SETTINGS IN SOCIAL NETWORKS: A MEASUREMENT MODEL[†]

Michael Schweinberger[††]
Tom A.B. Snijders

A class of statistical models is proposed which aims to recover latent settings structures in social networks. Settings may be regarded as clusters of vertices. The measurement model builds on two assumptions. The observed network is assumed to be generated by hierarchically nested latent transitive structures, expressed by ultrametrics. It is assumed that expected tie strength decreases with ultrametric distance. The approach could be described as model-based clustering with an ultrametric space as the underlying metric to capture the dependence in the observations. Maximum likelihood methods as well as Bayesian methods are applied for statistical inference. Both approaches are implemented using Markov chain Monte Carlo methods.

1. Introduction

Links between entities are commonly studied in terms of networks. Examples are friendship ties between individuals, cooperation or competition between organizations, wars between nations, links between websites etc. This article focuses on social networks (Wasserman and Faust, 1994), where the entities typically correspond to (corporate) actors, such as individuals or organizations.

Since the range of human interaction is restricted by time, money, geographi-

[†]This article is based on the first author's Master's Thesis (2002).

[††]Authors' address: ICS/Statistics & Measurement Theory, Grote Rozenstraat 31, 9712 TG Groningen, the Netherlands.

cal constraints etc., social networks typically contain some local social neighborhoods, called *settings* by Pattison and Robins (2002). Settings may be regarded as close-knit clusters corresponding to actors which are strongly tied. Other terms commonly used in the literature are, e.g., groups, communities etc. The social science literature acknowledges that settings structures in social networks can have remarkable influence on how well social, economical, and political organizations can function. Classic and contemporary sociology (see, e.g., Durkheim, 1984, Simmel, 1968, Tönnies, 1955, Homans, 1950) - as well as anthropology - deals nearly inevitably with settings, be it only as an implicit restriction on social or economic action.

The sociological intuition about settings can be described by relatively small and highly cohesive groups with little overlap. Furthermore, the interaction within settings is expected to be stronger than the interaction between settings. Some settings structures have the additional property that within large settings smaller and even more cohesive settings can be distinguished, meaning that settings structures can be hierarchically nested.

In the past, many attempts have been made to model settings. We are doomed to be selective in citing work, since citing the whole body of work on settings models would be beyond the scope of this paper (many basic models are cited/sketched in Freeman, 1992, Wasserman and Faust, 1994). Some examples are the Freeman/Winship model (Freeman, 1992) building on Winship (1977), the Freeman/Granovetter model (Freeman, 1992) building on Granovetter (1973), *LS* sets (Seidman, 1983, Borgatti, Everett, and Shirey, 1990), etc.

Some models succeed better than others in capturing the sociological intuition about settings, but most of these models share one characteristic: they regard observations as outcomes of deterministic forces. This frequently leads to poor model fit and limits applicability, because often it appears to be difficult to recover settings when settings are assumed to have arisen from deterministic forces. Social network data typically exhibit - in addition to some structural characteristics - some randomness, or structural character-

istics which are not accounted for by the model. Hence stochastic models appear to be more appropriate for modeling settings than deterministic models. Furthermore, deterministic models usually neglect model uncertainty by doing as if there was a single true model, while typically there are many models which can predict the observed network reasonably; in a probabilistic framework - in particular in a Bayesian framework - model uncertainty and model selection can be addressed. A probabilistic framework additionally allows to study model parsimony, while deterministic models typically fail to summarize the observations as much as possible without missing essential information.

Some stochastic settings models have been proposed, such as the transitive graph models by Frank (1978, 1980) and Frank and Harary (1982), but statistical inference for these models is very limited.

There have been recent advances in statistical network modeling, contained in the models proposed by Pattison and Robins (2002), Hoff, Raftery, and Handcock (2002), Snijders and Nowicki (1997), and Nowicki and Snijders (2001). The Pattison and Robins (2002) models specify the dependence graphs in the p^* class of models in ways that incorporate substantive knowledge about settings structures. Though these models are appealing, the estimation procedures (pseudo maximum likelihood estimation) are suspect (Snijders, 2002). The Hoff, Raftery, and Handcock (2002) models assume that the actors can be represented in some latent space, which is assumed to have a Euclidean or an arbitrary metric, and the probability to observe a tie depends on the distance in this space. The block models by Snijders and Nowicki (1997), Nowicki and Snijders (2001) do not intend to model settings, but attempt instead to solve the related problem to model blocks containing equivalent actors.

We propose in this paper a measurement model where the observed network is assumed to reflect underlying latent settings structures, and the latent structures are specified such that the model captures the sociological intuition about settings. The measurement model assumes that the latent settings are

non-overlapping and cohesive, that the interaction within settings is stronger than the interaction between settings, and that settings may be hierarchically nested. The observed network will in most cases not perfectly match the latent settings structures, as expected by the model, since the complexity of social reality implies that the model misses almost surely some structural characteristics, and chance may additionally play a role in the evolution of settings. This is expressed by the assumption that the observed network is related to the latent settings structures by stochastic rather than deterministic processes.

The basic statistical framework is Bayesian (though maximum likelihood estimation is proposed as well), allowing to capture the uncertainty about the settings structures. The model provides simple ways to deal with randomly missing data, and can handle not only dichotomous data, but also discrete ordered data as well as continuous data. The model, as implemented now, can be applied to networks with hundreds of actors.

We introduce an ultrametric measurement model in Section 2. Statistical inference is treated in Section 3 and 4. The model is applied to data in Section 6, and problems concerning the measurement model and its implementation, as well as possible model extensions, are discussed in Section 7.

2. Measurement Model

We assume that one symmetric relation on some vertex set $\mathfrak{N} = \{1, 2, \ldots, n\}$ has been observed. The network is represented as valued graph $\mathfrak{G}(\mathfrak{N}, \mathfrak{E})$ with edge set $\mathfrak{E}$, where the edges e_{ij},

$$e_{ij} = \begin{cases} 1 & \text{if the relation has been observed for } (i,j) \in \mathfrak{N} \\ 0 & \text{otherwise,} \end{cases}$$

distinguish missing values from non-missing values. The values of the (observed) edges between the vertices (i, j) are regarded as random variables X_{ij} with outcomes x_{ij}. The variables X_{ij} may have dichotomous outcome spaces, or discrete and ordered outcome spaces, or continuous outcome spaces. It

is convenient to exclude self-loops by defining $e_{ii} = 0$. The outcomes x_{ij} are usually stored in the below-diagonal half of an $n \times n$ adjacency matrix $x = (x_{ij})$.

2.1. Latent Settings Structures

A simple model for settings structures is as follows. The observed graph is assumed to have emerged from an unobserved, i.e. latent, graph with adjacency matrix $z = (z_{ij})$ defined by

$$z_{ij} = \begin{cases} 1 & \text{if there is an edge between } i \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

We exclude self-loops by defining $z_{ii} = 0$ and assume that z is symmetric, $z_{ij} = z_{ji}$. Two main characteristics of settings will be modeled.

Assumption 1 *The latent graph exhibits a transitive structure, meaning that*

$$(z_{ij} = 1 \text{ and } z_{jk} = 1) \text{ implies } z_{ik} = 1$$

for all i, j, k *in* $\mathfrak{N}$.

If $z_{ij} = 1$, then i and j are said to share the setting. Consider the relation $i \sim j$ on $\mathfrak{N}$ defined by $\{z_{ij} = 1 \text{ or } i = j\}$. Since $\sim$ is reflexive, symmetric, and transitive, $\sim$ is an equivalence relation. Thus the transitive graph partitions $\mathfrak{N}$, meaning that each vertex is assigned to one setting and the settings do not overlap.

Beginning with Rapoport (1953a,b), and continuing in the work of Davis, Holland, and Leinhardt (see, e.g., Holland and Leinhardt, 1970, Davis, Holland, and Leinhardt, 1971, Holland and Leinhardt, 1972, 1976, Davis, 1979, Holland and Leinhardt, 1979), it has been claimed that transitivity is an important structural characteristic of social groups. This claim, interpreted in a non-deterministic way, has been confirmed by numerous studies, which show that especially friendship networks exhibit strong tendencies towards transitivity.

Assumption 2 *The latent transitive graph is partitioned such that the interaction within settings is denser than between settings.*

This claim has been made by Homans (1950, p. 84) in his classic work on human groups, and underlies the definition of *LS* sets (Seidman, 1983, Borgatti, Everett, and Shirey, 1990) in social network analysis.
While these two assumptions will not quite bring us into the position to explain the entire social universe, they capture important structural characteristics which have long been claimed to hold for human groups (see also Freeman, 1992).
To model latent transitive structures in social networks, ultrametrics can be used, as was first suggested by Freeman (1992), elaborating on Winship (1977). Ultrametrics lead to multiple nested transitive structures, as will be outlined below.

2.2. Ultrametrics

A metric in the set $\mathfrak{N}$ is defined by a real valued distance function $d : \mathfrak{N} \times \mathfrak{N} \rightarrow [0, \infty[$ which satisfies the axioms

$$\begin{array}{lcl} d(i,j) & = & 0 \Leftrightarrow i = j \quad \forall \; i,j \in \mathfrak{N} \quad \text{(reflexivity)} \qquad (A1) \\ d(i,j) & \leq & d(i,k) + d(j,k) \quad \forall \; i,j,k \in \mathfrak{N} \quad \text{(triangle inequality)} \quad (A2) \end{array}$$

It can be proven (e.g., Hu, 1966) that if $d : \mathfrak{N}^2 \rightarrow [0, \infty[$ is any metric in $\mathfrak{N}$, then

$$\begin{array}{lcl} d(i,j) & = & d(j,i) \quad \forall \; i,j \in \mathfrak{N} \quad \text{(symmetry)} \quad (A3) \\ d(i,j) & \geq & 0 \qquad \forall \; i,j \in \mathfrak{N}. \qquad\qquad\quad (A4) \end{array}$$

A metric is called an ultrametric iff in addition $d : \mathfrak{N}^2 \rightarrow [0, \infty[$ satisfies the ultrametric inequality

$$d(i,j) \leq \max\left[d(i,k), d(j,k)\right] \quad \forall \, i,j,k \in \mathfrak{N}. \quad (A5)$$

The ultrametric inequality can be traced back to Hausdorff (1934). The ultrametric inequality constrains the distances more than the triangle inequality,

since

$$\max[d(i,k), d(j,k)] \;\leq\; d(i,k) + d(j,k).$$

In the sequel, it is assumed that d is an ultrametric on $\mathfrak{N}$. It is convenient to represent d by the below-diagonal half of a $n \times n$ distance matrix with entries $d(i,i) = 0$ on the diagonal and entries $d(i,j) > 0,\ j < i$, below the diagonal.

It can readily be proven that if d is an ultrametric on $\mathfrak{N}$ and $d(i,j) = \min[d(i,j), d(j,k), d(i,k)]$, then $d(j,k) = d(i,k)$. Furthermore, it is not a restriction to assume that the distances take integer values $h \in \{0, 1, \ldots, H\}$, since the ultrametric axioms are invariant under monotonic transformations.

2.3. Relating Ultrametric Structures to Transitive Structures

Ultrametric structures imply transitive structures. This can be seen by defining settings as follows.

Definition *A subset $\mathfrak{S} \subseteq \mathfrak{N}$ can be defined as a setting if there exists a positive number $d_{\mathfrak{S}}$ such that*

$$\{i \in \mathfrak{S}, d(i,j) \leq d_{\mathfrak{S}}\} \Leftrightarrow j \in \mathfrak{S}.$$

Given level h of an ultrametric d, a transitive structure can be derived by establishing

$$z_{ij}^{(h)} = \begin{cases} 1 & \text{if } d(i,j) \leq h \\ 0 & \text{otherwise.} \end{cases}$$

□

To each threshold value $h \in \{0, 1, \ldots, H\}$ corresponds a partition of $\mathfrak{N}$ into settings. This is the partition corresponding to the equivalence relation $i \sim j$ on $\mathfrak{N}$ defined by $d(i,j) \leq h$. The fact that this indeed is an equivalence relation can be proven from the ultrametric inequality. In addition, it is trivial to prove that that the partition at level $h-1$ is finer than the partition at

level h.

This has two implications. First, we made a transition from one latent transitive graph to $H+1$ latent transitive graphs, corresponding to the levels $h \in \{0, 1, \ldots, H\}$. Second, the transitive graphs are hierarchically nested.

We illustrate this in Table 1 by an ultrametric d on $\mathfrak{N} = \{1, 2, 3, 4\}$ with levels $h \in \{0, 1, 2\}$. For the maximum distance 2, there is a single setting

Table 1: Ultrametric d on $\mathfrak{N} = \{1, 2, 3, 4\}$ with levels $\{0, 1, 2\}$ and the corresponding transitive graphs

	ultrametric d				graph level 2				graph level 1				graph level 0			
vertex	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	0				1				1				1			
2	1	0			1	1			1	1			0	1		
3	2	2	0		1	1	1		0	0	1		0	0	1	
4	2	2	1	0	1	1	1	1	0	0	1	1	0	0	0	1

$\{1, 2, 3, 4\}$. Moving to level 1, we find two settings, $\{1, 2\}$ and $\{3, 4\}$. We observe that these settings do not overlap and are nested in the setting at level 2. Decreasing the threshold value to 0, we obtain settings $\{1\}$, $\{2\}$, $\{3\}$, and $\{4\}$. We notice that the number of settings is non-decreasing when moving from level 2 to level 0, and ranges from 1 to $n = 4$.

Settings as defined by ultrametric structures exist in an exact way in hierarchical organizations, e.g., in political administration (address - housing block - neighborhood - municipality - etc.) and in firms (working group - department - branch). In an approximate way, such settings structures also can be seen in many other social networks. Section 2.4 proposes how such approximations could be modeled.

Ultrametric structures can be regarded as a mathematical expression of Mazur (1971, p. 308)'s proposition that "Friends are likely to agree, and unlikely to disagree; close friends are very likely to agree, and very unlikely to disagree". Mazur (1971), Davis, Holland, and Leinhardt (1971), Holland

and Leinhardt (1976) tested this proposition, and claimed empirical support for it.

2.4. Probability Models

The measurement model defines how the probability distribution of the observed adjacency matrix x depends on the unobserved, and hence latent, ultrametric d. We assume that, given the distance d, the variables X_{ij} for $i, j \in \mathfrak{N},\ j < i$, are independent identically distributed random variables and the conditional distribution of X_{ij} depends only on $d(i, j)$. The outcome spaces in applications will depend on how the relation on $\mathfrak{N}$ has been measured, and can be dichotomous $\{0, 1\}$, discrete with ordered outcome space $\{0, 1, 2, \dots\}$, or continuous. The relation between d and x is expressed as

$$E(X_{ij} \mid d(i, j) = h) = \theta_h.$$

The vector $(\theta_1, \dots, \theta_H)'$ is denoted by θ. The trivial level 0 of the ultrametric is discarded.
For the three outcome spaces considered, the probability distributions conditional on the distances $d(i, j)$ are as follows:

- Bernoulli for dichotomous outcomes;
- Poisson for ordered discrete outcomes;
- normal with variance σ^2 (independent of $d(i, j)$) for continuous outcomes.

To model the sociological expectation that the interaction density within settings is higher than between settings, we impose on θ the constraint

$$\theta_1 \geq \cdots \geq \theta_H. \tag{1}$$

To illustrate, take level h of the ultrametric, and three distinct vertices $i, j, k \in \mathfrak{N}$ with

$$d(i, j) \leq h \text{ and } d(j, k) = d(i, k) > h.$$

We observe that i and j share the same setting, given level h of the ultrametric, while k does not belong to this setting. Thus

$$E(X_{ij} \mid d) \geq E(X_{jk} \mid d) = E(X_{ik} \mid d).$$

This expresses that interaction is denser within settings than between settings. Though this expression is quite simple, it has an intuitive appeal, and keeps the measurement model analytically tractable.
For dichotomous outcome spaces, this measurement model yields the probability function

$$P(x \mid d, \theta) = \prod_{h=1}^{H} \theta_h^{s_h} (1 - \theta_h)^{m_h - s_h}, \tag{2}$$

where X_{ij} can take values $0, 1$, the additional restriction

$$1 > \theta_1 \geq \cdots \geq \theta_H > 0,$$

with

$$m_h = \sum_{i=2}^{n} \sum_{j=1}^{i-1} I[d(i,j) = h]\ e_{ij} \tag{3}$$

denoting the number of ordered pairs of vertices $(i, j),\ j < i$, with distance h, and

$$s_h = s_h(x) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} I[d(i,j) = h]\ e_{ij}\ x_{ij} \tag{4}$$

denoting the number of ordered pairs of vertices $(i, j),\ j < i$, with distance h where $X_{ij} = x_{ij}$ was observed. Missing values, indicated by $e_{ij} = 0$, are excluded from the calculation of m_h and s_h. The expression $I[.]$ denotes an indicator function, taking the value 1 when its argument is true, and 0 otherwise. For ordered discrete outcome spaces, the measurement model yields the probability function

$$P(x \mid d, \theta) = \prod_{h=1}^{H} c_h\ \exp\left[-m_h\ \theta_h\right]\ \theta_h^{s_h} \tag{5}$$

with X_{ij} assuming nonnegative integer values, the additional restriction $\theta_H > 0$, m_h and s_h as defined above, and $c_h = c_h(x)$ depending on the observation x and level h but not on θ. For continuous outcome spaces, the measurement model yields the probability density function

$$P(x \mid d, \theta) = c \prod_{h=1}^{H} \exp\left[-\frac{a_h - 2s_h\theta_h + m_h\theta_h^2}{2\sigma^2}\right] \tag{6}$$

where X_{ij} can take any real values,

$$a_h = a_h(x) = \sum_{i=2}^{n}\sum_{j=1}^{i-1} I[d(i,j) = h]\; e_{ij}\; x_{ij}^2 \tag{7}$$

and c denoting a constant. Missing values are excluded from the calculation of a_h.

Ultrametrics have been used extensively in the social sciences to model proximity data (Corter, 1996), but the ultrametric d was commonly regarded as parameter in deterministic estimation procedures based on optimization criteria, such as in De Soete (1986), which are stepwise maximizing procedures, easily trapped in local optima. Some references may be found in Wedel and DeSarbo (1998). Statistical estimation of ultrametrics was rare until recently, when Markov chain Monte Carlo (MCMC) (Gilks, Richardson, and Spiegelhalter, 1996) started to find widespread application. Wedel and DeSarbo (1998) proposed the EM-algorithm (Dempster, Laird, and Rubin, 1977) to estimate (constrained) ultrametrics, and in biology MCMC methods have been applied by Yang and Rannala (1997) to estimate (very small) phylogenetic trees. Advanced MCMC methods for estimating phylogenetic trees have been developped by Huelsenbeck and Ronquist (2001).[1]

We propose two distinct approaches to statistical inference, one Bayesian approach and one maximum likelihood approach, implemented by MCMC methods. These two approaches are treated in Section 3 and Section 4. The

[1]We must admit that we were not aware of this work when we elaborated and implemented this model in 2002.

proposed estimation techniques apply in analogous ways to the three probability models and can be discussed without referring to the probability model in question.

3. Bayesian Approach

Bayesian statistics (see, e.g., Press, 1989) treats the entities d and θ as unobserved random variables. Inference concerning d and θ is based on the posterior distribution

$$P(d, \theta \mid x) = \frac{P(x \mid d, \theta)\ P(d, \theta)}{\kappa}$$

where $P(d, \theta \mid x)$ is the posterior distribution, $P(x \mid d, \theta)$ denotes the likelihood function given here by Equation (2), Equation (5), or Equation (6), $P(d, \theta)$ is the prior distribution, and $\kappa = P(x)$ denotes a normalizing constant, involving the sum over all states in the state space. Since the state space is finite but very large, it is practically infeasible to calculate κ, and we are left with

$$P(d, \theta \mid x) \propto P(x \mid d, \theta)\ P(d, \theta).$$

This implies that we can calculate the posterior distribution only up to a multiplicative constant. This problem is solved by Markov chain Monte Carlo (MCMC) methods (Gilks, Richardson, and Spiegelhalter, 1996), as described in Section 3.3.

The prior distribution can be decomposed into

$$P(d, \theta) = P(\theta \mid d)\ P(d).$$

Since in most cases in advance we are completely uncertain about d, we assume a uniform prior distribution for d. This gives

$$P(d) = \begin{cases} c & \text{if } d \text{ is ultrametric,} \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where the constant $1/c$ is the number of ultrametrics with values $0, 1, \ldots, H$. A prior distribution $P(\theta \mid d)$ can be obtained by the uniform distribution on the set

- Bernoulli:

$$\{1 > \theta_1 \geq \cdots \geq \theta_H > 0\} \tag{9}$$

- Poisson:

$$\{+\infty > \theta_1 \geq \ldots \theta_H > 0\} \tag{10}$$

- normal:

$$\{+\infty > \theta_1 \geq \ldots \theta_H > -\infty\}. \tag{11}$$

Prior (9) is proper, while the priors (10) and (11) are improper. The posterior distributions are nonetheless proper if we demand, for the Poisson probability model and the Gaussian probability model, that in (8) at least one dyad must be placed at level 1. In other words, the prior is the uniform distribution on the class of ultrametric distances satisfying $\min\{d(i,j) \mid i \neq j\} = 1$. For the Gaussian probability model, we additionally have to demand that at least one vertex must be placed at level H. These restrictions are weak ones, since the ultrametric axioms are invariant under monotonic transformations anyway; of all dyads, one or more dyads must be closest, and one or more dyads must be most distant.

When we summarize our knowledge about θ by uniform priors, then $P(\theta \mid d)$, and hence $P(d, \theta)$, is constant. This implies that the posterior distribution

$$P(d, \theta \mid x) \quad \propto \quad \begin{cases} P(x \mid d, \theta) & \text{if } d \text{ is ultrametric} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

under the restriction $\theta_1 \geq \cdots \geq \theta_H$, is proportional to the likelihood function. Taking the natural logarithm of the posterior distribution shows that the nuisance parameter σ^2 in the Gaussian model is a multiplicative constant; this implies that σ^2 needs not be estimated.

3.1. *Sampling d Conditional on θ*

With the uniform prior for d, the conditional probability function of d given θ and x is proportional to $P(x \mid d, \theta)$ on the space of all ultrametrics d assuming values in $\{0, 1, \ldots, H\}$, but the proportionality constant is unknown. As an algorithm that asymptotically generates random draws from this conditional distribution, the Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller, 1953, Hastings, 1970) can be used. The MH algorithm generates a sequence of ultrametrics, and the iterative procedure for generating the next ultrametric is as follows. Some stochastic mechanism is employed for proposing a new element in the sequence. We denote the current ultrametric by d and the proposed ultrametric by $\tilde{d}$. The iterative procedure either moves from d to $\tilde{d}$, or stays at d. According to the Metropolis-Hastings algorithm, the probability that the algorithm moves from state d to $\tilde{d}$ equals

$$\alpha(d, \tilde{d}) = \min\left(1, \frac{P(x \mid \tilde{d}, \theta)\ q(d \mid \tilde{d})}{P(x \mid d, \theta)\ q(\tilde{d} \mid d)}\right),$$

where $q(d \mid \tilde{d})$ and $q(\tilde{d} \mid d)$ denote the probability to move from $\tilde{d}$ to d and from d to $\tilde{d}$, respectively, according to the proposal distribution. The proposal distribution is presented in Appendix A.

To obtain the posterior distribution $P(d, \theta \mid x)$ as the limiting distribution of this process, two regularity conditions must hold, irreducibility and aperiodicity. Irreducibility is proved in Appendix B. Aperiodicity is ensured by irreducible Metropolis-Hastings kernels (Nummelin, 1984, Section 2.4).

The usual point estimate in Bayesian statistics, the posterior mean, is not applicable to ultrametric distances, because the space of all ultrametric distances is not a linear space. To obtain Monte Carlo estimates of the distances $d(i, j)$, we observe that neither the arithmetic means nor the medians of the generated $d(i, j)$ over the post-burn-in iterations necessarily defines an ultrametric. In our experience, the matrix with the posterior medians is often ultrametric, or nearly so, if the adjacency matrix shows a tendency

to transitivity. Since the ultrametric axioms are invariant under monotonic transformations, the posterior medians are a natural choice anyway. Therefore, we use as Monte Carlo estimates of the distances the medians of the generated $d(i,j)$ over the post-burn-in iterations. One advantage is that doing so may yield overlapping settings. Hence the posterior medians alleviate the problem of assuming strict transitivity to some extent.

3.2. Sampling θ Conditional on d

For drawing $\theta = (\theta_1, \ldots, \theta_H)'$ from $P(d, \theta \mid x)$, note that the elements of θ are strongly dependent because of the order restriction. They are drawn successively from the posterior distribution. For each draw, since the cumulative distribution function is not readily invertible, the Acceptance Rejection method (Press, Flannery, Teukolsky, and Vetterling, 1986, Fishman, 1996) is used. This method generates values for θ as follows from the posterior distribution. Beginning with θ_1, from the uniform distribution on the interval $[\theta_2, K[$ - where K denotes the upper bound according to the prior distribution - some candidate point θ_1^* is sampled and accepted with probability

$$g_1(\theta_1^*) = \frac{p(d, \theta_1^*, \theta_2, \ldots, \theta_H \mid x)}{c}, \tag{13}$$

where c is the supremum given by

$$c = \sup_{\theta_1} \; p(d, \theta_1, \theta_2, \ldots, \theta_H \mid x). \tag{14}$$

This procedure is repeated until one candidate point is accepted. Thereafter, some candidate θ_2^* is sampled from the uniform distribution on $[\theta_3, \theta_1]$ and accepted with probability $g_2(\theta_2^*)$. This procedure is applied in an analogous manner to $\theta_3, \ldots, \theta_H$.

To obtain starting values for $(\theta_1, \ldots, \theta_H)'$, the PAVA algorithm (Section 4.2) is applied to the initial ultrametric.

A good Monte Carlo estimate of the posterior mean $E(\theta_h \mid x)$ is the average of θ_h over the post-burn-in iterations. The standard deviation of the sampled $\hat{\theta}_h$ can be regarded as an estimate of the posterior standard devation and may

be considered to be an approximation of the standard error of estimation. The matrix with the posterior means $E(\theta_{d(i,j)} \mid x)$ can be regarded as the predictive posterior value of the edge between vertices i and j.

3.3. A Hybrid MCMC Algorithm

The approach used above to sample from the posterior distribution is a hybrid Markov Chain Monte Carlo approach. This stems from the fact that we conditionally sample d, given x and θ, by using the Metropolis-Hastings algorithm, and then we conditionally sample θ from the posterior distribution, given x and d, using the Acceptance Rejection method. The corresponding transition kernels give a cycle kernel which is itself a transition kernel. The Markov chains defined by these transition kernels are irreducible and aperiodic, which was proved for the first kernel, and which is trivial to show for the second one. Hence the kernel of the cycle is irreducible and aperiodic as well (Tierney, 1994). According to Gamerman (1997, section 6.4.1), the stationary distribution of the Markov chain defined by the cycle kernel is the posterior distribution $P(d, \theta \mid x)$.

4. Maximum Likelihood Estimation

In the maximum likelihood framework, the entities d and θ are treated as latent but fixed parameters. The profile likelihood can be maximized over (d, θ) by the Simulated Annealing method.

4.1. Maximizing over d

The *Simulated Annealing* method (e.g., Press, Flannery, Teukolsky, and Vetterling, 1986, Pflug, 1996, Häggström, 2002) is well-suited to discrete optimization problems with complicated state spaces, as is the case with the model introduced in Section 2. The Simulated Annealing method is based on the work of Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) and maximizes an objective function defined on a discrete state space. The

state space is defined here as the set of all ultrametrics with values $0, 1, \ldots, H$. This state space is finite but very large. It is intended to find the ultrametric(s) d for which the profile likelihood

$$\mathfrak{L}_p(x \mid d) \;=\; \begin{cases} \max\limits_{\theta} P(x \mid d, \theta) & \text{if } d \text{ is ultrametric} \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

is maximal, where the specific form of $P(x \mid d, \theta)$ is given by Equation (2), Equation (5), or Equation (6), while the maximum over θ is taken, subject to the constraint $\theta_1 \geq \cdots \geq \theta_H$.

The Simulated Annealing method exploits the fact that extrema are preserved under monotonic transformations. A Markov chain is constructed with the so-called Boltzmann distribution with probability function

$$\pi(d) = \kappa_{\mathfrak{L}_p,T}\, \mathfrak{L}_p(x \mid d)^{\frac{1}{T}}$$

as unique stationary distribution, where T denotes the temperature, $T \longrightarrow 0$ as the estimation process approaches the stop criterion[2], and $\kappa_{\mathfrak{L}_p,T}$ denotes the normalizing constant. The Boltzmann distribution can be simulated by MCMC methods. The Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller, 1953) - the special case of the Metropolis-Hastings algorithm which assumes that the probabilities $q(d \mid \tilde{d})$ and $q(\tilde{d} \mid d)$ are equal - can be used, with the same proposal distribution as in Section 3.1. Using this proposal distribution implies that $q(d \mid \tilde{d})$ and $q(\tilde{d} \mid d)$ are not necessarily equal, which means that the Metropolis-Hastings algorithm is more appropriate; however, since we focus on the global mode of the likelihood function rather than on the whole Boltzmann distribution, we can save computation costs by using the (computationally less expensive) Metropolis algorithm. For this algorithm, the normalizing constant $\kappa_{\mathfrak{L}_p,T}$ - which is in practice infeasible to calculate - cancels, since the acceptance probability is the ratio of two Boltzmann distributions.

[2]The stop criterion is the number of iterations to be executed, and is determined beforehand.

The computational efficiency strongly depends on the annealing scheme, which depends on how fast the temperature T is reduced. We discuss this issue in Section 7 in more detail. The essential idea behind the annealing scheme is that by sometimes taking proposed downhill steps the algorithm is able to escape from local maxima.
The algorithm, when exploring the Boltzmann distribution, keeps track of the up to now maximal value of the profile likelihood and the corresponding d and θ.

4.2. Maximizing over θ

The profile likelihood is evaluated on each iteration. The maximum likelihood estimate of θ under the order restriction $\theta_1 \geq \cdots \geq \theta_H$ is needed to evaluate the profile likelihood. Maximization over θ is easy in case there is no order restriction and leads to $\hat{\theta}_h = s_h(x)/m_h$, with m_h and $s_h(x)$ defined as above and computed on the basis of $\tilde{d}$ and x. However, when there is an order restriction, then maximizing over θ is more complicated. The estimation of θ under the order restriction $\theta_1 \geq \cdots \geq \theta_H$ can be solved by using the Pool Adjacent Violators Algorithm (PAVA) (Barlow, Bartholomew, Bremner, and Brunk, 1972, Robertson, Wright, and Dykstra, 1988). This algorithm goes back to Ayer, Brunk, Ewing, Reid, and Silverman (1955) and uses antitonic regression to smooth the curve defined by $\theta_1, \ldots, \theta_H$. The algorithm starts with the estimates $\hat{\theta}_h = s_h(x)/m_h$. Each such estimate forms one so-called solution block. Then the PAVA algorithm checks the order of the solution blocks. If the order of the solution blocks is non-increasing, then the estimate $\hat{\theta}$ is the maximum likelihood estimate of θ under the order restriction $\theta_1 \geq \cdots \geq \theta_H$, and the algorithm stops. Otherwise the PAVA algorithm starts at the first solution block and proceeds down to the last solution block until it encounters the first solution block which violates the order by $\hat{\theta}_h > \hat{\theta}_{h-1}$. The estimates $\hat{\theta}_{h-1}, \hat{\theta}_h$ are then pooled into one solution block by computing

$$\hat{\theta}_{h-1} = \hat{\theta}_h = \frac{s_{h-1} + s_h}{m_{h-1} + m_h}.$$

If the obtained order of the solution blocks is non-increasing, then the obtained $\hat{\theta}$ is the final estimate and the algorithm halts. Otherwise the algorithm continues to pool adjacent solution blocks until the order of the solution blocks is no more violated.
The cited literature gives the proof that this algorithm maximizes the likelihood as a function of θ under the restriction of a non-increasing ordering.

5. Implementation and Model Determination

In this section, Markov chain mixing, convergence, model checking, as well as model selection are treated.

5.1. Mixing

It is important that the Markov chain constructed by the methods in Sections 3 and 4 mixes well. Heating the target distribution in the beginning can aid. When simulating the Boltzmann distribution, the temperature should exceed 1 in the beginning. When simulating the posterior distribution, an analogoue can be used by setting

$$\alpha(d, \tilde{d}) = \alpha(d, \tilde{d}) + u$$

where $\alpha(d, \tilde{d})$ is the probability to move from d to $\tilde{d}$, and $u = u(t)$ is a function of the current iteration number t, equals 0.5 on iteration $t = 0$ and tends to 0 as the end of the first half of the burn-in is approached. The rationale behind this heating scheme is that support from the posterior distribution for proposals is in the beginning less necessary than later. Heating the posterior of d, given x and θ, can be considered as an analogue to starting with an overdispersed distribution for the parameters of interest, which is advised by Gelman and Rubin (1992).

5.2. Convergence

Convergence in the case of Simulated Annealing can be checked by running multiple runs with starting points sampled from an overdispersed distribu-

tion. In the Bayesian approach, it usually is assumed that the marginal distribution of the Markov chain has converged to the posterior distribution after a good number of initial burn-in iterations. A well-known method to check convergence is to run multiple independent Markov chains with starting points obtained from an overdispersed distribution (Gelman and Rubin, 1992). A simple way to obtain such startings points is to heat the target distribution (see Section 5.1) and run the Markov chain some time. Given multiple Markov chains, the so-called *Estimated Potential Scale Reduction* (EPSR) (Gelman, 1996, Section 8.4) can be computed. This is, for a given parameter of interest, the ratio of the between-chains variance to the within-chains variance. Before the Markov chains converge to the stationary distribution, the between-chains variance is an overestimate of the posterior standard deviation, while the within-chains variance is an underestimate of the posterior standard deviation. When the Markov chains converged to the stationary distribution, the two quantities should be approximately the same. We compute the EPSR for each θ_h. The EPSR should be close to 1 for each θ_h. Times series of the θ_h can be used as an additional tool to detect non-convergence.

5.3. Model Selection

The estimations will be carried out conditional on the number of ultrametric levels H. Model selection with regard to H can be based on Bayes factors (Kass and Raftery, 1995). To account for model uncertainty, we take one baseline model M_K with K ultrametric levels, where K is considered an upper bound to the true number of levels, and compute Bayes factors

$$B_{Kk} = \frac{P(x \mid M_K)}{P(x \mid M_k)}. \tag{16}$$

The models M_k are models with $k < K$ ultrametric levels, which are compared to M_K. The probabilities $P(x \mid M_k)$ in Equation (16) can be approxi-

mated by

$$\hat{P}(x \mid M_k) = \left(\frac{1}{L} \sum_{l=1}^{L} \left[P(x \mid d^{(l)}, \theta^{(l)}) \right]^{-1} \right)^{-1} \tag{17}$$

where L denotes the number of ultrametrics sampled from the posterior distribution (Newton and Raftery, 1994, Kass and Raftery, 1995). Then the posterior probabilities of the models M_k can be calculated,

$$P(M_k \mid x) = \frac{B_{Kk}}{\sum_{s=1}^{S} B_{Ks}} \tag{18}$$

where S is the number of models M_k compared to M_K, and all prior odds equal 1. Models with low posterior probabilities should be removed. The remaining models can be regarded as plausible models.[3]

5.4. Model Checking

Since the Bayesian approach captures the uncertainty about d and θ, we focus in this section on the Bayesian approach. To investigate the uncertainty about d, the concept of entropy (Shannon, 1948) is suited par excellence. The uncertainty about dyad (i, j), $j < i$, is expressed by

$$u_{ij} = u_{ij}(q_{ij}^{(1)}, \ldots, q_{ij}^{(H)}) = -\sum_{h=1}^{H} \left[q_{ij}^{(h)} \, {}^2\log \, q_{ij}^{(h)} \right] \tau^{-1}$$

where $q_{ij}^{(h)}$ gives the relative frequency of $\{d(i, j) = h\}$ among the ultrametrics sampled from the posterior distribution, ${}^2\log$ is the logarithm to the base 2, and $\tau = -{}^2\log \, 1/H$ is the normalizing constant (cf. Mathai and Rathie, 1975). The quantity u_{ij} takes values in the interval $[0, 1]$. The value 1 indicates maximum uncertainty, the value 0 indicates minimum uncertainty. The uncertainty about the partition can be quantified by

$$-\binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i-1} u_{ij},$$

[3]As Raftery (1995) and Kass and Raftery (1995) pointed out, model uncertainty is neglected whenever one selects a single 'true' model; it is more sensible to focus on a class of reasonable models rather than on a single model.

and vertices which contribute exceptionally much to this uncertainty can be identified by

$$\frac{-1}{n-1} \sum_{j=1,\, j\neq i}^{n} u_{ij}.$$

We note that the entropies can in most cases, due to the restriction that d is ultrametric, not attain the lower bound 0.

Alternative means to check the model are the matrix with the posterior medians of the distances and the matrix with posterior means $E(\theta_{d(i,j)} \mid x)$.

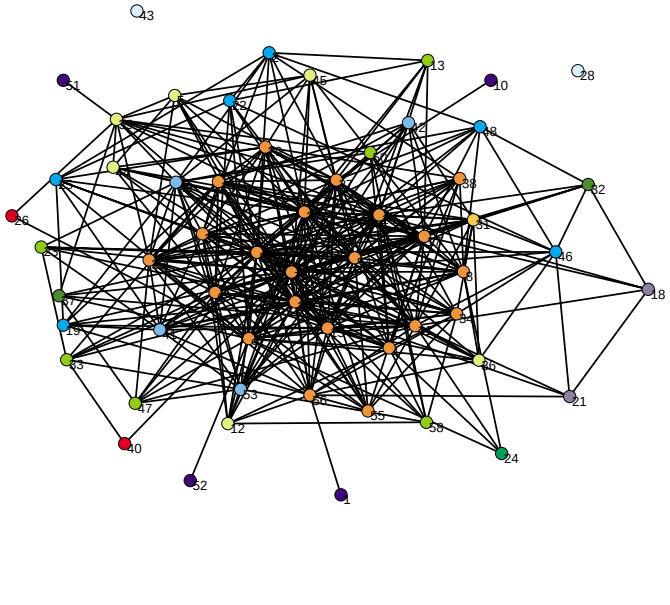
6. Application

We analyze the data collected by Bernard, Killworth, and Sailer (1980). Bernard, Killworth, and Sailer studied the interactions among 58 students living in a fraternity at a West Virginia college for at least 3 months. The intention was to study informant accuracy, which the research team did by recording how many times any two students had conversation within five days, and thereafter asking each student how much (s)he interacted with the other students in the five days.

We apply the model to the recorded interaction frequencies. This relation is symmetric. The network is given in Figure 1. Students are drawn as colored points; the colors correspond to partitions obtained by the core-routine in `Pajek` (Batagelj and Mrvar, 2003). The lines represent the number of conversations between students; the number of dyadic conversations varies between 0 and 51. Thicker lines correspond to more conversations. For reasons of clarity, a line is drawn only if the students had more than 2 conversations; this resulted in a 60% reduction in lines.

Since the data can be considered as count data, the Poisson probability model is the convenient choice. We begin by selecting the number of ultrametric levels H. Using the Poisson probability model, we run multiple sequences with $k = 20$ and with $k = 10$ ultrametric levels. Such high-dimensional spaces can cause convergence problems, even in the Bayesian case. Though our convergence checks gave no indication whatsoever to suspect non-convergence,

Figure 1: Fraternity data: observed network



convergence might still be doubted. Nevertheless we interpret the results, since we will see below that models with less dimensions point into the same direction. The posterior means $E(\theta^{(k)} \mid x)$ for k ultrametric levels are shown in Table 2. When comparing $E(\theta^{(20)} \mid x)$ with $E(\theta^{(10)} \mid x)$, it seems that the levels 6 - 20 could be merged to 2 or 3 levels without losing essential information. We therefore execute multiple sequences with $k = 9, 8, 7, 6, 5, 4, 3, 2$ ultrametric levels. Note that the model with $H = 1$ ultrametric level is trivial. Comparing $E(\theta^{(k)} \mid x)$ for $k = 9, 8, 7, 6$ ultrametric levels reveals that in particular $E(\theta_1^{(k)} \mid x)$ and $E(\theta_2^{(k)} \mid x)$ point roughly into the same direction. Moving from M_6 to M_5, and subsequently to M_4, changes the picture slightly. But when moving from model M_4 with $k = 4$ ultrametric levels to the model M_3 with $k = 3$ ultrametric levels, $E(\theta^{(k)} \mid x)$ changes considerably. The conditional probabilities $-\ln P(x \mid M_k)$ of the data conditional on the models with $k > 6$ ultrametric levels appear to vary between 2,810 and 2,830,

Table 2: Model selection: posterior means $E(\theta^{(k)} \mid x)$

model M_k	M_{20}	M_{10}	M_9	M_8	M_7	M_6	M_5	M_4	M_3	M_2
$E(\theta_1 \mid x)$	24.22	24.34	24.22	24.31	24.22	23.82	23.80	22.18	13.93	6.54
$E(\theta_2 \mid x)$	7.31	7.78	7.31	7.67	7.31	6.23	6.21	5.50	3.48	1.15
$E(\theta_3 \mid x)$	3.52	4.42	3.52	4.27	3.53	2.52	2.37	1.68	.95	
$E(\theta_4 \mid x)$	1.58	2.33	1.57	2.12	1.58	1.56	1.12	.49		
$E(\theta_5 \mid x)$	.74	1.47	.72	1.23	.74	.90	.30			
$E(\theta_6 \mid x)$	.28	.68	.22	.52	.22	.25				
$E(\theta_7 \mid x)$	.26	.21	.17	.17	.11					
$E(\theta_8 \mid x)$	.24	.16	.11	.09						
$E(\theta_9 \mid x)$	.22	.11	.06							
$E(\theta_{10} \mid x)$	.20	.05								
$E(\theta_{11} \mid x)$	.19									
$E(\theta_{12} \mid x)$	.17									
$E(\theta_{13} \mid x)$	.15									
$E(\theta_{14} \mid x)$	.13									
$E(\theta_{15} \mid x)$	.11									
$E(\theta_{16} \mid x)$	.09									
$E(\theta_{17} \mid x)$	.07									
$E(\theta_{18} \mid x)$	.06									
$E(\theta_{19} \mid x)$	.04									
$E(\theta_{20} \mid x)$	.02									

but even long runs do not agree about the exact values. Furthermore, the probabilities $-\ln P(x \mid M_7), \ldots, -\ln P(x \mid M_{10}), -\ln P(x \mid M_{20})$ must be non-decreasing, but we encounter deviations. Part of the problem could be that the estimator for $-\ln P(x \mid M_k)$ is unstable (Kass and Raftery, 1995). We suspect additionally that the posterior distribution is not well behaved, in the sense that for $k > 6$ ultrametric levels there is no single dominant posterior mode. This was found by inspecting output from multiple (long) runs for each possible number of ultrametric levels. We therefore seek an as low-dimensional representation as possible, with as few ultrametric levels as possible. According to the (natural) logarithmic scale in Kass and Raftery, there is very strong evidence against the models with $k \leq 6$ levels, as is shown in Table 3. The model M_6 with 6 levels has unit posterior probability within this set, but the log Bayes factor (see the row $2 \log B_{7k}$) indicates that model M_7 can predict the data considerably better than M_6. On the basis of

Table 3: Model selection: posterior modes $P(d^{(k)}, \theta^{(k)} \mid x)$, posterior probabilities $P(M_k \mid x)$

model M_k	M_6	M_5	M_4	M_3	M_2
$-$ mod $P(x \mid d, \theta)$	2,793.67	2,825.81	2,882.43	3,065.12	3,319.92
$-\log \hat{P}(x \mid M_k)$	2,840.72	2,865.31	2,918.30	3,143.65	3,340.82
$2 \log B_{7k}$	37.02	86.20	192.18	642.88	1,037.22
$P(M_k \mid x)$	1.0000	.0000	.0000	.0000	.0000

this evidence, we decide that the models with less than 7 ultrametric levels are clearly inappropriate.

We now focus on the model with $k = 7$ ultrametric levels to present information on the settings structures, since this model seems to be reasonable and is still estimable. Minus the logarithm of the posterior mode is 2,774.47, and minus the mean of the log likelihoods equals 2,797.51. Minus the logarithm of the initial likelihood - obtained by ordinary hierarchical cluster analysis (Johnson, 1967) - equals 4,394.87, demonstrating that the Bayesian estimation procedure yields in this case much better results than the classical clustering heuristic. Table 4 compares the posterior mean of θ, given the data, with the maximum likelihood estimate of θ. The posterior means are quite close to the ML estimates. The ultrametric which corresponds to the global maximum of the likelihood function is presented in Figure 2 by a Venn diagram. This two-dimensional drawing can be interpreted as social topological map: it shows a three-dimensional mountain drawn in two dimensions with the levels being represented by colors; darker colors indicate higher levels, with level 1 being the highest level, and level 7 being the lowest level. The expected interaction frequencies which correspond to the levels are given by the estimates $\hat{\theta}_1, \ldots, \hat{\theta}_7$ in Table 4. The students are represented by the integers 1 to 58. The settings can be derived as follows. Suppose the three-dimensional mountain is cut horizontally at some level. The result would be some mountain summits above the cut. Take any mountain summit and

Table 4: MLE $\hat{\theta}$ and posterior mean $E(\theta^{(7)} \mid x)$

level h	$E(\theta_h \mid x)$	posterior S.D.	MLE $\hat{\theta}_h$	S.E.
level-1	24.22	(1.44)	24.62	(1.38)
level-2	7.31	(.33)	7.45	(.28)
level-3	3.53	(.17)	3.52	(.13)
level-4	1.58	(.06)	1.60	(.05)
level-5	.74	(.05)	.80	(.05)
level-6	.22	(.04)	.40	(.05)
level-7	.11	(.06)	.16	(.03)

the students placed on it; the students on this mountain summit share the setting. That is, the mountain summits correspond to settings, and there are as many settings as mountain summits. Let us cut the mountain in Figure 2 at level 2. We obtain 8 settings, corresponding to the 8 subsets {3, 6, 7, 14, 15, 16, 17, 20, 27, 29, 30, 35, 54, 57}, {8, 31}, {4, 9}, {33, 53}, {11, 55, 56}, {19, 23, 41, 49}, {2, 13, 22}, and {5, 39, 45}. The expected number of conversations between students in the same setting at level 2 is 7.45 (according to Table 4). Now let us cut the mountain at level 1. We obtain 4 settings, corresponding to the 4 subsets {3, 6, 7, 20, 57}, {16, 17}, {29, 35}, and {5, 45}. The expected number of conversations between students in the same setting at level 1 is 24.62. The expected number of conversations between students in setting {3, 6, 7, 20, 57} and students in setting {16, 17} is 7.45. This is much less than the expected number of conversations within the settings at level 1, which is 24.62. Thus, there is much more interaction within settings than between settings. The settings are, in addition, non-overlapping and nested, as can readily be observed.

The matrix with the posterior medians of the distances (not shown) deviates slightly from the ML ultrametric, and is, not surprisingly, not completely ultrametric, but nearly so. While the posterior medians agree by and large

with the ML ultrametric, they slightly disagree about the settings (according to the ML ultrametric) $\{12, 50, 58\}$ (level 3) and $\{18, 21, 32, 46\}$ (level 4). Students 50 and 46 are mainly responsible for the disagreement. Student 37 is moved to student 47 to share the setting with 47, given level 3. In addition, the students 10, 26, 28, 43, 51 are moved one level upwards, to the contours at level 5 and 6, implying that the setting $\{26, 52\}$ (level 6) cancels and that level 7 is redundant.

The uncertainty about the partition can be more explicitely measured by the entropies. The mean entropy over all dyads equals .0590. Since in most cases the entropy cannot - due to the ultrametric restrictions - attain the lower bound 0, this low value is encouraging. The students whose position in the ultrametric space is somewhat in question are the students 10, 46, 48, 50. This was already (partly) suggested by the deviations of the posterior means from the ML distances.

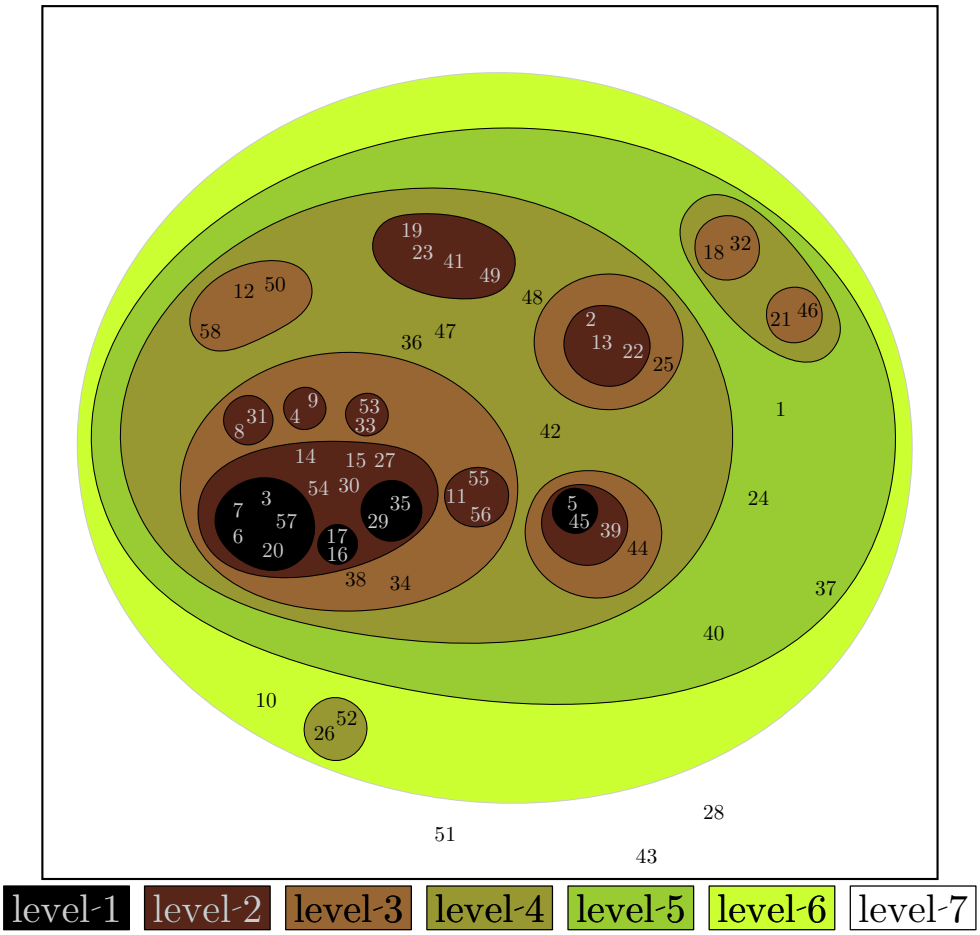
7. Discussion

A class of statistical models was described which models settings in social networks by assuming that the observed network has been generated by latent transitive structures, and the expected tie strength increases with decreasing ultrametric distance.

This method is flexible in the sense that it can be applied to a variety of data types (dichotomous, count, continuous) and that it easily accommodates randomly missing data.

From applications to empirical data sets as well as (simple) artificial data sets, it can be concluded that in particular the Bayesian approach performs very well. When the observed network exhibits strong tendencies towards transitivity, but some vertices are involved in more non-transitive triples than could be expected on this basis, then the Bayesian analysis will identify them. In such cases the Bayesian approach often hints that some settings overlap. When the structure in the network shows no tendency towards transitivity, then the Bayesian model will communicate this very clearly.

Figure 2: ML ultrametric for fraternity data



We now sketch some shortcomings, and possible model extensions which address these shortcomings. We begin by outlining two limitations concerning the applicability.

One limitation concerns the fact that the Simulated Annealing algorithm for ML estimation, applied to data sets with around 100 (or more) vertices, shows in low-dimensional spaces ($H < 4$) sometimes, and in high-dimensional spaces ($H \geq 4$) most times, no convergence. The algorithm settles down at local maxima. Running multiple sequences will yield multiple maxima, and typically, none is the global maximum. The basic convergence problem

stems from the combination of the annealing scheme, the proposal distribution, and the multimodal shape of the log-likelihood as a function of the ultrametric. According to certain statistical theorems (see, e.g., Geman and Geman, 1984), the Simulated Annealing algorithm converges in probability to the global maximizer when the annealing system meets certain regularity conditions. Unfortunately, using such an annealing system is infeasible in practice, since the number of iterations required is astronomically high, as was also noticed by Häggström (2002). For this reason, the annealing system must be built such that the algorithm converges within reasonable computation time, without meeting the regularity conditions. We have studied many annealing schemes and implemented one which works reasonably over a wide range of data sets. The temperature declines exponentially to zero. Even with this scheme, however, the process, for $n \geq 100$ or $H \geq 4$, often does not converge to the global optimum in practical amounts of computing time. On the other hand, the log-likelihoods of local maxima produced by multiple sequences are quite close together, suggesting that they are close to the global maximum.

The second limitation is that ultrametrics assume symmetry. While this will not affect the basic conclusions when edges tend to be reciprocated, which is known to be the case in many social networks, and in particular in friendship or collaboration networks, it may well affect the conclusions when reciprocity is low. It might be possible, as a model for directed graphs, to develop an ultrametric latent structure model that uses the dyads (X_{ij}, X_{ji}) as units of analysis like in Nowicki and Snijders (2001). Other extensions are also possible, e.g., with additional parameters for the degrees as in the p_1 model (Holland and Leinhardt, 1981) and for other structural effects and covariate effects as in the p^* model (Wasserman and Pattison, 1996).

Another interesting model extension is to model overlapping settings. That can be done by assuming that the observed network was generated from two or more ultrametrics, and assuming that the distribution of X_{ij} depends on the minimum ultrametric distance between i and j (Watts, Dodds, and New-

man, 2002).

Concerning model selection, we proposed to estimate models conditional on the number of ultrametric levels H, and base model selection with respect to H on Bayesian information criteria. An elegant alternative is to estimate the full Bayesian model with H being random, by employing a reversible jumb MCMC algorithm (Green, 1995), as proposed by Richardson and Green (1997) for mixture models with unknown number of components.[4] This involves taking Metropolis-Hastings steps between subspaces with varying numbers of levels H. Simulation output can then be used to select H.

Program This class of settings models is implemented in the program `Ultras` version 1.2, which can be downloaded (incl. manual) free of charge from `http://stat.gamma.rug.nl/stocnet` as part of the `StOCNET` program collection (Boer, Huisman, Snijders, and Zeggelink, 2003). `Ultras` is rapid and can handle missing data as well as huge data sets. The running time is in general $O(n^2)$, but for most parts of the estimation process running time is $O(n)$ instead of $O(n^2)$. Giant networks ($n \geq 1,000$) can be analyzed with the special version `UltrasXL` which reduces computation costs even more.

A. Proposal Distribution

Denote the current ultrametric on iteration t by d and the proposed ultrametric by $\tilde{d}$. Define $m(i) = \min_k[d(i,k)]$ $(k \neq i)$ and

$$
\begin{aligned}
i \prec j \quad & \text{if } m(i) = d(i,j) \ \wedge \ m(j) < d(i,j), \\
i \preceq j \quad & \text{if } m(i) = d(i,j), \\
i \approx j \quad & \text{if } m(i) = m(j) = d(i,j).
\end{aligned}
$$

In Section 6 we interpreted ultrametric structures in terms of mountains on which vertices are placed, like in Figure 2. Using this terminology, we can

[4] Kass and Raftery (1995) mention other possibilities to generate processes moving through the model space, or through the parameter space and the model space simultanously.

illustrate the relations as follows. The entity $m(i)$ gives the level at which vertex i is placed on the mountain. The relation $i \prec j$ means that vertex i is placed on the same mountain summit as vertex j, but that i is placed at a lower level than j; relation $i \preceq j$ means that vertex i is placed on the same mountain summit as vertex j, and that i is placed at a level which is not higher than the level at which j is placed; relation $i \approx j$ means that vertices i and j are placed on the same mountain summit at the same level.

The symbol p gives the probability of taking some action. The proposal generator is described in Table 5. It is used both in the Bayesian approach as well as in the maximum likelihood approach. In the Bayesian case, however, for the Poisson probability model and the Gaussian probability model, one Boolean condition has to be added, which ensures that at least one dyad remains at level 1. For the Bayesian case and the Gaussian probability model, a second Boolean condition is necessary to ensure that at least one vertex is placed at level H.

The proposal generator can be illustrated as follows. Two vertices are sampled at random. Suppose the current ultrametric looks like the ultrametric in Figure 2 which is displayed as mountain on a social topological map. Then one of the two sampled vertices is moved one level downwards (steps 1.1, 1.2.1, 2.1, 3.2.1, 3.2.2.1), or one of the two sampled vertices is moved one level upwards (steps 2.2, 3.2.2.2.1), or on the vertices' current plateau a new plateau is build on which the two vertices are placed (steps 1.2.2, 1.3, 3.2.2.2.2), or the positions of the two vertices in the ultrametric space are interchanged (step 3.1). Only one of the possible steps is taken, and the resulting ultrametric is stored in $\tilde{d}$.

The computation of the proposal distribution $q(\tilde{d} \mid d)$, given this proposal generator, is straightforward, and involves only the basic rules of probability.

B. Proof of Irreducibility

This appendix uses the same notation as Appendix A. The proof below is valid for both the Bayesian approach and the maximum likelihood approach,

Table 5: Proposal Generator

Sample two vertices $i \neq j \in \mathfrak{N}$ at random; if $j \prec i$, then interchange i and j.[5]

If $i \approx j$, then transform d into $\tilde{d}$ as follows:

1.1 if $d(i,j) = 1$, then set $\tilde{d}(i,k) = 2 \; \forall \; k$ with $k \approx i, \; k \neq i$;

1.2 if $2 \leq d(i,j) \leq H-1$, then set

- 1.2.1 with $p = .5$ $\tilde{d}(i,k) = d(i,k) + 1 \; \forall \; k$ with $i \preceq k, \; k \neq i$;
- 1.2.2 with $p = .5$ $\tilde{d}(i,j) = d(i,j) - 1$;

1.3 if $d(i,j) = H$, then set $\tilde{d}(i,j) = H - 1$;

else if $i \prec j$, then transform d into $\tilde{d}$ by setting

2.1 with $p = .5$ $\tilde{d}(j,k) = d(j,k) + 1 \; \forall \; k$ with $j \preceq k, \; k \neq j$;

2.2 with $p = .5$ $\tilde{d}(i,k) = d(i,k) - 1 \; \forall \; k$ with $i \prec k$ and $d(j,k) < d(i,k)$;

else if not$\{i \preceq j \; \vee \; j \preceq i\}$, then transform d into $\tilde{d}$ by setting

3.1 with $p = .5$ $\tilde{d}(i,k) = d(j,k) \; \forall \; k \neq i,j$ and $\tilde{d}(j,k) = d(i,k) \; \forall \; k \neq i,j$;

3.2 with $p = .5$

- 3.2.1 with $p = .5$ $\tilde{d}(i,k) = d(i,k) + 1 \; \forall \; k$ with $i \preceq k, \; k \neq i$;
- 3.2.2 with $p = .5$
  - 3.2.2.1 if $m(i) = 1$, then set $\tilde{d}(i,k) = 2 \; \forall \; k$ with $k \approx i, \; k \neq i$;
  - 3.2.2.2 else check whether there exists some k such that $i \prec k$;
    - 3.2.2.2.1 if yes, then sample one such vertex k at random and set $\tilde{d}(i,l) = d(i,l) - 1 \; \forall \; l$ with $i \prec l$ and $d(k,l) < d(i,l)$;
    - 3.2.2.2.2 if not, then sample one vertex k with $k \approx i, \; k \neq i$, and set $\tilde{d}(i,k) = d(i,k) - 1$;

Set $\tilde{d}(k,m) = d(k,m) \; \forall \; k \neq m \in \mathfrak{N}$ which are not updated yet.

[5] This (possible) interchange is only needed for 2.1 and 2.2.

but in the maximum likelihood approach $d^{\max}$ denotes the ultrametric with all entries equal to H, while in the Bayesian approach, for the Poisson probability model or the Gaussian model, $d^{\max}$ contains one entry equal to 1, and the remaining entries are equal to H.

Proof of irreducibility of $q(\tilde{d} \mid d)$. Given any ultrametric $d \neq d^{\max}$ on iteration t, an ultrametric $\tilde{d} \neq d$ might be proposed such that $\tilde{d}(i,j) \geq d(i,j)$ holds for all $i, j \in \mathfrak{N}$. This is true since on any iteration t there exist two distinct vertices $i, j \in \mathfrak{N}$ such that $\tilde{d}(i,k) = d(i,k) + 1$ for all vertices $k \neq i$ with $i \preceq k$ is proposed, whereby j may be interchanged with i. Since this is true for any ultrametric d and any iteration t, there exists a positive probability that the algorithm arrives within finitely many iterations at $d^{\max}$. We notice that there exists a positive probability that the two vertices (i,j), which were sampled on iteration t, are sampled on iteration $t+1$ again. Denote the ultrametric on iteration $t+1$ by b. We observe that the probability that on iteration $t+1$ an ultrametric $\tilde{d}$ with $\tilde{d}(i,k) = b(i,k) - 1$ for all k to which $\tilde{d}(i,k) = d(i,k) + 1$ was applied on iteration t, is proposed, with d as the result, is positive. Since this is true for any ultrametric d and any iteration t, any step the algorithm has taken on the way from d to $d^{\max}$ is invertible. This implies that there exists a positive probability that the algorithm goes back from $d^{\max}$ to d within finitely many iterations. Hence there exists a positive probability that the algorithm moves from an arbitrary ultrametric d to the ultrametric $d^{\max}$ and back again within finitely many iterations. Since d is arbitrarily chosen, this implies that there is a probabilistic path from any ultrametric to any other ultrametric via the ultrametric $d^{\max}$.□

Irreducibility of $q(\tilde{d} \mid d)$ is necessary but not sufficient, since the (ir)reducibility of the Metropolis-Hastings kernel depends on both $q(\tilde{d} \mid d)$ and $P(x \mid \tilde{d}, \theta)$. For the Metropolis-Hastings kernel to be irreducible,

$$P(x \mid \tilde{d}, \theta) > 0$$

must hold for all possible proposals $\tilde{d}$. This can readily be verified by observing that d enters $P(x \mid \tilde{d}, \theta)$ through θ, and that θ, restricted as in Section 2.4, necessarily leads to positive values of $P(x \mid \tilde{d}, \theta)$ for all $\tilde{d}$.

References

Ayer, M., H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics 26*, 641–647.

Barlow, R. E., D. J. Bartholomew, J. M. Bremner, and H. D. Brunk (1972). *Statistical Inference under Order Restrictions. The Theory and Application of Isotonic Regression.* Chichester: Wiley.

Batagelj, V. and A. Mrvar (2003). *Pajek. Program for Large Network Analysis.* Ljubljana: University of Ljubljana. To download from `http://vlado.fmf.uni-lj.si/pub/networks/pajek/doc/pajekman.htm`.

Bernard, H., P. Killworth, and L. Sailer (1980). Informant accuracy in social network data IV. *Social Networks 2*, 191–218.

Boer, P., M. Huisman, T. A. B. Snijders, and E. P. H. Zeggelink (2003). *StOCNET User's Manual. Version 1.4.* Groningen: ProGAMMA / ICS. To download from `http://stat.gamma.rug.nl/stocnet`.

Borgatti, S. P., M. G. Everett, and P. R. Shirey (1990). LS sets, lambda sets, and other cohesive subsets. *Social Networks 12*, 337–358.

Corter, J. E. (1996). *Tree Models of Similarity and Association.* London: Sage.

Davis, J. A. (1979). *The Davis/Holland/Leinhardt studies: an overview*, Chapter 4, pp. 51–62. New York: Academic Press.

Davis, J. A., P. W. Holland, and S. Leinhardt (1971). Comments on Professor Mazur's hypothesis about interpersonal sentiments. *American Sociological Review 36*, 309–311.

De Soete, G. (1986). Optimal variable weighting for ultrametric and additive tree clustering. *Quality and Quantity 20*, 169–180.

Dempster, A. P., N. M. Laird, and R. B. Rubin (1977). Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society. Section B 39*, 1–38.

Durkheim, E. (1984). *The Division of Labor in Society.* London: Macmillan.

Fishman, G. S. (1996). *Monte Carlo. Concepts, Algorithms, and Applications.* New York: Springer.

Frank, O. (1978). Inferences concerning cluster structure. In *Compstat 1978. Proceedings in Computational Statistics*, pp. 259–265. Wien: Physica-Verlag.

Frank, O. (1980). Transitivity in stochastic graphs and digraphs. *Journal of Mathematical Sociology 7*, 199–213.

Frank, O. and F. Harary (1982). Cluster inference by using transitivity indices in empirical graphs. *Journal of the American Statistical Association 77*, 835–840.

Freeman, L. C. (1992). The sociological concept of 'group': an empirical test of two models.

American Jounal of Sociology 98, 152–166.

Gamerman, D. (1997). *Markov Chain Monte Carlo. Stochastic Simulation for Bayesian Inference*. London: Chapman & Hall.

Gelman, A. (1996). *Inference and monitoring convergence*, Chapter 8, pp. 131–144. London: Chapman & Hall.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*, 457–472.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intellegence 6*, 721–741.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology 78*, 1360–1380.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*, 711–732.

Häggström, O. (2002). *Finite Markov Chains and Algorithmic Applications*. Cambridge: Cambridge University Press.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97–109.

Hausdorff, F. (1934). Über innere Abbildungen. *Fundamenta Mathematicae 23*, 279–291. Contained in F. Hausdorff (forthcoming). *Gesammelte Werke III. Deskriptive Mengenlehre und Topologie*. Heidelberg: Springer.

Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association 97*, 1090–1098.

Holland, P. W. and S. Leinhardt (1970). A method for detecting structure in sociometric data. *American Journal of Sociology 76*, 492–513.

Holland, P. W. and S. Leinhardt (1972). Some evidence on the transitivity of positive interpersonal sentiment. *American Journal of Sociology 77*, 1205–1209.

Holland, P. W. and S. Leinhardt (1976). Local structure in social networks. *Sociological Methodology*, 1–45.

Holland, P. W. and S. Leinhardt (1979). *Structural Sociometry*, Chapter 4, pp. 63–83. New York: Academic Press.

Holland, P. W. and S. Leinhardt (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association 76*, 33–65.

Homans, G. C. (1950). *The Human Group*. New York: Harcourt, Brace.

Hu, S. T. (1966). *Introduction to Contemporary Mathematics*. San Francisco: Holden-Day.

Huelsenbeck, J. P. and F. Ronquist (2001). MRBAYES: Bayesian inference of phylogenetic

trees. *Bioinformatics 17*, 754–755.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika 32*, 241–254.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Mathai, A. M. and P. N. Rathie (1975). *Basic Concepts in Information Theory and Statistics*. New Delhi: Wiley Eastern Limited.

Mazur, A. (1971). Comments on Davis' graph model. *American Sociological Review 36*, 308–311.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics 21*, 1087–1092.

Newton, M. A. and A. E. Raftery (1994). Approximate Bayesian inference with the weighted likelihoods bootstrap. *Journal of the Royal Statistical Society. Section B 56*, 3–48.

Nowicki, K. and T. A. B. Snijders (2001). Estimation and prediction for stochastic block-structures. *Journal of the American Statistical Association 96*, 1077–1087.

Nummelin, E. (1984). *General Irreducible Markov chains and Non-Negative Operators*. Cambridge: Cambridge University Press.

Pattison, P. and G. Robins (2002). Neighborhood-based models for social networks. In R. Stolzenberg (Ed.), *Sociological Methodology*, Volume 32, Chapter 9, pp. 301–337. Boston: Blackwell Publishing.

Pflug, G. C. (1996). *Optimization of stochastic models. The interface betweeen simulation and optimization*. Boston: Kluwer Academic.

Press, S. J. (1989). *Bayesian Statistics Principles, Models, and Applications*. Chichester: Wiley.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling (1986). *Numerical Recipes. The Art of Scientific Computing*. Cambridge: Cambridge University Press.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). *Sociological Methodology 15*, 111–196.

Rapoport, A. (1953a). Spread of information through a population with socio-structural bias: I. Assumption of transitivity. *Bulletin of Mathematical Biophysics 15*, 523–533.

Rapoport, A. (1953b). Spread of information through a population with socio-structural bias: II. Various models with partial transitivity. *Bulletin of Mathematical Biophysics 15*, 535–546.

Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Series B (Methodological) 59*, 731–792.

Robertson, T., F. T. Wright, and R. L. Dykstra (1988). *Order Restricted Statistical*

Inference. Chichester: Wiley.

Seidman, S. B. (1983). Internal cohesion of LS sets in graphs. *Social Networks 5*, 97–107.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J. 27*, 379–423, 623–656.

Simmel, G. (1968). *Conflict and the Web of Group Affiliations* (third ed.). New York: Free Press.

Snijders, T. A. B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure 3*.

Snijders, T. A. B. and K. Nowicki (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification 14*, 75–100.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics 22*, 1701–1728.

Tönnies, F. (1955). *Community and Association*. London: Routledge & Kegan Paul.

Wasserman, S. and K. Faust (1994). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Wasserman, S. and P. Pattison (1996). Logit models and logistic regression for social networks: I. An introduction to Markov graphs and p^*. *Psychometrika 61*, 401–425.

Watts, D. J., P. S. Dodds, and M. E. J. Newman (2002). Identity and search in social networks. *Science 296*, 1302–1305.

Wedel, M. and W. S. DeSarbo (1998). Mixtures of (constrained) ultrametric trees. *Psychometrika 63*, 419–443.

Winship, C. (1977). A distance model for sociometric structure. *Journal of Mathematical Sociology 5*, 21–39.

Yang, Z. and B. Rannala (1997). Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Molecular Biology and Evolution 14*, 717–724.