Automatic Generation of Social Network Data from Electronic-Mail Communications

Jason Yee[1], Robert F. Mills, Gilbert L. Peterson, and Summer E. Bartczak
Air Force Institute of Technology

Abstract

Most organizations have formal and informal elements. Formal structures are usually documented in organizational charts showing chain of command, levels of authority, and personnel resources. The actual effectiveness of the organization or specific individuals may actually depend on informal structures and internal communication networks. These are by definition personality-dependent and may provide significant insight into how work actually gets done within the organization.

Effective leaders will want insight into these informal structures for various reasons. Inefficient decision-making or staffing processes may result in unnecessary or redundant communications, chokepoints, or single points of failure, each of which can either delay decisions or degrade the quality of those decisions. Further, sudden changes in the informal structures may indicate underlying stresses within the organization, interpersonal conflicts, or behavioral problems that may significantly disrupt the mission effectiveness or morale of the organization.

Documenting these informal structures and networks can be achieved through a variety of means, often through personal interviews or direct observation, both of which are difficult and time consuming. In this paper, we describe a method of automatically generating social network data using electronic mail messaging logs. Performance is demonstrated using three months of real data from a medium sized organization.

I. Introduction

Interpersonal communication and behavior is important to the productivity and innovation of all organizations. Thus, any information or insight into understanding this communication and behavior is therefore very useful to an organization. Social network analysis is a relatively new field of psychology and sociology, and it is founded upon the idea that the relationships between people are just as important as the attributes of people. Thus, social network analysis provides a rigorous and standardized framework for analyzing internal communication patterns among individuals and groups and has become an increasingly powerful tool. With broad application, social network analysis has been used to help streamline business processes, improve internal organizational communication, and even position routers in a network topology with great success. Finally, social network data can be analyzed with graph theory concepts, allowing the speed and power of computers to be leveraged.

Unfortunately, generating social network data is time consuming and may require a large degree of cooperation from the subjects being studied. Additionally, a social network will also change over time, based on shifts in workload or project prioritization, which might render existing data obsolete. The purpose of this research then is to investigate means of efficiently building a social network map using automated tools/procedures and administrative logs of computer mediated communications.

[1]Student Author

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2005		**00-00-2005 to 00-00-2005**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Automatic Generation of Social Network Data from Electronic-Mail Communications	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Air Force Institute of Technology,Wright Patterson AFB,OH,45433**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**42**	

Our work focuses on usage of electronic mail logs, but the concepts could easily be extended to address instant messaging, web usage, chat groups, and electronic forums. Our research demonstrates that an automated system to generate useful social network data in a reasonable amount of time can be created. This system can then provide raw social network data to other tools that use social network analysis. The proof-of-concept tool generated is a form of middleware to process raw data into a form suitable for follow-on analysis.

II. Social Network Analysis

Social Network Analysis draws from the fields of Psychology and Sociology, to study people and the relationships between groups of people [1]. While it is generally easier for a sociologist to study individuals and their attributes, the ultimate goal of a sociologist is to understand the society itself: relationships within a group of social entities and how they affect the individuals.

Wasserman describes social network analysis as a "distinct research perspective within the social and behavioral sciences; distinct because social network analysis is based on an assumption of the importance of relationships among interacting units" [1]. Instead of focusing on the attributes of the individuals as is done in standard social analysis, social network analysis focuses on ties, the interactions and relationships between the individuals as a way of characterizing their behavior. When social network analysts study ties, they interpret their functioning in the light of the actors' relations with other network members [2].

A. Social Network Data

For example, a standard sociological study on the importance of individuals in an organization might count the number of phone calls each individual makes and receives and take into account the attributes of the callers, like age and gender as shown in Table I. The study then concludes by hypothesizing a relationship between the attributes and the measured importance of the sample.

On the other hand, a social network approach analyzes *who is calling whom* and the groups that are formed as a result. A secretary, for example, may make a lot of phone calls, but is not necessarily the most important person in the organization. The social network perspective looks at the relationships between the actors. Simple network data is composed of actors, the entities being studied, and ties (relationships) between those actors. Social network data is often displayed in an adjacency matrix as shown in Tables II and III, or in an edgelist as shown in Table IV.

Name	Gender	Age	Calls
Alice	F	34	3
Bob	M	32	13
Carol	F	49	11
Dan	M	19	9

TABLE I

Example of Standard Sociometric Data

	Alice	Bob	Carol	Dan
Alice	-	1	0	0
Bob	1	-	1	1
Carol	0	1	-	1
Dan	0	1	1	-

TABLE II

Example of an Undirected, Binary Sociogram, Social Network Data of an Organization

The information shown in these tables are often referred to as sociograms or social network maps. They depict the fundamental unit of study in social network analysis [3]. In Table II, the presence of a 1 at (x, y) in the matrix represents the presence of a tie between the two

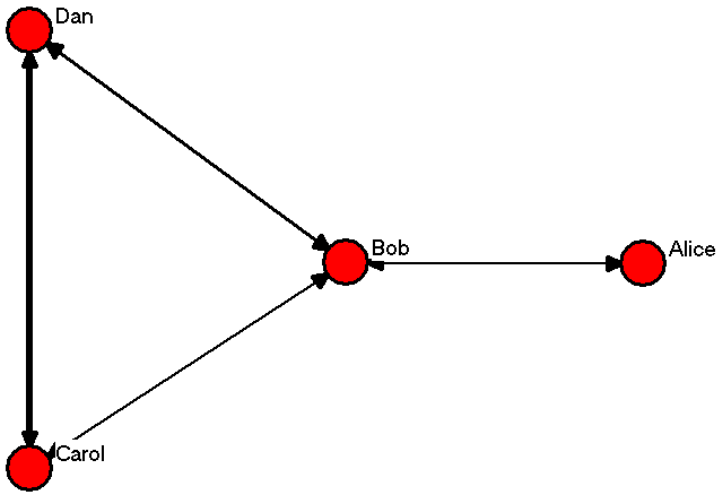


Fig. 1. Sociogram in Graphical Form

actors x and y. Since there is a 1 where row C and column A intersect, actor $Carol$ has a relationship with actor $Alice$. Notice that there is no mention of the attributes of the actors in this data.

In this example, the data is binary: each tie is either 1 or 0, meaning there is either a connection, or there is not. It is possible to have nonbinary connections, allowing the strength of the connections to be studied as well. For example, in Table III, actor $Carol$ is connected to actor Dan with a tie strength or weight of 7. These ties need not be reciprocal, meaning that the connection may be stronger in one direction than in the other. The edgelist in Table IV shows three columns to describe the relation between two actors: source actor, destination actor, and the strength of the relationship. Notice also how the data in the edgelist corresponds to the tie strengths in Table III.

	Alice	Bob	Carol	Dan
Alice	-	3	0	0
Bob	4	-	4	5
Carol	0	4	-	7
Dan	0	3	6	-

TABLE III

EXAMPLE OF A DIRECTED, WEIGHTED SOCIOGRAM, SOCIAL NETWORK DATA OF AN ORGANIZATION

```
dl
n = 4
format = edgelist
data:
Alice Bob 3
Bob Alice 4
Bob Carol 4
Bob Dan 5
Carol Bob 4
Carol Dan 7
Dan Bob 3
Dan Carol 6
```

TABLE IV

EDGELIST FORMATTED SOCIOGRAM

From the data, it is possible to construct a visualization of the actors' relationships with each other by making each actor a node in a graph and drawing an edge between them if a relation exists. This visualization can show the structure of relationships as shown in Figure 1.

As they are intrinsically graphs, sociograms adhere to the rules of graph theory. Graph theory is very useful in social network analysis because it provides a vocabulary that can be used to label and denote social structural properties. It also gives social network analysts the tools used in studying graph theory: mathematical operations and concepts that quantify and measure structural properties. Perhaps most importantly, giving social networks a rigorous, symbolic representation means that the power of computers can be leveraged to aid in research.

B. Social Network Analysis Capabilities

Before the widespread availability of computers, the fruits of labor of a lengthy interview or observational study were tedious arithmetic to calculate metrics of the social network data and analyze them. Even more work is required to create a visualization for the social structure. Computers have changed the speed of social network analysis dramatically. Social network analysis programs like UCINet [4], Pajek [5], and KrackPlot [6] can calculate the properties of a graph and visualize it with a few clicks of a mouse [2]. These programs use graph theory algorithms and concepts to quickly calculate social network metrics.

The basic metrics of social network analysis revolve around activity, betweenness, and closeness. Activity is often measured in terms of degrees, the number of ties that an actor has. Betweenness is a measure of how many shortest paths between two actors go through a specific actor. Closeness is a measure of how few connections are required to connect to other actors. These basic social network concepts are used to calculate almost all of the social network metrics and gain information about the entire network as well as the actors themselves.

Some social network analysis metrics that may be useful are Freeman Density, Bonacich Power, and k-cores. These metrics are standard in the social network analysis software programs UCINet and Pajek. The Freeman Density is a measure of the centrality of an actor based on the number of actors connected to it. Bonacich Power is a metric taken to demonstrate the ability of social network analysis to provide information about actor power and centrality. A k-core is a loose group of actors in which more tightly-knit groups of actors are found. Additionally, analyzing the changes to a specific actor's ego network also indicates changes to an actor's behavior. [7].

Currently, many different applications and uses of analysis on social network data have been proposed and implemented. One of applications that are most helpful to organizations is the use of social network analysis to analyze the communication networks within organizations.

1) Organizational Network Analysis: Rob Cross of the University of Virginia uses social network analysis to analyze organizational networks and help improve company productivity by increasing collaboration and information flow [8]. Cross calls this Organizational Network Analysis (ONA). Fundamental to ONA is the idea that people work well when they work together, and even better when the right people are connected. Social network analysis finds bottlenecks of information flow, organizational hermits who are not collaborating, and elitist groups that don't interact with those outside of the group. With the information he acquires, Cross recommends ways to restructure or join groups through meetings or the hiring of mediators. Among other things, ONA has been used to help integrate newly merged companies, improve strategic decision making in top leadership, and promote creative thought. Given its results, ONA is very effective, as some of Cross' most prominent clients are: American Express, Accenture, Asea Brown Boveri (ABB), Abbey National, A D Little, Aventis, Bank of Montreal, BP, Bristol-Myers Squibb, Capital One, Cardinal Healthcare,

Conoco, CSC, Eli Lilly, EnCana, FAA, Halliburton, IBM, Intel, Mars, Martha Jefferson Hospital, McKinsey, Microsoft, Nortel, Novartis, NSA, PriceWaterhouseCoopers [8].

2) Covert Network Analysis: Vladis Krebs used social network analysis to "uncloak terrorist networks" after the terrorist attacks of 2001 [9]. By gathering data from news articles that followed the attacks, Krebs was able to construct a sociogram representing the terrorist network. After some investigation by the US government, it was alleged that Mohammed Atta was the leader of the covert operation. Krebs took the network centrality metrics of degree, closeness, and betweenness, and found that Atta had the highest score for all three metrics. These social network metrics support the idea that he had been the leader of the operation. This does not mean that social network analysis can necessarily predict criminal activity; however, it may can help determine organizational structure and importance of members within a society or group.

C. Difficulty in Gathering Social Network Data

Gathering social network data can be a difficult and time consuming task. It is even more complicated when dealing with large populations and over protracted periods of time. Standard methods of data collection include conducting interviews/surveys, observing the actors, or extracting data from archived records.

The interview or survey-based approach to collecting data is extremely time consuming and not possible in all situations. Interviews and surveys inconvenience the people being studied (and can potentially invade their privacy), especially if they need to be repeated for a longitudinal study. Additionally, the questions asked are fairly simple and are only taken in one context. Further, these questions themselves often lead to bounding the number of connections, sometimes asking to name only a certain number [10]. In addition, these questions often do not capture the relative weight of the relationship and only the presence of a tie. Moreover, resources must be expended to carry out the interviews and surveys. The Network Roundtable at the University of Virginia developed a tool that can generate social network data from the results of customizable online surveys. While this expedites the process, it still requires user interaction and is only as accurate as the person filling out the survey [11].

Resources required to observe a social structure with people are especially high and time consuming. Observation often requires getting the permission of those studied, a permission that is not always granted. Moreover, the observers can only gather so much information, and this method of data collection works best when studying relatively small groups of people with close interaction [1].

On the other hand, the cost of gleaning social network data from archived information does not require direct interaction with live subjects. Information can be gathered from lots of different sources such as newspapers, attendance records, or email traffic [9]. Gathering data from recorded archives is done in a short period of time as opposed to gathering data as it happens, making it much easier to perform longitudinal studies. However, it does require time to read the archived data and extract the pertinent information. Moreover, as archives are records of the past, they often do not provide information about the current social structure.

D. Summary

The study of social networks has profoundly influenced the fields of mathematics, statistics, and economics as well as the fields of sociology, and psychology [1]. Social network analysis can provide useful information about groups and the actors within them. However, social

network analysis suffers from the lack of current, dynamics social network data of large organizations. This deficiency is addressed in is research.

III. Proof-of-Concept Tool Implementation

The objective of this research is to show that the creation of useful social network data from Simple Mail Transfer Protocol (SMTP) data in a time-efficient manner is possible. This data would then be read and analyzed by current social network analysis tools. It is believed the metrics gathered by these tools will prove useful to social network analysts in characterizing organizational behavior. These characterizations could then be used in a tool to provide leaders with more information about their organizations and internal communication patterns.

Creating social network data from readily available SMTP logs is cheaper, easier, and quicker than conducting surveys and relying on direct observation. The automatic creation of social network data also allows social network analysts to study the short-term dynamics of a large set of actors; this is extremely difficult—if not impossible—to do with current social network data gathering methods.

To create social network data from email logs, data is gathered, filtered and/or parsed, and mined for information. The overall process is illustrated in Figures 2 and 3. In our research, we also included an optional anonymization step to preserve privacy. Whether or not anonymization is needed depends on the intended application at hand. If general trends are being studied, then anonymization may be appropriate. Other applications may include personnel security or insider threat mitigation, in which case traceability to a particular individual is required. While this may seem to be a *Big Brother* issue, it is commonly accepted that most businesses and government organizations will routinely monitor email usage.

Several java programs were developed as shown in Figure 3. The programs are written in Java 1.50 and require Java JRE version 1.5. Queries are made with MySQL version 4.1. Microsoft Exchange and Javelina ADVantage generate the proxy list file `proxy.csv` and the SMTP logs. Although Exchange was used for our work, any electronic mail system that generates SMTP logs in the appropriate format can be used. The system is expected to be given a fresh set of SMTP logs every week or month to sanitize, process, and add to the database. This relatively short time interval ensures that the social network data recorded is current and thus presents a more accurate description of the behavior of the system users.

A. ProxyListToUID Component

The `ProxyListToUID` component resolves the multiple aliases of an actor to a specific actor. For example, a user in the system named *Jason Yee* might use both the email addresses

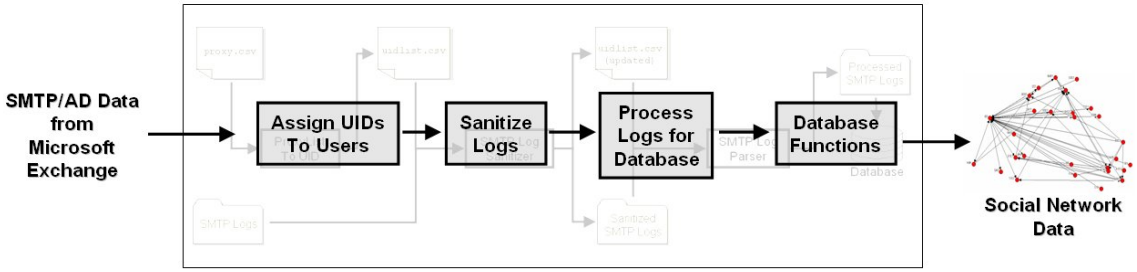


Fig. 2. Process of Generating Social Network Data from SMTP Logs

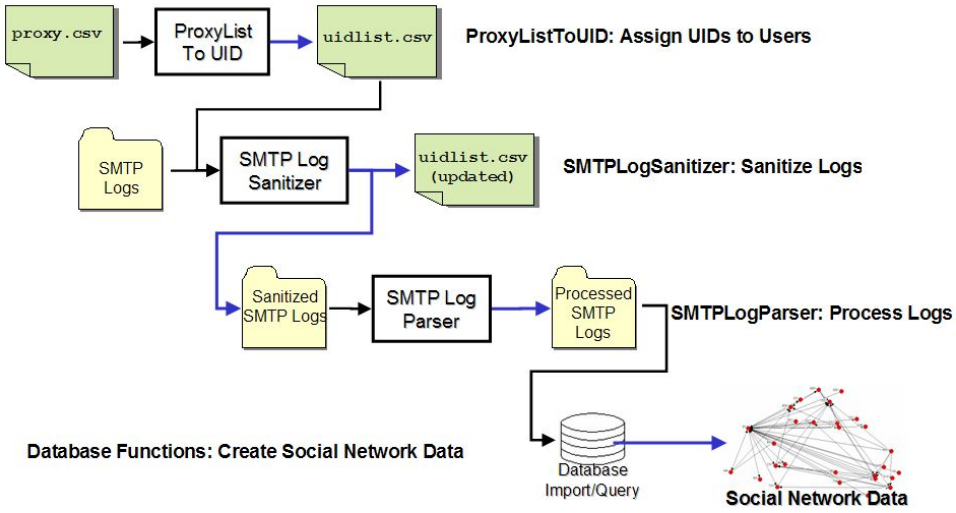


Fig. 3. System Implementation

`jyee@afit.edu` and `jason.yee@afit.edu`. This component assigns the same identity to the two different email addresses. It is important to understand that a UID is meant to correspond to a single human user. It is essential to identify unique users if social network analysis results are expected to be accurate. This is something that is taken for granted when carrying out interviews and surveys to gather social network data.

B. SMTPLogSanitizer Component

For the purposes of this research, specific identities were not required nor desired. Thus, before the SMTP logs are analyzed, they are sanitized. The sanitization process used in this research makes SMTP logs relatively safe for distribution by replacing the user name of an email address with a unique number, masking the identity of that user. It is important to note that NCSA formatted SMTP logs do not contain any information about the subject or content of the email message. Thus, an SMTP log is considered sanitized if the user names are anonymized.

The `SMTPLogSanitizer` program keeps a record of the email addresses already seen and the number of unique users identified. If the log sanitizer finds an email address that it has already seen, it replaces that email address with the UID associated with it. If the log sanitizer finds an email address that it has not seen, it generates a new UID, assigns it to the email address, and replaces the email address with the UID. The log sanitizer program `SMTPLogSanitizer` was built with those requirements.

C. SMTPLogParser Component

The `SMTPLogParser` program extracts data from SMTP logs that are imported into a database. Useful information from SMTP logs are the date and time an email was sent, the sender and recipient, if the sender and recipient are internal, and how many recipients received the same email. `SMTPLogParser` parses the contents of each sanitized log file and and extracts that data. This information is used by the database to generate social network data.

D. Database Functions Component

The processed information provided the needed information to create social network data. All that remains is to configure the database, import the data from the processed logs, and export the query results. Before importing the data, a table carrying the values is created. The fields in the table are the same as the fields created by the `SMTPLogParser` program including a primary key: MessageID (database key), Date/Time, UID of Sender, UID of Recipient, Internal Status of Sender, Internal Status of Recipient, and Number of Recipients of Message. When the database and required table is prepared, the data extracted by the `SMTPLogParser` is imported. Finally, social network data in UCINet-readable edgelist format (Table IV) is extracted from the `SMTPLogParser` generated data in the database.

E. Generated Sociograms

Sociograms are generated in the edgelist format as shown in Table IV. This format is standard for social network analysis applications and contains three columns: source actor, destination actor, and tie strength. As was shown in Table III, tie strength is a measurement of the strength of the relationship between the source and destination actors. There are many different ways to calculate tie strength, but for the purposes of this experiment, tie strength is calculated as the number of email messages sent from the sender to the recipient.

In conjunction with UCINet, the system is able to generate different types of sociograms based on different variables and parameters. While it is not within the scope of this research to determine the correct settings to make the sociograms correspond to the actual population, the flexibility and customizability of the tool will aid future research in this endeavor.

IV. Tool Performance

A. Evaulation Metrics and Techniques

The metrics used to evaluate the social network data generating system in this experiment are usefulness and timeliness. Usefulness is the ability for the generated data to be read by the standard social network analysis program UCINet. UCINet is a commonly used social network analysis program, and files that can be read by UCINet can be read by almost all other social network analysis programs such as Pajek and NetDraw. Thus, for our purposes, if UCINet can take some standard social network analysis metrics from the generated social network data, the tool is deemed useful. Whether or not the sociograms generated from email logs truly represent the organization's internal communication patterns is a topic of further study. Timeliness is measured by determining the amount of time elapsed in each stage of the log-parsing process. The full process of creating social network data from SMTP logs was repeated several times; the execution time of each stage was recorded via direct measurement.

B. System Setup and Workload

This testing is performed on a dedicated Dell Poweredge Pentium 4 hyperthreading-enabled 3.2 GHz computer with 2 GB RAM and the Windows XP Professional Edition operating system. The system tested is the same system implemented in the previous chapter, with components developed in Java 1.5, and a MySQL 4.0.21 server.

The proof-of-concept tool is tested and timed as it is expected to be run on an expected workload. The workload in this research consists of a list of users and their different email addresses and SMTP logs in the National Center for Supercomputing Applications (NCSA) format [12]. Data was collected from a medium sized organization (approximately 1,500

users) over an 86-day period from October-December 2004. The collected log data consisted of approximately 3.6 gigabytes worth of text as shown in Table V.

Month	Days	Size
October	25	1,034,273 KB
November	30	1,385,098 KB
December	31	1,359,860 KB

TABLE V
WORKLOAD DESCRIPTION

C. Validation and Verification

Verification that the components are performing correctly is tested by debugging and white-box code walkthroughs. Validation of the results is done by testing system with artificially-generated SMTP data. The social network data of the artificial SMTP data is gathered manually and is compared to the social network data automatically generated by the system. The components are correct if they provide the expected output. This method of evaluation is justified as no similar system with similar data has been implemented. This portion of testing is done during the development of the system and will not be reported in the results and findings.

V. RESULTS AND FINDINGS

A. Overall Test Results

The average time required for the components to process three months of SMTP data and generate social network data is about 80 minutes as shown in Table VI. On average, it takes about half an hour to add a month to the database and generate a set of social network data incorporating the new information.

Component	Runtime (One Month)	Runtime (Three Months)
ProxyListToUID	< 1	< 1
SMTPLogSanitizer	$\approx$20	$\approx$54
SMTPLogParser	$\approx$10	$\approx$24
Database Import	< 1	< 1
SQL Query	< 1	< 1
Total (three months)	≈ 30	≈ 80

TABLE VI
OVERALL COMPONENT RUNTIMES (MINUTES)

It is clear that this process can produce timely social network data. The bulk of the processing time occurs during sanitization and parsing of the SMTP logs.

B. SMTPLogSanitizer Test Results

The `SMTPLogSanitizer` component is given the workload of three months of SMTP data in the form of 12 separate log files in a directory. Over 170,000 total UIDs were assigned with an average time of about 55 minutes as shown in Table VII. This was the most time consuming of the components. The time spent sanitizing the SMTP logs from October was less than the other months because a server crash in October caused fewer days to be logged.

There were 25 days logged in October, 30 in November, and 31 in December as shown in Table V.

	Test 1	Test 2	Test 3	UIDs Added
October	15.4	14.5	17.7	57,232
November	21.3	20.1	20.1	49,590
December	20.6	19.7	18.7	67,902
Total	57.3	54.3	53.5	174,724

TABLE VII

SMTPLOGSANITIZER RUNTIMES (MINUTES) AND THE UIDS ADDED

The measured data suggests that the number of UIDs added when sanitizing a month's logs are independent of each other. This is unexpected, as the number of additional UIDs is expected to decrease as more email addresses were added to the UID list and are seen again in the logs. This behavior is explained by the many new email addresses (always senders) that are assigned to addresses like `BOUNCE-123@fedweek.sparklist.com`, usually generated by mailing lists to which network users subscribe.

C. SMTPLogParser Test Results

The `SMTPLogParser` component was given the workload of three months of sanitized SMTP data in the form of 12 separate sanitized log files resulting from the sanitization process. `SMTPLogParser` processed three months of data in the average time of about 24 minutes as shown in Table VIII. As is expected, the time spent parsing a month is independent. Each month, with about a million ties, took under 10 minutes to parse. The `SMTPLogParser` mined the sanitized data for over 1.8 million individual emails for a total of over 3 million connections between users as shown in Table VIII. These numbers are consistent with the missing emails from a week of server downtime in October.

To clarify the difference between an email and a connection, when $Alice$ sends an email to Bob and $Carol$, one email was created by $Alice$ and two connections, $Alice \rightarrow Bob$ and $Alice \rightarrow Carol$, are created. Connections are used to generate social network metrics.

	Test 1	Test 2	Test 3	Emails	Ties
October	6.9	6.9	6.2	464,895	883,024
November	9.1	8.7	8.3	675,411	1,149,305
December	9.3	8.0	7.9	689,071	1,065,800
Total	25.3	23.6	22.4	1,829,377	3,098,129

TABLE VIII

SMTPLOGPARSER RUNTIMES (MINUTES), AND THE EMAILS AND TIES RECORDED

D. Data Usefulness Test Results

All 12 generated sociograms were imported by UCINet successfully and the Freeman Degree, Bonacich Power, and k-core social network metrics were taken. The entire network and certain ego networks were also visualized in NetDraw. Thus, the data created by the system is considered usable. Screenshots of UCINet outputs for those metrics are shown in Figures 4-6 in the appendix.

The social network data can be visualized by NetDraw as shown in Figures 7-12 in the appendix. Figure 7 shows the egonet of Actor `59` in NetDraw, and Figure 8 shows a more readable subset of the December 2004 network. Figures 9-12 showcase different visualizations available in Netdraw for the egonet of Actor `1141`.

VI. Summary and Future Research

In summary, our proof-of-concept tool is able to efficiently generate social network maps from SMTP log data. Results show that useful social network data can be created by the system developed in a timely manner. The total time to convert three months of SMTP logs into social network data is about 80 minutes. The data created by this system is readable by UCINet and can be visualized by NetDraw.

The immediate impact of this tool is that it can be used to cheaply and quickly generate social network data from SMTP logs of an organization. The `SMTPLogSanitizer` component, while originally included as a privacy measure, may also facilitate sharing of such logs among the research community, allowing the construction of a rich social network analysis data set. Long-term benefits of this research include the ability to analyze social network data of medium-large organizations. Some potential applications include research on organizational efficiency and personnel security (insider threat) research.

The accuracy of data is defined as how well the data corresponds to the actual behavior of actors monitored. Future research should find the best parameters and restrictions with which to create the most accurate social network data. For instance, this research generates social network maps that only consider connections that are received by fewer than 20 recipients. This restriction may dampen the effect of broadcast "spam" on the data.

This proof-of-concept tool can also be extended to other forms of computer-mediated communication. For example, telephone call logs (facilitated by the adaptation of voice over internet protocol–VOIP), instant messaging logs, and web page access logs may be mined for information.

References

[1] S. Wasserman and K. Faust, *Social Network Analysis, Methods and Applications*. Cambridge, UK: Cambridge University Press, 1994.

[2] L. Garton, C. Haythornthwaite, and B. Wellman, "Studying online social networks," *Journal of Computer-Mediated Communication*, vol. 3, no. 1, 1997, http://www.ascusc.org/jcmc/vol3/issue1/garton.html.

[3] L. Freeman, "Visualizing social networks," *Journal of Social Structure*, vol. 1, no. 1, 2000.

[4] S. Borgatti, M. Everett, and L. Freeman, *Ucinet for Windows: Software for Social Network Analysis.* Harvard: Analytic Technologies, 2002.

[5] V. Batagelj and A. Mrvar, "Pajek - program for large network analysis," *Connections*, vol. 21, no. 2, 1998.

[6] D. Krackhardt, "Krackplot," 2003, http://www.andrew.cmu.edu/user/krack/krackplot/krackindex.html.

[7] R. Hanneman, *Introduction to Social Network Methods*. University of California, Riverside: Department of Sociology, 2001.

[8] R. Cross, "Robcross.org," 2005, http://www.robcross.org/.

[9] V. Krebs, "Uncloaking terrorist networks," 2001, http://www.firstmonday.org/issues/issue7_4/krebs/.

[10] C. McCarty, "Structure in personal networks," *Journal of Social Structure*, vol. 3, no. 1, 2002, http://www.cmu.edu/joss/content/articles/volume3/McCarty.html.

[11] U. of Virginia Network Roundtable, "Online survey," 2005, https://webapp.comm.virginia.edu/SnaPortal/Default.aspx?tabid=34.

[12] Microsoft, "Log formats," 2005, http://www.microsoft.com/resources/documentation/WindowsServ/2003/standard/proddocs/en-us/Default.asp?url=/resources/documentation/WindowsServ/2003/standard/proddocs/en-us/smtp_monitoring_log_formats.asp.

APPENDIX

UCINet Outputs

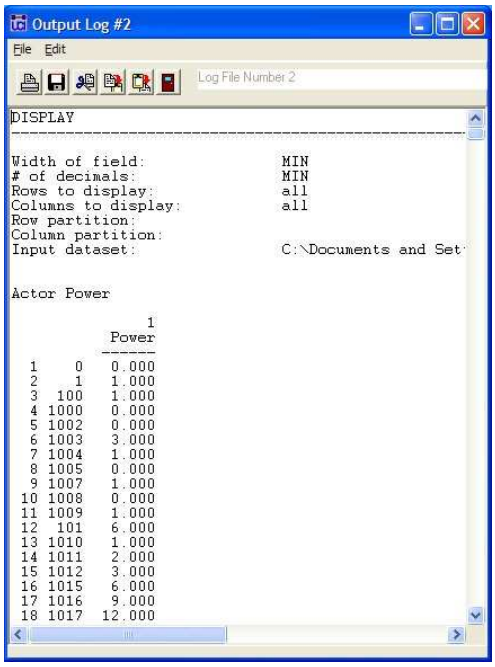


Fig. 4. Bonacich Power Metric

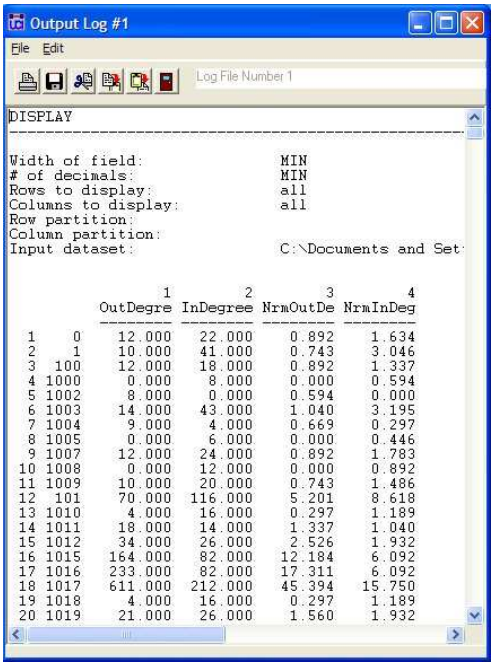


Fig. 5. UCINet Freeman Degree Metric

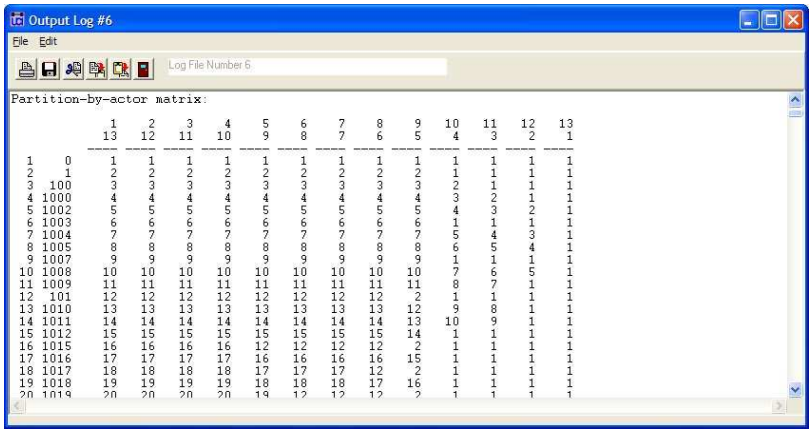


Fig. 6. UCINet Output for k-core Metric

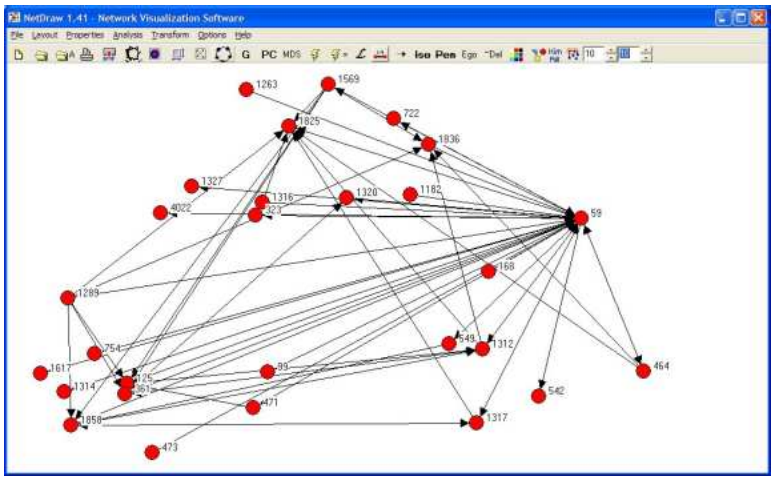


Fig. 7. Egonet of Actor with UID 59

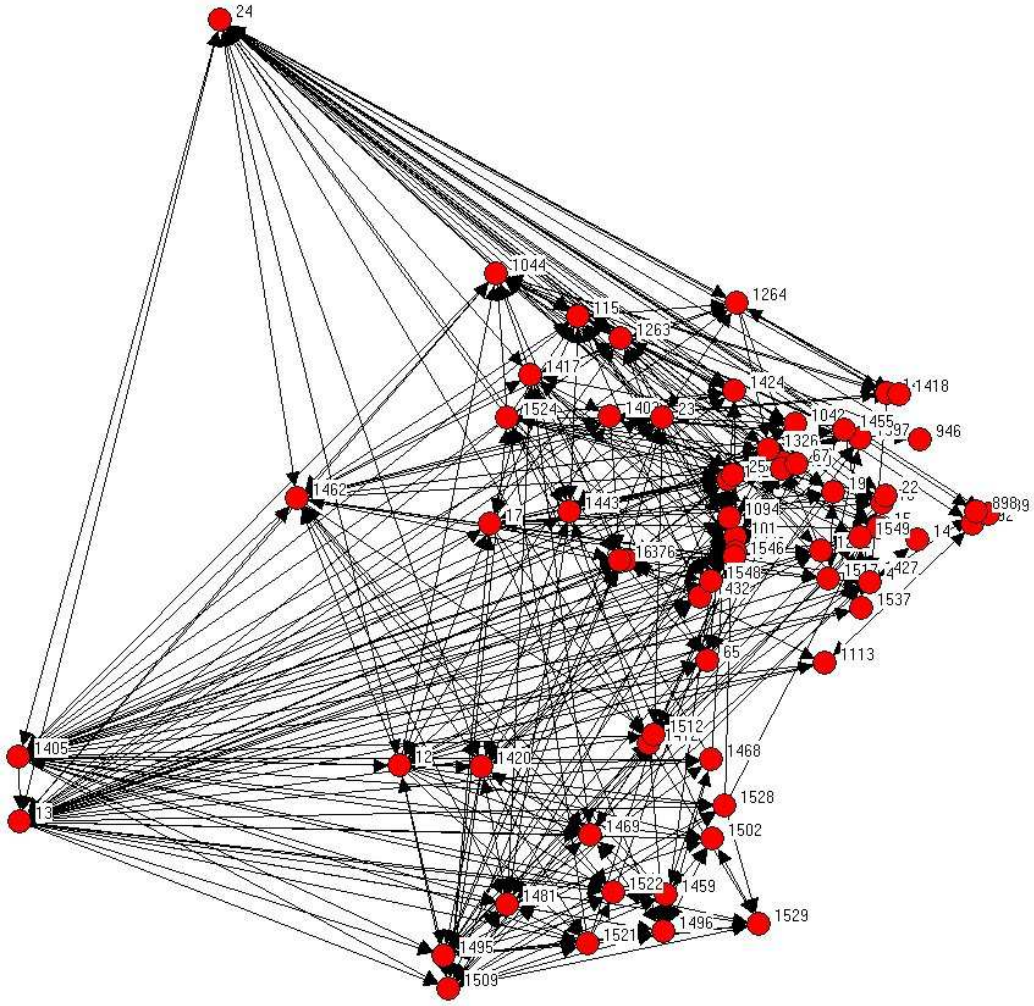


Fig. 8. Subset of December 2004 Sociogram

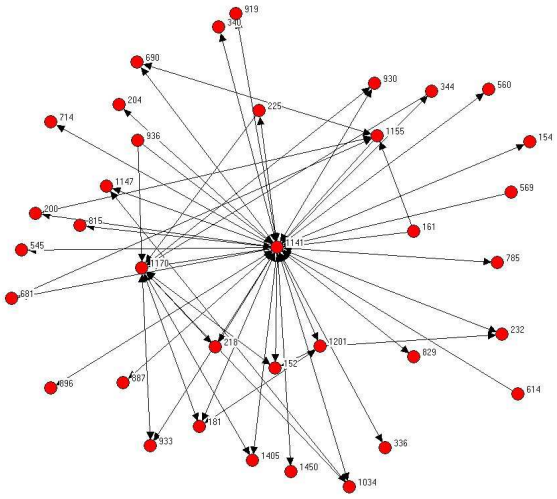


Fig. 9. Spring-Embedding Visualization

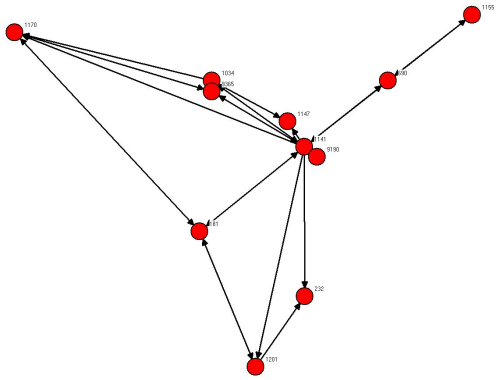


Fig. 10. Gower Visualization

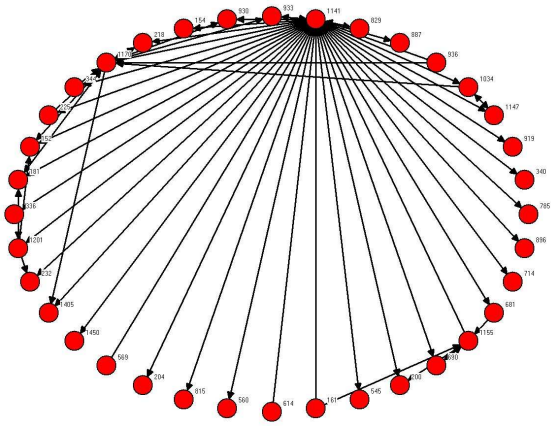


Fig. 11. Circular Visualization

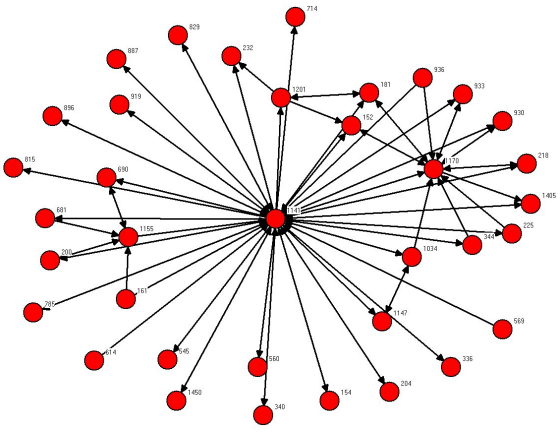


Fig. 12. Multi-Dimensional Scaling Visualization

Air Force Institute of Technology

Automatic Generation of Social Network Data from Electronic-Mail Communications

U.S. AIR FORCE

Jason W.S. Yee
Robert F. Mills
Gilbert L. Peterson
Summer Bartczak

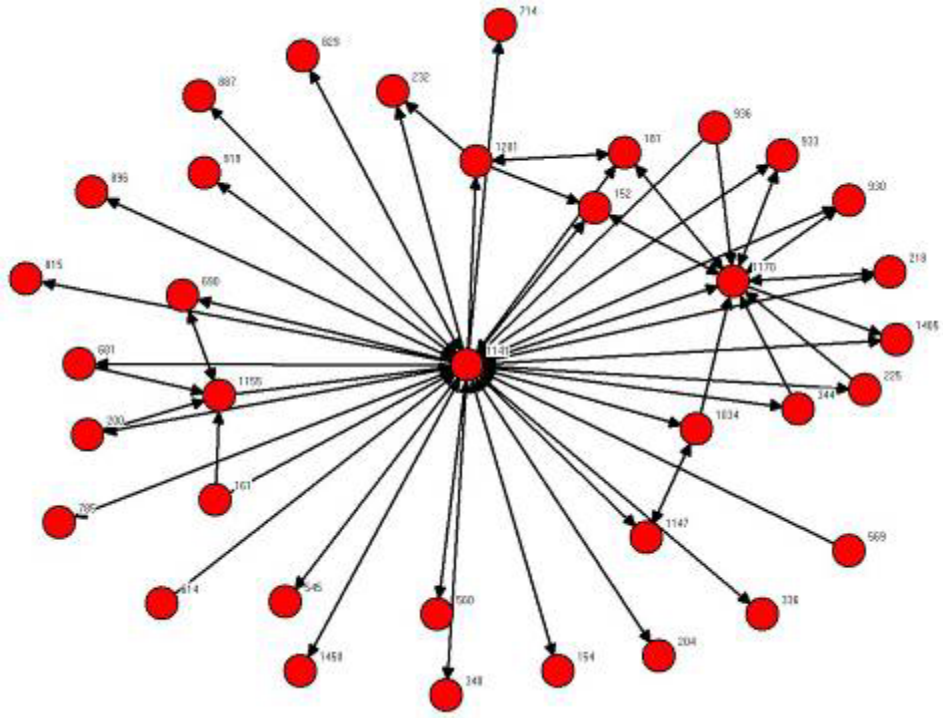
Air Force Institute of Technology
Wright-Patterson AFB OH

robert.mills@afit.edu

U.S. AIR FORCE

Overview

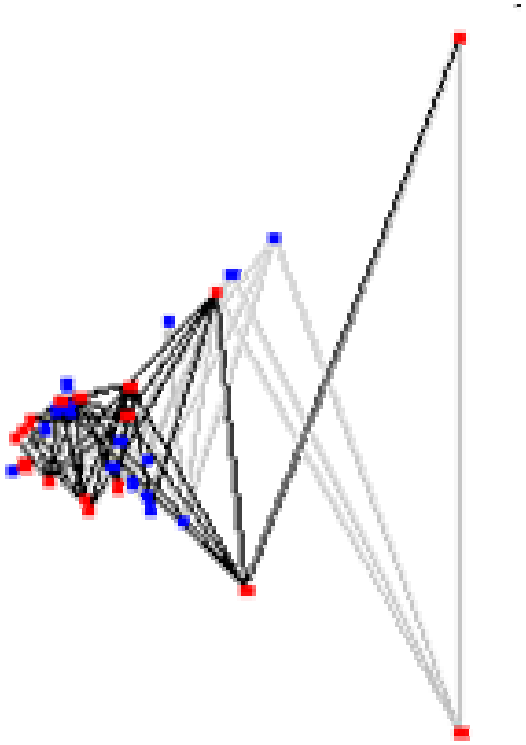
- Background
- System
- Experiment
- Conclusions
- Future Research



U.S. AIR FORCE

Social Network Analysis (SNA)

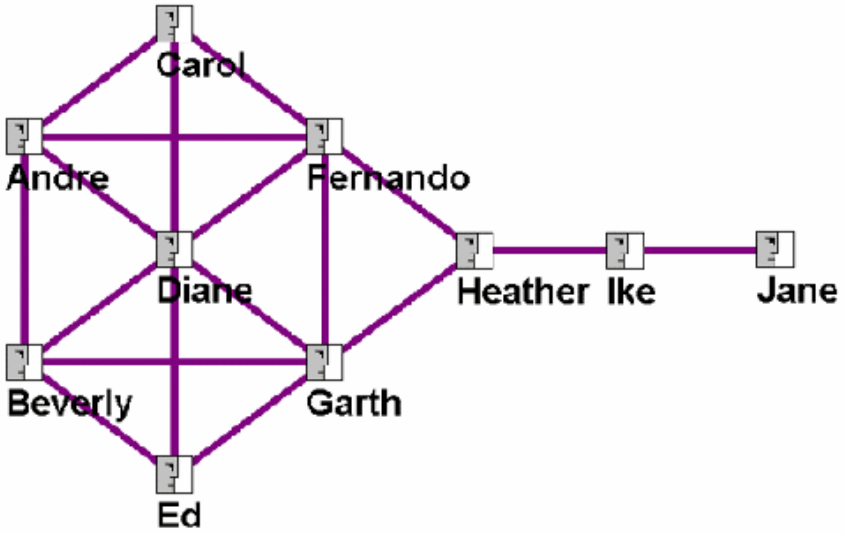
- Blend of Psychology and Sociology
- Who You Know vs What You Know
- Human Behavior Patterns
- Collaboration
- *Personnel Security*



Social Networking Relationships

U.S. AIR FORCE

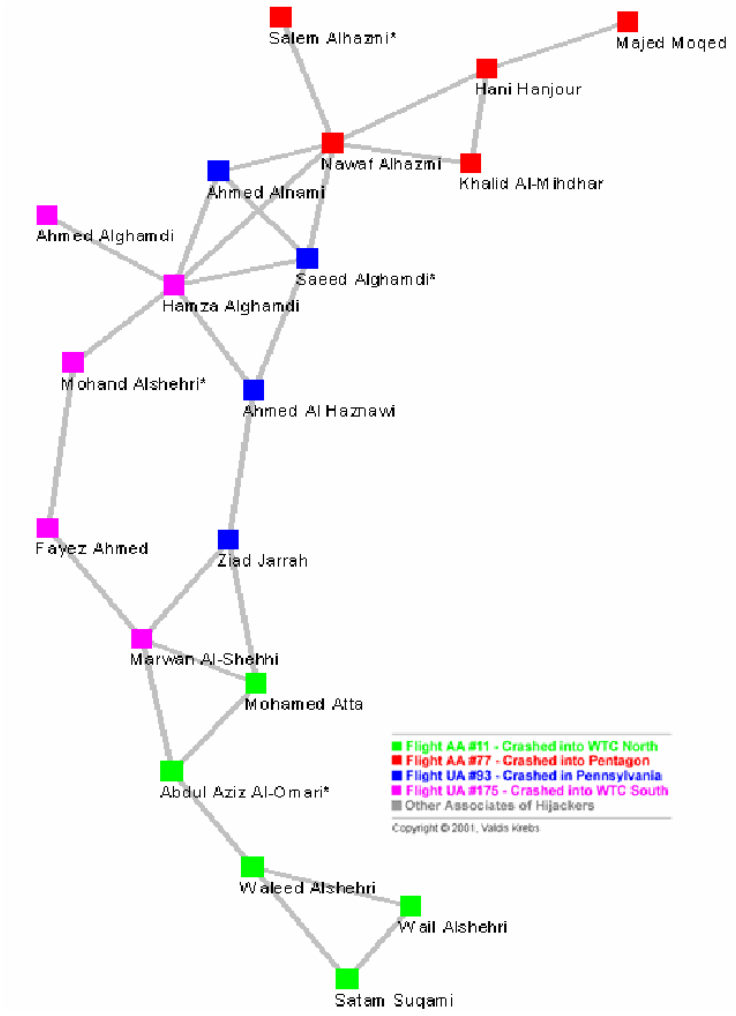
Name	Gender	Age	Calls
Andre	M	34	18
Beverly	F	32	17
Carol	F	49	9
Diane	F	19	32
Ed	M	54	10
Fernando	M	30	19
Garth	M	24	22
Heather	F	55	7
Ike	M	43	6
Jane	F	64	4



- Social Network Perspective
- Importance of *with whom* relations
- Graph Theory

U.S. AIR FORCE

SNA Capabilities

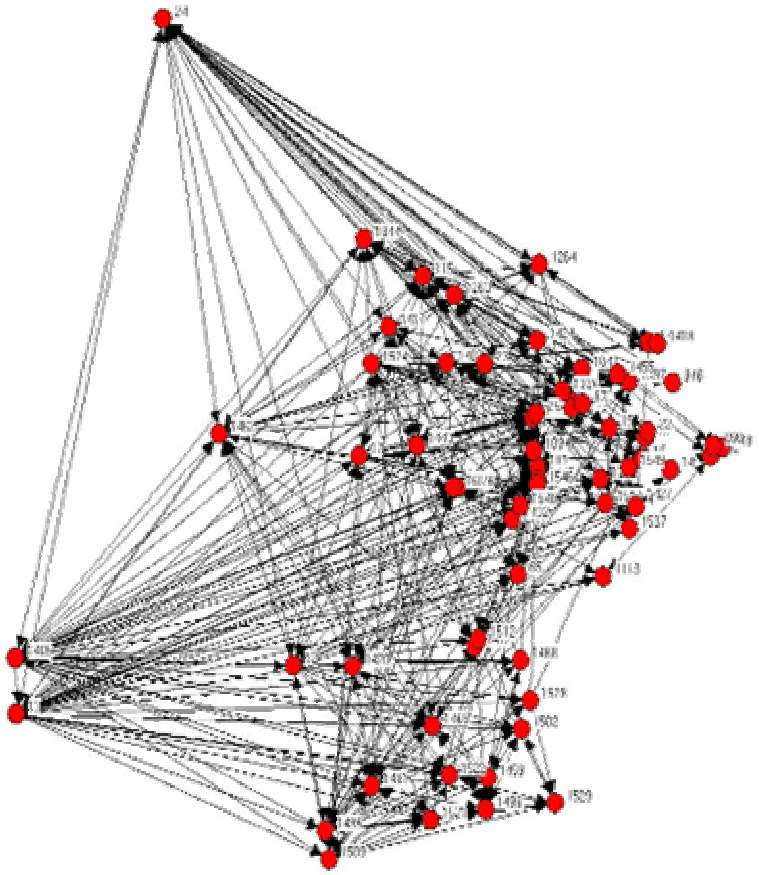


- Social Capital
- Spread of Epidemics
- Organizational Network Analysis
- Covert Terrorist Networks
- Communications Networks

U.S. AIR FORCE

Lack of Social Network Data

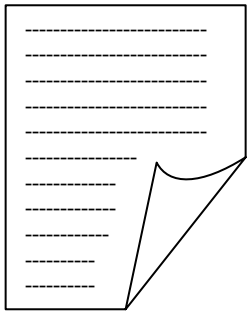
- Relatively New Field
- Available Data is Limited, Old
- Methods
  - Surveys, Observation, Archived Records
  - Costly, Time consuming
- Objective:
  - Gather data in automated fashion
  - Desire for longitudinal study
  - Changes in social network structures



U.S. AIR FORCE

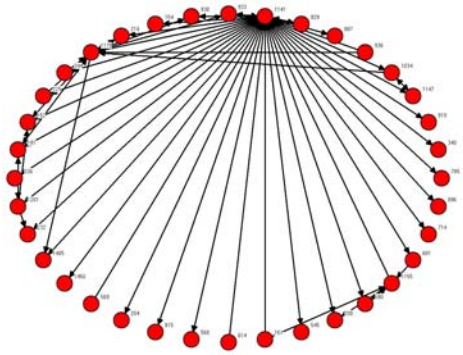
Research Goals

- Create a system that uses automated methods of generating useful social network data
- Automated:
  - E-mail, instant messaging, web browsing, web logs, online forums…anything that has logging capability
  - Our focus: e-mail
- Evaluate the execution timeliness of the system and usability of generated social network data
- Collect more data, at reduced cost, in shorter time



Raw Data

Social Network Data Generator



Social Network Data

U.S. AIR FORCE

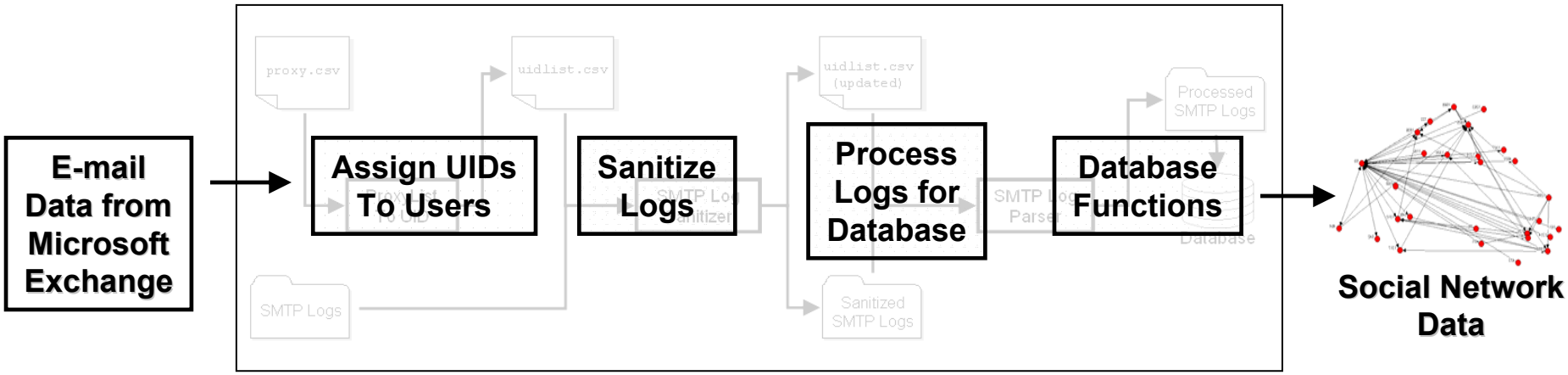
Solution Limitations and Scope

- E-mail Log Data used
  - National Center for Supercomputing Applications (NCSA) formatted e-mail logs
  - Simple mail transfer protocol (SMTP)
  - Collected by organization's e-mail servers
  - Raw data: timestamp, sender, recipient
  - Derived data: status of users (internal/external), number of recipients (one-to-many, one-to-one, etc.)

- Out of Research Scope
  - Content and Subject of Email
  - Validity of Social Network Data
  - Method and Meaning of Social Network Analysis

U.S. AIR FORCE

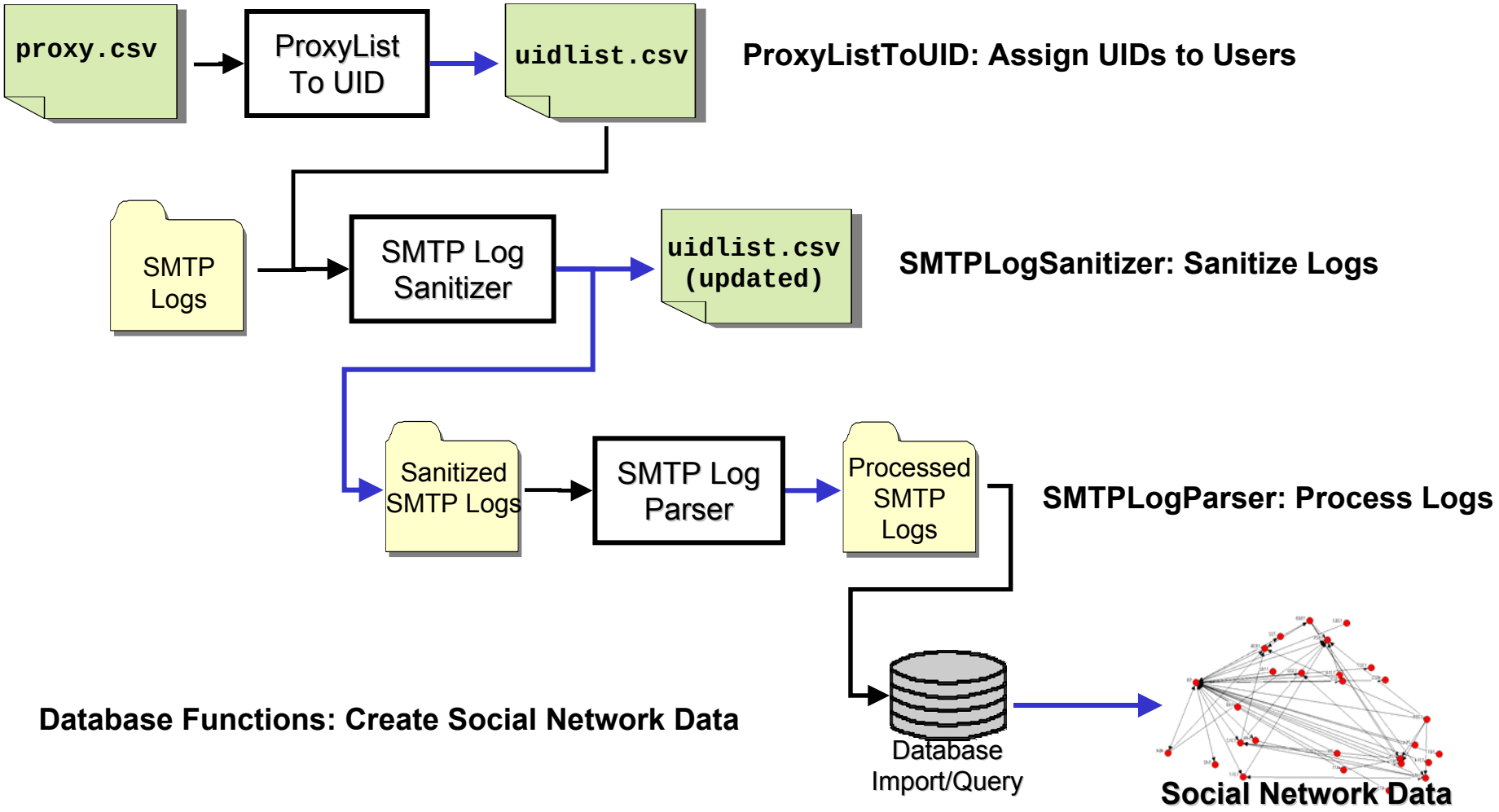
System Components



- Assign UIDs to Users – Attribution
- Sanitize Logs – Privacy
- Process Logs for Database – Mine for Data
- Database Functions – Extract and Format

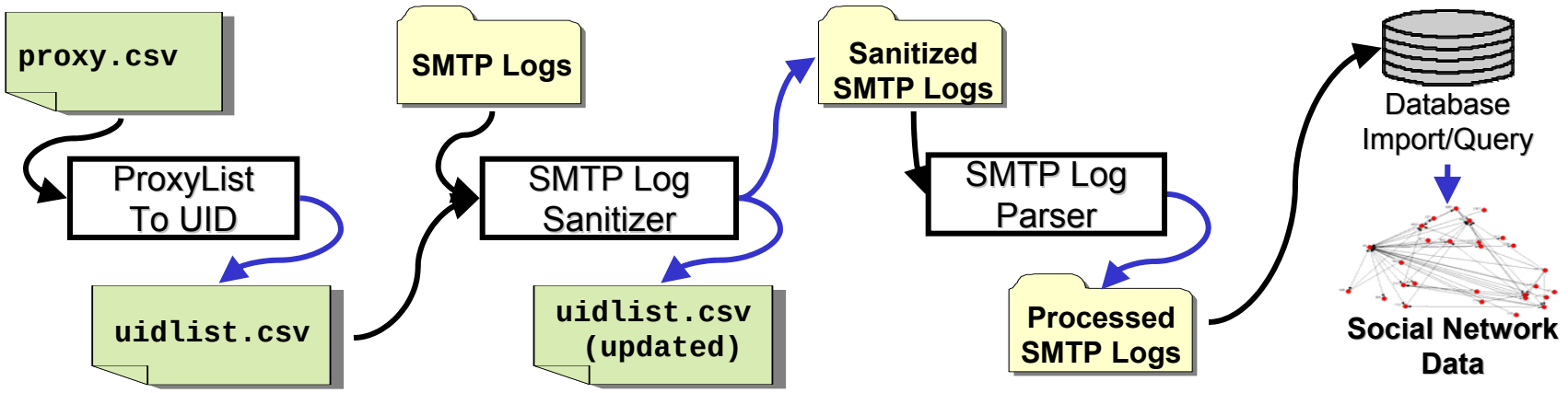
U.S. AIR FORCE

Implementation Overview



U.S. AIR FORCE

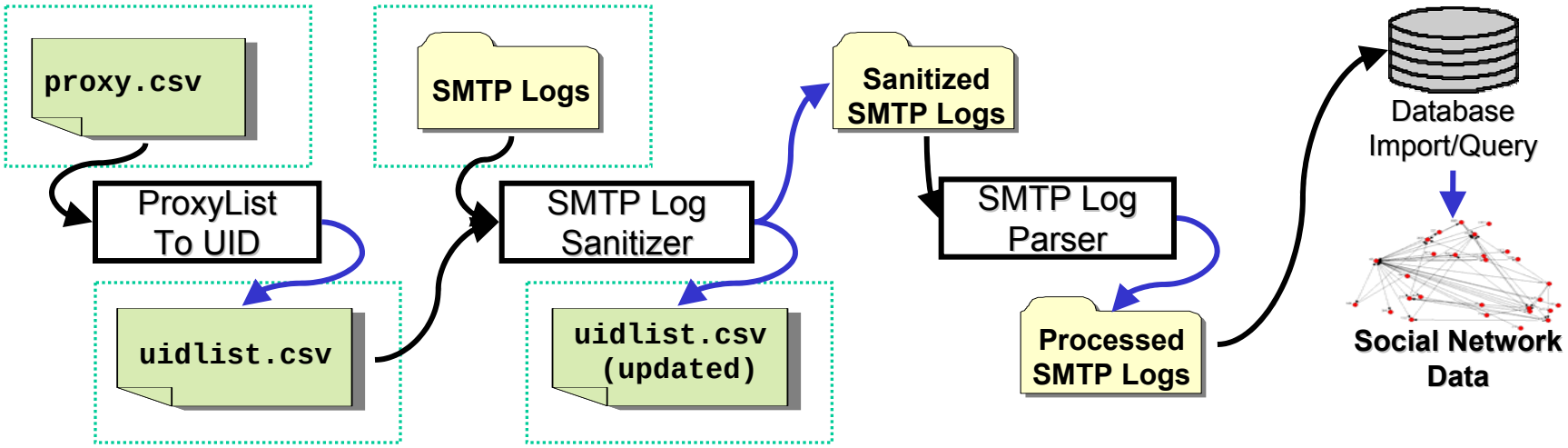
Implementation Overview



U.S. AIR FORCE

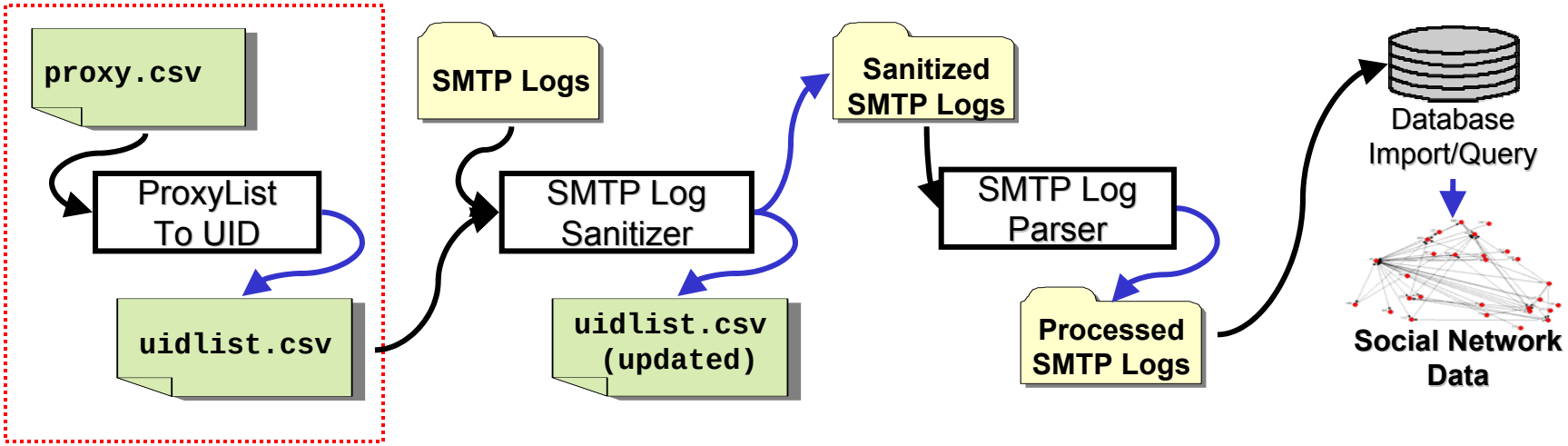
Privacy Protection

- Sanitization may be performed
  - Application dependent
- Sensitive Files
  - Initial Proxy List, Raw SMTP Logs, UID List
- Sanitized Data can be de-sanitized



U.S. AIR FORCE

ProxyListToUID



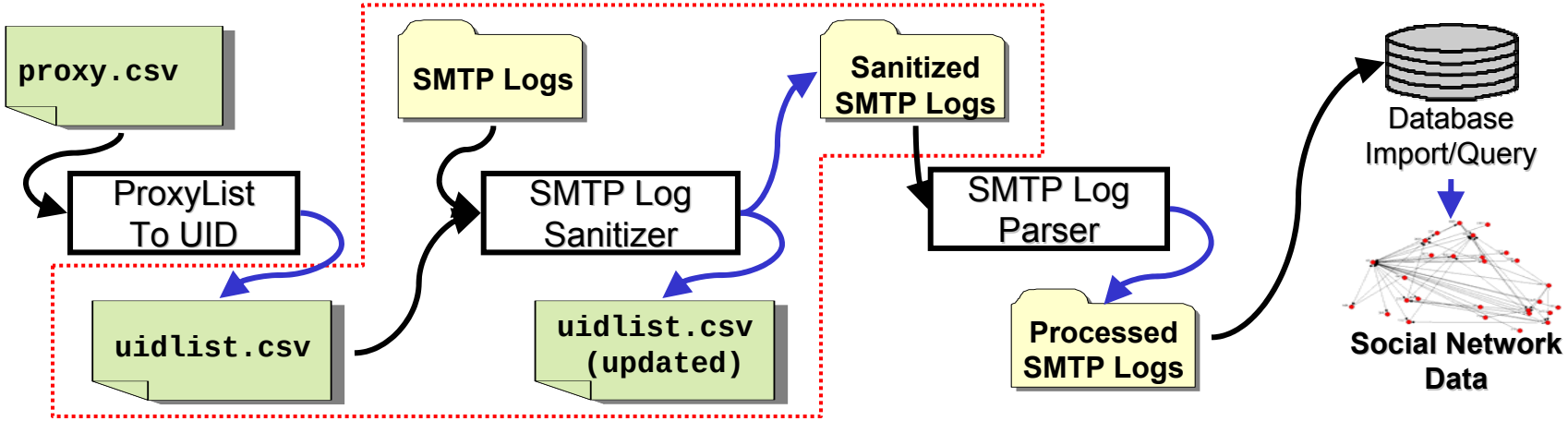

```
Yee Jason 2dLt AFIT/ENG,Jason.Yee@afit.edu,
"X400:c=US;a= ;p=AFIT;o=HANGAR;s=Yee;g=Jason;
smtp:jyee@afit.edu
SMTP:Jason.Yee@afit.edu"
"Smith John A Civ AFIT/SC",John.Smith@afit.edu,
"X400:c=US;a= ;p=AFIT;o=HANGAR;s=Smith;g=John;
smtp:jasmith@afit.edu
smtp:jsmith1@afit.edu
SMTP:John.Smith@afit.edu
SMTP:John.Smith.1@afit.edu"
```

→

```
0,Jason.Yee@afit.edu
0,jyee@afit.edu
1,John.Smith@afit.edu
1,jasmith@afit.edu
1,jsmith1@afit.edu
1,John.Smith.1@afit.edu
```

- Resolves Aliases to Unique Identification Numbers (UIDs)
- Action Attribution

SMTPLogSanitizer



SMTP Log

```
...-?TO:<John.Smith@afit.edu>...
...10 FROM:<jsmith@afit.edu>...
...-?TO:<Jane.User@domain.org>...
```

Sanitized SMTP Log

```
...-?TO:<15@afit.edu>...
...10 FROM:<15@afit.edu>...
...-?TO:<216@domain.org>...
```

UIDList Before

```
...
15,John.Smith@afit.edu
15,jsmith@afit.edu
```

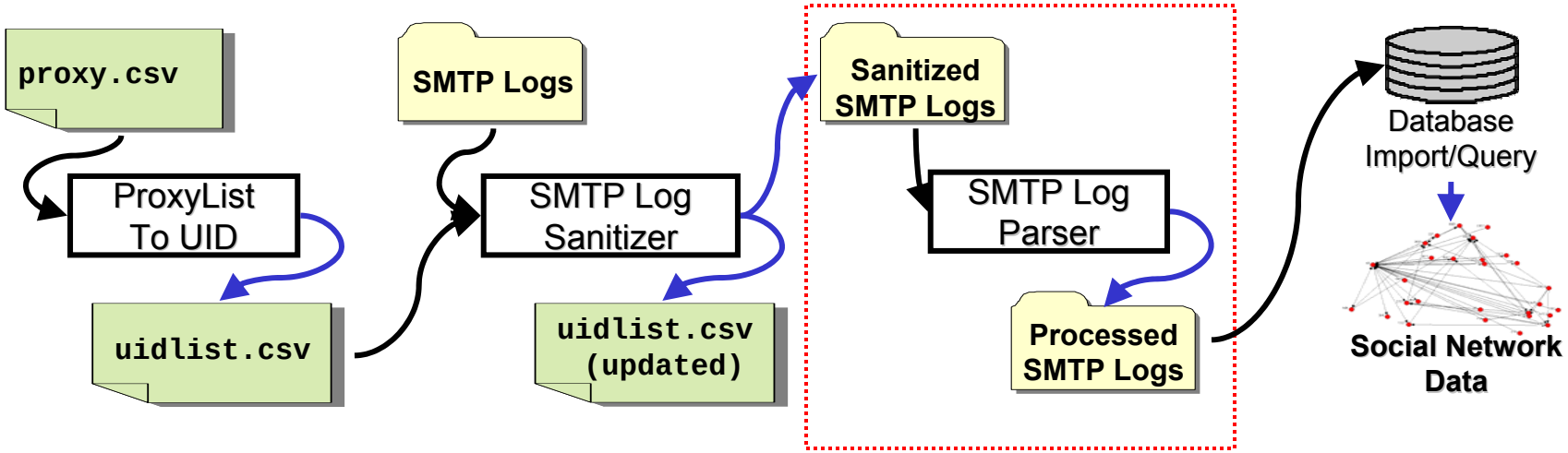
UIDList After

```
...
15,John.Smith@afit.edu
15,jsmith@afit.edu
216,Jane.User@afit.edu
```

- Replaces username portion of email addresses with UIDs
- External parties can determine insider/outsider status

U.S. AIR FORCE

SMTPLogParser




```
129.92.1.65 - OutboundConnectionCommand
[09/Dec/2004:08:01:17 -0500] "MAIL
-?FROM:<2718@afit.edu> SMTP" 0 4
129.92.1.65 - OutboundConnectionResponse
[09/Dec/2004:08:01:17 -0500] "-
-?250 2.1.0 2718@afit.edu....Sender OK SMTP" 0 43
129.92.1.65 - OutboundConnectionCommand
[09/Dec/2004:08:01:17 -0500] "RCPT
-?TO:<1828@afit.edu> SMTP" 0 4
129.92.1.65 - OutboundConnectionCommand
[09/Dec/2004:08:01:17 -0500] "RCPT
-?TO:<4590@ieee.org> SMTP" 0 4
```

→

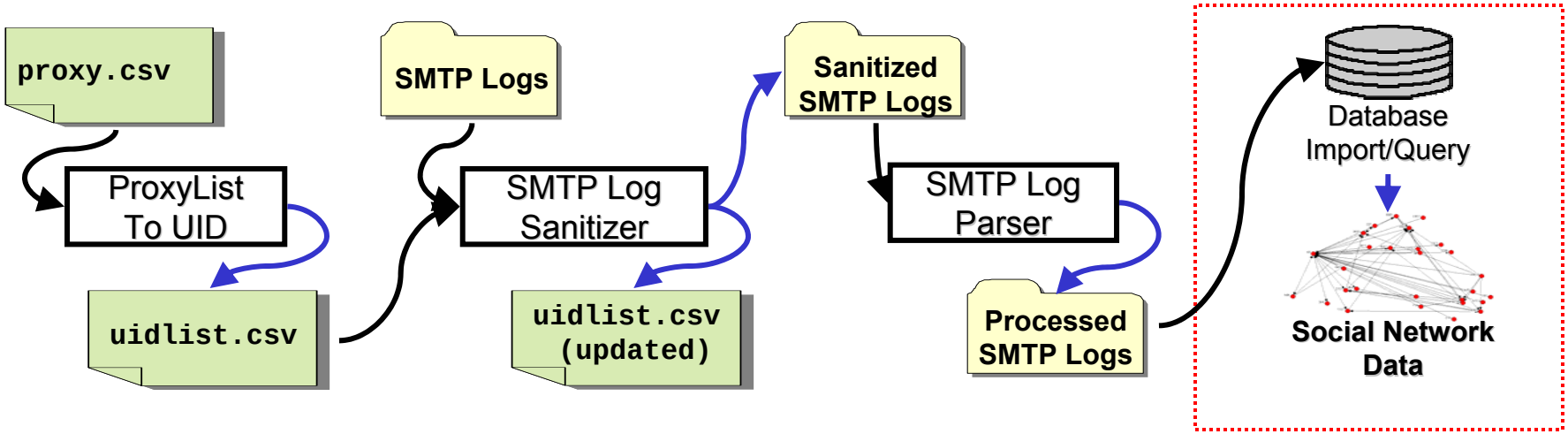
Date	Time	SUID	RUID	SI	RI	NR
2004/12/09	08:01:17	2718	1828	1	1	2
2004/12/09	08:01:17	2718	4590	1	0	2

Two *connections* made

- Process Sanitized Data
- Derive Internal Status and Number of Recipients

U.S. AIR FORCE

Database Functions



Date	Time	SUID	RUID	SI	RI	NR
2004/12/09	07:52:31	162	272	0	1	1
2004/12/09	08:01:17	272	162	1	1	2
2004/12/09	08:01:17	272	314	1	0	2
2004/12/09	08:01:18	314	162	1	1	3
2004/12/09	08:01:18	314	426	1	0	3
2004/12/09	08:01:18	314	272	1	0	3
2004/12/09	08:02:57	162	272	0	1	1
2004/12/09	08:11:33	314	162	1	1	1

→

Source	Dest	TieStrength
162	272	2
272	162	1
272	314	1
314	162	2
314	272	1
314	426	1
...		

- Import processed logs
- Generate social network data from logs

Experimental Methodology

- Operational Testing
  - Add Months Sequentially
- Use data gathered from AFIT
  - Oct-Dec 2004
- Parameters
  - XP Professional SP2
  - Pentium 4, 3.2 GHz HT, 2 GB RAM
  - Java 1.5
  - MySQL 4.1
- Direct measurement

Workload

- Proxy List from Active Directory
- SMTP logs from AFIT servers
  - Medium-sized organization (over 1500 users)
  - 86 days, over 3 GB of data
  - 1.8 million email messages, 3 million connections
  - 1550 internal actors

Month	Days	Size
October	25	1,034,273 KB
November	30	1,385,098 KB
December	31	1,359,860 KB

Timeliness Results

Component	Runtime
ProxyListToUID	< 1 min
SMTPLogSanitizer	54 min
SMTPLogParser	24 min
Database Import	< 1 min
SQL Query	< 1 min
Total (three months)	$\approx$ 80 min

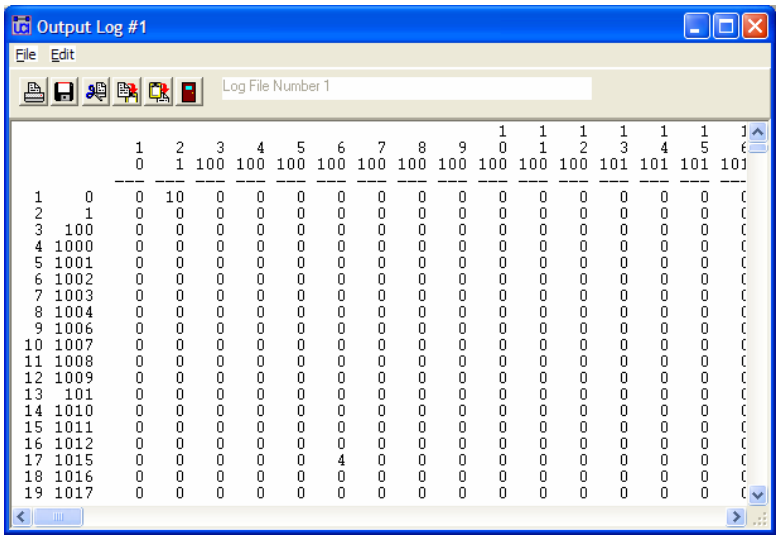
Component	Runtime
ProxyListToUID	< 1 min
SMTPLogSanitizer	$\approx$20 min
SMTPLogParser	$\approx$10 min
Database Import	< 1 min
SQL Query	< 1 min
Total (one month)	$\approx$ 30 min

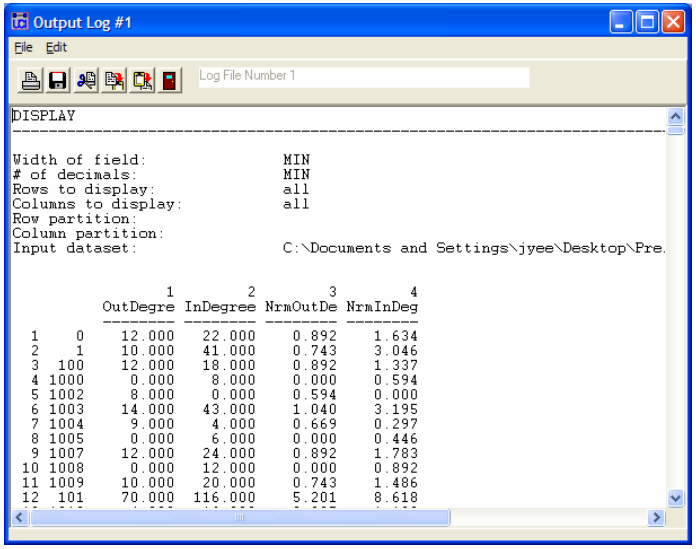
- **Fast**
- **Bottlenecks**
  - **Sanitization**
  - **Parsing (parallelizable)**

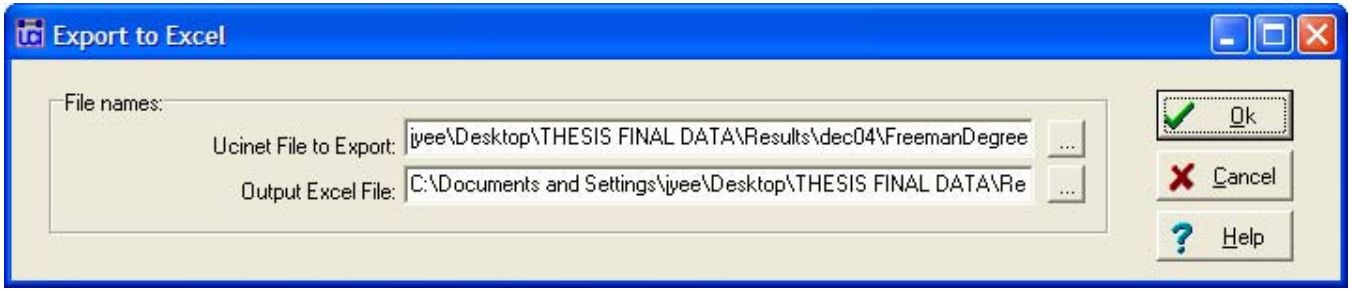
U.S. AIR FORCE

Usability Results

- **UCINet Readable**
  - **Centrality, Power, and Group Statistics Taken**



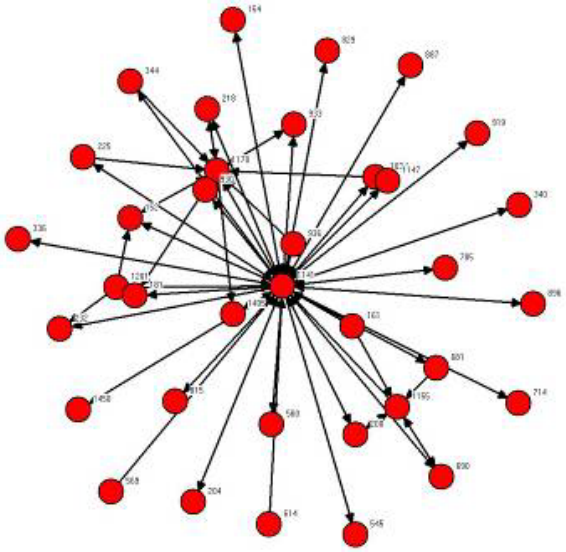


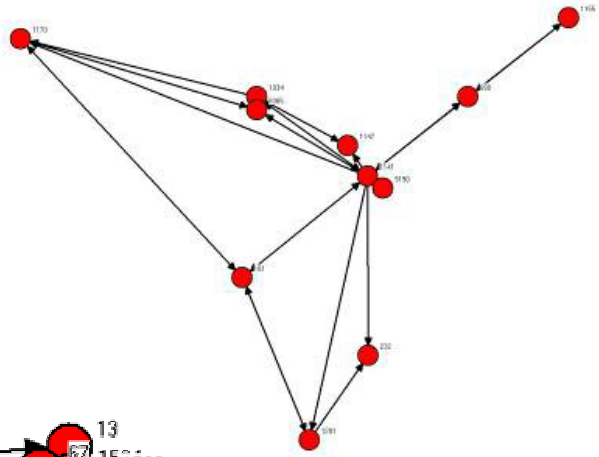


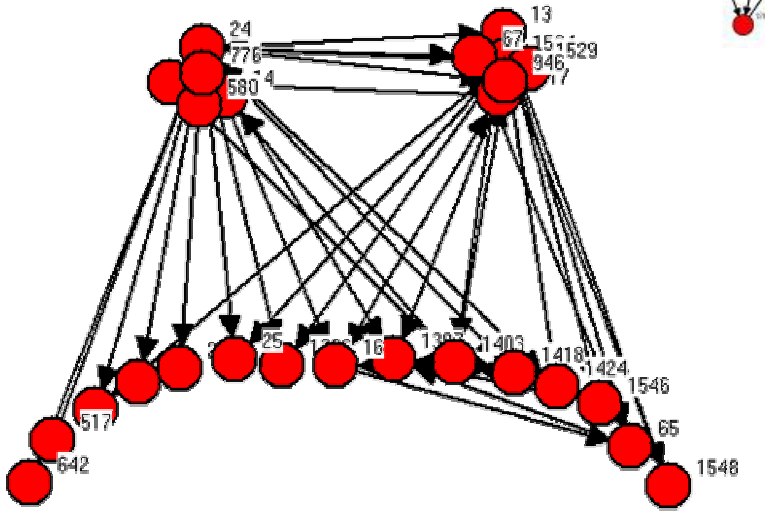
U.S. AIR FORCE

Usability Results

- NetDraw Visualization Tool



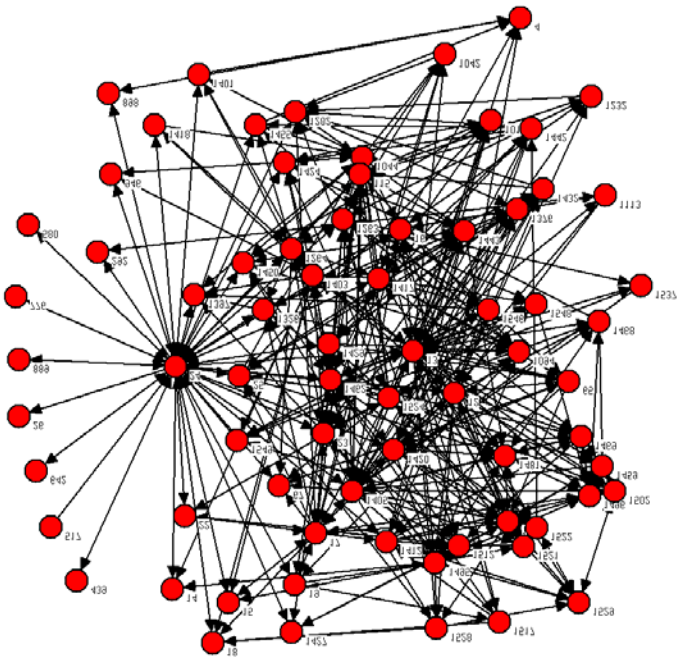




U.S. AIR FORCE

Impact

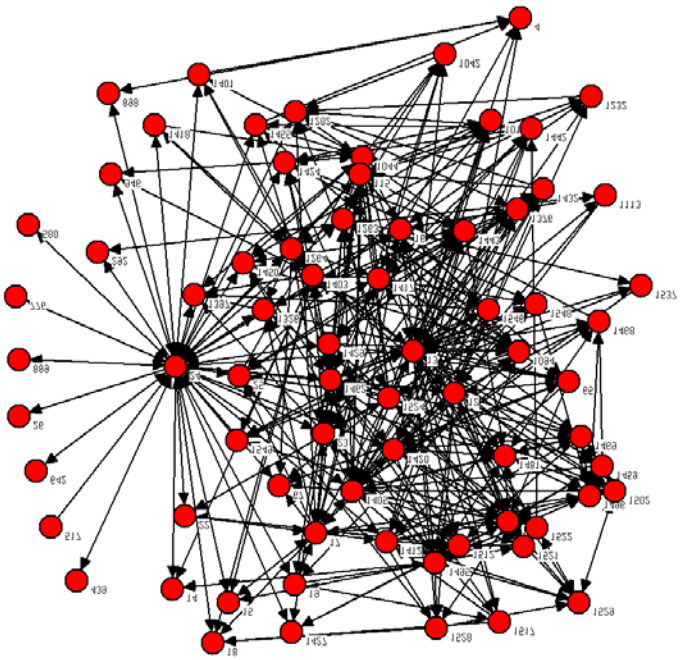
- Immediate
  - Time and Cost Reduction
  - More Data for Social Network Analysis
    - Sanitization
  - More Information for Managers
- Long Term
  - Long term studies
  - Understanding Employees
  - Potential Insider Threat Characterization
  - Possible Insider Threat Mitigation



U.S. AIR FORCE

Limitations

- Presumes that e-mail behavior adequately captures social interactions
- User “consent” to monitoring
  - Most organizations have an acceptable use policy
  - May routinely monitor employee communications
  - However, use of the system is somewhat voluntary



U.S. AIR FORCE

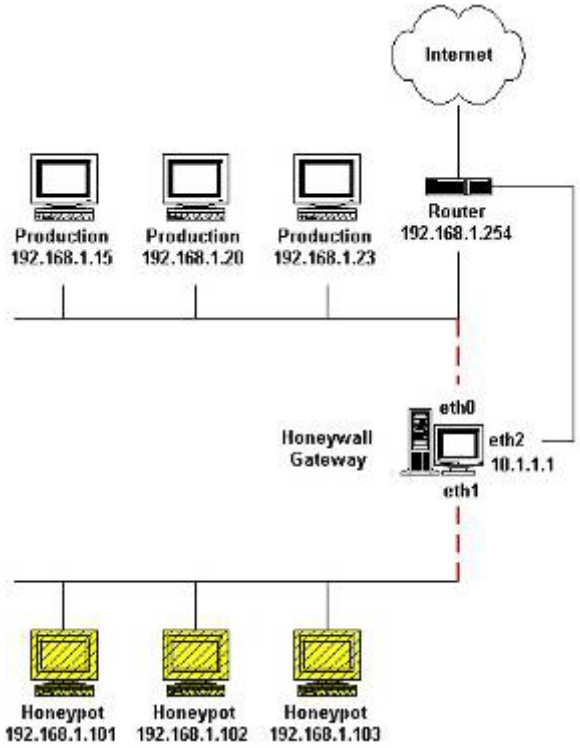
Future Research

- Data Collection
  - Extend to Other CMC Records – IM, Blogs, web browsing
  - Content of messages (subject, nature, contents, “reply-to”
- Analyze Social Network Data
  - Validation – does e-mail actually capture what we are hoping?
  - Analysis of collected data – longitudinal studies
  - Insider threat research, staff collaboration,
- Tools
  - Parallelization

Backup Slides

U.S. AIR FORCE

Insider Threat Mitigation

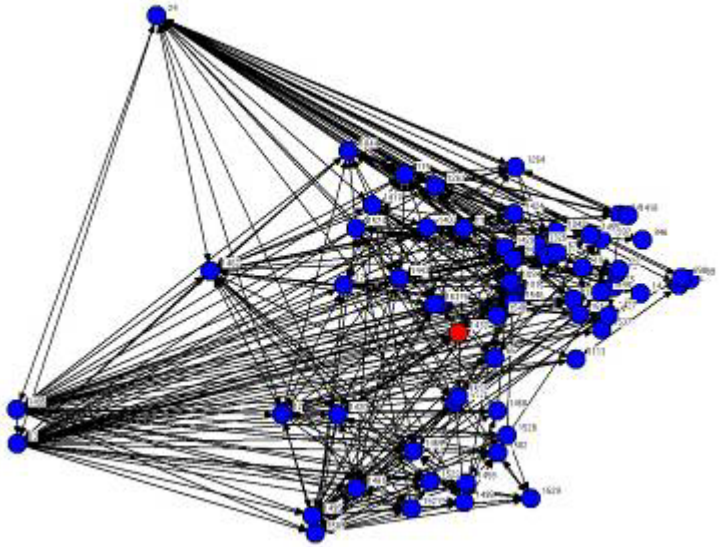


- Intrusion Detection Systems
  - Misuse Detection
- Honeypots
  - Detect Attacks on False Data
  - Studying Attackers
- Prevention
  - Leadership/Management
  - Policies
  - Deterrence
- Combination of the above
  - Defense in depth

U.S. AIR FORCE

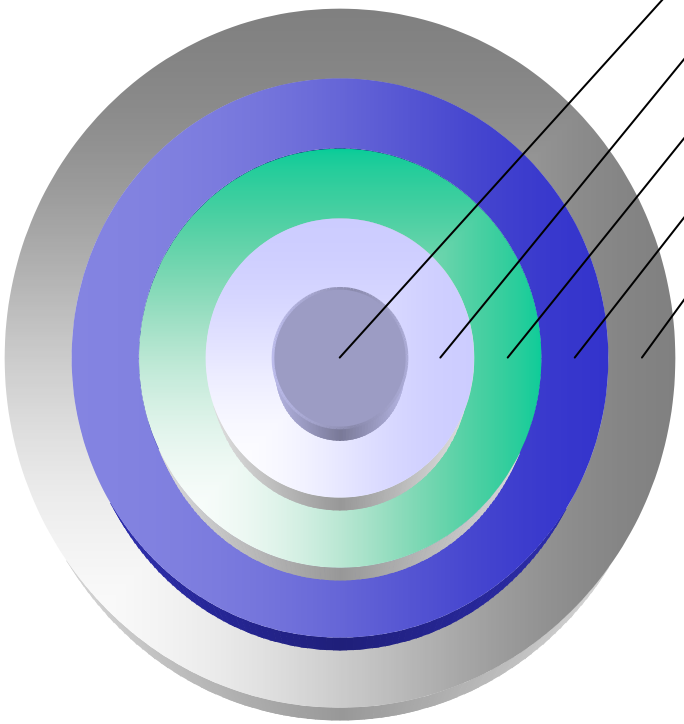
Insider Threat Profile

- System Knowledge, Privileges
- No attribute-based profile
  - Gender, age, background, marital status, position, income
  - No technical expertise needed
- Trends
  - *Planning* → Preventable
  - Behavior Change
- Behavior-based Profiling



U.S. AIR FORCE

Potential Insider Threat Mitigation



Insider Threat Mitigation

Information Tool for Managers

Understanding Employee Behavior

Social Network Analysis Of Real Data

Social Network Data Collection