# Automatic Text Analysis of Values in the Enron Email Dataset: Clustering a Social Network Using the Value Patterns of Actors

Yingjie Zhou
Barclaycard
yingjie.zhou@barclaycard.co.uk

Kenneth R. Fleischmann
University of Maryland
kfleisch@umd.edu

William A. Wallace
Rensselaer Polytechnic Institute
wallaw@rpi.edu

## Abstract

*This paper describes an automatic text analysis of values contained in the Enron email dataset that seeks to explore the potential to apply value patterns to cluster a social network. Two hypotheses are posed: individuals communicate more frequently with other individuals who share similar value patterns than with individuals with different value patterns; and people who communicate more frequently with each other share similar value patterns. The first hypothesis is supported: indeed, individuals were found to communicate more frequently with individuals who share similar value patterns, and further, the extent to which this is true appears to depend at least in part on the value patterns themselves. However, the second hypothesis is not supported – people who communicate more frequently with each other do not necessarily all fit into a particular value type. Thus, values have utility as a novel tool for social network analysis.*

## 1. Introduction

This paper poses the research question: is there a mutually shaping relationship between value patterns and communication patterns within a social network? That is, do value patterns shape communication patterns, and do communication patterns shape value patterns? The objective of this paper is to test two specific hypotheses based on the two parts of the research question: first, individuals communicate more frequently with other individuals who share similar value patterns than with individuals with different value patterns; and second, people who communicate more frequently with each other share similar value patterns. Testing these hypotheses will further understanding of the relationship between values and communication networks.

This paper focuses on automatic text analysis of values embedded in the Enron email dataset. The word count method is used to count the frequencies of a bag of keywords in each email. Since those emails are sent by a group of individuals, the frequencies of the bag of keywords can be aggregated into the individual level. As a result, each individual has a pattern of word usage in terms of these predefined keywords. If two persons have similar patterns of word usage, it is said they have similarity in certain characteristics depending on the meaning of the keywords. It is important to detect the relationship between people's characteristics and their communication frequencies, i.e., whether people sharing similar characteristics communicate more frequently; and whether people communicating more frequently share similar characteristics.

Studies show that word usage in people's writing and speaking correlates with their personalities [7, 11], and groups of people with distinct interests use words differently [5]. This study focuses on one particular aspect of personalities and interests: values. A value can be defined as "what a person or group of people consider important in life" [10, p. 349]. This paper describes an automatic text analysis of the relationship between values and communication. Individuals' values are measured through an analysis of their word use in their emails. Individuals' communication is measured through an analysis of their email network. Emails are used to form a work, professional, or friendship relationship between the sender and the recipient. An email network is one type of social network with people who send and/or receive emails as nodes and the email messages themselves as links [9]. Current technologies allow researchers to study huge email corpora [2, 4].

The study starts from value words selection, i.e., the bag of words of interest has to be defined first. The word frequencies in each email are counted by parsing the email content with these words. The word frequencies in each email are then aggregated according to the senders of the emails. Individuals' values are measured by this vector of word frequencies. The Enron email dataset constitutes a communication network with individuals as the nodes and emails from one individual to another as the links. Individuals' communication is measured

IEEE computer society

978-0-7695-3869-3/10 $26.00 © 2010 IEEE

through an analysis of this communication network. The detailed discussion of these analyses is described in the following sections.

## 2. Building the bags of words for values

The first step is to select the words for analysis. Although various dictionaries have been developed for a large number of categories – the LIWC2007 Dictionary, for example, is composed of almost 4,500 words and word stems for processing [15] – the words are not readily available to be considered as "values". Schwartz [17] developed ten value categories (power, achievement, hedonism, stimulation, self-direction, universalism, benevolence, tradition, conformity, security). Schwartz's ten value categories are used for analysis in this paper. The dictionary and statistical analysis described in this paper are a novel contribution that can be applied to other studies that use text analysis to study values.

Schwartz defined values as "desirable transsituational goals, varying in importance, that serve as guiding principles in the life of a person or other social entity" [17]. He developed ten motivational types of values and their definitions are as follows [17, p. 22]:

- **Power:** Social status and prestige, control or dominance over people and resources;
- **Achievement:** Personal success through demonstrating competence according to social standards;
- **Hedonism:** Pleasure and sensuous gratification for oneself;
- **Stimulation:** Excitement, novelty, and challenge in life;
- **Self-direction:** Independent thought and action – choosing, creating, exploring;
- **Universalism:** Understanding, appreciation, tolerance, and protection for the welfare of all people and for nature;
- **Benevolence:** Preservation and enhancement of the welfare of people with whom one is in frequent personal contact;
- **Tradition:** Respect, commitment, and acceptance of the customs and ideas that traditional culture or religion provide;
- **Conformity:** Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms; and
- **Security:** Safety, harmony, and stability of society, or relationships, and of self

Schwartz's ten value types are further broken into 56 values [17]. They were categorized into the ten value types applying a multidimensional scaling technique to survey data from 44 countries. The 56 values together with the ten value types they belong to are given as follows [17, p. 33]:

**Power (Pow)** social power, authority, wealth, preserving my public image, social recognition;

**Achievement (Ach)** successful, capable, ambitious, influential, intelligent, self-respect;

**Hedonism (Hed)** pleasure, enjoying life;

**Stimulation (Sti)** daring, a varied life, an exciting life;

**Self-direction (Sel)** creativity, curious, freedom, choosing own goals, independent;

**Universalism (Uni)** protecting the environment, a world of beauty, unity with nature, broad-minded, social justice, wisdom, equality, a world at peace, inner harmony;

**Benevolence (Ben)** helpful, honest, forgiving, loyal, responsible, true friendship, a spiritual life, mature love, meaning in life;

**Tradition (Tra)** devout, accepting portion in life, humble, moderate, respect for tradition, detachment;

**Conformity (Con)** politeness, honoring parents and elders, obedient, self-discipline; and

**Security (Sec)** clean, national security, social order, family security, reciprocation of favors, healthy, sense of belonging

First, each of the 56 values is converted to a single word that captures the meaning of that value, for example, "an exciting life" in Stimulation is converted to "excitement". The next step is to identify all the synonyms of the 56 words and their variations using a thesaurus. The thesaurus used is Roget's II: The New Thesaurus (3rd edition) which is available online at www.answers.com. The last step is to find all the variations (verb, noun, adjective, adverb, plurals, etc.) of the 56 words. The bags of words were edited to avoid any overlapping in words among values as well as words that are commonly used to convey a different meaning. As a result, for each value type, a unique bag of words has been developed for word count analysis.

## 3. Data preparation

The data used in this paper is the March 2, 2004 Version of Enron Corpus, which contains 517,431 messages organized into 150 folders. It has been found that 252,830 of them are unique messages. 156 employees are identified from these 150 folders, and the emails among them are defined as the research scope. Finding all the aliases for each employee and generalizing their formats reduces the 252,830 messages to the emails that are exchanged among these 156 employees. As a result, 22,050 emails are retained for this analysis [19].

For these 22,050 emails, the content was examined to count the frequencies for each word under each value type. The content needed to be cleaned in order to capture the real message delivered by the sender. Among these 22,050 emails, 2,987 of them are emails without content (forwarding emails), 11,095 of them are emails with content but the contents do not contain any value words, and 7,968 of them have at least one value word in the content. Table 1 shows the descriptive statistics of value types by 156 employees.

**Table 1. Descriptive statistics of the value types**

| | Pow | Ach | Hed | Sti | Sel | Uni | Ben | Tra | Con | Sec |
|---|---|---|---|---|---|---|---|---|---|---|
| avg | 26 | 22 | 1 | 3 | 20 | 4 | 20 | 3 | 7 | 21 |
| std | 58 | 44 | 2 | 6 | 36 | 10 | 37 | 7 | 13 | 40 |

The average number of "Power" words used by each employee is 26, which ranks first in the list; "Achievement", "Security", "Self-direction", and "Benevolence" are in the second tier; "Conformity", "Universalism", "Stimulation", "Tradition", and "Hedonism" are in the last tier. It shows that the frequency of value types in the third tier is much smaller than that in the first and second tier. The standard deviation of the count based on value types is relatively large. Some of the actors in the social network talked about "values" a lot, while others barely mentioned any value words in their emails. Such difference is partly caused by the huge difference in the number of emails sent by each employee. The descriptive statistics shows that on average, Enron employees have a culture of appreciation of power, achievement, and independence; and they care about safety and good relationships with their colleagues. The lack of words indicating conformity, stimulation, universalism, tradition, and hedonism also appear to indicate features of the organizational culture of Enron.

It is worth noting that the number of value words used in each employee's emails differs significantly. Twelve employees never used any value words, but employee 20 had 1,763 value words in his emails. As shown in Table 1, the average number of value words is 127 with a high standard deviation of 237. The employees are sorted by the frequencies of values in their emails as shown in Figure 1. One reason for the difference in number of values is that the number of emails sent by each employee varies significantly. Figure 2 shows the histogram of the employee counts for a series of value words. It shows that twelve employees used no value words at all; seven employees used one value word; ten employees used between two to five value words (inclusive), etc. A small number of value words cannot provide enough evidence about the value pattern for any individual. Therefore, the employees with few value words used were excluded from analysis. The cut off value was selected as 20, and as a result, 55 employees were removed, resulting in 101 employees for further analysis.

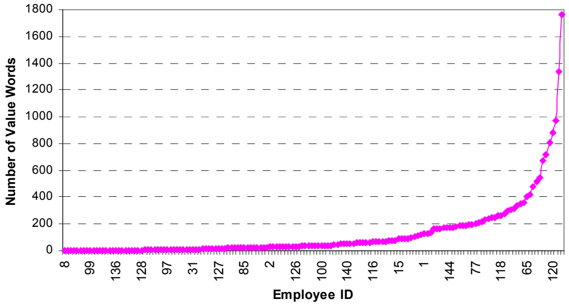

**Figure 1. Value words count for each employee**

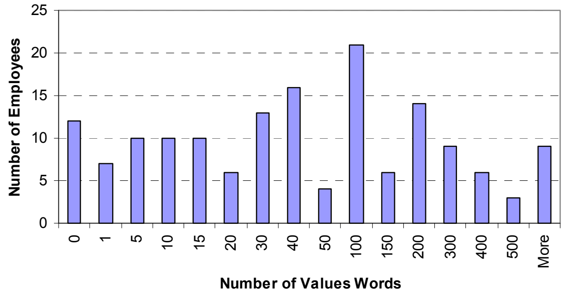

**Figure 2. Employee count for the value word bins**

Two initial matrices were prepared for analysis. The first matrix has emails as the rows, value types as columns, and the frequency of each value type in each email as the cell entry. The second is the sender and value type data. The frequencies for each value type in the first matrix are aggregated according to the senders. A matrix was created with employees as the rows, value types as columns, and the frequency of the value words from the employee in all the emails that s/he sent as the cell entry. The assumptions are first checked for validity before the analysis.

The value types reflect different aspects of the organizational culture, which means that value types should not be significantly correlated to each other. The email and value type matrix were analyzed to understand the relationship between value types. The correlation matrix **R**, which contains the correlation coefficient for every pair of value types ($r_{ij}$) was given in Table 2. The highest correlation coefficient is only 0.21 among all the pairs, which is moderately low and confirms that value types do indeed reflect different aspects of organizational culture.

**Table 2. Correlation matrix of the value types**

| | Ach | Ben | Con | Hed | Pow | Sec | Sel | Str | Tra | Uni |
|---|---|---|---|---|---|---|---|---|---|---|
| Ach | 1.00 | | | | | | | | | |
| Ben | .20 | 1.00 | | | | | | | | |
| Con | .08 | .10 | 1.00 | | | | | | | |
| Hed | .03 | .01 | .00 | 1.00 | | | | | | |
| Pow | .16 | .21 | .08 | -.03 | 1.00 | | | | | |
| Sec | .16 | .19 | .09 | .05 | .20 | 1.00 | | | | |
| Sel | .09 | .08 | .02 | -.02 | .15 | .07 | 1.00 | | | |
| Sti | .14 | .09 | .03 | .02 | .12 | .08 | .07 | 1.00 | | |
| Tra | .16 | .07 | .07 | -.01 | .19 | .15 | .08 | .07 | 1.00 | |
| Uni | .13 | .17 | .08 | .00 | .16 | .12 | .08 | .07 | .08 | 1.00 |

These Enron employees share similar value patterns. The correlation between the employees formed by the ten value types was checked from the sender and value type matrix. The 101×101 correlation matrix shows that most of the 101 employees are highly correlated, which means that they have similar value patterns. The Multidimensional Scaling (MDS) technique is used to visualize the similarities between employees. A two dimension MDS representation of 101 employees is shown in Figure 3 with Stress = 21.2%. The high stress value indicates that a three or more dimension space configuration is more appropriate. Nevertheless, the representation of the employees provides an idea of similarity among them. It shows that most of the employees are relatively close to each other, but a few of them are far from the groups.

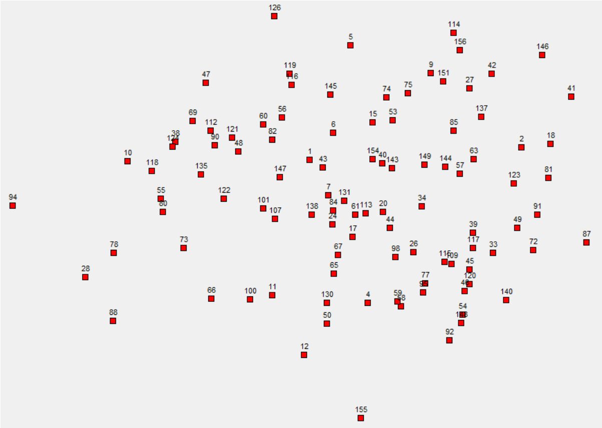

**Figure 3. A two-dimensional representation of 101 employees produced using MDS**

## 4. Analysis methods and results

The hierarchical clustering method is applied to identify groups of employees (clusters) with similar value patterns, and then relationship between clusters and value types is studied by correspondence analysis. The intra-cluster and inter-cluster communication density within and among clusters are compared to test the first hypothesis – individuals sharing similar value patterns communicate more frequently. The second analysis focuses on the second hypothesis, which is that people who communicate more frequently with each other share similar value patterns. This hypothesis is tested by identify overlapping clusters from communication frequencies and then label the value patterns for each cluster.

### 4.1 Hierarchical clustering method

Hierarchical clustering produces clusters either by successive divisions (divisive hierarchical methods) or successive mergers (agglomerative hierarchical methods) based on current clusters. The merger (division) decision is based on the measure of similarity (distance) between the clusters. If the number of items for clustering is n, agglomerative hierarchical methods will start with n clusters, and the nearest neighbors will merge to form a bigger cluster until all the n items are in a single cluster. The divisive hierarchical methods, on the other hand, start with a single cluster and divide the current clusters into smaller clusters by taking the farthest item as a separate cluster until each single item itself is one cluster.

Given a coordinate matrix, in which the rows are observations and columns are variables, the similarity or dissimilarity measure should be defined before applying any clustering algorithms. Various similarity measures such as Gower, correlation, Jaccard, and others, can be chosen depending on the data characteristics. The dissimilarity measure is opposite to the similarity measure, and it is equal to 1 minus the similarity measure. Or, the dissimilarity measure can be defined separately, such as Euclidean, Canberra metric, Czekanowski coefficient, etc.

Once an appropriate similarity (distance) measure is decided, various clustering algorithms are available for grouping the observations. For agglomerative hierarchical methods, there are simple linage, complete linkage, average linkage, centroid method, two-stage density linkage, Ward's minimum-variance method, etc. Each method emphasizes on different criteria of grouping two clusters based on the definition of distance. For example, simple linkage chooses the smallest distance between two cluster members as the distance of the two clusters, while complete linkage chooses the largest distance between two cluster members as the distance of the two clusters. One issue is to choose an appropriate clustering method for the data to be analyzed. It is essential to realize that "most methods are biased toward finding clusters possessing certain characteristics related to size (number of members), shape, or dispersion" [16].

Many studies have been performed to compare the clustering algorithms using artificial data obtained through pseudo random number generators [14]. The studies show that for well-separated data, all the clustering algorithms perform well. However, for poorly separated data, the performance of various clustering methods differs significantly. On average, the average linkage or Ward's minimum-variance method have the best overall performance. In this research, the Ward's minimum-variance method is applied to find clusters. In Ward's minimum-variance method, the distance between two clusters is the ANOVA sum of squares between the two clusters added up over all the variables. At each generation, the total of sum of squares over all the clusters is minimized over all partitions obtainable by merging two clusters from the previous generation [16, 18]. Ward's method tends to produce clusters with roughly the same number of observations and is very sensitive to outliers [13].

Another issue with the clustering technique is to determine the number of clusters. Clustering is distinct from the classification technique in that number of groups is unknown. No assumptions are made regarding the number of groups or the group structure. If the analysts know the data very well, they may have a preliminary idea about what the group structure would be. However, a lot of times the data are not well known. In addition, it shows that there are no completely satisfactory methods for determining the number of population clusters for any type of cluster analysis [3, 6].

The good news is that there are visualization tools and several measures available to help in determining the number of clusters. MDS, for example, can give a helpful visual presentation of the group structure of the data if the STRESS measure is low. The dendrogram is a two-dimensional diagram that illustrates the mergers or divisions made at each successive step. The measures include Sarle's cubic clustering criteria (ccc), pseudo F statistic (PSF), pseudo $T^2$ (PST2) test, $R^2$, and semi-partial $R^2$, etc. Here are some general rules when these measures are used to determine the number of clusters. Usually, an appropriate number of clusters should be retained is the number of clusters where ccc is at peak and ccc > 3 which means the sample data is not from a uniform distribution; PSF is at peak because PSF measures the separation among the clusters; $T^2$ statistic is large since it measures if the two clusters should be considered different; $R^2$ is steady; and the semi-partial $R^2$ should be small since two clusters cannot be joined otherwise.

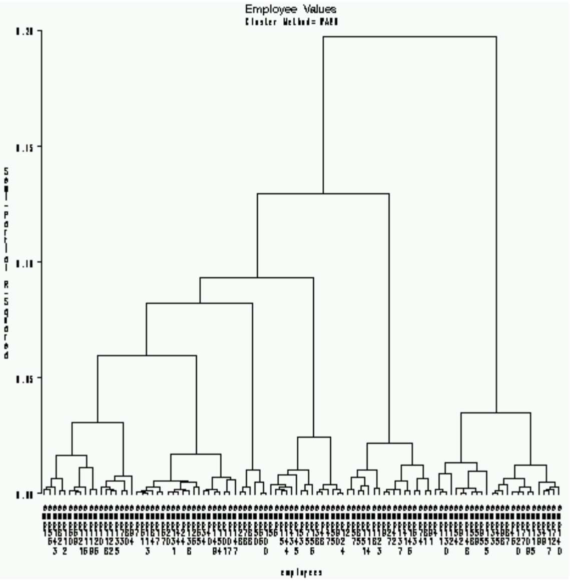

**Figure 4. Ward's minimum-variance dendrogram for distances between 101 employees**

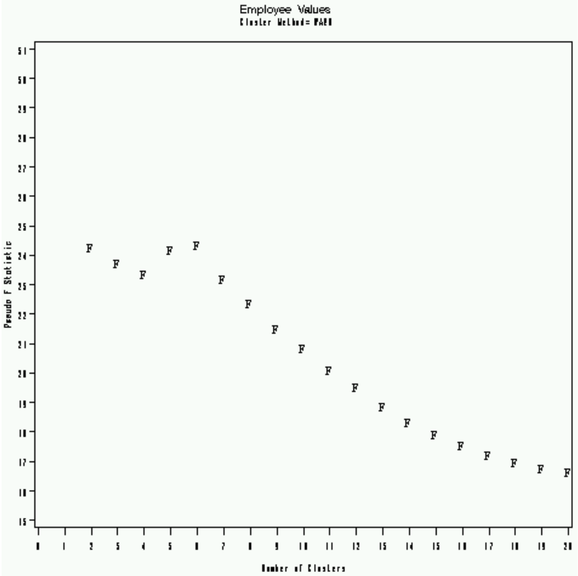

**Figure 5. The PSF measure versus number of clusters**

### 4.2 Clustering employees based on their value patterns

Employees who have similar value patterns have high correlation coefficients. The correlation matrix was used as the similarity measure because clustering the employees who have similar value pattern is the research interest. From the 101 × 10 coordinate matrix in which the rows are employees and the columns are value types, the correlation matrix between employees is calculated. Since SAS requires distance matrix as an input to perform cluster procedure, the distance $d_{ij} = \sqrt{1 - r_{ij}}$ is obtained as the distance measure for any two employees i and j. The Ward's minimum-variance clustering method is applied to find the groups of employees with similar value types. The ccc measure is not available since the distance matrix is used instead of the original coordinate matrix for the clustering analysis. The other measures are available for determining number of clusters. The dendrogram of the cluster results is given in Figure 4 and the plot of PSF versus number of clusters is given in Figure 5. Both of the figures and other measures suggest 6 clusters for this data set. The members in each of the six clusters are given in Table 3.

**Table 3. Employee clusters and their members**

| Cluster | Count | Employees |
|---|---|---|
| C1 | 25 | 4,11,12,33,39,45,46,49,54,58,59, 72,77,87,92,95,98,109,115,117, 120,130,140,148,155 |
| C2 | 15 | 5,6,15,38,43,47,48,53,55,75,90, 124,144,145,156 |
| C3 | 6 | 6 28,50,66,78,88,100 |
| C4 | 20 | 7,17,20,24,26,34,40,44,61,65,67, 84,101,107,113,131,138,147,149, 154 |
| C5 | 17 | 2,9,18,27,41,42,57,63,74,81,85,91, 114,12,137,146,151 |
| C6 | 18 | 1,10,56,60,69,73,80,82,94,112, 116,118,119,121,122,126,135,143 |

The number of employees in each cluster differs. Cluster 1 to 6 has a total of 25, 15, 6, 20, 17, 18 members respectively. With the member information in each cluster, a frequency matrix can be produced. The matrix has rows as clusters and columns as value types, and cell entries are the sum of the value word frequencies added up over all the employees within the cluster. The frequency matrix is shown in Table 4. To detect the associations between the clusters and value types, correspondence analysis is implemented. Correspondence analysis is a graphical procedure for representing associations in a table of frequencies [12]. It is a technique to map each row item and column item as a point in a two-dimension space such that the distance between the points reflects similarities and associations. Row points with small distance indicate they have similar profiles across the columns. Column points with small distance indicate they have similar profiles down the rows. If a row point is close to a column point, it means they have association relationship, i.e., they are correlated and expected to be together more frequently than an independence case. Other than the graphical representation of the data, the output from a correspondence analysis also includes a measure (called the inertia) of the amount of variation explained by each dimension. The analysis results of the cluster-value type frequency table are given in Figure 6.

**Table 4. Value type counts for each cluster**

| | Ach | Ben | Con | Hed | Pow | Sec | Sel | Str | Tra | Uni |
|---|---|---|---|---|---|---|---|---|---|---|
| C1 | 715 | 895 | 322 | 35 | 1945 | 1162 | 958 | 130 | 158 | 186 |
| C2 | 219 | 391 | 96 | 13 | 152 | 197 | 234 | 36 | 12 | 43 |
| C3 | 25 | 16 | 16 | 4 | 13 | 86 | 33 | 5 | 6 | 3 |
| C4 | 1468 | 1065 | 387 | 41 | 1433 | 1061 | 987 | 180 | 208 | 242 |
| C5 | 296 | 278 | 108 | 8 | 235 | 241 | 595 | 66 | 38 | 42 |
| C6 | 621 | 419 | 103 | 47 | 223 | 418 | 297 | 85 | 48 | 84 |

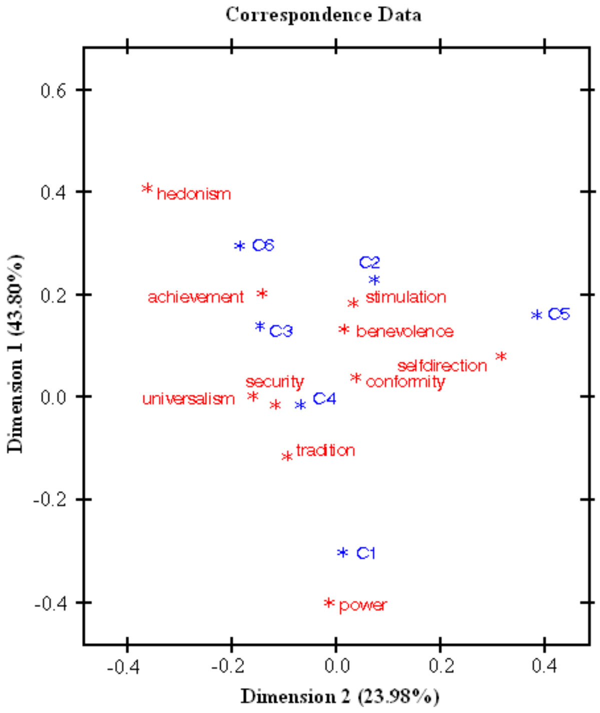

**Figure 6. Correspondence analysis plot of the cluster-value type data**

The correspondence analysis plot contains 16 points including 6 cluster points and 10 value type points. The plot indicates, for example, that clusters do not show similar profiles. Clusters C1 and C6 have very different profiles across value types. On the other hand, value type Security and Universalism, Stimulation and Benevolence have similar profiles, and value type Power and Hedonism have very different profiles across the clusters. When it comes to the association relationship between clusters and value types, it is concluded that Cluster C1 is associated with value type Power; Cluster C2 is associated with value types Stimulation and Benevolence; Cluster C3 and C4 are in the middle of the correspondence plot and their association with value types are not as clear as the other clusters; Cluster C5 is associated with value type Self-direction; and Cluster C6 is associated with value type Achievement. The inertia associated with Dimension 1 is 43.80% and with Dimension 2 is 23.98%. As a result, the two dimensions account for 43.80% + 23.98% = 67.78% of the total inertia, which means 67.78% of the information in the data is represented by the two-dimensional plot.

## 4.3 Testing the first hypothesis

The first hypothesis is: do employees who share similar value patterns talk to each other more frequently? To test this hypothesis, the intra-cluster and inter-cluster density for the six clusters are calculated. Let $n_a$ and $n_b$ denote the number of employees in Cluster a and b, and $x_{ij}$ denote number of emails between employees i and j. The intra-cluster density of Cluster a is defined as Equation (1). The inter-cluster density between Cluster a and b is defined in Equation (2). The results are shown in Table 5. The intra-cluster density of C1, C2, C4, and C5 is clearly larger then their inter-cluster densities. The intra-cluster density of C3 is smaller than all its inter-cluster densities. For Cluster C6, its intra-cluster density is smaller than the intra-cluster density with C4 and C2, but larger than that with C1, C3, and C5. It is concluded that people sharing similar value pattern do show a denser communication pattern within the groups, but it is not equally true for all of the clusters. The value type associated with the clusters may also affect the communication pattern.

$$D_a = \frac{\sum\limits_{i\in C_a}\sum\limits_{j\in C_a} x_{ij}}{n_a(n_a-1)} \tag{1}$$

$$D_{ab} = \frac{\sum\limits_{i\in C_a}\sum\limits_{j\in C_b} (x_{ij}+x_{ji})}{2n_a n_b} \tag{2}$$

**Table 5. Intra-cluster and inter-cluster densities between 6 clusters**

| | C1 | C2 | C3 | C4 | C5 | C6 |
|---|---|---|---|---|---|---|
| C1 | 9.15 | | | | | |
| C2 | 2.61 | 5.11 | | | | |
| C3 | 0.78 | 0.97 | 0.20 | | | |
| C4 | 5.51 | 2.29 | 0.34 | 13.85 | | |
| C5 | 1.13 | 2.52 | 0.96 | 1.72 | 4.34 | |
| C6 | 0.53 | 2.97 | 1.07 | 4.80 | 1.32 | 2.14 |

## 4.4 Testing the second hypothesis

The second hypothesis is: if people are clustered based on their communication frequencies, can the clusters be labeled with value types? – can people say this cluster is a "Power" cluster, while the other one is an "Achievement and Benevolence" cluster? This question is answered by labeling the people with value types, then compare the membership of each cluster with the membership of each value type to calculate the match ratio.

The overlapping cluster algorithm [1] is run to identify overlapping clusters based on the triples "senderID recipientID frequency" for the 101 employees. As a result, nine overlapping clusters are produced and their members are listed in Table 6.

**Table 6. Overlapping clusters and their members**

| Cluster | Count | Employees |
|---|---|---|
| OC1 | 23 | 2,5,6,41,42,43,55,63,66,69,74,78,85, 100,107,112,116,118,123,126,140,146 |
| OC2 | 46 | 1,2,4,6,7,11,15,24,26,33,40,42,45,49, 53,58,59,60,65,67,77,78,82,84,85,88, 92,95,98,100,107,109,113,116,120, 121,123,124,130,135,138,144,147, 148,154,155 |
| OC3 | 9 | 39,54,78,84,90,94,116,123,135 |
| OC4 | 7 | 18,44,87,114,119,137,151 |
| OC5 | 14 | 26,53,55,58,75,88,91,95,112,118,126, 143,144,149 |
| OC6 | 10 | 15,26,45,53,58,77,88,95,109,113 |
| OC7 | 46 | 1,2,6,7,11,15,20,24,33,40,42,45,46,53, 55,60,61,65,67,77,78,82,84,85,90,94, 98,100,107,109,113,116,121,122,123, 124,126,131,135,140,143,147,148, 154,155,156 |
| OC8 | 23 | 2,6,9,12,17,27,33,34,38,40,47,48,56, 57,72,73,80,81,101,107,115, 117,145 |
| OC9 | 34 | 5,18,20,33,40,43,44,46,55,61,66,69, 75,78,87,91,112,113,114,116,118,119, 122,124,126,131,137,140,143,146, 147,149,151, 156 |

People are labeled with value types based on their total value type frequencies. In the employee and value type matrix, it has employees as rows, value types as columns, and the cell entry is the total frequency of value type in the employee's emails. The percentage of each value type is calculated by dividing each cell entry with the row sum and then sorted. As a result, each employee is characterized by percentage of different value types. A rule is determined to label the employees with the value types. The rule has two considerations: one is that the value types with high percentage should be retained; and the other is the percentages between two consecutive value types should not differ significantly. If one of the value types dominates, that employee will be labeled with the specific value type even though other high percentages exist. The percentage difference threshold is set to be 5%. The value type with the highest percentage is put to the employee's value list first, then it is compared with the second highest percentage, two cases may happen: if the difference is larger than 5%, then stop; else put the second value type to the employee's value list, and compare with the third highest percentage, and so on. The number of value types may differ from employee to employee.

**Table 7. Value types and their members**

| Value Type | Count | Employees |
|---|---|---|
| Ach | 38 | 1,7,10,17,20,24,34,39,40,44,53,56,60,61,65,69,82,84,85,101,107,112,113,116,118,119,121,122,126,131,135,138,143,145,147,149,151,154 |
| Ben | 33 | 5,6,15,38,43,47,48,53,55,75,90,124,144,145,156 |
| Con | 3 | 17,143,145 |
| Hed | 1 | 143 |
| Pow | 38 | 4,7,17,20,24,26,33,34,39,44,45,46,49,54,58,59,61,65,67,72,77,84,87,92,95,98,109,113,115,117,120,130,131,140,143,148,149,155 |
| Sec | 31 | 7,10,11,17,20,24,28,39,44,50,58,61,65,66,67,73,74,78,80,82,84,88,94,100,112,113,122,130,131,143, 155 |
| Sel | 37 | 2,7,9,15,17,18,20,27,39,41,42,44,49,53,57,61,63,74,75,81,82,84,85,91,112,113,114,123,131,137,143,144,145,146,149,151,156 |
| Sti | 2 | 17,143 |
| Tra | 2 | 17,143 |
| Uni | 2 | 17,39 |

Table 7 shows the 10 value types with their members. Interestingly, the correspondence results are consistent with the value type members. For example, in the correspondence analysis, Cluster C1 is associated with value type Power. In this analysis, it is found that 23 out of 25 employees (except Employee 11 and 12) in C1 belong to the Power group. Also, C2 is associated with Stimulation and Benevolence, and all the 15 employees belong to the Benevolence group; C5 is associated with Self-direction, and all of its members belong to this value group; C6 is associated with Achievement, and 15 out of 18 of its members belong to the Achievement group. C3 and C4 are in the middle of the correspondence plot and surrounded by several value types, and their association with value types is not as clear as the boundary clusters.

The value type members and the overlapping cluster members are compared to detect if any match exists, i.e., can the clusters be labeled with one value type or a combination of value types? If $\{OC_i\}$ is denoted as the employee set of Cluster i and $\{VT_j\}$ is denoted as the employee set of the j$^{th}$ combination of value types, the match ratio is defined in Equation (3).

$$m_{ij} = \frac{\{OC_i\} \cap \{VT_j\}}{\{OC_i\} \cup \{VT_j\}} \qquad (3)$$

It turns out that the labeling is not successful. The maximum match ratio happens between Cluster 2 and the value type combination of "Achievement, Power, Security" with a value of 48.8%. The low match ratio indicates that the email communication frequency is not affected solely by the values of the employees. It is found that the employees with similar value patterns tend to communicate more frequently, but the reverse is not necessarily correct. The distribution of value type counts by the employees is listed in Table 8. It can be seen that the value type Achievement, Benevolence, Power, Security, and Self-direction are almost evenly distributed within the cluster, which makes the value labeling difficult. Another difficulty in labeling is that people sent emails for various reasons. For Enron email data set, a large number of the emails are work related. As a result, the recipients are not necessarily the person who share value pattern similar to the sender. Table 9 shows the distribution of employees from different departments across the overlapping clusters. The abbreviation of the department is used and the abbreviation list for the department name can be referred. It is reasonable to categorize "GC" (Gas Central), "GT" (Gas Texas), "GW" (Gas West), "GE" (Gas East) and "GF" (Gas Financial) as Gas

Trading group; “WP” (West Power and West Power Real Time) and “EP” (East Power) as Power Trading group; “EWS” (Enron Whole Services) as Executive group of the Enron Whole Services; “EO” as the Energy Operation group; “ETS” as the Enron Transportation Services group; “RG” as Regulatory and Government Affairs group; “L” as Legal group; and “E” as top Executives in Enron. Each overlapping clusters is composed of one or more of such function groups. It is found that the West Power, Legal, and Transportation groups dominate the communications in Cluster OC4, OC6, and OC8 respectively. OC1 and OC3 are Trading clusters; OC5 is a combination of West Gas and Legal cluster; Cluster OC2, OC7, and OC9 contains much more members than the other clusters do, therefore, they are a combination of several departments. The intra-department communications are very strong, but it also contains a significant amount of inter-department communications.

**Table 8. Distribution of value types across the overlapping clusters**

| | Ach | Ben | Con | Hed | Pow | Sec | Sel | Str | Tra | Uni |
|---|---|---|---|---|---|---|---|---|---|---|
| OC1 | 7 | 7 | 0 | 0 | 1 | 5 | 10 | 0 | 0 | 0 |
| OC2 | 18 | 11 | 0 | 0 | 22 | 14 | 12 | 0 | 0 | 0 |
| OC3 | 4 | 3 | 0 | 0 | 3 | 4 | 3 | 0 | 0 | 1 |
| OC4 | 3 | 1 | 0 | 0 | 2 | 1 | 5 | 0 | 0 | 0 |
| OC5 | 6 | 6 | 1 | 1 | 5 | 4 | 7 | 1 | 1 | 0 |
| OC6 | 2 | 3 | 0 | 0 | 7 | 3 | 3 | 0 | 0 | 0 |
| OC7 | 23 | 18 | 1 | 1 | 19 | 17 | 15 | 1 | 1 | 0 |
| OC8 | 7 | 7 | 2 | 0 | 6 | 3 | 7 | 1 | 1 | 1 |
| OC9 | 17 | 14 | 1 | 1 | 11 | 10 | 16 | 1 | 1 | 0 |

**Table 9. Distribution of departments across the overlapping clusters**

| | GC | GT | GW | GE | GF | WP | EP | EWS | EO | ETS | RG | L | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| OC1 | 1 | 4 | 9 | 0 | 2 | 0 | 3 | 0 | 4 | 0 | 0 | 0 | 0 |
| OC2 | 2 | 2 | 1 | 1 | 3 | 0 | 7 | 9 | 3 | 0 | 0 | 16 | 2 |
| OC3 | 1 | 2 | 0 | 5 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| OC4 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| OC5 | 1 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| OC6 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 | 0 |
| OC7 | 1 | 2 | 5 | 3 | 3 | 0 | 7 | 9 | 3 | 0 | 4 | 6 | 3 |
| OC8 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 18 | 0 | 0 | 0 |
| OC9 | 0 | 4 | 12 | 0 | 0 | 7 | 2 | 1 | 0 | 0 | 4 | 2 | 2 |

## 5. Limitations

As this study represents a novel and experimental approach to social network analysis, there are some limitations that may affect the reliability of the results. First, the dataset is an email dataset within an organization. Most messages tend to be work related and thus the use of values may reflect the organizational culture more than individuals’ values. It would be valuable to extend this research approach to other online discussions where participants might be discussing non-work-related issues.

Second, the list of value words is not perfect, and with the limitations of the current word count method, it is not possible for computers to take into consideration different meanings for the same word. Therefore, false positives or false negatives may appear occasionally due to actual word usage.

Further, since this approach used only automatic text analysis, there is no human annotation that could be used as a “ground truth” for comparison, which would allow for an assessment of the accuracy of the approach to automatic text analysis included here. Thus, it would be useful to compare the bag of words approach to human annotation to assess the degree of accuracy. However, the advantage of conducting macro-scale e-social science is that as long as the results are not biased in any ways at the macro scale, it is not important if there are errors at the micro scale [8], since the Enron dataset is large enough to potentially overcome localized errors. However, more analysis is needed to determine if this approach results in unbiased analysis. Thus, while this approach shows promise and potential, more work is needed to more conclusively test the hypotheses posed in this paper.

## 6. Conclusions

This paper describes how a word count method using a bag of value words is used to analyze Enron emails. Various statistical analyses, including hierarchical clustering, overlapping clustering, and correspondence analysis, are applied to identify the value profiles of the employees of at Enron and test two hypotheses. The first hypothesis, that people sharing similar value patterns tend to communicate more frequently, is supported by the results; the second hypothesis, that people who communicate more frequently with each other share similar value patterns is not supported by the results of analysis.

These research findings are significant because they help to explain the complex relationship between value patterns and communication patterns. People who share values are likely to communicate more than people who have different values, indicating that values can shape communication patterns. However, people who communicate frequently do not necessarily share similar value patterns, indicating that communication patterns do not appear to significantly shape values, at least within an online community involving professionals. These findings point toward the enduring nature of values, which are more likely to shape than to be shaped by communication. Thus, it is important for researchers studying social networks and human communication in general to consider the role of values in shaping communication patterns.

The methods used in this paper can be applied to other corpora and research questions as well. Thus, this approach has broad applicability as a novel means of analyzing social networks. Further, the findings of this study illustrate the overall potential for values to serve as a tool to assist social network analysis.

## 7. Acknowledgements

This material is based in part upon work supported by the National Science Foundation under Grant Nos. 0521834, 0646404, 0731717, and 0731718.

## 8. References

[1] J. Baumes, *Algorithms for Discovering Hidden Groups in Communications*, PhD Thesis, Rensselaer Polytechnic Institute, 2006.

[2] C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan, “Mining email social networks”, In *MSR’06: Proceedings of the 2006 international workshop on mining software repositories*, pp. 137–143, New York, NY, USA, 2006. ACM Press.

[3] H. H. Bock, “On Some Significance Tests in Cluster Analysis”, *Journal of Classification*, 2, 1985, pp. 77–108.

[4] J. Diesner, T. L. Frantz, and K. M. Carley, “Communication networks from the Enron email corpus”, *Computational & Mathematical Organization Theory*, 11(3), 2005, pp. 201–228.

[5] D. Dillon, D. Cottrell, and J. Reser, “Group Differences in Word Use and Meaning: A Text Analysis of the Abstract Word, ’Values’”, In *Proceedings of the 9th International Conference on the Statistical Analysis of Textual Data*, Lyon, France, March 2008.

[6] B. S. Everitt, “Unresolved Problems in Cluster Analysis”, *Biometrics*, 35(1), 1979, pp. 169–181.

[7] L. A. Fast and D. C. Funder, “Personality as Manifest in Word Use: Correlations with Self-Report, Acquaintance-Report, and Behavior”, *Journal of Personality and Social Psychology*, 94(2), 2008, pp. 334–346.

[8] K. R. Fleischmann, D. W. Oard, A.-S. Cheng, P. Wang, and E. Ishita, “Automatic Classification of Human Values: Applying Computational Thinking to Information Ethics”, In *Proceedings of the 72nd Annual Meeting of the American Society for Information Science and Technology*, Vancouver, British Columbia, Canada, November 6-11, 2009.

[9] L.C. Freeman, “The impact of computer based communication on the social structure of an emerging scientific speciality”, *Social Networks*. 6, pp. 201-221.

[10] B. Friedman, P. H. Kahn, and A. Borning, “Value Sensitive Design and Information System”, In P. Zhang and D. Galletta (Eds.), *Human–Computer Interaction in Management Information Systems: Foundations*, M.E.Sharpe, New York, 2006, pp. 348–372.

[11] C. J. Groom and J. W. Pennebaker, “Words”, *Journal of Research in Personality*, 36(6), 2002, pp. 615–621.

[12] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis, 5th Edition*, Prentice Hall, Upper Saddle River, New Jersey, 2002.

[13] G. W. Milligan, “An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms”, *Psychometrika*, 45(3), 1980, pp. 325–342.

[14] G. W. Milligan, “A Review of Monte Carlo Tests of Cluster Analysis”, *Multivariate Behavioral Research*, 16(3), 1981, pp. 379–407.

[15] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. *The Development and Psychometric Properties of LIWC2007*. LIWC.net, Austin, Texas, 2007.

[16] SAS Institute, *SAS/STAT 9.1 User’s Guide*, SAS Institute, 2004.

[17] S. H. Schwartz, “Are There Universal Aspects in the Structure and Contents of Human Values?”, *Journal of Social Issues*, 50(4), 1994, pp. 19-45.

[18] J. H. Ward Jr., “Hierarchical Grouping to Optimize an Objective Function”, *Journal of the American Statistical Association*, 58(301), 1963, pp. 236–244.

[19] Y. Zhou, M. Goldberg, M. Magdon-Ismail, and W. A. Wallace, “Strategies for Cleaning Organizational Emails with an Application to Enron Email Dataset”. In *Proceedings of the 5th Conference of North American Association for Computational Social and Organizational Science (NAACSOS 07)*, Atlanta, Georgia, June 7–9, 2007.