The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

Comparative Study on A Class of Evaluation Indices for Community Detection

Junhua Zhang* Shihua Zhang Xiang-Sun Zhang

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

Abstract Community detection and network partition are fundamental for uncovering the links between structure and function in complex networks. Recently Li et al. [11] introduced a novel quantitative function(D-value) for community detection which can overcome some drawbacks of the widely used modularity Q. We notice that although the modularity density D_λ has gained good performance for some networks, but how to determine a proper value of λ for any new network to be partitioned remains an open problem. This will certainly limit its further applications in practice to some extent. In this study, we propose a general form G of evaluation index for community detection from a perspective of intuition, and its three typical forms are given. The simplest one is the linear form G_L, which is just the D-value [11], the other two are the quadratic form G_Q and the entropy function form(or logarithmic form) G_E, respectively. By comparing the computational results on partitioning several real-world networks into communities we can conclude that G_Q is inefficient, but G_E is more powerful than G_L(i.e., D in [11]) to some extent. Moreover, the G_E can also overcome some drawbacks of Q, and it doesn't contain any parameters, so it is very convenient for using in practice.

Keywords Community detection; modularity; complex networks

1 Introduction

Many systems can be represented as networks composed of vertices and edges [1, 2, 3, 4]. For example, the Internet [5], social networks [6, 7], biological networks [8, 9] as well as the food webs [10] are all such systems. A common feature of many networks is "community structure" , which means the networks naturally decompose into groups, and within groups the connections are dense but between groups the connections are sparser [15, 16, 22, 21, 23].

Many community detection algorithms have been developed based on the optimization of a quantity called modularity Q introduced by Newman and Girvan [16], which is a quality index for a partition of a network into communities. In detail, given an undirected network $S(V,E)$ consisting of the node set V, the edge set E depicted by the symmetric adjacency matrix $A = [a_{ij}]_{n\times n}$, where $a_{ij} = 1$, if node i and node j are connected and

*Corresponding author: zjh@amt.ac.cn

otherwise $a_{ij} = 0$, n is the size of the network. The modularity function Q is defined as:

$$Q = \sum_{c=1}^{k} \left[\frac{l_c}{L} - \left(\frac{d_c}{2L} \right)^2 \right] \tag{1}$$

where the sum is over the k communities of the partition, l_c is the number of links inside community c, L is the total number of links in the network, and d_c is the total degree of the nodes in community c.

It's undoubted that maximization of the modularity Q over all the possible partitions of a network can provide, in many cases, a way to determine if a partition is valid to decipher the community structure in a network [24, 25, 26, 19], but the recent discoveries by some researchers tell us that we should pay more scrupulousness in using Q. In [12], Fortunato and BarthÍęlemy pointed out that modularity optimization may fail to identify modules smaller than a scale which depends on the total size of the network and on the degree of interconnectedness of the modules, even in cases where modules are unambiguously defined. Similar observations have also been raised in [13, 14]. To overcome this problem, recently a novel quantitative function(D-value) was introduced in [11]:

$$D = \sum_{c=1}^{k} \left[\frac{2l_c}{n_c} - \frac{t_c}{n_c} \right],$$

where t_c is the number of links between c and other communities, and n_c denotes the number of nodes in community c. We found that the D itself can sometimes decompose the network into small communities. In order to make the index work well, the authors in [11] must use the general modularity density D_λ:

$$D_\lambda = \sum_{c=1}^{k} \left[\frac{2\lambda \cdot 2l_c}{n_c} - \frac{2(1-\lambda) \cdot t_c}{n_c} \right], \quad 0 \le \lambda \le 1.$$

But how to determine a proper value of the parameter λ for a new network to be partitioned remains an open problem. Therefore, a more effective index is necessary for evaluating network partition in order to get more meaningful community structures.

In this study a general form of evaluation index G for community detection is proposed from a perspective of intuition, and its three typical forms are given. The simplest one is the linear form G_L, which is just the D-value [11], the other two are the quadratic form G_Q and the entropy function form(or logarithmic form) G_E, respectively. By comparing the computational results on partitioning several real-world networks into communities we can conclude that G_Q is inefficient, but G_E is more powerful than G_L(i.e., D in [11]) to some extent. The comprehensive comparison study between D and Q has been given in [11]. We notice that the performance of G_E in overcoming some drawbacks of Q is similar to D (for more details, please refer to Section 4). What is worthy to point out is that the new form G_E doesn't contain any parameters in it, so it is very convenient for using in practice.

2 A General Class of Evaluation Index for Community Detection

2.1 Presentation of the general evaluation index

Despite the profound meaning of the well known modularity Q in (1), which essentially measures the degree of correlation between the probability of having an edge joining two sites and the fact that the sites belong to the same community [15, 16], we can as well understand it from the intuitive point of view as follows. Because $l_c/L = 2l_c/(2L)$ represents the ratio of the inner degree of a community to the total degree of the network, we call it *the ratio of community inner degree to all*. similarly, we call $d_c/(2L)$ *the ratio of community degree to all*. It is easy to know $l_c/L \leq d_c/(2L) \leq 1$, so (1) means that, although the ratio of community inner degree to all is no more than the ratio of community degree to all, but it is required for a community that the square of the latter is less than the former.

Continuing using the notations above, set $d_c = 2l_c + t_c$. Because each community should be a connective subgraph, so the necessary condition is $(l_c + 1)/n_c \geq 1$. Furthermore,

$$1 \leq 2l_c/n_c \leq (2l_c + t_c)/n_c, \text{ for } l_c \geq 1 \tag{2}$$

($l_c = 0$ corresponds to a node with degree 1). Here $2l_c/n_c$ and $(2l_c + t_c)/n_c$ are called *community inner average density* and *community average density*, respectively. As mentioned above, a community is generally thought of as a part of a network where internal connections are denser than external ones. It is natural and reasonable to compare community inner average density with community average density. Although in general the relationship (2) holds for a connective subgraph, we imagine that a certain function $f(\cdot)$ of the former is larger than the latter for a community($f(\cdot)$ is called a *modulatory function*), with the goal to constrain the rapid increment of the outer degrees to some extent. Hence a general class of index for community detection is as follows:

$$G = \sum_{c=1}^{k} \left[f\left(\frac{2l_c}{n_c}\right) - \frac{d_c}{n_c} \right], \tag{3}$$

where we demand that $f(\cdot)$ satisfies

$$f(x) \geq x, \text{ for } x \geq 1. \tag{4}$$

Because for community detection our goal is to maximize G, thus in order to ensure each item of the sum in (3) is nonnegative, the relationship (4) must be satisfied.

2.2 Some concrete forms of the index

It is easy to understand that the main role of the modulatory function $f(\cdot)$ is to balance the inner average density of a subgraph and the average density of it. For any x, the larger the $f(x)$, the looser for the community, i.e., it allows of more outer degrees. To further investigate the application of the index G in detecting communities, here we give three concrete forms of it by indicating three typical forms of $f(\cdot)$. And in Section 3 we'll make some comparison studies on them.

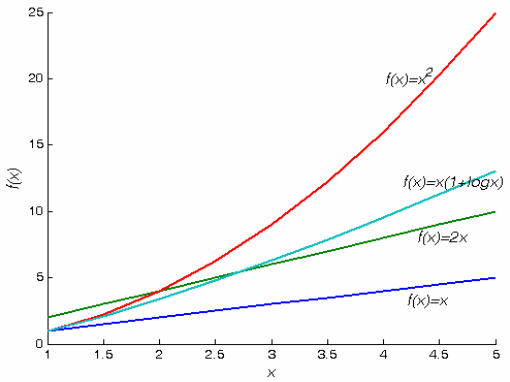

Figure 1: Comparison of three modulatory functions.

2.2.1 The linear form

According to (4), we can take $f(x) = 2x$. In this case (3) becomes

$$G_L = \sum_{c=1}^{k} \left[\frac{2l_c}{n_c} - \frac{t_c}{n_c} \right], \tag{5}$$

which is the D-value above recently carefully studied and discussed in [11].

2.2.2 The quadratic form

In this case we take $f(x) = x^2$, now (3) becomes

$$G_Q = \sum_{c=1}^{k} \left[\left(\frac{2l_c}{n_c} \right)^2 - \frac{2l_c + t_c}{n_c} \right]. \tag{6}$$

2.2.3 The logarithmic form

In this case we take $f(x) = x(1 + logx)$, and (3) becomes

$$G_E = \sum_{c=1}^{k} \left[\frac{2l_c}{n_c} \cdot \log \left(\frac{2l_c}{n_c} \right) - \frac{t_c}{n_c} \right]. \tag{7}$$

In view of the first term in the bracket of (7), we can also call G_E *the entropy function form*. In this paper we take *log* as the natural logarithm.

In order to easily understand the relationship of these three modulatory functions, please refer to Figure 1 for their graphs.

3 Experiments

We test the performance of the proposed indices here by applying them to several real-world networks. For comparison, for each data we give three partitions, corresponding to

G_L in (5), G_Q in (6) and G_E in (7), respectively. The algorithm for network partition we used here is genetic algorithm(GA) [17], the reason is that in most cases GA can search global optimum in possible solution domain.

It is known that there are some parameters need to be determined in using GA, such as the population size, the iteration count, as well as those related to the processes of cross-over, mutation, and clean-up. Here we choose these parameters according to [17]. For example, we use a value between 200 and 500 as iteration count and a value between 100 and 250 as population size. On the other hand, GA is a stochastic algorithm which reports different results with different initial solutions, here we select the best one through about 20 runs for each experiment.

3.1 The karate club network

The famous karate club network analyzed by Zachary [27] is widely used as a test example for methods of detecting communities in complex networks [15, 22, 30, 29, 28]. The network consists of 34 members of a karate club as nodes and 78 edges representing friendship between members of the club which was observed over a period of two years. Due to a disagreement between the club's administrator and the club's instructor, the club split into two smaller ones. The question we concern is that if we can uncover the potential behavior of the network, detect the two communities or multiple groups, and particularly identify which community a node belongs to.

Figure 2 shows the network and the corresponding community structure detected by G_L (in (a)), G_Q (in (b)) and G_E (in (c)). Maximizing either G_L or G_E can divide the network into three groups. The partition with G_L mislays one member (node 10) from one club to the other (see (a) of Figure 2). But using G_E, we can get completely consistent split with actual division of original club (see the thick curve in (c) of Figure 2), moreover, we can get more fine partitioning (the thin curve therein). At the same time, maximizing G_Q can divide the network into four groups (see (b) of Figure 2), it combines two unconnected nodes (10 and 12) each from one club into one community, this is meaningless. This indicates that the application of the index G_E to the empirically observed network not only can uncover its real situation, but also detect more complex substructure.

3.2 The scientific collaboration network

The scientific collaboration network collected by Girvan and Newman [15] is another widely used test example for methods of detecting communities in complex networks [15, 28]. This network consists of 118 nodes (scientists) and 200 edges. Maximizing G_L, the network is partitioned into 13 groups (see (a) of Figure 3). Among them we notice that there are two groups consisting of only two nodes, for which it is not proper to think them as communities. Maximizing G_Q, 9 groups are obtained (see (b) of Figure 3), but there is one group (with color blue) consisting of some unconnected nodes, this cannot also be as a community. Maximizing G_E, we can detect 11 communities (see (c) of Figure 3). Compared with the recent result in [19], which uses a new algorithm and the modularity Q, and 8 communities are detected, here these three additional communities are signed by curves 1, 2, and 3 respectively in (c) of Figure 3. Among them, the two communities with curve 1 and curve 2 are cliques, connected with other parts through fewer links. The

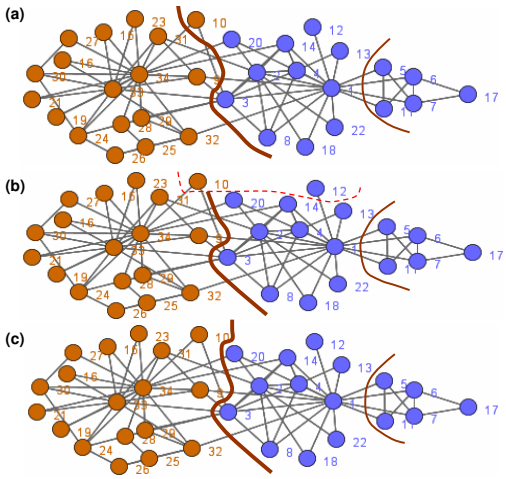

Figure 2: The community structure of the karate club network detected by G_L (in (a)), G_Q (in (b)) and G_E (in (c)).

one with curve 3 looks also more dense than outside. Anyway, all these 11 communities are visually quite reasonable.

3.3 The dolphin network

The dolphin social network reported by Lusseau et al. [20] and recently studied by Rosvall and Bergstrom [14] is also used here. This network consists of 62 nodes and 159 edges. The (o) of Figure 4 displays the division along which the actual dolphin groups were observed to split [20]. Maximizing G_Q, the network is partitioned into 5 groups (see (b) of Figure 4), with one group having only two nodes and another one (with color pink) consisting of unconnected nodes. Maximizing G_L and G_E, the network is partitioned into 4 and 3 groups, respectively (please refer to (a) and (c) of Figure 4), both of them can get completely consistent split with the actual division (see curve 1 therein). Furthermore, G_L splits the group with 42 nodes into three parts (dashed curves 2 and 3 in (a)), and G_E splits it into two (dashed curve 2 in (c)). Reminded of the results in [14], where the authors illustrated the partitions of the same dolphin network using four methods, i.e., their cluster-based compression, the edge-betweenness algorithm [15], the spectral analysis approximation [21], and maximizing the modularity Q, each of them split the network into two parts, but the first two methods mislaid one node, the third mislaid three nodes, and the fourth (i.e., maximizing Q) mislaid eight nodes. This indicates once more that the application of the index G_E not only can uncover the network's real situation, but also detect more complex substructure. Moreover, unlike G_L, which in some cases gives too many and too small groups, G_E can in general give more reasonable partitions.

4 Discussion

In this study, we propose a general form of evaluation index for community detection from a perspective of intuition, and three typical forms of it are given, i.e., the linear form

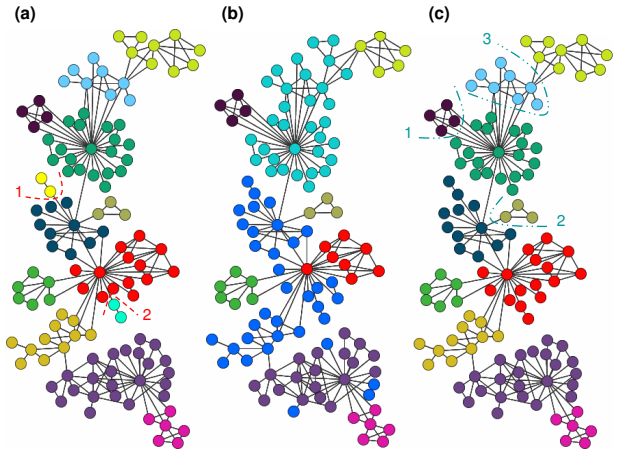

Figure 3: The community structure of scientific collaboration network obtained by G_L (in (a)), G_Q (in (b)) and G_E (in (c)).

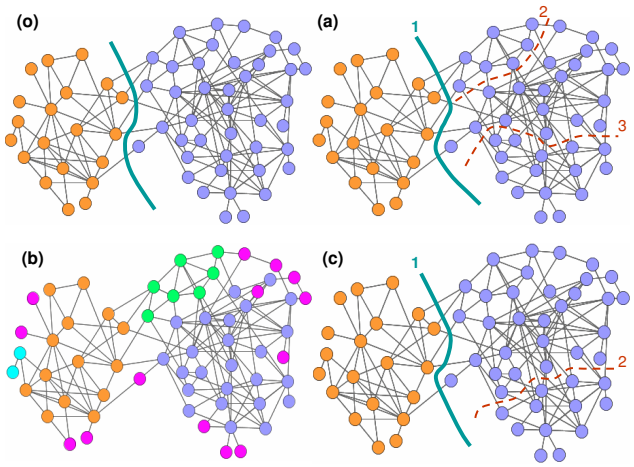

Figure 4: The original dolphin network (o) and the community structure obtained by G_L (in (a)), G_Q (in (b)) and G_E (in (c)).

G_L (it's in fact the D-value in [11]), the quadratic form G_Q, and the entropy function form G_E. For comparative study, we use these three indices to several real-world networks to detect communities. The conclusion is that in most cases G_Q is inefficient, and G_E can give more reasonable partitions than G_L.

The reason for the inefficiency of G_Q may attribute to the square term in (6), this is corresponding to the modulatory function $f(x) = x^2$. From Figure 1 we can see that when x becomes large x^2 goes up rapidly. Therefore in a partition of a network this may result in the phenomenon that some groups with very high $(2l_c/n_c)^2$ can compensate some other groups with unconnected nodes.

As stated in Section 2, the indices G_L and G_E are induced by $f(x) = 2x$ and $f(x) = x(1+logx)$, respectively. Comparing their corresponding graphs in Figure 1, we can find that for the natural logarithm log, if $x < e$, then $2x > x(1+logx)$; and if $x > e$, then $2x < x(1+logx)$. This means that for the partition of a network, when the community inner average density $2l_c/n_c$ is small G_L allows of more outer degrees than G_E, otherwise when $2l_c/n_c$ is large G_L allows of less outer degrees than G_E. We think that's why G_L detects two-node groups in the scientific collaboration network(Figure 3). We notice that the performance of G_E is accordant with the intuition that more outer degrees may be permitted when the community inner average density becomes larger. Furthermore, G_E ascends moderately, unlike G_Q going up rapidly, along with $2l_c/n_c$ becoming larger and larger. Well-tried results confirm G_E's ability as an evaluation index for partitioning networks into communities.

Moreover, like the D-value in [11], G_E can also overcome some resolution limits of the modularity Q. In [12], Fortunato and Barthĺelemy showed that Q contains an intrinsic scale that depends on the total number of links in the network, communities that are smaller than this scale may not be resolved. Typically, for the two schematic examples in Figure 5 with $\mathscr{A}$ consisting of a ring of n cliques (with n even) K_m connected through single links, and $\mathscr{B}$ consisting of two pairs of different sized cliques K_m's and K_p's ($p < m$) (here the clique K_m means a complete graph with m nodes and having $m(m-1)/2$ links), these networks have a clear modular structure where the communities correspond to single cliques, the authors demonstrated that in some cases (for instance, $m = 5$ and $n = 30$ for $\mathscr{A}$, and $m = 20$ and $p = 5$ for $\mathscr{B}$) maximizing the modularity Q would find the configuration with pairs of cliques (marked by dashed curves in Figure 5) rather than the actual communities.

Here if the index G_E is used, let $G_{E,single}$ and $G_{E,pairs}$ denote the partitions with single cliques and with pairs of them in Figure 5 $\mathscr{A}$, respectively, we have

$$\begin{aligned} G_{E,single} &= n\left[\frac{m(m-1)}{m}\ln\frac{m(m-1)}{m} - \frac{2}{m}\right] \\ &= n\left[(m-1)\ln(m-1) - \frac{2}{m}\right], \\ G_{E,pairs} &= \frac{n}{2}\left[\frac{2(m(m-1)+1)}{2m}\ln\frac{2(m(m-1)+1)}{2m} - \frac{2}{2m}\right] \\ &= \frac{n}{2}\left[\frac{m(m-1)+1}{m}\ln\frac{m(m-1)+1}{m} - \frac{1}{m}\right]. \end{aligned}$$

We can further prove that $G_{E,single} - G_{E,pairs} > 0$ always holds for any $m > 3$.

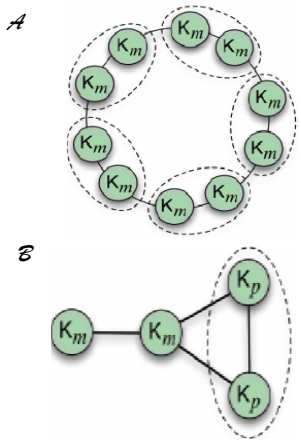

Figure 5: Schematic examples constructed by Fortunato and Barthĺelemy [12].

Similarly, for Figure 5 $\mathscr{B}$, let $G_{E,separate}$ and $G_{E,merge}$ respectively correspond to the partitions in which the two smaller cliques are separated and merged. Then

$$\begin{aligned} & G_{E,separate} - G_{E,merge} \\ = & -\frac{3}{p} + 2(p-1)\ln(p-1) - \frac{p(p-1)+1}{p}\ln\frac{p(p-1)+1}{p} \\ > & 0 \end{aligned}$$

can also be proved always correct for any $p > 3$.

Here it is important to point out that although some satisfying results are obtained in [11] for the general modularity density D_λ which depends on a parameter λ, but how to determine a proper value of λ for a new network to be partitioned retains an open problem.

The last thing about G_E to be mentioned is that through experiments we find that similar results can be obtained for natural logarithm and logarithm base 2, but if logarithm base 10 is taken, the partition prefers larger groups. So in this paper the natural logarithm is adopted for G_E in (7).

All these ensure that G_E can be an appropriate index for community detection.

5 Conclusion

In this paper, we get a new evaluation index G_E for community detection through comparative study. G_E can overcome some drawbacks of the widely used modularity Q, and it has good performance on partitioning several real-world networks. Although some other indices have been introduced in recent years [31, 32, 33, 11], we hope that our index G_E will be a helpful complementarity to this field. We expect that this new index will provide more promising results in the detection of communities in complex networks with practical significance.

Acknowledgments

This work was partly supported by the Ministry of Science and Technology, China, under Grant No. 2006CB503905, National Natural Science Foundation of China under Grant No. 10631070.

References

[1] S. H. Strogatz, Nature (London) 410, 268 (2001).

[2] R. Albertand A.-L. Barabĺćsi, Rev. Mod. Phys. 74, 47 (2002).

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

[4] S. N. Dorogovtsev and J. F. F. Mendes, Evolution of Networks: From Biological Nets to the Internet and WWW (Oxford University Press, Oxford, 2003).

[5] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. 29, 251 (1999).

[6] J. Scott, Social Network Analysis: A Handbook, 2nd ed. (Sage Publications, London, 2000).

[7] M. E. J. Newman and J. Park, Phys. Rev. E 68, 036122 (2003).

[8] E. Almaas, B. Kovĺćcs, T. Vicsek, Z. N. Oltvai, and A.-L. Barabĺćsi, Nature (London) 427, 839 (2004).

[9] F. Rao and A. Caflisch, J. Mol. Biol. 342, 299 (2004).

[10] J. A. Dunne, R. J. Williams, and N. D. Martinez, Proc. Natl. Acad. Sci. U.S.A. 99, 12917 (2002).

[11] Z. Li, S. Zhang, R.-S. Wang, X.-S. Zhang and L. Chen, Physical Review E, 77, 036109, (2008).

[12] S. Fortunato and M. Barthĺęlemy, Proc. Natl. Acad. Sci. USA 104 (1), 36-41 (2007).

[13] S. Muff, F. Rao and A. Caflisch, Phys. Rev. E 72, 056107 (2005).

[14] M. Rosvall and C.T. Bergstrom, Proc. Natl. Acad. Sci. USA 104, 7327-7331 (2007).

[15] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA 99, 7821-7826 (2002).

[16] M. E. J. Newman and M. Girvan, Phys. Rev. E 69, 026113 (2004).

[17] M. Tasgin and H. Bingol, arXiv.org: cond-mat/0604419 (2006).

[18] Thompson Scientific, Journal Citation Reports (Thompson Scientific, Philadelphia) (2004).

[19] J. Zhang, S. Zhang and X.-S. Zhang, Physica A, 387(7), 1675-1682 (2008).

[20] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten and S.M. Dawson, Behav Ecol Sociobiol 54, 396-405 (2003).

[21] M. E. J. Newman, Phys. Rev. E 74, 036104 (2006).

[22] M. E. J. Newman, Eur. Phys. J. B 38, 321-330 (2004).

[23] M. E. J. Newman, Proc. Natl Acad. Sci. USA 103(23), 8577-8582 (2006).

[24] R. Guimera and L. A. N. Amaral, Nature 433, 895-900 (2005).

[25] R. Guimera and L. A. N. Amaral, J. Stat. Mech. Theor. Exp. P02001 (2005).

[26] R. Guimera, M. Sales-Pardo and L. A. N. Amaral, Nature Physics 3, 63-69 (2007).

[27] W. W. Zachary, J. Anthropol.Res. 33, 452-473 (1977).

[28] F. Wu and B. A. Huberman, Eur. Phys. J. B 38, 331-338 (2004).