

Statistical Modeling of Social Groups on Communication Networks

Mark Goldberg
goldberg@cs.rpi.edu

Paul Horn
hornp@cs.rpi.edu

Malik Magdon-Ismail
magdon@cs.rpi.edu

William Wallace
wallaw@rpi.edu

Jessie Riposo
ripos@rpi.edu

David Siebecker
siebed@cs.rpi.edu

Bulent Yener
yener@cs.rpi.edu

February 28, 2003

Abstract

A communication network is a collection of social groups that communicate via an underlying communication medium (for example newsgroups over the Internet). Social groups evolve and as a result, the communication graph of the network evolves. We develop a probabilistic approach to modeling the evolution of social groups and communication networks that uses Hidden Markov models. We then develop a methodology for extracting the “laws” governing how a society behaves by *reverse engineering* these parameters from the data. We present preliminary results on some newsgroup societies.

1 Introduction

A social group is a collection of individual social units, or *actors* ([49]) that have some property in common. Groups within a social network may overlap; these groups may have a number of sub-groups, which may also overlap. Underlying a social network is its communication network. The Internet and its many separate domains are examples of such gigantic social networks containing numerous social groups and sub-groups of people, or organizations, united by common interests and activities. These groups evolve, expand, decline, or stabilize, to be eventually transformed into other groups of Internet users. We model the evolution of social groups in social networks in order to explain their history and predict their future.

We present new methodologies for the analysis of social groups that change over time. A combination of Social Network theories, Communication Theories, Graph Theory, Statistical Learning and Algorithms, are used to develop a comprehensive set of models, encompassed within a unified framework, to study the evolution of social groups. This framework provides tools that could be used for the identification, detection, and explanation of the behavior of social groups over time. The most paramount contribution, provided within this framework, is the introduction of machine learning methods for discovering what theories or combination of theories of individual behavior would lead to the observed evolution of a particular society/communication network. Traditionally, when studying group formation and evolution, some hypotheses are made about the laws that govern individual behavior. One then collects data and performs a hypothesis test to determine if the proposed laws are correct. Our approach is to start with the data, and then *learn* the individual behavior or “laws”. The learned behavior can then function as an established theory and be used to build models for predictive purposes. The proposed approach can be less time consuming and

has the potential to be more accurate as the number of governing laws increases in number and complexity. In addition, our models allow for the laws to change over time.

An understanding of the inner mechanisms that determine the functioning of social groups is instrumental in predicting and shaping their future development. Some of the applications include: *prediction for resource allocation*, for example, predicting how Internet groups are growing will allow for more efficient allocation of bandwidth and other infrastructures, such as cache; *identifying hidden or emerging communities* – 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 social network such as the Internet – our methods can provide for such a tool [32]; *non-intrusive control* – for example altering the micro-laws so as to achieve a desired effect such as altering a cost of membership so as to promote “good” groups and discourage “bad” groups; *modeling contagion spreading* such as email viruses, and developing reaction and immunization strategies.

The basis for our modeling is a set of probabilistic laws **micro-laws**, that govern the individual behavior of actors. The **micro-laws** largely determine the **macro-evolution** of the social groups. Macro-quantities of interest include the number of social groups, the distribution of their sizes, the number of communications, and the dynamics of these values over a period of time. Conversely, the macro-evolution (for example communication intensities) can be measured, and we develop a methodology for using the measured behavior to determine the micro-laws that best fit the observed macro-evolution.

We present a unified, general framework for describing and prediction the evolution of social networks. Within our general model, the evolution of social networks is described by probabilistic laws; the functioning of different social networks results from different settings for specific parameters of these laws. Our preliminary results demonstrate that choosing simple laws can reproduce the behavior of some small societies. Further, the observed evolution of those societies can be used to determine what the appropriate laws governing that society are.

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, [37, 38]; inferring the language of simple grammars [25]; computer vision, [12]; time series analysis, [24]; biological sequence analysis and protein structure prediction, [2, 3, 16, 14, 15]. 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 of the general theory of Hidden Markov models can be found in [37, 4, 23].

Related Work

In social network analysis there are many static models of, and static metrics for the measurement and evaluation of social networks [49]. 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 [50] to communication theories such as homophily and contagion [35]. 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 [13]. 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 [17, 18, 39] that attempts to develop dynamic models for social networks, none have employed the proposed approach and few incorporate sound probability theory or statistics [43] as the underlying model.

The majority of existing work on the evolution of social networks deals with the organization [7], or other particular network structures, but very few models are built independent of network type. There is also work that has been done on developing statistical methods for modeling and evaluating social networks over time (longitudinal social network data) [43]. A basic continuous-time Markov chain model for dichotomous social networks was elaborated by Wasserman [46, 48, 47] and further investigated by Leenders [30, 31] and Snijders [42]. This model is limited by its assumptions and, although it is computationally attractive, it does not leave much room for realistic statistical modeling. Other continuous-time models for social network evolution were proposed by Wasserman [47] and Mayer [34], but these models were also very restrictive in order to allow parameter estimation. Tom Snijders [43] proposes a Markov chain Monte Carlo (MCMC) method that can be used to develop statistical procedures for general probability models of the evolution of social networks. Tom Snijders' work involves establishing methods for statistically measuring and evaluating the networks over two time steps, but he does not invite the study of the social groups functioning over the network. We build upon this work by taking the next step in Markov Chain evaluation. We use simulation, and Hidden Markov Chain Models to learn or establish the parameters governing individual behavior, without suffering the restrictions of the aforementioned models. For more on dynamic modeling see [43, 41, 28, 39].

Our present test bed is the Internet newsgroup. Some work on modeling voluntary collectives from listserv data was done by Brian Butler, [13]. Butler considered a model of voluntary collectives that included the role played by the communication technology. He studied the interrelationship between structural change of the network and individual change. He came to conclusions like "individuals in larger structures tend to be less committed and less satisfied and hence less likely to join or remain members." These types of conclusions are paramount in building and validating the models we use. He described the development process of voluntary social collectives by modeling individual change in the form of member attitude shifts, and structural change, in the form of membership movement into and out of the collective, as we do. The proposed work will enrich and expand his models in many ways, in particular, more sophisticated statistical methodologies will be employed for analysis of the interaction between the structural (macro) and individual (micro) parameters.

Our work relies heavily on the foundations and recommendations that have been laid in Monge and Contractor's most recent book [35]. In this book they point out the most widely accepted and established communication theories, and how they are used to model social networks. They also point out "relatively few network studies utilize theories as the basis for formulating research hypotheses, those that do use only single theories." We build into our models the ability to represent a wide range theories through the choice of various model parameters, including principles such as social capital and role dynamics to model the individual's behavior.

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 the example, and we end with some concluding remarks.

2 Modeling and Simulation

In a society whose members (*actors*) communicate with each other, *communities, or social groups* emerge, change with time, and disappear, to be replaced by some other communities. The evolution of such groups is largely determined by the individual decisions of the actors, that are determined by the individual characteristics of these actors. While the evolution on a large scale is observable, the individual decision making processes is not easily observed. Time varying observable quantities might be: the average size and the average number of social communities; the distribution of their sizes; the stability of the communities (both with respect to their sizes and the identity of their members). Such observable quantities will be called **macro variables**, and statistical dependencies that the macro variables satisfy will be called **macro-laws**. The properties attributable to the individual members that ultimately give rise to the macro state will be called **micro variables** and the laws governing the evolution of micro variables will be the **micro-laws**.

The actors act autonomously. An example of an actor's action would be to leave a certain group and join another, or to choose to stay in the old group. An actor's *nature* determines his decisions. In our model, the preferences that the actors have include, among many, the tendency to belong to a large group, or a tendency to actively pursue a dominant position in a group. An actor's nature thus determines how that actor will act given its current *state*. For example, if a group size became small, as the result of group members leaving, an actor may choose to leave the group and join a larger one. Since the distribution and the sizes of the groups in a society evolve according to the actions of its actors, this evolution is ultimately determined by the individual natures of all its members. A model of such an evolution must include diverse *types* of actors, and associated to each is a set of numeric parameters that governing his probabilistic behavior. If these parameters are specified, one can then determine, through simulation, how this society would evolve. Our goal is that one can then test an hypothesis about the behavior of the actors in a particular social network by comparing the simulated evolution of the society with the actual evolution of that society. Traditionally this is the mode in which such research is performed. To illustrate, we introduce a specific example.

Example: Newsgroup Societies

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. An example newsgroup society evolving is illustrated in Figure 1. The upper part of Figure 1 shows the society's group structure. We do not observe the actual group composition, but rather the communications (who is posting and replying to posts in a given newsgroup). This is illustrated in the lower part of Figure 1, As the communications evolve with time, we observe a time series of node to node communications as illustrated in Figure 2, which shows the evolving communications of a hypothetical community. The individuals are represented by nodes in the graph. An edge between two nodes represents communication during that time period. The thickness (or weight) of the edge indicates the intensity of the communications. The entire communications of the society at a given time period are thus represented by the *weighted communication graph*.

Suppose, for illustration, that a newsgroup society is composed of actors all of who "like to be in large groups". The first step is to translate this heuristic statement of a "law" governing actor

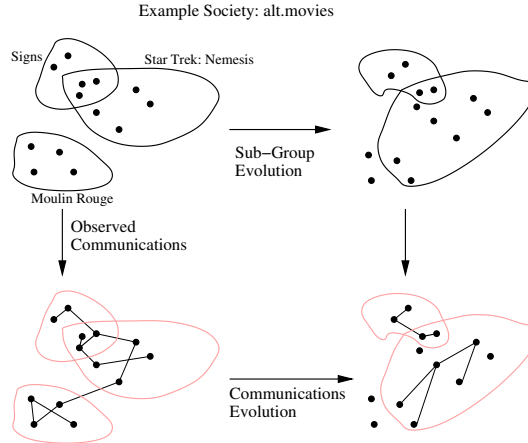


Figure 1: Evolving communications of a newsgroup society.

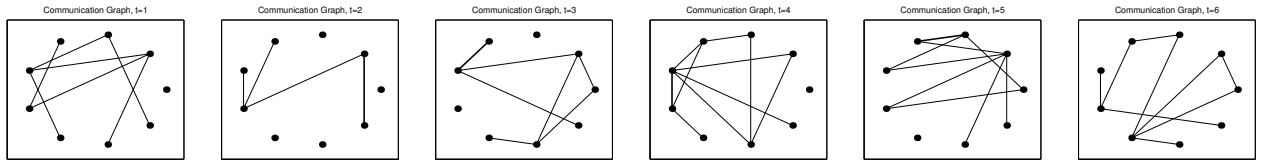


Figure 2: Communication time series of a hypothetical society.

behavior into a quantitative realization of that law. Without going into too many details, our model allows for a parameter that specifically governs the propensity of an actor for certain sized groups. So, for each actor, we set this particular parameter to have a high preference only for large groups. We set all the other parameters in the model to some reasonable values for illustrative purposes too. These other parameters can also be linked to heuristic laws, but we do not delve into the details here. Given an initial group structure, we can, evolve or *simulate* the groups according to this “law” that actors prefer large groups. Thus, we will get a time series for the group structure. As an initial group structure, we will take the connected components of the communication graph.

What we observe in the newsgroup society is the time series of communications, **not** the time series of groups. We thus need a model to connect the group structure at a given time step to its communication graph. For illustration we use a simple Poisson model. Let $\mathcal{P}(k; \lambda)$, where λ is the Poisson parameter, denote the Poisson probability distribution given by

$$\mathcal{P}(k; \lambda) = \frac{e^{-\lambda} \lambda^k}{k!}. \quad (1)$$

The Poisson distribution is often used to model “arrival” processes such as communications. Fix $\lambda > 0$. We will assume that the intensity of communication between two nodes has a Poisson distribution with Poisson parameter $K\lambda$ where K is the number of groups that both nodes are in. If $K = 0$ then we assume the Poisson parameter is λ/n where n is the total number of nodes. We

assume that each pair of nodes communicate in a statistically independent manner. Thus, Thus,

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

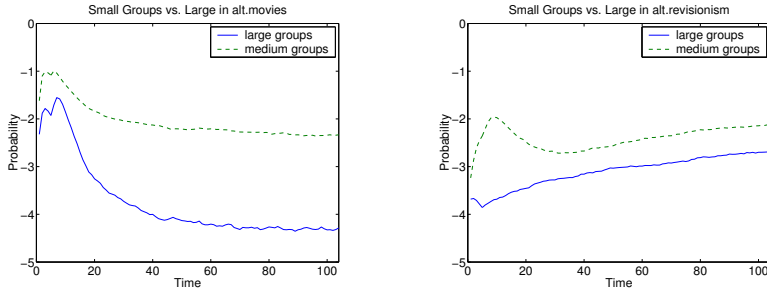
where x_i and x_j are two nodes, \mathbf{C}_{ij} is the communication intensity between these two nodes, and n is the total number of nodes in the society. Thus nodes in the same group are much more likely to communicate than nodes in different groups. Such a model derives from *homophily theory* [35] which asserts that nodes tend to communicate if they have some commonality (and hence are in the same group). Other social science theories could also be imposed on this model, such as the theory of *balance* [35] which asserts that one tries to keep one's communications balanced among the people one interacts with. For the moment, we use only this simple model emphasizing homophily. We will assume, once again for illustration, that the communication graphs at two different time periods are independent conditional upon knowing the group structures at those time periods. Having generated the group structure using the simulation (according to the hypothesized law), we can now generate the communications according to the Poisson model.

We now need a *metric* for comparing the communications obtained from the model with the true, observed, communications. One natural metric is the probability of obtaining the observed communications given the predicted group structure. Another metric that was suggested in [39] is the Hamming metric which measures the number of disagreements in the predicted communication structure and the observed one. Using this metric, we can compute the *distance* between the observed and simulated communications. This distance is a random variable depending on the particular realization of the group structure and communications that occurred in our simulation. Repeating this process many times, we can get an estimate of the expected distance between the true and simulated communications. If the metric is chosen so as to accurately represent the properties of the communications that are important, then if this expected distance is small, we can argue that our law adequately describes this society. We will use the probability of obtaining the communications given the model of the society as our metric. Let TC be the true communications and GS be the group structure. Then,

$$P[TC|\text{model}] = \sum_{GS} P[TC|GS]P[GS|\text{model}]. \quad (3)$$

The first term in the summation can be computed using the Poisson process assumption for how the group structure generates a communication graph. The second term in the summation may not be easy to compute for complicated models. However this summation is an expectation, and can be evaluated using Monte Carlo methods by sampling from $P[\text{group structure}|\text{model}]$.

Results. We apply this methodology to a pair of newsgroup societies – alt.movies and alt.revisionism. One is relatively peaceful, and one is relatively activist. These groups should behave in completely different ways. We test the hypotheses that actors like to join large groups against the hypothesis that actors like to join small groups. As a metric, we use the probability of observing the data, given that the society follows the hypothesized law. The results are summarized in the following figures.



As was expected, these two societies behave differently, but in both cases, it appears that the actors prefer to be in small groups for these two societies. This result is in accordance with the findings in [13], however, we have arrived at them using a general model, in an automated fashion.

3 Probabilistic Model

We present here the general framework for the model. The exact specification of the model contains many technical details and can be found in an accompanying technical report, [40] At time t , actors make decisions based on some information set or *micro-state* \mathcal{I}_t which is available to all the nodes at time t . In the newsgroup example above, the micro state at time t would be who is in a particular sub-group of the newsgroup at time t – in the alt.movies newsgroup, a particular subgroup might be all the people discussing a particular movie; such sub-groups can overlap. In principle, this information set could consist of the entire history of the society up to time t : what the groups were and who was in each group at every time period in the past. However, we will mostly assume that the only information needed for an actor to make decisions is the current state (group structure) of the society, and we also assume that all actors have access to this information or some approximation of it. These assumptions are reasonable and make the model more tractable.

When all the actors have taken their actions, the information set \mathcal{I}_{t+1} at time $t + 1$ is updated to reflect these new actions. Usually the update is in the form of some actors leaving certain groups and joining others. Since \mathcal{I}_{t+1} is determined by \mathcal{I}_t , we are assuming that the evolution is Markovian. Further, the actors act “stochastically”, i.e., given the same information on two separate occasions, an actor may act differently each time. Thus, the appropriate way to model the evolution is using the probabilistic setting of a **Markov process** - the actors are each following some probabilistic decision making process which causes the society to transition from state to state.

We assume that some parameters, which are *apriori* unknown but fixed, determine the exact way in which these probabilistic decisions are made. Collecting all these parameters in the vector θ , we then have that \mathcal{I}_{t+1} has a distribution dependent on \mathcal{I}_t and θ , given by

$$P[\mathcal{I}_{t+1}|\mathcal{I}_t, \theta] = Q(\mathcal{I}_{t+1}, \mathcal{I}_t, \theta). \quad (4)$$

Here Q is a function that takes as input the current micro-state \mathcal{I}_t , the parameters θ and the future target state \mathcal{I}_{t+1} and outputs the probability of obtaining that future state given the parameters and the current state. Specifying Q amounts to specifying the *model* for how the society is evolving, which are the micro-laws for the society. Specifying θ then amounts to specifying a particular realization of that model, which may or may not be appropriate to a given society. For the newsgroup example above, one of the parameters governing the model was the parameter that specified the propensity of an actor for a particular sized group – this parameter would be a part

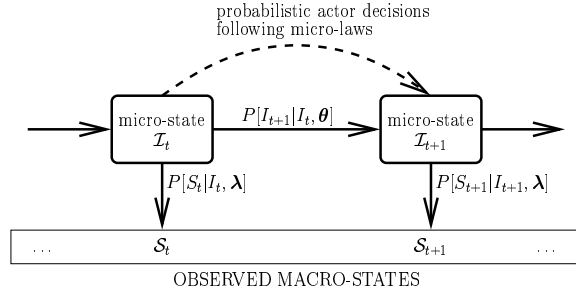


Figure 3: Pictorial representation of an evolving society.

of θ . \mathcal{Q} in the newsgroup example specifies exactly how the actors act given their propensities for different sized groups. *Different* societies may be described by different realizations of the *same* model. Our goal is to choose a general model that is capable of describing a wide variety of societies. Different societies will then be given by different realizations. In choosing the model, we use as a guideline many of the established theories from the social science literature.

The micro-state \mathcal{I}_t is not observed. The micro-state gives rise to a *macro-state* \mathcal{S}_t . In general this macro-state can depend probabilistically on the micro-state, and therefore is also specified by a probability distribution,

$$P[\mathcal{S}_t|\mathcal{I}_t, \lambda] = \mathcal{G}(\mathcal{S}_t, \mathcal{I}_t, \lambda), \quad (5)$$

Here, \mathcal{G} is a function that takes as input the current microstate \mathcal{I}_t , a set of society dependent parameters λ , and an observed macro-state \mathcal{S}_t and outputs the probability of observing that macro-state given the parameters and the current micro-state. λ are society dependent parameters that govern exactly how the macro-state results from the micro state – different societies may have different ways of communicating and this difference can be accommodated in different choices for λ . In the newsgroup example the micro-state is the group structure and the observed macro-state is the set of observed communications, which are governed by the Poisson parameter λ , thus λ was simply λ, β . Since \mathcal{I}_t follows a Markov process and \mathcal{S}_t is derived from \mathcal{I}_t , it is clear that \mathcal{S}_t follows a **hidden Markov process**. A hidden Markov model is appropriate when an observed process (in our case the macro-state communication structure) is naturally driven by an unobserved or hidden Markov process (in our case the micro-state group evolution). Hidden Markov models have been used extensively in speech recognition, [37, 38], but have found application in many other diverse areas such as: inferring the language of simple grammars [25]; computer vision, [12]; time series analysis, [24]; biological sequence analysis and protein structure prediction, [2, 3, 16, 14, 15]. Our interpretation of the group evolution giving rise to the observed macro-state evolution makes it natural to model the evolution of social networks using a hidden Markov process as well. Details about the general theory of hidden Markov models can be found in [37, 4, 23].

The pictorial representation of an evolving society is given in Figure 3. Our preliminary work is based on a particular interpretation of the general model described above. The society is made up of a set of groups $\{\mathcal{F}_i\}$. Each actor belongs to some subset of these groups at any given time t . Each actor has a “nature” – a set of parameters governing how it will behave – and a status which includes, among other things, a rank within each group, and an energy or budget which constrains the number of groups an actor can belong to. The micro-state \mathcal{I}_t is composed of the groups $\{\mathcal{F}_i\}$, including which actors are members of each group as well as the nature and the status of each actor.

The properties of the macro-state that we are interested in are the collection of groups, their sizes, their stabilities, and the resulting communication structure of the whole society. The actual groups in addition to the communications are observable for some societies, whereas for some others, such as communication networks, it is only possible to observe the pairwise communications (flakes) that are occurring.

The types of actions that an actor can take are to create new groups, leave some groups and/or join some other groups. The micro-laws that specify our model determine (probabilistically) the actions each actor makes. Since we do not claim to know how individuals in a society behave, we specify the micro-laws as parameterized functional forms which can be selected to exhibit a wide range of behaviors. These functional forms are selected because they appeared intuitive and satisfy certain intuitive properties, such as “*if one is a member of a community today, it is more likely that one is a member of that community tomorrow,*” in accordance with established theories in social science ([21, 49, 35]). The unspecified parameters θ, λ govern the detailed dynamics. A different choice for these parameters leads to different macro behavior. In fact, we have observed phase transitions where the qualitative behavior abruptly changes when a single parameter is slightly changed.

We can observe the behavior of a society and try to determine which values of these parameters are consistent with our observations. Having determined these parameters, we have a model of the society, which can give insight into its current state, and perhaps even lead to predictions on the future evolution of the society.

Model validation. It is not possible to observe the micro-laws directly, however, they are chosen so as to mimic established theories of social science. We would like to determine the parameters in the micro-laws by observing the society’s macro-state evolving. Having determined these parameters, if we can adequately predict the future macroscopic evolution of the society, we then have indirect confirmation of our micro-laws and parameters. This is the process that was illustrated in the newsgroup example. We tested the preference of actors for different sized groups.

4 Reverse Engineering

To illustrate one of the key contributions, we come back to the newsgroup example. The customary approach is to hypothesize on the “laws” of a society and then validate the hypothesis against the data, as was done in the example earlier using the small group/large group preferences. Why not let the data itself determine the “laws”, instead of hypothesizing them? To be more concrete, suppose we *did not* know what sized groups our society members preferred. We now describe methods to *reverse engineer* this information from the observed newsgroup postings. In other words, we propose to determine for a given society what the appropriate laws governing that society are, *from the data alone*. The size preferences for the members is only one parameter of the society, which is the parameter we picked for illustration. In principle, however, we could reverse engineer all the parameters, given enough data.

Formally, the goal is to determine what the appropriate parameters θ, λ are for a society, given the data describing how the macro-state evolves (in the news group data, this is the communication data). This falls under the general topic of parameter estimation for a statistical model given the data, and comes under the general paradigm of learning [33, 10, 26, 27, 45]. Generally, a data set is given, and the parameters of some model (for example neural networks, support vector machines,

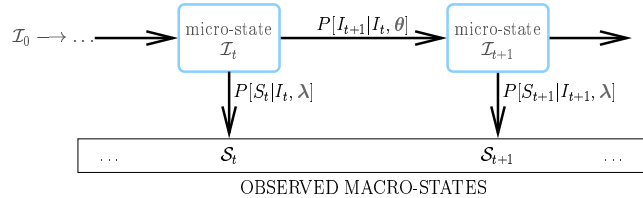


Figure 4: The reverse engineering task.

etc.) are to be tuned to *fit* the data. We describe an algorithm for performing this fit for the specific model we are dealing with. We postpone the discussion of other important issues such as avoiding overfitting and model selection to another presentation.

We wish to *learn* the parameters of a Hidden Markov model. A modification of the EM algorithm [22] leads to the well known Baum-Welch algorithm for parameter estimation [37], which is appropriate when the micro-state paths are not known (as in our case). The essential idea behind such learning algorithms is to maximize the likelihood of observing the data given the parameters,

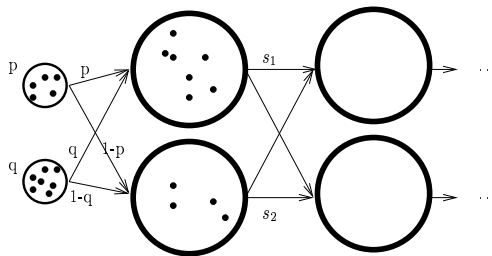
$$P[Data|\boldsymbol{\theta}, \boldsymbol{\lambda}].$$

While the parameters we wish to reverse engineer are $\boldsymbol{\theta}, \boldsymbol{\lambda}$, a number of other unknowns also exist, namely the actual path of micro-states, or actual group structure through time of the society, (remember that we only know, by observation, the macro-states, which in our example are the communications). This learning problem is complicated by the fact the some of the unknowns are discrete (some attributes of an actors nature, and the micro-state) and some of the parameters are continuous, thus making it a mixed discrete/continuous optimization problem. Some algorithms for discrete/combinatorial optimization problems are reactive search, [5, 6], and randomized approaches, see for example [36]. Continuous problems are often approached using derivative based methods such as gradient descent, conjugate gradients, Levenberg-Marquardt, etc., [10]. Mixed discrete/continuous problems have not been studied as intensely, and most methods are based upon simulated annealing [1] or genetic algorithms, [44]. For our preliminary results, we implemented a hybrid between simulated annealing (for the discrete parts) and a gradient based method (for the continuous parts), coupled with an EM style algorithm.

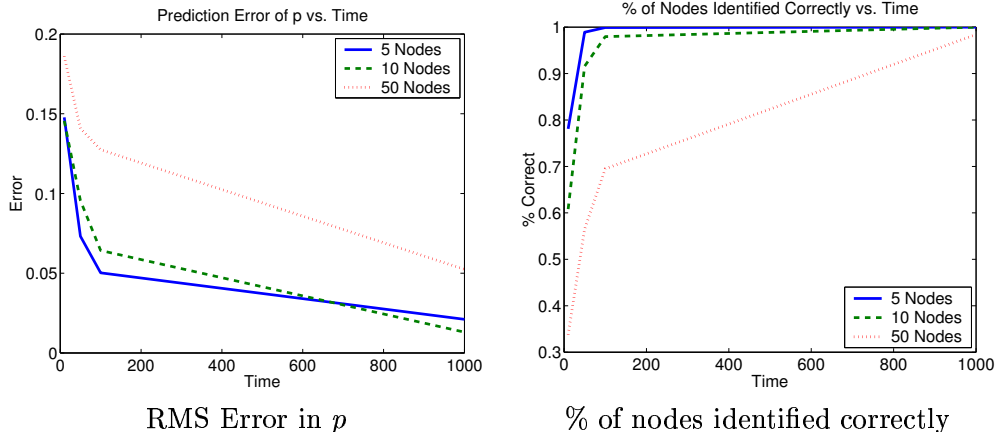
The task of reverse engineering is illustrated in Figure 4, where the dark quantities are known or observed, and the light ones are unknown and to be determined. Suppose that we have observations up to time period T . We would like to make predictions from then on, so we would also like to know the micro-state \mathcal{I}_T at time T , and $\boldsymbol{\theta}$. We thus need to estimate the hidden state space and the parameters, for example using some variant of the Baum-Welch algorithm [37] which maximizes the likelihood of the observed data. Alternatively, at much greater computational cost, and by assuming some prior distribution for the initial state, we can sample from the posterior distribution of the state space and parameters in a Bayesian formalism [11, 19]. Initially we focus on the maximum likelihood approach. The idea is that given $\{I_t\}, \boldsymbol{\theta}$, we can construct the likelihood $P[\{S_t\}|\{I_t\}, \boldsymbol{\theta}]$. The task is to now pick the assignment to $\{I_t\}, \boldsymbol{\theta}$ that maximizes this probability. Technically, since we only want \mathcal{I}_T , we can integrate over all other states (using Monte Carlo) to obtain maximum likelihood estimates of \mathcal{I}_T and $\boldsymbol{\theta}$. As already mentioned, we have a mixed discrete/continuous optimization problem, thus, we need to use techniques appropriate to such problems. Treating the

states as hidden variables, we approach the optimization problem using an EM style algorithm. In order to speed up the optimization and alleviate the local minimum problem, we begin by obtaining a good initial condition for the group structure, obtained by an algorithmic clustering of the society into initial groups. Our approach to getting an initial assignment to the group structure is based on the plausible assumption that the intensity of communications inside a group tends to exceed that of the average intensity. Thus, a collection of subsets that exhibit higher-than-average communication level will serve as an approximation for the true group structure. It turns out that existing algorithms for graph clustering [20] do not fit our goal. Our methodology is to repeatedly use graph-partitioning algorithms that minimize the communication level between the partitions, leaving within-group communication level higher than average. If an approximate size of the group is known, or conjectured, then the number of iterations of the partitioning can be determined as well. For our preliminary experiments, we used the classical Kernighan-Lin procedure ([29]), although we plan to use more involved algorithms ([8, 9]).

To illustrate the mechanics of the optimization process, we describe a simple society and the reverse engineering of *its* parameters. The more realistic society while technically more complicated (to account for more realistic behavior) is conceptually very similar. For this simpler society, suppose there are two types of actors, p 's and q 's. p and q denote probabilities associated with that particular actor. These actors can belong to one of two groups, F_1, F_2 . Actors move between these two groups according to functions of the probabilities, s_1, s_2 . s_1 (resp. s_2) is the average of the actor probabilities in F_1 (resp. F_2). Initially the actors choose F_1 with probability p (resp. q), and group F_2 with probability $1 - p$ (resp. $1 - q$). We have thus specified the micro-laws for *this* society, illustrated below.



The parameters that we would like to determine are p, q who is a p and who is a q . The observed macro state at any time t is the group assignment for every member. Thus, by observing how the actors move among the groups, we would like to determine the parameters. The observed data are the composition of the groups over time. We can compute $P[Data|p, q, \theta]$ where θ is an indicator variable which specifies which actors are of type p . As a preliminary learning algorithm, we used is a hybrid between a simulated annealing type algorithm for the discrete parameter θ , and a gradient descent for the continuous parameters, p, q . The figures below show the accuracy of determining the parameters as we vary the number of actors in the society and the number of time steps for which observations are available.



We can observe that the determination of the the parameters gets more accurate, the more data we have. Further, we can identify the natures of the individual actors by observing the collective behavior. The prediction errors display the expected $1/t$ behavior – as t increases, the number of samples available for estimation of the parameters increases, and for general learning systems, the error decreases a $1/t$ [33]. In general the more data one has, the more accurate the learning. Further, the more complicated the model (in this small society, the model gets more complicated when there are more actors), the harder it is to accurately determine the model parameters from the observed data. Thus given a fixed data set of observations, it is important to pick a model with an appropriate complexity. More details regarding such tradeoffs can be found in [10, 45].

We briefly look at the computational complexity of the reverse engineering process. The essential part of the optimization is the computation of $P[Data|p, q, \theta]$. If N is the number of nodes in the society, and T is the number of time steps for which data is available and F is the number of groups in the society, then obtaining $P[Data|p, q, \theta]$ is an $O((N^2 + NF)T)$ computation requiring $O(N^2 + NF)$ memory. Optimization for K iterations would introduce another factor of K .

4.1 Reverse Engineering the Newsgroups

We implemented our methodology to reverse engineer the parameters of two newsgroup societies. We then tested these reverse engineered parameters by predicting the future communications and comparing with the small and large group hypotheses. We implemented a prototype of the data collection module for collecting news messages from a news server which we used to collect data from alt.movies and alt.revisionism. We identified each node and constructed the communication graph by adding an edge between two nodes if one node replied to the post of another.

Using this data, we implemented an initial clustering to get initial groups, which we fixed. We then optimized with respect to θ to maximize $P[Data|\theta]$. We alternated between the discrete and continuous parameters in q , performing gradient ascent (hill climbing) on the continuous parameters and simulated annealing on the discrete parameters.

In order to compare the three different laws (large groups, small groups, learned preferences), we learned on half the data and then used all three models to predict future communications on the other half of the data. Using the metric of $P[\text{true communications}|\text{model}]$, we can then compare the three models. The results are illustrated in the figures below.

Acknowledgements

We would like to thank ***Bulents Students Names*** for collecting the data on which our simulations are based.

References

- [1] E. Aarts and J. Korst. *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing*. John Wiley & Sons Ltd., New York, 1989.
- [2] L. Allison, C. S. Wallace, and C. N. Yee. Finite-state models in the alignment of macromolecules. *J. Molec. Evol.*, 35(1):77–89, 1992.
- [3] L. Allison, C. S. Wallace, and C. N. Yee. Normalization of affine gap costs used in optimal sequence alignment. *J. Theor. Biol.*, 161:263–269, 1993.
- [4] P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, MA, 1998.
- [5] R. Battiti. Reactive search: Toward self-tuning heuristics. *Modern Heuristic Search Methods, Chapter 4*, pages 61–83, 1996.
- [6] R. Battiti and M. Protasi. Reactive local search for the maximum clique problem. Technical Report TR-95-052, Berkeley, ICSI, 1947 Center St.- Suite 600, 1995.
- [7] J. A. Baum and J. Singh, editors. *Evolutionary Dynamics of organizations*. Oxford Press, New York, 1994.
- [8] J. Berry and M. Goldberg. Path optimization and near-greedy analysis for graph partitioning: An empirical study. *Proceedings of Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1995.
- [9] J. Berry and M. Goldberg. Path optimization for graph partitioning problems. *Discrete Applied Mathematics (special issue on approximation algorithms)*, 90:27–50, 1999.
- [10] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.
- [11] BUGS: Bayesian inference using Gibbs sampling, 2003.
- [12] H. Bunke and T. Caelli, editors. *Hidden Markov Models*. Series in Machine Perception and Artificial Intelligence - Vol. 45. World Scientific, 2001.
- [13] B. Butler. The dynamics of cyberspace: Examining and modelling online social structure. Technical report, Carnegie Mellon University, Pittsburgh, PA, 1999.
- [14] C. Bystroff and D. Baker. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol*, 281(3):565–77, 1998.

- [15] C. Bystroff and Y. Shao. Fully automated ab initio protein structure prediction using I-sites, HMMSTR and ROSETTA. *Bioinformatics*, 18(1):S54–S61, 2002.
- [16] C. Bystroff, V. Thorsson, and D. Baker. HMMSTR: A hidden Markov model for local sequence-structure correlations in proteins. *Journal of Molecular Biology*, 301(1):173–90, 2000.
- [17] K. Carley and M. Prietula, editors. *Computational Organization Theory*. Lawrence Erlbaum associates, Hillsdale, NJ, 2001.
- [18] K. Carley and A. Wallace. Computational organization theory: A new perspective. In S. Gass and C. Harris, editors, *Encyclopedia of Operations Research and Management Science*. Kluwer Academic Publishers, Norwell, MA, 2001.
- [19] G. Casella and E. George. Explaining the Gibbs sampler. *Am. Stat.*, 46:167–174, 1992.
- [20] M. Charikar. Greedy approximation algorithms for finding dense components in a graph. In *APROX 2000*, pages 84–95, 2000.
- [21] J. Coleman. *Foundations of Social Theory*. The Belknap Press of Harvard University Press, Cambridge, MA, 1990.
- [22] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the E-M algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.
- [23] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge, new York, 2001.
- [24] T. Edgoose and L. Allison. MML Markov classification of sequential data. *Stats. and Comp.*, 9(4):269–278, September 1999.
- [25] M. P. Georgeff and C. S. Wallace. A general selection criterion for inductive inference. *European Conference on Artificial Intelligence (ECAI, ECAI84)*, pages 473–482, September 1984.
- [26] S. Haykin. *Neural Networks: A Comprehensive Foundation, 2nd Edition*. Prentice Hall, New Jersey, 1999.
- [27] G. Hinton and T. Sejnowski, editors. *Unsupervised Learning*. Foundations of Neural Computation. MIT Press, Cambridge, MA, 1999.
- [28] P. Holland and S. Leinhardt. Dynamic model for social networks. *Journal of Mathematical Sociology*, 5(1):5–20, 1977.
- [29] B. W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *The Bell system technical journal*, 49(1):291–307, 1970.
- [30] R. Leenders. Models for network dynamics: A markovian framework. *Journal of Mathematical Sociology*, 20:1–21, 1995.
- [31] R. Leenders. *Structure and Influence. Statistical Models for the Dynamics of Actor Attributes, Network Structure, and Their Interdependence*. Thesis Publishers, Amsterdam, 1995.

- [32] M. Magdon-Ismail, M. Goldberg, W. Wallace, and D. Siebecker. Locating hidden groups in communication networks using hidden markov models. In *International Conference on Intelligence and Security Informatics (ISI 2003)*, 2003. submitted.
- [33] M. Magdon-Ismail, A. Nicholson, and Y. S. Abu-Mostafa. Learning in the presence of noise. In S. Haykin and B. Kosko, editors, *Intelligent Signal Processing*. IEEE Press, 2001.
- [34] T. F. Mayer. Paries and networks: Stochastic models for relationship networks. *Journal of Mathematical Sociology*, 10:51–103, 1984.
- [35] P. Monge and N. Contractor. *Theories of Communication Networks*. Oxford University Press, 2002.
- [36] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, Cambridge, UK, 2000.
- [37] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [38] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, pages 4–15, January 1986.
- [39] A. Sanil, D. Banks, and K. Carley. Models for evolving fixed node networks: Model fitting and model testing. *Journal of Mathematical Sociology*, 21(1-2):173–196, 1996.
- [40] D. Siebeker, M. Goldberg, M. Magdon-Ismail, and W. Wallace. A Hidden Markov Model for describing the statistical evolution of social groups over communication networks. Technical report, Rensselaer Polytechnic Institute, 2003. Forthcoming.
- [41] T. Snijders. Analysis of longitudinal data using the hierarchical linear model. *QUALITY & QUANTITY*, 30(4):405–426, 1996.
- [42] T. Snijders. Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology*, 21(1-2):149–172, 1996.
- [43] T. Snijders. The statistical evaluation of social network dynamics. In M. Sobel and M. Becker, editors, *Sociological Methodology dynamics*, pages 361–395. Basil Blackwell, Boston & London, 2001.
- [44] M. Stelmack, N. N., and S. Batill. 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, September 1998.
- [45] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Series in Statistics. Springer Verlag, New york, 1982.
- [46] S. Wasserman. *Stochastic Models for Directed Graphs*. PhD thesis, Department of Statistics, Harvard University, 1977.

- [47] S. Wasserman. Analyzing social networks as stochastic processes. *Journal of the American Statistical Association*, 75(370):280–294, 1980.
- [48] S. Wasserman. A stochastic model for directed graphs with transition rates determined by reciprocity. In K. Schuessler, editor, *Sociological Methodology*, pages 392–412. Jossey-Bass, San Francisco, 1980.
- [49] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, 1994.
- [50] D. J. Watts. *Small Worlds: The dynamics of networks between order and randomness*. Princeton University Press, Princeton, NJ, 1999.