Locating Hidden Groups in Communication Networks Using Hidden Markov Models

Malik Magdon-Ismail[1], Mark Goldberg[1], William Wallace[2], and David Siebecker[1]

[1] CS Department, RPI, Rm 207 Lally, 110 8th Street, Troy, NY 12180, USA.
Email: {magdon,goldberg,siebed}@cs.rpi.edu.
[2] DSES Department, RPI, 110 8th Street, Troy, NY 12180, USA.
Email: wallaw@rpi.edu.

Abstract. A communication network is a collection of social groups that communicate via an underlying communication medium (for example newsgroups over the Internet). In such a network, a hidden group may try to camoflauge its communications amongst the typical communications of the network. We study the task of detecting such hidden groups given only the history of the communications for the entire communication network. We develop a probabilistic approach using a Hidden Markov model of the communication network. Our approach does not require the use of any semantic information regarding the communications. We present the general probabilistic model, and show the results of applying this framework to a simplified society. For 50 time steps of communication data, we can obtain greater than 90% accuracy in detecting both whether or not their is a hidden group, and who the hidden group members are.

1 Introduction

The tragic events of September 11, 2001 underline the need for a tool which is capable of detecting groups that hide their existence and functionality within a large and complicated communication network such as the Internet. In this paper, we present an approach to identifying such groups. Our approach does not require the use of any semantic information pertaining to the communications. This is preferable because communication within a hidden group is usually encrypted in some way, hence the semantic information will be misleading, or unavailable.

Social science literature has developed a number of theories regarding how social groups evolve and communicate, [1–3]. For example, individuals have a higher tendency to communicate if they are members of the same group, in accordance with homophily theory. Given some of the basic laws of how social groups evolve and communicate, one can construct a model of how the communications within the society *should* evolve, given the (assumed) group structure. If the group structure does not adequately explain the observed communications, but the addition of an extra, hidden, group does explain them, then we have grounds to believe that there is a hidden group attempting to camouflage its communications within the existing communication network. The task is to determine whether such a group exists, and identify its members. We use a maximum likelihood approach to solving this task.

Our approach is to model the evolution of a communication network using a Hidden Markov Model. A Hidden Markov model is appropriate when an observed process (in our case the macroscopic communication structure) is naturally driven by an unobserved, or hidden, Markov process (in our case the microscopic group evolution). Hidden Markov models have been used extensively in such diverse areas as: speech recognition, [4, 5]; inferring the language of simple grammars [6]; computer vision, [7]; time series analysis, [8]; biological sequence analysis and protein structure prediction, [9–13]. Our interpretation of the group evolution giving rise to the observed macroscopic communications evolution makes it natural to model the evolution of communication networks using a Hidden Markov model as well. Details about the general theory of Hidden Markov models can be found in [4, 14, 15].

In social network analysis there are many static models of, and static metrics for the measurement and evaluation of social networks [16]. These models range from graph structures to large simulations of agent behavior. The models have been used to discover a wide array of important communication and sociological phenomenon, from the small world principle [17] to communication theories such as homophily and contagion [1]. These models, as good as they are, are not sufficient to study the evolution of social groups and the communication networks that they use; most focus on the study of the evolution of the network itself. Few attempt to explain how the use of the network shapes its evolution [18]. Few can be used to predict the future of the network and communication behavior over that network. Though there is an abundance of simulation work in the field of computational analysis of social and organizational systems [2, 19, 3] that attempts to develop dynamic models for social networks, none have employed the proposed approach and few incorporate sound probability theory or statistics [20] as the underlying model.

The outline of the paper is as follows. First we consider a simplified example, followed by a description of the general framework. We also present some results to illustrate proof of concept on an example, and we end with some concluding remarks.

1.1 Example

A simple, concrete example will help to convey the details of our method. A more detailed formulation will follow. Consider the newsgroups, for example alt.revisionism, alt.movies. A posting to a newsgroup in reply to a previous posting is a communication between two parties. Now imagine the existence of a hidden group that attempts to hide its communications, illustrated in the figure below. Figure 1(a) shows the group structure. There are 4 observed groups. A fifth hidden group also exists, whose members are unshaded. We do not observe the actual group composition, but rather the communications (who is posting and

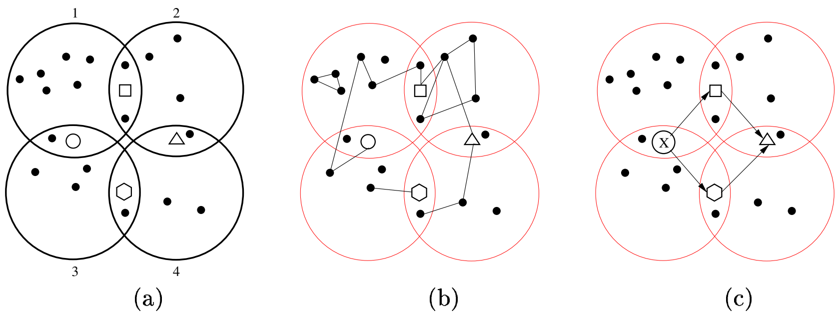

Fig. 1. Illustration of a society.

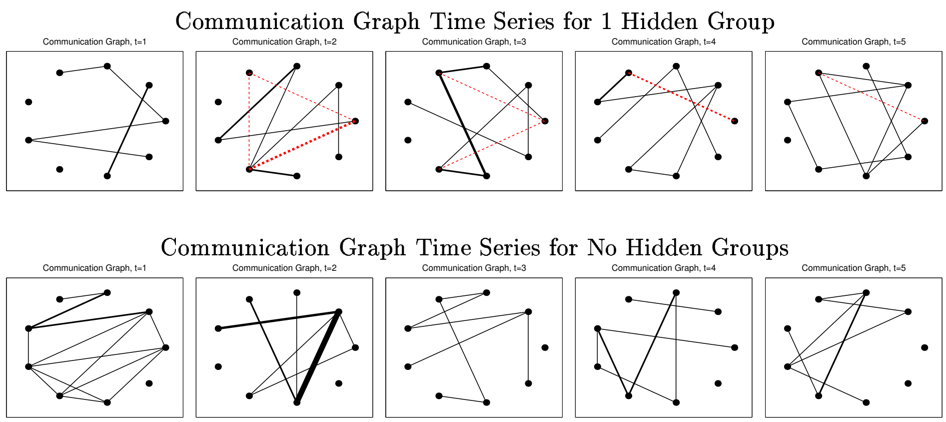

Fig. 2. Communication time series of two societies.

replying to posts in a given newsgroup). This is illustrated in Figure 1(b), where all the communications are between members of the same group. Figure 1(c) illustrates the situation when the hidden group members need to broadcast some information among themselves. The hidden group member who initiates the broadcast (say X) communicates with all the other hidden group members who are in the same visible groups as X. The message is then passed on in a similar manner until all the hidden members have received the broadcast. Notice that no communication needs to occur between members who are not in the same group, yet, a message can be broadcast across the whole group. In order to maintain the the appearance of being a bona-fide member of a particular newsgroup, a hidden node will participate in the "normal" communications of that group as well. Only occasionally will a message need to be broadcast through the hidden group, resulting in a communication graph as in Figure 1(c). The matter is complicated by the fact that the communications in Figure 1(c) will be overlayed onto the normal group communications, Figure 1(b). What we observe are a time series of node to node communications as illustrated in Figure 2, which shows the evolving communications of two hypothetical communities. The individuals are represented by nodes in the graph. An edge between two nodes represents communication during that time period. The thickness of the edge indicates the intensity of the communications. The dotted lines indicate communications between the hidden group members.

The task is to take the communication history of the community (for example the one above) and to determine whether or not there exists a hidden group functioning within this community, and to identify its members. It would also be useful to identify which members belong to which groups. The hidden community may or may not be functioning as an aberrant group trying to camouflage its communications. In the above example the hidden community trying to camouflage its broadcasts. However, the hidden group could just as well be a new group that has suddenly arisen, and we would like to discover its existence. We assume that we know the number of observed groups (for example the newsgroups societies are known), and we have a model of how the society evolves. We do not know who belongs to which news group, and all communications are aggregated into the communications graph for a given time period. We will develop a framework to determine the presence of a hidden group that *does not* rely on any semantic information regarding the communications. The motivation for this approach is that even if the semantics are available (which is not likely), the hidden communications will usually be encrypted and designed so as to mimic the regular communications anyway.

2 Probabilistic Setup

We will illustrate our general methodology by first developing the solution of the simplified example discussed above. The general case is similar, with only minor technical differences. The first step is to build a model for how individuals move from group to group. More specifically, let N_g be the number of observed groups in the society, and denote the groups by $F_1, \ldots, F_{N_g}$. Let n be the number of individuals in the society, and denote the individuals by $x_1, \ldots, x_n$. We denote by $\mathbf{F}(t)$, the *micro-state* of the society at time t. The micro-state represents the state of the society. In our case, $\mathbf{F}(t)$ is the membership matrix at time t, which is a binary $n \times N_g$ matrix that specifies who is in which group,

$$\mathbf{F}_{ij}(t) = \begin{cases} 1 & \text{if node } x_i \text{ is in group } F_j, \\ 0 & \text{otherwise.} \end{cases} \qquad (1)$$

The group membership may change with time. We assume that $\mathbf{F}(t)$ is a Markov chain, in other words, the members decide which groups to belong to at time $t+1$ based solely on the group structure at time t. In determining which groups to join in the next period, the individuals may have their own preferences, thus there is some transition probability distribution

$$P[\mathbf{F}(t+1)|\mathbf{F}(t), \boldsymbol{\theta}], \qquad (2)$$

where $\boldsymbol{\theta}$ is a set of (fixed) parameters that determine, for example, the individual preferences. This transition matrix represents what we define as the *micro-laws* of the society, that determines how its group structure evolves. A particular setting to the parameters $\boldsymbol{\theta}$ is a particular realization of the micro-laws. We will assume that the group membership is *static*, which is a trivial special case of a Markov chain where the transition matrix is the identity matrix. In the general case, this need not be so, and we pick this simplified case to illustrate the mechanics of determining the hidden group, without complicating it with the group dynamics. Thus, the group structure, $\mathbf{F}(t)$ is fixed, so we will drop the t dependence.

We do not observe the group structure, but rather the communications that are a result of this structure. We thus need a model for how the communications arise out of the groups. Let $\mathbf{C}(t)$ denote the communications graph at time t. $\mathbf{C}_{ij}(t)$ is the intensity of the communication between node x_i and node x_j at time t. $\mathbf{C}(t)$ is the "expression" of the micro-state $\mathbf{F}$. Thus, there is some probability distribution

$$P[\mathbf{C}(t)|\mathbf{F}(t), \boldsymbol{\lambda}], \tag{3}$$

where $\boldsymbol{\lambda}$ is a set of parameters governing how the group structure gets expressed in the communications. Since $\mathbf{F}(t)$ is a Markov chain, $\mathbf{C}(t)$ follows a Hidden Markov process governed by the two probability distributions $P[\mathbf{F}(t+1)|\mathbf{F}(t), \boldsymbol{\theta}]$ and $P[\mathbf{C}(t)|\mathbf{F}(t), \boldsymbol{\lambda}]$. In particular, we will assume that there is some parameter $0 < \lambda < 1$ that governs how nodes in the same group communicate. We assume that the communication intensity $\mathbf{C}_{ij}(t)$ has a Poisson distribution with parameter $K\lambda$, where K is the number of groups that both nodes are members of. If $K = 0$, we will set the Poisson parameter to $\lambda^2 \ll 1$. otherwise $K = \lambda$. Thus, nodes that are not in any groups will tend not to communicate. The Poisson distribution is often used to model such "arrival" processes. Thus,

$$P[\mathbf{C}_{ij} = k] = \begin{cases} \mathcal{P}(k; K\lambda) & x_i \text{ and } x_j \text{ are in } K > 0 \text{ groups together,} \\ \mathcal{P}(k; \lambda^2) & x_i \text{ and } x_j \text{ are in no groups together.} \end{cases} \tag{4}$$

Where $\mathcal{P}(k; \lambda)$ is the Poisson probability distribution function,

$$\mathcal{P}(k; \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}. \tag{5}$$

We will assume that the communications between different pairs of nodes are independent of each other, as are communications at different time steps. Suppose we have a broadcast hidden group in the society as well, as illustrated in Figure 1(c). We assume a particular model for the communications within the hidden group, namely that every pair of nodes that are in the same visible group communicate. The intensity of the communications, B is assumed to follow a Poisson distribution with parameter β, thus

$$P[B = k] = \mathcal{P}(k; \beta), \tag{6}$$

We have thus fully specified the model for the society, and how the communications will evolve. The task is to use this model to determine, from communication history (as in Figure 2), whether or not there exists a hidden group, and if so, who the hidden group members are.

2.1 The Maximum Likelihood Approach

For simplicity we will assume that the only unknown is $\mathbf{F}$, the group structure. Thus, $\mathbf{F}$ is static and unknown and λ and β are known. Let H be a binary indicator variable that is 1 if a hidden group is present, and 0 if not. Our approach is to determine how likely the observed communications would be if there is a hidden group, l_1 and compare this with how likely the observed communications would be if there was no hidden group, l_0. To do this, we use the model describing the communications evolution with a hidden group (resp. without a hidden group) to find what the best group structure $\mathbf{F}$ would be if this model were true, and compute the likelihood of communications given this group structure and the model. Thus, we have two optimization problems,

$$l_1 = \max_{\mathbf{F},\mathbf{v}} P[Data|\mathbf{F}, \mathbf{v}, \lambda, \beta, H = 1], \tag{7}$$

$$l_0 = \max_{\mathbf{F}} P[Data|\mathbf{F}, \lambda, H = 0], \tag{8}$$

where $Data$ represents the communication history of the society, namely $\{\mathbf{C}(t)\}_{t=1}^{T}$, and $\mathbf{v}$ is a binary indicator variable that indicates who the hidden and visible members of the society are. If $l_1 > l_0$, then the communications are more likely if there is a hidden group, and we declare that there is a hidden group. As a by product, of the optimization, we will obtain $\mathbf{F}$ and $\mathbf{v}$, hence we will identify not only who the hidden group members are, but also the remaining group structure for the society. In what follows, we will derive this likelihood function that needs to be optimized for our example society. What remains is to then solve the two optimization problems to obtain l_1, l_0.

The simpler case is when there is no hidden group, which we analyze first. Suppose that $\mathbf{F}$ is given. Let f_{ij} be the number of groups that nodes x_i and x_j are both members of,

$$f_{ij} = \sum_k \mathbf{F}_{ik}\mathbf{F}_{jk}. \tag{9}$$

Let λ_{ij} be the Poisson parameter for the intensity of the communication between nodes x_i and x_j,

$$\lambda_{ij} = \begin{cases} \lambda^2 & f_{ij} = 0, \\ \lambda f_{ij} & f_{ij} > 0. \end{cases} \tag{10}$$

Let $P(t)$ be the probability of obtaining the observed communications $\mathbf{C}(t)$ at time t. Since the communications between nodes are assumed independent, and each is distributed according to a Poisson process with parameter λ_{ij}, we have that

$$P(t) = P[\mathbf{C}(t)|\mathbf{F}, \lambda, H = 0] \tag{11}$$

$$= \prod_{i<j}^{n} \mathcal{P}(\mathbf{C}_{ij}(t); \lambda_{ij}) \tag{12}$$

Since the communications at different times are independent (given the group structure at that time), we have that

$$P[Data|\mathbf{F},\lambda,H=0]=\prod_{t=1}^{T}\prod_{i<j}^{n}\mathcal{P}(\mathbf{C}_{ij}(t);\lambda_{ij}). \tag{13}$$

Since l_0 is given by the maximum value of this function, we can equivalently maximize the logarithm. Further, the value of $\mathbf{F}$ that attains this maximum is the estimate of the group structure, assuming that there is no hidden group,

$$\log l_0 = \max_{\mathbf{F}} \sum_{t=1}^{T}\sum_{i<j}^{n} \log \mathcal{P}(\mathbf{C}_{ij}(t);\lambda_{ij}) \tag{14}$$

$$\mathbf{F}_0 = \operatorname*{argmax}_{\mathbf{F}} \sum_{t=1}^{T}\sum_{i<j}^{n} \log \mathcal{P}(\mathbf{C}_{ij}(t);\lambda_{ij}) \tag{15}$$

When a hidden group is present, we use an indicator variable $\mathbf{v}$ to identify who the hidden group members are. We assume that there are as many hidden members as there are groups for illustration, as in Figure 1. At time t, the observed communications can be broken down into the regular communications plus a hidden group broadcast. The intensity B of the broadcast can be $B=0,1,\ldots$, with B distributed according to a Poisson distribution, $P[B|\beta]=\mathcal{P}(B;\beta)$ This is our model of the hidden group communications. We can thus decompose the communication graph at time t into two components, a residual graph $\mathbf{R}(t)$, and the hidden graph $\mathbf{H}(t)$,

$$\mathbf{C}(t)=\mathbf{R}(t)+\mathbf{H}(t). \tag{16}$$

The hidden graph will be of the form in Figure 1(c) and is completely specified by the broadcast intensity B. This will then specify $\mathbf{R}(t)$ by $\mathbf{R}(t)=\mathbf{C}(t)-\mathbf{H}(t)$. Suppose that we know $\mathbf{F}$, $\mathbf{v}$. In this case, $P(t)$ is given by

$$P(t) = P[\mathbf{C}(t)|\mathbf{F},\mathbf{v},\lambda,\beta,H=1] \tag{17}$$

$$= \sum_{B=0}^{\infty} P[\mathbf{R}(t)|B]P[B] \tag{18}$$

Where $P[\mathbf{R}(t)|B]$ is given by an expression exactly analogous to (12),

$$P[\mathbf{R}(t)|B]=\prod_{i<j}^{n}\mathcal{P}(\mathbf{R}_{ij}(t;B);\lambda_{ij}) \tag{19}$$

where $\mathbf{R}(t;B)$ is the residual graph depending on B, and λ_{ij} is defined exactly analogously to (10) with $f_{ij}=\sum_k \mathbf{F}_{ik}\mathbf{F}_{jk}$. $\mathbf{v}$ places a constraint on what $\mathbf{F}$ can be, and serves to determine what the hidden group broadcast graph can be. Note that the sum in (18) gets truncated when B gets large enough so that the residual graph has negative edges, which is impossible, since it must be a communications graph. We will denote this maximum possible value of B by B^t_{max}. Then, using the fact that $P[B] = \mathcal{P}(B;\beta)$, we get that

$$P(t) = \sum_{B=0}^{B^t_{max}} \mathcal{P}(B;\beta) \prod_{i<j}^{n} \mathcal{P}(\mathbf{R}_{ij}(t;B);\lambda_{ij}) \tag{20}$$

Taking the logarithm and summing over t, we get that

$$\log l_1 = \max_{\mathbf{F},\mathbf{v}} \sum_{t=1}^{T} \log \sum_{B=0}^{B^t_{max}} \mathcal{P}(B;\beta) \prod_{i<j}^{n} \mathcal{P}(\mathbf{R}_{ij}(t;B);\lambda_{ij}) \tag{21}$$

$$\{\mathbf{F}_1, \mathbf{v}_1\} = \operatorname*{argmax}_{\mathbf{F},\mathbf{v}} \sum_{t=1}^{T} \log \sum_{B=0}^{B^t_{max}} \mathcal{P}(B;\beta) \prod_{i<j}^{n} \mathcal{P}(\mathbf{R}_{ij}(t;B);\lambda_{ij}) \tag{22}$$

Thus, in order to obtain $l_0, l_1, \mathbf{F}_0, \mathbf{F}_1, \mathbf{v}_1$, we need to solve two combinatorial optimization problems. Notice that the size of the search space is huge. When there is no hidden group, the size of the search space is 2^{nN_g}, and the evaluation of the objective function is $O(Tn^2)$. When there is a hidden group, the size of the search space is $2^{(n-N_g)N_g} n!/(n-N_g)!$ and the evaluation of the objective function is $O(C_{max}Tn^2)$, where C_{max} is the maximum communication intensity between any two nodes. If in addition the parameters of the model, namely λ, β are also not known, then we have to optimize with respect to these parameters as well, in which case, we have a mixed continuous/discrete optimization problem. Some algorithms for discrete/combinatorial optimization problems are reactive search, [21, 22], and randomized approaches, see for example [23]. Continuous problems are often approached using derivative based methods such as gradient descent, conjugate gradients, Levenberg-Marquardt, etc., [24]. Mixed discrete/continuous problems have not been studied as intensely, and most methods are based upon simulated annealing [25] or genetic algorithms, [26]. For illustration, we assume that the parameters are known, the purpose here is to set the framework for the problem. To illustrate, we have implemented a simulated annealing approach to the combinatorial optimization. We used 10,000 steps of Monte Carlo, where at each step, the current group structure $\mathbf{F}$ was randomly perturbed. The probability of perturbation decreased as a function of the step number.

Results We show results on a small society (9 nodes) with 3 groups. We picked this society so that it would be computationally efficient to run many simulations. We ran simulations to test both the false positive (declaring a hidden group when there isn't one) and false negative (declaring no hidden group when there is one) errors. For each, we generated a society group structure randomly, and then generated the communication time series. These communication time series were fed into the optimization algorithm to obtain $l_0, l_1, \mathbf{F}_0, \mathbf{F}_1, \mathbf{v}_1$. If $l_1 > l_0$ we declare a hidden group to be present and identify its members in $\mathbf{v}_1$ and the group structure in $\mathbf{F}_1$. If not, we declare no hidden group and identify the group structure in $\mathbf{F}_0$. The results are summarized in Table 1.

10 time steps

True H	Predicted H 1	0
1	0.73	0.27
0	0.19	0.81

% correct =84%

20 time steps

True H	Predicted H 1	0
1	0.78	0.28
0	0.04	0.96

% correct =89%

50 time steps

True H	Predicted H 1	0
1	0.88	0.12
0	0.03	0.97

% correct=94%

Table 1. Error matrices for different time periods. % correct is the percentage of nodes identified correctly (hidden or not) when a hidden group is present and is predicted correctly.

As can be seen, with just 50 time steps of data, the error rate in predicting the presence of a hidden group is lower than 0.1.

2.2 General Maximum Likelihood Formulation

In general, the group structure evolves according to the micro-law transition matrix for the Markov chain, $P[\mathbf{F}(t+1)|\mathbf{F}(t), \boldsymbol{\theta}]$, and, the group structure gets expressed as a communication graph according to $P[\mathbf{C}(t)|\mathbf{F}(t), \boldsymbol{\lambda}]$. In our example, $P[\mathbf{F}(t+1)|\mathbf{F}(t), \boldsymbol{\theta}]$ was the identity matrix, and $P[\mathbf{C}(t)|\mathbf{F}(t), \boldsymbol{\lambda}]$ based on modeling the communications using Poisson processes. A detailed description of a general model that describes an evolving society over a communication network is given in [27].

Let $\mathcal{N} = \{x_1, \ldots, x_n\}$ be the set of nodes and let $\mathcal{H} \subset \mathcal{N}$ be the subset of nodes that forms the hidden group. We assume that $\mathcal{H}$ does not change with time. The hidden group may have a communication pattern governed by a different probability distribution, $P[\mathbf{H}(t)|\mathcal{H}, \boldsymbol{\beta}]$, where $\boldsymbol{\beta}$ is a set of parameters that governs this distribution. The group structure of the society from $t = 1, \ldots, T$ is given by the time series of matrices $\{\mathbf{F}(t)\}_{t=1}^{T}$. In our example, this time series was specified by the constant matrix $\mathbf{F}$. If there is no hidden group, we can compute the likelihood of observing the communication data $\{\mathbf{C}(t)\}_{t=1}^{T}$ as follows. The probability of obtaining the evolution $\mathbf{F}(1), \mathbf{F}(2), \ldots, \mathbf{F}(T)$ is given by

$$P[\{\mathbf{F}(t)\}|\boldsymbol{\theta}] = P[\mathbf{F}(1)] \prod_{t=2}^{T} P[\mathbf{F}(t)|\mathbf{F}(t-1), \boldsymbol{\theta}]. \quad (23)$$

The likelihood of obtaining the observed communications given this evolution is then given by

$$P[\{\mathbf{C}(t)\}|\{\mathbf{F}(t)\}, \boldsymbol{\theta}, \boldsymbol{\lambda}] = \prod_{t=1}^{T} P[\mathbf{C}(t)|\mathbf{F}(t), \boldsymbol{\lambda}]. \quad (24)$$

Ideally, we would like to compute

$$l_0 = P[\{\mathbf{C}(t)\}|\boldsymbol{\theta}, \boldsymbol{\lambda}] = \sum_{\{\mathbf{F}(t)\}} P[\{\mathbf{C}(t)\}, \{\mathbf{F}(t)\}|\boldsymbol{\theta}, \boldsymbol{\lambda}] \tag{25}$$

$$= \sum_{\{\mathbf{F}(t)\}} P[\{\mathbf{F}(t)\}|\boldsymbol{\theta}]P[\{\mathbf{C}(t)\}|\{\mathbf{F}(t)\}, \boldsymbol{\theta}, \boldsymbol{\lambda}] \tag{26}$$

$$= \sum_{\{\mathbf{F}(t)\}} P[\mathbf{F}(1)]P[\mathbf{C}(1)|\mathbf{F}(1), \boldsymbol{\lambda}] \prod_{t=2}^{T} P[\mathbf{F}(t)|\mathbf{F}(t-1), \boldsymbol{\theta}]P[\mathbf{C}(t)|\mathbf{F}(t), \boldsymbol{\lambda}] \tag{27}$$

If $\boldsymbol{\theta}, \boldsymbol{\lambda}$ are known, then this summation can be computed using a Monte Carlo simulation. If not, then we find the values of $\boldsymbol{\theta}, \boldsymbol{\lambda}$ that maximize l_0. In this case, the optimization is computationally costly and an alternative is to simultaneously optimize with respect to $\{\mathbf{F}(t)\}_{t=1}^{T}, \boldsymbol{\theta}, \boldsymbol{\lambda}$, which is itself a non-trivial mixed discrete/continuous optimization problem.

When a hidden group $\mathcal{H}$ is present, we decompose the communications at time t to the hidden communications $\mathbf{H}(t)$ and the residual communications $\mathbf{R}(t)$, with $\mathbf{C}(t) = \mathbf{R}(t) + \mathbf{H}(t)$. Then,

$$P[\mathbf{C}(t)|\{\mathbf{F}(t)\}, \mathcal{H}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\beta}] = \sum_{\mathbf{H}(t)} P[\mathbf{R}(t)|\{\mathbf{F}(t)\}, \boldsymbol{\theta}, \boldsymbol{\lambda}]P[\mathbf{H}(t)|\mathcal{H}, \boldsymbol{\beta}], \tag{28}$$

where this summation is finite because both $\mathbf{R}(t)$ and $\mathbf{H}(t)$ must have non-negative edges. Taking the product over t gives us

$$P[\{\mathbf{C}(t)\}|\{\mathbf{F}(t)\}, \mathcal{H}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\beta}] = \prod_{t=1}^{T} \sum_{\mathbf{H}(t)} P[\mathbf{R}(t)|\{\mathbf{F}(t)\}, \boldsymbol{\theta}, \boldsymbol{\lambda}]P[\mathbf{H}(t)|\mathcal{H}, \boldsymbol{\beta}], \tag{29}$$

and finally multiplying by $P[\{\mathbf{F}(t)\}|\boldsymbol{\theta}]$ and summing over $\{\mathbf{F}(t)\}$, we get that

$$l_1 = \max_{\mathcal{H}} \sum_{\{\mathbf{F}(t)\}} P[\mathbf{F}(1)] \prod_{t=1}^{T} P[\mathbf{F}(t+1)|\mathbf{F}(t), \boldsymbol{\theta}]P[\{\mathbf{C}(t)\}|\{\mathbf{F}(t)\}, \mathcal{H}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\beta}], \tag{30}$$

where $P[\{\mathbf{C}(t)\}|\{\mathbf{F}(t)\}, \mathcal{H}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\beta}]$ is given in (29). The hidden group $\mathcal{H}$ at which the maximum is attained identifies who the hidden group members are. We assume that the Hidden Markov model and its parameters $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\beta})$ are known. If the parameters are not known, then they have to be optimized as well. For a relatively simple hidden group communication structure, for example the broadcast hidden group as in our example, the computation of the likelihood is tractable. For more complicated examples, one may need to use heuristic approaches to these combinatorial optimization problems.

3 Concluding Remarks

We have presented a framework for determining the members of a hidden group that attempts to camouflage its broadcasts within a functioning communication network. The basic idea is to first have a model for the society's evolutions. Then by examining the discrepancy between the observed and expected communications, one can draw conclusions regarding the presence or absence of a hidden group.

We focussed on a specific example, where we made a number of assumptions: static group structure; Poisson communication model; independence between communications at different times; the hidden group communications were only broadcasts; we used a maximum likelihood formulation. These restrictions were made primarily for expository and computational reasons, and are dropped in the general framework (resulting in more computationally intensive and complex optimization problems).

Ongoing research involves developing efficient heuristic algorithms that solve the combinatorial optimization problems faced in the more general framework, as well as applying our methodology toward finding hidden groups in real societies.

References

1. Monge, P., Contractor, N.: Theories of Communication Networks. Oxford University Press (2002)
2. Carley, K., Prietula, M., eds.: Computational Organization Theory. Lawrence Erlbaum associates, Hillsdale, NJ (2001)
3. Sanil, A., Banks, D., Carley, K.: Models for evolving fixed node networks: Model fitting and model testing. Journal oF Mathematical Sociology **21** (1996) 173–196
4. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77** (1989) 257–286
5. Rabiner, L.R., Juang, B.H.: An introduction to hidden Markov models. IEEE ASSP Magazine (1986) 4–15
6. Georgeff, M.P., Wallace, C.S.: A general selection criterion for inductive inference. European Conference on Artificial Intelligence (ECAI, ECAI84) (1984) 473–482
7. Bunke, H., Caelli, T., eds.: Hidden Markov Models. Series in Machine Perception and Artificial Intelligence - Vol. 45. World Scientific (2001)
8. Edgoose, T., Allison, L.: MML Markov classification of sequential data. Stats. and Comp. **9** (1999) 269–278
9. Allison, L., Wallace, C.S., Yee, C.N.: Finite-state models in the alignment of macro-molecules. J. Molec. Evol. **35** (1992) 77–89
10. Allison, L., Wallace, C.S., Yee, C.N.: Normalization of affine gap costs used in optimal sequence alignment. J. Theor. Biol. **161** (1993) 263–269
11. Bystroff, C., Thorsson, V., Baker, D.: HMMSTR: A hidden Markov model for local sequence-structure correlations in proteins. Journal of Molecular Biology **301** (2000) 173–90
12. Bystroff, C., Baker, D.: Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol **281** (1998) 565–77
13. Bystroff, C., Shao, Y.: Fully automated ab initio protein structure prediction using I-sites, HMMSTR and ROSETTA. Bioinformatics **18** (2002) S54–S61
14. Baldi, P., Brunak, S.: Bioinformatics: The Machine Learning Approach. MIT Press, Cambridge, MA (1998)
15. Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: Biological Sequence Analysis. Cambridge, new York (2001)

16. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press (1994)
17. Watts, D.J.: Small Worlds: The dynamics of networks between order and randomness. Princeton University Press, Princeton, NJ (1999)
18. Butler, B.: The dynamics of cyberspace: Examing and modelling online social structure. Technical report, Carnegie Melon University, Pittsburgh, PA (1999)
19. Carley, K., Wallace, A.: Computational organization theory: A new perspective. In Gass, S., Harris, C., eds.: Encyclopedia of Operations Research and Management Science. Kluwer Academic Publishers, Norwell, MA (2001)
20. Snijders, T.: The statistical evaluation of social network dynamics. In Sobel, M., Becker, M., eds.: Sociological Methodology dynamics. Basil Blackwell, Boston & London (2001) 361–395
21. Battiti, R.: Reactive search: Toward self-tuning heuristics. Modern Heuristic Search Methods, Chapter 4 (1996) 61–83
22. Battiti, R., Protasi, M.: Reactive local search for the maximum clique problem. Technical Report TR-95-052, Berkeley, ICSI, 1947 Center St.- Suite 600 (1995)
23. Motwani, R., Raghavan, P.: Randomized Algorithms. Cambridge University Press, Cambridge, UK (2000)
24. Bishop, C.M.: Neural Networks for Pattern Recognition. Clarendon Press, Oxford (1995)
25. Aarts, E., Korst, J.: Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing. John Wiley & Sons Ltd., New York (1989)
26. Stelmack, M., N., N., Batill, S.: Genetic algorithms for mixed discrete/continuous optimization in multidisciplinary design. In: AIAA Paper 98-4771, AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, St. Louis, Missouri (1998)
27. Siebeker, D., Goldberg, M., Magdon-Ismail, M., Wallace, W.: A Hidden Markov Model for describing the statistical evolution of social groups over communication networks. Technical report, Rensselaer Polytechnic Institute (2003) Forthcoming.