Clustering coefficients of protein-protein interaction networks

Gerald A. Miller,[1] Yi Y. Shi,[2] Hong Qian,[2] and Karol Bomsztyk[3]
[1]*Department of Physics, University of Washington Seattle, Seattle, Washington 98195, USA*
[2]*Department of Applied Mathematics, University of Washington Seattle, Seattle, Washington 98195, USA*
[3]*Department of Medicine, University of Washington Seattle, Seattle, Washington 98109, USA*
(Received 14 November 2006; published 16 May 2007)

The properties of certain networks are determined by hidden variables that are not explicitly measured. The conditional probability (propagator) that a vertex with a given value of the hidden variable is connected to k other vertices determines all measurable properties. We study hidden variable models and find an averaging approximation that enables us to obtain a general analytical result for the propagator. Analytic results showing the validity of the approximation are obtained. We apply hidden variable models to protein-protein interaction networks (PINs) in which the hidden variable is the association free energy, determined by distributions that depend on biochemistry and evolution. We compute degree distributions as well as clustering coefficients of several PINs of different species; good agreement with measured data is obtained. For the human interactome two different parameter sets give the same degree distributions, but the computed clustering coefficients differ by a factor of about 2. This shows that degree distributions are not sufficient to determine the properties of PINs.

DOI: 10.1103/PhysRevE.75.051910 PACS number(s): 87.10.+e, 89.75.Hc

I. INTRODUCTION

Physicists have recently shown that network analysis is a powerful tool to study the statistical properties of complex biological, technological, and social systems of diverse kinds [1–3]. Many networks exhibit a scale-free degree distribution in which the probability p_k that a vertex is connected to k other vertices falls as a power $p_k \sim k^{-\gamma}$. This property is not sufficient to completely describe natural networks because such systems also exhibit degree correlations—the degrees of the vertices at the end points of any given edge are not independent [4–7]. It is not surprising that natural systems depend on properties that do not appear explicitly in degree distributions. In particular, protein interaction networks depend on the availability of sufficient binding free energy [8] to cause interactions to occur (i.e., links between vertices to exist).

Caldarelli *et al.* [9] and Söderberg [10] proposed models in which vertices are characterized by a fitness parameter assigned according to a chosen probability distribution. Then pairs of vertices are independently joined by an undirected edge with a probability depending on the fitnesses of the end points. Reference [11] generalized these models as a class of models with hidden variables and presented a detailed formalism showing how to compute network properties using the conditional probability (propagator) that a vertex with a given value of a hidden variable is connected to k other vertices. This formalism, valid for any Markovian (binary) network, provides the generating function for the propagator, but not the propagator itself.

The purpose of this paper is twofold. We first use a mean field approximation to derive a general analytic formula for the propagator, therefore finding a general approximate solution to the inversion problem. This enables one to compute network properties without the use of a simulation procedure, thereby simplifying the computational procedure and potentially broadening the ability of scientists from all fields to use network theory. The validity of the method is assessed by comparing the results of using our approximation with published results. We then use this method to compute clustering coefficients of a specific hidden variable model for protein-protein interaction networks (PINs) from several organisms, developed by us [12], that previously had obtained degree distributions in agreement with measured data. We show that two models with the same degree distribution have very different clustering coefficients.

We outline this in more detail. Section II reviews the hidden variable formalism and our approximate solution to the inversion problem. We distinguish between sparse (which have been solved in Ref. [11]) and nonsparse networks (which are solved here). Section III studies the models of Refs. [9,13]. Our averaging procedure is found to work well for most situations. Our own model [12] is presented in Sec. IV. We present an analytic result for the average connection probability and extend the results of [12] to computing the clustering coefficients. The final section is reserved for a brief summary and discussion.

II. HIDDEN VARIABLE NETWORKS

We present the formalism for hidden variable models [11]. The probability that a node has a hidden continuous variable g is given by $\rho(g)$, normalized so that its integral over its domain is unity. This function is chosen to be an exponential in [9,12] and a Gaussian in [13]. The connection probability for two nodes of g, g' is defined to be $p(g,g')$. This is taken as a step function in [9,13], and a Fermi function in [12]. The two functions $\rho(g)$ and $p(g,g')$ can be chosen in a wide variety of ways to capture the properties of a given network. Reference [11] presents the probability generating function $G_0(x)$ that determines p_k in terms of the generating function for the propagator, $\hat{G}_0(z,g)$, as

$$G_0(z) = \int dg\, \rho(g) \hat{G}_0(z,g), \tag{1}$$

where

$$\ln \hat{G}_0(z,g) = N \int dg' \rho(g') \ln[1-(1-z)p(g,g')]. \tag{2}$$

The propagator $G_0(k,g)$ giving the conditional probability that a vertex of hidden variable g is connected to k other vertices is given implicitly by

$$\hat{G}_0(z,g) = \sum_{k=0}^{\infty} z^k G_0(k,g). \tag{3}$$

Knowledge of $G_0(k,g)$ determines the conditional probability $P(k'|k)$ that a node of degree k is connected to a node of degree k' [11] (as well as p_k), and those two functions completely define a Markovian network. Once $G_0(k,g)$ is determined, all of the properties of the given network are determined. The most well-known example is the degree distribution p_k:

$$p_k = \int_0^{\infty} dg\, \rho_\lambda(g) G_0(k,g). \tag{4}$$

It would seem that determining $G_0(k,g)$ from Eq. (2) is a simple technical matter, but this is not the case [11]. The purpose of the present section is to provide a simple, analytic, and accurate method to determine $G_0(k,g)$.

We obtain $G_0(k,g)$ from Eq. (2) by using the tautology

$$p(g,g') = \bar{p}(g) + [p(g,g') - \bar{p}(g)] \tag{5}$$

in Eq. (2), choosing $\bar{p}(g)$ so as to eliminate the effects of the second term, and then treating the remaining higher powers of $[p(g,g')-\bar{p}(g)]$ as an expansion parameter. Using Eq. (5) in Eq. (2) yields

$$\begin{aligned}\ln \hat{G}_0(z,g) &= \ln \hat{G}_0(z,g) \\ &= \ln[1-(1-z)\bar{p}(g)]^N \\ &\quad - N(1-z)\int dg' \rho(g') \frac{[\bar{p}(g) - p(g,g')]}{1-(1-z)\bar{p}(g)} \\ &\quad - N\sum_{n=2}^{\infty} \frac{(1-z)^n}{n} \int dg' \rho(g') \left(\frac{p(g,g') - \bar{p}(g)}{1-(1-z)\bar{p}(g)} \right)^n .\end{aligned} \tag{6}$$

In analogy with the mean-field (Hartree) approximation of atomic and nuclear physics, we find that the second term of Eq. (6) vanishes if we choose $\bar{p}(g)$ to be the average of $p(g,g')$ over $\rho(g')$:

$$\bar{p}(g) = \int dg' \rho(g') p(g,g'). \tag{7}$$

With Eq. (7) the effects of the term of first order in $[p(g,g')-\bar{p}(g)]$ vanish. We therefore obtain the result

$$\ln \hat{G}_0(z,g) = \ln[1-(1-z)\bar{p}(g)]^N - N\sum_{n=2}^{\infty} \frac{(1-z)^n}{n} \times \int dg' \rho(g') \left(\frac{p(g,g') - \bar{p}(g)}{1-(1-z)\bar{p}(g)} \right)^n, \tag{8}$$

with the putative term with $n=1$ vanishing by virtue of Eq. (7).

We treat the first term of Eq. (8) as the leading order (LO) term and regard the remainder as a correction. The validity of this approach can be checked by comparison with simulations, or (in certain cases) with analytic results. Numerical results for the PIN of current interest [12] indicate that the corrections to the LO terms induce errors in p_k of no more than a few percent and that the approximation becomes more accurate for large values of k. Therefore we use the LO approximation. Using exponentiation and the binomial theorem in the first term of Eq. (8) leads to the result

$$G_0^{(\mathrm{LO})}(k,g) = \binom{N}{k} [1-\bar{p}(g)]^{N-k} \bar{p}(g)^k, \tag{9}$$

which is of the form of a random binomial distribution in which the connection probability depends on the hidden variable g. Equation (9) is our central new general result that can be used for any hidden variable network. This binomial distribution has both the normal Gaussian and Poisson [$Np(g) \ll 1$] distributions as limiting cases.

A. Sparse and nonsparse networks

Reference [11] explained the difference between sparse and nonsparse networks. Sparse networks have a well-defined thermodynamic limit for the average degree, while this quantity diverges as the network size N approaches infinity. Reference [11] defines criteria for sparseness by pointing out the relevance of $\bar{p}$ of Eq. (7) in determining whether or not a network is sparse. Given this quantity, the average degree is

$$\langle k \rangle = \int dg\, \rho(g) \bar{p}(g) = \int dg \int dg' \rho(g) p(g,g') \rho(g'). \tag{10}$$

If $\rho(g)$ is independent of N the only way to obtain a nondivergent value $\langle k \rangle$ is for the connection probability to scale as N^{-1}:

$$p^{\mathrm{sparse}}(g,g') = \frac{C(g,g')}{N}, \quad \text{sparse network } t\ [11]. \tag{11}$$

Under the specific assumption that Eq. (11) holds, Ref. [11] finds a very interesting result. In our notation, this amounts to using Eq. (11) in Eq. (2) and taking the limit that N approaches infinity. Then

$$G_0^{\mathrm{sparse}}(z,g) = \exp(z-1) \int dg' \rho(g') C(g,g'). \tag{12}$$

This shows that the Poisson limit of Eq. (9) is obtained for the very special case of sparse networks in which the con-

nection probability scales as N^{-1}. None of the models of interest here [9,12,13] are sparse, so it is our present result (9) that is widely applicable.

B. General networks

Turning to the use of the propagator, we obtain the degree distribution as

$$p_k = \int dg\ \rho(g) G_0(k,g) \approx \int dg\ \rho(g) G_0^{(\mathrm{LO})}(k,g). \quad (13)$$

This expression can be thought of as averaging a binomial distribution over the hidden variable and is a natural generalization of classical graph theory. A similar expression for p_k has been obtained, in the Poisson limit, in Ref. [15]. In that work, p_k is presented as an integral of the Poisson distribution for $p(g)$ multiplied by the "P representation" of a density matrix. Comparing Eq. (9) with the result (3) of [15] shows that our propagator is proportional to the P representation, essentially our $\rho(g)$. Reference [15] shows how, under certain assumptions, to use $p(k)$ to determine the P representation. Our method allows underlying network properties, denoted by $\rho(g)$ and $p(g,g')$, to predict various network properties.

The clustering coefficient measures transitivity [3]: if vertex A is connected to vertex B and vertex B to vertex C, there is an increased probability that vertices A and C are connected. In graph theory, the clustering coefficient $c(k)$ is the ratio of the number of triangles to the number of pairs, computed for nodes of degree k. Reference [11] shows that

$$c(k) = \frac{1}{p_k} \int dg\ \rho(g) G_0(k,g) c(g), \quad (14)$$

$$c(g) = \int dg' \int dg'' \frac{\rho(g')p(g,g')}{\bar{p}(g)} p(g',g'') \frac{\rho(g'')p(g'',g)}{\bar{p}(g)}. \quad (15)$$

Our calculations replace G_0 by $G_0^{(\mathrm{LO})}$ of Eq. (9).

III. SIMPLE MODELS AND ANALYTIC RESULTS

One way to verify the LO approximation is to show that it reproduces analytic results for previously published models. We consider the models of [9,13] in this section. In both of these models $p(g,g')$ is taken as a step function (the zero-temperature limit of our model):

$$p(g,g') = \Theta(g + g' - \mu). \quad (16)$$

The two models differ in their choice of $\rho(g)$, but the use of Eq. (16) allows one to obtain compact general expressions for the generating functions $\hat{G}_0(z,g)$, $\hat{G}_0(k,g)$, p_k, and $c(k)$. We present these first and discuss specific details of the individual models in separate subsections.

The use of Eq. (16) in Eq. (2) yields

$$\ln \hat{G}_0(z,g) = N\left(\Theta(\mu - g)\int_{\mu-g}^{\infty} dg'\rho(g') + \Theta(g-\mu)\right)\ln(z)$$
$$= N\bar{p}(g)\ln(z), \quad (17)$$

so that

$$\hat{G}_0(z,g) = z^{N\bar{p}(g)}. \quad (18)$$

It is interesting to observe that Eq. (8) reduces to the above result. This is because powers of $p(g,g')^m = p(g,g')$ for Eq. (16), so that the integration appearing in Eq. (8) leads to an expression that is a function of $N, z, \bar{p}$. Then the use of the binomial theorem allows the second term of Eq. (8) to be expressed as a summable power series in $\bar{p}$ which ultimately leads to the result Eq. (18).

If we follow [11] and treat k as a continuous variable (which requires large values of k) we find

$$\hat{G}_0(k,g) = \delta(k - N\bar{p}(g)) \quad (19)$$

$$= \frac{\delta(g - g_N(k))}{N|\bar{p}'(g_N(k))|}, \quad (20)$$

where $g_N(k)$ is the solution of the equation

$$k = N\bar{p}(g). \quad (21)$$

Note that, for $k=N$, $g_N(k)$ can take on any value greater than μ. The result Eq. (19) is the same as Eq. (34) of [11], but written in a more compact form. The use of Eq. (19) in Eqs. (13) and (14) yields the results

$$p_k = \frac{\rho(g_N(k))}{N|\bar{p}'(g_N(k))|}, \quad (22)$$

$$\bar{c}(k) = \frac{c(g_N(k))}{N|\bar{p}'(g_N(k))|}. \quad (23)$$

Model of Caldarelli *et al.* [9]

This model is defined by using $\rho(g) = \exp(-g)$, but we generalize it to take the form

$$\rho_\lambda(g) = \lambda \exp(-\lambda g). \quad (24)$$

Reference [11] works out this model using the Green's function formalism. Our purpose here is to compare the results of our averaging approximation with their results. For this model the average interaction probability $\bar{p}(g)$ is given by

$$\bar{p}(g) = \int_0^{\infty} dg' \lambda \exp(-\lambda g')\Theta(g + g' - \mu)$$
$$= \Theta(g-\mu) + \Theta(\mu - g)\exp[-\lambda(\mu - g)]. \quad (25)$$

Then our approximation Eq. (13) for the degree distribution p_k is given by

$$p_k = \binom{N}{k} \int_0^{\mu} dg\ \lambda \exp(-\lambda g)\exp[-k\lambda(\mu - g)] \times \{1 - \exp[-\lambda(\mu - g)]\}^{N-k}. \quad (26)$$

Define the integration variable $t \equiv \exp[-\lambda(\mu - g)]$ so that

$$p_k = \binom{N}{k} e^{-\lambda\mu} \int_{t_0}^{1} \frac{dt}{t^2} t^k (1-t)^{N-k}, \quad t_0 \equiv e^{-\lambda\mu}, \quad (27)$$

$$p_{k>1} = \binom{N}{k} e^{-\lambda\mu} \left(\frac{\Gamma(N+1-k)\Gamma(k-1)}{\Gamma(N)} - B_{t_0}(k-1, N+1-k) \right), \quad (28)$$

$$p_{k=1} = N e^{-\lambda\mu} \frac{(1-t_0)^N}{N}\ {}_2F_1(1, N; N+1, 1-t_0), \quad (29)$$

where ${}_2F_1$ is the confluent hypergeometric function and B_{t_0} is the incomplete Beta function (and with $t_0 = 1$ the Beta function):

$$B_z(a,b) \equiv \int_0^z dt\ t^{a-1}(1-t)^{b-1}, \quad B_1(a,b) = B(a,b). \quad (30)$$

Consider the case

$$1 < k, \quad \lambda\mu \approx 10 \quad (31)$$

(the latter is typical of our biological model) so that the second term of Eq. (28) can be neglected. Evaluating the remaining Gamma functions gives

$$p_k = e^{-\lambda\mu} \frac{N}{k(k-1)}. \quad (32)$$

Reference [11] computes the degree distribution for this model in an analytic manner, using the approximation Eq. (19) in which k is treated as a continuous variable and therefore "is expected to perform poorly for small values of k." The result of [11] (p_k^{BPS}) is

$$p_k^{\mathrm{BPS}} = e^{-\lambda\mu} \frac{N}{k^2} + e^{-\lambda\mu} \delta(k - N) \quad (33)$$

which corresponds to agreement (for $k \neq N$) within the stated domain of accuracy of Ref. [11]. The confluence of Eqs. (32) and (33) provides a verification of the accuracy of the averaging approximation.

The results for $k = N$ seem to disagree, so we examine this more closely. Use Eq. (18) directly to obtain the generating function $G_0(z)$ as $G_0(z) = \int dg\ \rho(g) z^{N\bar{p}(g)}$. One obtains a result z^N for all values of g ($g > \mu$) such that $\bar{p}(g) = 1$. Using this generating function yields the result

$$p_{k=N} = \int dg\ \rho(g)\Theta(g - \mu). \quad (34)$$

The specific value of the integral depends on the choice of $\rho(g)$, but the result is a finite number for any choice of $\rho(g)$ that satisfies the normalization condition that its integral over its domain is unity. Thus we believe that the correct result of using the propagator [Eq. (34) of [11] in their Eq. (11)] is

$$p_k^{\mathrm{BPS}} = e^{-\lambda\mu} \frac{N}{k^2} \quad (35)$$

instead of Eq. (33), which is in agreement with our result.

Our approximation works very well in reproducing the computed clustering coefficient of [11]. In particular, we evaluate $c(g)$ of Eq. (15) to find that

$$\bar{c}(k) = \frac{1}{p_k} \left(\int_0^{\mu/2} \exp(-g) G_0^{(\mathrm{LO})}(k,g) + \int_{\mu/2}^{\mu} \exp(-g) G_0^{(\mathrm{LO})}(k,g)(2g - \mu + 1) \right). \quad (36)$$

Numerical evaluation of this approximate expression accurately reproduces the result of Fig. 3 of Ref. [11]. Thus our mean field approximation is accurate for both our model [12] and the model of Ref. [9],

TABLE I. Parameters obtained in Ref. [12].

Species	N	λ	μ
H. pylori	732	0.88	7.06
P. falciparum	1310	0.93	7.77
S. cerevisiae	4386	1.18	7.94
C. elegans	2800	1.29	8.19
D. melanogaster	2806	1.53	8.89
Human [21]	1494	0.64	10.6
Human [22]	1705	0.67	10.2

IV. PROTEIN-PROTEIN INTERACTION NETWORK MODEL OF SHI *et al.* [12]

Our principal application is to the the PIN of Ref. [12]. This model is based on the concept of free energy of association. For a given pair of proteins the association free energy (in units of RT) is assumed to deviate from an average value a number contributed by both proteins additively as $g + g'$. This is a unique approximation to first order in g and g'. Thermodynamics and the assumption that the interaction probability is independent of concentration allow us to write

$$p(g,g') = 1/(1 + e^{\mu - g - g'}), \quad (37)$$

which reduces to a step function in the zero-temperature limit, but otherwise provides a smooth function. Increasing the value of μ weakens the strength of interactions, and previous results [12] showed the existence of an evolutionary trend to weaker interactions in more complex organisms. The probability that a protein has a value of g is given by the probability distribution

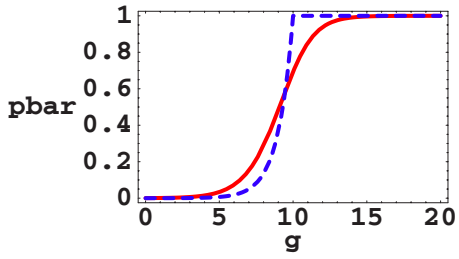

FIG. 1. (Color online) Average connection probability $\lambda=1$, $\mu=10$. Solid (red), result of Eq. (40); dashed (blue) (containing the step function), result of Eq. (25). The approach to unity is smooth for Eq. (40).

$$\rho_\lambda(g) = \frac{\lambda}{e} e^{-\lambda g}, \quad -1 \leqslant \lambda g \leqslant +\infty, \tag{38}$$

where the positive real value of λ governs the fluctuations of g. We previously chose the species-dependent values of λ and μ so as to reproduce measured degree distributions obtained using the yeast two-hybrid method that reports binary results for protein-protein binding under a controlled setting [16]. Those parameters are displayed in Table I. The impact of the parameters λ and μ is explained in Ref. [12] and displayed in Fig. 3 of that reference. Increasing the value of λ causes a more rapid decrease of p_k—the slope of p_k increases in magnitude. Increasing the value of μ decreases the magnitude of p_k without altering the slope much for values of k greater than about 10. The ability to vary both the slope and magnitude of p_k gives this model flexibility that allows us to describe the available degree distributions for different species.

We obtain an analytic form for $\bar{p}(g)$ [Eq. (7)] of this model. Given Eqs. (38) and (37) we find an analytic result:

$$\bar{p}(g,\lambda) = {}_2F_1(1,\lambda;\lambda+1;-\exp(\mu-g)), \tag{39}$$

where ${}_2F_1$ is the confluent hypergeometric function. The special case $\lambda=1$ yields a closed form expression

$$\bar{p}_1(g) = e^{g-\mu}\ln(1+e^{\mu-g}). \tag{40}$$

A smooth average connection probability is obtained in contrast with the result of the sharp cutoff model Eq. (25). This shown in Fig. 1.

It is useful to define the variable

$$\xi \equiv \exp(\mu-g) > 0, \tag{41}$$

and note that an integral representation [14]

$${}_2F_1(n,\lambda;\lambda+1;-\xi) = \lambda\int_0^1 dt\, t^{\lambda-1}(1+\xi t)^{-n} \tag{42}$$

is convenient for numerical evaluations.

Knowledge of the propagator Eq. (9) allows us to compute the clustering coefficients of diverse species. The resulting degree distributions of p_k (shown for the sake of completeness) and the newly computed clustering coefficients $\bar{c}(k)$ for the yeast *S. cerevisiae* [17], the worm *C. elegans*

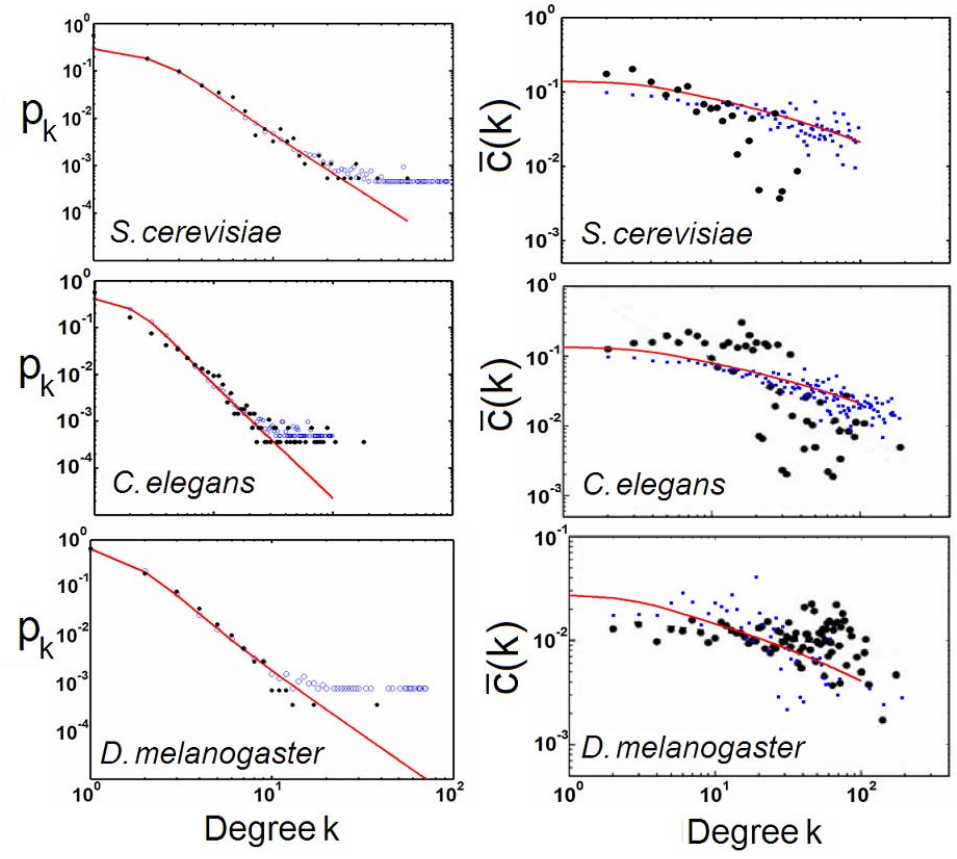

FIG. 2. (Color online) Degree distributions p_k and clustering coefficients $\bar{C}(k)$ of diverse species. Degree distributions p_k: The solid (red) curves are derived from the LO theory; the black dots are the results of experimental data as referenced in the text; the small (blue) circles are the results of a numerical simulation using the procedure of [12]. Clustering coefficients $\bar{C}(k)$: The solid (red) curves are derived from the LO theory; the small (blue) dots are the results of a numerical simulation using the procedure of [12]; the heavy (black) dots represent the measured data.

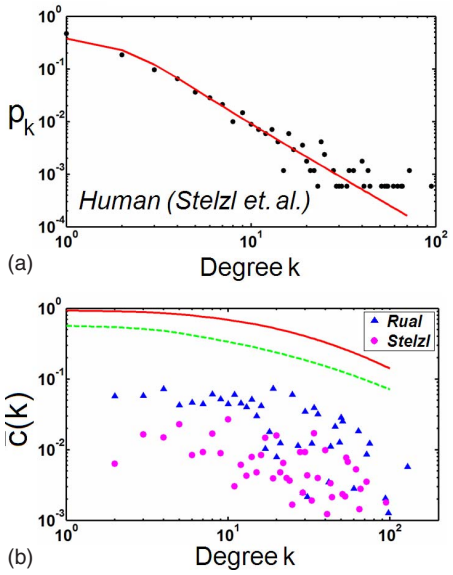

FIG. 3. (Color online) Human degree distribution p_k; the solid (red) curve is obtained using both set A, $\lambda=0.67$, $\mu=10.6$, and set B, $\lambda=0.94$, $\mu=8.3$. The black dots represent the experimental data. The data set is that of [21], but nearly identical data are obtained from [22]. Human cluster coefficient $\bar{c}(k)$: The solid (red) curve is computed using set A, $\lambda=0.67$, $\mu=10.6$; and the dashed (green) using set B, $\lambda=0.94$, $\mu=8.3$. Measured human clustering coefficients are from [21] triangles (blue) and [22] heavy dots (pink).

[18], and the fruit fly *D. melanogaster* [19] are shown in Fig. 2. The parameters λ and μ are those of [12], so the calculations of the clustering coefficients represent an independent major prediction of our model. Results of numerical simulations and our analytic procedure are presented. The excellent agreement between the two methods verifies the LO approximation. More importantly, the agreement between our calculations and the measured clustering coefficients is generally very good, so our model survives a very significant test. This bolsters the notion that the properties of a PIN are determined by a distribution of free energy. The clustering coefficient for yeast drops rapidly for large values of k (where statistics are poor), a feature not contained in our model.

It is worthwhile to compare our model with that of [13]. That work chooses a Gaussian form of $\rho(g)$, based on hydrophobicity, and a step function form of $p(g,g')$, and is applied only to yeast. We found [12] that p_k of [13] is scale-free only for a narrow range of parameters, and we could not reproduce the data for diverse species using that model.

The human interactome is of special interest. Figure 3(a) shows the human degree distributions computed with two sets of parameters, one from Ref. [12] (Table 1) and the other using values of $\lambda=0.94$, $\mu=8.27$ shown in the caption. The degree distributions are essentially identical, so only one curve can be shown. Each is approximately of a power law form and each describes the measured degree distribution very well [20]. Calculations of degree correlations allows one to distinguish the two parameter sets. Figure 3(b) shows that the cluster coefficients differ by a factor of 2. We find that $\bar{c}(k)$ decreases substantially as λ increases. The increase in λ reduces the allowed spread in the value of g and reduces the value of the integrand of Eq. (14). It is interesting to note that the two existing measurements of the human $\bar{c}(k)$ differ by a factor of about an order of magnitude, with the measurements of Ref. [22] giving much smaller values than those of [21]. The results of [21] are closer to our computed $\bar{c}(k)$ results for $\lambda=0.94$, $\mu=8.3$. In contrast with the results for other species, our $\bar{c}(k)$ lie significantly above the data. However, the two data sets disagree substantially (by a factor of as much as 100 for certain values of k), and both show a clustering coefficient that is generally significantly smaller than those of the other species. Several possibilities may account for the discrepancies between these two measurements of $\bar{c}(k)$ in humans and also for the differences between our model predictions and the experimental results. (i) The human studies sample a limited subset of links of the complete network and this could bias the results. (ii) The human protein subsets used in the two studies differ. (iii) The human interactome is truly less connected than that of other species. This demonstrates the importance of measuring degree correlations to determine the underlying properties of the network. The current model and these considerations suggest the need for better design of future PIN studies that will include not only other species, but also comparisons between the PINs of different organs of a given species. Furthermore, comparisons between normal and malignant tissues could also be very fruitful.

V. SUMMARY AND DISCUSSION

In summary, this work provides a method to obtain the properties of hidden variable network models. The use of the approximation Eq. (7), used to obtain the propagator Eq. (9), provides an excellent numerical approximation to exact results for the models considered here. If necessary, the method can be systematically improved through the calculation of higher-order corrections. Our principal example is the PIN of Ref. [12]. Not only does the use of Eq. (9) provide an accurate numerical result, but the model correctly predicts the clustering coefficients of most species. For the human interactome, two different parameter sets yield nearly the same degree distribution but very different clustering coefficients, showing the importance of measuring degree correlations to determine the underlying nature of the network.

ACKNOWLEDGMENTS

This work was supported in part by National Institutes of Health Grants No. GM45134 and No. DK45978 (to K.B.). We thank the authors of Refs. [21,22] for providing tables of their data.

[1] S. H. Strogatz, Nature (London) **410**, 268 (2001).
[2] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[3] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[4] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).
[5] A. Vázquez, R. Pastor-Satorras, and A. Vespignani, Phys. Rev. E **65**, 066130 (2002).
[6] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701, (2002).
[7] M. E. J. Newman, Phys. Rev. E **67**, 026126 (2003).
[8] B. Alberts *et al.*, *The Cell* (Garland Science, New York, 2002).
[9] G. Caldarelli, A. Capocci, P. DeLosRios, and M. A. Muñoz, Phys. Rev. Lett. **89**, 258702 (2002).
[10] B. Söderberg, Phys. Rev. E **66**, 066121 (2002).
[11] M. Boguñá and R. Pastor-Satorras, Phys. Rev. E **68**, 036112 (2003).
[12] Yi Y. Shi, G. A. Miller, H. Qian, and K. Bomsztyk, Proc. Natl. Acad. Sci. U.S.A. **103**, 11527 (2006).
[13] E. J. Deeds, O. Ashenberg, and E. I. Shakhonovich, Proc. Natl. Acad. Sci. U.S.A. **103**, 311 (2006).
[14] *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1970).
[15] S. Abe and S. Thurner, Phys. Rev. E **72**, 036102 (2005); Int. J. Mod. Phys. C **17**, 1303 (2006).
[16] S. Fields and S. Song, Nature (London) **340**, 245 (1989).
[17] http://www.nd.edu/networks/resources/protein/bo.dat.qz
[18] S. Li *et al.*, Science **303**, 540 (2004).
[19] L. Giot *et al.*, Science **302**, 1727 (2003).
[20] The value of p_0 is a testable result of our model, even though experimentalists do not measure this quantity. The predicted number of proteins with no interactions is p_0N, where the value of N is given in Table I. The experimentalists conventionally normalize their distributions as $\Sigma_{k=1}^{\infty}p_k=1$, so we multiply our computed p_k by a factor of $1/(1-p_0)$ so that the computed sum $\Sigma_{k=1}^{\infty}p_k$ is unity.
[21] J. F. Rual *et al.*, Nature (London) **437**, 1173 (2005).
[22] U. Stelzl *et al.*, Cell **122**, 957 (2005).