Cycles and clustering in bipartite networks

Pedro G. Lind,[1,2] Marta C. González,[1] and Hans J. Herrmann[1,3]
[1]*Institute for Computational Physics, Universität Stuttgart, Pfaffenwaldring 27, D-70569 Stuttgart, Germany*
[2]*Centro de Física Teórica e Computacional, Av. Prof. Gama Pinto 2, 1649-003 Lisbon, Portugal*
[3]*Departamento de Física, Universidade Federal do Ceará, 60451-970 Fortaleza, Brazil*
(Received 11 April 2005; published 22 November 2005)

We investigate the clustering coefficient in bipartite networks where cycles of size three are absent and therefore the standard definition of clustering coefficient cannot be used. Instead, we use another coefficient given by the fraction of cycles with size four, showing that both coefficients yield the same clustering properties. The new coefficient is computed for two networks of sexual contacts, one bipartite and another where no distinction between the nodes is made (monopartite). In both cases the clustering coefficient is similar. Furthermore, combining both clustering coefficients we deduce an expression for estimating cycles of larger size, which improves previous estimations and is suitable for either monopartite and multipartite networks, and discuss the applicability of such analytical estimations.

DOI: 10.1103/PhysRevE.72.056127 PACS number(s): 89.75.Fb, 89.75.Hc, 89.65.−s

I. INTRODUCTION

One important statistical tool to access the structure of complex networks arising in many systems [1,2] is the clustering coefficient, introduced by Watts and Strogatz [3] to measure "the cliquishness of a typical neighborhood" in the network and given by the average fraction of neighbors which are interconnected with each other. This quantity has been used for instance to characterize small-world networks [3], to understand synchronization in scale-free networks of oscillators [4] and to characterize chemical reactions [5] and networks of social relationships [6,7]. One pair of linked neighbors corresponds to a triangle, i.e., a cycle of three connections.

While triangles may be abundant in networks of identical nodes they cannot be formed in bipartite networks [6–8], where two types of nodes exist and connections link only nodes of different types. Thus, the standard clustering coefficient is always zero. However, different bipartite networks have in general different cliquishnesses and clustering abilities [7], stemming for another coefficient which uncovers these topological differences among bipartite networks. Bipartite networks arise naturally in, e.g., social networks [8,9] where the relationships (connections) depend on the gender of each person (node), and there are situations, such as in sexual contact networks [10], where one is interested in comparing clustering properties between monopartite (identical nodes) and bipartite (two types of nodes) compositions.

In this paper, we study the cliquishness of either monopartite and bipartite networks, using both the standard clustering coefficient and an additional coefficient which gives the fraction of squares, i.e. cycles composed by four connections. As shown below, such a coefficient retains the fundamental properties usually ascribed to the standard clustering coefficient in regular, small-world and scale-free networks. As a specific application, two examples of networks of sexual contacts will be studied and compared, one being monopartite and another bipartite.

Furthermore, we will show that one can take triangles and squares as the basic units of larger cycles in any network, monopartite or multipartite. The frequency and distribution of larger cycles in networks have revealed its importance in recent research for instance to characterize local ordering in complex networks from which one is able to give insight on their hierarchical structure [11], to determine equilibrium properties of specific network models [12], to estimate the ergodicity of scale-free networks [13], to detect phase transitions in the topology of bosonic networks [14], and to help characterize the Internet structure [15]. Since the computation of all cycles in arbitrarily large networks is unfeasible, one uses approximate numerical algorithms [13,16,17] or statistical estimations [18,19]. Here, we go a step further and deduce an expression to estimate the number of cycles of larger size, using both clustering coefficients, which not only improves recent estimations [19] done for monopartite networks, but at the same time can be applied to bipartite networks and multipartite networks of higher order.

We start in Sec. II by introducing the expression which characterizes the cliquishness of bipartite networks, comparing it with the usual clustering coefficient. In Sec. III we use both coefficients to estimate cycles of larger size and show how it is applied to bipartite networks, while in Sec. IV we apply both coefficients to real networks of sexual contacts. Conclusions are given in Sec. V.

II. TWO COMPLEMENTARY CLUSTERING COEFFICIENTS

The standard definition of clustering coefficient C_3 is the fraction between the number of triangles observed in one network out from the total number of possible triangles which may appear. For a node i with a number k_i of neighbors the total number of possible triangles is just the number of pairs of neighbors given by $k_i(k_i-1)/2$. Thus, the clustering coefficient $C_3(i)$ for node i is

$$C_3(i) = \frac{2t_i}{k_i(k_i-1)}, \tag{1}$$

where t_i is the number of triangles observed, i.e., the number of connections among the k_i neighbors. As in other studies,

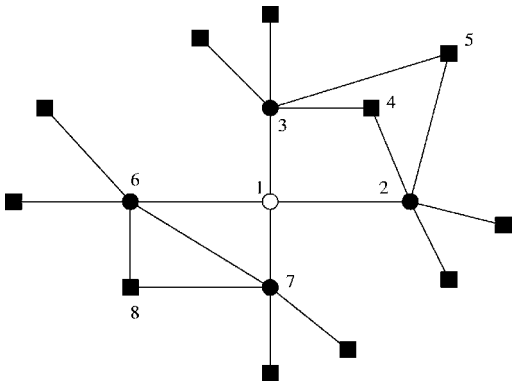

FIG. 1. Illustration of the neighborhood of a central node (○) compose by its first neighbors (●) and its second neighbors (■), i.e., the neighbors of its neighbors. First and second neighbors are used to compute the complementary clustering coefficient C_4 (see text).

here and throughout the paper multiple connections between the same pair of nodes are not allowed.

Similarly to $C_3(i)$, a cluster coefficient $C_4(i)$ with squares is the quotient between the number of squares and the total number of possible squares. For a given node i, the number of observed squares is given by the number of common neighbors among its neighbors, while the total number of possible squares is given by the sum over each pair of neighbors of the product between their degrees, after subtracting the common node i and an additional one if they are connected. Explicitly, for a given node i the contribution of a pair of neighbors, say m and n, to $C_4(i)$ reads

$$C_{4,mn}(i) = \frac{q_{imn}}{(k_m - \eta_{imn})(k_n - \eta_{imn}) + q_{imn}}, \tag{2}$$

where q_{imn} is the number of common neighbors between m and n (not counting i) and $\eta_{imn}=1+q_{imn}+\theta_{mn}$ with $\theta_{mn}=1$ if neighbors m and n are connected with each other and 0 otherwise. The numerator in Eq. (2) gives the number of squares containing nodes i,m, and n, while the denominator counts the total possible number of squares containing these three nodes.

To illustrate the definition giving in Eq. (2), we show in Fig. 1 a simple sketch of a node (○) neighborhood composed by its first and second neighbors (● and ■, respectively). Considering the neighbors 2 and 3, one has $q_{123}=2$ squares containing nodes 1, 2, and 3 and there are $k_2=5$ and $k_3=5$ neighbors of nodes 2 and 3, respectively. Since nodes 2 and 3 are not connected with each other $\theta_{23}=0$, yielding $\eta_{123}=3$ and a denominator in Eq. (2) which equals six possible squares, two squares which are observed and other four squares corresponding to the possible combinations of all pairs of noncommon neighbors. For neighbors 6 and 7 a similar calculation can be done, this time with $\theta_{67}=1$ since the neighbors are connected with each other. The clustering coefficient $C_4(i)$ is easily obtained from Eq. (2) just by summing the numerator and denominator separately over the neighbors of i.

While $C_3(i)$ gives the probability that two neighbors of node i are connected with each other, $C_4(i)$ is the probability that two neighbors of node i share a common neighbor (different from i). Averaging $C_3(i)$ and $C_4(i)$ over the nodes yields two complementary clustering coefficients, $\langle C_3\rangle$ and $\langle C_4\rangle$, characterizing the contribution for the network cliquishness of the first and second neighbors, respectively. For simplicity we write henceforth C_3 and C_4 for the averages of $C_3(i)$ and $C_4(i)$, respectively.

An important point to stress concerns the denominators in the definitions of both clustering coefficients. The possible number of triangles in Eq. (1) does not take into account the topology of the neighborhood, in particular the number of second neighbors. Instead, the standard way [3] to compute C_3, given by Eq. (1), is to assume that all possible triangles are observed when the neighbors are fully interconnected. Consequently, possible degree-correlation biases may appear. The same occurs for the definition of C_4. Recently [20] another expression for C_3 was proposed with the aim to filter out these degree-correlation biases by taking into account the minimum number of neighbors of each pair of nodes considered. A similar approach could be done for C_4, substituting the denominator in Eq. (2) by a suitable function of the minimum number of neighbors of n and m. However, here we consider C_4 as defined above, since it is our purpose to establish a parallel between C_4 and the the standard definition of C_3, which itself does not take into account either the correlation removal proposed in Ref. [20].

Figure 2 shows both clustering coefficients C_3 and C_4 in several topologies. In all cases C_3 and C_4 are plotted as dashed and solid lines, respectively, and are averages over samples of 100 realizations. As an example of regular networks, we use networks with boundary conditions where each node has n neighbors symmetrically disposed, i.e., when placed in a chain, each node has an even number of neighbors, half of them placed on one side and the other half placed on the other side. In particular, for $n=2$ one obtains a chain of nodes connected to its nearest neighbors. For these regular networks, Fig. 2(a) shows the dependence of the clustering coefficients on the fraction n/N of neighbors, with $N=10^3$ the total number of nodes. As one sees $C_4<C_3$ and for either small or large fractions of neighbors both coefficients increase abruptly with n. In the middle region C_3 is almost constant, while C_4 decreases slightly. Our simulations have shown that in regular networks the coefficients depend only on n/N, i.e., for any size of the regular network, similar plots are obtained.

Figure 2(b) shows the coefficients for small-world networks with $N=10^3$ nodes, constructed from a regular network with $n=4$ neighbors symmetrically disposed. The coefficients are computed as functions of the probability p to rewire short-range connections into long-range connections and they are normalized as usual [3] to the clustering coefficients $C^0_{3,4}$ of the underlying regular network. As one sees, C_4 yields approximately the same spectrum as the standard clustering coefficient C_3 being therefore able to define the same range of p for which small-world effects are observed. While here the small-world networks were constructed with rewiring of short-range connections into long-range ones, the

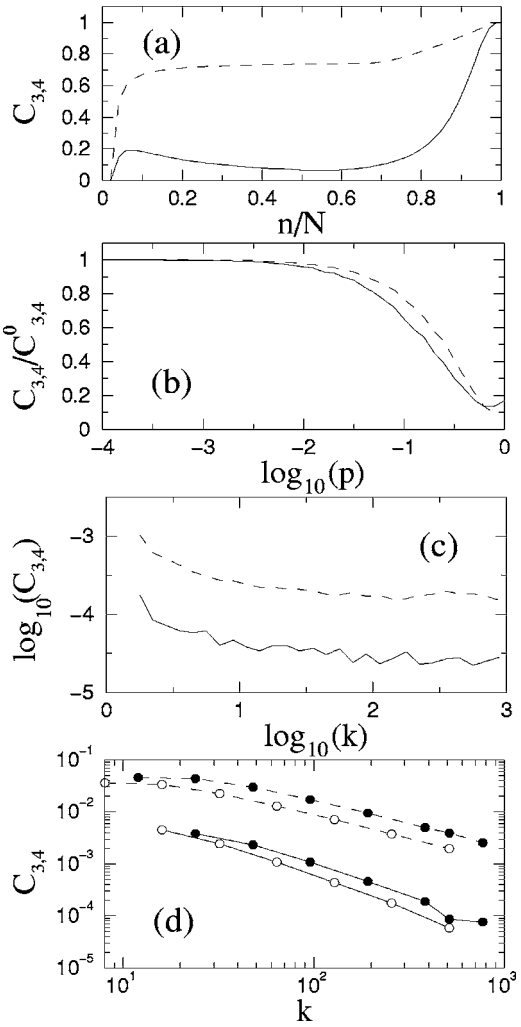

FIG. 2. Comparisons between the standard clustering coefficient C_3 in Eq. (1) (dashed line) and the clustering coefficient C_4 in Eq. (2) (solid line) for different network topologies, (a) in one regular network with n neighbors symmetrically placed ($N=10^3$), (b) in small-world networks where long-range connections occur with probability p ($N=10^3$ and $n=4$), and (c) in random scale-free networks where the distribution of the clustering coefficients is plotted as a function of the number k of neighbors ($N=10^5$ and $m=2$). In all cases samples of 10^2 networks were used. The distributions $C_3(k)$ and $C_4(k)$ are also plotted for (d) Apollonian networks [21] with $N=9844$ nodes (●) and pseudofractal networks [23] with $N=9843$ nodes (○).

same features are observed when using the construction procedure introduced in Ref. [24] where instead of rewiring one uses addition of long-range connections.

To construct scale-free networks, we use the standard procedure of Albert and Barabási with growing and preferential attachment proportional to number of neighbors (see, e.g., Ref. [2] for details). For such scale-free networks, which we call henceforth random scale-free networks, we plot in Fig. 2(c) the distribution of both coefficients as functions of the number k of neighbors, using networks with $N=10^5$ nodes and by given initially $m=2$ connections to each node. Here, one observes that $C_4(k)$ is almost constant as k increases, reproducing the same known feature as the standard $C_3(k)$ apart a scaling factor, $C_4(k)/C_3(k)$ is approximately constant for any k. In Fig. 2(d) we plot the clustering distributions for two different deterministic scale-free networks recently studied, namely Apollonian networks [21], represented with solid circles ●, and pseudofractal networks [23], represented with circles ○. In both cases, the same power-law behavior already known for $C_3(k)\sim k^{-\alpha}$ in these hierarchical networks is also observed for the coefficient $C_4(k)$ with the same value of the exponent α.

All networks in Fig. 2 are monopartite, i.e., no distinction between nodes is made, to aim the straightforward comparison between both clustering coefficients, C_3 and C_4. Of course, in the case that bipartite counterparts are considered, the standard clustering coefficient C_3 vanishes, and only C_4 is suitable to measure the clustering between nodes.

In short, the results shown in Fig. 2 give evidence that C_4 is also a suitable coefficient to characterize the topological features in several complex networks commonly done with the standard clustering coefficient C_3. Furthermore, since C_4 counts squares instead of triangles, it is particularly suited for bipartite networks. Next, we will use this coefficient to compare different models for networks of sexual contacts, where both monopartite and bipartite networks arise naturally.

III. ESTIMATING THE NUMBER OF LARGE CYCLES WITH SQUARES AND TRIANGLES

Recent studies have attracted attention to the cycle structure of complex networks, since the presence of cycles has important effects, for example, on information propagation through the network [25] and on epidemic spreading behavior [26]. In order to avoid numerical algorithms for counting the number of cycles with arbitrary size which implies long computation times, an estimate of the fraction of cycles with different sizes was proposed [19], using the degree distribution $P(k)$ and the standard cluster coefficient distribution $C_3(k)$. However, this estimation yields a lower bound for the total number of cycles and cannot be applied to bipartite networks, as shown below. The aim of this section is twofold. First, to show that by using both C_3 and C_4 one is able to improve that estimation, being suitable at the same time to either monopartite and bipartite networks. Second, to explicitly show some limitations of the estimations below and discuss their applicability.

The estimation in Ref. [19] considers the set of cycles with a central node, i.e., cycles with one node connected to all other nodes composing the cycle. Figure 3(a) illustrates one of such cycles, where the central node and each pair of its consecutive neighbors forms a triangle, in a total amount of four adjacent triangles. In such set of cycles, to estimate the number of cycles with size s one looks to the central node of each cycle which has a number, say k, of neighbors. The number of different possible cycles to occur is

$$n_0(s,k)=\binom{k}{s-1}\frac{(s-1)!}{2},$$

since one has

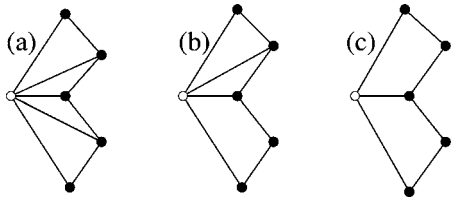

FIG. 3. Illustrative examples of cycles (size $s=6$) where the most connected node (○) is connected to (a) all the other nodes composing the cycle, forming four adjacent triangles. In (b) the most connected node is connected to all other nodes except one, forming two triangles and one subcycle of size $s=4$, while in (c) the same cycle $s=6$ encloses two subcycles of size $s=4$ and no triangles (see text).

$$\binom{k}{s-1}$$

different groups of s nodes and in each one of these groups there are $(s-1)!/2$ different ways in ordering the s nodes into a cycle. The fraction of $n_0(s,k)$ of cycles which is expected to occur is $p_0(s,k)=C_3(k)^{s-2}$, since the probability of having one edge between two consecutive neighbors is $C_3(k)$ and one must have $s-2$ edges between the $s-1$ neighbors. Therefore, the number of cycles of size s is estimated as

$$N_s = N g_s \sum_{k=s-1}^{k_{\max}} P(k) n_0(s,k) p_0(s,k), \tag{3}$$

where $P(k)$ is the degree distribution and g_s is a factor which takes into account the number of repeated cycles. This geometrical factor can be computed for each particular case of s cycles but the estimations can be carried out without the explicit computation of the factor [19].

The estimation in Eq. (3) is a lower bound for the total number of cycles since it considers only cycles with a central node. For instance, in Fig. 3(b) while cycles of size $s=4$ can be estimated with Eq. (3)), the cycle $s=6$ cannot since it has no central node, and in Fig. 3(c) the above equation cannot estimate any cycle of any size. In fact, Fig. 3(c) illustrates the type of cycles appearing in bipartite networks, where no triangles are observed. For such cycles $C_3(k)=0$ and therefore all terms in Eq. (3) vanish yielding a wrong estimation of the number of cycles.

To take into account cycles without central nodes [Figs. 3(b) and 3(c)], one must consider the clustering coefficient $C_4(k)$ defined in Eq. (2). One first considers the set of cycles of size s with one node (○) connected to all the others except one as illustrated in Fig. 3(b). In this case since there are $s-2$ nodes connected to node ○ one has

$$n_1(sk) = \binom{k}{s-2}(s-2)!/2$$

different possible cycles of size s with k the number of neighbors of node ○. The fraction of the $n_1(sk)$ cycles which is expected to be observed is given by $p_1(sk)=C_3(k)^{s-4}C_4(k)[1-C_3(k)]$ since the probability of having $s-4$ connections among the $s-2$ connected nodes is $C_3(k)^{s-4}$ the probability that a pair of neighbors of node ○ share a common neighbor (different from node ○) is $C_4(k)$ and the probability that these same pairs of neighbors are not connected is $[1-C_3(k)]$. Writing an equation similar to Eq. (3) where instead of $n_0(sk)$ and $p_0(sk)$ one has $n_1(sk)$ and $p_1(sk)$ respectively and the sum starts at $s-2$ instead of $s-1$ one has an additional number N_s' of estimated cycles which are not considered in estimation (3). Notice that, since for N_s' one considers at least one subcycle of size $s=4$, this additional estimation contributes only for the estimation of cycles with size $s\geq 4$. We call henceforth subcycle, a cycle which is enclosed in a larger cycle and which do not enclose in itself any shorter cycle.

Still, the new estimation N_s+N_s' is not suitable for bipartite networks, since it yields nonzero estimation only for $s=4$. To improve the estimation further one must consider not only cycles composed by one single subcycle of size $s=4$, as done in the preceding paragraph, but also cycles with any number of subcycles of size $s=4$. Figure 3(c) illustrates a cycle of size $s=6$ composed by two subcycles of size 4. In general, following the same approach as previously, for cycles composed by q subcycles of size 4 one finds

$$n_q(s,k) = \frac{(s-q-1)!}{2}\binom{k}{s-q-1}$$

possible cycles of size s looking from a node with k neighbors and a fraction $p_q(s,k)=C_3(k)^{s-2q-2}C_4(k)^q[1-C_3(k)]^q$ of them which are expected to be observed. For $q=0$ one considers cycles as the one illustrated in Fig. 3(a), while for $q=1$ and $q=2$ one considers the set of cycles with one and two subcycles with size 4, as illustrated in Figs. 3(b) and 3(c), respectively. Summing up over k and q yields our final expression

$$N_s = N g_s \sum_{q=0}^{[s/2]-1} \sum_{k=s-q-1}^{k_{\max}} P(k) n_q(s,k) p_q(s,k), \tag{4}$$

where $[x]$ denotes the integer part of x. In particular, the first term ($q=0$) is the sum in Eq. (3). The upper limit $[s/2]-1$ of the first sum results from the fact that the exponent of $C_3(k)$ in $p_q(s,k)$ must be non-negative, $s-2q-2\geq 0$. The estimation in Eq. (4) not only improves the estimated number computed from Eq. (3), but also enables the estimation of cycles up to a larger maximal size. In fact, since in the binomial coefficient

$$\binom{k}{s-1}$$

of Eq. (3) one must have $s-1\leq k\leq k_{\max}$, one only estimates cycles of size up to $k_{\max}+1$, while in Eq. (4) the maximal size is $2k_{\max}$, as can be concluded using both conditions $s-2q-2\geq 0$ and $s-q-1\leq k_{\max}$.

Figure 4 compares two cases treated in Ref. [19], both with a degree distribution $P(k)=P_0k^{-\gamma}$ and coefficient distributions $C_3(k)=C_3^{(0)}k^{-\alpha}$, using one value of $\alpha<1$ [Fig. 4(a)] and another one $\alpha>1$ [Fig. 4(b)]. Dashed lines indicate the estimation done with Eq. (3), while solid lines indicate the estimation done with Eq. (4). In both cases, the latter estima-

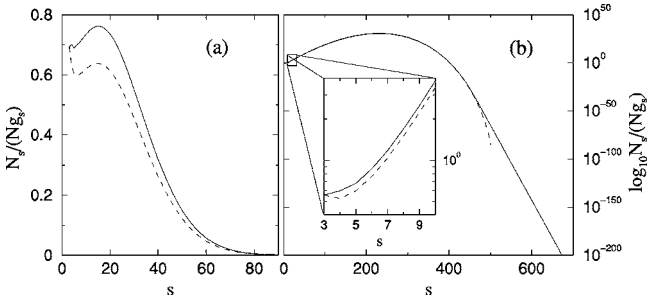

FIG. 4. Estimating the number of cycles using Eq. (3), dashed lines, and Eq. (4), solid lines. Here we impose a degree distribution $P(k)=P_0k^{-\gamma}$ with $P_0=0.737$ and $\gamma=2.5$, and coefficient distributions $C_{3,4}(k)=C_{3,4}^{(0)}k^{-\alpha}$ with (a) $C_3^{(0)}=2, C_4^{(0)}=0.33, \alpha=0.9$ and (b) $C_3^{(0)}=1, C_4^{(0)}=0.17, \alpha=1.1$. In all cases $k_{\max}=500$.

tion is larger. For $\alpha<1$ the difference between both estimations decreases with the size s of the cycle. For $\alpha>1$ the difference between the estimations increases with s beyond a size $s^* \lesssim k_{\max}$. Clearly, from Fig. 4(b) one sees that $k_{\max}+1$ is the larger cycle size for which Eq. (3) can give an estimation, while for Eq. (4) the estimation proceeds up to $2k_{\max}$ (partially shown). In both cases, the typical size for which N_s attains a maximum is numerically the same for both estimations, as expected. Moreover, for $\alpha>1$ [Fig. 4(b)], beyond a size of the order of $k_{\max}, N_s/(Ng_s)$ in Eq. (4) decreases exponentially with s, and not as a cutoff as observed for Eq. (3). In fact, the deviation of Eq. (3) from the exponential tail, is due to the fact that for very large cycle sizes ($s\sim k_{\max}$) Eq. (3) can only consider very few terms in its sum.

Another advantage of the estimation in Eq. (4) is that it estimates cycles in bipartite networks. For bipartite network there are no connections between the neighbors, i.e., all subgraphs are similar to the one illustrated in Fig. 3(c). Therefore all terms in Eq. (4) vanish except those for which the exponent of $C_3(k)$ is zero, i.e., for $s=2(q+1)$. Consequently, since q is an integer, Eq. (4) shows clearly that in bipartite networks there are only cycles of even size, as already known [8]. Moreover, substituting $q=(s-2)/2$ in Eq. (4) yields a simple expression for the number of cycles in bipartite networks, namely

$$N_s^{\text{Bipart}} = Ng_s \sum_{k=s/2}^{k_{\max}} P(k) \frac{(s/2)!}{2} \binom{k}{s/2} C_4(k)^{s/2-1}. \qquad (5)$$

A simple example to illustrate the validity of Eq. (4) is the fully connected network, where each node is connected to everyone else. In this case the number of cycles with size s is given by

$$N_s = \binom{N}{s} \frac{(s-1)!}{2}$$

The factor $(s-1)!$ counts for the arrangements between $s-1$ nodes in each combination of s nodes, while the division by two is due to the undirected links. To compute N_s from Eq. (4) one has for the particular case of fully connected network, $P(k)=C_3(k)=C_4(k)=\delta_{k-N+1}, k_{\max}=N-1$, and $g_s=1/s$. Consequently the only nonzero term in the first sum is the one for $q=0$, while the nonzero term in the second sum is the one for $k=N-1$, yielding the same result as above.

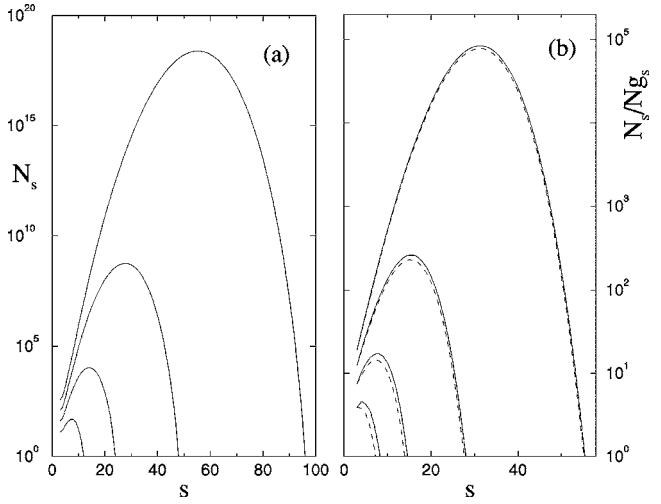

FIG. 5. (a) The exact number of cycles as a function of the size for the pseudofractal network [23] compared with (b) N_s/Ng_s of the analytical estimations in Eqs. (3), dashed lines, and (4), solid lines. From small to large curves one has pseudofractal networks with $m=2, 3, 4, 5$ generations (see text).

Both the estimation in Eq. (3)) and the one in Eq. (4) are particularly suited for networks or subnetworks where nodes are highly connected with each other, since in those situations there is a very large number of centrally connected cycles as the ones illustrated in Fig. 3. Highly connected subnetworks appear, for instance, in social networks which are composed by communities [22]. In Ref. [19], for instance, the estimation of small cycles from Eq. (3) is compared with the true values computed for several empirical networks, namely the Internet, the coauthorship web and semantic networks. While for $s=3$ and 4 the estimation is clearly good, for $s=5$ there is a clear underestimation, due to the appearance of no centrally connected cycles. Of course one expects that, similarly to what is observed in Fig. 4, the estimation in Eq. (4) improves the one used in Ref. [19] for such situations. However, one should stress that the drawback of such estimations for larger cycles both estimations get worse.

Next we illustrate this point using a particular network, so-called pseudofractal network, introduced by Dorogovtsev and co-workers [23]. This network is scale free and is constructed starting with three nodes connected with each other (generation $m=0$), and iteratively adding new generations of nodes such that in generation $m+1$ one new node is added to each previous edge and it is connected to the two nodes joined by that edge. For this network, the exact number of cycles with size s can be written iteratively [27] as

$$N_s^{(m+1)} = \sum_{l=3}^{s} \binom{l}{s-l} N_s^{(m)}, \qquad (6)$$

for $s\geqslant 4$ and $N_3^{(m+1)}=N_3^{(m)}+3^m$.

Figure 5 shows the real number of cycles of the pseudof-

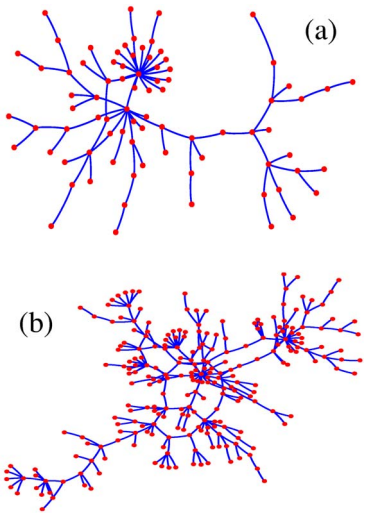

FIG. 6. (Color online) Sketch of two real sexual contact networks having (a) only heterosexual contacts ($N=82$ nodes and $L=84$ connections) and (b) homosexual contacts ($N=250$ nodes and $L=266$ connections). While in the homosexual network triangles and squares appear, in the heterosexual network triangles are absent (see Table I).

ractal network [Fig. 5(a)] with the quantity $N_s/(Ng_s)$ [Fig. 5(b)] for the pseudofractal network. In Fig. 5(b) solid lines indicate our estimation, while dashed lines indicate the previous estimation in Ref. [19]. In both cases, the underestimation is very significant when compared with the exact number from Eq. (6). Nevertheless, even in this case, the estimations predict the shape of the cycle distributions. Up to our knowledge, complex networks for which the exact number of cycles may be computed having most nodes highly connected are not known.

It is important to notice that triangles and squares may appear in any multipartite network (except in bipartite ones, where triangles are absent). Therefore, the estimation described and studied in this section can be applied not only to bipartite networks but to any multipartite network of any order. In the next section we will focus on the applicability of the clustering coefficient C_4 in empirical sexual networks (monopartite and bipartite) with the aim to compare simulated results for such networks.

IV. CYCLES AND CLUSTERING IN SEXUAL NETWORKS

In this section we apply both coefficients C_3 and C_4 in Eqs. (1) and (2) to analyze two real networks of sexual contacts. One network is obtained from an empirical data set, composed solely by heterosexual contacts among $N=82$ nodes, extracted at the Cadham Provincial Laboratory and is a 6-month block data [28] between November 1997 and May 1998. The other data set is the largest cluster with $N=250$ nodes in the records of a contact tracing study [29], from 1985 to 1999, for HIV tests in Colorado Springs (USA), where most of the registered contacts were homosexual. Figure 6 sketches these two networks, where one can see that cycles of different sizes appear. While the network with only heterosexual contacts is clearly bipartite, the network with homosexual contacts is monopartite.

For the two networks in Fig. 6, Table I indicates the number T of triangles, the number Q of squares and the coefficients C_3 and C_4. As one sees, although the heterosexual network has less squares than the homosexual network due to its smaller size, C_4 is much larger. Another feature common for both networks is $L/N\sim 1$, i.e., an effective coordination number of $2L/N\sim 2$.

In order to ascertain possible nontrivial features in these empirical networks, we compare the topological measures of them with the ones of a null model having the same degree distribution. The null model is a randomized version of the empirical networks, constructed by rewiring connections randomly selected [31]. Namely, whenever one pair of links is selected, say $i\leftrightarrow j$ and $k\leftrightarrow l$, we substitute this link by two others, one connecting i and k and another connecting j and l. While in the heterosexual network the number of squares and consequently the value of C_4 is overestimated, for the homosexual network the null model yields reasonable results for both clustering coefficients, although there is a large discrepancy in the number of triangles and squares. In order to compare the number of squares without the effect of the number of triangles in the network, we consider also the case of a null model where additionally to the degree distribution, the number T of triangles is also the same. In this case, Table I shows still an underestimation of C_3 and a much larger number Q of squares. Notice that, while the total number of triangles is the same, the standard clustering coefficient $\langle C_3\rangle$ can be nevertheless different, since it is an average over the local clustering coefficient of each node, which depends not only on the number of triangles the node belongs to but also on its degree.

Recently, we introduced [10] a model to simulate the statistical features of these networks of sexual contacts. The model is a sort of granular system with low density composed by N mobile particles representing persons and collisions between them representing their sexual contacts. Initially, all agents have a randomly chosen position and moving direction with the same velocity modulus $|\vec{v}_0|$ and no connections. When for the first time two agents collide, the corresponding collision is taken as the first connection in the network. Through time, more and more collisions occur giving rise to new connections and enabling the network growth, until the number of connected agents attains the required network size, at which the simulation is stopped and the accumulated number of connections is determined. As a particular feature of our model, we choose a collision rule where the velocity of each agent increases with the number of previous contacts. More details concerning this model are given in Ref. [10].

Using the same number of nodes as in the real networks illustrated in Fig. 6 and considering two types of nodes for the heterosexual (bipartite) case, we obtain with the agent model similar results for L,T,Q,C_3, and C_4, as shown in Table I where values represent averages over samples of 100 realizations. As one sees, in general, the agent model yields values much closer to the ones for the empirical networks, than the two null models considered above. Remarkably, for the bipartite case not only the number of connections and the

TABLE I. Clustering coefficients and cycles in two real networks of sexual contacts (top), illustrated in Fig. 6, one where all contacts are heterosexual [Fig. 6(a)] and another with homosexual contacts [Fig. 6(b)]. In each case one indicates the values of the number N of nodes, the number L of connections, the number T of triangles, the number Q of squares and both clustering coefficients C_3 and C_4 in Eqs. (1) and (2), respectively. The values of these quantities are compared with the ones of a null model (see text) with the same degree distribution for two cases, one where the number of triangles is found and another where this restriction is not imposed, and also with networks constructed with the agent model recently introduced [10]. Samples of 100 realizations were used in each case.

	N	L	T	Q	$\langle C_3 \rangle$	$\langle C_4 \rangle$
Heterosexual [Fig. 6(a)]	82	84	0	2	0	0.00486
Homosexual [Fig. 6(b)]	250	266	11	6	0.02980	0.00192
Heterosexual (Null model)	82	84	0	8.47	0	0.0451
Homosexual (Null model)	250	266	6.94	16.2	0.011	0.00373
Heterosexual (Null model, same T)	82	84	0	8.47	0	0.0451
Homosexual (Null model, same T)	250	266	11.0	21.462	0.0145	0.00477
Heterosexual (Agent model)	82	83.63	0	1.45	0	0.01273
Homosexual (Agent model)	250	287.03	8.23	10.52	0.02302	0.01224

number of squares are numerically the same, but also C_4 is of the same order of magnitude. Similar values of the topological quantities are also obtained for the monopartite case, with the exception of C_4.

In Fig. 7 we plot the degree and clustering coefficient distributions for the monopartite network of sexual contacts sketched in Fig. 6(b), while in Fig. 8 we plot the distributions for the bipartite network [Fig. 6(a)]. In both figures bullets indicate the distributions of the empirical data, while solid lines indicate the distributions of the networks obtained with the agent model, imposing the same size as the real network, i.e., stopping the simulation when the number of connected agents equals the size of the corresponding empirical network, and taking averages over a sample of 100 realizations.

The results above concern small empirical networks. To improve the particular study of sexual networks reproduced by our model, larger networks of sexual contacts should be also studied and comparisons with a null model [31] must be carried out to validate the agent model. These points are being further studied and will be presented elsewhere [30]. The main point here is that the results above show already that the complementary clustering coefficient C_4 is suitable for comparing the cliquishness of neighborhoods in either monopartite and bipartite counterparts of the same complex networks, while the standard clustering coefficient is not.

With the agent model one is able to construct larger networks than the empirical ones. In such large networks cycles

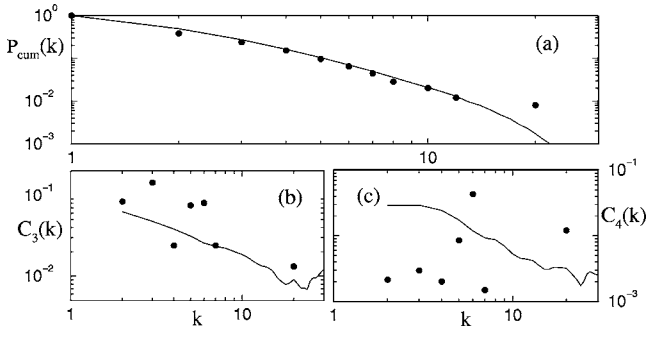

FIG. 7. Comparing topological features between networks obtained from the agent model (solid lines) used to reproduce one real monopartite network of sexual contacts (bullets), (a) cumulative degree distribution $P_{cum}(k)$, (b) standard clustering coefficient $C_3(k)$ in Eq. (1), and (c) clustering coefficient $C_4(k)$ in Eq. (2)). Here $N=250$ and samples of 100 realizations were used.

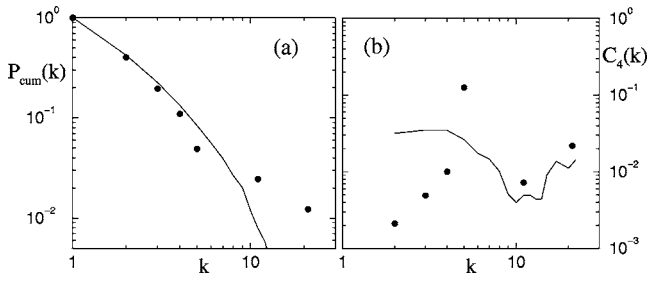

FIG. 8. Distributions for one real bipartite network of sexual contacts (bullets) compared with the one of networks obtained from the agent model (solid lines), (a) cumulative degree distribution $P_{cum}(k)$, (b) clustering coefficient $C_4(k)$ in Eq. (2). Here $C_3(k)=0$ always, $N=82$ and samples of 100 realizations were used.

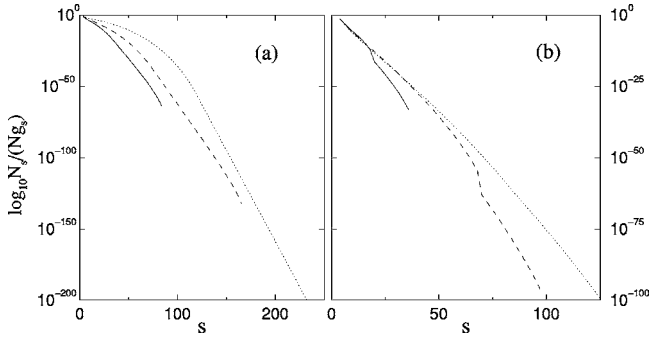

FIG. 9. Estimating the number of cycles for the agent model using Eq. (4) for $N=1000$ (solid lines), $N=5000$ (dashed lines), and $N=10\,000$ (dotted lines) in (a) a monopartite network and in (b) a bipartite network, both obtained with the agent model.

of different size may then appear and one important question is to know the frequency of cycles of any order. Using the agent model for large networks, and computing only their degree distribution and the two clustering coefficients we can then estimate the distribution of cycles in those networks. Figure 9 shows the distribution of the fraction $N_s/(Ng_s)$ of cycles as a function of their size s, for a monopartite network [Fig. 9(a)] and a bipartite network [Fig. 9(b)] composed by $N=1000$, 5000, and 10 000 nodes. Here, while monopartite networks show an exponential tail preceded by a region where the number of cycles is large, bipartite networks are composed by cycles whose number depends exponentially of their size. Furthermore one observes a clear transition for a characteristic size, which seems to scale with the network size.

V. DISCUSSION AND CONCLUSIONS

We introduced a clustering coefficient similar to the standard one, which instead of measuring the fraction of triangles in a network measures the fraction of squares, and showed that with this clustering coefficient it is also possible to characterize topological features in complex networks, usually done with the standard coefficient. We showed explicitly that the range of values of the probability to acquire long-range connections in small-world networks and the typical clustering coefficient distributions of either random scale-free and hierarchical networks are approximately the same. In addition, we showed that this second clustering coefficient enables one to quantify the cliquishness in bipartite networks where triangles are absent. Thus, one should take triangles and squares simultaneously as the two basic cycle units in any network.

An application of both clustering coefficients was proposed, namely to estimate the number of cycles in any network, either monopartite or multipartite. Using a recent estimation which yields a lower bound of the number of cycles in monopartite network up to a size $s<k_{\max}+1$ where $k_{\max}$ is the maximum number of neighbors in the network, we deduce a more general expression which not only improves the previous estimation but is also suitable for bipartite networks and enables one to estimate cycles of size up to $2k_{\max}$. Furthermore, in the particular case of bipartite networks our estimation yields as a natural consequence that only cycles of even size may appear.

To illustrate the applicability of the complementary clustering coefficient in bipartite networks, we studied a concrete example of two sexual networks, one where only heterosexual contacts occur (bipartite network) and another with homosexual contacts (monopartite). The results obtained with the two real networks were found to be similar to the ones obtained with an agent model recently introduced.

All in all, our analytical estimation gives a simple way to extract information concerning the distribution of cycles in multipartite networks, and in particular the clustering coefficient C_4 can be regarded as a suitable measure of neighborhood cliquishnesses in bipartite networks.

ACKNOWLEDGMENTS

One of the authors (M.C.G.) thanks Deutscher Akademischer Austausch Dienst (DAAD), Germany, and one of the authors (P.G.L.) thanks Fundação para a Ciência e a Tecnologia (FCT), Portugal, for financial support.

[1] B. Bollobás and O. M. Riordan, *Handbook of Graphs and Networks: From the Genome to the Internet* (Wiley-VCH, Weinheim, 2003).
[2] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[3] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[4] P. N. McGraw and M. Menzinger, Phys. Rev. E **72**, 015101 (2005).
[5] P. F. Stadler, A. Wagner, and D. A. Fell, Adv. Complex Syst. **4**, 207 (2001).
[6] M. E. J. Newman, Soc. Networks **25**, 83 (2003).
[7] P. Holme, C. R. Edling, and F. Liljeros, Soc. Networks **26**, 155 (2004).
[8] P. Holme, F. Liljeros, C. R. Edling, and B. J. Kim, Phys. Rev. E **68**, 056107 (2003).
[9] R. Guimerà, X. Guardiola, A. Arenas, A. Díaz-Guilera, D. Streib, and L. A. N. Amaral (unpublished).
[10] M. C. González, P. G. Lind, and H. J. Herrmann, physics/0508145 (unpublished).
[11] G. Caldarelli, R. Pastor-Satorras, and A. Vespignani, Eur. Phys. J. B **38**, 183 (2004).
[12] E. Marinari and R. Monasson, J. Stat. Mech.: Theory Exp. (2004), P09004.
[13] H. D. Rozenfeld, J. E. Kirk, E. M. Bollt, and D. ben-Avraham, J. Phys. A **38**, 4589 (2005).
[14] G. Bianconi and A. Capocci, Phys. Rev. Lett. **90**, 078701 (2003).
[15] G. Bianconi, G. Caldarelli, and A. Capocci, Phys. Rev. E **71**,

066116 (2005).
[16] C. P. Herrero, Phys. Rev. E **71**, 016103 (2005).
[17] S.-J. Yang, Phys. Rev. E **71**, 016107 (2005).
[18] G. Bianconi and M. Marsili, cond-mat/0502552 (unpublished).
[19] A. Vázquez, J. G. Oliveira, and A.-L. Barabási, Phys. Rev. E **71**, 025103(R) (2005).
[20] S. N. Soffer and A. Vázquez, Phys. Rev. E **71**, 057101 (2005).
[21] J. S. Andrade, Jr., H. J. Herrmann, R. F. S. Andrade, and L. R. da Silva, Phys. Rev. Lett. **94**, 018702 (2005).
[22] G. Palla, I. Derényi, I. Farkas, and Tamás Vicsek, Nature (London) **435**, 814 (2005).
[23] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Phys. Rev. E **65**, 066122 (2002).
[24] M. E. J. Newman and D. J. Watts, Phys. Rev. E **60**, 7332 (1999).
[25] H.-J. Kim and J. M. Kim,Phys. Rev. E **72**, 036109 (2005).
[26] T. Petermann and P. De Los Rios, Phys. Rev. E **69**, 066116 (2004).
[27] K. Klemm and P. F. Stadler, cond-mat/0506493 (unpublished).
[28] J. L. Wylie and A. Jolly, Sex Transm. Dis. **28**, 14 (2001).
[29] J. J. Potterat, L. Phillips-Plummer, S. Q. Muth, R. B. Rothenberg, D. E. Woodhouse, T. S. Maldonado-Long, H. P. Zimmerman, and J. B. Muth, Sex Transm. Infect. **78**, i159 (2002).
[30] M. C. González, P. G. Lind, and H. J. Herrmann (unpublished).
[31] S. Maslov and K. Sneppen, Science **296**, 910 (2002).