



Figure 16.5: This graph implies that $E(a, b, c, d, e, f)$ can be written as $E_{a,b}(a, b) + E_{b,c}(b, c) + E_{a,d}(a, d) + E_{b,e}(b, e) + E_{e,f}(e, f)$ for an appropriate choice of the per-clique energy functions. Note that we can obtain the ϕ functions in figure 16.4 by setting each ϕ to the exponential of the corresponding negative energy, e.g., $\phi_{a,b}(a, b) = \exp(-E(a, b))$.

form of the energy function from an undirected graph structure. One can view an energy-based model with multiple terms in its energy function as being a **product of experts** (Hinton, 1999). Each term in the energy function corresponds to another factor in the probability distribution. Each term of the energy function can be thought of as an “expert” that determines whether a particular soft constraint is satisfied. Each expert may enforce only one constraint that concerns only a low-dimensional projection of the random variables, but when combined by multiplication of probabilities, the experts together enforce a complicated high-dimensional constraint.

One part of the definition of an energy-based model serves no functional purpose from a machine learning point of view: the $-$ sign in equation 16.7. This $-$ sign could be incorporated into the definition of E . For many choices of the function E , the learning algorithm is free to determine the sign of the energy anyway. The $-$ sign is present primarily to preserve compatibility between the machine learning literature and the physics literature. Many advances in probabilistic modeling were originally developed by statistical physicists, for whom E refers to actual, physical energy and does not have arbitrary sign. Terminology such as “energy” and “partition function” remains associated with these techniques, even though their mathematical applicability is broader than the physics context in which they were developed. Some machine learning researchers (e.g., Smolensky (1986), who referred to negative energy as **harmony**) have chosen to omit the negation, but this is not the standard convention.

Many algorithms that operate on probabilistic models do not need to compute $p_{\text{model}}(\mathbf{x})$ but only $\log \tilde{p}_{\text{model}}(\mathbf{x})$. For energy-based models with latent variables \mathbf{h} , these algorithms are sometimes phrased in terms of the negative of this quantity,