Identifying Community Structures from Network Data

Jernej Čopič, Matthew O. Jackson, and Alan Kirman [*]

Draft: October 16, 2005

Abstract

We describe, axiomatize, and implement, a new method for identifying community structures (groups of structurally equivalent nodes) from network data. The method is based on maximum likelihood estimation, a standard statistical tool.

---

[*] Čopič and Jackson are at the Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, California 91125, USA, emails: jacksonm@hss.caltech.edu and jernej@hss.caltech.edu, web sites: http://www.hss.caltech.edu/∼jacksonm/Jackson.html and http://www.hss.caltech.edu/∼copic. Kirman is at GREQAM, Université de la Mediterranée, 2 rue de la Charité, 13236 Marseille CEDEX 02, France, email: kirman@ehess.cnrs-mrs.fr . We gratefully acknowledge financial support under NSF grant SES-0316493, as well as from the Lee Center for Advanced Networking and the Social Information Sciences Laboratory at Caltech. We thank Matteo Marsili for helpful conversations and Rik Pieters and Hans Baumgartner for making their data available.

1 Introduction

In many networks there are groups of nodes that are similar in nature. If nodes are people, they may belong to the same club, be of the same ethnicity, profession or have other possibly unobservable characteristics in common. Based on the network interactions, we can infer which nodes are similar and group them into communities of nodes which are similar or equivalent according to how they interact in the network. There are at least two reasons for wanting to partition networks into groups of nodes. First, this provides a simplification of the network, where now one has derived a smaller network between groups of "equivalent" objects. Second, this is a way of using network data to infer the equivalence and to categorize nodes or objects. For instance, as an application based on cross citations among journals, we can learn which journals should be grouped into communities based on how they cite each other. This can be useful in, among other things, developing rankings of journals.

Given the importance of partitioning networks and deriving community structures among nodes, there is a rich literature proposing a variety methods for doing so. This began with definitions of structural equivalence, defining equivalence between nodes based on their relationships to other was identical (see the seminal work on block modeling and positional analysis by Lorrain and White (1971) and White, Boorman and Breiger (1976)). As more applications arose, a burgeoning set of algorithms for partitioning networks has emerged (e.g., see Newman (2003) for a brief but very useful overview). These include methods that examine correlations of relationships between nodes (e.g., CONCOR)[1]; methods based on (repeated) bisection of the network (e.g., Kernighan and Lin (1970), Fiedler (1973), and Pothen, Simon and Liou (1990)); methods based on building hierarchies of similar nodes (for instance starting with all nodes separate and then matching the two most similar according to various measures, and then iterating upwards (e.g., Lance and Williams (1967)), methods based on identifying cliques (Bron and Kerbosch (1973)), and methods based on edge removal based on betweeness measures (Girvan and Newman (2002), Tyler, Wilkinson and Huberman (2003)) or loops (Radicchi et al (2003)).

As one might expect with such a variety of techniques, different methods can end up producing very different partitions from the same network data. This obviously means that the methods are identifying different things. Without some systematic study of them, it can

---

[1] See White, Boorman and Breiger (1976) and Boorman and White (1976). Such methods examine correlations of the adjacency matrix, so that two nodes are highly correlated if their relationships with other nodes are very similar, and then try to partition the nodes based on some algorithm that makes use of the correlation structure. CONCOR iterates this procedure, performing correlations on the correlation matrix, etc., until the matrix converges to 1's and -1's. See Schwartz (1977) for some discussion of the properties.

be difficult to know which method to use for any given problem, or even to know exactly what the resulting community structure means. While the algorithms are formal, clever, and at times complex, identifying community structures is still more of an art than a science.

In this paper, we introduce a method of identifying community structures based on the well-founded statistical method of maximum likelihood estimation. Moreover, the method not only identifies a "best" or "highest likelihood" community structure, it also provides a complete ranking of community structures. In fact, the maximum likelihood method provides explicit relative likelihoods of different community structures rather than just identifying a single "best" structure. Perhaps most importantly from the scientific perspective, we provide axioms that uniquely identify maximum likelihood as a method of identifying community structures. This then gives a complete characterization of the method in terms of its properties. If such characterizations are eventually provided for other methods, we will have much better understandings of the contrasts between, and relative strengths and weaknesses of different approaches.

It is critical to emphasize that we approach the problem of identifying community structures from a very different perspective than the previous literature. Rather than starting with the proposal of an algorithm based on some idea of what we wish to identify, we start with a model of what a community structure is and how it generates networks. Having such a model then suggests exactly how one should identify community structures. Have such a tie between a model of how network data is generated and a method of identifying community structures is a crucial step in understanding the foundations behind community structure identification techniques.

Before proceeding, let us describe the basic ideas behind the likelihood approach and the model on which it is based. We describe this for a special case where links are either present between two nodes or not, and leave the description of weighted and/or directed networks for the formal presentation. The model starts with a given set of nodes. There is some true underlying community structure which is a partition of the nodes into groups. Links between nodes are then formed at random, but in a way that depends on the underlying community structure.[2] The key is that links between nodes within a given community are more likely to form than links between nodes that lie in different communities. In particular, there is some probability p_{in} of a given link forming between members of the same community, and

---

[2] The randomness here can come from two different sources. It might be stochastics directly in the interaction patterns. Or instead, it could be that the interaction patterns are deterministic, but that there is measurement error so that randomness is introduced through the observation by the scientist. In most cases, it will actually be that both forms of stochastics are present. As is usual in statistical analysis, distinguishing between these two can be important. For our analysis here, the distinction is inconsequential.

another probability p_{out} of a link forming between members of different communities. For any given p_{in} and p_{out}, where $1 \geq p_{in} > p_{out} \geq 0$, one can calculate the probability of a given network g being observed under a given community structure. Thus, for each possible community structure, there is a well-defined probability of observing the given network data that was observed. Maximum likelihood does exactly what it says, namely it selects the community structure that led to the maximum probability of observing the data that was in fact observed. This method not only provides one with a "best" or most likely community structure, but also provides a complete ordering over all community structures and in fact provides a relative likelihood that any one could have generated the data.[3]

In terms of implementing this method, the challenges are twofold. First, the p_{in} and p_{out} described above are not known to the observer. This is easily overcome as these can be estimated from the data itself. There are some issues related to the fact that these estimates depend on the community structure identified, but we show that there is a nice consistency property that leads to accurate estimation of these parameters. Second, a challenge faced by any method for identifying community structures is that the number of possible partitions into communities is exponential in the number of nodes. Searching across all possible partitions then becomes impossible with even a moderate number of nodes. This means that some sort of approximation algorithm needs to be used. Most previous methods are simply defined directly in terms of their algorithms, and then if the algorithm is a polynomial time algorithm it will generally be feasible to practically implement. This is more of an issue for us, as instead of defining the method by specifying an ad hoc algorithm, we have defined it by a model of how networks are randomly generated. Thus, we need to find a practical algorithm for deducing the structure from the observed network. Here we show that one can adapt algorithms from various related problems that operate in polynomial time and provide nearly optimal (according to the maximum likelihood ordering) community structures.

Given the prominence of likelihood techniques in statistical analysis, there is much that can be drawn on. The most closely related work to ours is an important application of maximum likelihood techniques to the clustering of Potts variables by Giardi and Marsili (2005). The Potts model comes from statistical physics where Potts spins within groups are correlated but across groups are not. Giardi and Marsili explore grouping objects based on correlation patterns in the observed attribute vectors, based on an underlying Gaussian model and maximum likelihood. As Giardi and Marsili point out, this technique can be applied to settings beyond physics, such as to the stock market, where returns of stocks

---

[3] It is then also straightforward to develop statistical measures of fit, confidence, and significance, through the standard tools associated with likelihood estimation.

within groups are presumed correlated but across groups are not. Our analysis is similar in spirit to that of Giardi and Marsili in terms of using maximum likelihood on a data set to uncover an underlying grouping of objects. However, our model and applications are different.[4] Moreover, our axiomatization of a maximum likelihood method is the first of which we are aware.

The paper proceeds as follows. We first define the likelihood approach, and also describe orderings of community structures. Next, we provide axioms which uniquely characterize likelihood techniques as a method of identifying and ordering community structures. Then, we discuss the large sample statistical properties of the likelihood methods and possible algorithms for implementing the approach. Finally, we illustrate the method with an application to a network of citations among economics journals.

2 Definitions

2.1 Networks and Sizes

Let $N = \{1, 2, ..., n\}$ be a set of *nodes* or vertices.

A *(weighted and directed) network* on a set of nodes N is matrix $g \in I\!R^{n \times n}$.

We allow networks to be weighted and/or directed, although our methods also apply equally well to special cases where the network is undirected and/or unweighted.

A set of network sizes or capacities on a set of nodes N is a matrix $s \in I\!R^{n \times n}$.

A network g on a set of nodes N with sizes s is *feasible* if

$$g_{ij} \leq s_{ij}$$

for each ij.

Let $G(s)$ denote the set of feasible networks on nodes N given a size matrix s.

A special case of sizes is where $s_{ij} = s_i s_j$, with the interpretation that node i is composed of s_i units, and that each unit in every node can have at most one directed connection with every other unit in every other node. For instance, consider an application where the network is a set of citations among journals during some time period, where each node is a journal. As any article can cite any other article at most once, the most that journal i can

---

[4] For instance, they examine situations where objects have correlated vectors of attributes, while we examine situations where there are direct relationships between objects where the probability structure of the relationships depends on the groupings.

cite journal j is $s_i s_j$, where s_i is the number of articles in journal i during the time period and s_j is the number of articles appearing in journal j during the time period.

As another example, consider a network of coauthorships, where each node is an individual. Here one possibility is that $s_{ij} = 1$ for all ij, as there is a potential for two nodes to be co-authors. This would apply if the network simply keeps track of whether two individuals have ever written a paper together. If instead, the network keeps track of how many papers two individuals have written together over some time period, then s_{ij} would be the potential capacity of papers that could be written involving individuals i and j during the time period.

2.2 Community Structures

A community structure is a partition of the set of nodes.

Let $\Pi(N)$ be the set of all partitions of N.[5]

For any $\pi \in \Pi(N)$ and any $i \in N$, let $c_\pi(i)$ be the component of π containing i.

2.3 Orderings over Community Structures

We now present a new concept to the community structure literature, that of an ordering over community structures. This comes from our viewpoint on community structures. While there may be some "true" or "best" community structure, at least hypothetically, the network data that are observed will usually be generated with some inherent randomness and/or will be subject to measurement error. Thus, in general, the network that we observe as scientists may only allow us to imperfectly infer an underlying community structure. As such, it generally makes sense rather than to try to uncover a single "best" community structure, to instead be able to rank all possible community structures. The interpretation being that higher ones in the rankings look more likely to be the "true" underlying community structure that generated the data; while it can still be possible that lower ones in the rankings could turn out to be the "true" one.

In fact, when we turn to the maximum likelihood method below, it will not only allow us to rank the possible community structures, but will also tell us the relative likelihood with which each one is the true community structure! This is one very attractive by-product of the approach we take.

Given N, let R be the set of all functions that map the set of possible feasible networks on N into the set of weak orders on $\Pi(N)$.

---

[5] Thus an element $\pi \in \Pi(N)$ is a collection of subsets of N such that $\cup_{c \in \pi} c = N$ and if $c, c' \in \pi$ and $c \neq c'$ then $c \cap c' = \emptyset$.

We let $\succeq$ denote a generic element of R Thus, $\succeq_{s,g}$ is a weak ordering over the community structures on the set of nodes N with sizes described by s having observed the data of of the network $g \in G(s)$.

We call the set R the set of *community structure orderings*.

The interpretation is that given some observed network g, $\pi \succeq_{s,g} \pi'$ if π is somehow a (weakly) "better" or more likely partitioning of the nodes of N according to a criterion that is embodied in $\succeq$.

Let $\succ_{s,g}$ denote the associated strict relationship associated with $\succeq_{s,g}$, so that $\pi \succ_{s,g} \pi'$ if and only if *not* $\pi' \succeq_{s,g} \pi$.

3 The Maximum Likelihood Method

We now present the maximum likelihood method for identifying community structures.

A community is a group of similar nodes in terms of their probabilities of interaction with other nodes. The probability of any two given nodes inside a community interacting is p_{in} and the probability of two given nodes from different communities interacting is p_{out}, where $1 \geq p_{in} > p_{out} \geq 0$.

If s_{ij} is the potential size of the interaction between i and j, then the chance of seeing exactly g_{ij} interactions between i and j is

$$p_{in}^{g_{ij}} (1 - p_{in})^{s_{ij} - g_{ij}}$$

if i and j are in the same community, and

$$p_{out}^{g_{ij}} (1 - p_{out})^{s_{ij} - g_{ij}}$$

if i and j are in different communities.

From this, we can calculate what would be the probability of observing any given $g \in G(s)$ if some $\pi \in \Pi$ were the true community structure. This likelihood is[6]

$$L_{s,g}(\pi) = \times_{i \in N} \left[\left(\times_{j \in c_\pi(i)} (p_{in})^{g_{ij}} (1 - p_{in})^{s_{ij} - g_{ij}} \right) \left(\times_{j \in N \setminus c_\pi(i)} (p_{out})^{g_{ij}} (1 - p_{out})^{s_{ij} - g_{ij}} \right) \right]. \quad (1)$$

Under a maximum likelihood method, the partitions are simply ranked according to the likelihood that they generate. That is, we rank π as being more likely than π' given the sizes s and observed network g, if $L_{s,g}(\pi) > L_{s,g}(\pi')$. This provides a maximum likelihood ordering, $\succeq_{s,g}^{ML(p_{in}, p_{out})}$, on $\Pi(N)$.

---

[6] We adopt the convention that $0^0 = 1$.

Moreover, given the relative likelihood function L, we not only have an ordering, but we actually have relative likelihoods. In addition to ranking community structures, this will also allow us to say how close two community structures are to each other in a very precise sense.

It is important to note that this maximum likelihood method, as described up to this point, presumes knowledge of p_{in} and p_{out}. These can also be estimated from the data, as we describe in our section on estimation.

3.1 Alternative representations of the Likelihood Function

Before proceeding further, let us rewrite the likelihood function a bit, in ways that make it very clear that the maximum likelihood ordering has some simple and attractive properties.

Given $\pi \in \Pi$, let

$$In(\pi) = \{ij \mid i \in N, j \in c_\pi(i)\},$$

and

$$Out(\pi) = \{ij \mid i \in N, j \notin c_\pi(i)\}.$$

Thus, $In(\pi)$ is the set of all pairs of links that are in the same component of the partition π and $Out(\pi)$ is the set of all pairs of links that are in different components under the partition π. Let

$$T(g) = \sum_{ij \in N \times N} g_{ij}$$

and similarly define $T(s)$. These are the total weighted links in the network g and the total sizes, given s. Let

$$T^{In(\pi)}(g) = \sum_{ij \in In(\pi)} g_{ij}$$

and similarly define $T^{In(\pi)}(s)$, $T^{Out(\pi)}(g)$, and $T^{Out(\pi)}(s)$.

By taking logs of L, we still preserve the ordering over partitions, so let $\ell_{s,g}(\pi) = log(L_{s,g}(\pi))$. We can then write

$$\ell_{s,g}(\pi) = log(L_{s,g}(\pi)) = k_1 T^{In(\pi)}(g) + k_2 T^{In(\pi)}(s) + k_3 T^{Out(\pi)}(g) + k_4 T^{Out(\pi)}(s), \qquad (2)$$

where $k_1 = log(p_{in}/(1 - p_{in}))$, $k_2 = log(1 - p_{in})$, $k_3 = log(p_{out}/(1 - p_{out}))$, and $k_4 = log(1 - p_{out})$.

This representation of the log-likelihood makes it clear that the maximum likelihood method only depends on the total number of "inside connections" relative to their total capacity, and the total number of "outside connections" relative to their total capacity. Thus, this method works by counting total numbers of (weighted) connections and capacities inside and the same for outside, and it is the comparison across "ins" and "outs" that matters, but not how they are spread within the "ins" or "outs."

Noting that $k_1 > k_3$ but $k_2 < k_4$, we see why an optimum will not always be an extreme partition where all nodes are either grouped together or all nodes are grouped apart. As more nodes are grouped together, things are shifted from $T^{Out(\pi)}(g)$ to $T^{In(\pi)}(g)$, where the second one hits a higher factor, however things are also shifted from $T^{Out(\pi)}(s)$ to $T^{In(\pi)}(s)$, where the second one now hits a lower factor. It is this tradeoff that determines the optimal partition.

Since we can write $T^{out}(\cdot) = T(\cdot) - T^{in}(\cdot)$, we can rewrite the expression above as

$$\ell_{s,g}(\pi) = (k_1 - k_3)T^{In(\pi)}(g) + (k_2 - k_4)T^{In(\pi)}(s) + k_3T(g) + k_4T(s). \tag{3}$$

Thus, if we compare two partitions π and π', we have $\pi \succ_{s,g}^{ML(p_{in},p_{out})} \pi'$ if and only if

$$(k_1 - k_3)T^{In(\pi)}(g) + (k_2 - k_4)T^{In(\pi)}(s) > (k_1 - k_3)T^{In(\pi')}(g) + (k_2 - k_4)T^{In(\pi')}(s). \tag{4}$$

4 Axioms on Community Structure Orderings

Based on the simple representations of the maximum likelihood ordering over community structures, we can derive some basic properties that characterize the maximum likelihood method for ordering community structures.

The first axiom is a very simple one that implies that we are maximizing something rather than minimizing it.

AXIOM **1** WEAK MONOTONICITY. *Consider (g, s), $i, j \in N$, and π, π' such that $ij \in In(\pi) \cap Out(\pi')$ and such that $g_{kl} = s_{kl} = 0$ for all $kl \neq ij$.*

- *If $0 < g_{ij} = s_{ij}$ then $\pi \succ_{s,g} \pi'$,*
- *if $0 = g_{ij} < s_{ij}$ then $\pi' \succ_{s,g} \pi$, and*
- *if $0 = g_{ij} = s_{ij}$ then $\pi \sim_{s,g} \pi'$.*

The essence of the axiom is that we are just comparing two node networks, as $g_{kl} = s_{kl} = 0$ for all $kl \neq ij$. In this light, it says that if the amount of interaction is equal to the capacity, then a partition that groups the two nodes together is preferred to one that has them apart; while if the amount of interaction is 0 then it is better to have the nodes in separate communities than together. The last part simply says that if all capacities are 0, then we cannot order partitions so all of them are indifferent to each other.

The next axiom, independence, states that the rankings between two partitions does not depend on the links that are classified similarly across the two partitions (either both "in" or both "out").

Axiom **2** Independence. *Consider s and $g \in G(s)$. An ordering $\succeq \in R$ satisfies independence if*

$$\pi \succeq_{s,g} \pi' \iff \pi \succeq_{s',g'} \pi', \tag{5}$$

for every s, g, s' and g' such that $g_{ij} \neq g'_{ij}$ or $s_{ij} \neq s'_{ij}$ implies that $ij \in In(\pi) \cap In(\pi')$ or $ij \in Out(\pi) \cap Out(\pi')$.

This condition is clearly satisfied by the maximum likelihood as we saw in (2) that only the totals of ins and outs matter, and so interactions that enter in similar ways in both partitions are irrelevant to the ordering, we only need to keep track of which interactions change as we change partitions.

The next axiom, Neutrality, while a bit tedious to write out, is quite simple. It states that we can linearly rearrange errors. That is, take two partitions π, π', and take $i, j, k, l \in N$. SN says that if under π, i, j are in the same community, and k, l are in different communities, while under π', i, j are in different communities and k, l are in the same community, then if we increase the number of links between i and j by $x > 0$ and increase the number of links between k and l by x, the ranking between π and π' doesn't change. Similarly, the ranking shouldn't change if instead of increasing, we decrease the number of links between i, j and k, l.

Axiom **3** Neutrality *Consider s, $g, g' \in G(s)$, $i, j, k, l \in N$, $x > 0$ and $\pi, \pi' \in \Pi(N)$.*

1. *If $ij \in In(\pi) \cap Out(\pi')$ and $kl \in Out(\pi) \cap In(\pi')$, and $g_{ij} = g'_{ij} + x$, $g_{kl} = g'_{kl} + x$, and $g_{hm} = g'_{hm}$ for $hm \notin \{ij, kl\}$, then*

$$\pi \succeq_{s,g'} \pi' \iff \pi \succeq_{s,g} \pi'.$$

Similarly, if $s_{ij} = s'_{ij} + x$, $s_{kl} = s'_{kl} + x$, *and* $s_{hm} = s'_{hm}$ *for* $hm \notin \{ij, kl\}$, *then*

$$\pi \succeq_{s',g} \pi' \iff \pi \succeq_{s,g} \pi'.$$

2. *If* $ij \in In(\pi) \cap Out(\pi')$ *and* $kl \in In(\pi) \cap Out(\pi')$, *and* $g_{ij} = g'_{ij} + x$, $g_{kl} = g'_{kl} - x$, *and* $g_{hm} = g'_{hm}$ *for* $hm \notin \{ij, kl\}$, *then*

$$\pi \succeq_{s,g'} \pi' \iff \pi \succeq_{s,g} \pi'.$$

Similarly, if $s_{ij} = s'_{ij} + x$, $s_{kl} = s'_{kl} - x$, *and* $s_{hm} = s'_{hm}$ *for* $hm \notin \{ij, kl\}$, *then*

$$\pi \succeq_{s',g} \pi' \iff \pi \succeq_{s,g} \pi'.$$

This neutrality axiom is a key one in identifying the maximum likelihood method, as it really is the essence of the fact that we do not care where interactions in the "in" or "out" occur, but just the total numbers of them.

The last axiom, Scaling, implies that if we rescale the networks, say by multiplying all of the network interactions and sizes by a constant, then we won't change the ordering over community structures.

AXIOM **4** SCALING *There exists* $\gamma > 0$ *such that whenever there are* g, s, g', s', ij, *and* x *such that* $g'_{ij} = g_{ij} + x$ *and* $s'_{ij} = s_{ij} + \gamma x$, *and* $g'_{kl}, s'_{kl} = g_{kl}, s_{kl}$ *otherwise, then* $\pi \succeq_{s,g} \pi'$ *if and only if* $\pi \succeq_{s',g'} \pi'$.

4.1 Two Characterization Theorems

We can now state the following characterization theorem.

THEOREM **1** *A community structure ordering* $\succeq$ *satisfies monotonicity, independence, neutrality, and scaling, if and only if there exist* $p_{in}, p_{out} \in (0, 1)$ *such that* $\succeq$ *is a maximum likelihood-based ordering with probabilities* p_{in} *and* p_{out}.

In terms of proving Theorem 1, it is useful to prove an auxiliary theorem, which replaces independence and neutrality with another condition.

Let

$$D(g, \pi, \pi') = T^{In(\pi)}(g) - T^{In(\pi')}(g)$$

and similarly let

$$D(s, \pi, \pi') = T^{In(\pi)}(s) - T^{In(\pi')}(s)$$

Axiom **5** Internal Differences *If there are* (g,s), (g',s'), π,π',π'',π''' *such that*

$$D(g,\pi,\pi') = D(g',\pi'',\pi''')$$

and

$$D(s,\pi,\pi') = D(s',\pi'',\pi''')$$

then $\pi \succeq_{s,g} \pi'$ *if and only if* $\pi'' \succeq_{s',g'} \pi'''$.

Theorem **2** *A community structure ordering* $\succeq$ *satisfies monotonicity, internal differences, and scaling, if and only there exist* $p_{in}, p_{out} \in (0,1)$ *such that* $\succeq$ *is a maximum likelihood-based ordering with probabilities* p_{in} *and* p_{out}.

The proof of both theorems appears in the appendix.

It is natural to ask whether the above characterizations of the Maximum- Likelihood ordering are tight. It is clear that scaling and neutrality imply a form of linearity. Relaxing these will yield other orderings with non-linear functional forms. Relaxing monotonicity, we would get for example a “Minimum- Likelihood" ordering as well as anything in between the two. An example of a ordering that doesn’t satisfy independence is one similar to a likelihood ordering, but where the comparison between each two partitions depends on the number of common components.

5 Communities and Consolidation

One of the natural interpretations of communities is that they are collections of “equivalent” nodes. How does one formalize this notion? One way to formalize the notion is to say that if we have a community of a given set of nodes, and we combine those nodes to be one larger node, then the overall structure between communities would be preserved. We formalize this as follows.

Consolidations

We call a situation where we combine two nodes to become one a consolidation. The formal definition is long, but simple and straightforward.

Consider N, s, and $g \in G(N,s)$, and some $i, j \in N$. Define the *consolidation of* g *by combining* i *into* j, $g^{(i\sim j)} \in G(N \setminus i, s^{(i\sim j)})$, as follows.

First, the sizes associated with the new node are simply the sum of the sizes of the combined nodes, while other nodes keep their sizes.

$$s^{(i\sim j)}_{jk} = s_{ik} + s_{jk},$$

$$s^{(i\sim j)}_{kj} = s_{ki} + s_{kj},$$

$$s^{(i\sim j)}_{jj} = s_{ii} + s_{jj} + s_{ji} + s_{ij},$$

$$s^{(i\sim j)}_{kl} = s_{kl}$$

if $\{k,l\} \cap \{i,j\} = \emptyset$.

Next, the connections of the new node are simply the sum of the connections the previous nodes had

$$g^{(i\sim j)}_{jk} = g_{ik} + g_{jk},$$

$$g^{(i\sim j)}_{kj} = g_{ki} + g_{kj},$$

for $k \notin i, j$ and

$$g^{(i\sim j)}_{jj} = g_{ii} + g_{jj} + g_{ji} + g_{ij},$$

while the connections among other nodes remains unchanged:

$$g^{(i\sim j)}_{kl} = g_{kl},$$

if $j \notin \{k,l\}$.

Similarly, when we combine nodes, we have a corresponding collapsing of a community structure. Given $\pi \in \Pi(N)$ such that $j \in c_\pi(i)$, let $\pi^{(i\sim j)} \in \Pi(N \setminus i)$ be defined by

$$c_{\pi^{(i\sim j)}}(j) = c_\pi(j) \setminus i$$

and

$$c_{\pi^{(i\sim j)}}(k) = c_\pi(k)$$

for $k \notin c_\pi(j)$.

The next axiom is *consolidation*, and it imposes that a community of communities is also a community.

Axiom **6** Consolidation. *A community structure ordering $\succeq$ satisfies consolidation if for every (N,s), $g \in G(N,s)$ and consolidation of g combining i into j, $g^{(i\sim j)} \in G(N \setminus i, s^{(i\sim j)})$,*

$$\pi \succeq_{N,s,g} \pi' \iff \pi^{(i\sim j)} \succeq_{N\setminus i, s^{(i\sim j)}, g^{(i\sim j)}} (\pi')^{(i\sim j)}$$

for every $\pi \in \Pi(N)$ and $\pi' \in \Pi(N)$ such that $i \in c_\pi(j)$ and $i \in c_{\pi'}(j)$.

LEMMA **1** *For any given p_{in} and $p_{out} \in [0,1]$ (with $p_{in} \geq p_{out}$), the corresponding maximum likelihood ordering satisfies consolidation.*

As the proof follows directly from (4), we omit it.

6 Implementation and Estimation

In this section we present an application of the likelihood-based ordering $\succ_L$.

In order to apply the method, we have to resolve several estimation issues.

6.1 Estimation of Probabilities

First, the probabilities p_{in} and p_{out} are often unknown and have to be estimated. In general, given any partition π we can directly derive estimates $\hat{p}_{in}(\pi), \hat{p}_{out}(\pi)$ depend on the partition π, from which they are estimated. Formally, let

$$\hat{p}_{in}(\pi) = \frac{T^{In(\pi)}(g)}{T^{In(\pi)}(s)}, \text{ for any } \pi \in \Pi(N) \setminus \{\pi_d\}, \text{ and similarly,}$$

$$\hat{p}_{out}(\pi) = \frac{T^{Out(\pi)}(g)}{T^{Out(\pi)}(s)}, \text{ for any } \pi \in \Pi(N) \setminus \{\pi_t\},$$

where π_t is the trivial partition (with all nodes grouped together in one element) and π_d is the discrete partition (with all nodes being their own elements). Similarly, let

$$\hat{p}_{in}(\pi_d) = \max_{ij \in N \times N} \frac{g_{ij}}{s_{ij}} \text{ and } \hat{p}_{out}(\pi_t) = \min_{ij \in N \times N} \frac{g_{ij}}{s_{ij}}.$$

An obvious requirement is then that the partition π be optimal under $\hat{p}_{in}(\pi), \hat{p}_{out}(\pi)$. Call this requirement *Consistent Optimality.* However, there hypothetically could exist more than one set of a partition and estimated probabilities, satisfying this requirement.

Pinning down the optimal partition uniquely is possible in the limit, as our data sample becomes large. In the following Proposition we show that if the the sizes become large, then the estimation is consistent in the limit and there exist a unique optimal $(\pi, \hat{p}_{in}(\pi), \hat{p}_{out}(\pi))$. Proving that the true partition satisfies consistent optimality is a trivial consequence of the law of large numbers. But proving that no other partition satisfies consistent optimality is a little less obvious. It follows from the following Proposition.

PROPOSITION **1** *Let $g \in G(s)$, let $\bar{\pi} \in \Pi(N)$ be the true partition, and $p_{in}, p_{out} \in (0,1), p_{in} > p_{out}$, the true probabilities. Let s be such that $\frac{g_{ij}}{s_{ij}} = p_{in}, \forall ij \in In(\pi)$, and $\frac{g_{ij}}{s_{ij}} = p_{out}, \forall ij \in Out(\pi)$. Then $\pi \in \Pi(N)$ satisfies CO if and only if $\pi = \bar{\pi}$.*

The law of large numbers implies the following corollary.

Corollary **1** *Let N be fixed, let for each $t = 1, 2, ...,$ s^t be a matrix of sizes, $s^t \in I\!R^{n\times n}$, and such that $\lim_{t\to\infty} s^t_{ij} = \infty, \forall ij \in N \times N$. For each $t = 1, ...,$ assume that a network $g^t \in G(s^t)$ is generated according to $p_{in} > p_{out}$, and given some true partition of the nodes π. As $t \to \infty$, with probability one, there exists a unique $\hat{\pi}$, which is optimal according to the ordering $\succ_{L(\hat{p}_{in}(\hat{\pi}),\hat{p}_{out}(\hat{\pi}))}$, and $\hat{\pi} \to \pi$, $(\hat{p}_{in}(\hat{\pi}), \hat{p}_{out}(\hat{\pi})) \to (p_{in}, p_{out})$.*

6.2 Algorithms for Implementing the Likelihood Method

The maximum likelihood method allows us to use a variety of algorithms for finding a partition that is approximately optimal.

In order to understand how to adapt algorithms to this problem, it is helpful to simplify the problem. In particular, we show the following result: If two partitions are close to each other according to a simple metric that counts how many objects are in different groups across the two partitions, then the likelihoods of the two partitions are also close to each other. This is useful, since then if we can examine a grid of partitions and identify the likelihood of each element of the grid, we can bound the likelihood of any partition that we do not directly examine by knowing that is close to something on the grid. By being sure to search a fine enough grid, we are sure to find a community structure close to the maximum likelihood structure.

Recall that

$$\ell_{s,g}(\pi) = log(L_{s,g})(\pi) = k_1 T^{In(\pi)}(g) + k_2 T^{In(\pi)}(s) + k_3 T^{Out(\pi)}(g) + k_4 T^{Out(\pi)}(s), \qquad (6)$$

Let us consider a situation where $1 > p_{in} > p_{out} > 0$, and all sizes s_{ij} are bounded above by s_1, and the true partition is given by $\widehat{\pi}$.

As the likelihood L is a number strictly between 0 and 1, ℓ is a negative number. Let us a loose upper bound on the difference between $\ell_{s,g}(\pi)$ and $\ell_{s,g}(\pi')$, for two arbitrary partitions.

Let

$$Dist(\pi, \pi') = |In(\pi)\Delta In(\pi')|$$

where Δ is the symmetric difference. Note that $|In(\pi)\Delta In(\pi')| = |Out(\pi)\Delta Out(\pi')|$.

Thus, from 6 we easily obtain an upper bound

$$|\ell_{s,g}(\pi) - \ell_{s,g}(\pi')| \leq Dist(\pi, \pi') \left[|k_1 - k_3| + |k_2 - k_4|\right] s_1. \qquad (7)$$

Thus, the difference in log-likelihoods is bounded by a linear factor of $Dist(\pi, \pi')$. Thus, we have deduced the following proposition.

Proposition **2** *Consider any* $1 > p_{in} > p_{out} > 0$ *in a situation where* s_{ij}*'s are bounded above. Then independent of* n*, there exists a constant* K*, such that*

$$|\ell(\pi) - \ell(\pi')| < KDist(\pi, \pi').$$

It is important that the constant K be independent of n. Otherwise, fixing any n there are only finitely many partitions, and so obtaining such a relationship would be trivial. The fact that this works for any n, means that there is truly a relationship between the distance between partitions and the difference in log-likelihoods.

Bounding the absolute difference in log-likelihoods may not be so informative in some problems, as we might not know how much variation in log-likelihoods there is to begin with. Thus, it is also useful to bound the relative difference. To do this, we begin by defining a relative measure of distance between partitions.

Let $dist(\pi, \pi')$ be the normalized distance between two partitions. That is,

$$dist(\pi, \pi') = \frac{Dist(\pi, \pi')}{n(n-1)/2},$$

where $n(n-1)/2$ is the maximal possible distance between any two partitions (i.e., the distance between the discrete and degenerate partitions which differ in the grouping of all $n(n-1)/2$ pairs of nodes).

Next, we need some measure of how much variation in log-likelihoods we should expect. We do this by computing the distance between the optimal partition and the worst possible grouping of nodes, which is essentially the opposite of the optimal partition (all ij's in $In(\pi)$ are switched to Out and vice versa). This is not necessarily a partition, but it gives us a measure of the order of magnitude of how much the log-likelihood varies as we change from the optimal partition to the worst possible grouping of nodes. Let $M(n)$ denote this magnitude, which can vary with n. In fact, in the appendix, we bound this below by a factor that is proportional to $n(n-1)/2$. From this and Proposition 2 we then deduce the following.

Proposition **3** *Consider any* $1 > p_{in} > p_{out} > 0$ *and suppose that each* $s_{ij} = s_1 > 0$ *for some* s_1*. There exists a constant* k*, such that for any* n *and partitions* π *and* π'

$$\frac{|\ell(\pi) - \ell(\pi')|}{M(n)} < kdist(\pi, \pi'),$$

This now provides us with a technique for developing algorithms. If we can find a grid of partitions that comes close enough to any given partition, then we are sure to get within some distance of the optimal log-likelihood. Thus we have simplified the problem of approximating the optimal log-likelihood into a problem of approximating partitions.

Before proceeding, we also make another note on simplifying the problem. If we loosen the problem, so that instead of searching over partitions, we instead search over groupings of pairs of nodes into the In and Out categories, then it is very easy to derive an optimal log-likelihood. Essentially, this is a problem where we can group to nodes into being "in the same group" or "in different groups", without worrying about whether there is consistency across nodes. For example, in doing this we might end up classifying i in the same group with j, and j in the same group with k, but i and k in different groups. Thus, this would not correspond to an actual community structure, but instead a new graph over nodes. By doing this, we obtain the absolute highest possible log-likelihood score. If we are lucky and this turns out to be a partition, then it is surely the optimal partition. In general it will not be, but then starting from this grouping, if we can find a partition that is close to it, then by the above propositions, we are sure to have an approximately optimal partition.

The method for finding this upper bound, or starting point for an algorithm, is simple and is as follows. Given any estimates for p_{in} and p_{out}, we have estimates for the parameters $k_1, \ldots, k_4$ in (2). Then we set ij in the In category, if and only if $k_1 g_{ij} + k_2 s_{ij} > k_3 g_{ij} + k_4 s_{ij}$, or

$$\frac{g_{ij}}{s_{ij}} > \frac{k_4 - k_2}{k_1 - k_3}.$$

The following proposition then shows that if the sizes grow, this method will yield a partition with a probability converging to 1.

Note that this procedure only requires n^2 steps.

Proposition 4 *Let N be fixed, let for each $t = 1, 2, ...,$ s^t be a matrix of sizes, $s^t \in I\!R^{n \times n}$, and such that $\lim_{t\to\infty} s^t_{ij} = \infty, \forall ij \in N \times N$. For each $t = 1, ...,$ assume that a network $g^t \in G(s^t)$ is generated according to $p_{in} > p_{out}$, and given some true partition of the nodes π. As $t \to \infty$, with probability one, the above grouping technique leads to a partition $\hat{\pi}$, which is optimal according to the ordering $\succ_{L(\hat{p}_{in}(\hat{\pi}), \hat{p}_{out}(\hat{\pi}))}$, and $\hat{\pi} \to \pi$, $(\hat{p}_{in}(\hat{\pi}), \hat{p}_{out}(\hat{\pi})) \to (p_{in}, p_{out})$.*

The proposition follows from the fact that as $s^t \to \infty$, the probability that any two nodes are misidentified as being in the same group or separate groups goes to 0 by the law of large numbers. As the set of nodes is fixed, the proposition follows directly and so the proof is omitted.

6.3 An Application

We now proceed to an application to illustrate the method. The data are of cross-citations of 42 economic journals for the period of 1995-1997, from Pieters and Baumgartner (2002). The nodes are journals and the number of articles, and the entries g_{ij} are the number of citations by articles published in journal i during the time period of 1995-1997 of articles in journal j (of any previous time). We take a short-cut in estimating the sizes s_{ij}. The exact count should be the sum across articles published in i during 1995-1997 of the total number of articles that were ever published in j prior to the respective article in i. We simply estimate the size s_{ij} to be a factor proportional to the relative sizes of journals i and j in terms of articles published per year.[7] So, we estimate these based on the respective number of articles published in 2003, as reported on the ISI web of science journal performance metrics.

Our methodology for finding partitions with relatively high likelihoods is the following. We use an approximate greedy algorithm. We start by using the self-citations for an estimate of $\hat{p}_{in}$, and we make a grid for p_{out}. We then generate many random orderings of the set of journals, and for each ordering we find approximately optimal partitions for each p_{out}. We do that by at each step taking each consecutive pair of the elements of the current partition, and if by consolidating them the improvement in the likelihood is above a threshold, then we construct new partition where these two nodes (or groups at later steps) are consolidated. Given the new partition, we recompute the $\hat{p}_{in}$ and $\hat{p}_{out}$. We check that the final partition satisfies Consistent Optimality.

In Table 1 below we present a summary of the sample.

---

[7] So, this approximation does not account for the fact that some journals have been around for more years than others or that the number of articles per year may have varied differently across journals over time. Noting that there is a strong bias towards more recent articles in citations, this estimate should not be far off.

	journals	Size	out total	self	p_{in}
1	American Economic Review (AER)	176	10222	694	0.0025
2	American Journal of Agricultural Economics (AJAE)	124	3269	950	0.0069
3	Brooking Papers of Economic Activity (BPEA)	5	1239	89	0.3956
4	Cambridge Journal of Economics (CJE)	44	993	162	0.0093
5	Canadian Journal of Economics (CJE)	45	2487	163	0.0089
6	Economic Geography (EG)	19	216	91	0.0280
7	Econometrica (E)	61	8851	713	0.0213
8	Economic History Review (EHR)	9	1109	354	0.4856
9	Economic Inquiry (EI)	50	1796	122	0.0054
10	Economic Journal (EJ)	75	4189	318	0.0063
11	European Economic Review (EER)	52	2919	185	0.0076
12	Exploration of Economic History (EEH)	14	807	102	0.0578
13	Health Economics (HE)	82	305	81	0.0013
14	International Economic Review (IER)	54	2574	115	0.0044
15	Journal of Comparitive Economics (JCE)	37	731	135	0.0110
16	Journal of Development Economics (JDE)	73	2095	220	0.0046
17	Journal of Econometrics (JE)	79	3942	648	0.0115
18	Journal of Economic History (JEH)	29	1434	315	0.0416
19	Journal of Economic Literature (JEL)	21	2261	93	0.0234
20	Journal of Economic Perspectives (JEP)	37	1950	91	0.0074
21	Journal of Economic Theory (JET)	105	5311	907	0.0091
22	J of Environmental Econ and Management (JEEM)	66	1720	344	0.0088
23	Journal of Health Economics (JHE)	53	813	139	0.0055
24	Journal of Human Resources (JHR)	44	1627	162	0.0093
25	Journal of International Economics (JIE)	58	1992	237	0.0078
26	Journal of Labor Economics (JLE)	32	1521	100	0.0109
27	Journal of Law and Economics (JLE2)	24	1146	124	0.0239
28	Journal of Law, Economics and Organization (JLEO)	20	723	115	0.0319
29	Journal of Mathematical Economics (JME)	41	1362	252	0.0167
30	Journal of Monetary Economics (JME)	64	2962	281	0.0076
31	Journal of Political Economy (JPE)	42	7428	359	0.0226
32	Journal of Public Economics (JPE2)	116	3701	491	0.0041
33	Journal of Urban Economics (JUE)	58	1434	394	0.0130
34	Land Economics (LAE)	39	1296	210	0.0153
35	National Tax Journal (NTJ)	47	849	228	0.0115
36	Oxford Bulletin of Economics and Statistics (OBES)	40	1019	56	0.0039
37	Oxford Economic Papers (OEP)	32	1450	79	0.0086
38	Quarterly Journal of Economics (QJE)	40	4905	234	0.0163
39	Rand (Bell) Journal of Economics (RJE)	30	2365	274	0.0338
40	Review of Economic Studies (RES)	37	3676	158	0.0128
41	Review of Economics and Statistics (RES2)	92	3581	217	0.0028
42	World Development (WD)	120	2603	749	0.0058

Table 1. Sizes are as reported by ISI web of science for 2003, so that $\tilde{s} = 3sizes$. Out total is the total of times that the journal cites other journals, plus the number of times it is cited by other journals. Self is the number of times the journal cites itself. For each journal, $p_{in} = \frac{self}{9size^2}$.

Table 2 presents the estimated maximum likelihood partition.

2	American Journal of Agricultural Economics (AJAE)
13	Health Economics (HE)
6	Economic Geography (EG)
4	Cambridge Journal of Economics (CJE)
10	Economic Journal (EJ)
37	Oxford Economic Papers (OEP)
15	Journal of Comparitive Economics (JCE)
8	Economic History Review (EHR)
12	Exploration of Economic History (EEH)
18	Journal of Economic History (JEH)
16	Journal of Development Economics (JDE)
23	Journal of Health Economics (JHE)
25	Journal of International Economics (JIE)
24	Journal of Human Resources (JHR)
26	Journal of Labor Economics (JLE)
32	Journal of Public Economics (JPE2)
33	Journal of Urban Economics (JUE)
22	J of Environmental Econ and Management (JEEM)
34	Land Economics (LAE)
35	National Tax Journal (NTJ)
7	Econometrica (E)
14	International Economic Review (IER)
17	Journal of Econometrics (JE)
21	Journal of Economic Theory (JET)
29	Journal of Mathematical Economics (JME)
39	Rand (Bell) Journal of Economics (RJE)
40	Review of Economic Studies (RES)
36	Oxford Bulletin of Economics and Statistics (OBES)
1	American Economic Review (AER)
3	Brooking Papers of Economic Activity (BPEA)
5	Canadian Journal of Economics (CJE)
11	European Economic Review (EER)
19	Journal of Economic Literature (JEL)
20	Journal of Economic Perspectives (JEP)
30	Journal of Monetary Economics (JME)
31	Journal of Political Economy (JPE)
38	Quarterly Journal of Economics (QJE)
9	Economic Inquiry (EI)
27	Journal of Law and Economics (JLE2)
28	Journal of Law, Economics and Organization (JLEO)
41	Review of Economics and Statistics (RES2)
42	World Development (WD)

Table 2. The estimated maximum-likelihood partition.

This data set was studied by Pieters and Baumgartner [2002], and community structures were derived using a hierarchical clustering algorithm and specifying that there should be seven communities. The partition that they obtain is different than the partition that we obtain. Most notably, our estimated maximum likelihood partition has nineteen communities, while theirs has an assumed seven. Of course, their partition also has a lower likelihood score according to our measure (insert numbers).

Our estimated partition is also pictured below where the nodes of the same color are in the same community and the thickness of the lines represents the weights of the citations.

It is important to emphasize that it can be very difficult to identify the community structures without a methodology such as the one developed here. The graph above makes it fairly clear how the communities are structured, but obtaining such a graph depends on the method. In order to illustrate this, below is the graph of the raw data, where line thickness is again an indication of relative citations, but where we have not colored or grouped nodes according to communities. From such a graph, it is essentially impossible to determine the community structure without a careful method.

7 Concluding Remarks

There are some interesting issues for further development, as suggested by our approach and analysis here.

First, given the plethora of different algorithms for analyzing community structures, it is important to begin to systematically study their properties. Providing characterizations of different methods, as we have done for the likelihood approach, would be very helpful in being able to compare methods. In the past, comparisons across methods have usually been done simply by applying them and seeing which one seems to give a more subjectively sensible partition.

Second, along with this, it makes sense to see if one can rationalize some of the more popular algorithms through some model. That is, is there any explanation of how network data is generated for which given algorithms are finding the true or most likely to be true partition?

Third, as we have started from a model and derived conditions characterizing the optimal community structures, rather than starting from an algorithm, we face the task of finding algorithms that can search for approximately optimal structures. We used a greedy algorithm for the application above, and have presented some results that suggest different techniques for finding algorithms. But there is still much more that can be done in terms of deriving algorithms that will find approximately optimal community structures quickly.

Fourth, there are some obvious extensions that can be made to the model we have suggested here. The model we specified has a single p_{in} that indicates the probability of linkings between nodes in the same community, and a single p_{out} for the probability of linkings between nodes of different communities. Many applications may have more heterogeneity than this, where p_{in}'s and p_{out}'s vary across communities of nodes. Most notably, things might tend to be asymmetric. For instance, in the journal citation data, it is clear that some communities have much higher rates of being cited by other communities, than they do of citing other communities. This suggests that a hierarchical model might be worth exploring, where there are different levels of communities, and the chance of lower level communities attaching to higher level communities would be higher than the reverse. Such a model is an easy extension of what we have done here, and then there is a direct analog in terms of the likelihood expressions. There is essentially no conceptual difference, only some extra parameters (more p_{in}'s and p_{out}'s) to estimate.

8 References

References

[1] Boorman, S.A. and H.C. White (1976) "Social Structure from Multiple Networks II: Role Structures," *American Journal of Sociology*, 81: 1384- 1446.

[2] Bron, C. and J. Kerbosch (1973) "Finding all cliques of an undirected graph," *Comm. ACM*, 16: 575-577.

[3] Fiedler, M. (1973) "Algebraic connectivity of graphs" *Czech. Math. J.* 23: 298-305.

[4] Giada, L. and M. Marsili (2005) "Data clustering and noise undressing of correlation matrices," revision of arXiv:cond-mat/0101237 v1 16 Jan 2001.

[5] Girvan, M. and M. E. J. Newman, (2002) "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences* 99: 7821-7826.

[6] Kernighan B.W. and S. Lin (1970), "An efficient heuristic procedure for partitioning graphs," *Bell System Technical Journal* 49: 291-307.

[7] Lance, G.N. and W.T. Williams (1967) "A General Theory of Classificatory Sorting Strategies," *Computer Journal* 9:373-380.

[8] Lorrain F. and H.C. White (1971) "Structural Equivalence of Individuals in Social Networks," *Journal of Mathematical Sociology*, 1: 49-80.

[9] Newman, M.E.J. (2004) "Detecting Community Structure in Networks," *Physical Review E* 69, 066133.

[10] Pieters, F.G.M. and H. Baumgartner (2002) "Who talks to whom? Intra- and interdisciplinary communication of economics journals ," *Journal of Economic Literature*, 40(2):483-509.

[11] Pothen, A., H. Simon, and K.-P. Liou, (1990) "Partitioning sparse matrices with eigenvectors of graphs," *SIAM J. Matrix Anal. Appl.*, 11: 430-452.

[12] Radicchi, F., C. Castellano, F. Cecconi, V. Loreto, and D. Parisi (2003) "Defining and identifying communities in net- works," cond-mat/0309488.

[13] Schwartz, J.E. (1977) "An Examination of CONCOR and related methods for blocking sociometric data," in *Sociological Methodology 1977*, edited by D.R. Heise, Jossey-Basse: San Francisco, 255-282.

[14] Tyler, J.R., D.M. Wilkinson, and B.A. Huberman (2003) "Email as spectroscopy: Automated discovery of com- munity structure within organizations," In M. Huysman, E. Wenger, and V. Wulf (eds.), *Proceedings of the First International Conference on Communities and Technolo- gies*, Kluwer: Dordrecht.

[15] White, H.C., S.A. Boorman, and R.L. Breiger (1976) "Social Structure from Multiple Networks I: Blockmodels of roles and positions," *American Journal of Sociology*, 81: 730-779.

9 Appendix

The proof of Theorems **??** and 2 is as follows.

First, it is easy to check that the axioms hold for $\succeq_{s,g}^{L(p_{in},p_{out})}$, by direct inspection of 4. Thus, in each case, we simply prove the converse; namely that any ordering satisfying the given axioms must be a maximum likelihood ordering for some p_{in} and p_{out}.

We start with the proof of Theorem 2, as this is then used to prove Theorem **??**

Proof of Theorem 2 : The likelihood-based ordering $\succeq^{ML}$ satisfies the axioms as argued above.

For the converse, note that by internal differences, for any given rule $\succeq$, there exists $H : I\!R^2 \to I\!R$ such that $\pi \succeq_{s,g} \pi'$ if and only if $H(D(g,\pi,\pi'),D(s,\pi,\pi')) \geq 0$.

Note that weak monotonicity implies that $H(w,0) > 0$ and $H(0,w) < 0$ whenever $w > 0$.

Let γ be defined by scaling. Consider two cases.

Case 1. $z - \frac{1}{\gamma}y > 0$

By scaling, $H(z, y) > 0$ if and only if $H(z - \frac{1}{\gamma}y, 0) > 0$. Thus, since $H(w, 0) > 0$ whenever $w > 0$ it follows that $H(z, y) > 0$ whenever $z - \frac{1}{\gamma}y \geq 0$.

Case 2. $z - \frac{1}{\gamma}y < 0$

By scaling, $H(z, y) > 0$ if and only if $H(0, y - \gamma z) > 0$. Thus, since $H(0, w) < 0$ whenever $w > 0$ it follows that $H(z, y) < 0$ whenever $z - \frac{1}{\gamma}y < 0$.

Case 3. $z - \frac{1}{\gamma}y = 0$

By scaling, $H(z, y) = 0$ if $H(0, y - \gamma z) = H(0, 0) = 0$. Thus, since $H(0, 0) = 0$ it follows that $H(z, y) = 0$ whenever $z - \frac{1}{\gamma}y = 0$.

Thus, we have shown that $\pi \succeq_{s,g} \pi'$ if and only if $\gamma D(g\pi, \pi') \geq D(s, \pi, \pi')$. From (4) we know that this corresponds to a maximum likelihood method. ▮

Proof of Theorem 1 : Again, the likelihood-based ordering $\succeq^L$ satisfies the axioms as argued above.

Given Theorem 2, we need only show that a $\succeq \in R$, satisfying independence, neutrality, and scaling satisfies internal differences.

So, consider (g, s), (g', s'), π, π', π'', π''' such that

$$D(g, \pi, \pi') = D(g', \pi'', \pi''')$$

and

$$D(s, \pi, \pi') = D(s', \pi'', \pi''').$$

We need to show that $\pi \succeq_{s,g} \pi'$ if and only if $\pi'' \succeq_{s',g'} \pi'''$.

First, note that by independence, if $\pi \succeq_{s,g} \pi'$ the same is true regardless of the values of g_{ij} and s_{ij} such that either $ij \in In(\pi) \cap In(\pi')$ or $ij \in Out(\pi) \cap Out(\pi')$. So, without loss of generality suppose that $g_{ij} = 0$ and $s_{ij} = 0$ whenever either $ij \in In(\pi) \cap In(\pi')$ or $ij \in Out(\pi) \cap Out(\pi')$.

Next, by the second part of Neutrality if there is more than one link ij that is in $In(\pi) \cap Out(\pi')$, then we can equivalently consider a network g^1 such that $g^1_{kl} = 0$ and $s^1_{kl} = 0$, for all $kl \in In(\pi) \cap Out(\pi')$ such that $kl \neq ij$ and $g^1_{ij} = \sum_{kl \in In(\pi) \cap Out(\pi')} g_{kl}$ and $s^1_{ij} = \sum_{kl \in In(\pi) \cap Out(\pi')} s_{kl}$. Similarly, either $Out(\pi) \cap In(\pi') = \emptyset$ or we can find a link $hm \in Out(\pi) \cap In(\pi')$ and then set $g^1_{kl} = 0$ and $s^1_{kl} = 0$, for all $kl \in Out(\pi) \cap In(\pi')$ such that $kl \neq hm$ and $g^1_{hm} = \sum_{kl \in Out(\pi) \cap In(\pi')} g_{kl}$ and $s^1_{hm} = \sum_{kl \in Out(\pi) \cap In(\pi')} s_{kl}$.

So, $g^1_{kl} = s^1_{kl} = 0$ for all kl except at most two links, $ij \in In(\pi) \cap Out(\pi')$ and $hm \in Out(\pi) \cap In(\pi')$.

Consider the case where $g^1_{ij} \geq g^1_{hm}$, as the other is analogous.

By the first part of neutrality, this is equivalent to a network g^2 where $g^2_{ij} = \sum_{kl \in In(\pi) \cap Out(\pi')} g_{kl} - \sum_{kl \in Out(\pi) \cap In(\pi')} g_{kl}$ and $g^2_{hm} = 0$. Since $g_{ij} = 0$ whenever either $ij \in In(\pi) \cap In(\pi')$ or $ij \in Out(\pi) \cap Out(\pi')$, it follows that

$$g^2_{ij} = D(g^2, \pi, \pi'), \tag{8}$$

and $g^2_{kl} = 0$ for all $kl \neq ij$.

Case 1: $s^1_{ij} > \gamma g^2_{ij}$.

By scaling, we can consider s^3 and g^3 such that $s^3_{ij} = s^1_{ij} - \gamma g^2_{ij}$ and $g^3_{ij} = 0$ and other entries are as before.

Case 1a: $s^1_{ij} - \gamma g^2_{ij} > s^1_{hm}$

By neutrality and weak monotonicity, it follows that $\pi \succ \pi'$. Note that the condition of 1a is equivalent to $D(s, \pi, \pi') > \gamma D(g, \pi, \pi')$.

Case 1b: $s^1_{ij} - \gamma g^2_{ij} < s^1_{hm}$

By neutrality and weak monotonicity, it follows that $\pi' \succ \pi$. Note that the condition of 1b is equivalent to $D(s, \pi, \pi') < \gamma D(g, \pi, \pi')$.

Case 2: $s^1_{ij} \leq \gamma g^2_{ij}$.

Again by the first part of neutrality, we can equivalently consider s^2 such that $s^2_{ij} = \gamma g^2_{ij}$ and $s^2_{hm} = s^1_{hm} + \gamma g^2_{ij} - s^1_{ij}$. Then by scaling, we can consider s^3 where $s^3_{ij} = g^3_{ij} = 0$. By weak monotonicity if follows that $\pi' \succ \pi$ provided $s^2_{hm} > 0$. Note that in this case, $s^2_{hm} = -D(s, \pi, \pi') + \gamma D(g, \pi, \pi')$. So we are in the case where $D(s, \pi, \pi') < \gamma D(g, \pi, \pi')$.

We are only left with a case where $s^2_{hm} = 0$, which corresponds to a case where $D(s, \pi, \pi') = \gamma D(g, \pi, \pi')$. Here, weak monotonicity implies indifference.

If we perform the same analysis for (g', s') we also find that $\pi' \succ \pi$ whenever $D(s', \pi, \pi') < \gamma D(g', \pi, \pi')$ $\pi \succ \pi'$ whenever $D(s', \pi, \pi') > \gamma D(g', \pi, \pi')$ and $\pi \sim \pi'$ when $D(s', \pi, \pi') = \gamma D(g', \pi, \pi')$.

Thus, if the D's are the same across (g, s) and (g', s'), then the ranking of π and π prime are the same. ▮

Proof of Proposition 1: To see that $\bar{\pi}$ satisfies consistent optimality is straightforward.

We have to prove that $\bar{\pi}$ is unique partition satisfying consistent optimality. We prove this by showing that for every $\pi \in \Pi(N), \pi \neq \bar{\pi}$, $\bar{\pi} \succ_{L(\hat{p}_{in}(\pi), \hat{p}_{out}(\pi))} \pi$.

So take $\pi \in \Pi(N)$, and denote $p = \hat{p}_{in}(\pi), q = \hat{p}_{out}(\pi)$. Clearly, $p \leq p_{in}$, and $p = p_{in}$ if and only if $In(\pi) \subset In(\bar{\pi})$. Similarly, $q \geq p_{out}$, and $q = p_{out}$ if and only if $Out(\pi) \subset Out(\bar{\pi})$. Denote by $\bar{L}[.] = L(p_{in}, p_{out})[.]$ by $L[.] = L(p, q)[.]$, and $l[.] = \log(L[.])$, $\bar{l}[.] = \log(\bar{L}[.])$.

Consider first the case when $Out(\pi) \subset Out(\bar{\pi})$. Denote by $K = Out(\bar{\pi}) \cap In(\pi)$. Then,

$$l[\pi] - l[\bar{\pi}] = \left(\sum_{ij \in K} g_{ij} \right) \log p + \left(\sum_{ij \in K} s_{ij} - \sum_{ij \in K} g_{ij} \right) \log(1-p)$$
$$- \left(\sum_{ij \in K} g_{ij} \right) \log q - \left(\sum_{ij \in K} s_{ij} - \sum_{ij \in K} g_{ij} \right) \log(1-q).$$

Note that this expression is increasing in p. To see this, compute

$$\frac{d(l[\pi] - l[\bar{\pi}])}{dp} = \frac{\sum_{ij \in K} g_{ij}}{p} - \frac{\sum_{ij \in K} s_{ij} - \sum_{ij \in K} g_{ij}}{1-p} = \sum_{ij \in K} s_{ij} \left[\frac{p}{q} - \frac{1-q}{1-p} \right],$$

where the last inequality follows from the assumption that $g_{ij} = q s_{ij}, \forall ij \in Out(\bar{\pi})$. Clearly, $p < p_{in}$ (since p is some weighted average of p_{in} and $q < p_{in}$), and $\bar{l}[\pi] - \bar{l}[\bar{\pi}] < 0$. Thus, we have that $l[\pi] - l[\bar{\pi}] < 0$, which proves the claim for this case. Similarly, for the case when $In(\pi) \subset In(\bar{\pi})$ (then the expression will be monotonic in q, and the proof is almost identical).

If neither $Out(\pi) \subset Out(\bar{\pi})$ nor $In(\pi) \subset In(\bar{\pi})$, then we can construct a finite sequence of partitions $\{\pi_k\}_{k=1}^{K}$, such that $\pi_1 = \bar{\pi}$ and $\pi_K = \pi$, and such that either $Out(\pi_{k+1}) \subset Out(\pi_k)$ or $In(\pi_{k+1}) \subset In(\pi_k), \forall k = 1, ..., K$. By the same argument, for each $k = 1, ...K-1$, π_k is more preferred to π_{k+1} than π_{k-1} is to π_k. ▮

Proof of Proposition ??: Given Proposition **??**, it is enough to show that there exists k such that $M(n) > kn(n-1)/2$.

First, it is straightforward to check that $k_1 p_{in} + k_2 > k_3 p_{in} + k_4$ and that $k_1 p_{out} + k_2 < k_3 p_{out} + k_4$.

Let x denote the minimum of $(k_1 p_{in} + k_2) - (k_3 p_{in} + k_4)$ and $(k_3 p_{out} + k_4) - (k_1 p_{out} + k_2)$.

The log-likelihood of the optimal partition $\widehat{\pi}$ under the expected g is

$$s\frac{n(n-1)}{2} \left[(k_1 p_{in} + k_2)\, \phi + (k_3 p_{out} + k_4)(1 - \phi) \right] \tag{9}$$

where ϕ is the fraction of links in $In(\widehat{\pi})$.

The worst possible grouping of nodes would lead to a log-likelihood under the expected g of

$$s\frac{n(n-1)}{2} \left[(k_1 p_{out} + k_2)(1 - \phi) + (k_3 p_{in} + k_4)\, \phi \right]. \tag{10}$$

The difference between (9) and (10) is at least

$$s\frac{n(n-1)}{2} x,$$

and thus $M(n) \geq s\frac{n(n-1)}{2} x$, as claimed. ▮