Chapter 10
Graphical Models

A number of the probabilistic models seen in this book are members of the general class of graphical models. In this chapter, we briefly introduce this class of models and examine its relationship with some of those models encountered heretofore.

10.1 Introduction

We have seen a variety of techniques throughout this book for analyzing network data. Many of these techniques – especially in Chapters 6 through 9 – are based on probabilistic models. And among these, a number of them can be viewed profitably as specific instances of a particular class of probabilistic models called *graphical models*. We have chosen not to emphasize this connection previously, as it is not necessarily vital to the exposition of these models individually, nor to their direct application. Nevertheless, this connection is quite important at the level of conceptual, theoretical, and algorithmic foundations. Moreover, some understanding of the general graphical modeling framework is crucial for those interested in developing nontrivial extensions of these types of probabilistic models.

Therefore, in this final chapter we provide a brief introduction to the basic elements of the graphical modeling formalism. Our goal primarily is to make clear the connections between the class of graphical models and those instances of it that we have introduced previously, and to point out some of the implications of these connections. Graphical models have been studied in significant depth in statistics and related areas of computer science, and it is not our intent to provide a comprehensive overview of the topic. For a more in-depth coverage, there are various sources that the interested reader may consult. Useful survey articles include Jordan [216] and Heckerman [195], for example, while full-length treatments may be found in the books by Edwards [129], Lauritzen [248], and Whittaker [403].

E.D. Kolaczyk, *Statistical Analysis of Network Data*, Springer Series in Statistics,
DOI 10.1007/978-0-387-88146-1_10, © Springer Science+Business Media, LLC 2009

10.2 Defining Graphical Models

A *graphical model* is a representation for a joint probability distribution in terms of a graph and a corresponding set of functions defined with respect to that graph. The graph encodes a set of conditional independence relations pertaining to the underlying random variables, which allows – under appropriate conditions – for the joint distribution of these random variables to be decomposed in product form (i.e., to be factorized). The specific nature of the individual terms in this product is dictated by the specified functions. The graphical models formalism provides a useful (and general) approach to handling joint probability distributions, even frequently those of high dimension, by enabling the application of techniques and results from graph theory to probability theory. The synergy that emerges from the union of these two areas has been found to be particularly useful with respect to questions of representation and algorithms.

To characterize graphical models as a subtopic of statistical analysis of network data (the topic of this book) would be unfair. Rather, graphical models are better viewed as a complementary topic – one which is distinct, but at times quite useful for the purposes of analyzing network data. The distinction lies in how we interpret the underlying graph. As an illustration, consider the simple three-vertex graph in Figure 10.1. From the perspective adopted in the first nine chapters of this book, we would associate the three vertices, the two edges, and the flows across these vertices and edges with a system of entities, such as people or computers, linked with respect to some notion of 'relationship,' such as friendship or physical cables, exchanging information of some type. Random variables would then be defined in correspondence with some measurement(s) on this system, and the graph would facilitate the indexing of these variables. Probability modeling might then be used to describe the behavior of these variables, and the graph-based indexing would enter the expressions for these models accordingly. On the other hand, from the perspective of graphical models, we would associate with the vertices a set of three arbitrary random variables – say X, Y, and Z, respectively, from left to right – and the graph would dictate an assumed conditional independence structure of a probability model for these random variables. In particular, the graph in Figure 10.1 is read to indicate that the random variable X is conditionally independent of Z, given Y.

In order to distinguish between these two notions in which graphs may enter into probability modeling, in this chapter we will use $\tilde{G}$ to indicate a graph corresponding to a graphical model, and G to indicate a network graph, in the sense that it has been used throughout the rest of the book. Broadly speaking, there are two main sub-classes of graphical models, namely *directed graphical models* and *undirected graphical models,* which are based on directed acyclic graphs and undirected graphs $\tilde{G}$, respectively. The model mentioned just above is a simple example of an undirected graphical model. We will look at each of these model classes separately, beginning with that of directed graphical models.

Fig. 10.1 A simple three-vertex graph.

10.2.1 Directed Graphical Models

Let $X_1, \ldots, X_N$ be a collection of N random variables. A directed graphical model for this collection allows for the decomposition of the joint distribution $\mathbb{P}$ of $X_1, \ldots, X_N$, with respect to a set of conditional independence relations indexed by parent-child vertex combinations in a directed acyclic graph (DAG), say $\tilde{G} = (\tilde{V}, \tilde{E})$, where $\tilde{V} = \{1, \ldots, N\}$. This decomposition in turn implies a certain Markov property among the random variables X_i.

More precisely, let $\tilde{G}$ be a DAG and let $\{ f(x_i|\mathbf{x}_{pa(i)}) \}_{i \in \tilde{V}}$ be a set of conditional probability density (mass) functions for each random variable X_i, given its parents[1] $\mathbf{X}_{pa(i)}$. A directed graphical model defines a joint probability density (mass) function for $\mathbb{P}$ in the form

$$f(x_1, \ldots, x_N) = \prod_{i \in \tilde{V}} f\left(x_i|\mathbf{x}_{pa(i)}\right) \quad . \tag{10.1}$$

Conversely, given a joint distribution function $\mathbb{P}$, we say that a pre-specified graphical model represents $\mathbb{P}$ if (10.1) holds. More formally, we say in this case that $\mathbb{P}$ admits a recursive factorization with respect to $\tilde{G}$.

An important implication of this factorization is that every variable X_i is conditionally independent, given its parents $\mathbf{X}_{pa(i)}$, of all other variables X_j that are neither the parents nor descendants of X_i. This property is known formally as the *Markov condition*, and in fact it can be shown that the factorization and this condition hold in an if-and-only-if manner. See, for example, Lauritzen [248, Ch. 3.2.2]. Informally, the Markov condition usually is interpreted as saying that each X_i, given its parents, is conditionally independent of its other ancestors.

[1] Recall that $pa(i)$ denotes the index set of the parents of i in $\tilde{V}$.

The practical implication of factorizations like that in (10.1) is to indicate for each variable X_i a certain subset $pa(i)$ of indices j whose random variables X_j directly 'influence' X_i. Since often in applications the size of this subset is, either by nature or design, substantially smaller than the total number of random variables N, this feature can have important algorithmic consequences when performing inference in graphical models. We will revisit this point below in Section 10.3.

Specific directed graphical models are distinguished from each other by (i) the topology of their underlying DAG and (ii) the form of their conditional density (mass) functions $f(x_i|\mathbf{x}_{pa(i)})$. A common – and algorithmically attractive – choice for the graph topology is a tree. As for the conditional density (mass) functions, for discrete random variables, multinomial conditional distributions are popular, while for continuous random variables, conditional Gaussian distributions are frequently used. We have already seen some examples of such directed graphical models in earlier chapters.

Example 10.1 (Tomographic Inference of Tree Topologies). From Section 7.4.1, recall the problem of inferring the topology of a tree $G_T = (V_T, E_T)$, with root r, from measurements taken at its leaves $R \subset V_T$. Our primary illustration of this problem was the task of inferring computer network topologies from multicast probe measurements. The outcome of whether or not a probe reached each of the N_l leaves in R, when sent from the root r, was represented by an N_l-tuple $\{X_1, \ldots, X_{N_l}\}$. Moreover, to represent the full history of the probe within the tree G_T, we defined a cascade process $\{X_j\}_{j \in V_T}$, where $X_j = 1$ if vertex j received a copy of the probe, and $X_j = 0$, otherwise. This cascade process was then equipped with a particular probabilistic model dictating the manner in which the process was assumed to evolve from the root to the leaves.

Specifically, we assumed that $X_r = 1$ and, for each internal vertex k, that

$$\mathbb{P}(X_j = 1 \,|\, X_k = 1) = 1 - \mathbb{P}(X_j = 0 \,|\, X_k = 1) = \alpha_j \quad , \tag{10.2}$$

where $0 < \alpha_j < 1$ for all j. Otherwise, if $X_k = 0$, then $X_j = 0$ for all descendants j of k. This model can be expressed in the form of a directed graphical model as follows. Let $\tilde{G} = (\tilde{V}, \tilde{E})$ be a directed version of the graph G_T, where $\tilde{V} = V_T$ and $\tilde{E}$ has a directed edge if and only if E_T has an edge, with the direction of each such edge being away from the root and towards the leaves. Then we can write

$$\mathbb{P}\left(\mathbf{X} = \mathbf{x}\right) = \prod_{j \in \tilde{V}} \mathbb{P}\left(X_j = x_j \,|\, X_{pa(j)} = x_{pa(j)}\right) \quad , \tag{10.3}$$

where

$$\mathbb{P}\left(X_j = x_j \,|\, X_{pa(j)} = x_{pa(j)}\right) = \begin{cases} 1, & \text{if } j = r \text{ and } x_j = 1 \ , \\ \alpha_j^{x_j} \left(1 - \alpha_j\right)^{1 - x_j}, & \text{if } x_{pa(j)} = 1 \ , \\ 1, & \text{if } x_{pa(j)} = 0 \text{ and } x_j = 0 \quad . \end{cases} \tag{10.4}$$

The expression for the conditional probabilities in (10.4) is defined in analogy to the term in (7.52). Note that $X_{pa(j)}$ in (10.3) is a random variable – instead of a random

vector, as in (10.1) – as a consequence of the fact that non-root vertices have only single parents in a tree. □

Example 10.2 (Kalman Filtering). Recall that a traffic matrix summarizes the volume of traffic flowing between origin-destination pairs on a network graph $G = (V,E)$. In Section 9.3 we studied various methods for inferring characteristics of traffic matrices from measurements of link volumes $\mathbf{X} = (X_e)_{e \in E}$. Among the dynamic methods of traffic matrix estimation we discussed was the method of Kalman filtering. Under slightly modified conditions, the model underlying this method can be represented as a graphical model with respect to a particular type of binary tree.

To see this, recall that our model specified that the time-indexed link volumes $\mathbf{X}^{(t)}$ and the expected origin-destination traffic volumes $^{(t)}$ evolve together with respect to the equations

$$^{(t+1)} = \Xi^{(t)} \ ^{(t)} + \eta^{(t)} \tag{10.5}$$

$$\mathbf{X}^{(t)} = \mathbf{B}^{(t)} \ ^{(t)} + \varepsilon^{(t)} \ . \tag{10.6}$$

These equations are simply (9.44) and (9.45), repeated here for convenience. Our previous assumptions on the noise terms $\eta^{(t)}$ and $\varepsilon^{(t)}$ were that they be zero-mean random vectors, with covariances $\Psi^{(t)}$ and $\Sigma^{(t)}$, respectively, and that they be uncorrelated both with each other and among themselves across times t. If we now make the additional assumption that these terms have multivariate Gaussian distributions, then this lack of correlation is equivalent to independence, which in turn induces a set of conditional independence statements among the variables $^{(t)}$ and $\mathbf{X}^{(t)}$.

Specifically, we have that $^{(t+1)}$ is conditionally independent of $^{(0)}, \ldots, \ ^{(t-1)}$, given $^{(t)}$, and furthermore, that given $^{(t)}$, $\mathbf{X}^{(t)}$ is conditionally independent of all other terms $^{(s)}$ and $\mathbf{X}^{(s)}$. Therefore, formally, we can write

$$f\left(\ ^{(0)}, \ldots, \ ^{(\tau)}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\tau)} \right) = f\left(\ ^{(0)} \right) \prod_{t=1}^{\tau} f\left(\ ^{(t)} \mid \ ^{(t-1)} \right) f\left(\mathbf{x}^{(t)} \mid \ ^{(t)} \right) \ , \tag{10.7}$$

where the third probability density function on the right-hand side corresponds to a multivariate Gaussian distribution, with parameters

$$\mathbb{E}(\mathbf{X}^{(t)} \mid \ ^{(t)}) = \mathbf{B}^{(t)} \ ^{(t)} \quad \text{and} \quad \mathbb{V}(\mathbf{X}^{(t)} \mid \ ^{(t)}) = \Sigma^{(t)} \ , \tag{10.8}$$

and the second, to a multivariate Gaussian distribution, with parameters

$$\mathbb{E}(\ ^{(t)} \mid \ ^{(t-1)}) = \Xi^{(t-1)} \ ^{(t-1)} \quad \text{and} \quad \mathbb{V}(\ ^{(t)} \mid \ ^{(t-1)}) = \Psi^{(t-1)} \ . \tag{10.9}$$

The graph $\tilde{G}$ corresponding to this factorization can be represented as a binary graph for which, among the left- and right-hand children of any parent, only the right-hand child has descendants. Figure 10.2 provides a visualization of this model for the case $\tau = 2$. More generally, this type of graph corresponds to the class of *(first order) hidden Markov models* (HMMs), of which our model is just one instance. For background on HMMs and related models, see MacDonald and Zucchini [266]. □

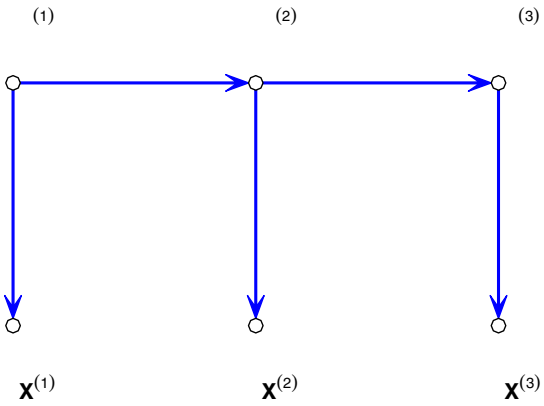

Fig. 10.2 Graphical representation of the model underlying Kalman filtering.

In considering these examples, two points are worth noting. First, the graph $\tilde{G}$ defining a graphical model may or may not correspond to the 'network graph' G, as we have referred to it throughout this book. In Example 10.1, $\tilde{G}$ and G are essentially the same. Indeed, one could argue that a sense of directedness was implicit throughout the exposition of Section 7.4.1. On the other hand, in Example 10.2, the graphs $\tilde{G}$ and G are not at all the same, in that $\tilde{G}$ corresponds to the temporal evolution of the relevant traffic volume random variables, while the network graph G embodies the 'spatial' system over which the traffic actually flows.

Second, although the graphical modeling formalism may not have been necessary to our previous exposition of the models in these two examples, it can be useful in conceptualizing extensions of the models. For example, in the multicast setting, we might treat the parameters α_j as random variables. In the simplest case, we could specify that the α_j be independent random variables, and independent of all X_j as well. As a result, since the X_j are then dependent not just on their parents but also on the corresponding α_j, our original DAG would be augmented to include additional vertices for each α_j and directed edges from each of these vertices to that of its respective X_j. More sophisticated models for the α_j would lead to graphs $\tilde{G}$ with more involved topologies.

10.2.2 Undirected Graphical Models

Let $\tilde{G} = (\tilde{V}, \tilde{E})$ be an undirected graph. Recall that a clique is a complete sub-graph H, and that H is a maximal clique if no other such sub-graph contains it. Let $\mathscr{C}$ be the set of maximal cliques in $\tilde{G}$. An undirected graphical model for a set of discrete random variables $X_1, \ldots, X_N$ is a model that specifies a probability mass function for their joint distribution $\mathbb{P}$ in the form

$$f(x_1, \ldots, x_N) = \left(\frac{1}{\kappa}\right) \prod_{H \in \mathscr{C}} \psi_H(\mathbf{x}_H) \quad , \tag{10.10}$$

where $\psi_H(\mathbf{x}_H)$ is a so-called *potential function*, which is a positive (but otherwise arbitrary) function of the variables $\{x_i\}_{i \in H}$, and

$$\kappa = \sum_{\mathbf{x}} \prod_{H \in \mathscr{C}} \psi_H(\mathbf{x}_H) \tag{10.11}$$

is a normalization factor. The model is defined analogously when the X_i are continuous random variables.

The factorization in (10.10) can be shown to imply that $\mathbb{P}$ satisfies a so-called *global Markov property.* That is, for $A, B, C \subset \tilde{V}$, the vector of random variables $\mathbf{X}_A = (X_i)_{i \in A}$ is independent of $\mathbf{X}_B = (X_i)_{i \in B}$, given $\mathbf{X}_C = (X_i)_{i \in C}$, if and only if the subset C separates the subsets A and B in the graph $\tilde{G}$, in the sense that every path between a vertex in A and a vertex in B intersects a vertex in C. This property in turn is equivalent to the so-called *pairwise Markov property,* which states that X_i is conditionally independent of X_j, given the other X_k, for $k \in \tilde{V} \setminus \{i, j\}$, if and only if there is no edge between i and j in $\tilde{G}$. In fact, the converse can be true as well here, in that, under appropriate conditions, if a distribution $\mathbb{P}$ satisfies the pairwise Markov condition, it factorizes with respect to $\tilde{G}$, as in (10.10). This latter result is generally attributed to Hammersley and Clifford [191], who proved it for the case of discrete random variables. See Lauritzen [248, Ch. 3.2] for details on this and the rest of the above discussion.

Analogous to the case of directed graphical models, specific undirected graphical models are distinguished by the choice of the undirected graph $\tilde{G}$ and the potential functions ψ_H. We have seen a number of examples of undirected graphical models in previous chapters.

Example 10.3 (Markov Random Fields). In Section 8.3.1 we introduced a class of Markov random field models on a network graph $G = (V, E)$, for a vector of attributes $\mathbf{X} = (X_i)_{i \in V}$ on the vertices $i \in V$. Specifically, a discrete random vector $\mathbf{X}$ was said to be a Markov random field on G if its probability mass function obeyed the condition

$$\mathbb{P}\left(X_i = x_i \,|\, \mathbf{X}_{(-i)} = \mathbf{x}_{(-i)}\right) = \mathbb{P}\left(X_i = x_i \,|\, \mathbf{X}_{\mathscr{N}_i} = \mathbf{x}_{\mathscr{N}_i}\right) \quad , \tag{10.12}$$

where recall that $\mathbf{X}_{(-i)}$ is the vector $(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_{N_v})^T$ and $\mathbf{X}_{\mathscr{N}_i}$ is the vector of all X_j for $j \in \mathscr{N}_i$, with $\mathscr{N}_i$ defined to be the neighborhood of i in G.

The condition in (10.12) is equivalent to the global Markov condition. Furthermore, under appropriate conditions, the Hammersley-Clifford theorem establishes the equivalence of the full joint distribution of $\mathbf{X}$ with the so-called Gibbs distribution in (8.5). See Besag [36]. In particular, recall that under such conditions we have

$$\mathbb{P}(\mathbf{X}=\mathbf{x}) = \left(\frac{1}{\kappa}\right)\exp\{U(\mathbf{x})\} \quad , \tag{10.13}$$

where $U(\cdot)$ is the energy function and κ is a normalization factor (often called the partition function), and where the energy function can be decomposed as a sum over all (maximal) cliques $H \in \mathscr{C}$, in the form

$$U(\mathbf{x}) = \sum_{H\in\mathscr{C}} U_H(\mathbf{x}) \quad . \tag{10.14}$$

In other words, the Markov random field model can be expressed in the form of (10.10), with $\tilde{G} = G$ and potential functions defined as

$$\psi_H(\mathbf{x}_H) = \exp\{U_H(\mathbf{x})\} \quad . \tag{10.15}$$

Several choices of such potential functions were provided in Example 8.2. We note that undirected graphical models are often referred to synonymously with Markov random fields. □

Example 10.4 (Gaussian Graphical Models). Recall from Section 7.3.3 our discussion of Gaussian graphical models for continuous attributes X_i corresponding to elements $i \in \{1,\ldots,N\}$. In particular, suppose that $\mathbf{X} = (X_i)_{i=1}^N$ has a multivariate Gaussian distribution, with expectation $\mathbb{E}(\mathbf{X}) =$ and variance $\mathbb{V}(\mathbf{X}) = \Sigma$, and let $\Omega = \Sigma^{-1}$ denote the precision matrix. The graph $G = (V,E)$ was defined so that $V = \{1,\ldots,N\}$ and

$$E = \left\{ \{i,j\} \in V^{(2)} : \rho_{ij|V\setminus\{i,j\}} \neq 0 \right\} \quad , \tag{10.16}$$

where

$$\rho_{ij|V\setminus\{i,j\}} = \frac{-\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}} \tag{10.17}$$

is the partial correlation of X_i and X_j adjusted for all other $X_k, k \in V \setminus \{i,j\}$. Since a partial correlation of zero is equivalent to conditional independence in the Gaussian setting, the graph G – the conditional independence graph – encodes the set of pairwise Markov relations among the X_i.

The impact of these conditional independence relations is evident, for example, in the log-likelihood, the quadratic component for which, without loss of generality, in the case of $= 0$ takes the form

$$\mathbf{x}^T \Omega \mathbf{x} = \sum_{i \in V} \omega_{ii} x_i^2 + 2 \sum_{\{i,j\} \in E} \omega_{ij} x_i x_j \ . \tag{10.18}$$

However, this expression does not itself directly yield a factorization like that in (10.10). Some additional work is necessary here, through which it can be shown that it is possible to write

$$f(\mathbf{x}) = \frac{\prod_{H \in \mathscr{C}} f(\mathbf{x}_H)}{\prod_{S \in \mathscr{S}} [f(\mathbf{x}_S)]^{\nu(S)}} \ , \tag{10.19}$$

where $f(\mathbf{x}_A)$ is the joint density of $\mathbf{X}_A = (X_i)_{i \in A}$, $\mathscr{S}$ is a set of so-called separators indexed according to a particular ordering (i.e., a so-called perfect sequence) of the cliques H in $\mathscr{C}$, and $\nu(S)$ is the number of times the separator S occurs in the sequence. See Lauritzen [248, Ch. 5.3.1]. □

Example 10.5 (Exponential Random Graph Models). The exponential random graph models (ERGMs) introduced in Section 6.5 provide a general approach to modeling the dependencies among edges in a network graph. Recall that these models specify an exponential family distribution for a random matrix $\mathbf{Y}$, where the latter describes the adjacencies in an undirected random graph G. And note that the expression for $\mathbb{P}(\mathbf{Y} = \mathbf{y})$ in (6.24) is qualitatively quite similar to that in (10.10). However, importantly, the summation in (6.24) is over an arbitrary set of 'configurations' (i.e., possible sets of edges among subsets of vertices in the network graph G), whereas that in (10.10) is specifically over maximal cliques in a conditional independence graph $\tilde{G}$. Under certain specifications, an ERGM will satisfy the appropriate Markov conditions needed to be equivalent to an undirected graphical model. On the other hand, many ERGM formulations do not satisfy these conditions, yet continue – by definition – to retain an exponential family form.

Frank and Strauss [155] introduced the notion of a pairwise Markov property to the modeling of random network graphs G, specifying that the edge random variables Y_{ij} and $Y_{i'j'}$ be conditionally independent, given the other random variables $Y_{k,l}$, if and only if they do not share a vertex. This condition can be represented using a conditional independence graph $\tilde{G} = (\tilde{V}, \tilde{E})$, such that each unordered pair of vertices $\{i, j\} \in V^{(2)}$ corresponds to a vertex in $\tilde{V}$, and pairs of vertices in $\tilde{V}$ have an edge between them in $\tilde{E}$ if and only if the corresponding variables in $\mathbf{Y}$ are conditionally dependent. Using the Hammersley-Clifford theorem and the type of homogeneity assumption discussed in Section 6.5, Frank and Strauss establish that the distribution of $\mathbf{Y}$ satisfies this pairwise Markov property if and only if

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \left(\frac{1}{\kappa}\right) \exp\left\{ \sum_{k=1}^{N_v - 1} \theta_k S_k(\mathbf{y}) + \theta_\tau T(\mathbf{y}) \right\} \ , \tag{10.20}$$

where $S_1(\mathbf{y}) = N_e$ is the number of edges, $S_k(\mathbf{y})$ is the number of k-stars, and $T(\mathbf{y})$ is the number of triangles. Recall that such Markov random graphs also include the Bernoulli random graph, discussed in Example 6.6, as a special case.

More generally, other ERGM formulations, such as those based on the alternating k-stars and alternating k-triangles defined in (6.31) and (6.33), respectively, do not satisfy the pairwise Markov condition and hence cannot be represented as in (10.10). Nevertheless, it is possible to conceptualize these models – and others like them – within the framework of a somewhat broader notion of conditional independence, that of 'partial conditional dependence.' See Pattison and Robins [311] and Robins and Pattison [330] for details. □

As in the case of directed graphical models, the framework of undirected graphical models can be useful in facilitating extensions of existing models. See, for example, Robins, Pattison, and Elliot [331], who offer an extension of the Hammersley-Clifford theorem in the context of joint and conditional models for a network adjacency matrix $\mathbf{Y}$ and a vector $\mathbf{X}$ of vertex attributes.

10.3 Inference for Graphical Models

There are a variety of tasks in graphical modeling that can be referred to, broadly speaking, as 'inferential' tasks – or perhaps more generally, as tasks of statistical learning. Such tasks can range from being relatively straightforward to intractable. The complexity of inferential tasks varies as a function of both graph topology (i.e., model complexity) and the task itself. From a computational point of view, statistical inference in this area is looked at as being fundamentally a problem of computing relevant conditional – or, more generally, marginal – probabilities. That is, given a graphical model, with graph $\tilde{G} = (\tilde{V}, \tilde{E})$ and joint probability mass (density) function f, we wish to compute

$$f(\mathbf{x}_A) = \sum_{\mathbf{x}' : \mathbf{x}'_A = \mathbf{x}_A} f(\mathbf{x}') \ , \tag{10.21}$$

for some $A \subset \tilde{V}$. Quantities like means and other values involving expectations can be obtained from such probabilities accordingly.

There are three main categories of methods for graphical model inference: exact methods, sampling methods, and variational methods. We characterize each of these methods briefly in the paragraphs that follow. For more details, see the survey article by Jordan [216], the edited volume by Jordan [215], and the many various references therein.

Exact methods of inference, as their name indicates, seek to compute the requisite marginal probabilities exactly. The primary hurdle to this computation is the summation (or, in the continuous case, integration) over all possible combinations of values for the variables $\{X_i\}_{i \in \tilde{V} \setminus A}$ not appearing in $f(\mathbf{x}_A)$. Counting terms in the summation, it is easy to see that if each X_i has a range of r values, the computational complexity of naively evaluating just one such probability can be seen to scale like $O(r^m)$, where $m = |\tilde{V} \setminus A|$. Therefore, algorithms in this area seek to increase computational efficiency through exploitation of the structure of the independence graph

$\tilde{G}$. The computational gains realized by such techniques are particularly important when more than one probability must be computed, for instance, when the computation is just one element of a larger optimization algorithm.

One well-known example of such algorithms is the *sum-product* algorithm, which allows for exact computation on tree-structured graphs. This algorithm leverages the fact that the joint probability mass function $f(\mathbf{x})$ can be written in product form (i.e., as in (10.1) or (10.10)), by employing a pattern of inter-changes of summation and product in marginalizing $f(\mathbf{x})$ to obtain $f(\mathbf{x}_A)$, so as to eliminate redundancy entailed by re-calculation of intermediate partial sums. A related algorithm, and essentially a special case of the sum-product algorithm, is *belief propagation* (Pearl [312] and Lauritzen and Spiegelhalter [249]), which performs exact inference on directed acyclic graphs. Also related is the *max-sum* algorithm, which uses an analogous strategy to maximize a probability function and return the corresponding argument. These types of algorithms are based on principles of dynamic programming and are sometimes called *message-passing* algorithms, since the computational process can be diagrammed as a sequence of steps in which 'messages' are 'passed' locally between adjacent vertices. A gentle introduction to these algorithms may be found in Bishop [39, Ch. 8.4], who also talks briefly about their extension for doing approximate inference in more general graphs (i.e., other than trees or DAGs).

Implementation of exact algorithms can require a certain amount of preliminary preparation. For example, they tend to be defined in terms of undirected graphs; their usage with directed graphs actually involves an initial conversion of the graph to a particular undirected graph called the *moral graph*. In addition, generally there is a choice in the manner in which the calculations are organized. For example, there can be multiple possible orders in which the alternating sums and products are performed in the sum-product algorithm. Finding an optimal ordering that minimizes the overall computational complexity is in general an *NP*-hard problem, but various heuristics exist. Jordan [216, Sec. 3.1] provides some discussion.

Exact algorithms are always in principle available, but in practice may be infeasible, depending on the order and structure of the graph underlying the graphical model. In such settings, methods of approximate inference become attractive. *Sampling methods* offer one such alternative, and Gibbs sampling algorithms are a standard choice in this area. Recall that the basic Gibbs sampling algorithm involves only the computation of univariate conditional probabilities of the form $f(x_i \,|\, \mathbf{x}_{(-i)})$. Because of the Markov properties associated with graphical models, these computations typically reduce to computing $f(x_i \,|\, \mathbf{x}_{\mathcal{N}_i})$. That is, the graphical model structure dictates a usually reduced set of variables to condition upon, for each variable X_i. We saw a detailed discussion of Gibbs sampling for inference of parameters and predictions in Markov random field models in Section 8.3.1. See Jordan [216, Sec. 3.2] for additional details and references.

Another class of methods of approximate inference in graphical models, and one that has seen some notable research activity in the past decade, is that of *variational methods*. These methods can be viewed as characterizing the calculation of probabilities as an optimization problem and proposing to solve a perturbed (and

more tractable) form of that problem, rather than the original problem itself. Such perturbed problems can be obtained, for example, by ignoring certain of the dependencies indicated by the graphical model and working instead with a probability distribution of simpler dependency structure. For details, we refer the reader to Jordan [216, Sec. 3.3] and the references therein.

10.4 Additional Related Topics and Reading

There are other representations for graphical models encountered in the literature besides the directed and undirected graphs we have focused upon. Two important examples are the class of *chain graphs*, which combine undirected and directed edges within a single graph, and *factor graphs*, which are bipartite graphs that provide a more nuanced factorization in the undirected case. Also, it should be noted that terminology varies somewhat in this area. For example, directed graphical models often are called *Bayesian networks*, although they are used in both frequentist and Bayesian settings. And, as was mentioned previously, undirected graphical models sometimes are referred to alternatively as Markov random fields. Finally, we mention that there are various related classes of probabilistic models that are relevant to the analysis of network data. The class of *probabilistic relational models,* for example, as developed largely by Koller and colleagues, can be viewed as an extension of directed graphical models to include relations as variables. A comprehensive overview of this more relatively recent class of models can be found in the volume edited by Getoor and Taskar [168].