, if one considers mixtures of GMM distributions on two trees T_1 and T_2, the pair $\{T_1, T_2\}$ can be recovered from the mixed distribution.

Recently, algebraic tools have been used to prove the consistency of phylogenetic reconstruction methods for more complex models, including methods that reconstruct the species tree from gene trees according to the multispecies coalescent model. This is a very active area of research with important biological implications. For instance, Allman et al. (2018, 2019) make use of deep algebraic tools.

Finally, in recent years there have also been incursions of algebraic statistics into questions about identifiability of phylogenetic and hybridization networks (Gross & Long 2018, Baños 2019, Chifman & Kubatko 2019), but many open problems remain.

4.3. Model Selection

Algebraic statistics has also been applied to phylogenetics to select the evolutionary model that best fits the data. As mentioned above, a range of evolutionary models has been described, from JC69 to GMM [see, for example, the Felsenstein hierarchy of Pachter & Sturmfels (2005)]. Among them, the ones that have been deeply studied from an algebraic viewpoint are JC69, K80 (Kimura 1980), K81 (Kimura 1981), SSM (Casanellas & Sullivant 2005), and GMM. Following Kedzierska et al. (2012), we explain here how the model invariants for these models can be used in model selection within the framework of phylogenetic mixtures.

Model invariants on trees of n leaves for an evolutionary model $\mathcal{M}$ are algebraic equations satisfied by all distributions arising from any set of parameters on the model $\mathcal{M}$ on any phylogenetic tree on n leaves. Equations 4 are an instance of model invariants for $\mathcal{M} = \text{JC69}$ on trees of four leaves. As they are linear equations, they are also satisfied for any mixture of a collection of distributions on these trees. In general, the space of mixtures of distributions on trees on n leaves evolving under $\mathcal{M}$, i.e., the set of mixtures of any number of distributions on trees on n leaves evolving under $\mathcal{M}$, is determined by the collection of linear model invariants. Moreover, the linear model invariants for $\mathcal{M} = \text{JC69}$, K80, K81, and SSM are generated by binomial equations for any n, analogous to Equations 4 and computed by Casanellas et al. (2012), which leads to the exact computation of the likelihood maximum for data points coming from mixtures of distributions on a particular $\mathcal{M}$. Finally, these likelihoods can be combined into an information criterion for model selection such as corrected AIC or BIC. These tools were applied by Kedzierska et al. (2012) to DNA data from the PANDIT (Protein and Associated Nucleotide Domains with Inferred Trees) database: While the usual model selection method chooses the most complex model GRT+Γ (+ invariable sites) and gives a tree incompatible with the accepted phylogeny, the method presented there selects a mixture of JC69 and leads to the accepted phylogenetic tree.

5. DISCUSSION

Since its beginning in the late 1990s, the field of algebraic statistics has grown rapidly. The development of new theory and algorithms for data analysis inspired by algebra, combinatorics, and algebraic geometry has brought together previously disconnected communities of algebraists and statisticians. By now, algebraic methods have touched on all major themes in statistics, including parameter identifiability and estimation, hypothesis testing, model selection, and Bayesian inference. Conversely, problems and models from statistics have inspired significant new pure developments in algebraic combinatorics, high-dimensional commutative algebra, and computational algebraic geometry. Various textbooks have been written on algebraic statistics, including those of Pachter & Sturmfels (2005), Drton et al. (2009), Sullivant (2018), and Aoki et al. (2012), and for readers interested in using algebraic tools for statistical analysis, there is a package `algstat` implemented in R (Kahle 2014).

We have provided an overview of developments made possible through the use of algebraic methods in three areas related to networks. However we did not touch upon many interesting developments of algebraic statistics. For example, significant contributions related to Markov bases have been applied to disclosure limitation (Fienberg & Slavkovic 2004) and genetics (Malaspinas & Uhler 2011). Another recent direction is the use of commutative algebra for experimental design in system reliability (Sáenz-de Cabezón & Wynn 2015). Finally, another domain where algebraic techniques have been very fruitful is for the analysis of chemical reaction networks [see, e.g., Müller et al. (2016) and the work cited therein]. It has been an exciting two decades for algebraic statistics, which has had a major impact on statistical theory and its applications, and we expect that it will expand into many further application domains.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

M.C. was partially supported by AGAUR Project 2017 SGR-932 and MINECO/FEDER Projects MTM-2015-69135 and MDM-2014-0445. C.U. was partially supported by NSF (DMS-1651995), ONR (N00014-17-1-2147 and N00014-18-1-2765), IBM, and a Sloan Fellowship. S.P. was partially supported by NSF (DMS-1522662) and IIT CISC (Center for Interdisciplinary Scientific Computation). We thank Steffen Lauritzen, Bernd Sturmfels, Piotr Zwiernik, Orlando Marigliano, Eliana Duarte, Yassine El Maazouz, and an anonymous reviewer for helpful discussions and/or feedback on initial versions of this paper.

LITERATURE CITED

Agrawal R, Squires C, Yang KD, Shanmugam K, Uhler C. 2019. ABCD-Strategy: budgeted experimental design for targeted causal structure discovery. *Proc. Mach. Learn. Res.* 89:3400–9

Airoldi EM, Blei DM, Fienberg SE, Xing EP. 2009. Mixed membership stochastic blockmodels. In *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, ed. D Koller, D. Schuurmans, Y Bengio, L Bottou, pp. 33–40. **https://papers.nips.cc/paper/3578-mixed-membership-stochastic-blockmodels**

Allman ES, Ane C, Rhodes JA. 2008. Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. *Adv. Appl. Probab.* 40:229–49

Allman ES, Degnan JH, Rhodes JA. 2018. Species tree inference from gene splits by unrooted star methods. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15:337–42

Allman ES, Long C, Rhodes JA. 2019. Species tree inference from genomic sequences using the log-det distance. *SIAM J. Appl. Algebr. Geom.* 3:107–27

Allman ES, Matias C, Rhodes J. 2009. Identifiability of parameters in latent structure models with many observed variables. *Ann. Stat.* 37:3099–132

Allman ES, Petrović S, Rhodes JA, Sullivant S. 2011. Identifiability of two-tree mixtures for group-based models. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 8:710–22

Allman ES, Rhodes JA. 2007. Phylogenetic invariants. In *Reconstructing Evolution*, ed. O Gascuel, MA Steel, pp. 108–41. Oxford, UK: Oxford Univ. Press

Allman ES, Rhodes JA. 2008. Phylogenetic ideals and varieties for the general Markov model. *Adv. Appl. Math.* 40:127–48

Allman ES, Rhodes JA, Sturmfels B, Zwiernik P. 2015. Tensors of nonnegative rank two. *Linear Algebr. Appl.* 473:37–53

Allman ES, Rhodes JA, Taylor A. 2014. A semialgebraic description of the general Markov model on phylogenetic trees. *SIAM J. Discrete Math.* 28:736–55

Andersson SA. 1975. Invariant normal models. *Ann. Stat.* 3:132–54

Aoki S, Hara H, Takemura A. 2012. *Markov Bases in Algebraic Statistics*. New York: Springer

Bailey RA. 1981. A unified approach to design of experiments. *J. R. Stat. Soc. A* 144:214–23

Banerjee D, Ma Z. 2017. Optimal hypothesis testing for stochastic block models with growing degrees. arXiv:1705.05305 [math.ST]

Baños H. 2019. Identifying species network features from gene tree quartets under the coalescent model. *Bull. Math. Biol.* 81:494–534

Barndorff-Nielsen OE. 1978. *Information and Exponential Families in Statistical Theory*. New York: Wiley

Carnegie NB, Krivitsky PN, Hunter DR, Goodreau SM. 2015. An approximation method for improving dynamic network model fitting. *J. Comput. Graph. Stat.* 24:502–19

Carr MP, Devadoss SL. 2006. Coxeter complexes and graph-associahedra. *Topol. Appl.* 153:2155–216

Casanellas M, Fernández-Sánchez J. 2010. Relevant phylogenetic invariants of evolutionary models. *J. Math. Pures Appl.* 96:207–29

Casanellas M, Fernández-Sánchez J, Kedzierska A. 2012. The space of phylogenetic mixtures for equivariant models. *Algorithms Mol. Biol.* 7:33

Casanellas M, Sullivant S. 2005. The strand symmetric model. In *Algebraic Statistics for Computational Biology*, ed. L Pachter, B Sturmfels, pp. 305–21. Cambridge, UK: Cambridge Univ. Press

Catanese F, Hoşten S, Khetan A, Sturmfels B. 2006. The maximum likelihood degree. *Am. J. Math.* 128:671–97

Cavender JA, Felsenstein J. 1987. Invariants of phylogenies in a simple case with discrete states. *J. Classif.* 4:57–71

Chang JT. 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.* 137:51–73

Chatterjee S, Diaconis P, Sly A. 2011. Random graphs with a given degree sequence. *Ann. Appl. Probab.* 21:1400–35

Chickering DM. 2002. Optimal structure identification with greedy search. *J. Mach. Learn. Res.* 3:507–54

Chifman J, Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30:3317–24

Chifman J, Kubatko L. 2015. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. *J. Theor. Biol.* 374:35–47

Chifman J, Kubatko L. 2019. An invariants-based method for efficient identification of hybrid speciation from large-scale genomic data. *BMC Evol. Biol.* 19:112

Chor B, Hendy MD, Snir S. 2006. Maximum likelihood Jukes–Cantor triplets: analytic solutions. *Mol. Biol. Evol.* 23:626–32

Cong L, Ran F, Cox D, Lin S, Barretto R, et al. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science* 339:819–23

Cussens J, Haws D, Studený M. 2017. Polyhedral aspects of score equivalence in Bayesian network structure learning. *Math. Program. Ser. A* 164:285–324

Davidson R, Lawhorn M, Rusinko J, Weber N. 2018. Efficient quartet representations of trees and applications to supertree and summary methods. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15:1010–15

Devitt TJ, Wright AM, Cannatella DC, Hillis DM. 2019. Species delimitation in endangered groundwater salamanders: implications for aquifer management and biodiversity conservation. *PNAS* 116:2624–33

Diaconis P, Sturmfels B. 1998. Algebraic algorithms for sampling from conditional distributions. *Ann. Stat.* 26:363–97

Dillon M. 2016. *Runtime for performing exact tests on the p_1 statistical model for random graphs*. PhD Thesis, Illinois Inst. Technol., Chicago

Dinh V, Matsen IV, Frederick A. 2017. The shape of the one-dimensional phylogenetic likelihood function. *Ann. Appl. Probab.* 27:1646–77

Drton M, Lin S, Weihs L, Zwiernik P. 2017. Marginal likelihood and model selection for Gaussian latent tree and forest models. *Bernoulli* 23:1202–32

Drton M, Robeva E, Weihs L. 2018. Nested covariance determinants and restricted trek separation in Gaussian graphical models. arXiv:1807.07561 [math.ST]

Drton M, Sturmfels B, Sullivant S. 2009. *Lectures on Algebraic Statistics*. Basel: Birkhäuser

Erdös P, Rényi A. 1961. On the evolution of random graphs. *Bull. Inst. Int. Stat.* 38:343–47

Evans RJ. 2018. Model selection and local geometry. arXiv:1801.08364 [math.ST]

Evans SN, Speed TP. 1993. Invariants of some probability models used in phylogenetic inference. *Ann. Stat.* 21:355–77

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27:401–10

Fernández-Sánchez J, Casanellas M. 2016. Invariant versus classical quartet inference when evolution is heterogeneous across sites and lineages. *Syst. Biol.* 65:280–91

Fienberg SE, Meyer MM, Wasserman SS. 1985. Statistical analysis of multiple sociometric relations. *J. Am. Stat. Assoc.* 80:51–67

Fienberg SE, Slavkovic AB. 2004. Making the release of confidential data from multi-way tables count. *Chance* 17:5–10

Fienberg SE, Wasserman SS. 1981a. Categorical data analysis of single sociometric relations. *Sociol. Methodol.* 12:156–92

Fienberg SE, Wasserman SS. 1981b. Discussion of Holland, P. W. and Leinhardt, S. "An exponential family of probability distributions for directed graphs." *J. Am. Stat. Assoc.* 76:54–57

Frank O, Strauss D. 1986. Markov graphs. *J. Am. Stat. Assoc.* 81:832–42

Gaither J, Kubatko L. 2016. Hypothesis tests for phylogenetic quartets, with applications to coalescent-based species tree inference. *J. Theor. Biol.* 408:179–86

Gao C, Lafferty J. 2017. Testing network structure using relations between small subgraph probabilities. arXiv:1704.06742 [stat.ME]

Gillispie SB, Perlman MD. 2001. Enumerating Markov equivalence classes of acyclic digraph models. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, ed. J Breese, D Koller, pp. 171–77. San Francisco: Morgan Kaufmann

Goldenberg A, Zheng AX, Fienberg SE, Airoldi EM. 2010. A survey of statistical network models. *Found. Trends Mach. Learn.* 2:129–233

Gross E, Long C. 2018. Distinguishing phylogenetic networks. *SIAM J. Appl. Algebr. Geom.* 2:72–93

Gross E, Petrović S, Stasi D. 2017. Goodness-of-fit for log-linear network models: Dynamic Markov bases using hypergraphs. *Ann. Inst. Stat. Math.* 69:673–704

Handcock MS. 2003. *Assessing degeneracy in statistical models for social networks*. Work. Pap. 39, Cent. Stat. Soc. Sci., Univ. Wash., Seattle

Hendy MD, Penny D. 1989. A framework for the quantitative study of evolutionary trees. *Syst. Zool.* 38:297–309

Hendy MD, Penny D, Steel M. 1994. A discrete Fourier analysis for evolutionary trees. *PNAS* 91:3339–43

Hoff PD, Raftery AE, Handcock MS. 2002. Latent space approaches to social network analysis. *J. Am. Stat. Assoc.* 97:1090–98

Holland PW, Laskey KB, Leinhardt S. 1983. Stochastic blockmodels: first steps. *Soc. Netw.* 5:109–37

Holland PW, Leinhardt S. 1981. An exponential family of probability distributions for directed graphs (with discussion). *J. Am. Stat. Assoc.* 76:33–65

Huelsenbeck JP. 1995. Performance of phylogenetic methods in simulation. *Syst. Biol.* 44:17–48

Hunter DR, Goodreau SM, Handcock MS. 2008. Goodness of fit of social network models. *J. Am. Stat. Assoc.* 103:248–58

Hunter DR, Handcock MS, Butts CT, Goodreau SM, Morris M. 2008. ergm: a package to fit, simulate and diagnose exponential-family models for networks. *J. Stat. Softw.* 24:1–29

Jaakkola T, Sontag D, Globerson A, Meila M. 2010. Learning Bayesian network structure using LP relaxations. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, ed. YW Teh, M Titterington, pp. 358–65. **http://proceedings.mlr.press/v9/**

James AT. 1954. Normal multivariate analysis and the orthogonal group. *Ann. Math. Stat.* 25:40–75

Jensen ST. 1988. Covariance hypotheses which are linear in both the covariance and the inverse covariance. *Ann. Stat.* 116:302–22

Ji P, Jin J. 2016. Coauthorship and citation networks for statisticians. *Ann. Appl. Stat.* 10:1779–812

Jukes TH, Cantor C. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism*, Vol. III, ed. HN Munro, pp. 21–132. New York: Academic

Kahle D. 2014. `Algstat`. *R package for algebraic statistics*. **https://github.com/dkahle/algstat**

Kalisch M, Bühlmann P. 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* 8:613–36

Karrer B, Newman ME. 2011. Stochastic blockmodels and community structure in networks. *Phys. Rev. E* 83:016107

Karwa V, Pati D, Petrović S, Solus L, Alexeev N, et al. 2016. Exact tests for stochastic block models. arXiv:1612.06040 [stat.ME]

Karwa V, Petrović S. 2016. Coauthorship and citation networks for statisticians: comment. *Ann. Appl. Stat.* 10:1827–34

Kedzierska AM, Drton M, Guigo R, Casanellas M. 2012. SPIn: model selection for phylogenetic mixtures via linear invariants. *Mol. Biol. Evol.* 29:929–37

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16:111–20

Kimura M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *PNAS* 78:454–58

Kosta D, Kubjas K. 2019. Maximum likelihood estimation of symmetric group-based models via numerical algebraic geometry. *Bull. Math. Biol.* 81:337–60

Krivitsky PN, Kolaczyk ED. 2015. On the question of effective sample size in network modeling: an asymptotic inquiry. *Stat. Sci.* 30:198–98

Kubjas K, Robeva E, Sturmfels B. 2015. Fixed points of the EM algorithm and nonnegative rank boundaries. *Ann. Stat.* 43:422–61

Lake JA. 1987. A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. *Mol. Biol. Evol.* 4:167–91

Lauritzen SL. 1996. *Graphical Models*. Oxford, UK: Oxford Univ. Press

Lei J. 2016. A goodness-of-fit test for stochastic block models. *Ann. Stat.* 44:401–24

Lin S, Uhler C, Sturmfels B, Bühlmann P. 2014. Hypersurfaces and their singularities in partial correlation testing. *Found. Comput. Math.* 14:1079–116

Malaspinas AS, Uhler C. 2011. Detecting epistasis via Markov bases. *J. Algebr. Stat.* 2:36–53

Matsen FA. 2009. Fourier transform inequalities for phylogenetic trees. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 6:89–95

Mohammadi F, Uhler C, Wang C, Yu J. 2018. Generalized permutohedra from probabilistic graphical models. *SIAM J. Discrete Math.* 32:64–93

Morton J, Pachter L, Shiu A, Sturmfels B, Wienand O. 2009. Convex rank tests and semigraphoids. *SIAM J. Discrete Math.* 23:1117–34

Müller S, Feliu E, Regensburger G, Conradi C, Shiu A, Dickenstein A. 2016. Sign conditions for injectivity of generalized polynomial maps with applications to chemical reaction networks and real algebraic geometry. *Found. Comput. Math.* 16:69–97

Neyman J. 1923. Sur les applications de la théorie des probabilités aux experiences agricoles: essai des principes. *Roczniki Nauk Rolniczych* 10:1–51

Pachter L, Sturmfels B. 2005. *Algebraic Statistics for Computational Biology*. Cambridge, UK: Cambridge Univ. Press

Pearl J. 2000. *Causality: Models, Reasoning and Inference*. Cambridge, UK: Cambridge Univ. Press

Pearl J. 2012. The causal foundations of structural equation modeling. In *Handbook of Structural Equation Modeling*, ed. RH Hoyle, pp. 68–91. New York: Guilford

Pistone G, Riccomagno E, Wynn H. 2001. *Algebraic Statistics*. Boca Raton, FL: Chapman & Hall/CRC

Radhakrishnan A, Solus L, Uhler C. 2017. Counting Markov equivalence classes by number of immoralities. In *Proceedings of the 2017 Conference on Uncertainty in Artificial Intelligence*, art. 97. **http://auai.org/uai2017/proceedings/papers/97.pdf**

Radhakrishnan A, Solus L, Uhler C. 2018. Counting Markov equivalence classes for DAG models on trees. *Discrete Appl. Math.* 244:170–85

Ramsey J, Zhang J, Spirtes P. 2006. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, ed. R Dechter, T Richardson, pp. 401–8. Arlington, VA: AUAI Press

Ranwez V, Gascuel O. 2001. Quartet-based phylogenetic inference: improvements and limits. *Mol. Biol. Evol.* 18:1103–16

Raskutti G, Uhler C. 2018. Learning directed acyclic graphs based on sparsest permutations. *Statistics* 7:e183

Rhodes J, Sullivant S. 2012. Identifiability of large phylogenetic mixture models. *Bull. Math. Biol.* 74:212–31

Richardson TS, Evans RJ, Robins JM, Shpitser I. 2017. Nested Markov properties for acyclic directed mixed graphs. arXiv:1701.06686 [stat.ME]

Rinaldo A, Petrović S, Fienberg SE. 2013. Maximum likelihood estimation in the β-model. *Ann. Stat.* 41:1085–110

Robins JM. 1999. Association, causation, and marginal structural models. *Synthese* 121:151–79

Robinson RW. 1973. Counting labeled acyclic digraphs. In *New Directions in the Theory of Graphs*, ed. F Harary, pp. 239–73. New York: Academic

Roca-Lacostena J, Fernández-Sánchez J. 2018. Embeddability of Kimura 3ST Markov matrices. *J. Theor. Biol.* 445:128–35

Rubin DB. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66:688–701

Rubin DB. 2005. Causal inference using potential outcomes. *J. Am. Stat. Assoc.* 100:322–31

Sáenz-de Cabezón E, Wynn HP. 2015. Hilbert functions in design for reliability. *IEEE Trans. Reliab.* 64:83–93

Shalizi CR, Rinaldo A. 2013. Consistency under sampling of exponential random graph models. *Ann. Stat.* 41:508–35

Snir S, Rao S. 2010. Quartets maxcut: a divide and conquer quartets algorithm. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 7:704–18

Solus L, Wang Y, Matejovicova L, Uhler C. 2017. Consistency guarantees for permutation-based causal inference algorithms. arXiv:1702.03530 [math.ST]

Spirtes P, Glymour C, Scheines R. 2001. *Causation, Prediction and Search*. Cambridge, MA: MIT Press

Steel M. 2016. *Phylogeny: Discrete and Random Processes in Evolution*. Philadelphia: SIAM

Steel M, Faller B. 2009. Markovian log-supermodularity, and its applications in phylogenetics. *Appl. Math. Lett.* 22:1141–44

Strimmer K, von Haeseler A. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* 13:964–60

Sullivant S. 2018. *Algebraic Statistics*. Providence, RI: Am. Math. Soc.

Sumner JG. 2017. Dimensional reduction for the general Markov model on phylogenetic trees. *Bull. Math. Biol.* 79:619–34

Sumner JG, Jarvis PD, Fernández-Sánchez J, Kaine BT, Woodhams MD, Holland BR. 2012. Is the general time-reversible model bad for molecular phylogenetics? *Syst. Biol.* 61:1069–74

Sumner JG, Taylor AE, Holland BR, Jarvis PD. 2017. Developing a statistically powerful measure for quartet tree inference using phylogenetic identities and Markov invariants. *J. Math. Biol.* 75(6–7):1619–54

Swofford DL. 2003. *PAUP *: Phylogenetic analysis using parsimony (* and other methods), version 4.0b10*. Sunderland, MA: Sinauer Assoc.

Uhler C, Raskutti G, Yu B, Bühlmann P. 2013. Geometry of faithfulness assumption in causal inference. *Ann. Stat.* 41:436–63

Verma TS, Pearl J. 1990. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, ed. PP Bonissone, M Henrion, LN Kanal, JF Lemmer, pp. 220–27. Amsterdam: Elsevier

Votaw DF. 1948. Testing compound symmetry in a normal multivariate distribution. *Ann. Math. Stat.* 19:447–73

Wang Y, Solus L, Yang KD, Uhler C. 2017. Permutation-based causal inference algorithms with interventions. *Adv. Neural Inform. Proc. Syst.* 30:5824–33

Wang Y, Squires C, Belyaeva A, Uhler C. 2018. Direct estimation of differences in causal graphs. *Adv. Neural Inform. Proc. Syst.* 31:3774–85

Wang YR, Bickel PJ. 2017. Likelihood-based model selection for stochastic block models. *Ann. Stat.* 45:500–28

Wilks SS. 1946. Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *Ann. Math. Stat.* 17:257–81

Wright S. 1921. Correlation and causation. *J. Agric. Res.* 10:557–85

Wright S. 1934. The method of path coefficients. *Ann. Math. Stat.* 5:161–215

Yan T, Jiang B, Fienberg SE, Leng C. 2018. Statistical inference in a directed network model with covariates. *J. Am. Stat. Assoc.* 114:857–68

Yan T, Leng C, Zhu J. 2016. Asymptotics in directed exponential random graph models with an increasing bi-degree sequence. *Ann. Stat.* 44:31–57

Yan X, Shalizi C, Jensen JE, Krzakala F, Moore C, et al. 2014. Model selection for degree-corrected block models. *J. Stat. Mech. Theory Exp.* 2014:P05007

Yang KD, Katcoff A, Uhler C. 2018. Characterizing and learning equivalence classes of causal DAGs under interventions. *Proc. Mach. Learn. Res.* 80:5537–46

Zhang J, Spirtes P. 2003. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, ed. C Meek, U Kjærulff, pp. 632–39. San Francisco: Morgan Kaufmann

Zwiernik P. 2015. *Semialgebraic Statistics and Latent Tree Models*. Boca Raton, FL: Chapman & Hall/CRC

Zwiernik P, Smith JQ. 2011. Implicit inequality constraints in a binary tree model. *Electron. J. Stat.* 5:1276–312