

## A Research Infrastructure for Machine Learning in Social Networks<sup>1</sup>

Matthew Francisco  
Department of Science &  
Technology Studies  
([francm@rpi.edu](mailto:francm@rpi.edu))

Hung-Ching Chen  
Department of Computer Science  
([chen3@cs.rpi.edu](mailto:chen3@cs.rpi.edu))

Mark Goldberg  
Department of Computer Science  
([goldberg@cs.rpi.edu](mailto:goldberg@cs.rpi.edu))

Malik Magdon-Ismail  
Department of Computer Science  
([magdon@cs.rpi.edu](mailto:magdon@cs.rpi.edu))

William Wallace  
Department of Decision Support &  
Engineering Sciences  
([wallace@rpi.edu](mailto:wallace@rpi.edu))

One of the outcomes of having large time-series social network data sets is the emergence of machine-learned or reverse-engineered theories of actor behavior, which have been referred to as *micro-laws* (Goldberg, et al. 2003) and *temporal social positions* (Christley and Madey 2007). Micro-laws and temporal social positions are essentially ways of classifying actors in a network based on how the network changes and how an actor maneuvers within this history of change. We extend these machine-learned theories to encompass a broader program of research. We call these *network practices*. Network practice extends machine learning in social networks by theorizing the choices for structuring and bounding a machine-learning approach by using multi-agent modeling. Our modeling laboratory, ViSAGE (Baumes, et al. In Press), which we highlight in this paper, supports such theorizing.

This paper discusses two questions regarding network practices. First, what research strategies are available for thinking about network practices and how does one further formalize these strategies into a research infrastructure or experimental system? We argue that ViSAGE is a key tool/methodology in growing such an experimental system. Second, what can be done with these classifications and how are we to make sense of them in regard to specific questions about human behavior and practices in naturalistic settings? Identifying ways of bringing grounded case research to verify and add richness to network practices ought to be a goal for model building and machine learning.

The network practice concept is facilitated by two aspects of the current research infrastructure for machine learning in social networks. First is the software development work being done to make machine learning a possible technique in agent-based modeling. This includes developing software packages for use in agent-based modeling libraries such as Repast (Gieseler 2005) and developing frameworks for integrating agent-based modeling and machine learning (Rand 2006). The second part of the

---

<sup>1</sup> This material is based upon work partially supported by the National Science Foundation under Grants No. 0324947, No. 0323324, No. 0634875, and by the ONR Grant No. N00014-06-1-0466. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or U.S. Government.

infrastructure is data and online repositories that facilitate community experimentation and analysis of the data. One example here is the Open Source Software Research Collaboratory (Gao, Antwerp, Christley and Madey 2007), which is used for machine learning studies, among other approaches, on data from Sourceforge.net.

ViSAGE, which stands for the Virtual Laboratory for Simulation and Analysis of Social Group Evolution, is a modeling and simulation tool for building multi-agent models of communities. These models are all built around the notion of the social group. The social group (Campbell 1958) is a key unit in the development or “evolution” of a community because groups mediate the interaction between activity at the individual level and change at the community level. Once specified, ViSAGE models are used as a framework for learning the categories and qualities of a community through machine-learning techniques. Depending on how a ViSAGE model is specified the machine learning process ought to categorize the community in different ways.

The contribution that ViSAGE provides is a way of exploring categories using a theoretical model building approach. Earlier we used ViSAGE to develop several models of social capital practices, i.e. how individuals in communities with different value systems for engaging groups produce social capital (Francisco, et al. In Press). Each of these models can be used to classify the individuals and groups in time-series, social network data. Because these categories arise from a social theorized machine-learning framework they should provide a good base for interpretation. We can ask, for example, does the network practice adequately describe real practices? For the mathematician conducting machine-learning experiments, the merit is in testing which framework learns rules and categories that best predict how a community evolves. That is, which network practice best reproduces the network patterns and predicts future network properties. This robust way of categorizing and interpreting, we argue, provides a key stepping-stone towards understanding the dynamics of human communities. Just as important, it provides a space where mathematical modelers interested in advancing machine learning can collaborate with researchers seeking to better understand differences in practices across case studies that are embedded in a single network.

Interpreting ViSAGE categories or micro-laws and developing theories of network practices require two activities. The first is to create theoretical bounds in the data using different arguments for what constitutes a community. For example, is SourceForge.net a single community or several communities? If SourceForge.net has sub-communities how are those communities distinguished? For example, Christy and Madey (2007) show that project administrators in SourceForge.net are “similarly embedded” in the network, which means they have a similar local social network context. We may further ask how the network practices and positions of project administrators in “scientific modeling” projects differ from project administrators in “game development” projects. Such a question, however, requires theorizing the difference between these two communities. Once these boundaries are postulated and specified each community can be categorized using the ViSAGE reverse engineering models.

The second step in interpreting ViSAGE models and the associated learned categories is achieved by gathering observational data on the practices of the social groups, which can be selected through the categorization process. The idea here is that modeling and machine learning research guide and direct case study research. Case studies research could be designed around a number of network practices. A simple example could be categorizing actors based on the size of groups they have a propensity for joining. One then could design a case study to see if the actor is aware of joining groups based on size. A case study strategy could also be designed to discover what other factors that may have caused such a propensity to emerge in the machine learning analysis based on what practices and activities occur locally in the field.

We are currently developing a study of the open source scientific modeling community to formalize and develop ViSAGE and the network practices concept. Our goal is to grow the research infrastructure for machine learning in social networks by further developing the available infrastructure of data and machine learning tools. Overall, the impact of formalizing and developing an infrastructure for discovering network practices is in bringing advanced methods in mathematical modeling and computer science (machine learning) together with theorizing and researching social phenomenon.

Baumes, J., et al. (In Press), "ViSAGE: A Virtual Laboratory for the Simulation and Analysis of Social Group Evolution," *ACM Transactions on Autonomous and Adaptive Systems*.

Campbell, D. T. (1958), "Common Fate, Similarity, and Other Indices of the Status of Aggregates of Persons as Social Entities," *Behavioral Science*, 3, 14-25.

Christley, S., and Madey, G. (2007), "Social Positions at Sourceforge.Net," in *Third International Conference on Open Source Systems (OSS 2007)*, Limerick, Ireland, p. 12.

Francisco, M., et al. (In Press), "Valuing Social Structure: An Agent-Based Simulation of Social Capital," *Social Networks*.

Gao, Y., Antwerp, M. V., Christley, S., and Madey, G. (2007), "A Research Collaboratory for Open Source Software Research," in *Proceedings of the 29th International Conference on Software Engineering + Workshops*, Minneapolis, MN, p. 5.

Gieseler, C. (2005), "A Java Reinforcement Learning Module for the Recursive Porous Agent Simulation Toolkit," Masters of Science, Iowa State University, Computer Science.

Goldberg, M., et al. (2003), "Statistical Modeling of Social Groups on Communication Networks," in *First Conference of the North American Association for Computational Social and Organizational Science*, Pittsburgh, PA.

Rand, W. (2006), "Machine Learning Meets Agent-Based Modeling: When Not to Go to a Bar," in *Agent 2006 on Social Agents: Results and Prospects*, Chicago, Illinois