ZL Technologies at TREC 2009 Legal Interactive

Comparing Exclusionary and Investigative Approaches for Electronic Discovery using the TREC Enron Corpus

John Wang
jwang@zlti.com

Cameron Coles
ccoles@zlti.com

Rob Elliot
relliot@zlti.com

Sofia Andrianakou
sandrianakou@zlti.com

ZL Technologies
San Jose, CA

Abstract

Organizations responding to requests to produce electronically stored information (ESI) for litigation today often conduct information retrieval with a limited amount of data that has first been culled by custodian mailboxes, date ranges, or other factors chosen semi-arbitrarily based on legal negotiations or other exogenous factors. The culling process does not necessarily take into account the composition of the data set; and may, in fact, impede the expediency and cost-effectiveness of the eDiscovery process as ESI not initially identified may need to be collected later in the eDiscovery process. This exclusionary eDiscovery approach has been recommended by search and information retrieval technology providers in the past, in part, based on the state of technology available at the time; however, the technology now exists to perform an inclusive, content-based, investigative eDiscovery across a large document collection without the introduction of semi-arbitrary exclusion factors. In this paper, we investigate whether limited document retrieval based on custodian email mailboxes results in lower recall and produces fewer responsive documents than a broader, inclusive search process that covers all potential custodians. In order to compare the two approaches, we designed an experiment with two independent teams conducting electronic discovery using the different approaches. We found that searching across the entire data set resulted in finding significantly more responsive documents and more initial custodians than implementing an approach that relies on custodian-based culling. Specifically, investigative eDiscovery found 516% more relevant documents and 1825% more initial custodians in our study. Based on these results, we believe organizations that employ an exclusionary, culling-based methodology may require subsequent collections, risk under production and sanctions during litigation, and will ultimately expend more resources in responding to eDiscovery production requests with a less comprehensive result.

1 Introduction

In 2009, the TREC Legal Track continued to provide avenues for research in modeling "more completely and accurately the task of reviewing documents for responsiveness to a request for production in civil litigation."[i] One widely adopted search and retrieval methodology used by lawyers is the use of *exclusionary* approaches to remove large amounts of content from review for electronic discovery (ED) as opposed to more comprehensive, content-based, *investigative* approaches.[ii] The exclusionary approach is a newer, but widely-popular approach that seeks to reduce the amount of content being analyzed for responsiveness by culling the set of data using semi-arbitrary exclusion factors such as specific persons of interest (custodians), date ranges, and other factors. In these cases, the parties assume that all or most responsive documents are contained within this subset of data. Since the exclusionary approach provides for a semi-arbitrary sampling of data, the validity of this assumption can play a significant role in under production and spoliation. The investigative approach attempts to analyze all data available for relevance using matter-based culling and had its roots in the computer forensics field of litigation practice. While the exclusionary approach was adopted as volumes of electronically stored information (ESI) grew, new technologies are now available that enable the application of an investigative approach to electronic discovery.

The investigation into well-established search protocols is timely as the growing volume of ESI has led to inadequate application of keyword search and other technologies. In a recent Federal civil case, Victor Stanley, Inc. v. Creative Pipe, Inc., US Magistrate Judge Paul Grimm wrote in his opinion, "all keyword searches are not created equal; and there is a growing body of literature that highlights the risks associated with conducting an unreliable or inadequate keyword search."[iii] Wide-spread adoption of the exclusionary approach is one way keyword searches may be

significantly compromised in the effort to locate relevant information. One reason for this practice may be the exponential growth in informational records grows at exponential rates which may contribute to higher overall discovery costs for organizations.[iv] The large volume of ESI needed to be handled has also been known to lead to sub-optimal performance with traditional IR solutions that may need to search hundreds or thousands of individual search indexes when performing an investigative search. Some newer and evolving IR technologies are designed to search billions of documents quickly similar to the Google search engine and may enable large-scale investigative approaches without the possible performance limitations of traditional approaches.

The availability of such search and IR technology combined with recent court decisions and the recommendation of The Sedona Conference for courts to be alert to new and evolving methods[v] makes investigation of the adequacy of exclusionary approaches very timely.

To date, the effectiveness of the exclusionary approach has not been sufficiently evaluated to confirm the assumption that this widely adopted approach does not increase the risk of underproduction, leave behind significant amounts of relevant, exculpatory material, and increase costs by requiring multiple rounds of collection from custodians, when compared with an investigative, inclusionary approach. Research in the use of IR for law has largely focused on aspects of IR and legal discovery other than the effectiveness of the exclusionary approach. Hogan, Brassil, et al. explored the use Relevance Feedback.[vi] Amati, Bianchi, et al. explored the use of query construction and weighting.[vii] Zhang, Scholer, and Turpin investigated the effects of OCR error minimization.[viii] Given the availability of search and IR technology advances to enable wide-spread inclusionary ED, we added to the body of research for civil litigation by investigating the effectiveness of the current practice of using narrow, exclusionary approaches to information retrieval in law when compared to a broader, content-based, inclusionary approach by simulating the effect of performing search and IR against a discrete number of custodians vs. the entire population. Two non-overlapping experimental groups were created to represent the exclusionary and inclusionary approaches.

- **Exclusionary ED Team (culling or custodian methodology):** this team evaluated email for responsiveness by first identifying several potentially relevant custodians, excluding email from other custodians, and then running searches and information retrieval techniques against the email from those custodians.
- **Investigative ED Team (broad or enterprise methodology):** this team evaluated the email for responsiveness by running content-based searches and information retrieval techniques directly against the entire corpus based on Boolean keyword and concept searches without excluding data based on custodian.

We expected the investigative approach would yield superior results by examining a larger amount of data. Our assumption was that responsive documents would not be limited to only a few custodians who were likely to be involved based on their job description, but would rather be spread out across many custodians. Since this study was primarily concerned with the effects of discovery scope on the number of responsive documents, both experimental groups were given the same resources and utilized the same information retrieval techniques.

Until now there has not been a lot of attention on the use of exclusionary electronic discovery approach and its effects on electronic discovery, and we believe our results will have implications for electronic discovery practices.

2 Test Collection

For our experiment, we prepared two sources of data, an email corpus and a custodian list.

2.1 Enron Email Corpus

For the email corpus, we used the Enron data set prepared and distributed by the 2009 TREC Legal Interactive Track, referred to as the TREC Enron Corpus. This corpus was created as a way to address some of the short comings of the Enron data set prepared by MIT, SRC and CMU, referred to as the CMU Enron Corpus since CMU is the current distribution provider for the collection.[ix]

Item	Value
Raw Emails (partially de-duplicated)	1,231,904
Unique Emails (complete de-duplicated)	569,034
Total Emails (fully re-duplicated)	2,965,103
Number of Custodians	104
Average Emails per Custodian, including duplicates	28,511
Median Emails per Custodian, including duplicates	13,385

Table 1. TREC Enron Corpus Details.

The two primary shortcomings that were addressed are the size of the corpus and the presence of attachments which are not included in the CMU Enron Corpus. TREC Enron Corpus was made available as a partially de-duplicated collection in the EDRM XML format. To create a data set for our investigation, we re-duplicated all the email in the collection and organized it by custodian, yielding 2,965,103 email messages spread across 104 custodians. The messages were then de-duplicated per custodian.

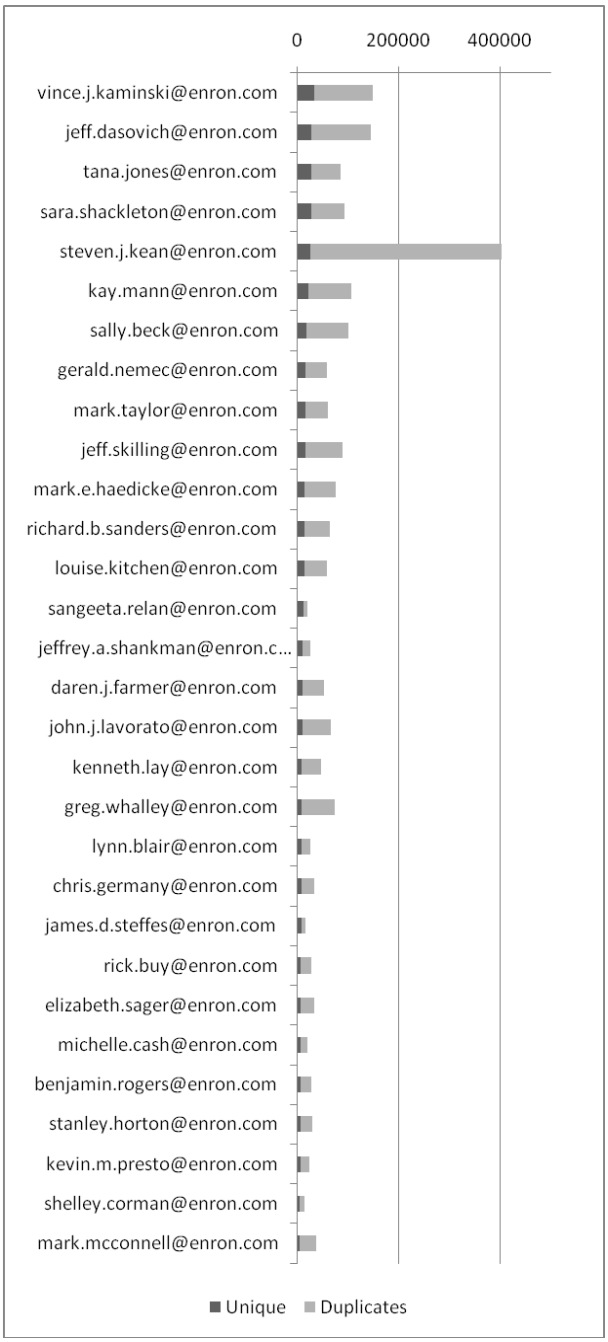

Figure 1. Email distribution across top 30 custodians

In order to more closely approximate an eDiscovery investigation, the exclusionary ED group culled the entire data set by custodian and selected only a subset of custodians as a first-pass cull. To estimate the number of custodians to include in the initial cull, we analyzed the collection process utilized by the original Enron collection during the period of litigation from 2000 to 2002. We then based the number of custodians for the exclusionary ED group on the number and percentage of custodians we identified by analyzing collection dates present in the Enron Corpus.

Collection (dates are from the TREC Enron Corpus)	**Custodians in Each Collection**	**Aggregate Custodians in Overall Collection**	**% of Enron Employees**
Aug 2000	2	2	0.009%
Dec 2000	44	46	0.209%
Jun 2001	59	67	0.305%
Jul 2001	1	67	0.305%
Oct 2001	16	67	0.305%
Nov 2001	67	84	0.382%
Jan 2002	55	95	0.432%
Mar 2002	53	98	0.445%
Total Identified Custodians in TREC Enron Corpus	104	104	0.473%
Total Identified Custodians CMU Enron Corpus	148	148	0.673%

Table 2. Collection statistics over time and acrross the entire collection.

We identified eight distinct dates associated with individual data collection times spanning the period beginning in August 2000 and ending in March 2002. August 2000 coincides with the month the San Diego Gas & Electric Company filed a complaint against Enron alleging market manipulation; an event that likely triggered the initial collection. March 2002 is one month after FERC began their investigation into Enron's involvement in the Western U.S. Energy Crisis. The distribution of custodians across the eight collection dates is particularly instructive as it may be interpreted to indicate organizational knowledge of the issue is growing over time. This may indicate an exclusionary eDiscovery may miss custodians earlier in the

process when there is an imperfect state of knowledge of the matter.

While we were not able to identify dates for all custodians, the identified dates covered 98 of the 104 custodians we identified. We compared the number of custodians per collection date with the estimated Enron staff count of 22,000 at the end of 2001.[x]

This data was used to provide a test of reasonableness for the number of custodians selected by the exclusionary eDiscovery team as an absolute number and percentage of employees for the full Enron staff. The total number of custodians for the TREC and CMU Enron Corpuses was also considered.

It should be noted that while the TREC Enron Corpus has more email than the CMU Enron Corpus, many are duplicates and the collection is still a small collection when compared to the actual email at Enron or other large organizations. The TREC Enron corpus covers 104 custodians while Enron had a worldwide staff count of 22,000 at the end of 2001. We recognize this and compensated by allowing the inclusive team access to all 104 custodians' emails while the exclusionary team selected four key custodians and only after the names were selected, were they given access to the custodian's email. Recognizing the limited email collection made available, subsequent studies with larger corporate email data sets would be useful in future studies of this type.

2.2 Enron Custodian List

For the Enron custodian list, we started with the "ex employee status report" created by Shetti and Adibi of USC.[xi] We then enhanced this list for the 104 custodians so we had a more complete list of titles, job descriptions, and Enron departments to work with. From here, the exclusionary approach team was able to choose their custodians for eDiscovery collection and review.

3 Experiment Description

This year we participated in Topic 203 of the TREC Interactive Task. The hypothesis we were testing is that responsive documents would not be isolated to custodian mailboxes of people who would discuss whether the company would meet its financial obligations as identified by external factors such as their business title and job description. To test this hypothesis, we created two non-overlapping teams that independently identified and reviewed email for responsiveness. One team selected a group of custodians to review to simulate the exclusionary approach while the other team performed search and IR on the entire data set to simulate the investigative approach.

3.1 Two Team Structure

In order to properly evaluate the differences between an exclusionary and investigative eDiscovery process, we conducted the experiment with two teams which operated independently and approached the TREC Interactive Task differently.

- **Exclusionary ED Team:** The first team conducted their electronic discovery by culling the collection to four custodians similar to how custodians would be identified during the litigation process. The custodians were Enron employees who were determined to have responsive documents based on their job title or department before viewing messages in the data set.

- **Investigative ED Team:** The second team conducted the project across the entire Enron data set as if the entire organization's data was available for search and information retrieval. This is how ED can be conducted today with advances in search and IR technology.

Because the process of searching documents and iteratively refining search criteria depends in large part on the search results, we felt it was necessary to maintain strict separation between the two teams during the identification and review process. Thus, we could not only compare the number of responsive documents identified by each team, but also evaluate how the approaches determined each team's ability to refine their search criteria, discover new avenues of inquiry, and review documents.

	Exclusionary (Narrow) eDiscovery	Coverage	Investigative (Broad) eDiscovery	Coverage
Custodians	4	3.8%	104	100%
Total Emails Covered	557,077	18.8%	2,965,103	100%

Table 3. Team Custodian and Email Coverage

3.2 Information Retrieval Techniques

3.2.1 Exclusionary ED Team's Custodian-based Culling

Before both teams applied the search and IR techniques listed below, the exclusionary ED team identified potentially relevant custodians on an iterative basis to approximate the typical eDiscovery process. As a starting point, an initial set of four custodian mailboxes were chosen by examining Enron's corporate structure and identifying custodians who, by the team's judgment, likely had knowledge of Enron's financial projections and performance. The number of custodians selected was chosen in order to approximate the number of custodians, in proportion to the population, that are reviewed during an actual eDiscovery investigation. The exclusionary team arrived at their custodian figure by analyzing the data in the TREC Enron Corpus, taking into account additional data, and choosing a more conservative figure to partially account for the variability in different cases and organizations.

	Enron FERC Production Corpus	**Exclusionary Team study**
Comparison 1: initial custodian selection v. number of total employees in Enron for FERC production v. exclusionary ED experiment		
Initial custodians / Enron employees	2 / 22,000	4 / 22,000
Percentage of total employees	0.009%	0.018%
Comparison 2: final custodian selection v. number of total available custodians for FERC production v. exclusionary ED experiment		
Final custodians / total available custodians	148 / 22,000	4 / 104
Percentage of available custodians	0.673%	3.846%

Table 4. Enron FERC Production and Exclusionary ED Team custodian coverage

We used two comparison figures in determining the number of custodians in our study: (a) the initial number of custodians across an enterprise of Enron's size and (b) the percentage of custodians across all potential custodians in the collection. In both cases, we elected to use a more conservative figure than the numbers would indicate, in the sense that we chose a higher number of custodians in order to more conservatively test our hypothesis that exclusionary ED would yield substantially fewer responsive documents than investigative ED.

In the first analysis, we sought to understand how many custodians were initially selected in original Enron Western U.S. Energy Crisis case. Table 2 shows the number of custodians per collection we identified in the collection based on dates used in the collected PST and NSF filenames. We recovered collection date data for 98 of the possible 104 custodians which indicated two custodians were selected in the initial August 2000 collection, a date which corresponds to the litigation action from the San Diego Gas & Electric Company. This analysis suggested that in order to accurately model a civil litigation in an organization's of Enron's size, two or more custodians would be included during the culling process.

Next, we analyzed the total number of custodian mailboxes collected in the case as a percentage of the total number of possible custodians, with the latter being the set used for an investigative eDiscovery. For this analysis, we sought to confirm that the number of custodians selected by the exclusionary ED team would reflect the percentage of custodians collected during the Enron litigation. While the TREC Enron Corpus had 104 custodians, we took a broader view and considered the number of custodians identified in other distributions of the Enron Corpus, namely the CMU Enron Corpus which is also based on the FERC release. In this corpus, 148 custodians were identified, or 42% more custodians, to provide a more accurate figure for the total number of custodians covered. We compared this with the total possible number of custodians, 22,000 in this case, and then applied the ratio to the collection we were working with. With this higher custodian count, the FERC production covered 0.673% of possible custodians. For the TREC Enron Corpus we were working with, this represents less than 1 custodian of the 104 identified in the TREC Enron Corpus we were studying, or 0.962%. From this analysis, we wanted to study one or more custodians in our research.

Combining the two sets of analyses of the production from the actual case with our desire to model a more conservative exclusionary eDiscovery, we decided to collect and analyze four custodian mailboxes, twice the figure indicated by the first analysis and four times the figure indicated by the second. We felt this better approximated the effects of actual litigation and was very conservative, for the purposes of testing our hypothesis, when compared to the actual numbers of custodians in both the TREC Legal Track and CMU Enron Corpuses.

An additional consideration in choosing a larger percentage of custodians for this study is that the TREC Enron Corpus represents email that has been deemed responsive and produced for the FERC Western U.S. Energy Crisis investigation. One could argue that this would indicate a higher percentage of custodians would have relevant information. We considered this and noted that the complaint and topic for the TREC study are separate and distinct from the FERC requests under which the productions was originally made.

3.2.2 Search and IR Techniques Common to Exclusionary and Investigative Teams

Both teams had access to the same search and information retrieval tools. The system comprised several IR techniques which refined the data set and narrowed down the documents for manual review. The TREC Enron corpus contains variability in the types of patterns that are considered responsive for this topic. Due to this variability, the teams found that a standard Boolean keyword search alone was not sufficient to identify all of the types of responsive documents. However, the keyword search technique was useful in producing an initial subset of documents that contained many responsive documents. A series of Boolean searches was used by each team to generate this document subset.

The subset was indexed separately and stemming was applied to the search tokens to decompose them to their root form. A concept semantic index was first built using the full-text index, and then augmented with latent aspect using various IR cluster algorithms, e.g. PLSI, LDA, and KNN. [xii, xiii, xiv, xv] Documents were clustered according to similar semantic meaning or theme using a hybrid vector and latent space model. The topic clusters produced additional documents with novel keyword combinations that were used to repeat the process iteratively. We also carefully studied the correlation between documents, topics, and words, using advanced visualization graphics to help validate and re-train the model to discover hidden relation between entities.

The overall process is outlined in Figure 2.

- **Entire TREC Enron Corpus:** The modified custodian-based Enron corpus was used as a starting data set for both teams.
- **Cull by custodian mailbox:** This step was performed by the Exclusionary ED Team only and created a data set for investigation consisting of 4 of the 104 possible custodians.
- **Lexical Analysis:** Keyword searches and lexical analysis was applied to identify an initial set potentially relevant documents.
- **Manual Assessment:** Reviewers in both teams manually reviewed the results of the initial, independently created keyword searches and lexical analyses.
- **Hybrid Semantic Vector Analysis:** Once the initial assessment was approved, a concept index was created to provide further analysis of the data.
- **Cross Check:** A final manual review was performed using the concept index to identify relevant documents.
- **Relevant Document ID List:** A document list by unique ID was generated. Additional reports were generated mapping relevant documents to custodian mailboxes for the purposes of reviewing relevant email coverage in the potential custodian population.

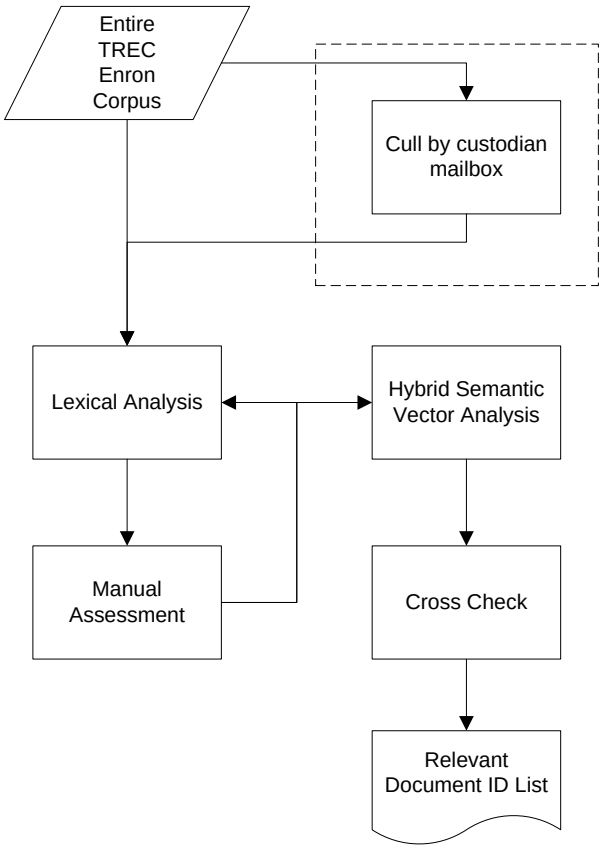

Figure 2. Search and IR Methodology

4 Document assessment

In this year's TREC Interactive Task, both teams met independently with the Topic Authority for our topic. The teams did not share the contents of their discussions with the TA to the other team until after the project was completed, to avoid any cross-pollination of feedback that would jeopardize the experimental validity. The discussions clarified the TA's perspective of document relevance. For instance, these discussions clarified the positions on the following:

- Any general discussion about creating or generating forecasts or models without a reference to the performance is not within scope according to discussions with the Topic Authority.
- Discussions about making any kind of financial forecasts, company-wide, or any way up the company are all within scope.
- Comparing a financial performance metric to the past or to another company is not responsive.
- Even indirect reference to a performance metric such as "we beat the street" or "our earnings are going to make Wall Street ecstatic" is responsive because these statements compare performance against an implied model.
- Statements made by individuals including indirect discussions of performance such as "we didn't close as many deals as we thought we would" or "we didn't do as well as we thought we would" are potentially responsive. To be responsive, the comparison needs to be more than just a vague idea, it needs to have been modeled or projected and quantified.

Based on independent discussions with the Topic Authority, both teams were instructed that items would need to include both a measure of performance from a projection and an indication of whether or not company would meet the metric previously set forth in the projection in order to be considered responsive. Therefore, a statement of actual performance must be compared to a prior projection, forecast, or model in order to meet the responsiveness criteria. Based on this strict requirement, items that contain only metrics associated with projections, or references to actual performance would not be considered responsive.

5 Results

In this section we describe our results. Our overall evaluation method seeks to prove the hypothesis that a more inclusive electronic discovery produces a higher number of responsive documents than an exclusionary process.

The results of the experiment after accounting for the TREC topic authority's final assessments are listed in table 5. The exclusionary team reviewed documents from four custodian mailboxes and identified 49 unique responsive messages after adjusting for the final assessment. The inclusive team performed information retrieval across all messages and identified 302 unique responsive messages after adjusting for the final assessment. The exclusionary team's results comprise 16% of responsive documents identified by the inclusive team as a whole. Or put another way, the inclusive team identified 616% documents more than the exclusionary team.

	Exclu-sionary ED	**Investi-gative ED**	**Exclu-sionary / Investi-gative %**	**Investi-gative / Exclu-sionary %**
Custodian Coverage	4	104	4%	2600%
Unique Relevant Emails	49	302	16%	616%
Relevant Custodians	4	77	5%	1925%

Table 5. Summary results for both teams

All four custodians selected by the exclusionary team contained responsive messages. Of the 104 custodians across the data set, the investigative team identified 77 custodian mailboxes containing responsive messages in total. Therefore, the exclusionary team examined 5.2% of the individuals whose mailboxes contained responsive messages according to the investigative approach. Distributions of the responsive documents are shown in Figures 3 and 4 below and indicate that relevant documents were well distributed across the custodian population.

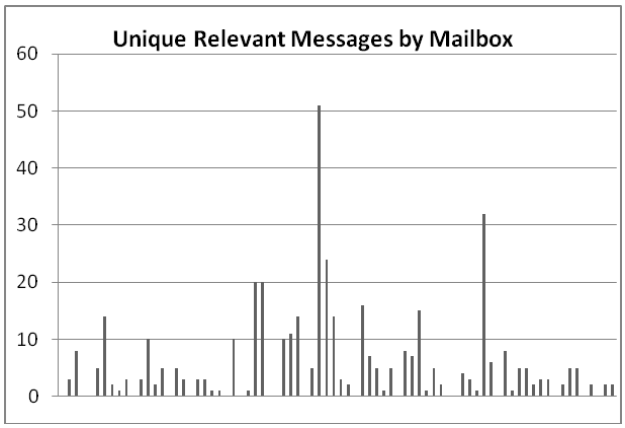

Figure 3. Distribution of unique messages across custodian mailboxes

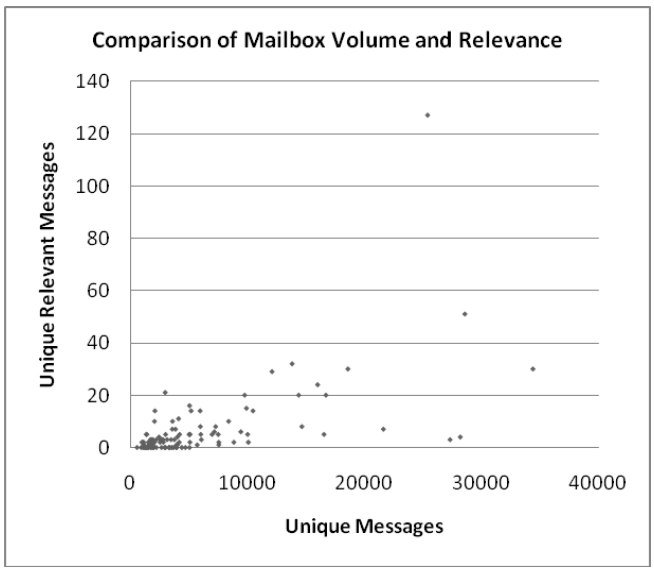

Figure 4. Comparison of Mailbox Volume and Relevant Message Volume

Table 6 examines the email volume for the custodians selected by the exclusive ED team. The team examined 4 of 104 custodians which comprised 3.8% of the individuals in the data set. Those individuals were involved in 18.8% of email conversations. All four custodians they selected ex ante were in the top 50% percentile for email volume. While the exclusive ED team was able to find a number of relevant custodians and messages, the inclusionary approach was able to find more relevant messages with less iteration.

Name	Title	Unique Email Volume	Email Volume Rank	Investigative ED Relevance Rank
Kenneth Lay	Chairman, CEO	9,902	18	17
Jeff Skilling	CEO	15,988	10	12
Steven Kean	VP, Chief of Staff	25,391	5	1
Rod Hayslett	VP, CFO, Treasurer	3,834	51	44

Table 6. Exclusionary ED Team's per-Custodian Results

6 Discussion

Based on the results, the investigative ED team identified a higher number of responsive documents across a higher number of custodians which supports our initial hypothesis. Even if the exclusionary team had selected the four custodians with the highest number of responsive documents, this approach would overlook over half of the responsive documents identified by the investigative team resulting in a lower recall rate and potentially under producing relevant documents. The higher number of responsive documents for the broader scope suggests that organizations conducting electronic discovery may miss a substantial number of documents by relying on custodian searches.

Both teams received equal resources, not only in terms of IR tools but also in number of reviewers. Therefore, even though the inclusive team had a larger amount of data to search across than the exclusionary team, they had the same number of people and resources for the review of documents. Further studies may examine whether an investigative approach to electronic discovery can reduce review times and / or produce a qualitatively different result in addition to a quantitative result, in light of the search and IR technology available. That is outside the scope of this paper.

The exclusionary team only examined 18.8% of the documents while they identified 16% of all responsive documents identified by the investigative team as a whole across the population, suggesting that the custodians chosen contained proportionally more responsive documents than the data set as a whole. Our results were somewhat dependent on the composition of the data set. We assumed that responsive documents would be distributed more broadly than an initial custodian selection

would suggest; something which is confirmed by the distribution of responsive documents across 78 mailboxes.

One factor we considered was possible selection bias due to the nature of the TREC Enron corpus being a subset of email data produced by FERC as part of its investigation into the company. The corpus is not simply a random sample of email data from Enron. The custodians and their data were requested precisely because they were deemed responsive to the investigation. As a result of the selection mechanism used to generate the corpus, we hypothesized that one may expect a higher number of responsive documents to be distributed among the custodian mailboxes we evaluated in the TREC Enron corpus than in the Enron Corporation as a whole. However, we concluded that this was not a significant factor in our experiment because the FERC investigation dealt with a wholly different topic from the TREC interactive task.

In some data sets, responsive documents may be isolated in a few custodian mailboxes whereas other data may be more homogenous with responsive documents distributed evenly among more custodians. Obviously this has implications depending on the number of custodians selected for exclusionary ED, and whether or not those custodians reflect the location of responsive documents. The selection of custodians themselves is also of interest as some custodians that have responsive documents may be overlooked. The nature of ED is that the understanding of the matter often evolves over time and the imperfect state of knowledge at the time of culling may increase the risks of under production and spoliation when making fundamental decisions in the response, including choosing which custodians to include and exclude. Therefore, we expect that results would vary depending on the data and state of knowledge of the matter at the point of culling. While our experiment presents a novel investigation by examining the effectiveness of two ED approaches, it would be interesting to further explore the results of this experiment on a different data set with perhaps a larger amount of data as well as to examine the effect of varying levels of knowledge at the time of culling.

7 Conclusion

Mapping the messages in the Enron data set to the original employee mailboxes and conducting two eDiscovery approaches is a novel approach to the TREC Interactive Task that we feel more closely resembles real-world eDiscovery. Our initial findings are that investigative approaches can provide significantly more responsive documents across a wider number of custodians and that the exclusionary approach may result in underproduction. Specifically, investigative ED located 516% more relevant documents and 1825% more custodians in the initial review than exclusionary ED in our study. Given advances in search and IR techniques that enable wide-spread adoption of investigative ED, we feel this area can benefit from additional research to provide information for ED researchers, ED practitioners, and the courts.

---

[i] Interactive Task Guidelines – TREC 2009 Legal Track, 2009. Availabile at http://trec-legal.umiacs.umd.edu/LT09_ITG_final.pdf.

[ii] George J. Socha Jr., Esq (2009). Bringing e-Discovery in-house: Risks and Rewards, Socha Consulting LLC, Saint Paul, Minnesota.

[iii] P. Grimm. Victor Stanley, Inc. Plaintiff vs. Creative Pipe, Inc., et al. Defendant, Civil Action No. MJG-06-2662. In the United States District Court for the District of Maryland. May 2008.

[iv] J. R. Baron et al (2007). The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery. August 2007. Available at http://www.thesedonaconference.org/dltForm?did=Best_Practices_Retrieval_Methods___revised_cover_and_preface.pdf

[v] J. R. Baron et al (2007). The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery. August 2007. Available at http://www.thesedonaconference.org/dltForm?did=Best_Practices_Retrieval_Methods___revised_cover_and_preface.pdf

[vi] C. Hogan, D. Brassil, S. Rugani, et al. H5 at TREC 2008 Legal Interactive: User Modeling, Assessment & Measurement. In *The Seventeenth Text Retrieval Conference (TREC 2008) Proceedings*, November 2008.

[vii] G. Amati et al. CNIPA, FUB and University of Rome "Tor Vergata" at TREC 2008 Legal Track. In *The Seventeenth Text Retrieval Conference (TREC 2008) Proceedings*, November 2008.

[viii] Y. Zhang, F. Scholer, and A. Turpin. RMIT University at TREC 2008: Legal Track. In *The Seventeenth Text Retrieval Conference (TREC 2008) Proceedings*, November 2008.

[ix] Interactive Task Guidelines – TREC 2009 Legal Track, 2009. Availabile at http://trec-legal.umiacs.umd.edu/LT09_ITG_final.pdf.

[x] Beth MacLean and Peter Elkind. Smartest Guys in the Room: The Amazing Rise and Scandalous Fall of Enron, 2003, ISBN 1591840082.

[xi] Jitesh Shetty and Jafar Adibi. Enron ex employee status report. Available at http://www.isi.edu/~adibi/Enron/Enron.htm.

[xii] Thomas Hofmann. Probabilistic Latent Semantic Indexing, In *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99)*, 1999.

[xiii] Blei, David M.; Andrew Y. Ng, Michael I. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* **3**: 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

[xiv] Girolami, Mark; Kaban, A. (2003). On an Equivalence between PLSI and LDA. In *Proceedings of SIGIR 2003*. New York: Association for Computing Machinery. ISBN 1581136463.

[xv] Okolica, J., Peterson, G. and Mills, R. (2006) *Using PLSI-U to Detect Insider Threats from Email Traffic*, Springer-Verlag, New York, NY, pp. 91-104