Discovery of Email Communication Networks from the Enron Corpus with a Genetic Algorithm using Social Network Analysis

Garnett Wilson and Wolfgang Banzhaf

During the legal investigation of Enron Corporation, the U.S. Federal Regulatory Commission (FERC) made public a substantial data set of the company's internal corporate emails. This work presents a genetic algorithm (GA) approach to social network analysis (SNA) using the Enron corpus. Three SNA metrics, degree, density, and proximity prestige, were applied to the detection of networks of high activity and presence of important actors with respect to email transactions. Quantitative analysis revealed that density and proximity prestige captured networks of high activity more so than degree. Subsequent qualitative analysis reveals that there are trade-offs in the selection of SNA metrics. Examination of the discovered social networks revealed that density and proximity prestige isolated networks involving key actors to a greater extent than degree. In particular, density picked out interesting patterns in terms of email volume, while proximity prestige better isolated key actors at Enron. The roles of the particular actors picked out by the networks as reasons for their prominence are also discussed.

I. Introduction

Enron corporation filed for bankruptcy in December 2001 due to a mixture of corruption, fraudulent accounting, and poor regulation. Once one of the world's largest electricity and natural gas companies with over 22 000 employess worldwide [1], its stock plummeted from heights of over $90 a share to $0.05 (Sept. 2003) during the ensuing scandal [2]. The Enron data set of emails was made public by the U.S. Federal Regulatory Commission (FERC) during the legal investigation of Enron. This original corpus disclosed to the public included over 619 446 email messages belonging to 158 users that were sent or saved in folders during the collapse of Enron [3], spanning from 1998 to 2002 [4]. The corpus is now considered a valuable resource for research in link analysis, social network analysis, and natural language processing. This paper presents a genetic algorithm (GA) approach to analysis of emails during the fall of Enron that uses fitness metrics from the field of social network analysis (SNA) to rank the email activity of Enron employees and examine the possible key players at Enron. This work thus presents the first use (to the authors knowledge) of the Enron corpus in the field of evolutionary computation (EC), and the first instance of social network analysis (SNA) used with an evolutionary algorithm. Through the use of a GA with a SNA-based fitness measure, all senders and receivers of the Enron data set can be examined for relationships. Previous techniques have largely restricted SNA analysis of the Enron data set to the 151 employees who had sent or stored the emails.

The following section examines related previous work. Section 3 explains the creation and composition of the data set used in our experiments, and Section 4 describes the SNA measures used by the GA for examination of the emailing activities of the Enron employees. Section 5 provides details on the SNA-based GA and experiment parameterization. Section 6 provides quantitative results, and visualizations of the final networks corresponding to the highest value for each SNA metric are then examined in Section 7. Conclusions follow in Section 8.

II. Previous Work

The Enron data set has been used extensively for research including data mining, text analysis, and natural language processing. To provide a few examples, Berry and Browne [5] detected topics and clustered messages using sparse term-by-message matrices and a low rank non-negative matrix factorization algorithm. Priebe et al. [6] used a technique called "scan statistics," which slides a moving window over portions of data to find outlying points corresponding to deviations from normal communications among individuals. Keila and Skillicorn [7] examined structural features of emails, such as message length and word usage and frequency, to detect patterns of unusual communication.

The Enron corpus has also been subject to SNA analyses of varying rigor. Shetty and Abidi [4] produced a social network from their cleaned version of the data set involving the 151 employee accounts possessing the email accounts, where a link was only established between employees if 5 or more emails were exchanged and the exchange of emails was reciprocal. Chapanoid et al. [8] produce both a directed and undirected social network from the data set where a link is only considered if 30 or more emails have been exchanged between any of 150 employees with the accounts where the emails were stored. (Cleaning or relevance-based design decisions cause the number of main employees considered to range from 147 to 151 throughout the literature.) Chapanoid et al. provide particular SNA measures on the graphs including degree, distance, and betweenness. McCallum et

This work was supported by a PRECARN Postdoctoral Fellowship and Memorial University of Newfoundland.

G. Wilson is with the Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, A1B 3X5, Canada (e-mail: gwilson@cs.mun.ca).

W. Banzhaf is head of the Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, A1B 3X5, Canada (e-mail: banzhaf@cs.mun.ca).

al. [9] describe an ART (Author-Recipient-Topic) model based on a Bayesian network to simultaneously model message content and the directed social network in which the messages are sent. The authors only consider a social network among pre-selected employees in particular divisions of the Enron company, with those being a subset of a total of 147 from the main 151 employees. Duan et al. [10] use a social network only involving 150 of the employees that were the original focus of the dataset, with edges weighted according to the emails sent between users. They use a link analysis algorithm to rank those 150 employees according to their implied importance based on email communication. Diesener et al. [11] provided a more in-depth social network-based examination of the network across the time frame of the emails included in the corpus. Expanding on previous work, they included some senders of emails as graph nodes even if they were not members of the list of Enron employees whose emails constituted the data set. In particular, they added 525 previously unaccounted employees of Enron and Andersen (the accountancy firm associated with the Enron scandal) and their associated email addresses to raise the number of email addresses considered from 151 to 1 234. As in some previous work, they also weighted the edges according to the number of emails sent and used a directed graph (digraph). They found that during the actual crisis period, communication across employees became more diverse in terms of contacts and corporate roles and previously disconnected employees established ongoing mutual communication. The authors also found that interpersonal communication generally intensified and spread widely throughout the network as the collapse of the company progressed, leading to increased density of the network. Frantz and Carley [12] examine 20 000 actors in the Enron database using 165 weekly snapshots of their activity. The authors report on the five SNA metrics of betweenness, degree (in and out), closeness, and eigenvector.

This work goes beyond previous SNA-based analyses of the Enron corpus, as the work introduces intelligent search appropriate for very large search spaces. Most researchers [4, 8, 9] only analyze a social network based on the approximately 151 primary employees of the Enron data set. Deisener et al. [11] and Frantz and Carley [12] were the only researchers to use more than the primary 151 employees in the data set. Deisener et al. added only a select number of extra individuals they felt were pertinent, whereas Frantz and Carley examined 20 000 email addresses, but produced manageable social networks by only examining weekly snapshots of the emailing behaviour. Rather than addition of arbitrary individuals to the social network or using short windows of time, the SNA GA introduced in this work uses an evolutionary approach to intelligently add appropriate individuals selected from the entire set of sender or recipient addresses in the Enron database to a selected social subnetwork. That is, the GA search mechanism is efficient at exploring the search space of possible networks by retaining the most relevant partitions of account networks (called "building blocks" in evolutionary computation models) and then attempting to add associated accounts of interest in an intelligent search. The algorithm builds networks of interest by keeping interesting sub-graphs as building blocks in the evolutionary process and applying mutation to explore other connections to produce a sub-graph of even greater interest.

A computational intelligence approach such as this is an appropriate alternative to exhaustive search or selection of arbitrary accounts to determine SNA networks of interest given the size of the data set when all senders and receivers are considered. The search space corresponding to the Enron data set used in this work (that of Shetty and Adibi [4]) involves 17 568 unique addresses and 252 759 emails which link them. For instance, in this work we examine networks of interest featuring 50 unique emails (up to 100 nodes) in a GA individual. Given a total of 252 759 unique emails, this is a search space of 252 759 choose 50, or $\frac{252\,759!}{50!(252\,759-50)!}$, which is greater than 10^{300}. The SNA GA approach of this work provides a way of isolating areas of the entire network of the Enron database participants (senders and recipients) for further inspection, as the data set in its entirety could not be searched exhaustively , let alone comprehended or analyzed for patterns by a user. By using evolutionary computation, all participants in the Enron data set and their behaviour can be examined.

While Langdon et al. [13] and Luthi et al. [14] have both used social network analysis metrics to examine the relationship among authors in the genetic programming (GP) community, to the authors' knowledge evolutionary computation techniques themselves have not yet been applied to social network analysis. Graph theory and link analysis, however, are well-established tools for the detection of fraud across a number of domains. For instance, Galloway and Simoff [15] describe the use of the NetMap commercial software tool to determine an actual case of insurance fraud. Data mining techniques have been used along with link analysis search for fraud detection. Cortes *et al.* [16] developed a method to analyze large dynamic graphs of telecommunications transactions to identify fraud by examining small subgraphs of interest called "communities of interest." Rather than using social network analysis metrics to determine subnetworks of interest, the authors examine subgraphs consisting of nodes connected by the highest-valued edges to a central node within some low radius. This work is the first instance of a GA using SNA measures as a fitness metric.

III. Data Set

The Enron corpus used in this work is that provided by Shetty and Adibi [4]. The dataset was cleaned by the authors to remove duplicate emails, messages determined to contain junk data, blank messages, and emails generated by the mail system as email transaction failures. A MySql database was formed from the cleaned corpus, and contained four tables for the entities of employees, messages, recipients, and reference information. For the purposes of

reproducibility, the names used to denote tables and fields are those used in the MySQL database Shetty and Abidi [4] provide for download on the associated web site. In total, the dataset contained 252 759 messages (in the *message* table) stored in the folders of 151 Enron employees (*employeeList* table), with a total of 17 568 distinct senders for all messages (including the 151 employees in the *employeeList* table).

For the purposes of the SNA-based GA algorithm, a particular employee ID (*eid*) is select by a GA instruction to specify an employee from the *employeeList* table containing 151 employee entities. From the *recipientInfo* table, all messages (*mid*) for which the employee email (*Email_id*) address corresponding to *eid* was the recipient (*rvalue*) are retrieved, that is, messages are retrieved where *rvalue* = *Email_id*. Finally, from these emails, a particular email from the *message* table is specified by the GA (as an algorithm defined value of *mid*). The sender of this message (*sender*) is then retrieved from the *message* table to provide an employee, sender pair joined by an email transaction. The selection of the relevant database entities specified by the GA is depicted in Figure 1. For faster retrieval, a hash table with recipient emails (*Email_id*) as keys, and a vector of the sender email addresses (*sender*) for each message as values, is created. This hashtable is kept in memory during execution. The additional details of the interpretation of a GA instruction are provided in Section 5.

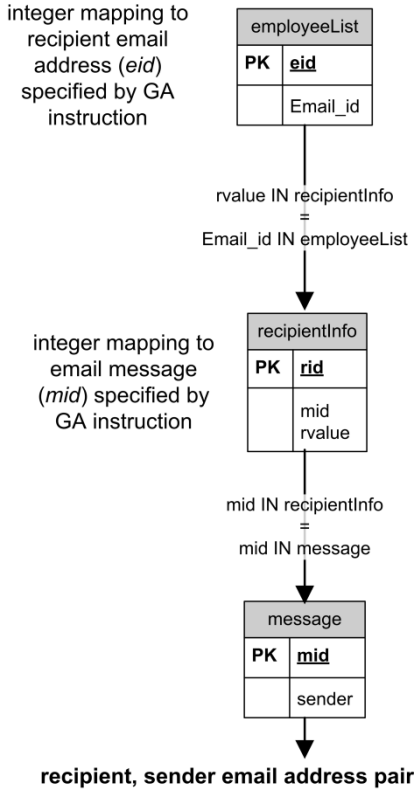

Fig. 1. Database entities used by the SNA GA system to create an employee recipient email address, sender email address pair connected by an email message. Only table columns relevant to this analysis are shown in the diagram.

IV. Social Network Analysis

Three SNA metrics are used to identify networks of interest from the larger social network corresponding to the Enron email corpus: degree, density, and proximity prestige. Degree and density are rooted in the identification of the importance of agents in an interacting network using the SNA concept of *centrality*. Centrality assumes that important actors usually occupy strategic locations in the network without relation to the direction of the transactions (edges). While directed edges are displayed in the network so end users are aware of direction of emails sent, calculations for both these metrics (described in this section) are based on undirected number of messages sent in either direction between two email accounts. Proximity prestige, in contrast, provides the study with a metric that involves the direction of edges in its calculation.

A. Degree

Degree is the average number of edges incident with each node in the graph [17]. For the purpose of AML, it is a measure of the overall transaction-based activity of the nodes in the network. While valued versions of this metric are possible, the non-valued version is used in these experiments to contrast the valued alternative of the density metric (explained immediately in the next section). The mean degree of the network is given by the equation

$$\overline{d} = \frac{\sum_{i=1}^{g} d(n_i)}{g}$$

where g is the number of nodes in the network, n is a node in the graph, and $d(n_i)$ is the degree of a node (number of lines incident with that node). To provide a visualization of the concept of degree in a graph, hypothetical networks of high and low degree are shown in Figure 2.

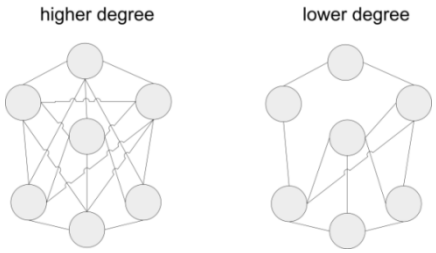

Fig. 2. Hypothetical networks of high and low degree.

B. Density

Density is the proportion of possible lines that are actually present in the graph [17]. For a valued graph, as is the case here where number of emails are construed as edges connecting the accounts construed as nodes, the density (Δ) is measured as

$$\Delta = \frac{\sum v_k}{g(g-1)}$$

where v_k are the values over all k for the values $\{v_1, v_2, ..., v_L\}$ attached to the set of lines (edges) L, and g is the number of nodes. The density metric was expected to perform well in detection interesting email networks, as it measures the interconnectedness of a network while incorporating the number of email transactions. Note that for a non-valued graph the value of the density will fall in the range [0, 1]; however, there is no such restriction on the valued version of the density metric. Hypothetical networks of high and low density are depicted in Figure 3.

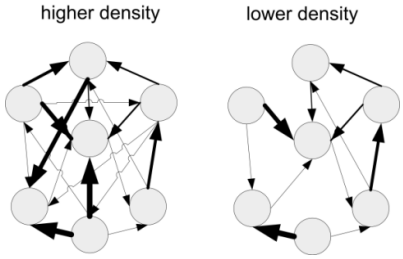

Fig. 3. Hypothetical networks of high and low density.

C. Proximity Prestige

The *proximity prestige* measure is used to determine what is known as the account's *influence domain*, where the influence domain of an actor is the set of actors who are both directly and indirectly linked to that actor. This set includes all actors that are reachable to i, where two nodes are reachable from one another if there exists a *path* between them. A path is a *walk* in which all nodes and lines are distinct, where a walk is a sequence of nodes and lines, starting and ending with nodes, where each node is incident with lines preceding and following it in the sequence. The influence domain for actor i consists of all actors whose entries in the ith column of a distance matrix are finite; the number of actors in the influence domain for an actor i is denoted I_i.

In practice, *proximity prestige* is used to measure closeness using distances to, rather than from, each actor in a directed graph. The proximity measure for the problem domain of this work, from [17], is the ratio of the proportion of actors who can reach i to the average distance of these actors from i. This proximity prestige measure is

$$P_P(n_i) = \frac{I_i/(g-1)}{\sum d(n_j, n_i)/I_i}$$

where I_i is the number of actors in the influence domain of node n_i, g is the total number of nodes in the graph, and $d(n_j, n_i)$ is the *distance* that actor j is from actor i. *Distance*, or more precisely *geodesic distance*, between nodes n_i and n_j is found by using power matrices. Distance from one node to another is simply the length of the shortest path between them, where in a directed graph distances from n_i to n_j and n_j to n_i can be different. The length of the shortest path (the distance) from n_i to n_j is the first integer power p of the original sociomatrix (with $p = 1$) for which the (i,j) element is non-zero:

$$d(n_i, n_j) = min_p x_{ij}^{[p]} > 0 \text{ , p > 0}$$

When actors who can reach actor i become closer, then the P_p ratio becomes larger and actor i has greater prestige. P_p has a maximum value of 1 when all actors are adjacent to n_i, and a minimum value of 0 when n_i is unreachable. A group level measure of proximity prestige is simply the average of actor proximities:

$$\bar{P}p = \sum_{i=1}^{g} \frac{P_P(n_i)}{g}$$

A network with individual members of high and low prestige is shown in Figure 4 to provide a visualization of the concept.

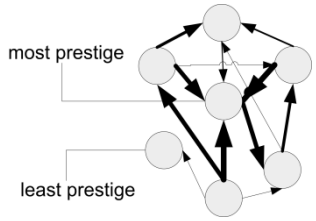

Fig. 4. A hypothetical network containing members of high and low proximity prestige.

V. Genetic Algorithm and Experimental Setup

A. Individual Representation

The individuals in the genetic algorithm consist of binary strings composed of 50 instructions, with each instruction consisting of 22 bits. The first 8 bits of an instruction correspond to the database entity that is to be queried, which is one of the employees in the employee list table. The binary number corresponding to the first 8 bits is interpreted as its integer equivalent, modulo the number of entities (primary keys) in the database (151). The largest number of emails sent to an individual in the employee table is 9 052, which can be specified by 14 bits (2^{14} = 16 384). The following 14 bits thus map to an email sent to the individual specified in the first portion of the instruction, modulo the number of emails present for that entity. In particular, for the purpose of building the SNA network, the main sender of the received email is of interest: all individuals CCed are not considered per individual instruction. (However, an individual that receives an email in virture of being CCed can appear in an SNA graph by being specified as the receiver in the first segment of an instruction.) Each

instruction thus corresponds to a receiver, sender pair in the real world. The entire GA individual is a list of such pairs, which represents a network of possibly interacting email accounts (nodes) connected by sent emails (edges).

The 50 instructions comprising an individual are converted to a *sociomatrix* to allow fitness evaluation by the GA in terms of the SNA metrics. A sociomatrix is a matrix indexed by the set of originating actors (rows) and the set of receiving actors (columns). The total unique emails between the originating accounts and receiving accounts are the sociomatrix values for the ties between the actors. The actors in rows or columns can be any of the 151 employees (receivers) from the main employee table or any of the other total 17 417 senders who emailed messages to the 151 employees (for a total 17 569 email addresses).

Once an individual sociomatrix of interest is found by the GA, it is displayed as its equivalent graph of nodes and edges. The set of actors (all sending or receiving accounts) are used as the nodes of the graph, and the set of email messages is used as the collection of edges. Edges are weighted according to number of unique emails sent from one actor to the other, where edge thickness grows in integer increments of one email. The interpretations of a simplified, hypothetical GA individual are shown in Figure 5.

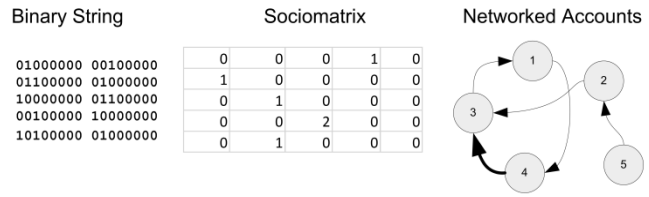

Fig. 5. Interpretation of a GA individual as bit string, sociomatrix, and network of email senders and recipients.

B. Algorithm Details

Tournaments were steady state and consisted of 1500 rounds, which was determined to be of reasonable duration across all three chosen SNA metrics based on preliminary experiments. During each tournament round, four individuals were chosen to compete. In the competition, each binary string individual is interpreted by querying the database and producing a sociomatrix corresponding to transactions as described in the previous section. The top two individuals' sociomatrices are declared winners, and the binary strings (genotype) of the last two individuals are replaced with the winning individuals' genotype (becoming the children). Following selection, mutation was applied using an XOR mask on an individual instruction segment of the binary string. Mutation occurred with a threshold of 0.9, after choosing a particular instruction from the individual's instruction set using a uniform random distribution. Crossover was not performed, as retaining groups of related instructions as a sequence is not appropriate for this domain: In this representation, each instruction represents potentially unique emails for particular email accounts. Mutation serves to explore new account, email pairings in the candidate networks. At the end of the tournament, the individual with the highest value for the SNA metric based on its sociomatrix is displayed to the user as a directed, cyclic graph with edges weighted to reflect the number of email(s) sent between them. A similar random, greedy search was conducted where the best two of four individuals in a tournament round were kept and the two losing individuals were replaced with randomly generated candidate networks.

Fifty trials were performed using an Apple IMac with an Intel Core 2 Duo CPU 2.8 GHz and 4.00 GB of RAM, running OS X Leopard version 10.5.4. The solution was implemented using Java version 1.6.0 and MySQL. The JAMA library [18] was used for the matrix manipulations required for the evaluation of social network-based fitness metrics. All visualizations were produced using a customized version of the Prefuse framework [19]. The GA parameters used are summarized in Table 1.

TABLE I
EXPERIMENTAL GENETIC ALGORITHM PARAMETERS

Experimental Trials	50
Tournament Rounds	1500 per trial
Instruction format	22 bits total (8 bits account, 14 bits transaction)
Genotype structure	50 instructions per individual
Mutation	instruction-level XOR mask, threshold = 0.9
Fitness metric	degree, density, proximity prestige
Objective	Find network that maximizes metric.

VI. QUANTITATIVE SNA PERFORMANCE

It should be noted that the quantitative metrics used in these networks due not necessarily measure "success" *per se*. The actual benefit of the SNA-based GA search is actually qualitative in nature, where the aim of the search is to provide the end users with subnetworks of interest whereby highly connected or active accounts can be discovered (discussed in Section 7). The quantitative results provide a feel of the performance of the GA, SNA metric combinations in general, and confirm that GA search finds networks corresponding to higher SNA values than a random, greedy alternative (thus GA provides a means of intelligently searching the space). Greedy, exhaustive (as opposed to random) search is not a viable option due to the enormous size of the search space (see Section 2).

The results shown in Figures 6, 7, and 8 indicate the best final degree, density, and proximity prestige measures at the end of each trial for fifty trials of 1500 generations comparing GA and random greedy search. In the boxplot figures, each box indicates the lower quartile, median, and upper quartile values. If the notches of two boxes do not overlap, the medians of the two groups differ at the 0.95 confidence interval. Points represent outliers to whiskers of 1.5 times the interquartile range. To accompany the boxplots, Table 2 provides precise measures for the GA and random greedy search algorithms.

TABLE II
ALGORITHM MEAN, MINIMUM, AND MAXIMUM SNA METRICS

	Genetic Algorithm			Greedy Search		
	Mean	Min	Max	Mean	Min	Max
Degree	1.36	1.19	1.61	0.68	0.70	0.67
Density	0.055	0.040	0.087	0.010	0.011	0.0098
Prestige	0.37	0.32	0.42	0.034	0.041	0.030

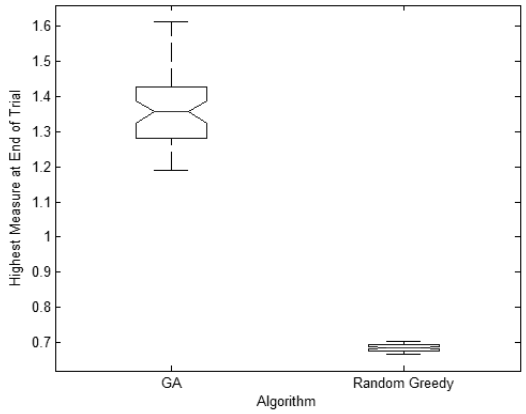

Fig. 6. Boxplot of highest degree found by the GA at the end of 1500 rounds over 50 trials for a social network of up to 100 nodes.

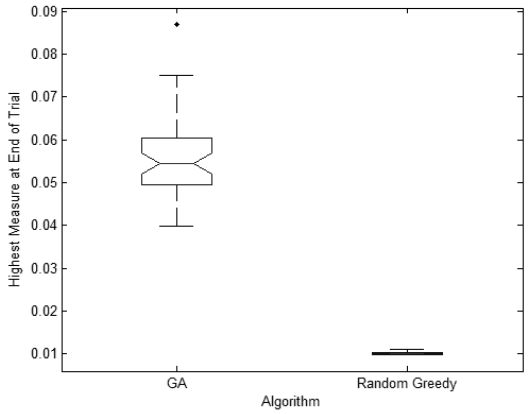

Fig. 7. Boxplot of highest density found by the GA at the end of 1500 rounds over 50 trials for a social network of up to 100 nodes.

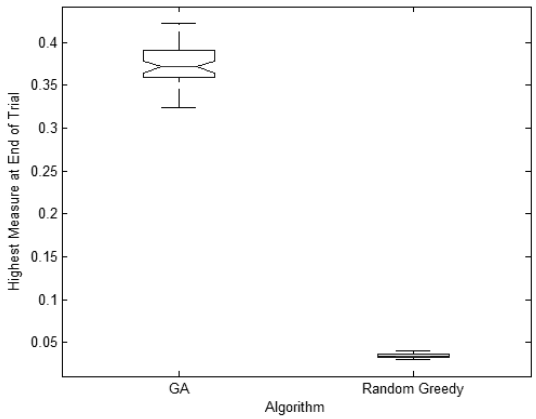

Fig. 8. Boxplot of highest proximity prestige found by the GA at the end of 1500 rounds over 50 trials for a social network of up to 100 nodes.

Examining Figures 6, 7, and 8 it is evident that the SNA-based GA search outperforms random, greedy search with very high statistical significance (no notches of the boxplots overlap or are even in close proximity). Examining Figure 6, we see that the degree measure indicates that, on average, approximately 1.19 to 1.61 edges (Table 2) are attached to each node. This means that the degree metric is finding sparse networks. Density in Figure 7 ranges from 0.040 to a maximum 0.087 (Table 2), indicating that the GA can choose networks that possess up to 9% of *all possible* email connections among nodes. Even though density is valued, the proportions can roughly be construed as percentages when edges represent single emails (as is the case for the 9% best network, Figure 11). In practical terms, this means that the density networks represent high email volume and active communication among individuals. Given that proximity prestige, Figure 8, reflects what proportion of actors can reach a particular actor, 0.32 to 0.42 (Table 2) in the networks discovered by the GA indicate the likely presence of key actors. We now complement these quantitative observations with qualitative analysis of the networks.

VII. QUALITATIVE ANALYSIS: EMAIL NETWORKS

A. Degree

The network interpretation for the best individual (highest fitness level) for the degree metric is shown in Figure 9. The degree metric expresses the overall activity in the network. A high degree (non-valued in this work) network is composed of accounts that are involved in a large number of incoming or outgoing emails to different individuals, independent of number of emails sent between those individuals. The edges are weighted in Figure 9 as an indication of the number of emails sent, but the degree metric is left as non-valued to provide greater contrast with the valued version of the density metric used (see Section 4).

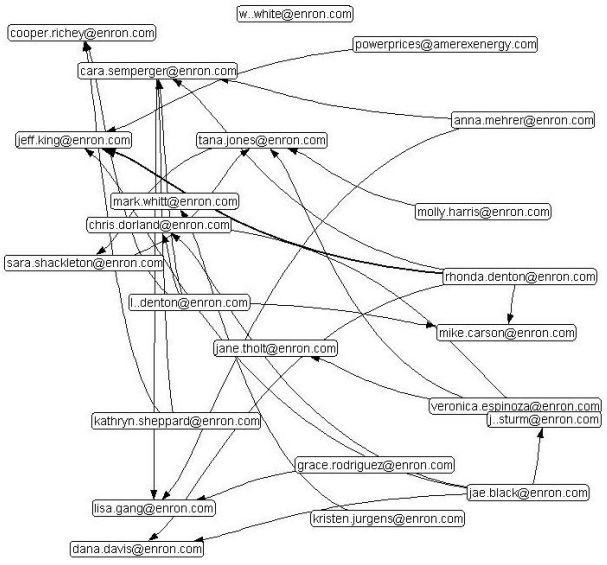

Fig. 9. Highest degree network found by the GA. Thickness of edges reflects number of emails sent.

While this network represents a good attempt at detecting employee email accounts that are quite active, we can see that the metric only picks up one edge that involves more than one email between two parties. While degree can pick up accounts with much activity, it is by chance that it will find individuals that are sending large numbers of emails between one another. That is, the fact that thicker edges are detected in a non-valued degree-based network is a matter of chance.

B. Density

The network interpretation for the best individual (highest fitness level) is shown in Figure 10 for the density metric. This network represents a balance between the number of emails sent among individuals and the overall activity of the email accounts.

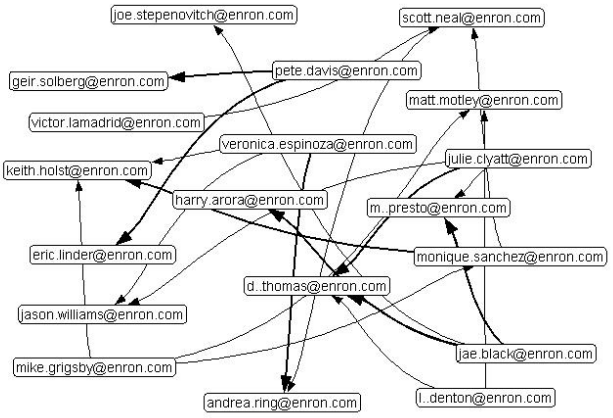

Fig. 10. Highest density network found by the GA. Thickness of edges reflects net amount of transaction.

Comparing Figures 9 and 10, it is readily evident that the density metric finds a network incorporating thinker edges. This corresponds to a network where it can be seen that particular employees are sending larger volumes of email between one another. Since each GA individual / final network consists of a fixed number of email transactions, the presence of the thicker edges for the density metric (Figure 10) results in a network of less nodes compared to degree (Figure 9). In both degree and density, the degree of an individual node ranges from 1 to 4, and there does not actually appear to be a great deal of difference in the average level of email activity per node expressed in each of the networks. (That is, while degree simply has more nodes than density, the individual nodes are not more active in their email activity.)

C. Proximity Prestige

The network interpretation for the highest proximity prestige group is shown in Figure 11. The network features a number of nodes that are connected by to a high number of other nodes with (incoming or outgoing) emails (edges) compared to the density and degree networks. Whereas other networks have nodal degrees of up to 4, there are nodes in this network with nodal degee 5 (*eric.bass@enron.com*), nodal degree 6 (*mike.grigsby@enron.com*), and nodal degree 10 (*veronica.espinoza@enron.com*). Proximity prestige finds a network composed of high activity (individual) nodes, but it does not necessarily pick up large volume connections between individuals as occurs with the density metric (compare Figures 10 and 11). Proximity prestige appears to be the best metric for locating individuals of interest in terms of high degree (active emailing behavior), even better than the degree metric (compare Figures 9 and 11). It is likely that this occurs because the presence of high degree individuals in a GA's candidate network is rewarded implicitly by the group level proximity prestige metric, whereas the group level degree metric rewards a network with high degree overall while not necessarily favoring networks where a few individuals have very high degree.

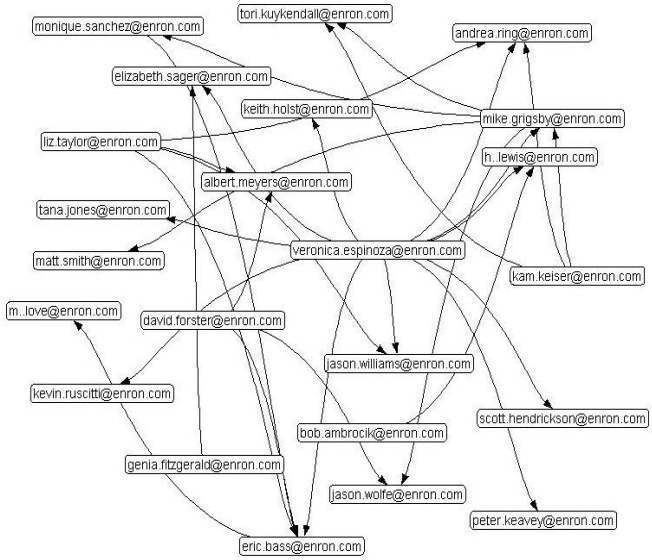

Fig. 11. Highest proximity prestige network found by the GA. Thickness of edges reflects net amount of transaction.

D. Key Actor Analysis

In terms of relevant of individuals picked up by the network, the job titles associated with the email addresses are provided as available from [4] and additional information about employees is otherwise cited. The density metric picks up a number of important actors, one being *mike.grigsby@enron.com*. Grigsby was a manager at Enron, and McCallum et al. [9] in their topic analysis note that he was an active emailer in virtue of being involved in the "sports pool." The density network also picks up two of Enron's directors, *matt.motley@enron.com* and *keith.holst@enron.com* and two Vice Presidents (*joe.stepenovitch@enron.com* and *scott.neal@enron.com*). In contrast, the degree network picks up less nodes of interest. No directors are present, but one vice president (*jane.tholt@enron.com*) is present in the density network. However, the presence of the vice president is only due to a single email connection to a key actor

(*veronica.espinoza@enron.com*).

The proximity prestige network involves the presence of three key actors, plus an additional two traders (*tori.kuykendall@enron.com* and *kevin.ruscitti@enron.com*) and a director (*keith.holst@enron.com*). Proximity prestige is the only metric to pick up any traders in the best network, where this was an important role at Enron. The key players *mike.grigsby@enron.com* (nodal degree 6) and *eric.bass@enron.com* (nodal degree 5) are present in virtue of their participation in the sports pool. As mentioned, Mike Grigsby was a manager who was also prominent in the density network. Eric Bass seems to have been the coordinator of a fantasy football league for Enron employees [9]. The highest nodal degree of 10 in the proximity prestige network belongs to *veronica.espinoza@enron.com*. Her email account is actually present in all three networks (Figures 9, 10, and 11). She is mentioned in the study of Frantz and Carley [12] as one of the top five key actors in their weekly snapshots. Although her official job title was not readily discernable, examination of her emails reveals that she seems to have been active in the administration of "credit worksheets" involved in Enron's trading activities.

VIII. Conclusion

This work presents the analysis of three SNA-based metrics for the examination of the Enron email dataset: degree, density, and proximity prestige. The quantitative results for each SNA metric were compared between a GA and random greedy search. It was found that GA definitively outperformed a random greedy search, meaning that GA is an appropriate application of computational intelligence to search the large Enron emails dataset space. It was expected, based on quantitative analysis, that density and prestige metrics would provide the most useful networks compared to degree, and this was evidenced by qualitative analysis. The social networks corresponding to the highest measure for each SNA metric in the GA were then examined using a visualization tool. Particular SNA metic-based networks provided specialized information, sometimes to the detriment of other metrics. For instance, high measure of the group-level degree metric came at a cost of the absence of individual nodes with higher nodal degree or edges reflecting large numbers of emails transferred. Also, while higher measures of the proximity prestige metric provided nodes of high nodal degree, there were also an absence of edges with multiple email transfers. Overall, the density and proximity prestige SNA metrics were found be useful in isolating the most interesting email accounts and their associated activities. Future work will examine in greater detail the relationship among the three metrics as GA evolution is directed by a particular metric. Also, consensus building using multiple SNA metrics will be explored to mitigate the trade-off costs of one metric over another in order to isolate networks of increased interest.

References

[1] B. McLean and P. Elkind, *The Smartest Guys in the Room: The Amazing Rise and Scandalous Fall of Enron*. New York, USA: Portfolio, 2003.

[2] R. Bryce, *Pipe Dreams: Greed, Ego, and the Death of Enron*. New York, USA: PublicAffairs, 2003.

[3] B. Klimt and Y. Yang, "Introducing the Enron Corpus," in Fifth Conference on Email and Anti-Spam (CEAS) 2008, Mountain View, USA, 2008.

[4] J. Shetty and J. Adibi, "The Enron Email Dataset Database Schema and Brief Statistical Report," in *http://www.isi.edu/~adibi/Enron/Enron.htm*, 2004.

[5] M. Berry and M. Browne, "Email Surveillance Using Non-negative Matrix Factorization," *Computational & Mathematical Organization Theory*, vol. 11, no. 3, pp. 249-264, 2005.

[6] C. Priebe, J. Conroy, D. Marchette, and Y. Park, "Scan Statistics on Enron Graphs," *Computational & Mathematical Organization Theory*, vol. 11, no. 3, pp. 229-247, 2005.

[7] P. Keila and D. Skillicorn, "Structure in the Enron Email Dataset," *Computational & Mathematical Organization Theory*, vol. 11, no. 3, pp. 183-199, 2005.

[8] A. Chapanond, M. Krishnamoorthy, and B. l. Yener, "Graph Theoretic and Spectral Analysis of Enron Email Data," *Computational & Mathematical Organization Theory*, vol. 11, no. 3, pp. 265-281, 2005.

[9] A. McCallum, X. Wang, and A. Corrade-Emmanuel, "Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email," *Journal of Artificial Intelligence Research*, no. 30, pp. 249-272, 2007.

[10] Y. Duan, J. Wang, M. Kam, and J. Canny, "A Secure Online Algorithm for Link Analysis on Weighted Graph," in Proceedings of the Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005, Newport Beach, 2005, pp. 71-81.

[11] J. Diesner, T. Frantz, and K. Carley, "Communication Networks from the Enron Email Corpus ?It's Always About the People. Enron is no Different?," *Computational & Mathematical Organization Theory*, vol. 11, no. 3, pp. 201-228, 2005.

[12] T. Frantz and K. Carley.Dynamics of Organizational Chatter. presented at North American Association for Computational Social and Organization Sciences (NAACSOS) 2006.[PowerPoint]. Available: http://www.cs.cmu.edu/~terrill/docs/NAACSOS2006_Frantz_OrganizationChatter_Presentation.pdf

[13] W. B. Langdon, R. Poli, and W. Banzhaf, "An eigen analysis of the GP community," *Genetic Programming and Evolvable Machines*, pp. upcoming, 2008.

[14] L. Luthi, M. Tomassini, and M. Giacobini, "The Genetic Programming Collaboration Network and its Communities," in Proceedings of GECCO 2007, London, UK, 2007, pp. 1643-1650.

[15] J. Galloway and S. Simoff, "Network Data Mining: Discovering Patterns of Interaction Between Attributes," in Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) 2006, Singapore, 2006.

[16] C. Cortes, D. Pregibon, and C. Volinsky, "Computational Methods for Dynamic Graphs," *Journal of Computational and Graphical Statistics*, vol. 12, pp. 950-970, 2003.

[17] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. New York, USA: Cambridge University Press, 1999.

[18] J. Hicklin, C. Moler, P. Webb, R. Boisvert, B. Miller, R. Pozo, and K. Remington, "JAMA: A Java Matrix Package," in *http://math.nist.gov/javanumerics/jama/*, 2008.

[19] J. Heer, S. K. Card, and J. Landay, "Prefuse: A Toolkit for Interactive Information Visualization," in Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2005), Portland, USA, 2005, pp. 421-430.