Statistical Modeling of Social Groups on Communication Networks

Mark Goldberg
Rensselaer Polytechnic Institute
goldberg@cs.rpi.edu

Paul Horn
Rensselaer Polytechnic Institute
hornp@cs.rpi.edu

Malik Magdon-Ismail
Rensselaer Polytechnic Institute
magdon@cs.rpi.edu

Jessie Riposo
Rensselaer Polytechnic Institute
ripos@rpi.edu

David Siebecker
Rensselaer Polytechnic Institute
siebed@cs.rpi.edu

William Wallace
Rensselaer Polytechnic Institute
wallaw@rpi.edu

Bulent Yener
Rensselaer Polytechnic Institute
yener@cs.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). 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.

Contact:
Prof. Malik Magdon-Ismail,
Dept. of Computer Science,
Rensselaer Polytechnic Institute,
Troy, NY 12180.

Tel: 1-518-276-4857
Fax: 1-518-276-4033
Email: magdon@cs.rpi.edu

Key Works: Hidden Markov, reverse engineering, learning, micro-laws, probabilistic evolution

Acknowledgement: We would like to thank Seyit Camtepe and Fikret Sivrikaya who were instrumental in collecting the data on which our simulations are based.

Statistical Modeling of Social Groups on Communication Networks

Mark Goldberg, Paul Horn, Malik Magdon-Ismail, Jessie Riposo,
David Siebecker, William Wallace and Bulent Yener

A social group is a collection of individual social units, or *actors* ([6]) 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 present new methodologies for the analysis of social groups that 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; identifying hidden or emerging communities; non-intrusive control; modeling contagion spreading.

An important contribution, provided within our framework, is the introduction of machine learning methods for discovering the laws governing a societies evolution. Traditionally, one hypothesizes laws and then validates them against data. Our approach is to start with the data, and then *learn* the individual behavior or "laws". The learned behavior can then function as a theory and be used to build models for predictive purposes.

We model a communication network using a Hidden Markov Model, which 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). Details of the general theory of Hidden Markov models can be found in [3]. In developing our model, we rely heavily on the foundations that have been laid in Monge and Contractor's most recent book [2], where they point out the most widely accepted and established communication theories.

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. Observable quantities such as the average communication density 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 variables will be called **micro variables** and the laws governing the evolution of micro variables will be the **micro-laws**.

The actors act autonomously, according to individual preferences. An example would be to leave a certain group and join another. An actor might (for example) have an affinity for large groups, or may have 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, an actor may choose to leave that group and join a larger one. A society's evolution is ultimately determined by the individual actions and hence preferences of all its members. A model of such an evolution must include diverse *types* of actors, each with a set of numeric parameters that govern her particular probabilistic behavior. One can test a particular set of such parameters by comparing a simulated evolution of the society (according to these parameters) with the actual evolution of that society. Traditionally this is the mode in which such research to discover laws is performed. To illustrate, we introduce a specific example.

Example: Newsgroup Societies. 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, which shows the evolving time series of communications for 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 communication. 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 behavior into a

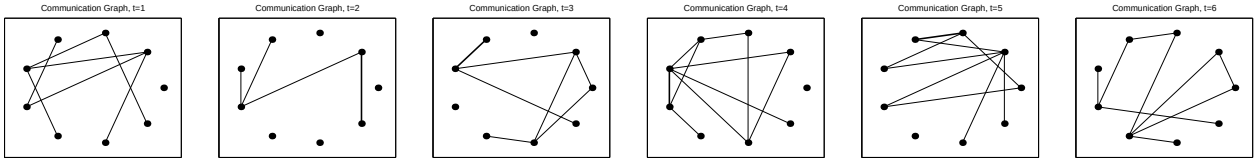
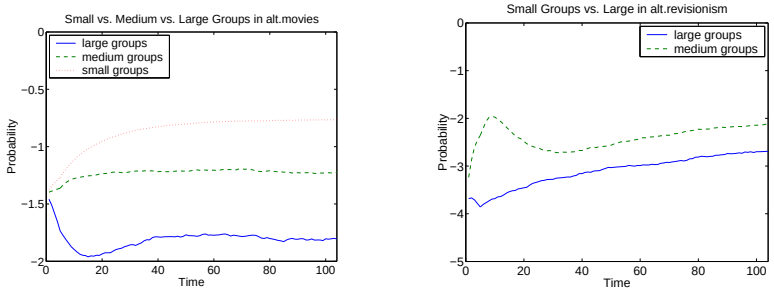


Figure 1: Communication time series of a hypothetical society.

quantitative realization of that law. Our model allows for a parameter that specifically governs the propensity of an actor for certain sized groups, which we set so that each actor has a high preference for large groups. We can now evolve or *simulate* the groups according to this "law" that actors prefer large groups.

What we observe in the newsgroup society is the time series of communications, **not** the time series of groups. We connect the group structure to its communication graph using a Poisson model for the communication intensity. In this way we can simulate a time series of communications. We compare the simulated communications with the true, observed, communications using a *metric*. One natural metric is the probability of obtaining the observed communications given the predicted group structure. Another metric that was suggested in [4] is the Hamming metric. Using this metric, we can compute the *distance* between the observed and simulated communications. The metric is chosen so as to accurately represent the properties of the communications that are important. If the simulated and actual communication evolutions are similar (according to the metric), then 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.

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 [1], however, we have arrived at them using a general model, in an automated fashion.

General Probabilistic Model

The exact specification of the model contains many technical details which we postpone. 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 the group structure – 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. 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. An 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 set of parameters, $\boldsymbol{\theta}$, which are *a priori* unknown but fixed, determine how these probabilistic decisions are made. Then, $\mathcal{I}_{t+1}$ has a distribution dependent on $\mathcal{I}_t$ and $\boldsymbol{\theta}$, given by

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

Here $\mathcal{Q}$ is a function that takes as input the current micro-state $\mathcal{I}_t$, the parameters $\boldsymbol{\theta}$ 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 $\mathcal{Q}$ amounts to specifying the *model* for how the society is evolving, which are the micro-laws for the society. Specifying $\boldsymbol{\theta}$ 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, one of the parameters specified the propensity of an actor for a particular sized group – this parameter would be a part of $\boldsymbol{\theta}$. $\mathcal{Q}$ 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.

The micro-state $\mathcal{I}_t$ is not observed; rather, it induces a *macro-state* $\mathcal{S}_t$, which depends probabilistically on the micro-state, and is also specified by a probability distribution,

$$P[\mathcal{S}_t|\mathcal{I}_t, \boldsymbol{\lambda}] = \mathcal{G}(\mathcal{S}_t, \mathcal{I}_t, \boldsymbol{\lambda}), \tag{2}$$

Here, $\mathcal{G}$ is a function that takes as input the current micro-state $\mathcal{I}_t$, a set of society dependent parameters $\boldsymbol{\lambda}$, 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. $\boldsymbol{\lambda}$ 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 $\boldsymbol{\lambda}$. In the newsgroup example, the observed macro-state is the set of observed communications, which are governed by the parameters of the Poisson process. Since $\mathcal{I}_t$ is a Markov process and $\mathcal{S}_t$ is derived from $\mathcal{I}_t$, $\mathcal{S}_t$ follows a **Hidden Markov process**.

In our model, the types of actions that an actor can take are to create new groups, leave some groups and/or join some other groups, and improve her rank in her current groups. The micro-laws that specify our model determine (probabilistically) the actions each actor makes. 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 ([6, 2]).

Model validation. It is not possible to observe the micro-laws directly. If the micro-law parameters adequately predict the future macroscopic evolution of the society, we then have indirect confirmation of our micro-laws and parameters. This process was illustrated in the newsgroup example, where we tested the preference of actors for different sized groups.

Reverse Engineering

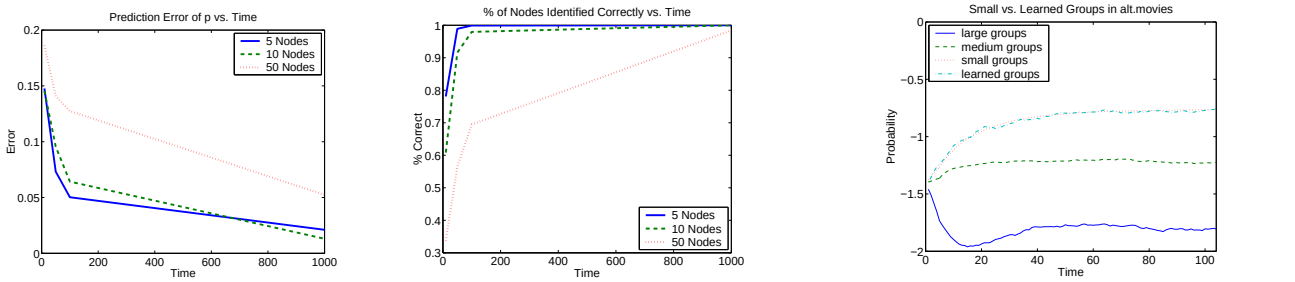
To illustrate one of the key contributions, we come back to the newsgroup example. The customary approach would be to hypothesize the "laws" of a society, and then validate the hypothesis against data, as was done in the newsgroup example with small/large group preferences. Why not let the data itself determine the "laws", instead of hypothesizing them? To be more concrete, suppose we *did not* know what sized groups our society members preferred. In other words, we propose to determine for a given society what the appropriate laws governing that society are by *reverse engineering* them *from the data alone.*

The goal is to determine what the appropriate parameters $\boldsymbol{\theta}, \boldsymbol{\lambda}$ 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, and comes under the general paradigm of learning [5]. We wish to *learn* the parameters of our Hidden Markov model, which we can do using Baum-Welch type algorithm for parameter estimation [3]. Such algorithms maximize the likelihood of observing the data given the parameters, $P[Data|\boldsymbol{\theta}, \boldsymbol{\lambda}]$.

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. We use 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, we use graph-partitioning algorithms that minimize the communication level between the partitions, leaving the within-group communication level higher than average.

Results. First, we investigate the accuracy of reverse engineering on a small "simulated" society. Figure 2(a) and (b) 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. The task was to classify each actor into its correct type, and determine a parameter p that determined individual actor behavior. We observe that



(a) RMS Error in parameter p (b) % of nodes identified correctly (c) Comparing the group preference laws.

Figure 2: Reverse engineering results.

the accuracy increases with more data, which is not surprising. Further, the more complicated the model (more actors), the harder it is to accurately determine the model parameters, also not surprising.

Next we reverse engineered the the parameters of two newsgroup societies using half the data, and predicted the future communications on the other half. We compared these predictions with the small/large group hypotheses. The data we used was collected from the news servers alt.movies and alt.revisionism. Using this data, we implemented an initial clustering to get initial groups, which we fixed. We then obtained the optimal group size propensity parameter, $\boldsymbol{\theta}$.

Figure 2(c), shows preliminary results comparing the different laws. It is clear that the only competitor with the learned law is small group preference law, which is slightly worse. More detailed simulations are currently underway to compare these, and use more efficient learning.

Concluding Remarks

We were lucky to "guess" the small group preference law. Our learned law was capable of explaining the data better than the best guessed law. using learning. Further, the law obtained using the learning algorithm was arrived at in a completely automated fashion. This is the key contribution of this work, namely the ability to automatically deduce the laws that may govern a society using only the data.

References

[1] B. Butler. The dynamics of cyberspace: Examing and modelling online social structure. Technical report, Carnegie Melon University, Pittsburgh, PA, 1999.

[2] P. Monge and N. Contractor. *Theories of Communication Networks.* Oxford University Press, 2002.

[3] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[4] 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.

[5] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer Series in Statistics. Springer Verlag, New york, 1982.

[6] S. Wasserman and K. Faust. *Social Network Analysis.* Cambridge University Press, 1994.