



Figure 16.4: This graph implies that $p(a, b, c, d, e, f)$ can be written as $\frac{1}{Z} \phi_{a,b}(a, b) \phi_{b,c}(b, c) \phi_{a,d}(a, d) \phi_{b,e}(b, e) \phi_{e,f}(e, f)$ for an appropriate choice of the ϕ functions.

and $E(\mathbf{x})$ is known as the **energy function**. Because $\exp(z)$ is positive for all z , this guarantees that no energy function will result in a probability of zero for any state \mathbf{x} . Being completely free to choose the energy function makes learning simpler. If we learned the clique potentials directly, we would need to use constrained optimization to arbitrarily impose some specific minimal probability value. By learning the energy function, we can use unconstrained optimization.⁵ The probabilities in an energy-based model can approach arbitrarily close to zero but never reach it.

Any distribution of the form given by equation 16.7 is an example of a **Boltzmann distribution**. For this reason, many energy-based models are called **Boltzmann machines** (Fahlman *et al.*, 1983; Ackley *et al.*, 1985; Hinton *et al.*, 1984; Hinton and Sejnowski, 1986). There is no accepted guideline for when to call a model an energy-based model and when to call it a Boltzmann machine. The term Boltzmann machine was first introduced to describe a model with exclusively binary variables, but today many models such as the mean-covariance restricted Boltzmann machine incorporate real-valued variables as well. While Boltzmann machines were originally defined to encompass both models with and without latent variables, the term Boltzmann machine is today most often used to designate models with latent variables, while Boltzmann machines without latent variables are more often called Markov random fields or log-linear models.

Cliques in an undirected graph correspond to factors of the unnormalized probability function. Because $\exp(a) \exp(b) = \exp(a + b)$, this means that different cliques in the undirected graph correspond to the different terms of the energy function. In other words, an energy-based model is just a special kind of Markov network: the exponentiation makes each term in the energy function correspond to a factor for a different clique. See figure 16.5 for an example of how to read the

⁵For some models, we may still need to use constrained optimization to make sure Z exists.