

An Experiment on Graph Analysis Methodologies for Scenarios

A. J. Brothers
P. D. Whitney
K. E. Wolf

O. A. Kuchar
G. Chin

September 2005



Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.** Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY

operated by

BATTELLE

for the

UNITED STATES DEPARTMENT OF ENERGY

under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161
ph: (800) 553-6847
fax: (703) 605-6900
email: orders@ntis.fedworld.gov
online ordering: <http://www.ntis.gov/ordering.htm>



This document was printed on recycled paper.

(9/2003)

An Experiment on Graph Analysis Methodologies for Scenarios

A. J. Brothers
P. D. Whitney
K. E. Wolf

O. A. Kuchar
G. Chin

September 2005

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Summary

Visual graph representations are increasingly used to represent, display, and explore scenarios and the structure of organizations. The graph representations of scenarios are readily understood, and commercial software is available to create and manage these representations. The purpose of the research presented in this report is to explore whether these graph representations support quantitative assessments of the underlying scenarios. The underlying structure of the scenarios is the information that is being targeted in the experiment and the extent to which the scenarios are similar in content.

An experiment was designed that incorporated both the contents of the scenarios and analysts' graph representations of the scenarios. The scenarios' content was represented graphically by analysts, and both the structure and the semantics of the graph representation were attempted to be used to understand the content. The structure information was not found to be discriminating for the content of the scenarios in this experiment; however, the semantic information was discriminating. A Clustering Analysis/Chi Square metric was able to distinguish graphs drawn for the same scenario by different analysts from graphs based on different scenarios. It was not able to make the finer distinction between types of scenarios. While in principle the ontologies were meant to be directly related to the graphs, in practice issues arose that would be difficult to resolve as part of an algorithmic process. Future research could identify additional measures and analytic methodologies that could better discriminate graphs based on either structural or semantic information or the combination of the two.

Acknowledgements

We are grateful to T. Katy Bragg, Cynthia L. Henderson, Caroline F. Varley, and Katherine E. Henderson for reading the scenarios and constructing graphs without which this experiment would not have been possible.

Contents

Summary	iii
Acknowledgments.....	iv
1.0 Introduction.....	1
2.0 Description of Experiment.....	3
2.1 Description of the Four Scenarios and How They were Generated.....	3
2.2 Description and Process for Generating Graphical Elements.....	4
2.3 Purpose of the Experiment: The Analytical Challenge.....	5
3.0 Structural Analysis.....	6
3.1 Triad Topology	6
3.2 Structural Analytical Challenge.....	7
3.3 Analytical Methods.....	8
3.3.1 Profile Analysis.....	8
3.3.2 Euclidean Distance with Sign Test	9
3.3.3 F-Statistic	12
4.0 Ontological Analysis.....	15
4.1 Ontologies	15
4.2 Semantic Distance between Nodes	16
4.3 Semantic Distance Between Graphs	16
4.4 Measuring Graph Separation Using Cluster Analysis	17
4.4.1 Cluster Analysis.....	18
4.4.2 Chi Square Tests	19
4.5 Results of Ontological Cluster Analysis	20
5.0 Discussion.....	23
5.1 Issues that Arose in the Construction of the Ontologies.....	23
5.1.1 Richness of the Graphs as Compared to the Ontology	23
5.1.2 Variations Across Analysts.....	23
5.1.3 Use of External Knowledge.....	24
5.1.4 Limitations of WordNet.....	24
5.1.5 Semantic Distance Weights	25
5.2 Conclusions.....	25
5.3 Directions for Future Research	25
6.0 References.....	27
Appendix A, Scenarios	A-1
Appendix B, Analyst Generated Graphs.....	B-1
Appendix C, Profiles of Triad Counts	C-1

Figures

1.	Example of Analyst-Generated Graph.....	2
2.	Expected Distance Relationships among Graphs	3
3.	Two Graphs Each Consisting of a Single Triad.....	7
4.	Profiles for all Sixteen Graphs.....	8
5.	Comparison of Idealized Versus Actual Euclidean Distance Displayed as Relative Brightness.....	11
6.	Distance Distributions for Categories of Scenario Relationships 3.3.3 F-Statistic	12
7.	Distribution of F-statistics and P-values Across Elements of Signature Vector	14
8.	Simple Semantic Net with Weighted Connections	16
9.	Cluster Dendrogram for a Pair of Graphs	19
10.	Comparison of Idealized Versus Actual Ontological Distance Displayed as Relative Brightness as Revealed Through Chi Square Scores.....	21
11.	Chi Square Distributions for Categories of Graphs	22
B.1.	Analyst 1's Graph of Animal 1 Scenario.....	B-1
B.2.	Analyst 2's Graph of Animal 1 Scenario.....	B-2
B.3.	Analyst 3's Graph of Animal 1 Scenario.....	B-3
B.4.	Analyst 4's Graph of Animal 1 Scenario.....	B-4
B.5.	Analyst 1's Graph of Animal 2 Scenario.....	B-5
B.6.	Analyst 2's Graph of Animal 2 Scenario.....	B-6
B.7.	Analyst 3's Graph of Animal 2 Scenario.....	B-7
B.8.	Analyst 4's Graph of Animal 2 Scenario.....	B-8
B.9.	Analyst 1's Graph of Geeks 3 Scenario.....	B-9
B.10.	Analyst 2's Graph of Geeks 3 Scenario.....	B-10
B.11.	Analyst 3's Graph of Geeks 3 Scenario.....	B-11
B.12.	Analyst 4's Graph of Geeks 3 Scenario.....	B-12
B.13.	Analyst 1's Graph of Geeks 4 Scenario.....	B-13
B.14.	Analyst 2's Graph of Geeks 4 Scenario.....	B-14
B.15.	Analyst 3's Graph of Geeks 4 Scenario.....	B-15
B.16.	Analyst 4's Graph of Geeks 4 Scenario.....	B-16
C.1.	Plots of Raw Data for Each Scenario Type	C-1
C.2.	Plots of Raw Data for Each Analyst	C-2
C.3.	Plots of Data Normalized by Rows for Each Scenario	C-2
C.4.	Plots of Data Normalized by Rows for Each Analyst.....	C-3
C.5.	Plots of Data Normalized by Columns for Each Scenario.....	C-3
C.6.	Plots of Data Normalized by Columns for Each Analyst	C-4
C.7.	Plot of Raw Data for All Sixteen Graphs.....	C-4
C.8.	Plots of All Sixteen Graphs; Data Normalized by Rows	C-5
C.9.	Plots of All Sixteen Graphs; Data Normalized by Columns.....	C-5

Tables

1.	Euclidean Distance between Graphs Based on Triad Structure.....	10
2.	ANOVA Summary Table Showing Average F-statistics and P-values.....	13
3.	Contingency Table for Graph Shown in Figure 7.....	19
4.	Chi Square Scores for Contingency Tables Derived from Cluster Analysis Using Ontological Distance	20

1.0 Introduction

Visual graph representations are increasingly used to represent, display, and explore scenarios and the structure of organizations. The graph representations of scenarios are readily understood, and commercial products such as Analyst's Notebook (AN) (i2 Ltd 2005) are providing impetus for such a representation by supporting the creation and organization of graph representations of scenarios. The purpose of the research presented in this report is to explore the stability of quantitative assessments of those representations. We focused on representations of scenarios. An experiment was designed that incorporated both the contents of the scenarios and analysts' graph representations of the scenarios. The underlying structure of the scenarios is the information that is being targeted in the experiment and the extent to which the scenarios are similar in content. This information was represented graphically by our analyst subjects. Scenarios and information associated with particular "cases" are represented by graphs or similar structures. Figure 1 shows an example of a graph representation of a scenario; Appendix A provides the contents. This particular graph representation focuses on the relationships among individuals, organizations, and a few events that the analyst decided were "key." This type of representation is readily understood. A natural question is: to what extent can this representation be the basis for cross-scenario analyses? That is, are there similar scenarios, or scenarios with similar components, in the case files? Another question involves the extent to which the creator of a particular graph representation influences the view.

The research presented in this report involves first an experiment based on four scenarios (presented in Appendix A). A graph representation was drawn for each scenario by four analysts. The graphs were analyzed to assess whether the similarities in the scenarios were reflected in the analyses of the graphs and to assess the effect of the individual analyst/artist. This report explores statistical techniques for uncovering the underlying structure and content of the scenarios as revealed in the graphs. The graph representations under study here are fairly free-form. No restrictions are enforced for labeling, as opposed to restricting the graphs to particular forms of information present in the scenarios.

Section 2 contains a more detailed description of the experiment, gives an overview, and contrasts structural and semantic analytic methods for the graphs. Sections 3 and 4 provide more details on the structural and semantic methods, respectively, and present the results of each methodology. Section 5 discusses the conclusions, hypotheses, and recommended next research that emerges from the analysis of this experiment.

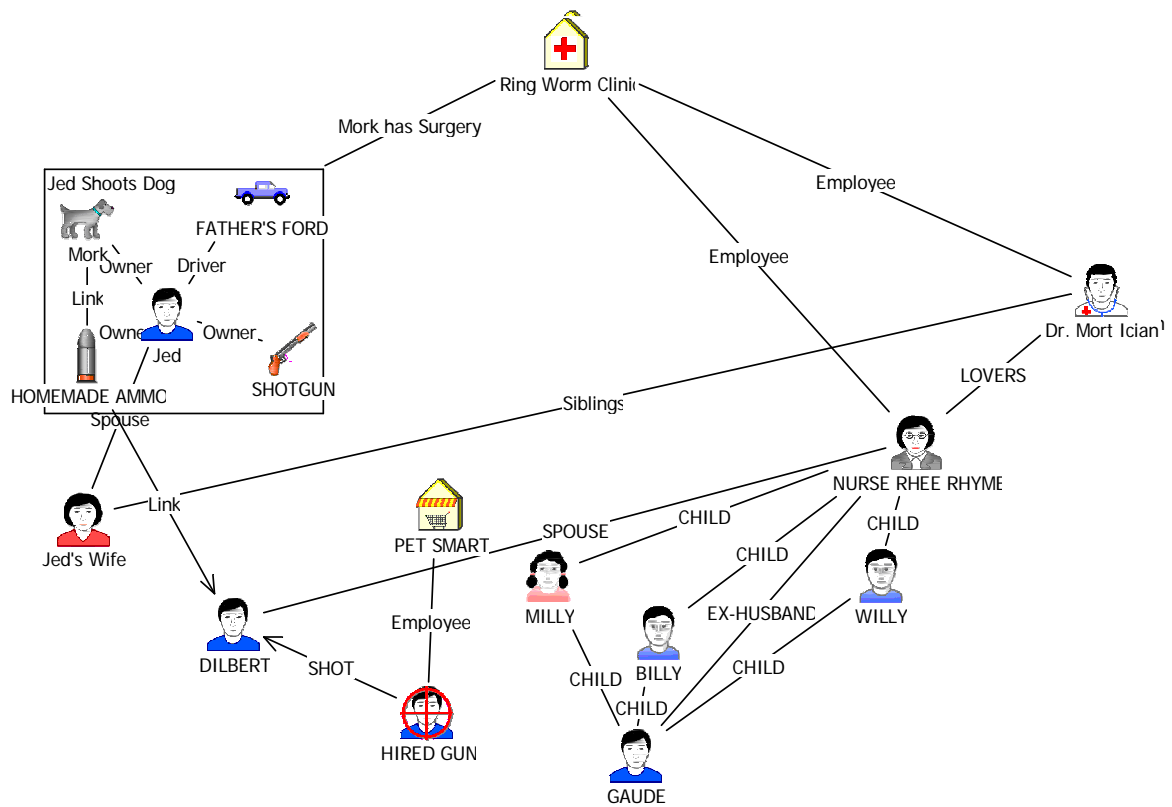


Figure 1. Example of Analyst-Generated Graph

2.0 Description of Experiment

Scenarios were generated such that their contents would be related in a specific way. These scenarios were given to analysts to read and process into semantic graphs. Statistical tests were carried out on the information in the graphs in an attempt to uncover the relationships among the scenarios.

2.1 Description of the Four Scenarios and How They were Generated

Four one-page scenarios were generated and given to four analysts to read, dissect, and portray in AN, creating a total of 16 analyses products to use as a test data set for the Scenario Graph Analysis experiment. Two of the four one-page scenarios were animal/pet-based and were generally different but contained overlaps in storyline and characters. The other two scenarios revolved around homeless individuals moving between two locations, these scenarios were also generally different but had overlap in characters.

The idea was that when each of the analysts charted out each of the scenarios, the resulting animal charts should be similar to each other but different from the homeless charts, and the homeless charts should be similar to each other but different from the animal charts. Further, within the animal chart category, we expected to see two mini-clusters of four charts each, with each mini-cluster revolving around either Animal Scenario 1 or Animal Scenario 2. Also, within the homeless chart category, we expected to see two mini-clusters of four charts each, with each mini-cluster revolving around either Homeless Scenario 3, or Homeless Scenario 4. These relationships are depicted in Figure 2.

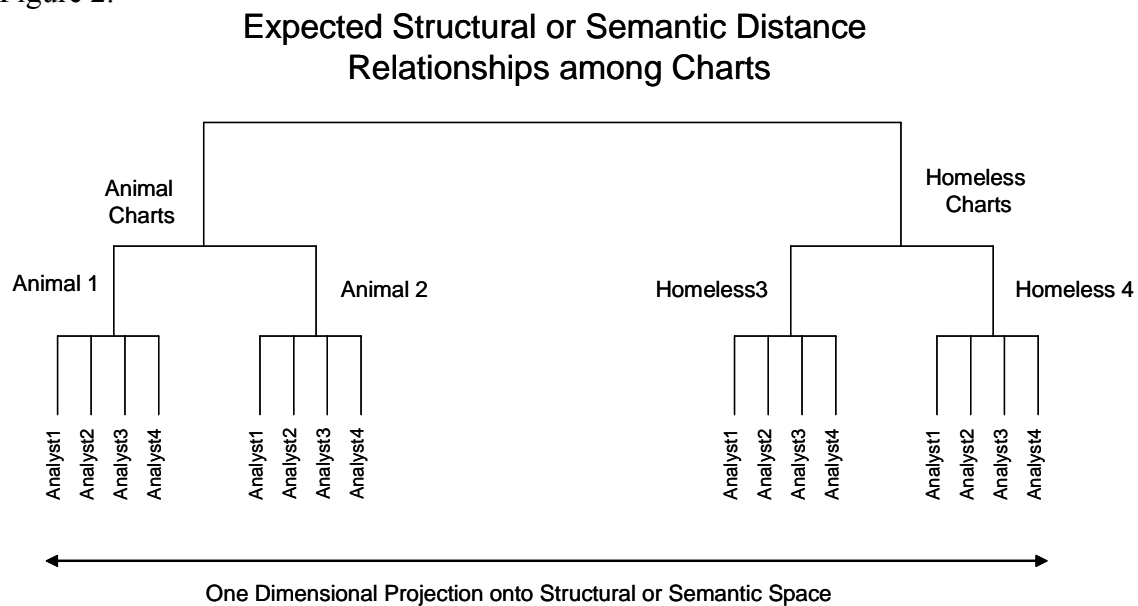


Figure 2. Expected Distance Relationships among Graphs

2.2 Description and Process for Generating Graphical Elements

Icon selection and naming conventions selected by each of the participating analysts impacted not only the visual appearance of each of their charts but also the results of the Scenario Graph Analysis. Since AN was the tool chosen for the analysts to use to visually depict their interpretations of the scenario, analysts were limited to the icons within AN. However, hundreds of icons are available within AN, and no guidance was given to limit the selection. As a result of this lack of control over icon selection, the four participating analysts used a wide variety of icons to represent people, places, pets, and events; that is, to represent what they viewed as the important elements in each scenario. While each individual analyst's icon selections were generally consistent across their individual four graphs, no two analysts used entirely similar conventions between each other¹.

Naming conventions used to identify scenario characters within AN also differed between analysts. For example, a fictional character, Susie Smith, could be labeled as either 1) Susie Smith, 2) Smith, Susie, 3) Smith Susie, 4) Susie S, 5) S. Smith, etc. How each analyst chose to name icons within AN impacted the results of the scenario graph analysis.

The level of detail of link descriptions also impacted how similar the AN charts appeared. If one analyst chose to label a relationship between two entities as "shooting" and a second analyst elected to label a relationship as "Fred shot Betty," while both labels were entirely accurate, the difference in the details represented in the link labels could also impact the results of the scenario graph analysis.

Sometimes external knowledge was brought into the graphs in ways that were simply incorrect, or at best not justified. For example, the firearm used in the accidental shooting of the dog was referred to as a shotgun. This is inconsistent with the use of bullets as stated in the scenario, because shotguns do not fire bullets; they fire shot. As another example, one analyst designated a particular individual to be a murderer, and while there might have been some evidence to suggest that, it was not conclusive evidence, and it was subsequently revealed not to be the case in a follow-on scenario.

The analysts also made assumptions that, while not explicitly stated, could be inferred from the choice of graphical icons. For example, the vehicle used to transport Mork, the dog, to the veterinarian was referred to as Father's Ford. The icon universally chosen to represent this was a pickup truck, yet it was never stated to be a truck in the scenario. Presumably, sedan icons were also available, but possibly a truck seemed to be a more logical choice for transporting a wounded animal, or it could be that the analysts simply associated pickup trucks with dogs.

¹ The analysts were prohibited from discussing the written scenarios or their interpretations of them until after all analysts had completed all scenarios

Perhaps future experiments of this sort could implement greater controls on naming conventions, icon selection, and link labels; however, these would be constraints that most analysts do not face in the “real word.”

2.3 Purpose of the Experiment: The Analytical Challenge

Sixteen graphs of the type shown in Figure 1 were generated by four analysts. The complete set of 16 graphs is reproduced in Appendix B. The purpose of this experiment was to discover, through statistical analysis, the underlying content relationships of the scenarios based upon the graphs as depicted in Figure 2. The analysis fell into two categories: techniques that focused primarily on the topological structure of the graphs independent of content, and techniques that focused on content or ontological information independent of structure. For any graph, there are 64 possible relationships between any three nodes, or triads. These triads are based on the structural relationships among the nodes independent of meaning. For each graph, one possible structural signature is the vector of the number of triads of each type. The analytical challenge for the structural analysis was to uncover the semantic relationships among the graphs just using this structural information. The method and results of the structural analysis is described in Section 3.

It is also possible to characterize each graph based on the semantic information independent of structure. An ontology of the semantic content based on the node labels can be constructed for each graph. The ontological challenge was to characterize the semantic distance relationships among the graphs based upon their ontologies. Section 4 describes the method and results of the ontological analysis.

3.0 Structural Analysis

For the purpose of this report, we take a fairly literal translation of Figure 1 as an information representation appropriate for scenarios². We use a highly labeled graph³ as a representation of a scenario. In particular, the representation is a collection of triples $n_i E_k n_j$ where:

$$\begin{aligned} n_i, n_j & - \text{are nodes} \\ E_k & - \text{is an edge} \end{aligned}$$

and each edge and node that comprises the graph has an associated label, free text, or structured information. This representation allows for the structures evident in existing scenario analysis tools. The proposed structure has no necessary temporal structure (although that type of information could be incorporated in the meta-data associated with nodes and edges). Also, the structure has no intrinsic support for discovery. However, in the next sections we will describe how the structure can take advantage of some existing information representations and indicate how the analytic activities of summarization and retrieval can be supported.

Let \mathcal{N} be a set of nodes and for each r , let \mathcal{L}_r be a relation on \mathcal{N} , for $r=1, \dots, R$. Then a multivariate graph is the combination of the node set and relations: $\mathcal{G} = (\mathcal{N}, \mathcal{L}_r)$. A measure in a graph is a mapping from the graph to the real numbers. The measures described below can be expressed in this notation. In the context of a multivariate graph, the degree could be specialized for various relations and subsets.

With this strategy, the challenge becomes how to create signature vectors for the graphs that represent scenarios. Several potentially useful measures on graphs and nodes of graphs have been proposed. See Whitney et al. (2004) for examples. Our research strategy is to select measurements/features that are fairly well established with perhaps some adjustments for the particular objectives at hand. In particular, we would like signature vectors for a pair of graphs that differ only by a single node or edge to be close to each other.

3.1 Triad Topology

The particular strategy for this experiment and in Whitney et al. (2004) is to use features that are primarily triad-based. For any two nodes, A and B, there are four possible links between them:

1. No connection
2. A link directed from A to B

² A more flexible representation would be a collection of n-ary relationships, which turns out to be an acceptable input that leads to the creation of graphics like Figure 1, when using Analyst's Notebook.

³ The information in Figure 1 is not, technically, a graph. See Schmidt et al., (1993) page 6. The figure represents can be viewed as a labeled, multivariate directed graph; see Wasserman and Faust (1994).

3. A link directed from B to A
4. A bi-directional link between A and B.

Thus, for each triple of nodes there are 4^3 or 64 possible connections. The structural characterization of a graph based upon triads is to consider all possible triples of nodes and count the number of occurrences of each of the 64 possible connections. So, a numeric signature vector of length $4^3=64$ can be constructed based on the occurrence of the 64 triad patterns. Two such triads appear in Figure 3. With a straight triad-count signature strategy, two graphs that are the triads in Figure 3 would have signatures that are orthogonal unit vectors. The first adjustment to the signature that was made was that the appearance of a triad not only resulted in an incremental count for that triad's coordinate, but also for all the coordinates corresponding with triads that contain the particular triad. Thus, the occurrence of the right triad of Figure 2 within a graph would increment both its coordinate and the coordinate for the upper triad.

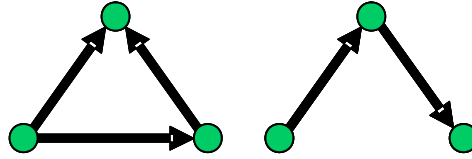


Figure 3. Two Graphs Each Consisting of a Single Triad

The modified triad-based signatures can be used to screen for whether one graph is a sub-graph of another. A useful characteristic of the modified signatures is that they are *sub-graph preserving*. Note that there are multiple objects that can be visually represented as a graph, for which the line of reasoning below would hold.

Definition: A measure m on a graph is *sub-graph preserving* if:
 G_1 is a sub graph of G_2 implies that $m(G_1) \leq m(G_2)$.

The modified triad signatures are sub-graph preserving. Additional examples of sub graph preserving measures include the order of a graph (number of nodes), the size of graph (number of edges), and the degree of a particular node (number of edges attached to a node) that is in each graph.

3.2 Structural Analytical Challenge

The structural analytical challenge is to find a metric that captures the similarities and differences among the graphs based upon the configuration of the nodes and links independent of semantic content. The metric should not exploit the particular data set from this experiment, and it should not be limited to a particular topology. It would take as input the vector of numbers used to define the structure, and its output should be some measure of the similarity or differences among the graphs.

3.3 Analytical Methods

A number of measures were considered as candidates to meet the structural challenge, and three were explored. These were profile analysis, Euclidean distance, and F-statistics.

3.3.1 Profile Analysis

A profile is a graph obtained by plotting the vector frequencies versus position number and joining the points. An intuitive description of the steps in a profile analysis is 1) plot the profiles on a single page, 2) check to see if the plots are parallel, and 3) if they are parallel, test to see if the heights are significantly different (Seber 1984). Profile plots were made of the data in three forms: 1) raw data of triad counts, 2) data normalized by dividing by row totals (sum of the vector elements), and 3) data normalized by dividing by columns totals (sum of the n^{th} tuples over all graphs). The complete sets of profiles are shown in Appendix B. The profiles for all 16 graphs using the raw data are shown in Figure 4. A noticeable characteristic of the plots is that, while the profiles are not parallel, they appear to be ratio-scaled relative to each other, and a pattern exists with few tall and medium spikes and the majority of entries being zero. The spikes represent triad patterns that are repeated in the graphs. The test for significant differences between graph types was further explored analytically with an F-statistic as described below.

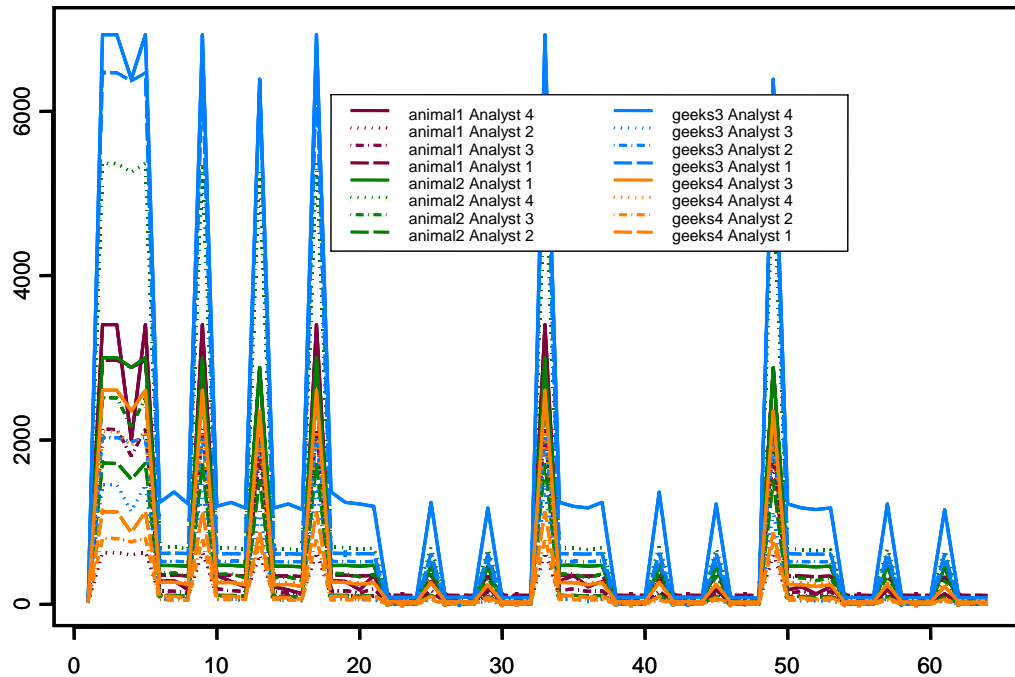


Figure 4. Profiles for all Sixteen Graphs

3.3.2 Euclidean Distance with Sign Test

A natural measure of the relative closeness of the triad vectors is their distance apart as points in 64 dimensional space. The Euclidean distance between the vectors is given by:



$$(1) \quad d(X, Y) = \sqrt{\sum_{i=1}^{64} (x_i - y_i)^2}$$

where X and Y are the triad vectors for two graphs and x_i and y_i are their respective values for the i^{th} elements. The pairwise Euclidean distances between all possible pairs of vectors was calculated and are shown in Table 1.

To better understand the pattern of these numbers, a matrix in which shade intensity corresponds to distance is shown in Figure 5. The top panel in Figure 5 shows the pattern that would be expected if Euclidean distance between the triad vectors reflected the relationship between the scenarios as designed into the experiment and depicted in Figure 2. The bottom panel in the figure shows the pattern that corresponds to the Table 1 distances that were actually found. This shading pattern implies that the same scenarios described by different analysts, and scenarios that are similar, are not any closer in distance than different types of scenarios. To further illustrate these concepts, the distributions of distances for different categories of pairwise relationships are shown in Figure 6. The top panel in the figure is the distribution of distances between graphs of the same scenario but described by different analysts. The middle panel shows the distribution of distances between graphs when the analysts are the same but there is a different scenario, and the bottom panel shows this information when both the scenario and analyst are different. In all cases, the distributions are skewed to the right, and the bottom panel in addition to being skewed is somewhat bimodal. These tables and figures show that Euclidean distance does not seem to be an effective measure for uncovering the underlying patterns of how the content of the scenarios are related as revealed through the triad vectors. However, it is possible that for other structural or semantic characteristics of the graphs Euclidean distance could be a more effective indicator.

Table 1. Euclidean Distance between Graphs Based on Triad Structure

	A1 Anal1	A1 Anal2	A1 Anal3	A1 Anal4	A2 Anal1	A2 Anal2	A2 Anal3	A2 Anal4	G3 Anal1	G3 Anal2	G3 Anal3	G3 Anal4	G4 Anal1	G4 Anal2	G4 Anal3	G4 Anal4
A1.Anal1	0	7088	2998	2028	844	4114	1703	7366	10589	2981	5051	12317	5879	6665	1445	2751
A1.Anal2		0	4232	7254	7275	3102	5490	14440	17649	4706	2239	19304	1315	586	5759	4365
A1.Anal3			0	3177	3271	1161	1515	10302	13465	1946	2083	15243	2955	3762	1583	631
A1.Anal4				0	2217	4285	2297	7773	10850	3704	5096	12609	5964	6805	2073	3201
A2.Anal1					0	4368	1877	7203	10454	2863	5296	12050	6085	6870	1765	2943
A2.Anal2						0	2589	11442	14603	2441	956	16382	1825	2613	2722	1468
A2.Anal3							0	8980	12214	1570	3488	13845	4279	5085	667	1162
A2.Anal4								0	3363	9997	12381	5159	13225	14015	8724	10083
G3.Anal1									0	13283	15540	3375	16410	17197	11903	13295
G3.Anal2										0	3150	14701	3685	4406	2111	1348
G3.Anal3											0	17308	936	1733	3658	2372
G3.Anal4												0	18117	18913	13672	14974
G4.Anal1													0	848	4510	3152
G4.Anal2														0	5309	3945
G4.Anal3															0	1431
G4.Anal4																0

 Same Scenario Different Analyst
 Same Analyst Different Scenario

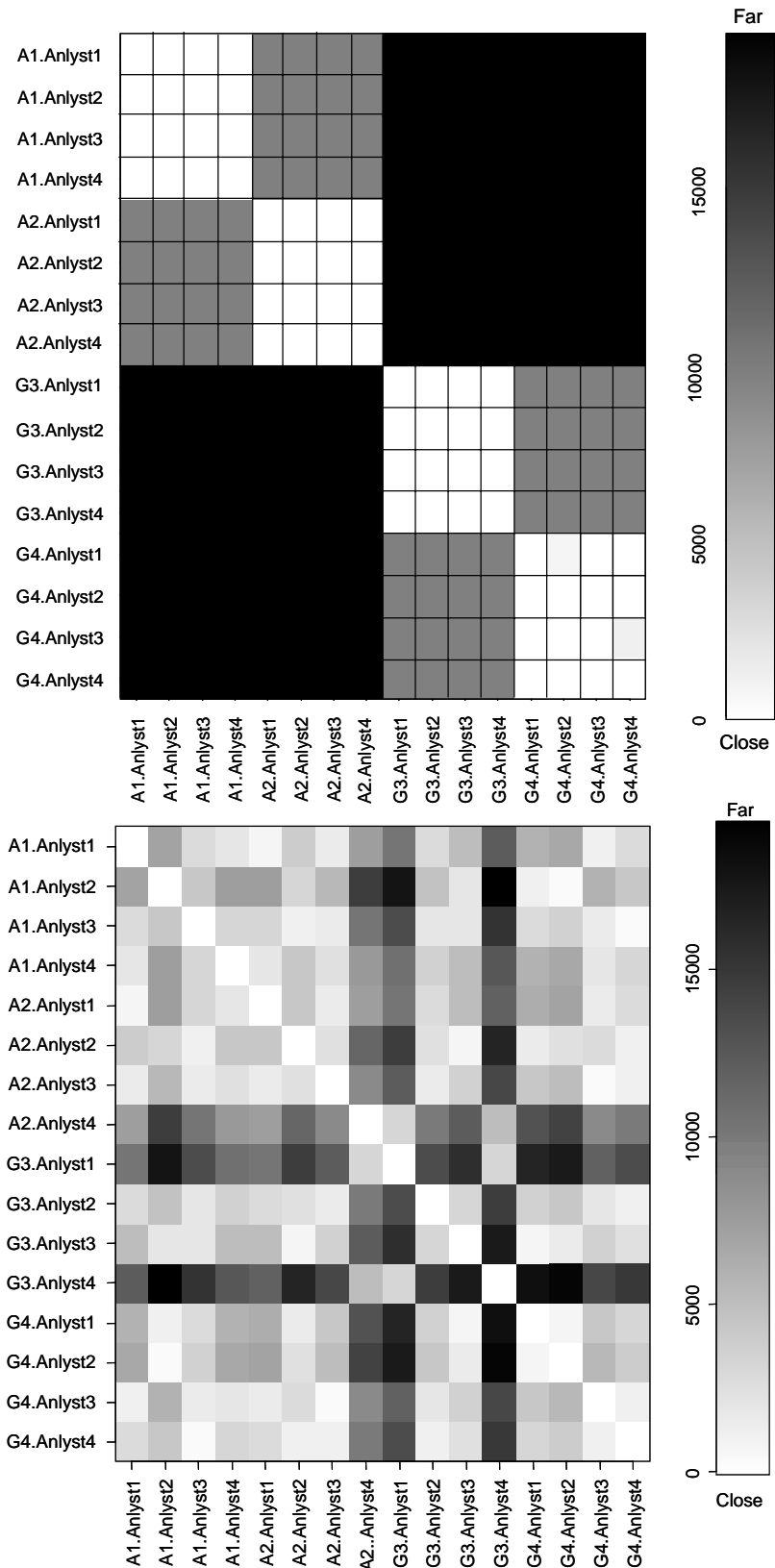


Figure 5. Comparison of Idealized Versus Actual Euclidean Distance Displayed as Relative Brightness

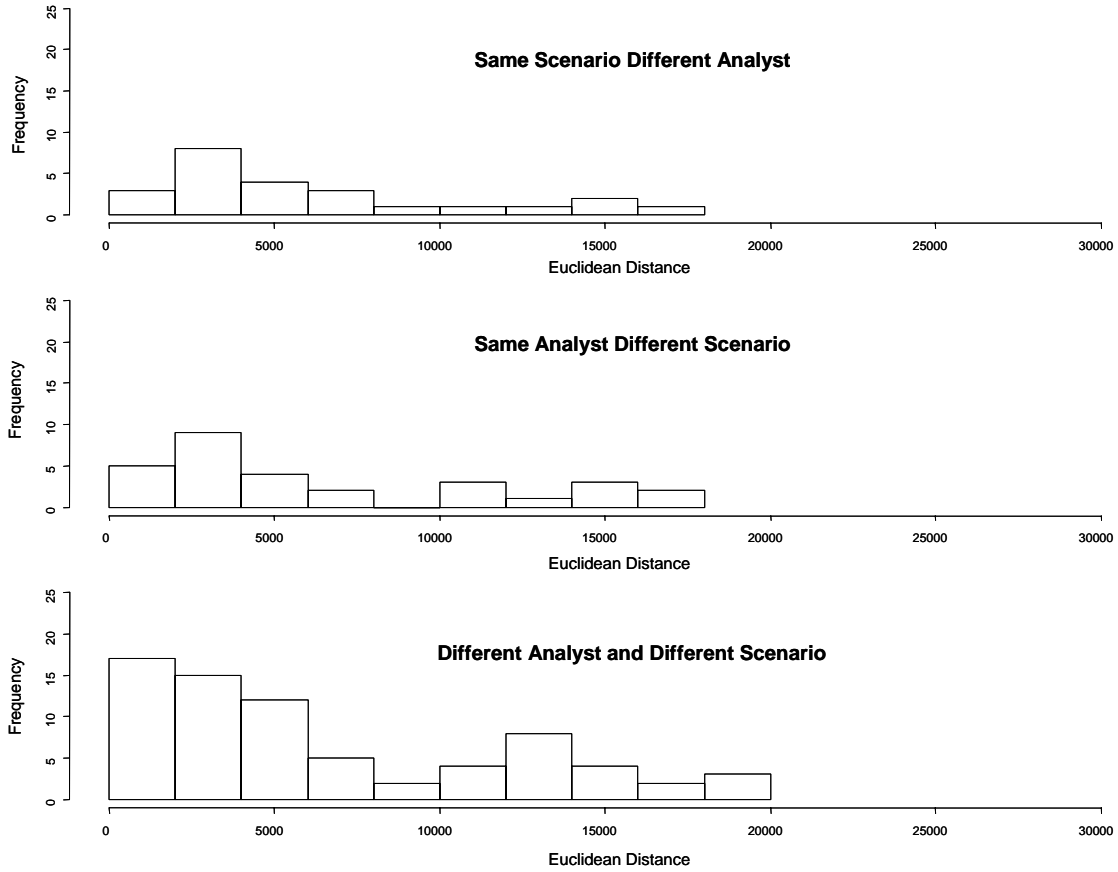


Figure 6. Distance Distributions for Categories of Scenario Relationships

3.3.3 F-Statistic

Recall from the profile analysis that we got spikes of varying heights. We would like to know whether the difference in heights corresponds to the underlying information structure in the scenarios. Restating this question in terms of the analytical challenge; we would like to show that the mean values of the vectors triads are different depending on the type of scenario and/or analyst. This is analogous to an experiment in which we want to show a difference among means for dependent variables for groups receiving different treatments. A statistical test for differences among group means is analysis of variance (ANOVA), which generates an F statistic. From a statistical perspective, the F statistic is distributed as the ratio of two chi squares. Intuitively, it can be thought of as a ratio of variances:

$$(1) \quad F = \frac{\text{Variance_Between_Scenarios}}{\text{Variance_Within_Scenarios}}.$$

Thus, the F-statistic measures the extent to which we can distinguish graphs of same type from the collective of all graphs. The more similar the value of the corresponding elements in the vectors within a particular scenario category, relative to the variation in

values between categories, the larger the F-statistic. Because the triad vectors are 64-tuples, we could have done a multiple analysis of variance (MANOVA), possibly just on the spikes shown in Figure 4, but we choose instead to carry out separate ANOVAs for each of the 64 triad types and average the F-statistics across triads. This is general procedure that does not exploit this particular data set and would generalize beyond the triad signatures. Two-way ANOVAs were carried out for each triad type using scenario type and analysts as grouping factors. With few exceptions, none of the F values were statistically significant. P-values of less than .05 are generally considered significant, but for 20 analyses that would be expected to occur once by chance. Here we had 64 analyses, and there was only one Scenario p-value less than .05 and six analyst values less than .05. More than half (33) of the Scenario p-values were moderate (≤ 0.10), and more than 20% of the analyst p-values were moderate (≤ 0.10). The distribution of F statistics and p-values for scenarios and analysts are shown in Figure 7.

Table 2 shows the average of the F-statistics and p-values. Interestingly, while the F-statistics are not generally statistically significant, the average values are higher for the analysts than the scenarios indicating more similarity within analysis across the scenario types than within scenarios across analysts. A paired comparison of the F-statistics shows that analyst's values were significantly greater ($t=3.2$, $df=63$, $p\text{-value}=.0021$).

Table 2. ANOVA Summary Table Showing Average F-statistics and P-values

Source	Df	Average F Value	Average Pr(F)
Scenario	3	1.93	0.345
Analysts	3	2.23	0.212
Residuals	9		

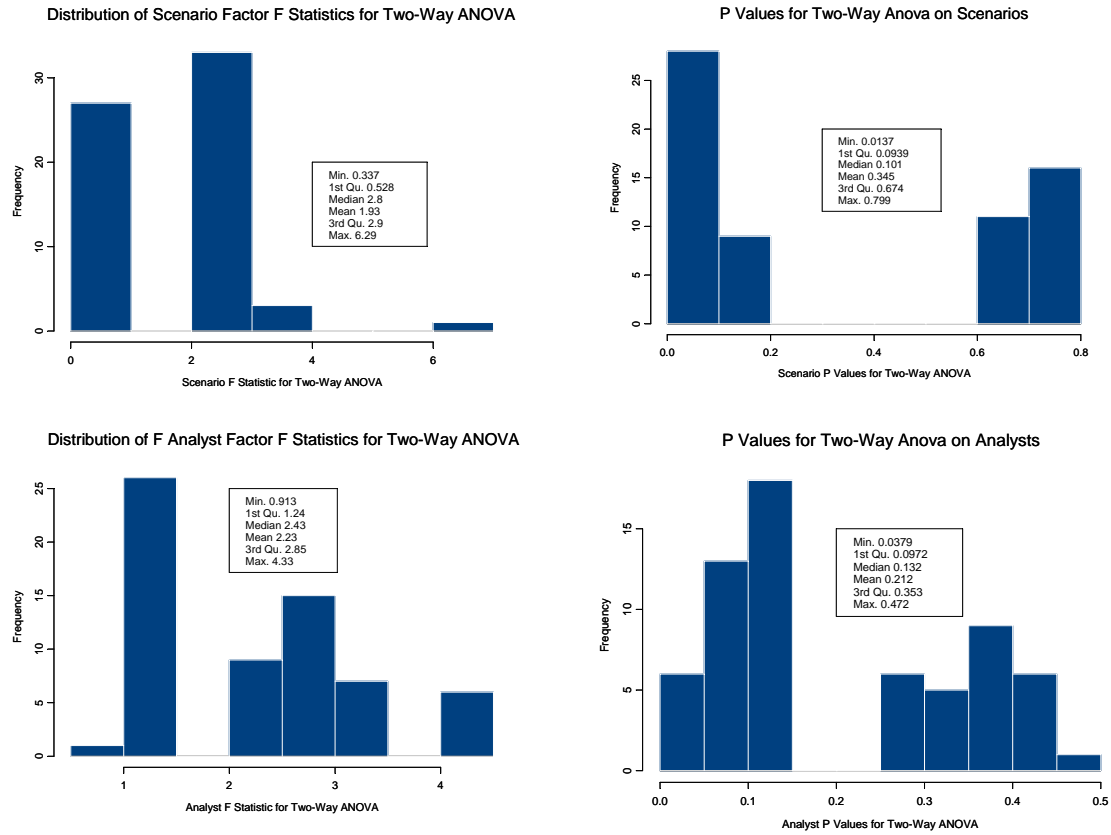


Figure 7. Distribution of F-statistics and P-values Across Elements of Signature Vector

4.0 Ontological Analysis

The purpose of the ontological analysis is to find a method to distinguish between the graphs based on the ontologies. Whereas the structural challenge is based on the topology of the graphs without consideration of semantics or content, the ontological challenge is to use only content information as defined by the ontologies. From a content perspective, graphs will be different to the extent that their respective objects of discussion, or nodes, form distinct non-intersecting sets. A possible measure of the distance between graphs is how far apart these nodes are, on average, in semantic space. The greater the disparity in content, the farther apart the nodes should be, on average, in semantic distance between each other. Consequently, we developed ontologies for each of the graphs as a first step in measuring the semantic distance between the nodes.

The content information in the graphs is captured by the labels associated with the nodes and edges in a graph representation of scenarios. For purposes of this experiment, we restricted the labels to individual concepts. For further convenience, and to take advantage of existing work, we adapted a subset of WordNet (2005) viewed as a semantic net.

4.1 Ontologies

For our purposes, an ontology is a collection of the concepts, and their specification and relationships, within a semantic net for a particular knowledge domain (after Gruber 2005). Ontologies were constructed for each of the four scenarios based on the four graphs produced by the four analysts using WordNet. WordNet was chosen because it is freely available and widely known and used. WordNet has been developed under the direction of George Miller at Princeton University's Cognitive Science Laboratory. It is inspired by current psycho-linguistics theories of human lexical memory (WordNet 2005). The method used in the construction of the ontologies was intended to follow a mechanical algorithm and be reproducible by anyone having access to WordNet. The leaves (end points) in the ontology correspond to the nodes of the graphs, and the branches consist of a sequence of hypernyms obtained from WordNet. Each of the words is separated by a number indicating the "semantic distance" between words. Generally, a default distance of 1 was used. The hypernym ordering is from general to specific, and the ending is the label or name of the leaf/node. The hypernym just before the leaf node (penultimate node) specifies how the leaf node is being characterized (e.g., father or spouse), and in general depends on how the links to the node are labeled. This penultimate node is the word that is used to enter WordNet, and it determines the sequence of hypernyms that make up the branch.

While the use of WordNet provides the advantage of an existing net, it also forces some limitations on label choices. WordNet only provides a net for nouns and verbs, and the verb net is often very limited, at least compared with the organization available for nouns. Whenever possible, nouns are used as labels in our examples to provide maximum possible information ("works for" is replaced with "employee," for example). In some

cases, we have terms that do not appear in WordNet’s lexicon and where no appropriate synonym can be found. A frequent example of this is proper nouns. In these cases, an appropriate parent for the term is selected such that the parent is in WordNet’s lexicon. For example, “Bob” is made an element of “male.” Additionally, for our examples, the word sense is selected manually where multiple senses are available.

A key assumption is that the net is a rooted tree, as there must be a path between two entries to determine distance. At present, our software ensures this by creating a generic “root” entry that is made the parent of all entries that have no natural parents. The ontologies constructed consist of a collection of hypernym lists—one for each concept. Consequently, the semantic net is implicitly defined by common root node in each hypernym list in the ontologies, but was not actually constructed.

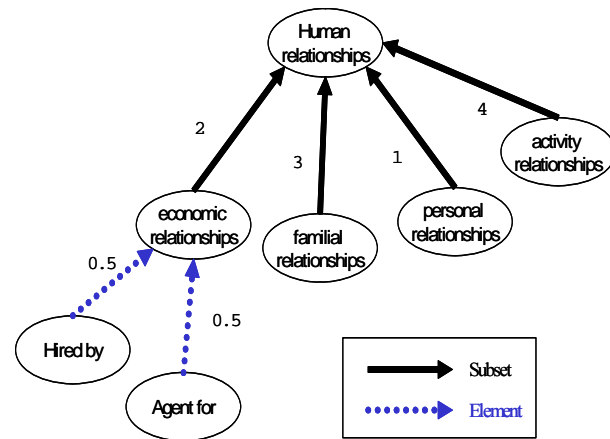


Figure 8: Simple semantic net with weighted connections

4.2 Semantic Distance between Nodes

Figure 8 shows a simple example of a semantic net. This type of supporting information is used to account for the contents of the graph node and edge labels in graph comparisons. Accordingly, our primary interest is in how the information in the semantic net can contribute to the analysis objectives for a scenario.

For single-word labels, the hypernym structure of WordNet is used to calculate distances between individual labels. Some distance measures between labels in WordNet are described in Budanitsky and Hirst (2001). The measure used in this analysis is the total path length distance between the nodes. This measure is equivalent to one described in Budanitsky and Hirst (2001), and unlike other measures he describes, this is a true distance measure in that the triangle inequality holds. Using this measure, the distance in Figure 8 between “Hired by” and “familial relationship” is $0.5 + 2 + 3 = 5.5$. Because the semantic net was implicitly defined by the ontologies rather than actually constructed, the distance was found by comparing the hypernym lists for two concepts to find the most specific word in common, and the distance was taken as the sum of the distances to this first common word.

4.3 Semantic Distance Between Graphs

The ability to measure distance between specific nodes is critical to the ontological analysis, but is not sufficient for measuring distance between graphs. The individual measures can be combined in a variety of ways to create an overall measure between collections of nodes. Some candidates for $d_{label}(A,B)$ are:

$$\text{average } d(a,b), \quad \max_{a \in A, b \in B} d(a,b), \quad \min_{a \in A, b \in B} d(a,b)$$

where $d(a,b)$ is the ontological distance between the labels a and b , and A and B are two collections of nodes, such as a graphs. Note that these measures are similar to set distances used in hierarchical clustering (Everitt 1993). Consequently, we used cluster analysis to indicate the semantic distance between graphs. Recall from the general strategy that a measure of ontological distance between two graphs is the extent to which their respective nodes are distinct and located far apart in semantic space. The farther apart their nodes are on average in semantic space, the greater the disparity in the subject content of the graphs.

Cluster analysis is a method of grouping entities into categories based upon distance. If cluster analysis is used to group the union of the two sets of nodes from two graphs into categories, the following results would be expected. For two graphs that are similar in content, we would expect there to be many nodes in common and the semantic distance between nodes to be close; consequently, individual clusters of nodes would likely have proportional representation from both graphs. On the other hand, for graphs that are very different in content, we would expect the individual clusters to consist primarily of nodes from only one of the two graphs. Thus, the extent to which nodes cluster membership is independent of the graph from which the nodes originated is a measure of ontological distance.

4.4 Measuring Graph Separation Using Cluster Analysis

A Chi Square statistic derived from a cluster analysis was used to measure the extent of separation between graphs. Chi Square provides a measure of the extent of dependency in a contingency table. A large Chi Square value would indicate that nodes with clusters tend to come from the same graph, and the graphs are dissimilar in content. A small Chi Square would indicate a proportional representation from graphs within the clusters and that the graphs were similar in content. The procedure for measuring the semantic distance between graphs using cluster analysis and Chi Square is as follows:

- The semantic distance between all pairs of nodes used in all graphs was calculated. Note that while the content information is primarily based on nodes and does not include link labels, the labels entered into the distance measure in that they influenced the structure of the ontologies upon which distance was based through the choice of the penultimate node that defined the hypernym lists that made up the ontologies.

- A distance matrix was created for each of the 120 pairs of the 16 graphs that contained the pairwise distance between the union of their nodes labels. Each node label was tagged to indicate the graph with which it was associated. If the same label appeared in both graphs, it was listed twice.
- A hierarchical cluster analysis was carried out on each of the 120 distance matrices to obtain their respective dendrograms.
- Each dendrogram was cut into five clusters.
- Contingency tables were constructed of frequencies of labels by graph membership versus cluster membership.
- Chi Square tests were performed on contingency tables to indicate the extent of graph label separation in the clusters.

4.4.1 Cluster Analysis

For each of the 120 pairs of graphs, a distance matrix consisting of the semantic distance between all words in the union of nodes for the two graphs was used to perform a hierarchical cluster analysis. The analysis was carried out using the “hclust” function in R language for statistical computing. “This function performs a hierarchical cluster analysis using a set of dissimilarities for the n objects being clustered. Initially, each object is assigned to its own cluster and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters, continuing until there is just a single cluster. At each stage distances between clusters are recomputed by the Lance-Williams dissimilarity update formula...” (The R Foundation for Statistical Computing 2004).

Figure 9 is a typical dendrogram showing the label clustering for Graphs 6 and 16. The cut line shows the separation for five clusters. The nodes are labeled by graph number suffix so one can visually inspect the extent of clustering of graphs based on labels. However, the R code was used to actually carry this out for the analysis.

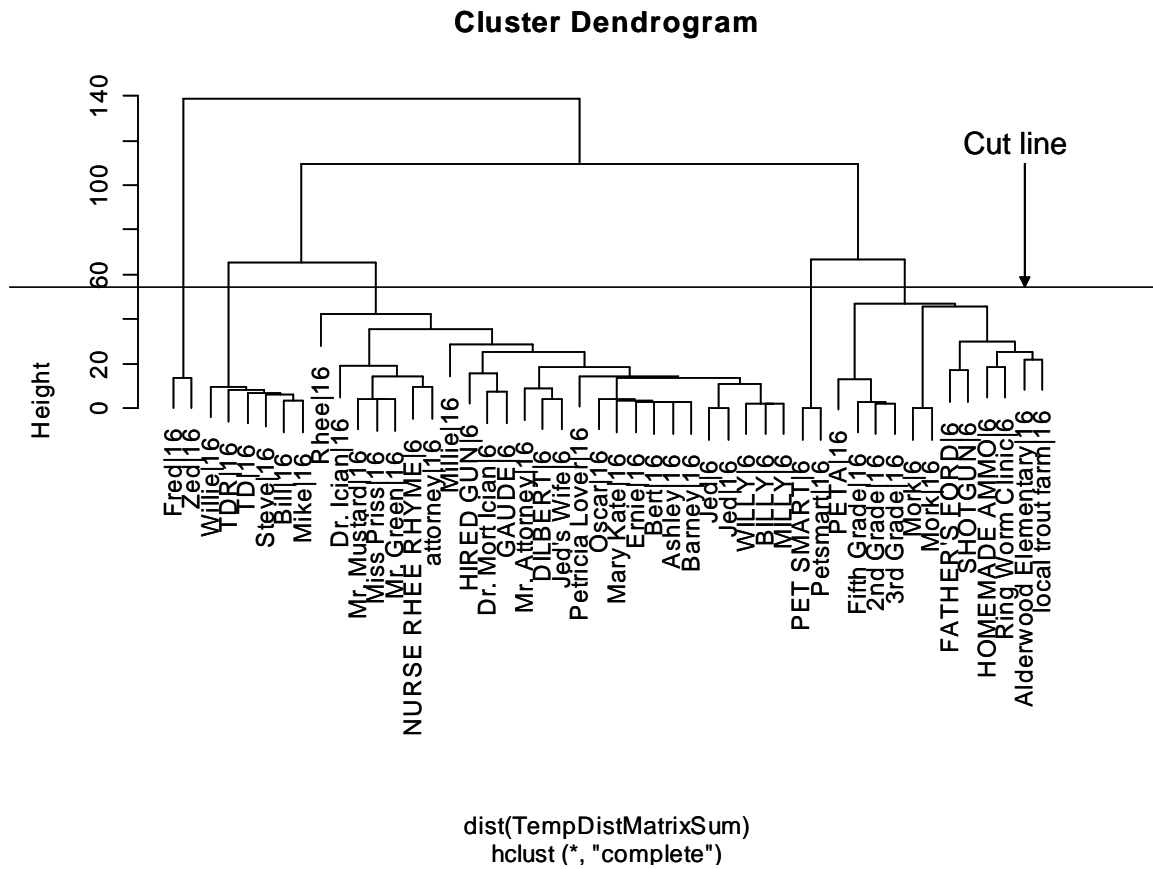


Figure 9. Cluster Dendrogram for a Pair of Graphs

4.4.2 Chi Square Tests

As shown in Figure 9, the cluster analysis was cut at the level of five clusters. The R program was used to construct a contingency table in which the five clusters formed one dimension of the table, and the pair of graphs formed levels of the other dimension. A Chi Square statistic was computed on the contingency table as a measure of the independence of row and column frequencies.

Table 3. Contingency Table for Graph Shown in Figure 7

Cluster	A1.Analyst2	A2.Analyst1
1	10	16
2	5	7
3	1	1
4	0	6
5	0	2



4.5 Results of Ontological Cluster Analysis

Independence of rows and columns in the contingency table would mean that label clusters of the two graphs are proportionally represented in the contingency table. A large value of the Chi Square statistic is an indication that graphs are disproportionately represented in the clusters meaning that their content is distinguishable based on semantic distance. The contingency table corresponding to Figure 9 is shown in Table 3. Its Chi Square statistic is 4.93. The Chi Square statistics for all 120 pairs of graphs are shown in Table 4. The pattern of these numbers, illustrated by a relative brightness matrix in which shading corresponds to distance, is shown in Figure 10. The top panel in the figure shows the pattern that would be expected if ontological distance between the graphs reflected the experimental design as depicted in Figure 2. The bottom panel shows the Table 4 Chi Square scores that were actually found. As seen in the bottom panel, the pattern shows that graphs representing the same scenario were quite similar across analysts. On the other hand, comparing the patches in the lower panel that correspond to the gray and black patches in the upper panel shows that there is little difference between different scenarios of the same class, e.g., Animal 1 versus Animal 2, and different scenarios within the same class, e.g., Animal 1 versus Geek 3, respectively.

The results from this analysis are further illustrated in Figure 11, which shows the distributions of Chi Square values for different types of relationships in the experimental design. The top panel in this figure shows the distribution of Chi Square statistics for pairs of graphs with the same scenario constructed by different analysts. These graphs should be most similar and should have the smallest Chi Square values, which can be seen to be the case by comparing this panel with the two panels below, which represent different scenarios/same analyst in the center panel and different analyst/different scenario in bottom panel.

Table 4. Chi Square Scores for Contingency Tables Derived from Cluster Analysis Using Ontological Distance

	A1 Anal1	A1 Anal2	A1 Anal3	A1 Anal4	A2 Anal1	A2 Anal2	A2 Anal3	A2 Anal4	G3 Anal1	G3 Anal2	G3 Anal3	G3 Anal4	G4 Anal1	G4 Anal2	G4 Anal3	G4 Anal4
A1.Anal1		1.03	4	1	10.38	11.9	11.06	11.31	18.45	16.52	8.62	26.77	15	12.07	11.74	14.03
A1.Anal2			2.28	1.62	4.93	11.8	6.26	11.51	6.7	5.53	9.64	19.1	3.92	7.62	15.72	14.44
A1.Anal3				0.54	9.64	9.6	7.34	8.71	13.79	20.96	10.35	9.13	5.16	4.68	1.63	1.45
A1.Anal4					7.46	14.17	10.86	5.19	24.62	42.24	14.26	24.25	7.7	15.93	8.54	9.94
A2.Anal1						3.24	1.12	0.23	10.2	8.6	3.95	11.53	6.36	3.64	7.72	6.67
A2.Anal2							0.47	5.73	19.34	26.56	10.72	17	9.89	8.09	14.19	12.8
A2.Anal3								2.01	20.68	29.23	14.4	18.23	12.86	9.09	11.33	9.93
A2.Anal4									27.34	16.19	12.87	11.33	10.82	22.82	11.83	12.46
G3.Anal1										1.06	1.44	1.36	16.29	9.46	18.13	20.88
G3.Anal2											2.14	1.03	10.06	10.28	20.43	18.74
G3.Anal3												4.25	3.31	9	12.47	11.37
G3.Anal4													9.93	6.34	16.56	15.67
G4.Anal1														3	2.27	0.67
G4.Anal2															4.07	5.16
G4.Anal3																1.08
G4.Anal4																

 Same Scenario Different Analyst
 Same Analyst Different Scenario

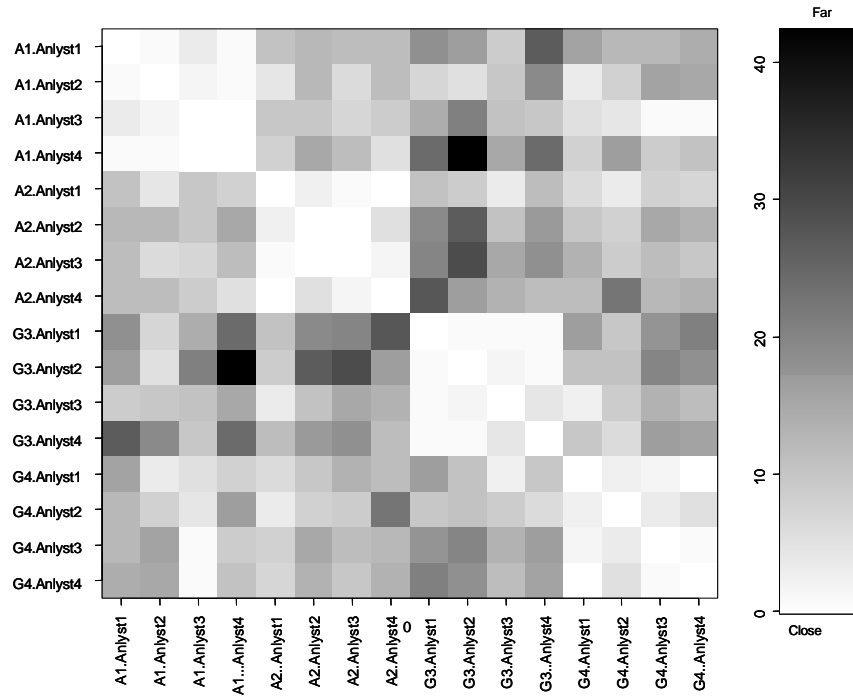
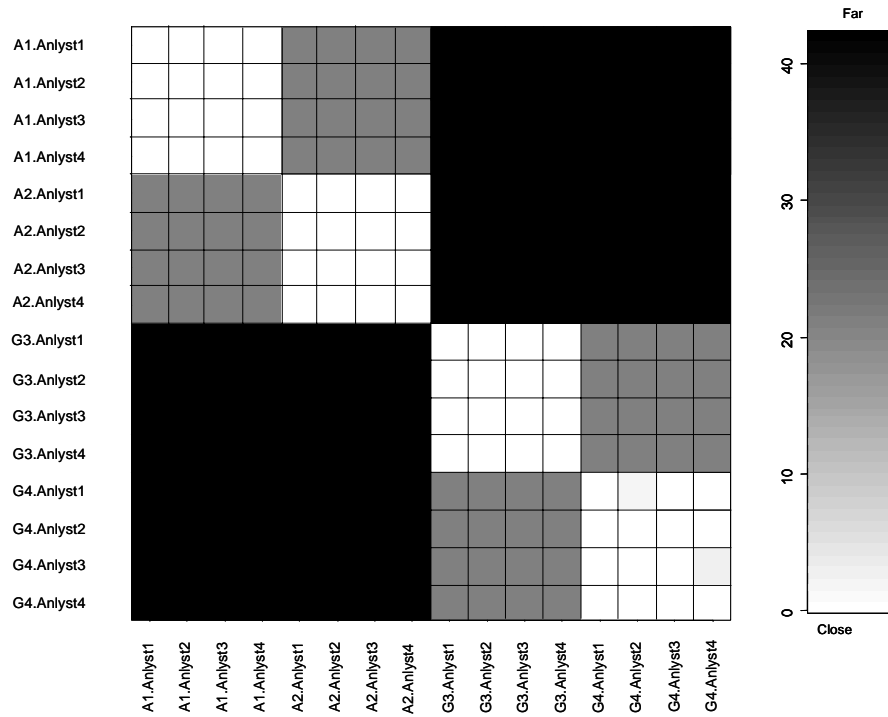


Figure 10. Comparison of Idealized Versus Actual Ontological Distance Displayed as Relative Brightness as Revealed Through Chi Square Scores

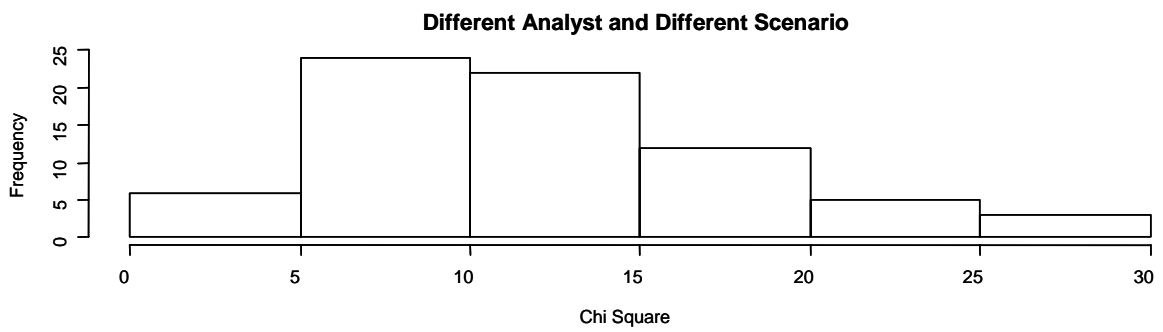
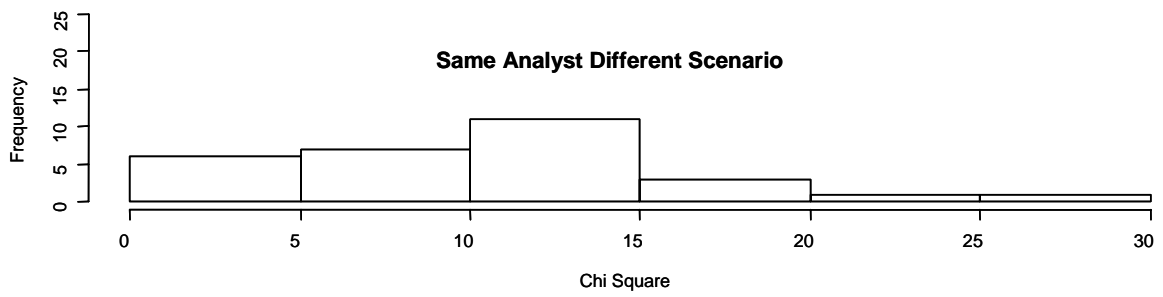
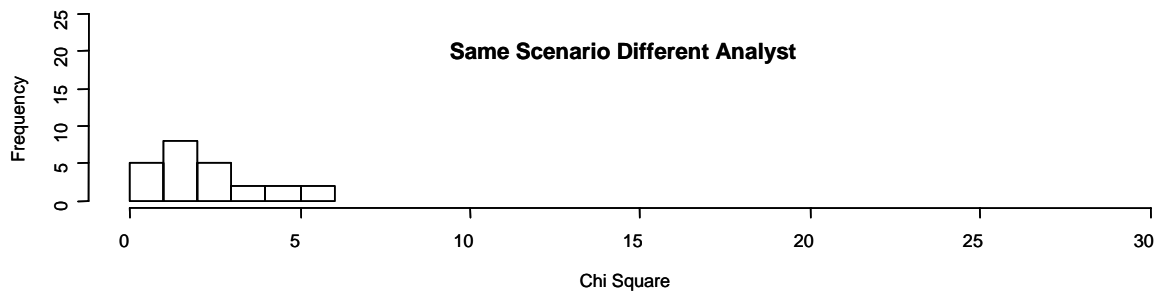


Figure 11. Chi Square Distributions for Categories of Graphs

5.0 Discussion

This section addresses issues that arose during the construction of the ontologies that have implications for reproducibility of the ontologies and strengthen the connection between the ontologies and graphs. This section also contains a summary of the experimental findings and possible directions for further research.

5.1 Issues that Arose in the Construction of the Ontologies

There were a number of issues that arose during the construction of the ontologies that are important to document. For some of these issues, their resolution could help standardize the process for creating the ontologies. For others, if ways were found to address them, there would be an increased isomorphism between the ontologies and the graphs. This could result in finer distinctions between the graphs based on semantic information in the ontologies.

While in principle the construction of the ontologies was meant to be mechanical and reproducible, in practice there were many instances in which subjective judgment was part of algorithmic process.

5.1.1 Richness of the Graphs as Compared to the Ontology

The graphs capture a variety of relationships among the graphical entities. Because a single node can have multiple connections, it will have multiple roles in the graph corresponding to the different links. This leads to multiple ways in which a node can be characterized. In contrast, the ontology was constructed as a conceptual chain of hypernyms, from general to specific, in which the node was given a unique characterization. Consequently, it was necessary to make a choice from among the multiple roles of a graphical entity for the purposes of its unique specification as captured in the ontology. No underlying principle was identified as a basis for this choice, and it is doubtful whether such a principle could be found. A possible solution would have been to have had multiple hypernym lists for each of the labels, in which case the semantic distance discussed below would have to have been somewhat modified.

5.1.2 Variations Across Analysts

The graphs reflect the different analysts' styles in terms of amount of detail and the types of information chosen to be represented. Different analysts often selected from the scenarios a different subset of the potential nodes for inclusion in their graphs. Also, the labels attached to nodes and links tended to vary somewhat among analysts. For example, a link could be labeled "spouse," "wife," or "married." The ontology for a given scenario was constructed to be inclusive all the nodes across all analysts, so the fact that an individual analyst included nodes omitted by others and vice versa was not a problem. However, the use of different labels for nodes and edges representing the same entity or relationship are issues that needed to be addressed in constructing the ontology.

The issue of different labels for the same node was solved by designating a primary name and constructing clones of the hypernym lists for each of the multiple labels in which the penultimate node was the primary label and the semantic distance between it and the end node was zero. This has the effect of creating a unique branch for each name and an equivalence class of names for the underlying entity/concept in the scenario. The choice of a primary name did not make a substantial difference in the ontology because the clone branches create an equivalence class of names, which in effect define the entity.

The issue of different labels for edges was not as cleanly solved. The edge labels were often the basis for how the node would be characterized by the penultimate link in the hypernym branch. Just as there could be multiple links to a node within a given graph, and a choice was needed as to how the best characterize the node, the diversity across graphs simply increased the size of the choice set from which a characterization was to be selected. How the node was characterized depended somewhat on the consistency of edge labels across analysts and also on which role the entity played that was judged to be most central to the scenario. This was somewhat subjective and often required the use of external knowledge. A possible solution would have been to expand the ontology to have duplicate hypernym lists for each label, as needed, as suggested in the previous section.

5.1.3 Use of External Knowledge

In choosing how to characterize a given node, sometimes the best characterization, or at least a viable contender for best characterization, was information never formally captured by the graphs or even explicitly stated in the scenario. For example, when an animal is taken for medical treatment to an individual that is given the title of doctor, it can be reasonably inferred that the individual is a veterinarian rather than a medical doctor. When the individual works at an animal hospital, this could be doubly inferred. Furthermore, the designation of the clinic as an animal hospital was itself inferred from the name, “Ring Worm Clinic,” as was the fact that it was an animal being taken there for treatment. The links for this individual indicate he is also an employee, a lover, and a father. That he is also a veterinarian, while never explicitly stated, may still be the most central role of the individual in the scenario, and hence the best choice as the defining characterization to use in the ontology. So the best characterization for use in the ontology may depend on the defining role that can only be inferred or guessed at based on knowledge that emerges from a complete understanding of the scenario.

5.1.4 Limitations of WordNet

There were some issues based on limitations in WordNet. In particular, WordNet does not necessarily contain the specific words or phrases used in the graphs or have precise meanings of those words. For example, “deceased spouse” is not a searchable term. So an exact adherence to WordNet would require that the branch be defined based on either spouse or deceased. Yet other word pairs in WordNet, such as “animal disease,” are defined. So it would seem consistent to construct a branch from “deceased spouse” in which the penultimate node is either deceased or spouse. Such constructions are in the spirit of WordNet but cannot be made by strictly adhering to WordNet.

5.1.5 Semantic Distance Weights

Currently, the semantic distance between hypernyms was given a default of 1, except when creating equivalence classes of names and zero was used. Casual observation of the branches in the ontology reveals that the distances are not all the same. Consider the following example for Mad Cow Disease: State, Condition, Pathological State, Ill Health, Illness, Disease, Animal Disease, and Mad Cow Disease. Clearly, some transitions represent a larger “jump” in semantic distance than do others. For example, Ill Health and Illness are probably closer in semantic space than Condition and Pathological State, or even Disease and Animal Disease. There is not, at this time, any structured system for determining weights. Some of the findings in Budanitsky and Hirst (2001) would seem to suggest that relative frequencies of terms in some broad lexicon could be useful for this purpose. Possible topics for exploration are to what extent a refinement of the ontologies to reflect these differences would impact the analysis, and what methods are available to assign the distances.

5.2 Conclusions

The metrics we identified for measuring the distance between graphs based upon the structural information in the triads were not able to differentiate among the graphs. We were not exhaustive in our search for measures, and it is still possible that some measure exists that could distinguish among the graphs based on the triad structure. It is also possible that one of the metrics we explored, or some other metric, would differentiate among the graphs for a structural signature other than the one based on triad topology.

The analysis based on ontologies appears to be more useful. The Clustering Analysis/Chi Square metric was able to distinguish graphs drawn for the same scenario by different analysts from graphs based on different scenarios. It was not able to make the finer distinction between types of scenarios. So for example, it was unable to distinguish a greater semantic distance between an animal scenario and a geek scenario than between two animal scenarios or two geek scenarios. Improving the richness of the ontologies, as discussed in the previous subsection, could make it possible better measure the underlying relationships among the graphs.

5.3 Directions for Future Research

A logical next step would be to expand the ontological analysis to explicitly include link information. This richer characterization of the graphs by the ontologies might allow the Cluster Analysis/Chi Square metric to make the finer distinction between types of graphs. Currently, we were unable to distinguish two different graphs of the same type (for example, animal) from two graphs of different types. A richer ontology might make this possible.

Another possibility for making finer distinctions is to vary the number of clusters used in the analysis. We chose to cut the cluster hierarchy at five clusters, which is a reasonable number, and we were able to get positive results. Worth exploring is whether we could

have a more sensitive metric with a larger number of clusters. It would also be worth analyzing the effect of varying the number of clusters to see how stable the metric is using either more clusters or fewer clusters.

6.0 References

- Budanitsky, A and G Hirst. 2001. *Semantic Distance in WordNet: An experimental, application-oriented evaluation of five measures*. North American Chapter of the Association for Computational Linguistics, Pittsburgh, Pennsylvania.
<http://citeseer.nj.nec.com/budanitsky01semantic.html>.
- Everitt, BS. 2005. Webpage accessed 2005, <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>.
- i2 Ltd (Analyst's Notebook). 2005. Webpage accessed 2005, http://www.i2.co.uk/Products/Analysts_Notebook/default.asp.
- R Foundation for Statistical Computing. 2004. R : Copyright 2004, Version 1.9.1 (2004-06-21), ISBN 3-900051-00-3.
- Seber, GAF.1984. *Multivariate Observations*, John Wiley & Sons. New York.
- Schmidt, G and T Ströhlein. 1993. *Relations and Graphs*. Springer-Verlag, London.
- Wasserman, S and K Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge England.
- Whitney, P, G Chin, O Kuchar, M Powers, KE Johnson, and M Sloughter. 2004. *A Data Signature Approach for Analyzing, Manipulating and Understanding Collections of Graphical Summaries of Scenarios*. PNNL-SA-42047, Pacific Northwest National Laboratory, Richland, WA.
- WordNet 2005. Webpage accessed 2005, <http://wordnet.princeton.edu/w3wn.html>.

Appendix A

Scenarios

Four one-page scenarios were generated for use in this experiment. Two of the scenarios were animal/pet based and were generally different, but that contained overlaps in storyline and characters. These were the Animal Scenarios. The two other scenarios revolved around homeless individuals moving between two locations. These two scenarios were also generally different, but with overlap in characters, many of whom had computer skills. These were the Geeks Scenarios.

Animal 1 Scenario

Let me tell you the story of a man named Jed, he was so poor he barely kept his family (wife and 2 kids) fed. And then one day when he was shooting up some food (with homemade ammo), he heard the eerie sound of a hound in a bad mood. Pain that is. And lots of it. Well Jed walked ahead and here is what he said, “you poor damn dog I didn’t mean to shoot you in the head! I’ll call you Mork! And I’ll take you to the vet and try to save you out of guilt, but if you up and die I’m going to eat you as food kilt!”

So Jed scoops up the dog and put him in his father’s ford, and away Jed drove, as the mighty engine roared. They went to the vet, the cheapest one in town, by the name of Ring Worm Clinic it could make a doggy frown. But the vet at the clinic (Dr. Mort Ician) was his wife’s father’s only son, and Jed was fairly sure he wouldn’t ask about the gun!

When the vet looked at the hound he thought it was quite a sorry sight; A bullet in the head, but still putting up a fight. So he called in Nurse Rhee Rhyme and asked her for a dime and exclaimed “I don’t want its brain to drain and we can stop it just in time!” So into surgery they went, the Doctor’s finger in the hole, the nurse with the dime, as she figured out the goal. That dime they did done solder onto the skull of that damn dog, the nurse held its head as the doctor gleaned instructions from an online blog. The surgery completed, they moved him to the recovery room, joining company with three dogs (Chuckie, Pluckie, and Clifford) and a balding baby baboon, and then the doctor and the nurse squeezed in a quickie before noon.

Well the nurse went home and fed her children (Billy, Willy, and Milly) some food fast, born from a marriage she was glad was in the past. Her ex-husband, Gaude, had sired children well, then bought each child a cat (Clawed, Flawed, and Felix, respectively) and said he would see them all in hell. Now Gaude was a cop, from the station to the west, and when it came to the kids, he always thought that he knew best. He thought their step dad Dilbert needed a bullet in the chest, so he paid a hired gun to figure out the rest!

This gun, he worked at PetSmart, to feed his wife and kids (twins). With the money insufficient, his connections greased the skids. His boss he was a Christian, almost generous like a fairy, told this desperate father to feed his family all the dog food he

could carry. It was this desperation that turned this man astray; he shot the Nurse's husband, and dumped his body in the bay. Now the seagulls drew attention to his rotting fleshy stink, the Medical examiner's autopsy revealed, that the bullet was unique. And time passed...

While reading stories to her children, Nurse Rhee Rhyme thought of her dead spouse. She thought of the unique bullet and then the man who was a louse. She knew she'd seen another, and she thought that she knew where, she dropped her kids off with her Brother Jed's wife before she got too scared. Flawed (the cat) was chosen for her ruse as she returned to work, asked Dr. Mort Ician to declaw him, and said she would check on Mork. The recovery room was dark and hot, her nausea successfully fought. Her eyes were on the trash can, confirming her suspicions weren't for not (double negative). For in there lay a bullet, its appearance causing fear, to the unique round that was found that had killed her husband dear.

Animal 2 Scenario

Jed and his sister Rhee have had a fairly close relationship all of their lives, and once they became adults, elected to live and raise their families in the same town, Alderwood, so that their respective children could experience the familial bonds they treasured in their childhood memories. Even tragedy (Jed was almost wrongly convicted for the murder of Rhee's second husband, Dilbert, by her first husband and father of her children, Gaude) only strengthened the family tie. Jed's kids, Tweedle Dum and Tweedle Dummer, and Rhee's kids, Bill, Willie, and Millie, were all close in age, and were the best of friends.

And now that his sister Rhee had a new man in her life, Dr. Ician, things were looking up. Their children all attended Alderwood Elementary, and though in different grades, maintained close contact not only through family ties, but school activities. Tweedle Dummer was a 5th grader, along with Willie, and they both adored their teacher, Mr. Mustard. Millie was in 2nd grade, along with Dr. Ician's son, Steve, and they also adored their teacher, Miss Priss. However, neither Steve nor Millie really cared for their classmate, Mike, because Mike's Mom, Petricia Lover, worked at PetSmart with Jed, and Jed says she is mean. Both Bill and Tweedle Dum are in Mr. Green's 3rd grade classroom.

All of the children have been very much looking forward to a planned field trip to the local PetSmart for the past few weeks, a trip arranged by Petricia Lover, Mike's mom. However, a requirement to go on the field trip was that each child had to be a responsible pet owner, as part of PetSmart's national campaign promoting responsibility in children through pet ownership. Now, Tweedle Dum and Tweedle Dummer, had twin hamsters, Zed and Fred, that they planned to take to PetSmart for the field trip. And Mike had a dog named Mork that his dad had given him that he could take to on the field trip, now that Mork was fully recovered from his life threatening injuries. (Side Note: Steve's mom Matilda and dad divorced because of this dog... her being more of a cat lover and all).

Bill, Willie, and Millie had no pets, since their father took away their gerbils (Penny, Polly, and Peter) and used them for target practice. Petricia, when she heard of this, was appalled that the children would allow this to happen to their pets (how irresponsible) and banned them from the field trip. However, Mr. Green, Mr. Mustard, and Miss Priss all thought this was highly unfair, given the circumstances under which the gerbils died, and arranged an alternative field trip for children without pets, to go fishing at the local trout farm. Petricia, a PETA member, was extremely enraged when she learned of the planned field trip to murder fish, and contacted an attorney to see if it could be halted. She also called Mr. Green and expressed her frustrations by making threats, which resulted in Mr. Green obtaining a restraining order against Petricia.

Other students known to be pet free, and planning on attending the alternative fishing field trip, include Oscar, Bert, Ernie, Barney, Mary Kate and Ashley, but details regarding which classroom they are members of is unknown. What is known, however, is that Mr. Green successfully got Petricia's lawyer removed from the case under conflict of interest rules, since Petricia's Lawyer's husband owns the local trout farm.

Geeks3 Scenario

For some reason, the East Bay Homeless shelter appeared to be the most popular in town. Located next to the San Jose public library, East Bay has become the home of homeless "geeks" who were victims of the "dot bomb" economy. Many of these geeks appeared to never leave the city block which housed the library and the East Bay shelter, since this block also contained a Soup Kitchen which provided 2 meals per day to the needy, as well as providing 12 beds for individuals willing to assist with Soup Kitchen chores. These beds had been filled by the same individuals for the past 2 months (Bob, Allen, Dennis, Miller, Don, Bradley, George, Paul, Olivia, Kay, Marta, and Stewart), until recently when "bed turn over" was occurring more frequently. Now, all but Mary, Katie, and Olga had departed, according to the rumor mill, for some hick town in Washington State. Olivia was particularly distraught by Stewart's departure, since he was leaving her (his girl friend) and the baby boy they were expecting behind.

Olivia told the staff manager (Cliff) at the Soup Kitchen that some unknown rich dude from the Hick Town in Washington was promising jobs, opportunities, and bus tickets, for geeks willing to relocate. Because Cliff was a nice guy, he agreed to go to Lamaze classes with Olivia at the local community center. After 9 of the residents moved to BFE, the new residents who moved into the Soup Kitchen all came from the East Bay Shelter. These included John, Scott, Curtis, Doug, Jim, Bill, Ryan, Jason, and Justin. When these folks moved out of the East Bay Shelter, Rolf, Glen, Rick, Andrew, Alex, Sarah, Emma, and Abby, and Glenda moved in.

During the day, many of the shelter residents spent their time at the public library, researching job opportunities in Hick Village, Washington. The free Internet connections from public terminals were nice, but the geeks could also connect over

wireless to the library network, and “surf” from the street. Rolf, Glen, and Rick were big wireless guys, having kept their laptops even while homeless.

The network security guru at the library, Puria, (who also worked part time at the unemployment office) noticed that Rick was engaging in some odd traffic, and when he looked a little further, found Rick’s research into Mike Bubba, a politician local to Hick Town, and a lot of information on bus schedules. Because Mike is a snoop, he dug a little further, including reading some of Rick’s emails stored on a local server. These documents revealed that Mike Bubba was funding homeless geeks to relocate to Hick Town, Washington, and that Rick was serving as a local recruiter for the relocation effort. Mike would wire money to Rick, and Rick would arrange to meet homeless geeks at the local Bus Station, and pay cash for one way tickets for the geeks to Hick Town. Of interest, Rick’s girlfriend, Matilda, worked the ticket counter at the bus station, and used her connection to ensure that Rick got first priority for open seats to Hick Town. According to the emails, Sarah, Alex, and Andrew would be departing from the bus station within the next few weeks to go to Hick Town. Additionally, as soon as Olivia has her baby, she will relocate as well to be with her boy friend.

Also, as attachments to the emails, Puria found numerous voter registration forms for geeks who had previously departed for Hick Town, with registration dates from when they still lived in the Bay Area. This confused Puria. He knew specifically that Justin, Jason, and Jim couldn’t have lived in Hick Town when the forms said.

Geeks 4 Scenario

Mike Bubba, a low level assistant to the Mayor of Alderwood, Bert Renoldswrap, has been tasked to oversee the migration of technically proficient homeless individuals from California to Alderwood. Alderwood has suffered financially from the decline in local tourism that resulted from the Mad Cow outbreak at Rick’s Rodeo and Rendering Plant (RRRP) (2 miles outside of Alderwood), and was looking for new skills to boost the local economy. While RRRP had been the crown jewel in the cattle rancher tourism industry, discovery of mad cow also damaged several other local businesses, including the Cowboy Cutlery Beef and Pork Rendering Plant (CCBPRP), the Davis Dude Ranch (DDR), Selma’s Bovine Burger Bar (SBBB) (a favorite local water hole), and the Alderwood Mostly-Milk Cheese Factory (AMMCF). Mike Bubba was really upset by the damage done to the local economy because of the mad cow case, especially because his sister Selma, owner of SBB, had lost a lot of money because of it, meaning she could no longer afford to pay for his daughter Casey’s tuition at Alderwood State College.

So when the mayor approached Mike about how to turn the town’s economy around, he was fully supportive of any plan that could get his darling child back into school. The mayor provided Mike with sufficient funding to buy an undisclosed number of one way bus tickets from California to Alderwood, to attract intelligent programmers and networkers to the Alderwood region. So far Rick has recruited dozens of individuals, including Paul, Allen, Bob, and George, all of whom have taken positions in the ailing

cattle/tourism industry. Paul is working with RRRP, Allen at SBBB, Bob at AMMCF, and George at the CCBPRP. It is unclear at this point, however, what specifically their job duties entail.

George was immediately glad he had decided to move to Alderwood, because he promptly fell head over heels for a local girl, Casey, and they started dating, however Casey's father, the Mayor, was not impressed by this turn of events. He kept trying to set up Casey with Dan, the son of Pete, owner of RRRP. Little did their fathers know, however, that Dan and Casey dated in high school, and didn't much like each other now. Casey had a plan to thwart this however, by trying to fix up her girl friend Shawna with Dan instead.

The first thing all the new folks would do when they got to town was register to vote. This process was made simple by Mike, whose brother Greg ran the local voter registration office. Greg, who likes surfing the Internet and wishes there was a local ISP in town, is generally very enthused about getting such smart people registered to vote, even if a little earlier than the law generally likes. He carries this out with the help of his wife and a county file clerk working out of the County Registration office, Betty. Conveniently, the Voter Registration office, the County Registration office, and the Alderwood Bus Station are all located side by side.

Analyst Generated Graphs

Four analysts generated graphs for each of the four scenarios creating a total of 16 graphs. Analyst's Notebook (AN) was used to visually depict their interpretations of the scenario. Analysts were limited to the icons within AN; however, there are hundreds of icons available within AN, and no guidance was given to limit the selection. Also, the analysts were prohibited from discussing the written scenarios or their interpretations of them until all the graphs were completed. Given freedom in the choice of icons and the lack of communication, the four analysts used a wide variety of icons to represent what they viewed as the important elements in each scenario. While each individual analyst icon-selections were generally consistent across their individual four graphs, no two analysts used entirely similar conventions. The graphs reflect the different analysts' styles in terms of amount of detail and the types of information chosen to be represented. The graphs are ordered by scenario type and then analyst.

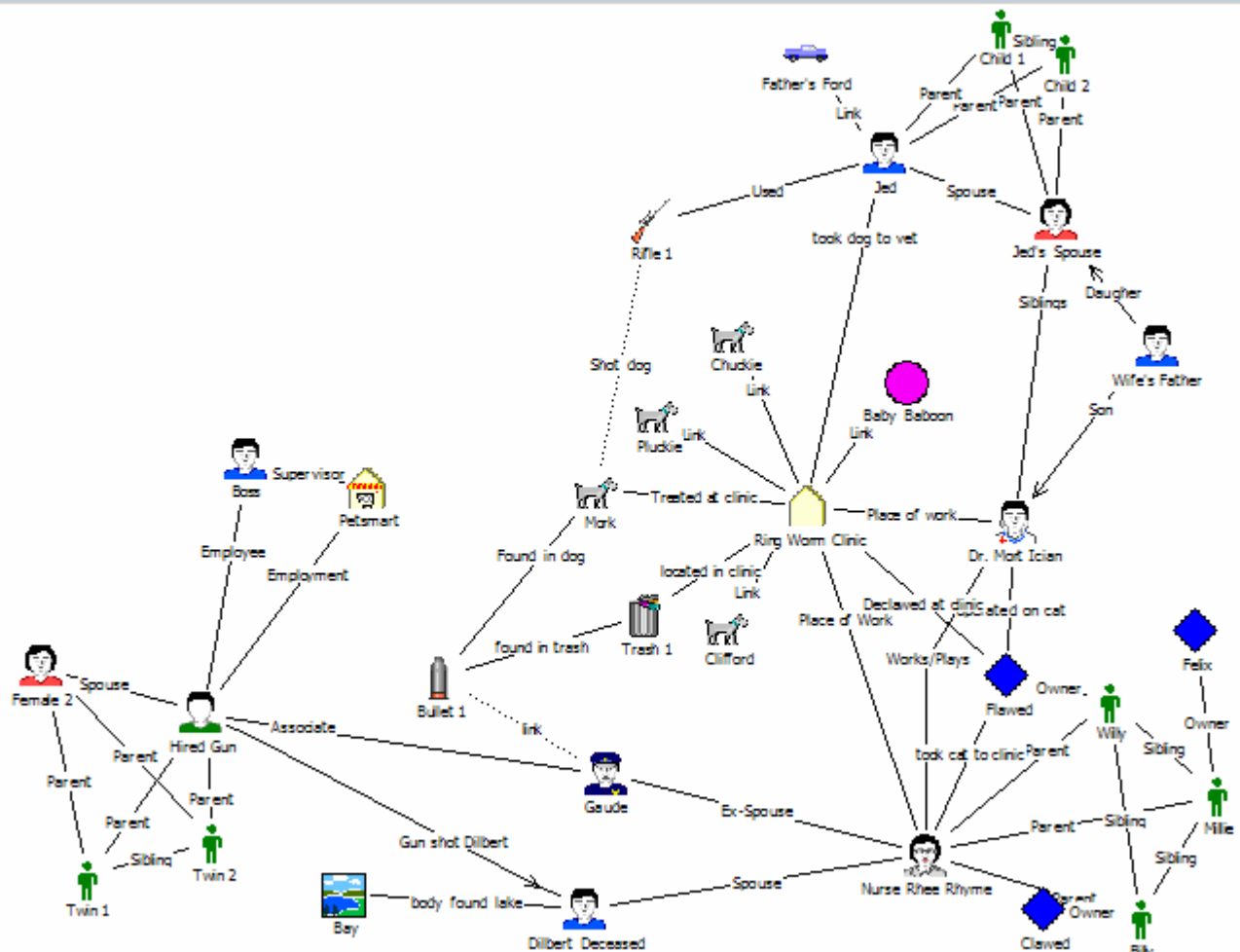


Figure B.1. Analyst 1's Graph of Animal 1 Scenario

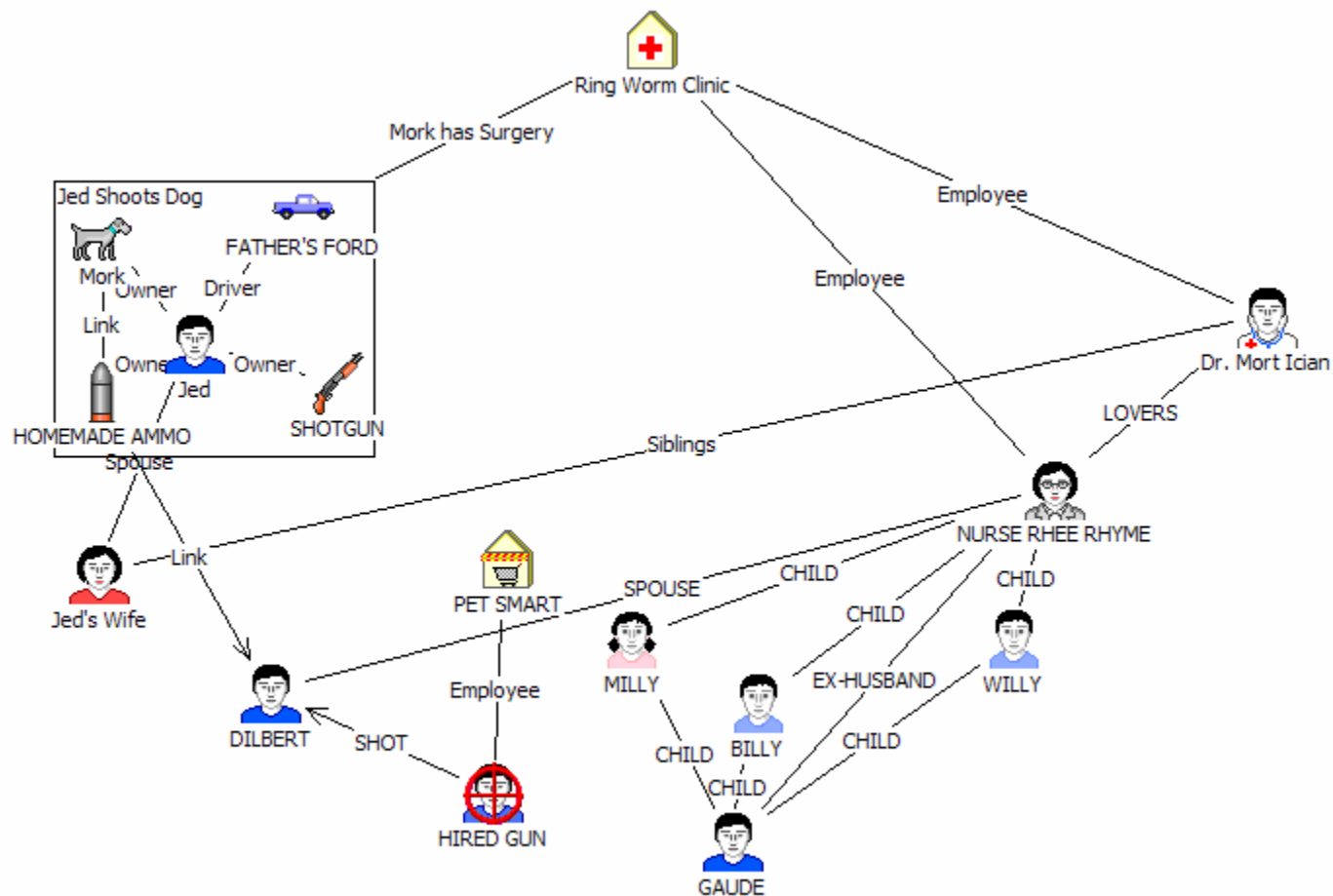


Figure B.2. Analyst 2's Graph of Animal 1 Scenario

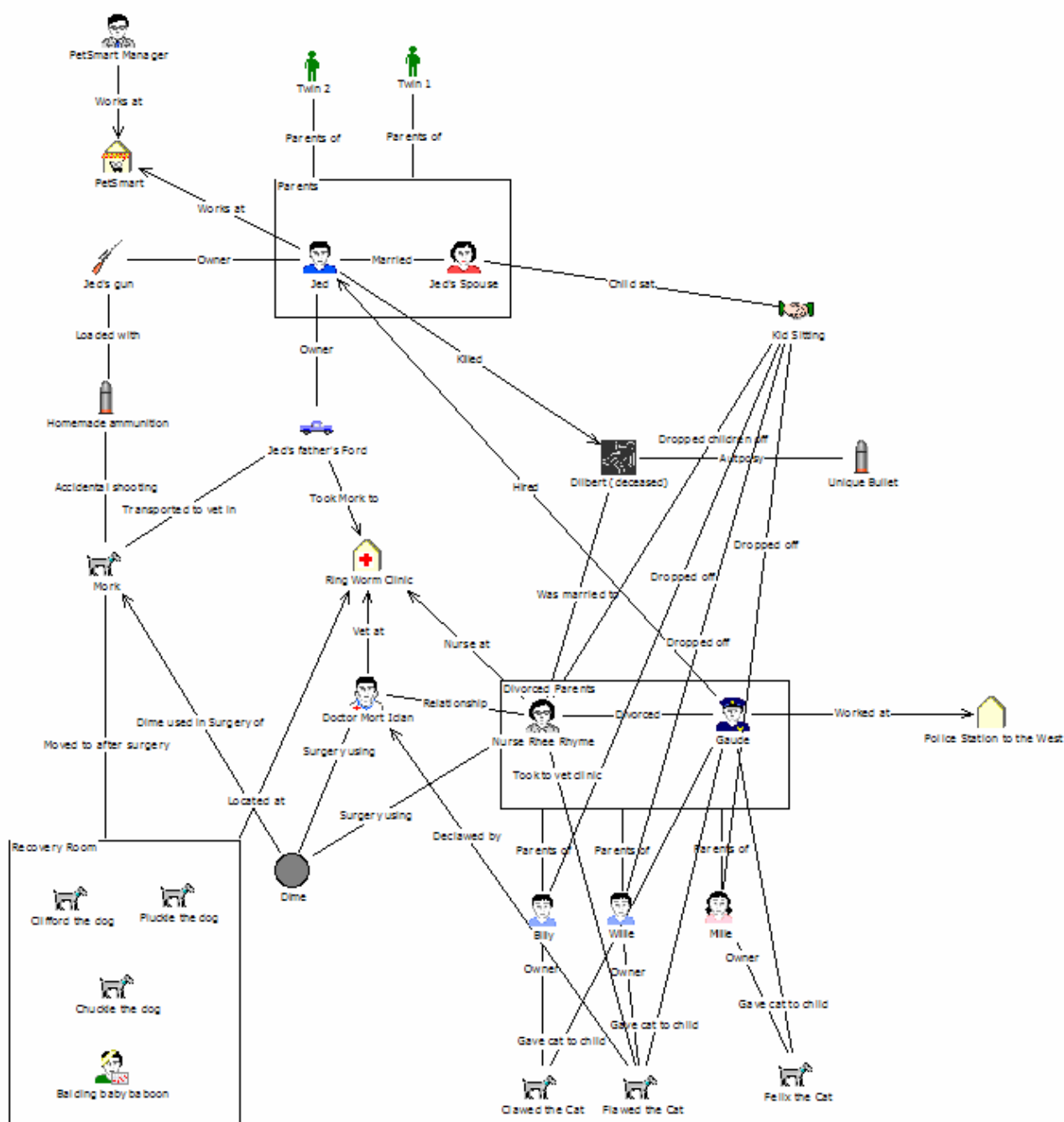


Figure B.3. Analyst 3's Graph of Animal 1 Scenario



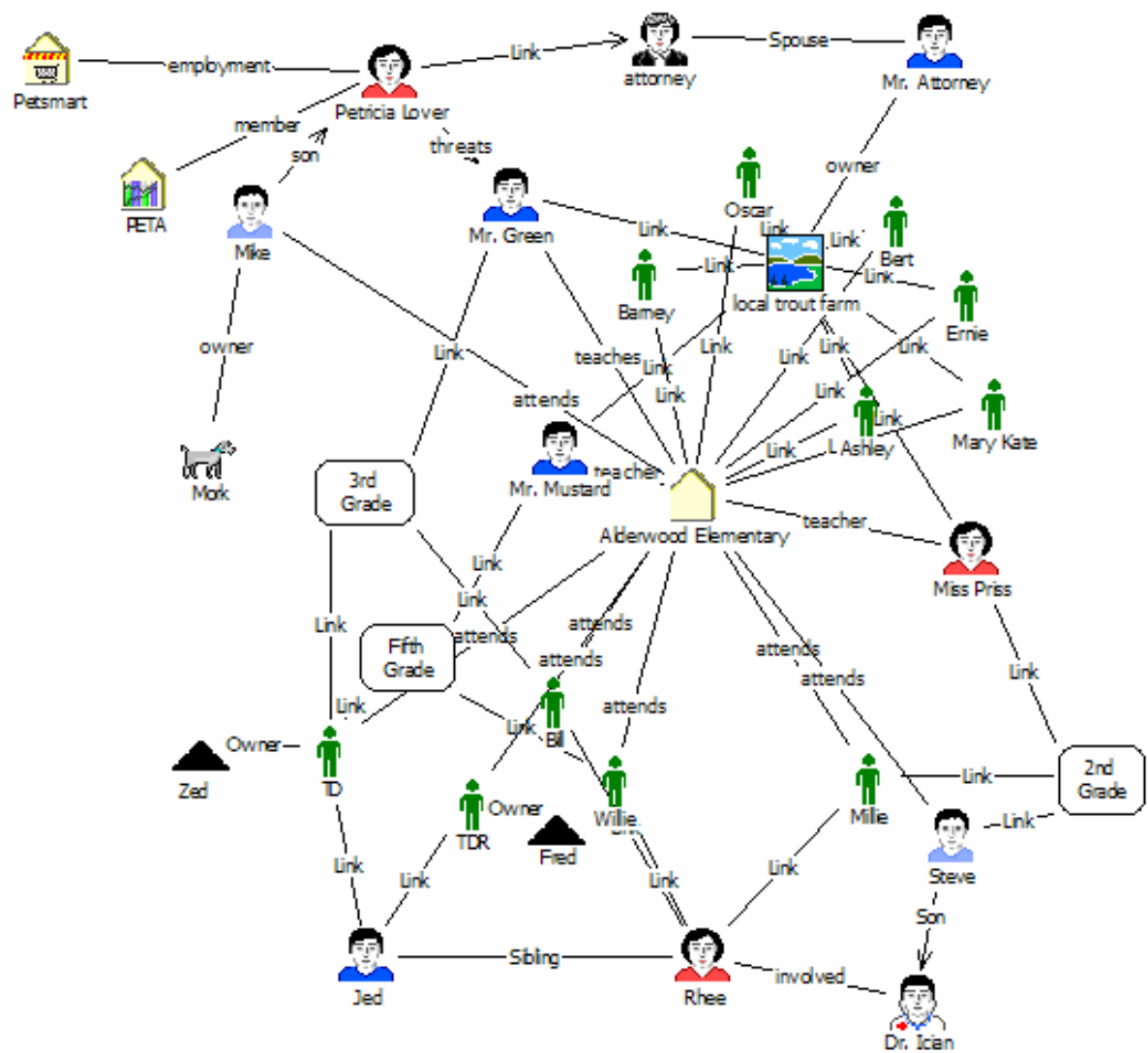


Figure B.5. Analyst 1's Graph of Animal 2 Scenario

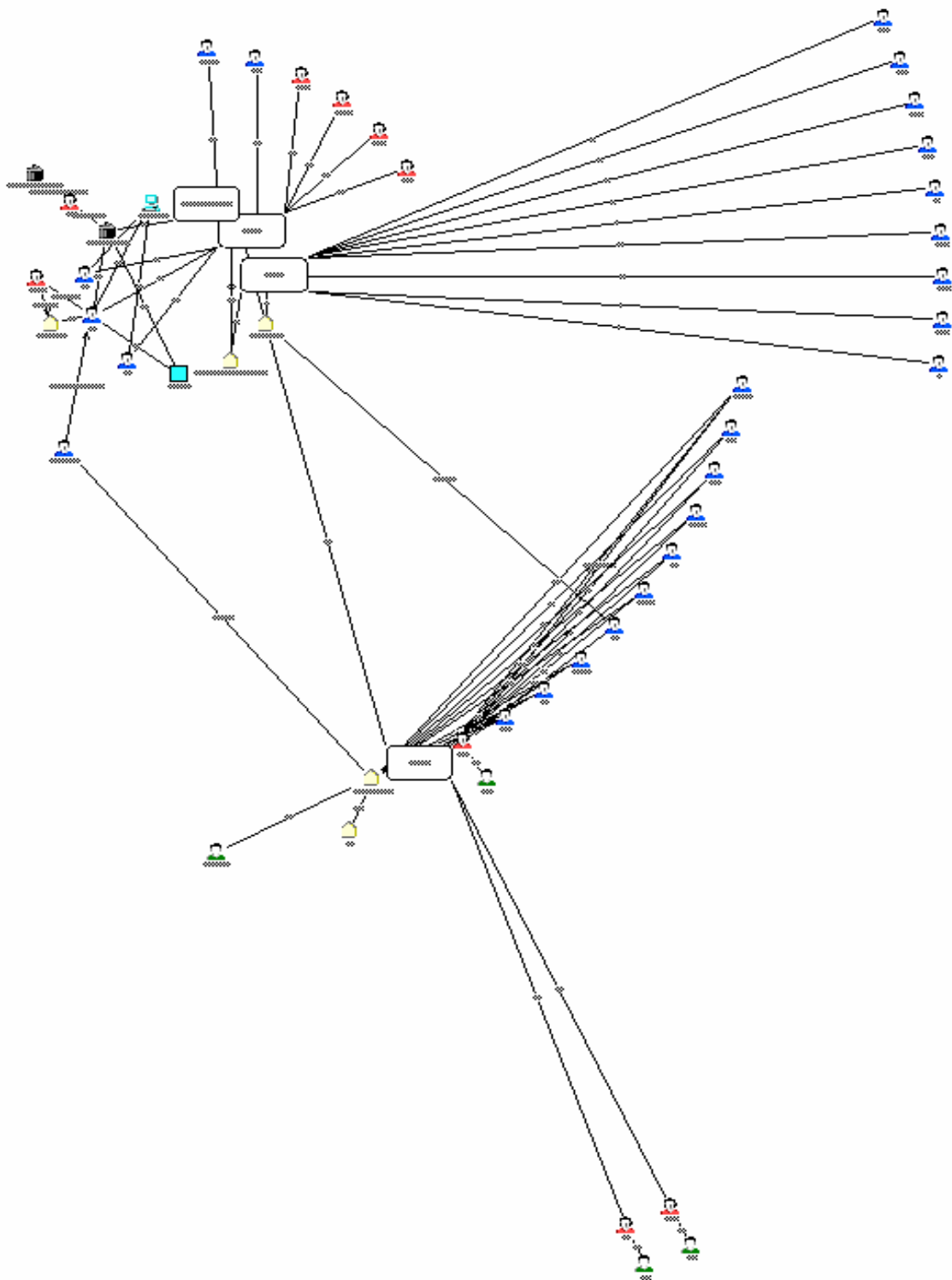


Figure B.9. Analyst 1's Graph of Geeks 3 Scenario

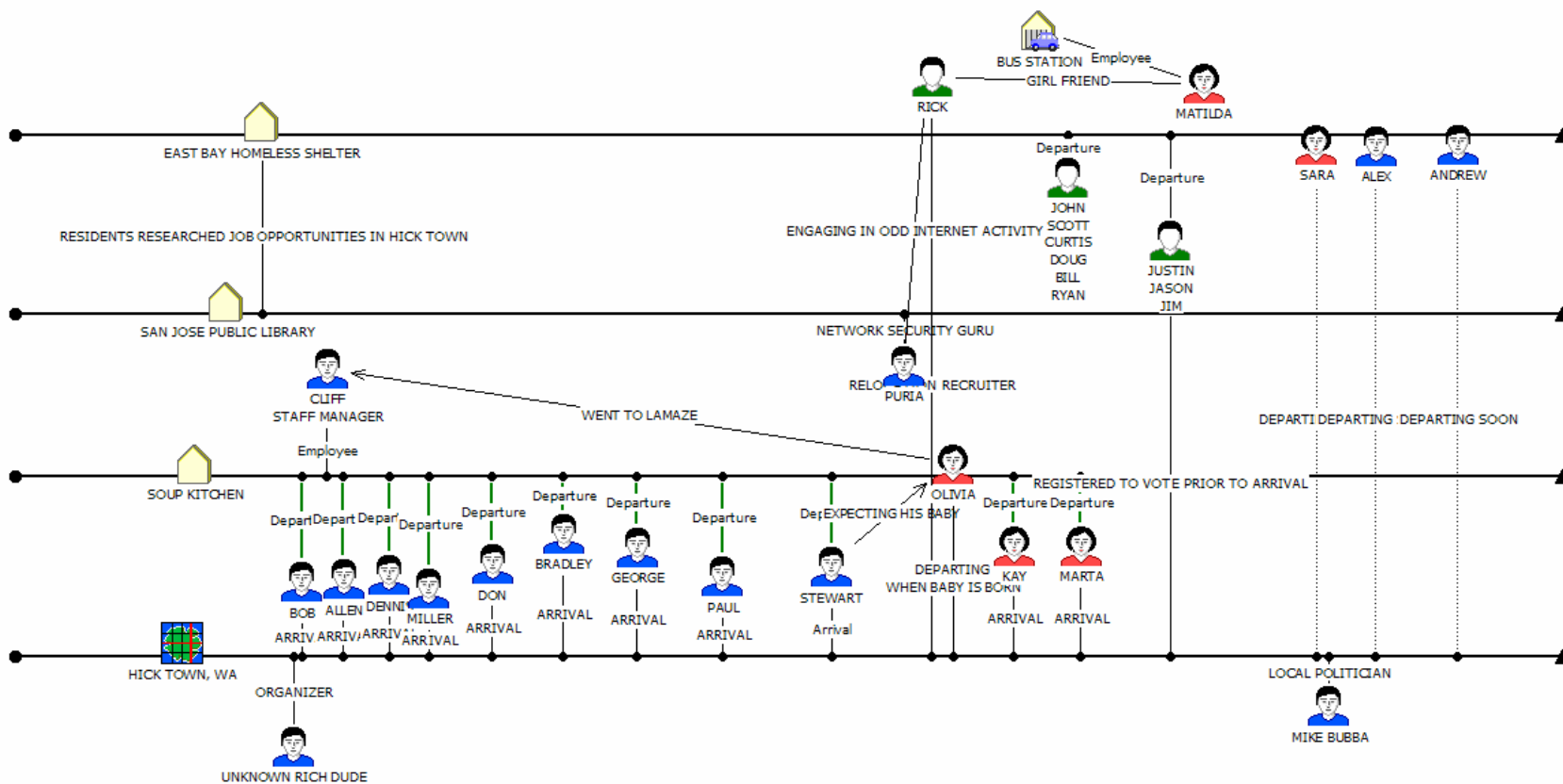


Figure B.10. Analyst 2's Graph of Geeks3 Scenario

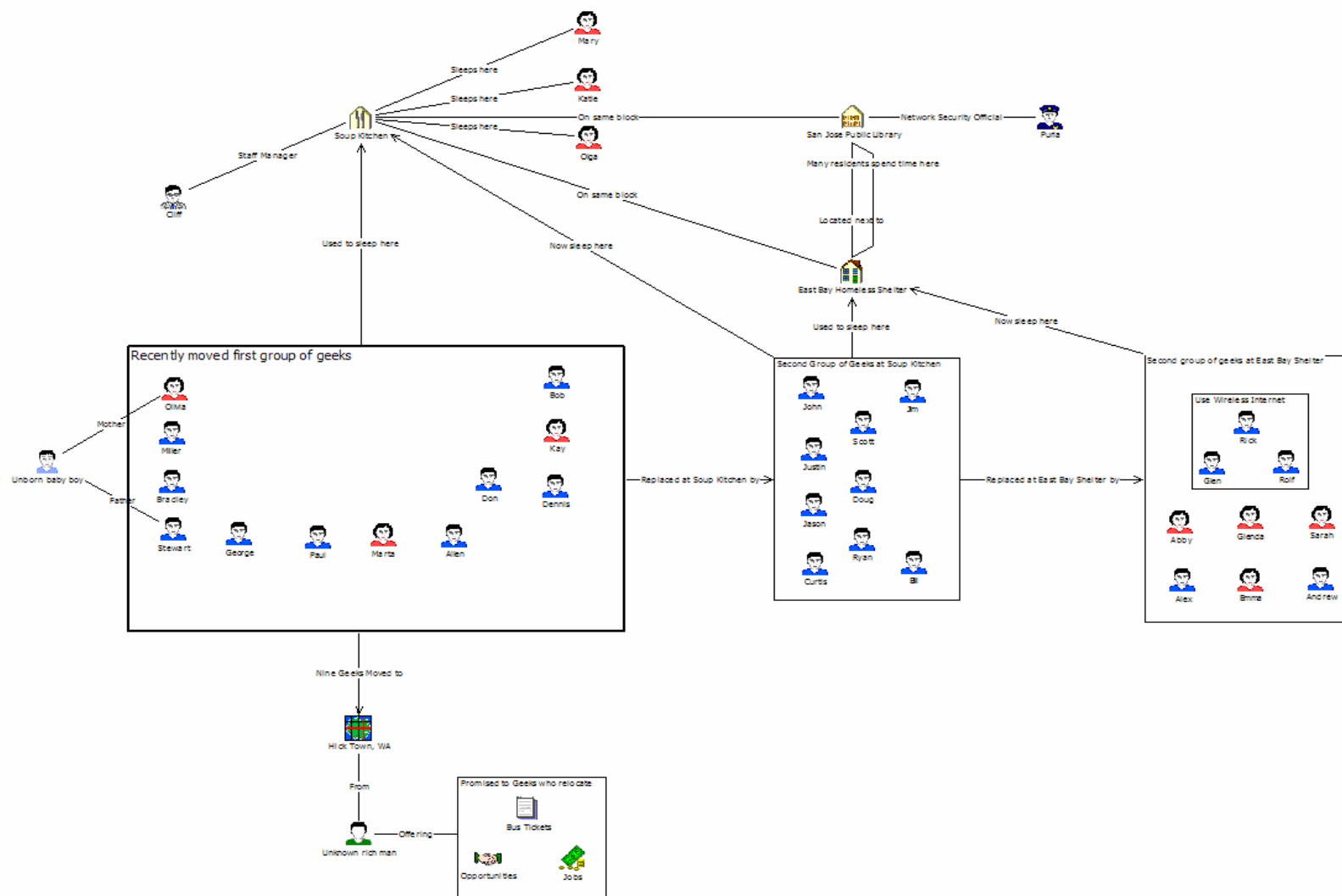


Figure B.11. Analyst 3's Graph of Geeks 3 Scenario



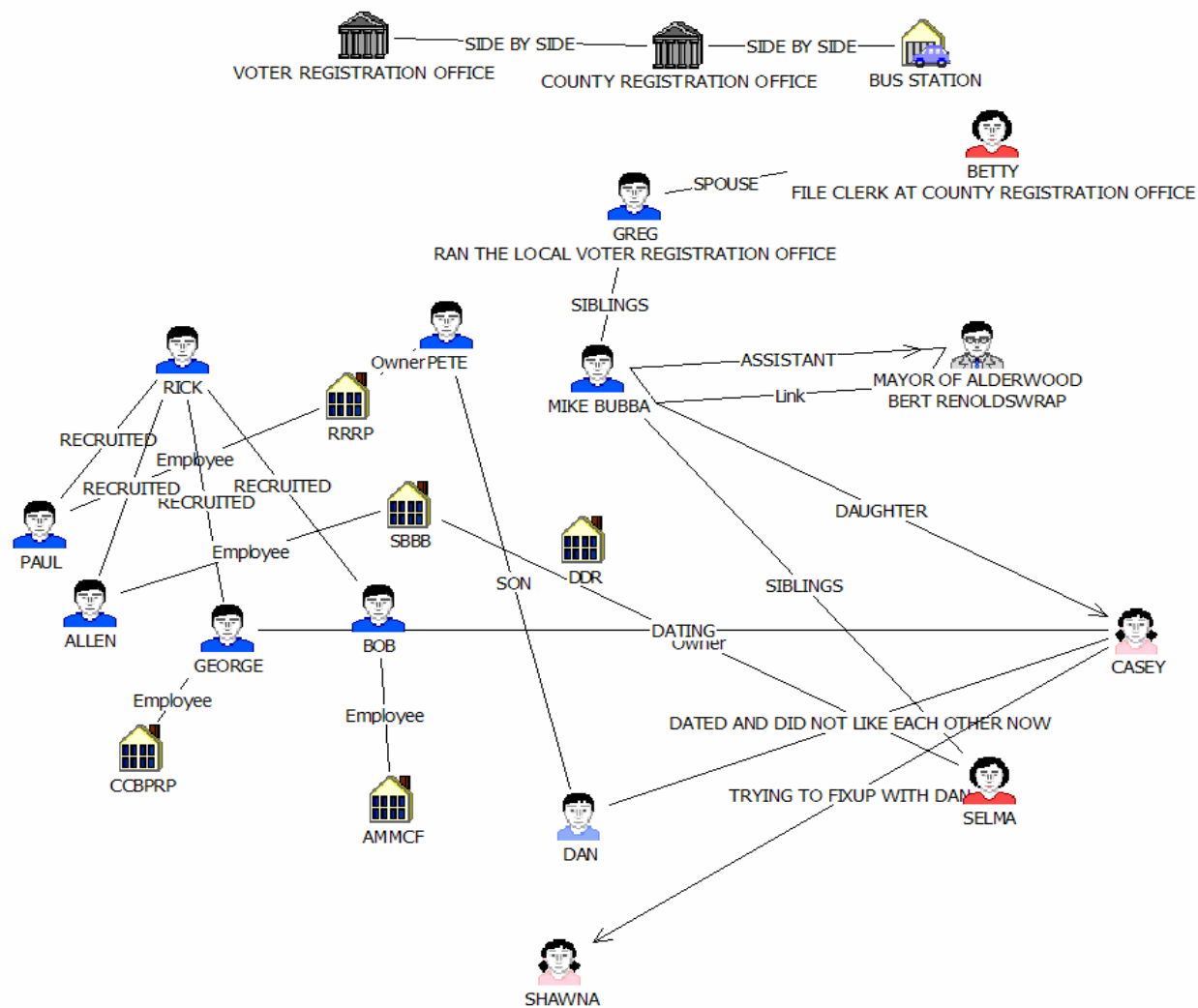


Figure B.14. Analyst 2's Graph of Geeks 4 Scenario

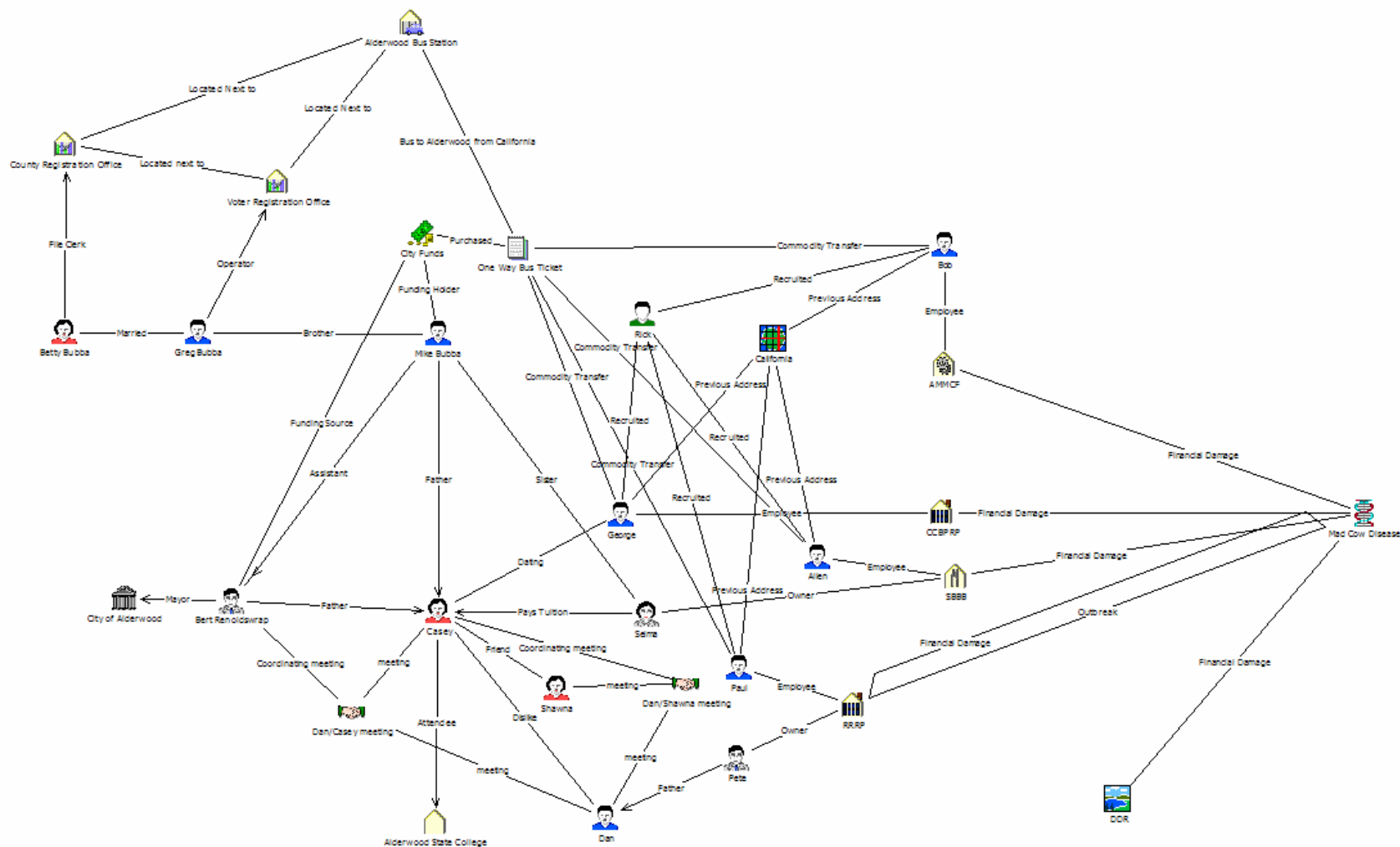


Figure B.15. Analyst 3's Graph of Geeks 4 Scenario

Appendix C

Profiles of Triad Counts

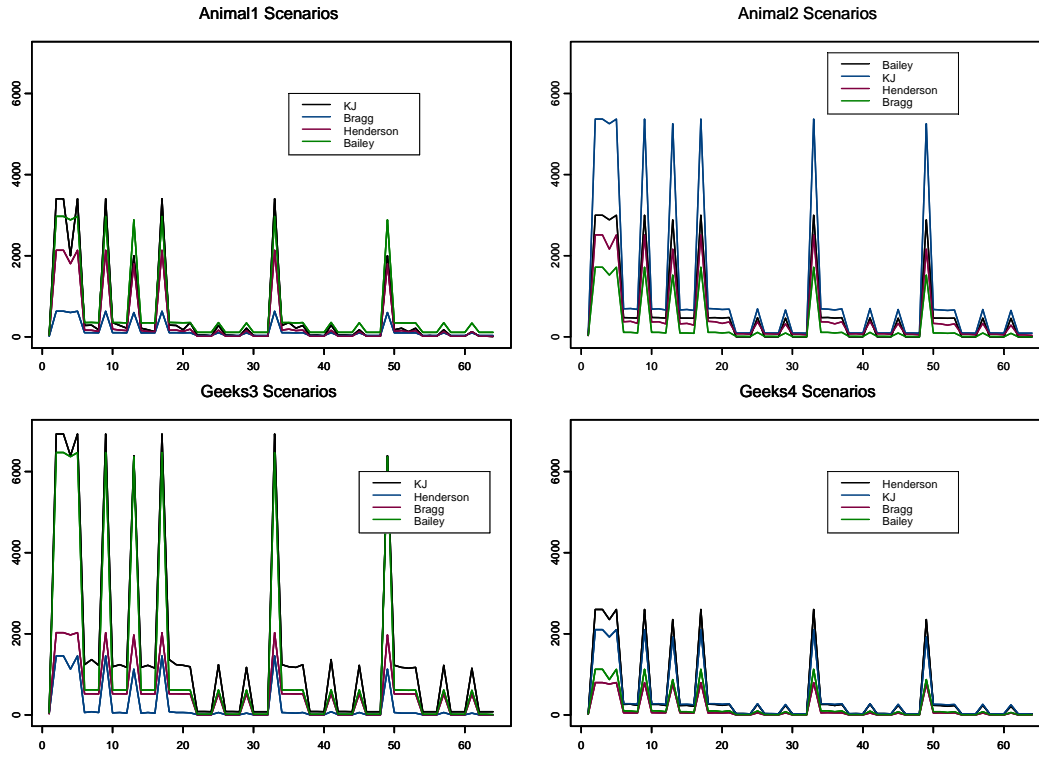


Figure C.1. Plots of Raw Data for Each Scenario Type

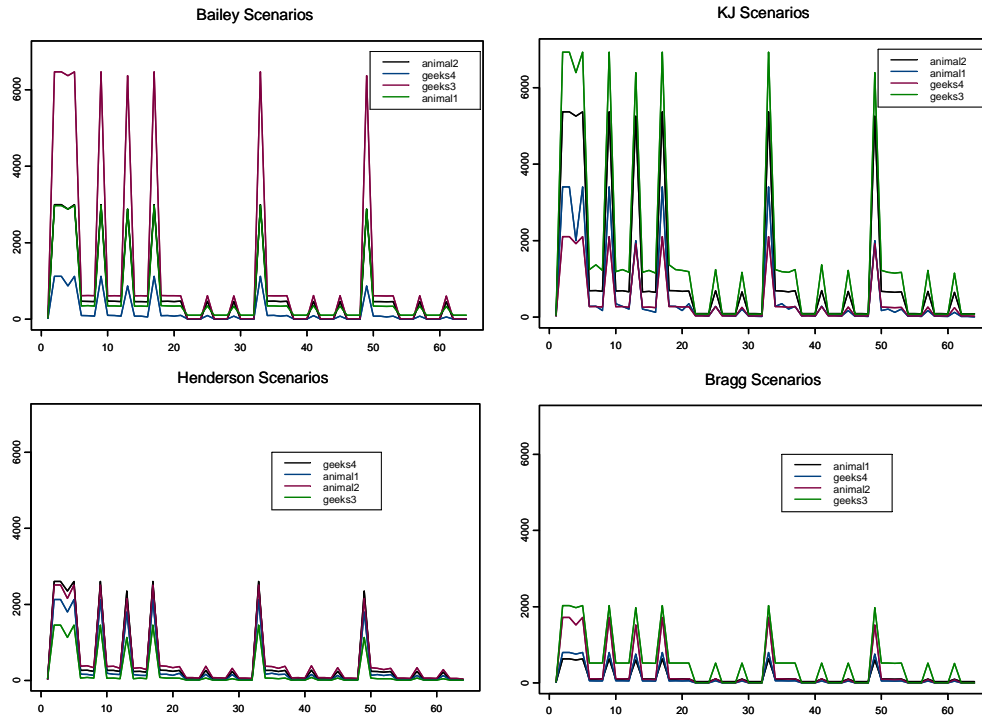


Figure C.2. Plots of Raw Data for Each Analyst

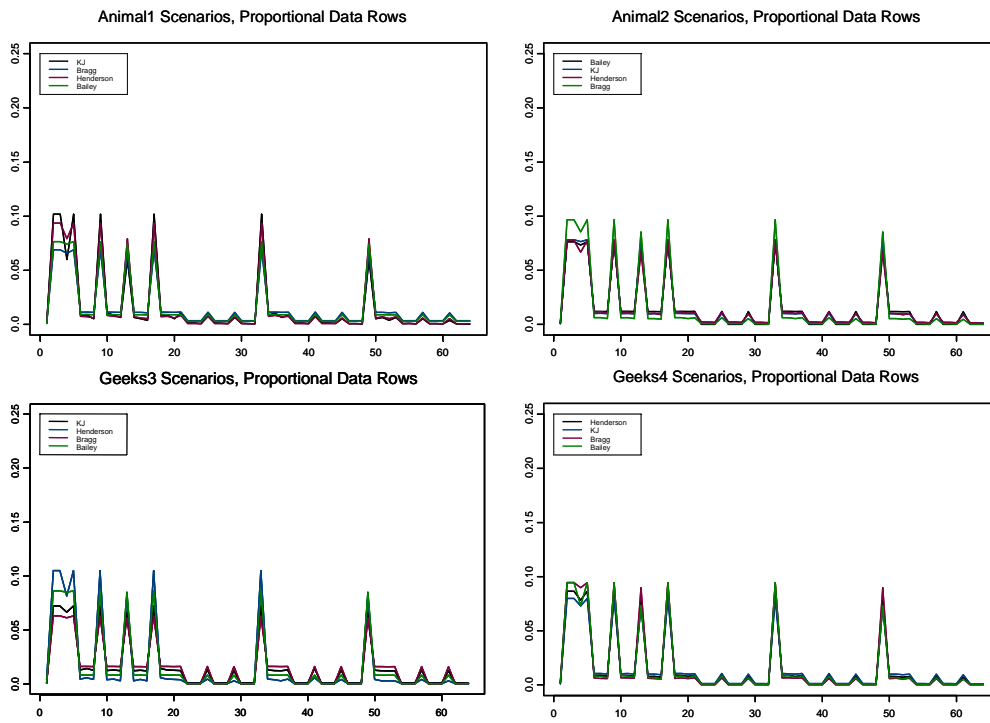


Figure C.3. Plots of Data Normalized by Rows for Each Scenario

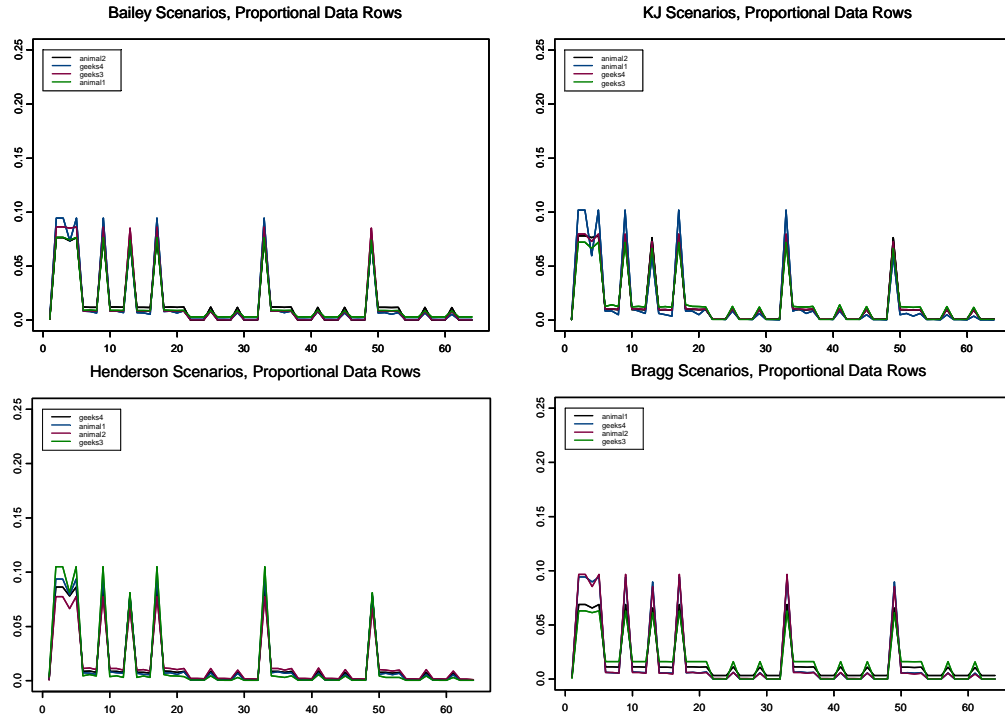


Figure C.4. Plots of Data Normalized by Rows for Each Analyst

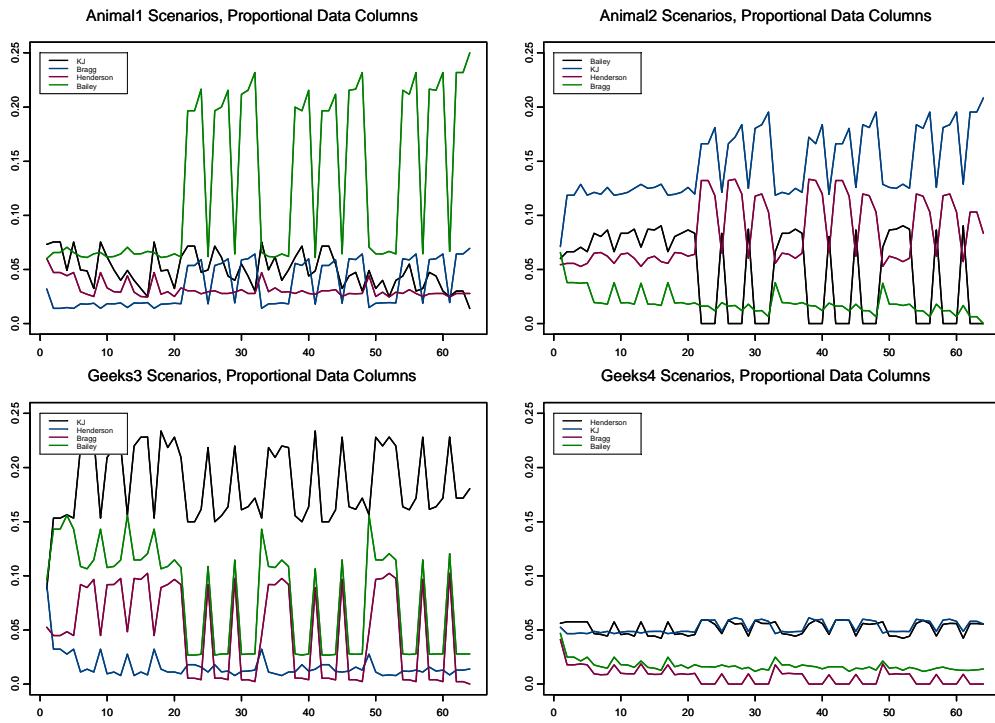


Figure C.5. Plots of Data Normalized by Columns for Each Scenario

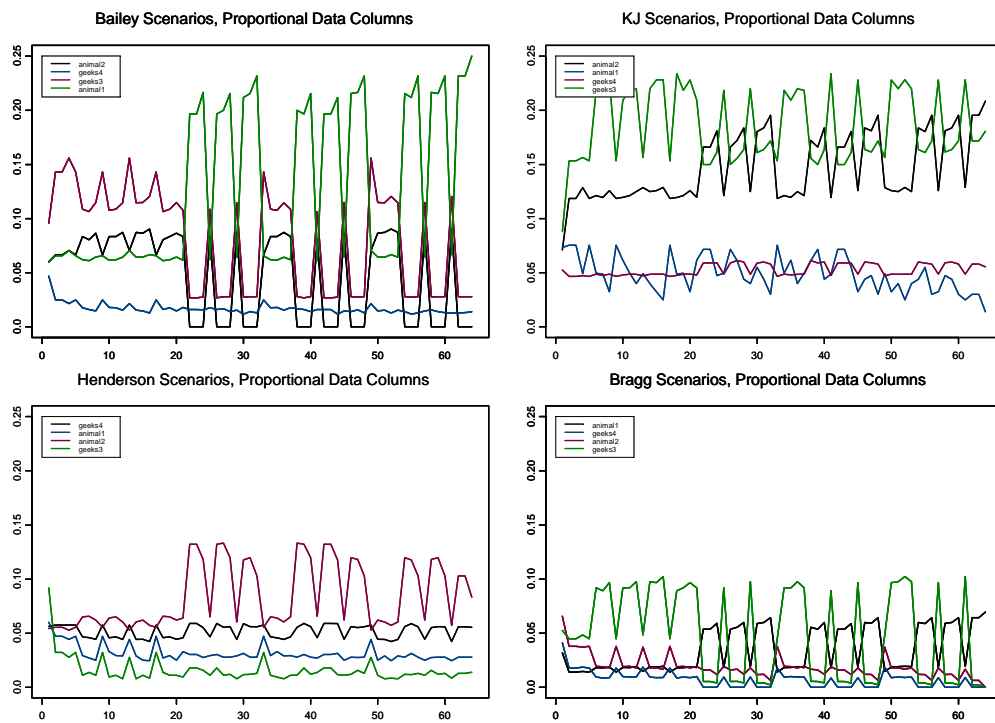


Figure C.6. Plots of Data Normalized by Columns for Each Analyst

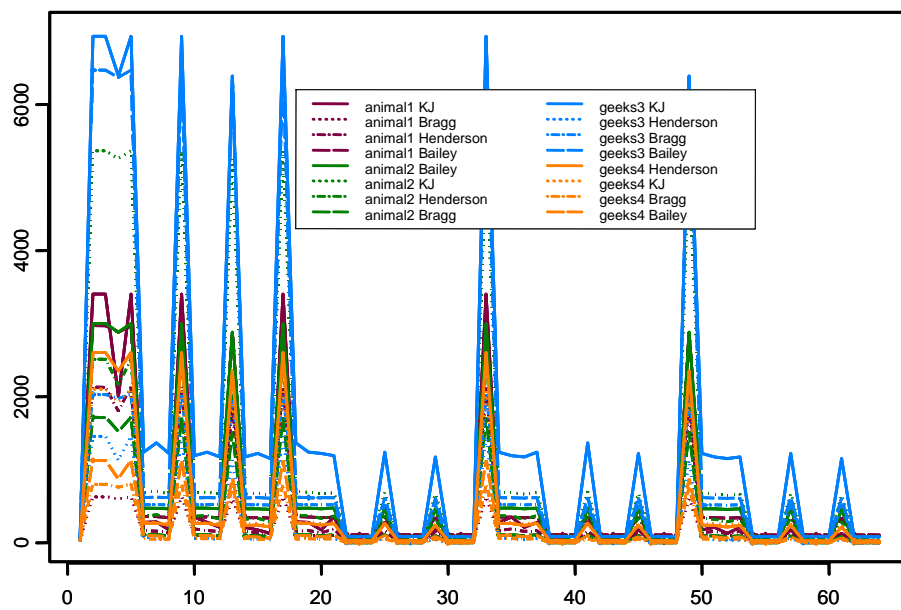


Figure C.7 Plot of Raw Data for All Sixteen Graphs

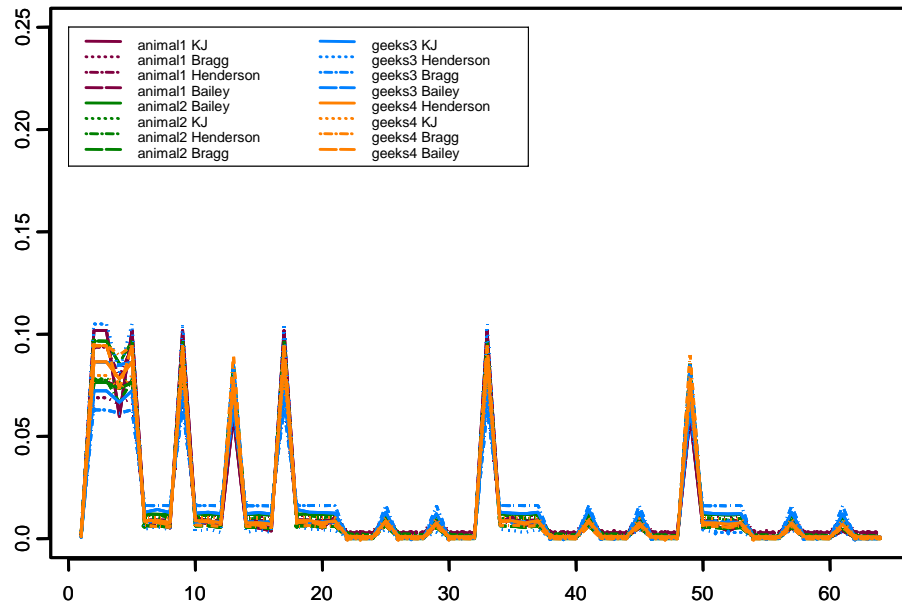


Figure C.8. Plots of All Sixteen Graphs; Data Normalized by Rows

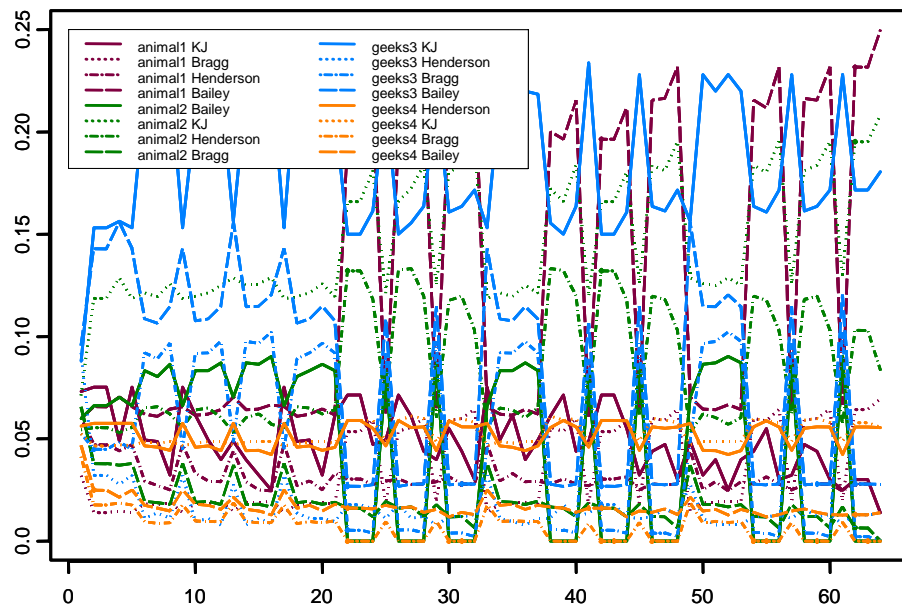


Figure C.9. Plots of All Sixteen Graphs; Data Normalized by Columns

Distribution

No. of Copies

ONSITE

14 Pacific Northwest National Laboratory

T.K. Bragg	K8-41	C. Posse	K6-08
A.J. Brothers	K6-08	R. Quadrel	K7-10
G. Chin	K7-90	A.P. Sanfilippo	K7-22
C.L. Henderson	K8-41	J.J. Thomas	K7-10
E.G. Hetzler	K7-22	C.F. Varley	K8-41
O.A. Kuchar	K7-22	P.D. Whitney	K6-08
D.K. Lemon	K8-40	K.E. Wolf	K8-41