Data Farming and Defense Applications

Gary Horne
Naval Postgraduate School
gehorne@nps.edu

Ted Meyer
Naval Postgraduate School
temeyer@nps.edu

Data farming uses simulation modeling, high performance computing, experimental design, and analysis to examine questions of interest with large possibility spaces. This methodology allows for the examination of whole landscapes of potential outcomes and provides the capability of executing enough experiments so that outliers might be captured and examined for insights. It can be used to conduct sensitivity studies, to support validation and verification of models, to iteratively optimize outputs using heuristic search and discovery, and as an aid to decision-makers in understanding complex relationships of factors. In this paper we describe efforts at the Naval Postgraduate School in developing these new and emerging tools. We also discuss data farming in the context of application to questions inherent in military decision-making. The particular application we illustrate here is social network modeling to support the countering of improvised explosive devices.

1.0 INTRODUCTION

Data farming uses simulation modeling, high performance computing, experimental design, and analysis to examine questions of interest with large possibility spaces. This methodology allows for the examination of whole landscapes of potential outcomes and provides the capability of executing enough experiments so that outliers might be captured and examined for insights. In this paper we will provide an overview of data farming and describe the six domains of data farming. We will also illustrate data farming in the context of application to questions inherent in military decision-making, in particular social network analysis related to countering improvised explosive devices.

1.1 Overview of Data Farming

Data farming uses simulation in a collaborative and iterative team process (Horne 1997, Horne and Meyer 2004). This process normally requires input and participation by subject matter experts, modelers, analysts, and decision-makers.

Data Farming focuses on a more complete landscape of possible system responses and progressions, rather than attempting to pinpoint an answer. This “big picture” solution landscape is an invaluable aid to the decision maker in light of the complex nature of the modern battle space. And while there is no such thing as an optimal decision in a system where the enemy has a role, data farming allows the decision maker to more fully understand the landscape of possibilities and thereby make more informed decisions. Data farming also allows for the discovery of outliers that may lead to findings that allow decision makers to no longer be surprised by surprise.

Data farming continues to evolve from initial work in a USMC effort called Project Albert (Hoffman and Horne 1998) to the work documented in the latest edition of the *Scythe* (Horne and Meyer 2010) documenting International Data Farming Workshop (IDFW) 20 held in March 2010 in Monterey, California. *The Scythe* is the publication of the International Data Farming Community that contains the proceedings of the IDFWs. IDFW 21 is scheduled to take place in Lisbon, Portugal in September 2010.

1.2 The Six Domains of Data Farming

The discovery of surprises and potential options are made possible by data farming. But many disciplines are behind these discoveries and their use in the overall data farming process evolved over a period of

time. In this section we give a brief account of this development.

Six realms or domains were incorporated into the data farming methodology from 1997 to 2002. Initial data farming efforts in the 1997-98 time frame relied upon two basic ideas:

1. Developing models, called distillations, which may not have a great deal of verisimilitude but could be focused to specifically address the questions at hand. (Horne 1999)

2. Using high performance computing to execute models many times over varied initial conditions to gain understanding of the possible outliers, trends, and distribution of results

The models need not be agent-based models, but because of the ease with which they can be prototyped, agent-based models were used during this beginning time period. This rapid prototyping facilitated the iterative nature of the approach the use of high performance computing to execute models many times over varied initial conditions to gain understanding of the possible outliers, trends, and distribution of results. Also, the huge volume of output from the simulations made possible by the high performance computing resulted in a need to develop visualization tools and methods commensurate with this tremendous amount of data. Thus, visualization of simulation data and rapid prototyping of scenarios became important to data farming efforts in the 1999-2000 time frame.

The simulations that defense analysts use are often large and complex. An evaluation of complete landscapes is extremely time consuming, sometimes not even possible. Also, even the smaller more abstract agent-based distillations referred to above can have many parameters that are potentially significant and that could take on many values. Thus, even with high performance computing and the small models used in data farming, gridded designs, where every value is simulated, are unwieldy.

Thus, using efficient experimental designs is essential and The Naval Postgraduate School in Monterey, California joined Project Albert researchers in the early 2000s with their expertise in this area. And NPS researchers have collaborated with others worldwide as well (see Kleijman, Sanchez, Lucas, and Cioppa 2005).

Finally, collaboration must take place at many levels if the full power of data farming is to be brought upon any question. Collaborative processes help to integrate the other five domains of data farming through interdisciplinary work in creating models and data farming infrastructure and during the iterative process of prototyping scenarios and examining output from model runs. Collaboration also takes place between people from different organizations and nations sharing information and perspectives at various points in approaching common questions.

With the addition of design of experiments and collaborative processes in 2001-2002 to data farming efforts, much attention then focused on the defense applications discussed in the next section. The six realms, or domains, discussed above that contribute to the data farming process are depicted in Figure 1.

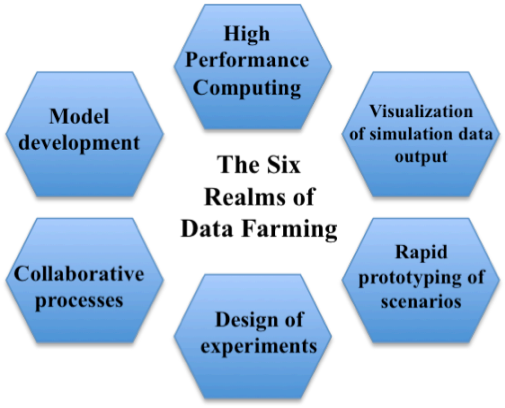

Figure 1. The Six Domains of Data Farming

1.3 Defense Applications

Since the incorporation of the above six domains into the process we call data farming, several articles have captured the fundamentals of data farming (e.g. Horne and Meyer 2005). But the key tenet in the data farming process has been the focus on the questions and since 2002 many application efforts have been documented. For example, at the Naval Postgraduate School many theses have been completed which have used data farming. And over the past decade, over 150 international work teams have formed around questions at International Data Farming Workshops.

These 150 work teams fall into areas, or themes, which include: Joint and Combined Operations (e.g. C4ISR Operations, Network Centric Warfare, Networked Fires, and Future Combat Missions), Urban Operations, Combat Support (e.g. UAV Operations, Robotics, Logistics, and Combat ID), Peace Support Operations, the Global War on Terrorism, Homeland Defense, Disaster Relief, and others.

The types of questions in these areas typically do not have precisely defined initial conditions and a complete set of algorithms that describe the system being considered. These questions address open systems that defy prediction. Data farming is used to provide insight that can be used by decision-makers. As an illustrative example, we now describe how data farming is being integrated with other techniques in the context of countering improvised explosive devices.

2.0 ILLUSTRATIVE APPLICATION: SOCIAL NETWORK MODELING TO SUPPORT THE COUNTER-IED FIGHT

This work represents results from an ongoing study to examine the utility of distillation modeling in the Counter-IED (Improvised Explosive Devices) fight. Understanding social networks, their nature in insurgencies and IED networks, and how to impact them, is important to the Counter-IED (C-IED) fight. This study, conducted as a team effort with international and inter-agency participation, is exploring methods of extracting, analyzing, and visualizing dynamic social networks that are inherent in models with agent interaction. This effort is being conducted in order to build tools that may be useful in examining and potentially manipulating insurgencies. The team started with a simple scenario that evolves cliques via interactions based on shared attributes. This simple model is the initial basis for the team's investigations and is being used to examine the types of network statistics that can be used as MOEs and pointers to unique and emergent behaviors of interest.

The team's initial goals were to extend this very basic scenario with simple variations and to test candidate tools and prototype methods for data farming the scenario, extracting network data, analyzing end-of-run network statistics, and visualizing network behaviors.

Social Network Analysis (SNA) techniques were explored in detail to determine which network metrics would be most beneficial for analyzing the types of networks produced by the agent-based scenario. Developing these tools and methods, and delineating applicable metrics will allow the exploration of questions regarding C-IED issues—including insurgent network evolution and adaptation.

Insurgent networks can be categorized into two groups of interest to C-IED efforts: IED Emplacement Networks (consisting of personnel that are directly involved with IED usage) and IED Enabling Networks (consisting of communities that indirectly support and enable the IED Emplacement networks). This study is identifying tools that can be used to explore patterns that might provide valuable insights into emergent behaviors of interest for both of these classes of networks.

2.1 Background

In previous work related to the use of agent-based modeling in the C-IED work, task plans aimed at addressing specific C-IED questions were developed. The current work is aimed at producing capabilities that can address these tasks. Tasks topics included: methods of indirect network attack; identifying important link layers for impacting the insurgent networks in specific environments, identifying important individuals, emergence of insurgent cells, eroding popular support for insurgent networks.

From this set of tasks the study team selected a set of candidate tasks for follow–up study and analysis. It was concluded that both data farming and SNA concepts and techniques needed to be applied to address the candidate tasks and that the current set of tools and methods available in these domains was not up to the task required.

The study team is working on developing the necessary tools and methods. In this effort we have:

- Demonstrated the ability to extract social network data from an existing scenario that included agent interaction, but that did not explicitly define a network. In this scenario the network "emerged" or evolved from the basic agent interactions.
- Data farmed this initial scenario and established the need to simplify the target scenario in order to more closely examine cause and effect relationships to SNA statistics.
- Developed a new base scenario, delineated a simple illustrative design of experiment (DOE), and data farmed the model to provide a sample data set for further exploration.
- Examined the utility of and approach to applying specific SNA statistics, methods, and concepts using the data farming output provided from previous work.
- Delineated the data requirements for the various types of networks that might be extracted from various modeling.
- Established and documented software and processes for applying these capabilities to detecting and analyzing emergent networks.

This work has lead to the study team's conviction that additional work needs to be accomplished in order to address C-IED-oriented problems. Generalized SNA/data farming tools that can be applied to output from various model types should have the capability to:

- Detect the presence of a network or networks.
- Distinguish different networks and different classes of networks.
- Determine if and when networks achieve equilibrium.
- Determine which model inputs have significant impact on the state and behaviors of the network.

Specifically, the intent is to use these capabilities to be able to address a variety of social network questions such as:

- What do insurgent networks look like? Who is in the network? Who is not?
- How do we distinguish networks that should be attacked, networks that should be attritted or that should be co-opted?
- Who are the High Value Individuals (HVI) and what are their identifiable characteristics?
- Will removing specific nodes destabilize a network?

- What are the 2nd and 3rd order effects of network manipulation?

- What are the potential unintended consequences?

2.2 Abstracted Illustrative Scenario and DOE

The Pythagoras agent-based model development environment was used for the initial scenarios. The first phase of activity was based on the Pythagoras distribution "Peace" scenario with some minor modifications of the source code to support the extraction of network interaction data. Data farming of this scenario demonstrated the ability to extract inherent emergent network data. Initial analysis of the results led to the development of a more basic scenario in order to test basic network concepts.

The illustrative "Clique Creator" (CC) scenario was developed using Pythagoras's "relative" color change capability as a tool for experimenting with SNA extraction and analysis. CC has a single agent class with 100 instantiated agents that are uniformly distributed across Pythagoras's red and blue color spaces. The agents' only "weapon" is "Chat" which induces a relative color change on other agents with which the agent interacts. As the scenario is executed, entities move through various color states, becoming "more" red or "more" blue depending on the interactions with other red-"ish" or blue-"ish" entities. States will change depending on whether two entities engage in "chatting" and form a connection. The more any two agents interact, the more "alike" they become.

The focus of the scenario selection was to represent dynamic homophily and use the results to explore the various analysis tools under study. Multiple excursions and replications of the Pythagoras-developed Clique Creator scenario were used to produce the data for analysis with the candidate tools. This baseline provided a means for the team to experiment with various SNA measures and analysis techniques.

Pythagoras can provide multiple views of agent state data. A spatial view showed the physical relationship between entitities and where connections or bonds were formed. The inclination space view sorted the entities by colors. This color space view is used to illustrate the homophilic state of the participating entities in the simulation.

A very basic full-factorial design space was used to data farm the scenario.

Table 1. Experimental Design Matrix

Excursion	RelativeChange	InfluenceRng	FriendThresh	EnemyThresh
0	5	25	5	60
1	5	25	50	105
2	5	25	100	155
3	5	100	5	60
4	5	100	50	105
5	5	100	100	155
6	5	250	5	60
7	5	250	50	105
8	5	250	100	155
9	20	25	5	60
10	20	25	50	105
11	20	25	100	155
12	20	100	5	60
13	20	100	50	105
14	20	100	100	155
15	20	250	5	60
16	20	250	50	105
17	20	250	100	155
18	60	25	5	60
19	60	25	50	105
20	60	25	100	155
21	60	100	5	60
22	60	100	50	105
23	60	100	100	155
24	60	250	5	60
25	60	250	50	105
26	60	250	100	155

The design matrix (Table 1) reflects four input parameters that will influence the composition of the resulting networks:

- ***RelativeChange*** - Percentage relative change of color when "chatted."

- ***InfluenceRng*** - Maximum distance of chat.

- ***FriendThresh*** - Agents within this range are considered "linked."

- ***EnemyThresh*** – Dependent variable; is calculated as FriendThresh plus 55, in order to preserve the same Friend to Enemy Distance (equivalent to the "neutral" range) as was present in the base scenario.

The CC scenario can be considered as a metaphor for a group of people establishing relationships based on shared interests or desires (color space proximity) and physical proximity (relative agent location). Agents are drawn toward agents with similar color and move away from agents of disimilar color. The closer agents are in location, the more frequently they "chat" each other, and thus, the closer they grow in color space. Eventually, cliques of "like-interest" agent form and are impacted by other agents and cliques. The input parameters varied in the design matrix affect these behavioral processes in straightforward ways.

2.3 Visualizing the Dynamic Network State

Part of a toolset to examine social network dynamics is the ability to analyze the ongoing agent interactions, behaviors, and network responses. Co-visualizing the various aspects (layers) of network dynamics can potentially provide powerful insight into the network.

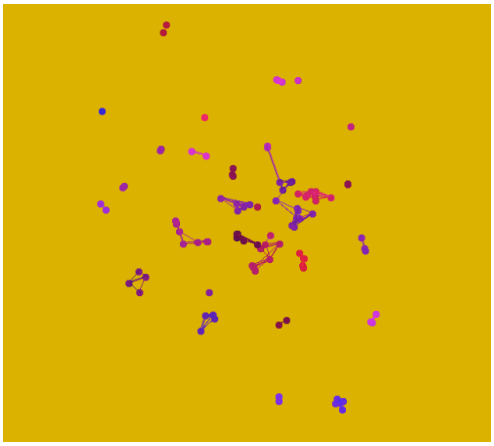

Figure 1. CC Scenario – Spatial View

The research team has done initial examination of the CC scenario using several visualization capabilites. Figure 1 is the spatial view provided by Pythagoras.

Figure 1 shows the agents at a time-step midway in the scenario. "Chats" are shown as lines between agents. This view, though, focuses on the location of the agent spatially.

Figure 2 shows four time-steps of an "inclination"-space view. In this image the location of the agents is based on their location in color space. The "redness" (0-255) of the agent is represented on the x axis. The "blueness" (0-255) of the agent is represented on the y axis. As the scenario proceeds left to right, top to bottom, note the congregation of agents into color groups. These groups do not represent the cliques formed though, because the spatial aspect is not represented.

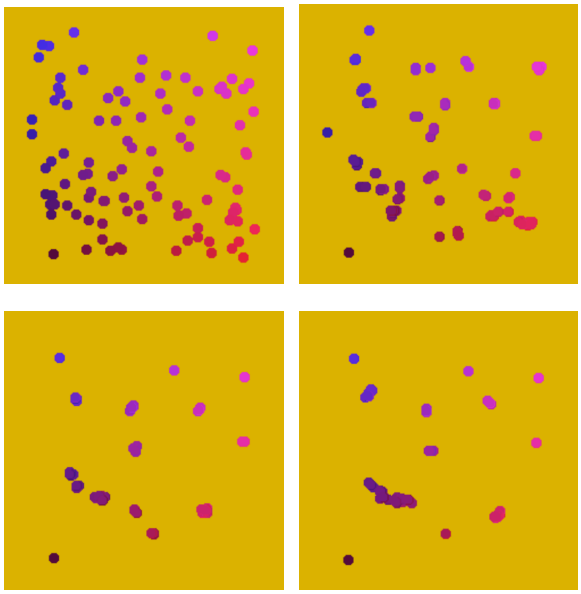

Figure 2. CC Scenario – Inclination Space View

FIgures 3 and 4 represent the same agent network , derived from the CC scenario, using the social network analysis "layout" generated by the R SNA plug-in and SoNIA software packages.

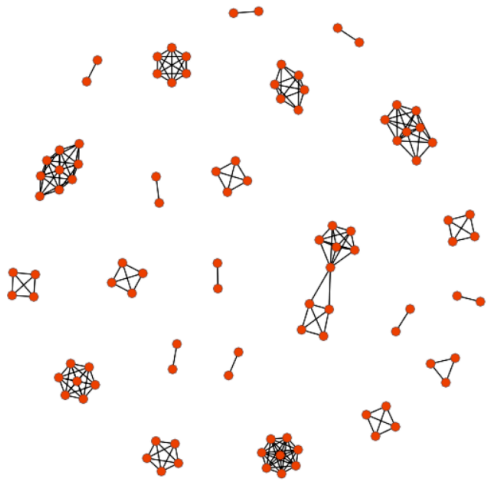

Figure 3. CC Scenario – Static Graph View

Figure 3 shows a static network layout representation of one of the CC time-steps using the default SNA layout algorithm. The SNA R package plots each time-step independently, not accounting for the layout defined in the previous time-step. The layout of each time-step is independent and as a result, the dynamic evolution is difficult to examine.

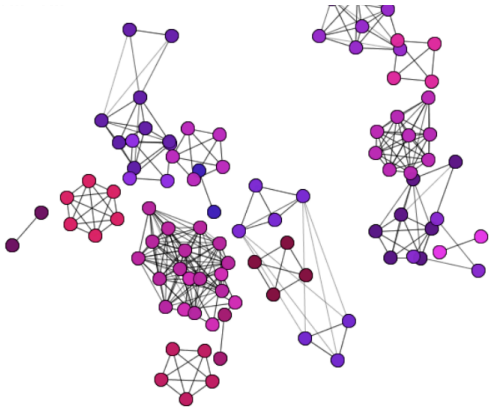

Figure 4. CC Scenario – Dynamic Graph View

Figure 4 shows a single time-step using the SoNIA application. SoNIA is designed to support dynamic time-series network data. As a result, the layout of any timestep can be based on the previous time step as a starting point. The result is a layout which displays the evolution of the network, but that can result in layouts that are not easily viewed statically.

It should be noted that Figures 2, 3 and 4 do not represent the spatial data shown in Figure 1 in any way... the "physical" location is ignored in these representations. In Figure 2 location represents color, and in Figures 3 and 4 the location is purely a function of the layout algorithm, which is designed to display the network in an uncluttered and easily-viewed manner, not the spatial location of the agents.

2.4 Social Network Analysis (SNA)

One of the team's goals is to begin to understand the utility of various SNA statistics in understanding the scenario dynamics and the result of data farming. Step one in this process was to delineate what outputs and analysis methods provide insight into network evolution and impact on agent behaviors.

SNA statistics fall into two classes: node statistics and network statistics. Node statistics include: betweenness, closeness, eigenvector centrality, and degree. Network statistics include: number of components, number of cliques, and average path length.

The study team decided to focus on node statistics initially and produced time-series output for every node of betweenness, eigenvector centrality and degree. Although data for 27 excursions of data farming was collected, it was decided to do an intial comparison of three excursions, where the primary variation was the color distance that defined what is considered a friend (a homophilic link). Excursions 0, 1, and 2 were examined.

Figure 5 represents one replication each from excusions 0, 1, and 2 as delineated in Table 1. The three plots represent the degree of each agent over time. The vertical axis is degree (the number of links associated with a node), the horizontal axis is time, and the axis going into the page is agent number. Figure 5 was generated using the PlotGL plugin to R.

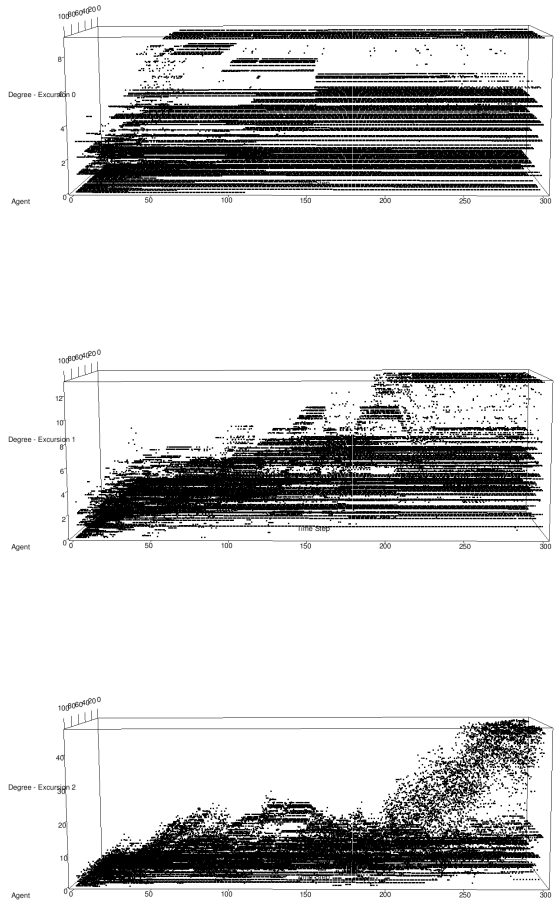

Figure 5. Centrality for Excursions 1-3

In Figure 5, various pattern differences, related to the evolution and devolution of cliques and components, can be discerned There are obvious differences between the excusions, with 0 and 1 appearing to reach covergence, but 2 never converging. It can be seen that some agents reach a steady-state and maintain it for some time, while other groups of agents particpate in behaviours which lead to the growth and reduction of degree for groups of agents.

2.5 Results

Two counter-intuitive results presented themselves. Excursion 2, in Figure 5c, shows that an increase in FriendThresh, that is, expanding the range and number of agents that an agent has homophilic links with in color space leads to increased instability in terms of clique formation. The initial assumption was that this would affect the size of the cliques and number of components. The unexpected result is that this increase prevents the stabilization of cliques and network components. Rather, it appears that this increase results in groups being able to "steal" members from other groups more easily.

Another interesting behavior is the Excursion 0 (Figure 5a) degree variation that occurs before equilibrium. In this case it appears that larger components are formed intially, but that they devolve into smaller groups over time. The team intends to investigate the set of replicates associated with this excursion to determine whether this behavior is consistent for this level of FriendThresh.

3.0 SUMMARY AND WAY AHEAD

Significant insight was gained by team members in delineating capabilities needed in a toolkit for the extraction and analysis of dynamic social data from models. The following capabilities will be needed for ongoing data farming research of basic social networks:

- **Synching of Visualization**: Various representations of the dynamic network are useful, but examining multiple views of the network time-step synced would provide powerful relational insights.
- **Equilibrium Time**: Determining whether equilibrium occurs and how long it takes is often the first step in analysis.
- **Data Farming Time Window Reduction Size**: Dynamic network analysis requires defining what constitutes a link, for example, a single interaction or multiple interactions over some time window. Being able to data farm this time window would provide analysts insight into network basics.
- **Node Statistic Capability**: Degree, betweenness, eigenvector, and

closeness need to be extractable for each node, time-step, replicate, and excursion and then represented effectively.

- **Network/Component Statistic Capability**: The number of cliques, and components, density, and others need to be acquired for each time step, replicate and excursion.

- **Newcomer/Leaving Effects**: Measure the effects of dynamic birth and death of agents.

- **Network Boundary Effects**: Data farm the impact of varying the size and extent of the network.

- **MOEs** (end-of-run vs. time-series) Both end-of-run and ongoing behaviors may be important.

The study team intends to continue to delineate tool capabilities for data farming social network models. We intend to accomplish the folowing tasks in the upcoming months:

- Document tools and methods identified in previous work.

- Define model output requirements for SNA analysis.

- Expand the toolkit to include additional network, node, and link statistics.

- Expand data farming methods for other network layers including weapon and resource interaction, spatial, communication, and multiple "inclination" parameters.

- Continue detailed analysis of CliqueCreator data farming results.

- Test use of tools and methods on other models (MANA, Netlogo scenarios).

- Begin delineating insurgent IED network scenario.

4.0 ACKNOWLEDGMENT

The authors would like to thank the members of Team 6 at International Data Farming Workshop 19 in Auckland, New Zealand and IDFW 20 in Monterey, California for their contributions, insights, and support.

5.0 REFERENCES

- Carrington, Peter J., Scott, John, Wasserman, Stanley, 2005, *Models and Methods in Social Network Analysis (Structural Analysis in the Social Sciences)*, Cambridge University Press
- Henscheid, Z., Middleton, D., and Bitinas, E. 2007. Pythagoras: An Agent-Based Simulation Environment, Scythe Issue 1: 40-44. Monterey, CA.
- Hoffman, F. and Horne, G. 1998. *Maneuver Warfare Science 1998*. United States Marine Corps Project Albert. Quantico, VA.
- Horne, G. 1997. Data Farming: A Meta-Technique for Research in the 21st Century, briefing presented at the Naval War College. Newport, RI.
- Horne, G. 1999. Maneuver Warfare Distillations: Essence Not Verisimilitude. Proceedings of the 1999 Winter Simulation Conference, eds. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 1147-1151. Phoenix, AZ.
- Horne, G. and Meyer, T. 2004. Data Farming: Discovering Surprise. Proceedings of the 2004 Winter Simulation Conference, eds. R. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 171-180. Washington, DC.
- Horne, G. and Meyer, T. 2005. Data Farming Architecture. Proceedings of the 2005 Winter Simulation Conference, eds. M. E. Kuhl, N. M. Steiger, F.B. Armstrong, and J. A. Joines, 1082-1087. Orlando, FL.
- Horne, G. and Meyer, T. January 2010. Scythe, Proceedings and Bulletin of the International Data Farming Community, Issue 7, Workshop 19, SEED Center for Data Farming, Monterey, CA.
- Horne, G. and Meyer, T. August 2010. Scythe, Proceedings and Bulletin of the International Data Farming Community, Issue 8, Workshop 20, SEED Center for Data Farming, Monterey, CA.
- Kleijnen, J., Sanchez, S., Lucas, T., and Cioppa, T. 2005, A User's Guide to the Brave New World of Designing Simulation Experiments, INFORMS Journal on Computing, 17(3): 263-289. Hanover, MD.
- PlotGL R Package (http://cran.r-project.org/web/packages/plotgl/index.html)
- SNA R Package (http://cran.r-project.org/web/packages/sna/index.html)
- SoNIA Social Network Image Animator (http://www.stanford.edu/group/sonia/index.html)
- Wasserman, Stanley, Faust, Katherine, 1994, *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*, Cambridge University Press