Comment

Joe R. Hill

The authors of these two papers are among the most active nodes in an ever growing hypergraph of interesting papers on statistical applications of graph theory. It is an honor to discuss these two new hyperedges.

My discussion is divided into four parts. Section 1 discusses statistical applications of graph theory. Section 2 briefly describes ways of leveraging parallels between probability and database theory. Section 3 highlights two important points made in each of the papers. Finally, Section 4 asks some specific questions.

1. STATISTICAL APPLICATIONS OF GRAPH THEORY

Graph theory has a lot to offer statisticians. Consequently, graph theory is quickly becoming an integral part of modern statistics. Graphs, both directed and undirected, and hypergraphs can be used to (a) represent qualitative multivariate relationships, (b) specify and visualize multivariate statistical models, (c) determine statistical properties of multivariate models and (d) develop computationally efficient algorithms for dealing with large multivariate models. The first two of these contribute to effective communication between applications experts and statisticians. The third helps statisticians develop appropriate statistical theory. The fourth makes computing feasible for more complicated problems.

Graphical models provide a flexible paradigm for describing multivariate statistical models. They can have *discrete* variables (as in Bayesian networks, graphical and recursive loglinear models for contingency tables, and influence diagrams for applied decision analysis), or *continuous* variables (as in covariance selection and structural equation models). Conditional Gaussian models (Lauritzen and Wermuth, 1989; Wermuth and Lauritzen, 1990) provide a framework for having both kinds of variables in a single graphical model. Graphical models can have *directed* edges (as in Bayesian networks, influence diagrams and regression models) or *undirected edges* (as in graphical and decomposable loglinear models, covariance selection models and Markov random field models for image restoration). Chain graphs provide a framework for having both kinds of edges in a single graphical model.

In their paper, Cox and Wermuth (CW) introduce, for multivariate normal models, the concept of *dashed* edges as a way to represent constraints on covariance matrices (i.e., to represent marginal independencies), complementing the use of *full* edges to represent constraints on concentration matrices (i.e., to represent conditional independencies). They illustrate the use of the new enriched class of models with a number of empirical examples.

Spiegelhalter, Dawid, Lauritzen and Cowell (SDLC) give a status report on their ongoing development of Bayesian networks for expert systems. They have carefully combined a number of methods. They elicit Bayesian graphical models from medical experts. They use graphical ideas to convert the model into a computationally efficient form. They apply Bayesian estimation techniques to "learn" probability parameters as additional data are observed, and they use significance testing methods to monitor and critique the model.

SDLC provide an effective method for eliciting the qualitative, the probabilistic and the initial quantitative aspects of an expert-defined model. The key to their method is to use a directed acyclic graph to represent the qualitative relationships between variables. Nearly everything else follows from this graph.

This graph determines a recursive factorization of the joint distribution with, for each variable, a factor that is the conditional distribution of that variable, given its parents. This representation of the joint distribution has two advantages. First, the number of probabilities that the expert has to specify is considerably less than for a general joint distribution that does not encode the implied conditional independencies as efficiently. Second, these probabilities are "easy" for an expert to specify for three reasons: (a) the expert has to think about the distribution of only one variable at a time, (b) each distribution is conditioned on the parents of the variable, which are the variables that directly influence it and (c) the conditioning events can be thought of as fixed scenarios. In short, it is easy for an expert to think about the probability distribution of a single "effect" given its immediate "causes." This second advantage contrasts sharply with the problems associated with directly specifying an overall joint distribution. In that case, the expert would not be able to think conditionally but would have to think in multiple dimensions simultaneously and would typically have to specify many very small probabilities.

Once the model has been specified, it is converted to a junction tree representation for efficient computation. This conversion is carried out in a series of steps

Joe R. Hill is R & D Manager, EDS Research, 5951 Jefferson St. NE, Albuquerque, New Mexico 87109.

guided and justified by three important ideas: (a) graph separation in the moral graph of an ancestral set determines conditional independence, (b) the cliques of a chordal graph form an acyclic hypergraph and only acyclic hypergraphs have junction trees [later in the paper, their method for specifying hyper-Markov prior distributions depends on the fact, proved by Vorob'ev (1962), that a consistent set of marginal distributions has an extension iff the margins they are defined on form an acyclic hypergraph] and (c) large problems can be made computationally more tractible by decomposing them into smaller, component problems that require communication between neighboring components only.

The major point of this section has been to emphasize the important role that graph theory is playing in both of these papers. It has helped in communicating with substantive experts. It has helped in specifying and understanding multivariate statistical models. And it has helped with the computational aspects of those models. It is time that we started teaching graph theory in statistics courses of all levels.

2. LEVERAGING PARALLELS TO DATABASE THEORY

2.1 A Problem

Not everything is bliss in the world of graphical models. They have some rather subtle properties. They also lack some properties that seem at first to be trivially true. Some of the more important of these problems arise when probabilities can be zero. Although this situation does not arise in either of the papers, newcomers to graphical models might be misled into thinking that some statements made in SDLC and CW are valid in more general settings.

For example, CW state that "for a trivariate normal distribution of Y, Z, X the hypothesis $Y \perp\!\!\!\perp X \mid Z$ *and* $X \perp\!\!\!\perp Z \mid Y$ corresponds to zero concentrations for pairs (Y, X) and (X, Z) and it implies $X \perp\!\!\!\perp (Y, Z)$." Nothing could be simpler. The conditional independency $Y \perp\!\!\!\perp X \mid Z$ splits X and Y and the conditional independency $X \perp\!\!\!\perp Z \mid Y$ splits X and Z, so the two of them together split X and (Y, Z), hence they imply the conditional independency $X \perp\!\!\!\perp (Y, Z)$. For multivariate normal models, which CW are dealing with, this reasoning is fine; in fact, it is valid for any family of strictly positive probability distributions. However, if probabilities can be zero, then the result is not true! For example, the distribution $p(0, 0, 0) = p(1, 1, 1) = 1/2$, $p(x, y, z) = 0$ otherwise, satisfies the first two of these conditional independencies, but does not satisfy the third. See Moussouris (1974) and Dawid (1979b) for other examples.

The problem is that the Gibbs–Markov theorem requires strictly positive probability distributions. This positivity condition limits the possible applications of the equivalence of graph-generated conditional independence models and factorizations of joint distributions. In particular, the theorem cannot be applied to Bayesian networks with functional constraints (Lauritzen and Spiegelhalter, 1988) or to contingency tables with structural zeros or to statistical mechanics systems with forbidden states (Moussouris, 1974).

In his discussion of Besag's paper on Markov random fields in spatial statistics, Hammersley (1974) explained why he and Clifford did not publish the result when they first discovered it in 1971. He wrote (pp. 230–231),

> In proving this result, we assumed a *positivity condition*, namely that no probability should be zero. . . . In many of the most important practical applications to statistical mechanics, the physical system is subject to constraints which prevent the system from assuming certain *forbidden states*. . . . So it seemed to us not only aesthetically desirable but also practically important to amend our proof in order to make the theorem independent of the positivity condition . . . The very good reason for our failure [to do so] was the unexpected discovery by a graduate student, Mr John Moussouris, of a counter-example!

In short, Hammersley and Clifford did not publish the result because they thought the positivity condition limited the theorem too much for it to be useful in practice. Now no one doubts the importance of the theorem even with the positivity condition. But it is still quite inconvenient that no result exists for distributions with zero probabilities.

2.2 A Solution

Here is a solution that was suggested by parallels to relational database theory. Table 1 summarizes basic database/probability parallels; see Hill (1991) for more details. To state the results, we need some terminology from graph theory. A *hypergraph* is a set of nodes together with a set of hyperedges; each *hyperedge* is a subset of the nodes of the hypergraph. The *2-section* of a hypergraph is an undirected graph with the same set of nodes as the hypergraph and an edge between each pair of nodes that belong to a common hyperedge. A hypergraph is *conformal* if its set of hyperedges equals the set of cliques of the edge set of its 2-section. A hypergraph is *acyclic* if it is conformal and its 2-section is chordal. It can be shown that a hypergraph is acyclic iff it has the running intersection property iff it has a junction tree.

We also need some terminology adapted from database theory. Graph separation in an undirected graph determines a set of conditional independencies. A set

TABLE 1
Basic database and probability parallels

Probability concepts	Database concepts
Set of random variables V	Set of attributes (column names) R
Distribution for V, $p[V]$, a probability function	Relation (table) over R, $r[R]$, an indicator function for a set of tuples (rows)
Marginal distribution of $X \subseteq V$, $p[X]$	Projection of r onto $X \subseteq R$, $r[X]$
Conditional distribution $p[V \mid X = x]$	Selection $r[R \mid X = x]$
Factorization constraint $\otimes\{V_1, \ldots, V_k\}$, $V_j \subseteq V$	Join dependency $\bowtie\{R_1, \ldots, R_k\}$, $R_j \subseteq R$
Conditional independency $X \perp\!\!\!\perp Y \mid Z$, binary factorization constraint $\otimes\{X \cup Z, \ Y \cup Z\}$	Multivalued dependency $Z \twoheadrightarrow X \mid Y$, binary join dependency $\bowtie\{X \cup Z, Y \cup Z\}$

of conditional independencies is said to be *graph-generated* if there exists a graph that generates it. A conditional independency $X \perp\!\!\!\perp Y \mid Z$ *splits* variables in X from variables in Y; the variable set Z is called the *kernel* of this conditional independency. The *split graph* generated by a set of conditional independencies has an edge between every pair of variables that is not split by any of the conditional independencies in the set. The *closure* of a set of conditional independencies is the set of conditional independencies implied by the original set. Two sets of conditional independencies are said to *cover* each other if their closures are equal. A set of conditional independencies is said to be *conflict-free* if it is graph-generated and it does not split any of its kernels. Two sets of constraints are said to be *equivalent* if the sets of probability distributions that satisfy them are equal. Similar definitions have been given for databases.

The Gibbs–Markov theorem can be stated in the following three ways, each providing insight into the relationships between graphs, sets of conditional independencies and factorization constraints.

THEOREM 1+. *Let $\mathcal{G}$ be an undirected graph over V. The set of conditional independencies generated by $\mathcal{G}$ is equivalent, for strictly positive distributions, to the factorization constraint generated by the cliques of $\mathcal{G}$.*

THEOREM 2+. *Let $\mathcal{V}$ be a hypergraph over V. The set of conditional independencies implied by the factorization constraint generated by $\mathcal{V}$ is equivalent, for strictly positive distributions, to the factorization constraint generated by $\mathcal{V}$ if and only if $\mathcal{V}$ is conformal.*

THEOREM 3+. *Let $\mathbf{C}$ be a set of conditional independencies defined on V. $\mathbf{C}$ is equivalent, for strictly positive distributions, to the factorization constraint generated by the cliques of the split graph of $\mathbf{C}$.*

Fagin, Mendelzon and Ullman (1982) and Berri et al. (1983) proved the following database theorems, which, after accounting for the different terminology, look a lot like the three theorems stated above. In fact, however, because relations are indicator functions (therefore allowing zero values), these theorems, which have stronger requirements on the underlying graphical structure, suggest a way to relax the positivity condition.

THEOREM DB1. *Let $\mathcal{G}$ be an undirected graph over R. The set of multivalued dependencies generated by $\mathcal{G}$ is equivalent to the join dependency generated by the cliques of $\mathcal{G}$ if and only if $\mathcal{G}$ is chordal.*

THEOREM DB2. *Let $\mathcal{R}$ be a hypergraph over R. The set of multivalued dependencies implied by the join dependency generated by $\mathcal{R}$ is equivalent to the join dependency generated by $\mathcal{R}$ if and only if $\mathcal{R}$ is acyclic.*

THEOREM DB3. *Let $\mathbf{M}$ be a set of multivalued dependencies defined on R. $\mathbf{M}$ is equivalent to the join dependency generated by the cliques of the split graph of $\mathbf{M}$ if and only if $\mathbf{M}$ has a conflict-free cover.*

By translating database terms into probability terms (Table 1) in these three database theorems, we get the following three probability theorems, the proofs of which will be given elsewhere.

THEOREM 1*. *Let $\mathcal{G}$ be an undirected graph over V. The set of conditional independencies generated by $\mathcal{G}$ is equivalent to the factorization constraint generated by the cliques of $\mathcal{G}$ if and only if $\mathcal{G}$ is chordal.*

THEOREM 2*. *Let $\mathcal{V}$ be a hypergraph over V. The set of conditional independencies implied by the factorization constraint generated by $\mathcal{V}$ is equivalent to the factorization constraint generated by $\mathcal{V}$ if and only if $\mathcal{V}$ is acyclic.*

THEOREM 3*. *Let $\mathbf{C}$ be a set of conditional independencies defined on V. $\mathbf{C}$ is equivalent to the factorization constraint generated by the cliques of the split graph of $\mathbf{C}$ if and only if $\mathbf{C}$ has a conflict-free cover.*

Although Theorems 1*, 2* and 3* do not require strictly positive distributions, they do impose stricter constraints on the underlying graphical structures than do Theorems 1+, 2+ and 3+. Theorem 1* re-

quires the graph to be chordal for there to be equivalence, whereas Theorem 1+ puts no requirements on it. Theorem 2* requires the hypergraph to be acyclic for there to be equivalence, whereas Theorem 2+ requires only that it be conformal. Theorem 3* requires the set of conditional independencies to have a conflict-free cover for there to be equivalence, whereas Theorem 3+ puts no requirements on it (actually, the closure with respect to strictly positive distributions of a set of conditional independencies is always graph-generated).

As far as I know, Theorems 1*, 2* and 3* are new, although, by now, they are probably not unexpected.

Parallel developments in the two fields have occurred in the past, with neither aware of the other, apparently. For example, Vorob'ev's (1962) results on extending consistent marginal distributions parallel similar results for the extension of consistent databases (Beeri et al., 1983). And Beeri and Kifer's (1986a, 1986b, 1987) work on fixing sets of multivalued dependencies that have intersection anomalies parallels Dawid's (1979b) method for fixing up sets of conditional independencies.

3. MODELS AND DATA

Two simple but important points, each mentioned in both papers and neither having to do directly with graph theory, deserve to be emphasized. First, both papers take the position that a model represents the substantive knowledge that an expert brings to the problem prior to seeing specifically relevant data. One practical consequence of such a position is that statisticians cannot work in a vacuum; rather, they must interact and communicate effectively with domain specialists. And, on a more philosophical note, this position highlights the fact that a scientifically meaningful model for the data is as much a subjective prior assessment of the relative likelihood of possible values as is a scientifically meaningful model for the parameters of such a model. Second, SDLC stress and CW mention that observed data allow us not only to estimate parameters in the model but also to monitor and, if need be, to critique the model. It is refreshing to see frequentists concerned about representing expert knowledge and Bayesians worried about model criticism.

4. SOME QUESTIONS FOR THE AUTHORS

Can you have discrete variables in chain graphs with dashed edges? Can you explain why the diagnostic ability of the Bayesian network was not as good as that of the CART-like algorithm? From Table 6, it appears that for 110 cases (of 168) the Bayesian network assigned the correct diagnosis the highest probability; what were the ranks of the correct diagnoses for the other 58 cases? Has anyone created Bayesian networks with both discrete and continuous variables? Of course, with mixed models the number of parameters in each distribution will not stay fixed after updating. Has anyone considered creating a "Bayesian chip" that could be used to create truly parallel "Bayesian machines"?

Reading and thinking about these papers has been a real pleasure.

Comment: What's Next?

David Madigan

These papers represent two of the many different graphical modeling camps that have emerged from a flurry of activity in the past decade. The paper by Cox and Wermuth falls within the statistical graphical modeling camp and provides a useful generalization of that body of work. There is, of course, a price to be paid for this generality, namely that the interpretation of the graphs is more complex. I cannot resist complementing the authors on the remarkable feat of finding an example for each of the different graphical models they propose.

The paper by Spiegelhalter, Dawid, Lauritzen and Cowell falls within the probabilistic expert system camp. This is a tour de force by researchers responsible for much of the astonishing progress in this area. Ten years ago, probabilistic models were shunned by the artificial intelligence community. That they are now widely accepted and used is due in large measure to the insights and efforts of the authors, along with other pioneers such as Judea Pearl and Peter Cheeseman.

I will confine my remaining comments to the Spiegelhalter et al. paper and explore some open questions that I believe will rapidly become important, now that

David Madigan is Assistant Professor, Department of Statistics, GN-22, University of Washington, Seattle, Washington 98195.