Network bipartivity

Petter Holme*
Department of Physics, Umeå University, 901 87 Umeå, Sweden

Fredrik Liljeros
*Department of Epidemiology, Swedish Institute for Infectious Disease Control, 171 82 Solna, Sweden
and Department of Sociology, Stockholm University, 106 91 Stockholm, Sweden*

Christofer R. Edling
Department of Sociology, Stockholm University, 106 91 Stockholm, Sweden

Beom Jun Kim
Department of Molecular Science and Technology, Ajou University, Suwon 442-749, Korea

(Received 14 February 2003; published 7 November 2003)

Systems with two types of agents with a preference for heterophilous interaction produce networks that are more or less close to bipartite. We propose two measures quantifying the notion of bipartivity. The two measures—one well known and natural, but computationally intractable, and the other computationally less complex, but also less intuitive—are examined on model networks that continuously interpolate between bipartite graphs and graphs with many odd circuits. We find that the bipartivity measures increase as we tune the control parameters of the test networks to intuitively increase the bipartivity, and thus conclude that the measures are quite relevant. We also measure and discuss the values of our bipartivity measures for empirical social networks (constructed from professional collaborations, Internet communities, and field surveys). Here we find, as expected, that networks arising from romantic online interaction have high, and professional collaboration networks have low, bipartivity values. In some other cases, probably due to low average degree of the network, the bipartivity measures cannot distinguish between romantic and friendship oriented interaction.

DOI: 10.1103/PhysRevE.68.056107 PACS number(s): 89.75.Fb, 89.75.Hc, 05.50.+q

I. INTRODUCTION

Any system, natural or man-made, consisting of entities that interact pairwise can be described in terms of a network. Networks in the real life often contain some degree of randomness, and have also some structure arising from the strategies or laws the entities follow to make new contacts. Such networks—that can only be described as having both randomness and structure—are called complex networks and have lately received much attention in the physicist community [1,2]. Among the most important developments in this recent surge of activity in network research is arguably the categorization and quantification of static network structures such as clustering [3], degree distribution [4], assortative mixing coefficient [5], grid coefficient [6], etc. A network with no circuit of odd length is called *bipartite*. Many systems are naturally modeled as bipartite networks: Biochemical networks can be described by vertices representing chemical substances separated by vertices representing chemical reactions [7]. As another example, we have the so called ‘‘two-mode’’ representation of affiliation networks where one kind of vertices represents, e.g., organizations and the other type represents individual actors, and the edges indicate to which organizations an actor belongs. But there are also networks that are not necessarily bipartite, but closer to bipartite than what can be expected from a completely random network. Examples of such networks are those that are formed by two types of agents with a preference for heterophilous interaction (human sexual contacts [8,12] and human romance or partnership networks [9] being two cases). In many cases one knows the type of the individual vertices (the gender of the actors in the examples above) [10], but in other cases such information might be lacking (see the data studied in Ref. [11] for a concrete example). Nevertheless, the ‘‘bipartivity’’—how far away from being bipartite a graph is—is a measurable structure, and therefore, we believe, deserves attention.

Bipartivity measures have some potentially interesting applications: Network-based studies of sexually transmitted diseases [12] is one such area as the transmission rates for homosexual and heterosexual contacts differ [13]. Apart from romantic and sexual networks, there are other areas where a bipartivity measure may prove useful: One can consider a trade network where some agents are more or less pronounced sellers and others are primarily buyers (cf. Ref. [14]); such networks would not have a neutral bipartivity. Another application is for the ‘‘genealogical’’ network of a disease outbreak: Some contagious diseases have a relatively stable duration between when an individual is infected and when he or she becomes infectious. Epidemics of these types of diseases can therefore roughly be divided into different generations of infected individuals [13]. A network consist-

[*] Electronic address: holme@tp.umu.se

ing of possible edges of infections, for an outbreak of this type of disease, should therefore have very few odd-length circuits. The reason is that the infection is only transmitted between succeeding generations, which generates only circuits of even length (in the reflexive closure of the network). When reconstructing the paths this kind of disease has taken in a population, a minimization of the bipartivity measures can be a method for excluding redundant infectious edges. Yet an area of potential applications is the study of food webs [15]—these are networks representing ecosystems where the vertices are species and edges represent predation by one species on another. In the simplest picture a food web consists of different ''trophic'' levels where species in one level predates upon species located in the level below. An undirected representation of such a graph would be bipartite. The reality is more complex (and the graph is not necessarily bipartite), so a bipartivity measure can be a way to estimate just how complex reality is compared to the simple picture.

How can we measure bipartivity? The idea we use in this paper is the following: We suppose that all agents of one type tried their best in forming a connection to an agent of the other type. Then we measure to what extent this assumption fails. We can assign a label $\sigma_v \in \{-1,+1\}$ to each vertex v and check for the maximal fraction of edges between vertices of different sign. This fraction will be equal to or higher than the actual fraction of edges between vertices of different type. But, at least for strong heterophilous preference in the network formation, the difference should be small. For weak heterophilous preference this approach will likely fail to produce a correct classification of the individual vertices. Still, the number of even circuits should be larger than in a network created under the same circumstances but with no heterophilous preference; and this will (as we will see) give a lower value of such a bipartivity measure. So even if we cannot reproduce the correct fraction of vertices of different type, we have a measure that is a monotonous function of the strength of the heterophilous preference. It is convenient (at least for people familiar with statistical mechanics) to phrase a problem like this in terms of the antiferromagnetic Ising model. Our bipartivity measure—the maximal fraction of edges between vertices of different sign—is directly related to the ground state energy of the antiferromagnetic Ising model (the relation is given in Sec. II A 1). Throughout the paper we will often use the terminology of such spin systems, such as the antiferromagnetic Ising model. For example, we talk of an edge between two vertices of the same tag as a ''frustrated'' edge.

The spin system analogy to combinatorial optimization problems such as the one we are facing—to find minimal fraction of frustrated edges—is nothing new. With this approach the fraction of frustrated edges defines a cost function corresponding to the energy of the spin system. The two most studied problems in this area are the p-coloring problem and the graph bisection problem. In the p-coloring problem the question is whether or not the vertices of a graph can be assigned one of p colors in such a way that no edge goes between two vertices of the same color. This problem is solvable in linear time for $p=2$, but NP complete[1] (i.e., in the general case not calculable in polynomial time [16]) for $p>2$. The graph bisection problem (also NP complete) is to partition the vertex set into two sets of equal size such that the number of edges between the two sets is minimized [17–19]. Both these problems can, just as ours, be phrased in terms of spin models with antiferromagnetic interaction. Our minimization problem is a little bit different from the bisection problem in that the two sections can have arbitrary sizes. However, as in the bisection and p-coloring problems, we are also faced with an NP-complete optimization problem. (Our aim—to find the ground state energy of antiferromagnetic Ising model—can be mapped to a min-flow max-cut problem [20] which is NP hard on general networks [21].)

As the spin models of statistical physics are familiar to statistical physicists, it is not surprising that topics like the Ising and XY models on various model networks [22,23] have received much attention in physicists' network literature. The motivation for such studies, as models of real-world systems, is that they can capture some features of opinion formation or similar social processes [24]. The present work can also be described as a study of a spin model on a complex network, but unlike the above mentioned studies, the spin model is used as a tool to measure a static network structure.

II. THE MEASURES

In the following sections we will go through the two bipartivity measures. We state the definitions, dissect the algorithms, and give analytic discussions about the limit properties.

We represent an undirected network by $G=(V,E)$ and a directed network by $G_{\mathrm{dir}}=(V,A)$, where V is the set of vertices, E is a set of edges (or undirected pairs of vertices), and A is a set of arcs (or ordered pairs of vertices). A *path* of length l is a sequence of vertices $v_1,\ldots,v_l$ such that $(v_i,v_{i+1})\in E$ [or $(v_i,v_{i+1})\in A$ for directed graphs]; a *circuit* is a path where the first and last vertex are identical. In an *elementary* path, or circuit, no vertex appears twice (except the first and last in case of circuits). In the present paper we will only talk about elementary paths and circuits—so, for brevity we omit the word ''elementary.'' Throughout the paper, when necessary, we let subscript or superscript ''dir'' denote directed versions of quantities. In many cases the generalization from undirected to directed networks is straightforward; in these cases we will pursue the discussion in the framework of undirected networks.

A. The measure b_1

1. Definition

The first measure we consider is simply the fraction of unfrustrated edges in the ground state of the antiferromag-

[1] NP is the class of problems that a nondeterministic Turing machine accepts in polynomial time. An NP-complete problem is an NP problem that does not belong to an easier class [49].

netic Ising model on the network. In terms of the antiferromagnetic Ising model the quantity can be written as

$$b_1 = 1 - \frac{M_{\mathrm{fr}}}{M} = \frac{1}{2} - \frac{E_0}{2M}, \tag{1}$$

where M_{fr} is the number of frustrated edges in the ground state (the usual cost function in the two-coloring problem). E_0 is the ground state energy

$$E_0 = \min_{\{\sigma_v\}} H, \tag{2}$$

where H is the Hamiltonian of the antiferromagnetic Ising model:

$$H = \sum_{(v,w)\in E} \sigma_v \sigma_w, \tag{3a}$$

$$H_{\mathrm{dir}} = \sum_{(v,w)\in A} \sigma_v \sigma_w. \tag{3b}$$

The directed quantity is obtained by substituting H by H_{dir} in Eqs. (1) and (2), and edges by arcs in the above discussion. The topology of the energy landscape is determined by the underlying network, and can in general be very complex [25].

2. Limit properties

The b_1 measure takes values in the interval $(1/2,1]$. The upper bound is attained for bipartite graphs. It is easy to see that b_1 cannot be lower than 1/2: Consider a ground state configuration for which the opposite is true. Then there must be at least one vertex with more than half of its edges frustrated. Flipping this spin would reduce the energy, which contradicts the fact that the system is in the ground state [26]. We do not know if this bound is realized for any finite graphs, but $b_1 = 1/2$ is the limit value for b_1 for a fully connected graph as $N \to \infty$: Partition the fully connected graph K_N of N vertices [and $M = N(N-1)/2$ edges] into one set of N' and one set of $N - N'$ vertices and assign opposite spins to the elements of these sets. The number of frustrated edges is precisely the number of edges within each set, which is

$$\begin{aligned} M_{\mathrm{fr}}(K_N) &= \frac{N'(N'-1)}{2} + \frac{(N-N')(N-1-N')}{2} \\ &= M - N'(N-N'). \end{aligned} \tag{4}$$

Thus the minimum number of frustrated edges is exactly $N^2/4 - N/2$ for $N' = N/2$, and the fraction of unfrustrated edges is

$$b_1 = \frac{1}{2 - 2/N} \to \frac{1}{2} \quad \text{as} \quad N \to \infty. \tag{5}$$

The above arguments can be generalized to directed networks straightforwardly.

3. Minimization by exchange Monte Carlo

The complexity of the ‘‘energy landscape’’ of the antiferromagnetic Ising model on an arbitrary network is difficult to judge *a priori*. There are indications that no natural network would be too hard for a regular simulated annealing approach [17,27]. To be on safer ground, we use a Monte Carlo scheme that is evidently very efficient to sweep even an extremely ‘rugged’’ energy landscape without getting stuck in local minima—the so called exchange Monte Carlo (XMC) [28]. The idea of exchange Monte Carlo is to run standard Metropolis Monte Carlo for N_T replicas of the system, each at a specific temperature. Then from time to time two replicas at adjacent temperatures are compared, and with a probability

$$P_{\mathrm{exch.}} = \begin{cases} 1 & \text{if } \Delta < 0 \\ e^{-\Delta} & \text{otherwise,} \end{cases} \tag{6}$$

where

$$\Delta = \left(\frac{1}{T} - \frac{1}{T'}\right)(E' - E), \tag{7}$$

and E is the energy of the configuration at temperature T (similarly for T' and E'), and $T < T'$. The two replicas are swapped between the temperatures. This condition is designed so that the Monte Carlo scheme preserves the Boltzmann distribution. This is not decisive for us as we are looking for the ground state energy, rather that performing a proper sampling of the configuration space, but anyway kept in our measurements. Besides just running the XMC scheme we also periodically quench the system, i.e., we sweep through all vertices of the network consecutively and flip spins that lower the energy. The sweeps are continued until a sweep with no spin flips has occurred. For later reference we introduce the notations t_{avg} for the total number of MC sweeps—we refer to the number of MC sweeps as ‘‘time’’—t_{quench} for the time between each quench, t_{exch} for the time between exchange trials, t_{measure} for the time between measurement sweeps (where the energy is sampled).

For the exchange Monte Carlo scheme to efficiently sample the configuration space all replicas needs to tour the whole range of temperatures in a reasonably short time. At the same time one would not like the exchange trials, at any neighboring temperatures, to be constantly affirmative—then the separation of the two temperatures would be of no use. We follow Ref. [28] and choose the temperature set

$$T_i = T_{\mathrm{low}} \left(\frac{T_{\mathrm{high}}}{T_{\mathrm{low}}}\right)^{(i-1)/(N_T-1)}, \tag{8}$$

where $1 \leqslant i \leqslant N_T$ enumerates the replicas. T_{low} is the lowest and T_{high} represent the highest temperatures, respectively. To find the actual parameter values (which will be stated in Secs. IV A and IV B) one has to check that the replicas travel throughout the temperature range with reasonable exchange ratios for all temperature gaps.

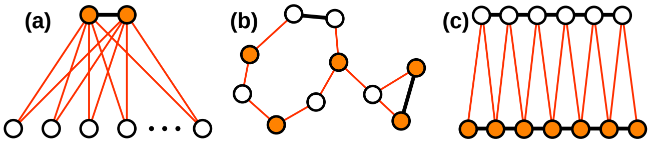

FIG. 1. Some graphs in the discussion of the b_2 quantity. The coloring of the vertices minimizes M_{fr}. Black edges indicate frustration. (a) An almost bipartite graph with many triangles. (b) A graph where all odd circuits contribute to the frustration. (c) A graph were only the shortest circuits contribute to the frustration.

B. The measure b_2

Apart from finding an approximative value of b_1, one can also define a quantity that is exactly solvable in polynomial time. Our intention is in the first hand not to make a heuristic algorithm for calculating b_1, but rather a quantity that captures the same structure, i.e., which grows monotonously with b_1.

That a graph contains no odd circuits is the defining property of bipartiteness [29]. It is thus natural that we base a bipartivity measure on an odd-circuit count in some way. Unfortunately, defining a quantity in this way becomes a little bit more complicated than at first expected. One complication is that a graph can be very close to bipartite and still contain many odd circuits [see Fig. 1(a)]. A way of dealing with this problem is to mark as few edges as possible such that each odd circuit contains at least one marked edge. In many cases a marked edge will correspond to a frustrated edge of the ground state of the antiferromagnetic Ising model. In Fig. 1(a) only the upper, horizontal edge needs to be marked. Another problem one faces is how to deal with odd circuits of different length—in a network with very few odd circuits a circuit of, say, length seven would contribute as much to the global frustration of the network as a triangle [a subgraph of three adjacent vertices—see Fig. 1(b)]. But in many real networks the total length of the odd circuits is very long (this is true for all networks we measure, see Sec. III B), much larger than M (the number of edges in the graph), in these cases the short circuits are in general the most important in determining the ground state configuration. For example, in Fig. 1(c) $M=23$, and while we have 11 triangles, summing the lengths of all odd circuits gives 218 (33 from the 11 triangles, 45 from the nine circuits of length five, and so on). However, only the triangles contribute to the ground state configuration in the sense that each triangle has the same configuration as the ground state of an isolated triangle, while all odd circuits of length larger than four (e.g., the periphery) do not have the best coloring for a circulant of that length. To deal with this we need to weigh short circuits higher than long ones. We will do this by assigning a cutoff length and neglect all circuits exceeding this length.

1. Definition

Now, we make an algorithm of the above ideas as follows. Let C_n be the set of odd circuits of length $\leqslant n$. Let $\Sigma(C_n)$ be the accumulated length of the circuits in C_n [so, for example, $\Sigma(C_3)=3$ in Fig. 1(b)]. Now we assign the cutoff $3M$ to $\Sigma(C_n)$, and let $\hat{n}$ be the smallest n such that $\Sigma(C_n)\geqslant 3M$. Next we turn to the marking procedure sketched above. Let $\nu(e)$ denote the number of circuits in $C_{\hat{n}}$ passing through the edge e. Clearly edges of high ν are likely to be frustrated in the ground state [viz. Fig. 1(a)]. We now estimate M_{fr} roughly as the number of edges that have to be marked so that each odd circuit of length $\leqslant\hat{n}$ is marked at least once. To be precise we perform the following algorithm.

```
(1) Start with $C=C_{\hat{n}}$.
(2) Sort the edges in order of $\nu$.
(3) Repeat the following while $C\neq\varnothing$.
(a) Mark the edge $e$ with highest $\nu$.
(b) Remove all circuits in $C$ containing $e$.
(c) Recalculate $\nu$ for each edge.
```

Then the number of iterations m' is the assessment of M_{fr}, and we define our bipartivity measure as

$$b_2 = 1 - \frac{m'}{M}. \tag{9}$$

This algorithm is not an attempt to actually identify the frustrated edges, rather it is supposed to give a high M_{fr} for a system with high (total) geometric frustration, and vice versa. First, it does not necessarily find the minimal number of edges needed to be marked for all odd circuits of length less than $\hat{n}$ to contain a marked edge. But we expect this steepest descent optimization to come close in most cases. Second, an odd circuit can in reality only have an odd number of frustrated edges, but in the algorithm there is no such restriction on the number of marked edges.

In case there is more than one edge with the highest ν [in step (3a) of the algorithm] we choose the edge to mark at random. The variance between different random seeds turns out to be negligible in most cases. We will run the algorithm for different seeds to choose the highest b_2 value, and get an idea about the error in b_2 from the selection of edge to mark. An alternative (and more ambitious) approach would be to iterate the whole calculation until the highest b_2 has reappeared a fixed number of times (cf. Ref. [30]).

If we assume a sparse network (i.e., $N\propto M$) the running time of the algorithm above is $O(M^2)$. To see this we first note that there can be at most $O(M)$ iterations at step (3). To find the edge with highest ν [in step (3a)] we do not need to sort all edges more than once [as done in step (2)]. Instead we can find this out while recalculating ν [in step (3c)]. Removing all circuits containing e [as in step (3b)] can be done in time bounded by the total length of circuits containing e, which cannot be larger than $3M$. Step (3c) also needs to go through all circuits passing e and thus needs the same running time as step (3b). To sum this up, the running time for this section of the algorithm is of order N^2.

2. Limit properties

In the $N\to\infty$ limit the b_2 measure lies in almost the same interval as b_1. The upper limit $b_2=1$ is attained if and only if the graph is bipartite. [If the graph is bipartite $C_{\hat{n}}$ is empty and $\nu(a)=0$ for all a, so $m'=0$ and $b_2=1$. If there exist odd circuits $m'\geqslant 0$, so $b_2<1$.] b_2 cannot be as low as 0 (if one marks all edges, every circuit must be marked). Since the b_2 definition is inspired by the ground-state configuration of the antiferromagnetic Ising model, we expect a similar lower bound to b_2 as to b_1. In Appendix A we argue that the lower bound on the b_2, as for the b_1 measure, is 1/2 in the $N \to \infty$ limit.

3. The complete algorithm

So far we have overlooked the central part in calculating the b_2 measure—namely, to find odd circuits. To do this we use a modified version of Johnson's algorithm [31]. In principle, Johnson's algorithm is a depth first search where, to avoid futile searching, some vertices are blocked while stepping down the search tree. The running time for Johnson's algorithm is $O(M(C+1))$ (if $M>N$) where C is the total number of circuits. Now C can grow fast with N which would make the finding of all odd circuits a quite intractable computation. In many cases the cutoff of the circuit length, which we introduced above to give less priority to long circuits, saves us by setting a limit on the search depth. To implement this we let $\overline{n}$ be the current upper bound on circuit length (or search depth), and $\overline{\Sigma}$ be the current sum of odd circuits $\leqslant \overline{n}$. As soon as $\overline{\Sigma} \geqslant M$ we iteratively decrease $\overline{n}$ by 2 and recalculate $\overline{\Sigma}$ until $\overline{\Sigma}<3$. If $\overline{\Sigma}<M$ when the search is over we rerun the procedure where we use $\overline{n}+2$ as our new (fixed) $\overline{n}$ [32]. When the search is over we assign $\hat{n}$ the value $\overline{n}$. For dense bipartite graphs the algorithm is intractable. In the worst case, the full bipartite graph $K_{N/2,N/2}$, there are

$$C(K_{N/2,N/2}) = \sum_{k=4}^{N} \frac{1}{2k}\left[\frac{(N/2)!}{(N/2-k/2)!}\right]^2 \tag{10}$$

circuits (where the sum is over even values of k) [33] giving a running time of $O(N^2 C(K_{N/2,N/2}))$. One can of course decide whether or not a graph is bipartite in linear time, but nonbipartite cases of similar complexity are easily constructed (by, e.g., adding an isolated triangle). In practice these worst cases are, probably, very rare—a, relatively speaking, very low density of odd circuits is needed to get a small $\hat{n}$—even in the real-world network with highest bipartivity we have $\hat{n}=3$. In this case ($\hat{n}=3$) all odd circuits are found in $O(M^2)$ time.

Now we turn to a more complete description of the algorithm. Johnson's algorithm takes the "least" (smallest in some enumeration) vertex in a strongly connected subgraph as its starting point. To find strongly connected components we use the algorithm in Ref. [34]. To sum up, the algorithm reads as follows:

(1) Mark all vertices as unchecked.

(2) While there are unchecked vertices, iterate the following.

(a) Pick an unchecked vertex v.

(b) Find the largest strongly connected component Λ_v containing v.

(c) Set $\Lambda := \Lambda_v$ and repeat the following steps as long as $\Lambda \neq \varnothing$.

(i) Pick the least vertex u of Λ.

(ii) Call a subroutine implementing the modified Johnson's algorithm. Recalculate $\overline{n}$ and add $C_{\overline{n}}^{-}$ to a list $\mathcal{C}$. Delete circuits longer than $\overline{n}$ from $\mathcal{C}$.

(iii) Delete u from Λ.

(3) Set $\hat{n} := \overline{n}$.

(4) Run the algorithm described above (in Sec. II B 1) to mark edges and calculate b_2.

In all cases, step (2) sets the limit on running time. As mentioned, in most applications we expect the running time of step (2) to be $O(M^2)$ [similarly to that of step (4)].

C. Clustering coefficients

In this section we state the definitions of clustering coefficients [11,22] c and d that will be used for comparison.

Let $\xi(n)$ denote the number of representations of circuits of length n, and let $\zeta(n)$ denote the number of representations of paths of length n. (By representations we mean different ways of listing adjacent vertices; so, for example, an undirected triangle has six representations, whereas a directed triangle has three representations.) Then we define

$$c = \frac{\xi(3)}{\zeta(3)} \quad \text{and} \quad d = \frac{\xi(4)}{\zeta(4)}. \tag{11}$$

For directed graphs we consider the directed versions of c and d and denote this by the subscript "dir."

Note that these definitions differ from the Watts-Strogatz clustering coefficient [35], defined as $M(\Gamma_v)/\binom{k_v}{2}$ [$M(\Gamma_v)$ is the number of edges in the neighborhood of the vertex v, and k_v is v's degree] averaged over V. Nevertheless, the definitions of c and d are common in physicists network literature (and have been used for a long time in social network analysis). They are also more natural measures for the whole graph's clustering since they can be interpreted as the number of triangles (or squares, i.e., four circuits), normalized by dividing by the maximal number of triangles (or squares) for a graph with the given number of paths of length three (or four). In this sense c and d can be thought of as the densities of triangles and squares in the graph.

III. THE NETWORKS

A. Test networks with tunable bipartivity

To test and compare the b_1 and b_2 quantities we construct three types of test networks where the bipartivity can be tuned by model parameters. The principle behind all models is to start from bipartite networks and add lesser or greater number of edges within a partition to create odd circuits.

One type (model 1) is a quite straightforward generalization of the Erdös-Renyi (ER) model [36]: We partition the vertices in two disjoint sets of sizes $\tilde{N}$ and $N-\tilde{N}$. Then we add $r_1 M$ edges randomly between vertices of the different sets, and $(1-r_1)M$ edges regardless of what set the vertices belongs to [see Fig. 2(a)]. In this way we interpret r_1 as the strength of the heterophilous preference in a model where bipartivity is the only structural bias. The choice of vertex

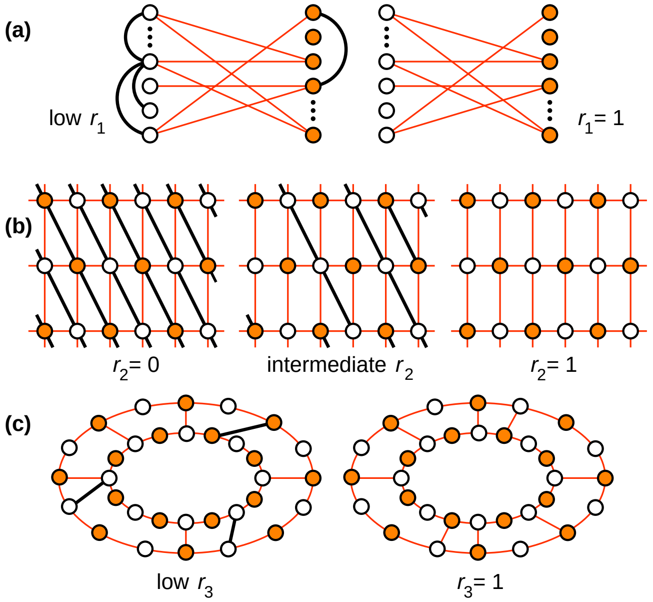

FIG. 2. Construction of the test networks. (a) shows the generalization of the ER model (model 1). (b) shows interpolation between quadratic and triangular lattices (model 2). (c) shows the model with predominantly longer circuits (model 3). All models are bipartite for $r_{1,2,3}=1$. Additional edges create odd circuits (frustration) for lower $r_{1,2,3}$ values. The black lines illustrate these additional edges. The white and nonwhite vertices symbolize a partition giving $b_1=1$ in the $r_{1,2,3}=1$ case (it is not meant to represent the optimal coloring when $r_{1,2,3}<1$).

pairs is done with randomness, the only restriction being that loops and multiple edges are not allowed. If $r_1=0$ the model reduces to the ER model, while for $r_1=1$ the networks are bipartite (cf. Ref. [37]). This model is probably the most random (i.e., having least structural biases) model with tunable bipartivity. The disadvantage is that the expectation values of b_1 and b_2 are hard to calculate (even in the frustrated limit $r_1=0$).

Model 2 interpolates between two-dimensional square and triangular lattices. We start, for $r_2=0$, with a triangular grid with periodic boundary condition. Let L, the linear dimension of the system (i.e., $N=L^2$), be even. For a nonzero parameter value we (by uniform randomness) delete r_1L^2 "diagonal" edges creating frustration as illustrated in Fig. 2(b). To be more precise, if we index the vertices as (i_x,i_y), $1\leqslant i_x,i_y\leqslant L$, then the edges are $[(i_x,i_y),(i_x+1,i_y)]$ and $[(i_x,i_y),(i_x,i_y+1)]$ (giving the square grid) plus r_1L^2 edges of the form $[(i_x,i_y+1),(i_x+1,i_y)]$ chosen by uniform randomness (addition is modulo L). This model has a high degree of short circuits. The extremes $r_2=0$ and $r_1=1$ represent two generic lattice types. The symmetries of the regular networks simplify the calculations of, e.g., limit properties for the bipartivity measures. If $r_2=1$ the system is bipartite (note that L has to be even for this to hold) so $b_{1,2}=1$. When $r_2=0$ we have $b_1=b_2=2/3$: For the lower limit of the b_1 quantity, see Ref. [38]. For the lower limit b_2 we note that $\Sigma(C_3)=6N$ (since each vertex can be associated with two triangles). This gives $\hat{n}=3$ and $\nu=2$ for all edges. Now it is enough to mark N edges (e.g., all $[(i_x,i_y),(i_x+1,i_y)]$ edges). In this case we note that each edge will have $\nu=2$ when it is marked, which means that the marking sequence is optimal and that the number of iteration cannot be less with another choice of edges to mark. So $b_2=1-N/3N=2/3$. The major disadvantage with model 2 is that the average degree is a function of r_2 $[M=(3-r_2)L^2]$. This change in the average degree can make it harder to separate effects of the shift in bipartivity from the shift in average degree.

In both model 1 and (even more) model 2 triangles will dominate the set of odd circuits. To test networks with predominantly longer circuits we construct a model 3 as follows [see Fig. 2(c)]: We make two circulants of size $N/2$ with the vertices $\{v_1^i,\ldots,v_{N/2}^i\}$ and edges $\{(v_1^i,v_2^i),\ldots,(v_{N/2-1}^i,v_{N/2}^i),(v_{N/2}^i,v_1^i)\}$, $i\in\{1,2\}$. Then we add M_{trans} transverse edges between the circulants. $M_{\text{trans}}/2$ of these edges are placed out separated by equal distance N/M_{trans} separating the double circulants into $M_{\text{trans}}/2$ "sectors." Then we fill up each sector with another transverse edge. With probability r_3 we add an (v_i^1,v_i^2) edge [such that (v_i^1,v_i^2) is none of the previously added transverse edges], otherwise we add a (v_i^1,v_i^2+1) edge (addition modulo $N/2$). We note, to a first approximation, that if $r_3=0$ marking (in the process of calculating b_2) one edge between every transverse edge on one of the circulants is needed to mark the shortest odd circuits. This will make $b_2\in O(1-M_{\text{trans}}/N)$.

B. Real-world networks

Physicists' networks studies have, in the spirit of statistical mechanics, emphasized the properties remaining when the system grows beyond any limit. Bipartivity, as discussed above, is well defined for all system sizes. Still it is a quantity that can potentially suffer from finite-size effects (from the fact that not all real neighbors of all actors in an empirically constructed social network are a part of the graph) and is therefore preferably measured for large networks. Now the problem is to find data for large-scale real-world networks of social interaction. In general two methods have been successful for this purpose—one either uses professional collaborations of some sort or data from interaction over the Internet (either in Internet communities [11,39] or through email exchange [40]).

1. Professional collaboration networks

In the professional collaboration networks we study the vertices as professionals of some field—networks of scientists and company directors are considered in this papers, the movie-actor network is another frequently studied example; the edges represent that two actors have been involved in the same professional collaboration. This is sometimes referred to as a "one-mode" representation of an affiliation network (as opposed to the bipartite two-mode representation discussed in Sec. I).

Professional collaboration networks are no doubt interesting in their own right as accounts for the interaction dynamics of the respective fields. Assuming that the formation of professional ties follows similar principles as general human interaction, we can use professional collaboration networks

TABLE I. Sizes, clustering coefficients, and bipartivity measures b_1 and b_2 for real-world social networks.

Network	N	M_{dir}	M	c_{dir}	c	d_{dir}	d	b_1^{dir}	b_1	b_2^{dir}	b_2
All contacts	29 341	174 662	115 684	0.012	0.0060	0.016	0.017	0.859	0.860	0.948	0.928
Messages	20 691	73 346	52 435	0.0052	0.0061	0.0081	0.0061	0.897	0.892	0.984	0.964
Guestbook	21 545	76 257	55 076	0.014	0.014	0.015	0.021	0.863	0.889	0.943	0.965
nioki.com	50 259	405 742	239 452	0.0076	0.0065	0.016	0.013	0.842	0.855	0.956	0.975
Emails	637	554	443	0.11	0.16	0.071	0.14	0.944	0.944	0.971	0.941
arxiv.org	52 909		490 600		0.45		0.35		0.630		0.623
Directors	7475		48 899		0.21		0.37		0.549		0.507
Karate club	34		78		0.26		0.26		0.782		0.782
Prison	64	182	85	0.19	0.31	0.089	0.14	0.786	0.878	0.918	0.847

to draw conclusions about the structure of more general social networks. However, at one point (at least) professional collaboration differs from general social interactions: A collaboration tie does not necessarily imply a strong personal acquaintance, but in these networks each collaboration constitutes a fully connected cluster. This leads to higher fraction of short circuits than, say, a friendship network.

One of the professional collaboration networks we use is of scientists who have uploaded manuscripts to the preprint repository arxiv.org. Two scientists are linked if their names (identified by surname and initials) appear together on at least one preprint. A detailed description of this network can be found in Ref. [41]. In the other professional collaboration network the vertices represent company directors from the Fortune top 1000 list of companies in U.S.A. in the year 2001. An edge (collaboration) in this network means that two directors are sitting in board of the same company. A detailed description of this network can be found in Ref. [42]. Sizes of the networks can be seen in Table I.

2. Online interaction networks

In online interaction networks, the vertices are users of Internet communities and an arc (A,B) is added if A contacts B, or if A adds B to his/her list of friends [11,39]. Another kind of online interaction networks are email networks [40], where an arc can be assigned if an email is sent, or if a person adds another to his/her address book. Just as for professional collaboration networks, one can argue that online interaction networks are representative as general social networks. One can assume that new contacts are formed through preference-matching searches to a larger extent, and introduction by mutual friends to a lesser extent, than in general friendship networks. Since the introduction of mutual friends to each other is believed to be the major cause of high clustering (large density of triangles, or, large transitivity) [43] one can expect a lower clustering in networks of online interaction (still the clustering in these network seems to be finite in the $N\to\infty$ limit [11]).

The specific online interaction networks we consider are constructed from the Internet communities nioki.com and pussokram.com. The nioki.com data are described in Ref. [39]. In these data an arc (A,B) means that B is listed as a friend by A, which allows A to see if B is online and send instant messages to B. In the pussokram.com data the arcs correspond to communication between the users. There are four different types of communication in this specific network (all described in detail in Ref. [11]). We use the networks obtained from two types of interaction (‘‘messages’’—like ordinary emails within the community, and ‘‘guest book’’—where one user contacts another by writing in his/her guest book), and the network of any of the four types. Network sizes can be found in Table I.

Another large difference between the pussokram.com and nioki.com data is that the former community has a very pronounced romantic profile, encouraging flirts and romantic correspondence. nioki.com has also a search engine to ‘‘‘trouve l’amour’’ (find love), but that is all.

Apart from the two Internet communities, we study another type of online interaction network based on the flow of email. For this network all incoming and outgoing email traffic to a server was logged for around three months [40]. The server handles undergraduate students’’ email accounts at Kiel University, Germany. Thus there are two categories of vertices—internal vertices, whose activity is accurately mapped; and external vertices, which only have edges leading to internal vertices. In this study we restrict ourselves to the network of internal-internal contacts. The reason we do not include external contacts is that we would miss the (probably many) circuits containing external-external edges which would bias the bipartivity.

3. Network from interview and field survey

Apart from the above networks, all obtained from databases, we also measure the bipartivity of two networks obtained from interview and field surveys. The first dataset is gathered by observations of interaction between members of a university karate club [44]. We also study the network of acquaintance ties in a prison [45]. The outgoing arcs from A correspond to prisoners listed by A in response to the question: ‘‘What fellows on the tier are you closest friends with?’’ Due to their acquisition methods these kind of real-world networks have to be rather small. This can, as mentioned, result in finite-size effects. On the other hand they, most likely, more truly reflect the structure of real acquaintance networks.

IV. RESULTS

In this section we present the results of the test networks and the measurement for the real-world social networks.

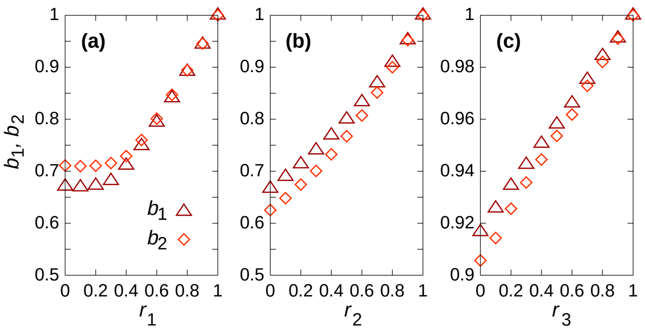

FIG. 3. The bipartivity measures vs the model parameters of the two models defined in Sec. III A. (a) shows the result for model 1, (b) shows the result for model 2, and (c) shows the result for model 3. All error bars would be smaller than the symbol size. The monotonous growth of the bipartivity measures shows that the measures behave expectedly.

A. Test networks

As expected, both b_1 and b_2 are monotonously increasing as functions of the r_1, r_2, and r_3 parameters of [almost [46] all our test network (see Fig. 3)]. This is encouraging and suggests that both b_1 and b_2 are quite relevant measures of bipartivity.

The model 1 measurements shown in Fig. 3(a) are made with the model parameters $N=2\tilde{N}=100$ and $M=800$. We have checked many other sizes too, but all have the characteristic appearance of Fig. 3(a)—a linear increase of b_1 and b_2 for larger r_1 and a flatter slope for r_1 close to zero. This shape is expected from the discussion in Sec. I—in networks where a heterophilous preference is the only structure-inducing force, only the strong preference limit gives a strong measurable effect: Close to the ER limit $r_1\approx 0$, the original two partitions will not be identified correctly, only when the different partitions (to a large extent) have different signs will the bipartivity be proportional to the strength of the heterophilous preference.

As seen in Fig. 3(b), model 2 shows an almost linear functional form of $b_{1,2}(r_2)$. In this case, triangles dominate the odd circuits even at small values of r_2. Tuning r_2 will give a proportional increase of the number of triangles. Thus a linear r_2 dependence of b_2 would be expected.

Also model 3 has linear $b_{1,2}$ versus r_3 curves. The model parameters used are $N=100$ and $M_{\mathrm{trans}}=10$. As mentioned in Sec. III A, we expect $b_2\approx M_{\mathrm{trans}}/N$ for $r_3=0$, which is confirmed in Fig. 3(c).

An interesting feature is that, in Fig. 3(a), b_2 is consistently higher than b_1; whereas in Figs. 3(b) and 3(c) the situation is reversed. To explain this we note that different parts of the b_2 algorithm overestimate or underestimate b_2 with respect to b_1. (As mentioned before, this does not matter much if one sees b_2 as another bipartivity measure, rather than an approximation of b_1.) To be more specific, the marking procedure in the b_2 algorithm neglects some constraints of the b_1 quantity, e.g., two edges can be marked in a triangle subgraph (in the b_2 calculation), but there cannot be two frustrated edges in a triangle in the antiferromagnetic Ising model. This lack of constraints in the b_2 calculation tends to make b_2 larger than b_1. On the other hand, the fact that we do not minimize the marked edges—most notably, if there is more than one edge with the highest ν, we choose the edge to mark at random, see Sec. II B 1—leads to $b_2<b_1$. Situations with many edges of equal (highest) ν are more frequent in more regular networks. This explains that the more regular models 2 and 3 have $b_2<b_1$, whereas model 1 (which lacks an underlying regular lattice) has $b_1<b_2$.

The measurements for both b_1 and b_2 are averaged over 100 network realizations. The XMC scheme for the b_1 quantity is ran at 24 temperatures in parallel, between temperatures 0.01 and 2. Other network parameters are $t_{\mathrm{avg}}=4\times 10^5$, $t_{\mathrm{measure}}=4$, $t_{\mathrm{quench}}=20$, and $t_{\mathrm{exch}}=1000$. These are more modest parameter values than we will use for the real-world networks, but the test networks are also much smaller, and since the distribution, of b_1 and b_2 are (likely) symmetric, the network average helps to reduce the error.

B. Real-world social networks

Now we turn to the result for the bipartivity measures of real-world networks. The values are presented in Table I. For comparison we also give values for the clustering coefficient c and the density of squares d in both directed and undirected versions. Undirected networks are constructed by taking the reflexive closure. At first glance at the table we arrive at the pleasing conclusion that the bipartivity for the pussokram.com networks is very high (as expected from a network of romantic interaction of mostly heterosexuals). But disappointingly, the bipartivity measures show similarly high values for the nioki.com and email networks. This can be explained by the fact that nioki.com, just like the pussokram.com, data have very low c and d values, and presumably very few circuits. Now branches (subgraphs without circuits that can be isolated by cutting one edge) not give a positive contribution to either b_1 or b_2, no irrespective of the gender of the agents. The email network does have a high clustering, but still rather high bipartivity. The reason is that the email network is rather heavily fragmented and contains many isolated subnetworks of two vertices and one edge, and three vertices and two edges. Such subnetworks do not affect the clustering coefficient but tend to decrease the bipartivity measures [47].

The collaboration networks consist of a number of fully connected clusters (corresponding to a specific collaboration) that are interconnected. It is thus natural that we see low bipartivity and a high density of short circuits. The lower bipartivity values for the company director network can be explained by smaller average size of such fully connected clusters. The average number of vertices per collaboration is 9.5 for the corporate director network and 2.5 for the scientific collaboration data [41,42].

The two small networks constructed from field surveys (the "karate club" and "prison" networks of Table I, discussed in Sec. III B 3) show midrange bipartivities and relative high values of c and d. From the above discussion we can expect that the bipartivity of large, real, acquaintance networks is somewhere between those of the collaboration networks and the Internet community networks (because they probably have higher clustering than Internet community networks, and lower number of fully connected clusters than the collaboration networks). Encouraging enough, this is exactly what we see in Table I. Of course, the very small system sizes might affect the results, but that the bipartivity measures of real-world acquaintance measures would be close to either the upper or lower limits seems hard to believe.

As a word of caution we note that the bipartivity often correlates with other structural measures and is dependent on the sizes (N and M) of the network. The bipartivity measures, as comparative tools, are most applicable to classes of networks with a constant average degree. Furthermore, if the network has a low clustering and high bipartivity, one has to carefully consider the network forming dynamics to be able to choose an appropriate picture of the situation—if the high bipartivity causes the low clustering, if the low clustering causes the high bipartivity, or if the two structures are independent.

We conclude this section by a note on the parameters for the XMC optimization. The measurement of b_1 for all real-world network (except the nioki.com data where we study the convergence more carefully) are done just once with the following simulation parameters: $N_T=24$ (with temperatures from 0.002 to 5), $t_{\text{avg}}=1\times10^7$, $t_{\text{measure}}=16$, $t_{\text{quench}}=40$, and $t_{\text{exch}}=2\times10^4$.

V. SUMMARY AND DISCUSSION

This paper concerns the quantification of the network structure ‘‘bipartivity’’—how close to bipartite a given graph is. Such measures are potentially interesting in the study of sexually transmitted diseases, genealogical maps of disease outbreaks, trade networks, and food webs. We propose two measures for this quantity. One quantity b_1 based on the optimal two coloring of the network—or, equivalently, the ground state of the antiferromagnetic Ising model on the network. The exact value of this quantity (which has been used in different roles elsewhere) is NP complete and thus in general not feasible to calculate exactly. Instead we seek an approximate solution by a simulated annealing approach. The simulated annealing is based on the exchange Monte Carlo scheme. We argue that this unorthodox minimization method helps us avoid local minima of the energy landscape of the antiferromagnetic Ising model. This method could also be useful in ground-state studies of traditional systems of statistical mechanics. Furthermore, we develop a measure b_2 based on the count of odd circuits that, for almost all networks, is calculable in polynomial time.

We propose three different random graph test models where one can interpolate between arguably nonbipartite and bipartite graphs by tuning a control parameter. Both our bipartivity measures are shown to increase monotonically with tuning the control parameters towards the bipartite extreme. From this we conclude that the bipartivity measures really quantify the notion of bipartivity.

By considering example networks we infer that bipartivity is a structure that cannot be measured by currently popular structural measures, such as the clustering coefficient. At the same time any sensible quantification of bipartivity probably has to have a positive correlation with the clustering coefficient for most networks [with exceptions for exotic cases like Fig. 1(a)]—so, in that case bipartivity and clustering are not independent.

We measure b_1 and b_2 of a number of real-world networks, constructed from online interaction, professional collaborations, and field surveys. As expected, we see high bipartivity values for data from the Internet community pussokram.com, where romantic contacts are encouraged, and hence a high degree of heterophilous interaction expected. We also see the expected low bipartivity values for the professional collaboration and empirical acquaintance networks we study. Disappointingly we cannot use our bipartivity measures to distinguish between the networks driven by romantic or friendship (or professional) contacts. To do this other structures and the network sizes have to be taken into account, in a more elaborate analysis (which is out of the scope of this study).

We conclude with an analogy to linear algebra—we have identified a dimension (structure) and proposed base vectors (measures), which unfortunately are not orthogonal to the other dimensions.

ACKNOWLEDGMENTS

We would like to thank Niklas Angemyr, Stefan Bornholdt, Gerald Davis, Holger Ebel, Michael Lokner, Stefan Praszalowicz, and Christian Wollter for help with data acquisition; and Jens Andersson, Johan Giesecke, James Moody, Mats Nylén, and Pontus Svenson for comments and suggestions. P.H. was partly supported by the Swedish Research Council through Contract No. 2002-4135. F.L. was supported by the National Institute of Public Health. C.R.E. was supported by the Bank of Sweden Tercentenary Foundation. B.J.K. was supported by the Korea Science and Engineering Foundation through Grant No. R14-2002-062-01000-0.

APPENDIX A: THE LOWER BOUND OF THE MEASURE b_2

In this appendix we argue that, in the $N\to\infty$ limit, the lower bound for b_2 is 1/2 (just like b_1). First we conjecture that the minimal value for b_2, just as for b_1, is attained for complete graphs. (This will be further motivated below.)

To assess b_2 for complete graphs, we note that [48]

$$\Sigma(C_n)=\sum_{\text{odd } 3\leqslant i\leqslant n}\frac{N!}{2(N-i)!}, \tag{A1a}$$

which implies

$$\Sigma(C_3)=\frac{N(N-1)(N-2)}{2}\geqslant\frac{N(N-1)}{2}=M, \tag{A1b}$$

so $\hat{n}=3$ which results in that $\nu=N-2$ for each edge.

Now we apply the marking procedure of Sec. II B 1. Marking an edge (u,v) makes $\nu(u,v)=\nu(v,u)=0$. Further-

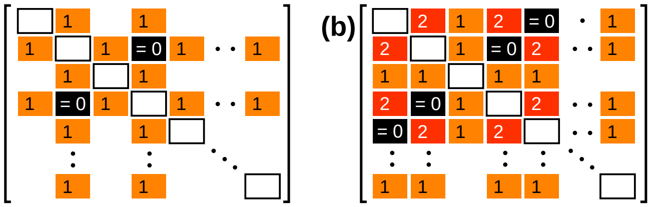

FIG. 4. Marking of edges (in matrix representation) while calculating the b_2 quantity for a fully connected graph. “-1” means that ν at that position is decreased by one unit, “$=0$” means that $\nu=0$ at that position.

more, every edge (u,w) and (v,w) $(w\neq u,v)$ will be decreased by one since the triangle $\{u,v,w\}$ now contains a marked edge. The discussion will be simplified by considering a matrix representation of $\nu(u,v)$. Marking (u,v) sets $\nu(u,v)=\nu(v,u)=0$ and decreases the uth and vth columns, and uth and vth rows by 1 [an example is given in Fig 4(a)]. Marking another edge (u',v') [u' and v' are different from both u and v, otherwise $\nu(u',v')$ would not be maximal] will have the same effect as marking the first. For positions like (u,v') the original ν are decreased by 2 [see Fig. 4(b)], since it has lost the two passing triangles $\{u,u',v'\}$. and $\{v',u,v\}$. Continuing this process we see that it takes $N/2+O(1)$ markings for ν of each edge to be decreased by two units, and thus $m'=N^2/4+O(N)$ markings to make $\nu=0$ for all edges. This gives $b_2=1/2$ in the $N\rightarrow\infty$ limit. Since the appropriateness of b_2 as a bipartivity measure is not really dependent on the limit values, we will not give a rigorous proof that the correction is of a lower order for all levels of the marking procedure [one level is the $N/2+O(1)$ edges needed to be marked for ν to be decreased by at least two units for each edge].

Now we argue that the b_2 takes its minimal value for complete graphs. First we note that the number of circuits of length n per edge, for any n, is largest in a complete graph [2]. So if we set $\hat{n}$ arbitrarily and discard circuits of length $\leqslant n$ in the calculation of $\nu(v)$, the fully connected graph would give the highest m' value and thus the lowest bipartivity measure. The strongest candidate for a lower bipartivity measure than that of a fully connected graph would thus be a graph such that the $\Sigma(C_n)<3M$ and $\Sigma(C_{n+2})$ is as big as possible for some n. But the number edges needed to be removed from a fully connected graph for $\Sigma(C_n)<3M$ to hold, which not only reduces the contribution to ν from circuits of length n but also from circuits of length $n+2$ to a similar extent. If one performs the approximate marking procedure outlined above for circuits of length five, one starts from $\nu=(N-2)(N-3)(N-4)$ and it takes $N/2+O(1)$ markings to decrease every ν with at least $2N^2$. This means that the number of edges needed to be marked to make $\nu=0$ for every edge is the same if circuits of length five are considered. It also means that a graph as outlined above [with $\Sigma(C_n)<3M$ and $\Sigma(C_{n+2})$ are as big as possible] probably does not have a lower b_2 than a complete graph.

To epitomize, the b_2 measure lies in the interval $[1/2,1]$ in the $N\rightarrow\infty$ limit. The finite-size corrections to b_2 for fully connected graphs, however, turn out to make b_2 slightly less than 1/2.

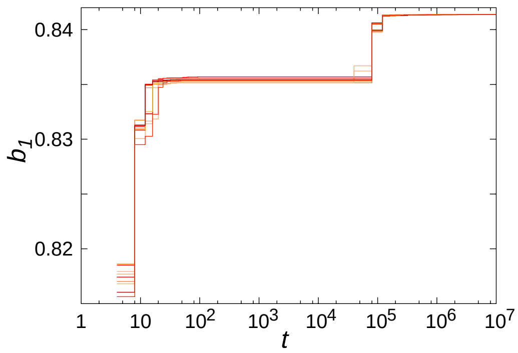

FIG. 5. The current value of b_1 (at the lowest-temperature level of the cooling) as a function of running time for ten independent measurements of the directed version of the nioki.com data.

APPENDIX B: CONVERGENCE OF THE SIMULATED ANNEALING

To analyze the convergence of the simulated annealing scheme we run ten independent calculations of the b_1 quantity (with the same parameter values as in Sec. IV B). The individual time evolutions of b_1 (at the lowest temperature $T=0.002$) for the different runs are shown in Fig. 5. We note that already after the first quench b_1 is only 3% away from the value at the end of the run, and after 50 time steps b_1 is 0.5% of the value after 1×10^7 time steps. We note that there is no way of constructing a statistically valid confidence interval for the true b_1 value since an arbitrary complex energy landscape could have a global minimum with a basin of attraction of measure zero. There are, however, indications that this is seldom a major problem, at least not for the bisection problem [17].

An interesting observation from Fig. 5 is the steplike structure. This is a result of the exchange trials: After $t\approx100$ the local minimum has been found, but at the temperature in question the system is, in principle, stuck in a confined part of the configuration space, and cannot enter lower lying energy valleys. In the time scale $t=10^5$ there is another jump in the b_1 value. This is related to that other replicas from other parts of the configuration space reache the lowest level. At around $t=10^6$ the current highest b_1 values (lowest energy) reache another plateau. At this time, each replica should have covered the whole temperature range several times. This second plateau gives two encouraging implications. First, that the correct value of b_1 probably is not very far off the measured value. Second, that the exchange steps really are helpful. If one wants to run this algorithm more efficiently the t_{exch} we use is far too large (but beneficial for separating the time scales in the discussion above). Ideally t_{exch} should probably be chosen to be of the same order as the first jump (from the regular Monte Carlo steps)—in the nioki.com network (displayed in Fig. 5) this would be $t\approx100$.

[1] S.H. Strogatz, Nature (London) **410**, 268 (2001); R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002); S.N. Dorogovtsev and J.F.F. Mendes, Adv. Phys. **51**, 1079 (2002).

[2] M.E.J. Newman, SIAM Rev. **45**, 167 (2003).

[3] D.J. Watts and S.H. Strogatz, Nature (London) **393**, 440 (1998).

[4] R. Albert and A.-L. Barabási, Science **286**, 509 (1999).

[5] M.E.J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).

[6] G. Caldarelli, R. Pastor-Santorras, and A. Vespignani, e-print cond-mat/0212026.

[7] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.-L. Barabási, Nature (London) **407**, 651 (2000).

[8] F. Liljeros, C.R. Edling, L.A.N. Amaral, H.E. Stanley, and Y. Åberg, Nature (London) **411**, 907 (2001).

[9] P.S. Bearman, J. Moody, and K. Stovel (unpublished).

[10] When the type of every vertex is known, this structure can be measured by Freeman's segregation index *S*, which is (roughly speaking) the fraction cross-type edges missing in a graph, compared with a completely random graph—a graph that is close to bipartite would then have $S<0$. L.C. Freeman, Sociolog. Methods Res. **6**, 411 (1978); See also: J.C. Mitchell, Connections **2**, 9 (1978); L.C. Freeman, *ibid.* **2**, 13 (1978).

[11] P. Holme, C.R. Edling, and F. Liljeros, e-print cond-mat/0210514.

[12] F. Liljeros, C.R. Edling, and L.A.N. Amaral, Microbes and Infection **5**, 189 (2003).

[13] R.M. Anderson and R.M. May, *Infectious Diseases of Humans* (Oxford University Press, Oxford, 1991).

[14] H.C. White, AJS **87**, 517 (1981).

[15] S.L. Pimm, *Food Webs*, 2nd. ed. (University of Chicago Press, Chicago, 2002).

[16] It should be noted that many NP-hard optimization problems display phase transitions between "easy" and "hard" regimes, e.g., the three-coloring problem is known to be hard in the small-world regime of the WS model [3]. T. Walsh, in *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, edited by T. Dean (Morgan Kaufmann, San Francisco, 1999). For general references, see, e.g., P. Cheeseman, B. Kanefsky, and W.M. Taylor, in *Proceedings of IJCAI-91*, edited by J. Mylopoulos and R. Reiter (Kaufmann, San Mateo, CA, 1991), pp. 331–337; T. Hogg, B.A. Huberman, and C.P. Williams, Artif. Intell. **88**, 1 (1996).

[17] M. Jerrum and G. Sorkin (unpublished).

[18] G.R. Schreiber and O.C. Martin, SIAM J. Optim. **10**, 231 (1999).

[19] Y. Fu and P.W. Anderson, J. Phys. A **19**, 1605 (1986).

[20] M.J. Alava, P.M. Duxbury, C.F. Moukarzel, and H. Rieger, in *Phase Transitions and Critical Phenomena*, edited by C. Domb and J.L. Lebowitz (Academic Press, London, 2001), Vol. 18, pp. 143–317.

[21] R.M. Karp, in *Complexity of Computer Computations*, edited by R. E. Miller and J.W. Thatcher (Plenum Press, New York, 1972), pp. 85–103.

[22] A. Barrat and M. Weigt, Eur. Phys. J. B **13**, 547 (2000).

[23] See, e.g., M. Gitterman, J. Phys. A **33**, 8373 (2000); P. Svenson, Phys. Rev. E **64**, 036122 (2001); B.J. Kim, H. Hong, P. Holme, G.S. Jeon, P. Minnhagen, and M.Y. Choi, *ibid.* **64**, 056135 (2001); C.P. Herrero, *ibid.* **65**, 066110 (2002); A. Aleksiejuk, J.A. Holyst, and D. Stauffer, Physica A **310**, 260 (2002); D. Boyer and O. Miramontes, Phys. Rev. E **67**, 035102 (2003); K. Medvedyeva, P. Holme, P. Minnhagen, and B.J. Kim, *ibid.* **67**, 036118 (2003); G. Bianconi, Phys. Lett. A **303**, 166 (2002); A. Aleksiejuk-Fronczak, e-print cond-mat/0206027.

[24] D.B. Bahr and E. Passerini, J. Math. Sociol. **23**, 1 (1998); D.B. Bahr and E. Passerini, *ibid.* **23**, 29 (1998); S.N. Durlauf, Proc. Natl. Acad. Sci. U.S.A. **96**, 10 582 (1999); H.P. Young, in *The Economy as an Evolving Complex System*, edited by L.E. Blume and S.N. Durlauf (Oxford University Press, Oxford, 2003).

[25] For an interesting discussion on this problem in a somewhat more complex Ising spin-glass model, see F. Barahona, J. Phys. A **15**, 3241 (1982).

[26] Note that our starting point is that there are no known attributes to the vertices. If one would know, e.g., the gender of the actors in a social network, the fraction of edges between actors of the same gender is in the interval [0,1].

[27] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, Science **220**, 671 (1983).

[28] K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604 (1996).

[29] See any introductory text on graph theory, for example, A. Tucker, *Applied Combinatorics*, 3rd ed. (Wiley, New York, 1995), p. 31.

[30] L.R. Walker and R.E. Walstedt, Phys. Rev. B **22**, 3816 (1980).

[31] D.B. Johnson, SIAM J. Comput. **4**, 77 (1975).

[32] The intuitive way to find a least upper bound might be to search at two levels simultaneously ($\bar{n}$ and $\bar{n}-2$) and decrease the bound $\bar{n}\mapsto\bar{n}-2$ when $\Sigma_{\bar{n}-2}\geqslant M$ ($\Sigma_{\bar{n}-2}$ denotes the sum of the length of all circuits shorter than or equal to $\bar{n}-2$). This would slow down the computation considerably since our modified Johnson's algorithm mostly finds circuits of the length of the search depth $\bar{n}$, and thus it takes a long time to increase $\Sigma_{\bar{n}-2}$.

[33] Consider $K_{N/2,N/2}=(V,U,E)$ where *V* and *U* are the two vertex sets. We write a circuit as a *k*-tuple $(v_1,u_1,\dots,v_{k/2},u_{k/2})$ where $v\in V$ and $u\in U$. Then there are $(N/2)(N/2)\cdots[N/2-(k/2-1)][N/2-(k/2-1)]=[(N/2)!/(N/2-k/2)!]^2$ distinct *k*-tuples; As for circuits, the choice of start vertex v_1 does not matter; neither does the direction matter. To compensate for this we divide by 1/2*k* to get the right number of circuits of length *k* in $K_{N/2,N/2}$.

[34] A.V. Aho, J.E. Hopcroft, and J.D. Ullman, *The Design and Analysis of Computer Algorithms* (Addison-Wesley, Reading, MA, 1974), pp. 189–195.

[35] D.J. Watts and S.H. Strogatz, Nature (London) **393**, 440 (1998).

[36] P. Erdös and A. Rényi, Publ. Math. Debrecen **6**, 290 (1959).

[37] M.E.J. Newman, S.H. Strogatz, and D.J. Watts, Phys. Rev. E **64**, 026118 (2001).

[38] G.H. Wannier, Phys. Rev. **79**, 357 (1950); R.M.F. Houtappel, Physica (Amsterdam) **16**, 425 (1950).

[39] R. Smith, e-print cond-mat/0206378.

[40] H. Ebel, L.I. Mielsch, and S. Bornholdt, Phys. Rev. E **66**, 035103 (2002).

[41] M.E.J. Newman, Phys. Rev. E **64**, 016131 (2001).

[42] G.F. Davis, M. Yoo, and W.E. Baker, Strategic Organization **1**, 301 (2003).

[43] M.E.J. Newman, Phys. Rev. E **64**, 025102 (2001).

[44] W. Zachary, J. Anthropolog. Res. **33**, 452 (1977).

[45] J. MacRae, Sociometry **23**, 360 (1960).

[46] Actually the b_1 value for model 1 is 0.2% lower (around three standard deviations) for $r_1 = 0.1$ than for $r_1 = 0$. We will not speculate in the reason for this since the effect is small and the overall picture is clear.

[47] A potential improvement would be to measure b_1 and b_2 on the two-core (the maximal subgraph with minimal degree 2) of G. This would eliminate circuit-free subgraphs that contain no information about the degree of heterophilous preference among the agents forming the network.

[48] In the K_N, the number of circuits of length i is the i permutations $N!/(N-i)!$ divided by $2i$ (a factor i to compensate for the overcounting since a circuit is independent of starting vertex; a factor 2 to compensate for the double counting of the two directions). For K_N the contribution of circuits of length i to the sum is i times the number of them, this gives Eq. (A1a).

[49] J.E. Hopcroft and J.D. Ullman, *Introduction to Automata Theory, Languages and Computation* (Addison-Wesley, Reading, MA, 1979).