-- THIS THESIS IS UNDER CONSTRUCTION –

IT IS POSTED AS AN EXPERIMENT IN

WEB-BASED INFORMATION SHARING

Comments are welcomed.

KNOWLEDGE DISCOVERY IN CORPORATE EMAIL:

THE COMPLIANCE BOT MEETS ENRON

by

K. Krasnow Waterman

Juris Doctorate, Benjamin N. Cardozo School of Law (1989)

Bachelor of Arts, University of Pennsylvania (1979)

Submitted to the Sloan School of Management in Partial

Fulfillment of the Requirements for the Degree of

Master of Science in the Management of Technology

at the

Massachusetts Institute of Technology

June 2006

Signature of Author:____

MIT Sloan School of Management

March x, 2006

Certified by:____

, Thesis Advisor

Certified by:____

, Thesis Reader

ABSTRACT

[TO BE WRITTEN]

Note: The data used in this thesis is pre-existing and publicly available; it is exempt from the federal regulation on the Protection of Human Subjects. 45 CFR § 46.101(b)(4).

TABLE OF CONTENTS

Preface....5
Introduction....7
I. Email: Population Explosion....8
II. Email Challenges for Compliance Managers....10
A. Workplace Emails are Usually Not Private....11
B. Obligation to Proactively Search for Violations....12
C. Personal Use of Corporate Resources....13
D. Evidence of a Hostile Environment....14
E. Other Employment Issues....14
III. Knowledge Discovery: Meaning from Chaos....15
A. What is "Knowledge Discovery"?....15
B. Timing of Knowledge Discovery Development....16
C. Value of Knowledge Discovery Industry....16
D. Knowledge Discovery Tools....17
1. Pre-processing....17
2. Processing....19
3. Presentation....21
IV. Knowledge Discovery Attacks Email Problems....21
V. Boon for Researchers: Enron Emails Made Available....22
A. The Number of Available Emails....22
B. Understanding the Cleansed Set....24
1. Unique record identifiers....25
2. Changes to Email Addresses....25
3. Conversion of Time Stamps....25
4. Duplicates in the Original Dataset....26
5. De-duplication and the Loss of Location Data....26
6. Summary Statistics....26
C. ENRON Knowledge Discovery Work to Date....27
1. Occurrence Counts....27
2. Automated Categorization According to Personal Preference....29
3. Categorization According to Single Standard....30
4. Social Network Analysis....31
VI. Thesis Experimentation....33
A. Occurrence Counts....33
1. Hostile Environment....34
2. Personal Use of Corporate Resource....34
3. Limitations....36
B. Boolean search....37
1. Evidence of a Hostile Environment....37
C. Thread search....38
D. Latent Semantic Indexing....38
E. [InBox?]....38

Preface

Over the last twenty-five years, I often have been responsible for the management of Information and Information Technology. During those years, I have observed the myriad of advances. Punch card systems became interactive systems; serial processors became multi-processors; the 32 Megabit $1million mainframe became the 2 Gigabit $1,000 laptop; the 300 baud suction-cup modem became the wireless multi-Gigahertz modem card; programming advanced from machine language in hexadecimal code to nascent natural language systems; and so on and so on. Generally, the Information Technology industry has made it significantly easier, faster, and cheaper to collect and store data. The result is a massive increase in available data; it has been estimated that the volume equivalent of the Library of Congress is created digitally every 15 minutes.[1] One of the major challenges today is how to make sense of so much data.

Email, in particular, poses a tremendous challenge for organizational knowledge management. Business transactional data remains in corporate databases, but "soft" business – planning, human resources, marketing, etc. – discussions occur increasingly through email and out of formal organizational records. These individual email accounts are not only another form of "silo" but are generally lost altogether when employees leave the company. Also, email fosters an informality that may reduce productivity or lead to corporate liability. Corporations need "knowledge discovery" tools that make it possible to understand what is in their email repositories. This would allow them to both extract higher business value for their daily work and to identify potential problems at early stages.

A pre-research review of the field indicated that most such tools have been built either for spam-filtering or for the purpose of retroactive analysis: support for litigation, intelligence, or archival activities. The author wished to know if the same technologies could support business managers in active management activities. In an effort to understand how knowledge discovery across an organization's stored email might support operations management, I surveyed existing research and performed some experimentation of my own. Specifically, I focused on the more than half a million emails of Enron management released by the Federal Energy Regulatory Commission. The release of that information has provided a robust real-world dataset for researchers to study.

Most of the time that I have not spent in Information Technology management or broader operations management has been time I have spent in the active practice of law. Because I have pre-existing professional experience as a trial attorney regarding employee misconduct, I focused on finding evidence of such misconduct that could lead to liability. Also, understanding the scope and volume of the personal use of the corporate email

[1] "Eternal Bits: How can we preserve our digital files and preserve our collective memory?" by Mackenzie Smith, published in *IEEE Spectrum*, p. 22, para. 1 (July 2005) (http://www.spectrum.ieee.org/WEBONLY/publicfeature/jul05/0705bit.html).

resource could help the operations manager understand losses due to system and system support costs as well as lost productivity.

These areas of inquiry are selected because they are topics of which I have knowledge. However, the purpose of the study is not only to determine the relative efficiency and effectiveness of the technologies studied and the ability to perform proactive compliance activities through email analysis. It is intended as a step along the road of inquiry regarding the effectiveness of cross-organizational access to email. The study is intended to provide insight into whether any person with knowledge of a particular category of work effort can supplement their knowledge – finding other existing projects on the topic, other employees with similar interests, or obtain legacy knowledge – through email knowledge discovery.

This thesis assumes that the reader is a business professional rather than a technical professional. I assume no prior knowledge of any of the technology discussed and provide explanations of all terms. I describe how the developments of email and knowledge discovery are driving changes in law and legal obligations. Then, I describe the Knowledge Discovery research performed on the Enron emails to-date. Based upon my own experience, I provide insights into the ways in which those tools or activities could be applied to operations management issues. Where others have provided their tools, I have tried to use them to further the understanding of Knowledge Discovery applied to these human resources issues. And, I have identified and used other tools that had not previously been used on the Enron data.

Introduction

This thesis addresses a confluence of law and technology in recent years. In one generation, employment law and email have both matured tremendously. Many people don't realize that there was very little law regulating employment before the Civil Rights Act of 1964 and that law in this area is still changing rapidly. And while email was first developed in the late 1960's, the global adoption of the medium really began with the introduction of the World Wide Web in the 1990's. As email gained dominance, business and personal communications migrated to this medium.

Today, the human resources manager is ultimately responsive to every inappropriate employee act, whether that act involves violating government regulations, company policy, or the rights of others. Today, a significant amount of that inappropriate conduct takes place in, or is evidenced through, email. How, then, is the human resources manager to become aware of such conduct? Should the human resources manager wait for one employee to turn in another employee? Does the human resources manager, or a compliance officer, have the obligation to proactively search for evidence of such inappropriate conduct?

A series of changes and clarifications in employment law appear to create an obligation to affirmatively search for inappropriate conduct. Luckily, another series of technical advances will make this possible. The field of Knowledge Discovery, which was formalized in the late 1980's and has been progressing ever since, provides tools that find and express meaning from very large collections of data. If that software could be harnessed as a "bot" – an automated program that performs like a person – what would it look for? How would it look?

In 2003, the Federal Energy Regulatory Commission released more than a half million emails of the senior managers of Enron Corporation. This was the first major repository of emails available to Knowledge Discovery researchers. This paper describes the history of all of the precursor factors and how they have been or could be applied to the emails of the senior managers of Enron. It uses the Enron email corpus to bring to life the concept of the "Compliance Bot."

I. Email: Population Explosion

Email is a relatively new phenomenon. In the 1960's, as people began to share access to computers, they realized that they could communicate with each other as well.[2] In 1971, the first inter-computer email was sent on ARPANET, a government-created precursor to the internet.[3] It has been suggested that because of the general cultural shifts of the 1970's – from the "Man in the Gray Flannel Suit" of the 1950's to the hippies of the 1970's – email is a medium in which informality has always been acceptable. Although both ARPANET and USENET (a university-funded internet precursor) were offered in a work environment, both had a significant percentage of email traffic not related to work activities, including topics such as chess, science fiction, recipes, jokes, rock and roll, and sex. One company participating in USENET complained that it was turning into "electronic graffiti." Email was a success from inception and grew rapidly. By the early 1980's, ARPANET email traffic was essentially equally in size to file transfer traffic. And, USENET creators had under-predicted the level of email traffic by about 2,000%.

By the mid-1980's, email had been adopted by other technology platforms. For example, by 1982, IBM had introduced a prototype of the Professional Office System (PROFS), a mainframe computer application that provided mail; PROFS was a major industry email application for many years.[4] In 1985 the Wang company, which sold word processing systems that were much less expensive than mainframes and accessible to smaller companies, introduced Wang OFFICE which integrated internal company email with word processing.[5] By 1988, Wang recognized that companies would need to connect multiple email systems and the software was offering gateways that would permit connections to the IBM and DEC mail systems.

Also in 1988, experimental commercial use of the internet began with connection of MCI Mail to NSFNET (another government project).[6] Compuserve began offering service in 1989. At about the same time, Sir Tim Berners-Lee created the World Wide Web and, in 1991, he posted the first website.[7] In 1993, AOL (America Online) began offering service.[8]

Email usage and storage has become so popular that Microsoft discovered that the size limit it had set for a personal email file was not big enough. Through 2002, the size

[2]

[3] "History of Electronic Mail," Richard T. Griffiths, Leiden University, *History of the Internet*, Chapter 3 (last update Oct. 11, 2002) (http://www.let.leidenuniv.nl/history/ivh/chap3.htm).

[4] "100 Years of IT," Frank Hayes, *Computer World* (April 5, 1999) (http://www.thocp.net/reference/info/100_years_of_it.htm).

[5] "Wang OFFICE," Vincent Flanders, *Access 88: The Magazine for Wang OFFICE Users* (Feb. 1988) (http://www.vincentflanders.com/2-88.html).

[6] "Email History," Dave Crocker, posted as part of "Living Internet" (http://www.livinginternet.com/e/ei.htm).

[7]

[8] "Email History," Dave Crocker, posted as part of "Living Internet"

limitation for an individual's email file on Microsoft's Outlook was 2 Gigabits,[9] roughly equivalent to the storage needed for more than 16,000 20 page documents[10] or 642 copies of the e-book version of Isaac Asamov's *I, Robot.*[11] Yet, individual power users were bumping up against that limit, getting locked out, and losing emails.[12] In its 2003 version, the storage limitation was increased by 1000% and now sits at a 20Gigabits.[13]

By 2005, one market study determined that corporate users were averaging 133 email messages (sent and received) per day, adding 294MB of storage requirements per month.[14] The same group[15] evaluated the cost of messaging in 1998 and again in 2003, finding an average total cost per user (e.g., administration, acquisition, training, storage) per year for Microsoft Exchange jumping from $64.93 to $221.42 during that 5 year period.[16] By 2003, storage costs alone were $0.07 per MB for a Microsoft Exchange user; in the 2005 environment this would equate to approximately $17.43 per user for a year's storage of a month's emails. In companies where law or policy require full archival, this equates to $113.29[17] per average user per year for data storage costs alone.

[9] *See*, "The .pst file has a different format and folder size limit in Outlook 2003," Microsoft Help and Support webpage (http://support.microsoft.com/?kbid=830336).

[10] *See*, "Chapter 5: Data Transfer Rates: A Primer," Texas State Library and Archives Commission, *Wireless Community Networks: A Guide for Library Boards, Educators, and Community Leaders*,(explanation in "Large Units" subsection that a 20 page word-processed document can take up to 60,000 bits) (http://www.tsl.state.tx.us/ld/pubs/wireless/chapter5.html).

[11] Calculated by dividing 2,000,000,000 by 3,111,000 based upon Amazon.com listing the ebook download as 3111 KB (http://www.amazon.com/gp/product/B0002CH6J4/ref=ase_ebookuniverse05-20/104-6419261-9984764?s=ebooks&v=glance&n=551440&tagActionCode=ebookuniverse05-20).

[12] *See, e.g.,* "FAQs & Tips for Outlook 2002," University of New Hampshire, Computing and Information Services webpage (last updated Aug. 9, 2005) (describing system lockout at 2GB) (http://www.outlook.unh.edu/faq/Faq2002.html); "Outlook 2002 Hotfix Addresses 2GB Size Limit," Sue Mosher, Contributing Editor, *Windows ITPro Magazine* (Sept. 13, 2001) (explaining that Microsoft had responded to user problems by releasng software that would keep users from reaching the maximum file size) (http://www.windowsitpro.com/Article/ArticleID/22509/22509.html); and .

[13] *See*, "The .pst file has a different format and folder size limit in Outlook 2003," Microsoft Help and Support webpage (http://support.microsoft.com/?kbid=830336).

[14] "Taming the growth of email: An ROI analysis," a white paper by The Radicati Group, Inc., for the Hewlett-Packard company (2005) (https://h30046.www3.hp.com/campaigns/ 2005/promo-evolution/1-1LRYR/images/Preview_Radicati.pdf).

[15] "Messaging Total Cost of Ownership," by Sarah Radicati & Laura Venutura, The Radicati Group, Inc., p. 4 (1998) (www.terracetech.com/jp/data/ Messaging%20Total%20Cost%20of%20Ownership.pdf) . and "Messaging Total Cost of Ownership -2003: in Enterprise and Service Provider Environments," The Radicati Group, Inc. (2003) (www.sun.com/aboutsun/media/ presskits/aiim**2003**/**2003**TCOSummary).

[16] *Id*., at p. 2 and fn. 2.

[17] Calculated as **[figure out how to show as** ∑ (n+1) from 1 to 12 = 6.5/mon];*see also,* (finding £80[17] per user for MS exchange email services in 2003) (http://www.computerweekly.com/Articles/2003/05/06/194340/Linuxe-mailset-upslashescoststo%c2%a38peruser.htm).

II. Email Challenges for Compliance Managers

Just as the email phenomenon began growing in the 1960's, so too did the field of employment law. After the Civil War, the first Equal Rights Act was passed, granting to all citizens the rights which had previously been exclusive to "white" citizens.[18] In the early part of the twentieth century, just a few laws were passed that regulated the overall employer/employee relationship: laws such as the Fair Labor Standards Act,[19] which sets overtime pay requirements, and the National Labor Relations Act,[20] which established the rights of employees to collective bargaining and establishing unions. With the passage of the Civil Rights Act of 1964,[21] the era of modern employment law began. As recently as the 1970's, employment law was not yet a subject taught in most law schools.[22]

Since 1964, Congress and the States have passed a flurry of laws regulating employer/employee relationships. The law now prohibits discrimination based upon race, national origin, gender, religion[23] and, in some circumstances, age,[24] disability, [25] pregnancy,[26] familial status,[27] or sexual orientation.[28] The law requires employers of a particular size to grant employees leave to handle family matters.[29] There are laws detailing the manner in which benefits, pensions, and insurance[30] can be provided. And,

[18] *See,* ch. 114, § 16, 16 Stat. 144 (enacted May 31, 1870) (precursor to 29 U.S.C. § 1981, enacted Nov. 21, 1991) (http://www.law.cornell.edu/uscode/html/uscode42/usc_sec_42_00001981----000-.html and http://www.law.cornell.edu/uscode/html/uscode42/usc_sec_42_00001981----000-notes.html).

[19] 29 U.S.C. § 201, *et seq.* (enacted June 25, 1938) (http://www.law.cornell.edu/uscode/html/uscode29/usc_sup_01_29_10_8.html) and http://www.law.cornell.edu/uscode/html/uscode29/usc_sec_29_00000201----000-notes.html).

[20] 29 U.S.C. § 151, *et seq.* (enacted July 5, 1935) (http://www.law.cornell.edu/uscode/html/uscode29/usc_sec_29_00000151----000-.html and http://www.law.cornell.edu/uscode/html/uscode29/usc_sec_29_00000151----000-notes.html).

[21] 42 USC § 2000a, *et seq.* (enacted July 2, 1964) (http://www.law.cornell.edu/uscode/html/uscode42/usc_sec_42_00002000---a000-.html and http://www.law.cornell.edu/uscode/html/uscode42/usc_sec_42_00002000---a000-notes.html).

[22] *See, e.g.,* "Introduction – Including a Brief History of Employment Law & Practice," William P. Bethke and James W. Griffin, *Personnel Practices and Policies: Understanding Employment Law* (Nov. 2000) ("When the oldest author of this handbook was going to law school - graduating in 1978 - there were no courses in "employment law." Legal digests and encyclopedias did not mention "employment," but instead, "Master and Servant." Labor law was treated as its own, rather arcane, subject. Some law schools had just begun offering employment discrimination courses. It was not that employment lacked an interesting, complex legal history - quite the opposite. But outside specialized areas - unionized work places, civil service systems, workers compensation and the nascent subject of discrimination - employees had few rights.") http://www.uscharterschools.org/gb/personnel/intro.htm.

[23] The Civil Rights Act of 1964 outlaws discrimination based upon race, national origin, religion, and gender. **(insert cite);**

[24] Age Discrimination in Employment Act outlawed discrimination against people over the age of 40 **(insert cite) (1967)**;

[25] Americans with Disabilities Act severely limits the circumstances under which disability may be considered in an employment decision **(insert cite) (1990)**;

[26] Pregnancy Discrimination Act (1978)

[27] **Cite to state laws**

[28] **Cite to state laws**

[29] Family Medical Leave Act **(insert cite)** (1993)

[30] Employment Retirement Income Security Act (ERISA) regulates all three.

there are laws regulating employment contracts, background investigations, termination procedures, payment of salary, and a myriad of other topics.[31] I have heard that, in recent years, the US Supreme Court has rendered more decisions on the subject of employment law than on any other subject.

In addition to all of these laws that regulate how an employing organization should treat its employees, there is quite a bit of new law allocating responsibility to the employer for the conduct of its employees. Since the 19xx's, stockholders have been able to bring lawsuits against companies for conduct which inappropriately diminishes the value of the corporation.[32] In the employment arena, employers can be sued for negligence, for failing to adequately check the background of a person they hire if that person goes on to cause harm.[33] Some states hold the employer liable, if a supervisory employee attempts to get subordinates to break the law – by doing such things as altering the financial records or **(add another example)**.[34]

As a matter of pure statistics, one would have to assume that the sheer size of email as a means of communication would have to be the repository of a certain amount of evidence of inappropriate conduct. In my experience, the numbers reflect more than pure chance. As courts began to find corporations liable, corporations have increasingly trained their management employees about what constitutes inappropriate conduct. Unfortunately, though, sometimes management employees take that instruction as a cautionary tale of what not to get caught doing rather than as what not to do. To some, email seemed to provide the equivalent of the private club, the locker room, the closed door – an apparently private place to continue conducting the same inappropriate acts.

A 2004 study of nearly more than 800 companies found that more than one in eight had been sued because of employee emails; these lawsuits included claims of sex and race discrimination and harassment as well as hostile environment.[35] The number could be significantly higher, as another quarter of the survey respondents did not know the answer to the question.

A. Workplace Emails are Usually Not Private

In casual conversations, people often tell me that their emails at work are private and, sometimes, proceed to describe a system of protection that parallels the criminal law on warrantless seizures. Generally, these people are mistaken. At present, there is no single federal law that addresses the question of privacy for workplace email.

[31] **Cite, eg. To AZ laws in these areas**

[32]

[33]

[34]

[35] "2004 Workplace E-mail and Instant Messaging Survey Summary," American Management Association, p.1 (2004) (http://www.amanet.org/research/pdfs/IM_2004_Summary.pdf)

The Electronic Communications Privacy Act of 1986 makes it illegal to intercept electronic communications between two people.[36] However, it has an exception for employees of the company that provides an electronic communications service, allowing them to intercept, use or disclose the communications as necessary to perform the service or protect the rights and property of the service provider.[37] At least one court has held that a company that provides email functionality is covered by this exception. Perhaps, more importantly, there is an exception for consent.[38] And, with few exceptions, an employer can condition employment on the waiver of a right someone otherwise possesses. This is not so unusual as it might sound at first. A person accepting a job that provides access to trade secrets, patient medical histories, or attorney-client secrets is required to agree to abridge his or her right of free speech to the extent of not ever talking about those facts without the employer's permission. And, employees at any number of convenience stores and restaurants have voluntarily waived possible privacy rights when they agreed to bring their personal possessions to work only in clear plastic purses and backpacks. In the case of email, employees are often told of email monitoring at new employee training, in an employment handbook, and/or frequently through a pop-up window at the time of log-on. In 2004, more than half of nearly 1,000 corporations surveyed provided email policy training to their employees.[39] In 2005, more than half were monitoring employee emails.[40]

In addition to voluntary access to employee emails, employers are also subject to involuntary searches. The 2004 survey indicates that more than 1 in 5 have had employee email subpoenaed by a court.[41] The number could be significantly higher, as another 20% did not know if they had been subpoenaed. The issue of access to corporate digital records (including emails) as part of the litigation process has become so important, that **[STOPPED HERE]**t .[42]

B. Obligation to Proactively Search for Violations

We know that employers may look at employee emails and sometimes do. Do employers have to look at emails? Are they obligated to attempt to find wrongdoing therein? While there may not be a single law or court decision which says that they must, there is definitely a trend in law to create such an obligation.

[36] 18 U.S.C. § 2511(1) (http://www.law.cornell.edu/uscode/html/uscode18/usc_sec_18_00002511----000-.html).

[37] 18 U.S.C. § 2511(2)(a)(i) (http://www.law.cornell.edu/uscode/html/uscode18/usc_sec_18_00002511----000-.html).

[38] *Id.*, at § 2511(2)(d).

[39] "2004 Workplace E-mail and Instant Messaging Survey Summary," American Management Association, pp.2 & 4 (2004) (http://www.amanet.org/research/pdfs/IM_2004_Summary.pdf)

[40] 2005 AMA survey

[41] "2004 Workplace E-mail and Instant Messaging Survey Summary," American Management Association, p.1 (2004) (http://www.amanet.org/research/pdfs/IM_2004_Summary.pdf).

[42]

In 1998, the United States Supreme Court issued a decision[43] that created new obligations for employers. In that case, the Court decided that female lifeguards who had been subjected to offensive touching (ranging from putting an arm around them to touching their buttocks), lewd remarks (including talking about sex and asking to have sex), and offensive comments about women (including comments about non-employees and women generally) had been victims of hostile environment sexual harassment and entitled to relief. There are a number of issues raised in that case which are relevant to the inquiry in this paper. The Court focused on whether such conduct had been by the employee's immediate supervisor and/or someone above that supervisor in the direct management chain. It found the employer liable even though the conduct had taken place at a location (lifeguard stations) away from the rest of the organization; and the employees had not made formal complaints. Under certain circumstances the employer could defend against the claim by showing that it had "exercised reasonable care to prevent and correct promptly any sexually harassing behavior." This raises the question: if software is available that can find lewd and offensive comments being mailed from supervisors (or above) to subordinates, is the employer failing to exercise reasonable care if it does not utilize the software?

In 2002, the Enron scandal propelled Congress to pass the Sarbanes-Oxley Act.[44]

Extend to other violations???

create a *de facto* obligation to use real-time (or near real-time) KDD for other hostile environment or harassment?

Does availability of email KDD for litigation discovery and SOX compliance tools

Nearly half of corporations are subject to legal or industry regulation but nearly half of them either do not comply or do not know if they comply with related email retention requirements.[45]

and If appropriate real-time (or near real-time) knowledge discovery tools were available, a compliance manager could search for many things in corporate email **Talk about personal use of corporate resources:**

C. Personal Use of Corporate Resources

[43] *Faragher v. City of Boca Raton*, 524 U.S. 775, 118 S. Ct. 995 (1998).

[44]

[45] "2004 Workplace E-mail and Instant Messaging Survey Summary," American Management Association, p.3 (2004) (http://www.amanet.org/research/pdfs/IM_2004_Summary.pdf)

Corollary to 80's when corporations began to control use of long-distance telephone (loss of productivity combined with increased infrastructure costs; discovery of latter pays for hunting for the former)

In a 2004 survey conducted by the American Management Association, nearly all employees claimed that they use corporate email for personal reasons less than 10% of the time.[46] However, one company that that mines corporate email for litigation discovery estimates that non-work-related emails make up approximately 1/3 of email traffic.[47]

Examples from legal experience:

switchboard employee patching sibling through to international numbers
DIRECT: Talk about shopping, social plans, sports pools, jokes, etc.

D. Evidence of a Hostile Environment

Talk about pejorative and derogatory terms and jokes; predatory behavior; …still adding ideas

From experience:

Emails about the boss who made female employees sit on his lap during the Xmas party

E. *Other Employment Issues*

From experience:

emails about the junior employee who got so drunk at the Xmas party she was catatonic on the restroom floor and paramedics had to be called. Does the firm have liability for serving that much? Does the employee suffer from untreated alcoholism?

Talk about court decisions that say corporate email is not private (and company pop-ups with explicit statements to this effect]

[46] "2004 Workplace E-mail and Instant Messaging Survey Summary," American Management Association, p.6 (2004) (http://www.amanet.org/research/pdfs/IM_2004_Summary.pdf

[47] **[get permission to list this telephone call** (Feb. 10, 2006).

In 2004, nearly 80% of corporations surveyed had email content policies and more than half provided email policy training to their employees.[48]

Searching for what everyone knows is there

III.Knowledge Discovery: Meaning from Chaos

A. What is "Knowledge Discovery"?

The Association for Computing Machinery (ACM), the first computing society, founded in 1947 and currently sustaining over 80,000 members[49], is one of the world's premier professional computing organizations. The term "Knowledge Discovery" was coined at a 1989 ACM workshop.[50] ACM says that Knowledge Discovery addresses the issue "how does one understand and use one's data"[51] in the context of massive data collection. More fully, it is the "process of finding new, interesting, previously unknown, potentially useful, and ultimately understandable patterns from very large volumes of data."[52] Most simply, it is described as the ability to convert "data" to "knowledge."[53]

Knowledge Discovery is a cross-disciplinary field that draws from "statistics, databases, pattern recognition and learning, data visualization, uncertainty modelling, data warehousing and [On Line Analytical Processing], optimization, and high performance computing."[54]

[48] "2004 Workplace E-mail and Instant Messaging Survey Summary," American Management Association, pp.2 & 4 (2004) (http://www.amanet.org/research/pdfs/IM_2004_Summary.pdf)

[49] Association for Computing Machinery home page (http://www.acm.org/).

[50] "From Data Mining to Knowledge Discovery in Databases," Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, *AI Magazine*, Vol. 17, No. 3 (Fall 1996) (http://www.aaai.org/Library/Magazine/Vol17/17-03/vol17-03.html) and "Systematic Knowledge Management and Knowledge Discovery" by Igor Jurisica, published in the *Bulletin for the American Society for Information Science*, Vol. 27, No. 1 (October/November 2000) (http://www.asis.org/Bulletin/Oct-00/jurisica.html).

[51] Charter of ACM Special Interest Group on Knowledge Discovery and Data Mining (http://www.acm.org/sigs/sigkdd/charter.php).

[52] Abstract of First ADBIS (Advances in Databases and Information Systems) Workshop on Data Mining & Knowledge Discovery (held in conjunction with 9th East-European Conference on ADBIS) at Tallinn, Estonia (Sept. 15-16, 2005), by Prof. Roman Slowinski, Institute of Computing Science, Poznan University of Technology (http://www.cs.put.poznan.pl/admkd05/).

[53] "A Survey of Data Mining and Knowledge Discovery Software Tools," Michal Goebel, University of Auckland, Department of Computer Science and Le Gruenwald, University of Oklahoma, School of Computer Science, *ACM SIGKDD Explorations Newsletter*, Vol. 1, No. 1 (June 1999) (http://portal.acm.org/citation.cfm?id=846172&coll=portal&dl=ACM&CFID=61582900&CFTOKEN=98899665)/

[54] Description of *Data Mining and Knowledge Discovery Journal*, Springer Science+Business Media website (includes definition of data mining and knowledge discovery) (http://www.springer.com/sgw/cda/frontpage/0,11855,4-0-70-35596293-0,00.html?referer=www.wkap.nl).

B. Timing of Knowledge Discovery Development

[Talk about how KDD timing relates to timing of email growth]

ACM began hosting Knowledge Discovery in Data (KDD) workshops in 1989 and conferences in 1995.[55] By 1995, interest in the topic had spread throughout the world, into governmental, commercial, and academic communities.[56] The *Data Mining and Knowledge Discovery Journal* was first published in 1997.[57] Topics of great interest to the membership are addressed by Special Interest Groups (SIGs) which often host conferences, publish journals, maintain libraries, and provide other venues for the exchange of expert knowledge.[58] Although there are only thirty-four SIGs to address the entire field of computing,[59] one devoted solely to KDD[60] was founded in 1998.[61] At that time, the SIG Charter described the understanding of Knowledge Discovery as approximately fifteen years behind the understanding of databases.[62]

[Edit to include non-ACM history/cites]

C. Value of Knowledge Discovery Industry

The "business analytics" market was been estimated at $13billion (US) in 2003.[63] A recent study revealed that companies reported a median Return on Investment of 112%., while a significant number saw a return of 1,000% or more.[64] The mean payback period was a swift 1.6 years, with the average project costing $4.5million.[65] "Business intelligence," which is largely knowledge discovery/data mining, is estimated to reach

[55] Chronology of ACM SIGKDD Conferences (http://www.acm.org/sigs/sigkdd/conferences.php).
[56] See, e.g., Program Committee List, The First International Conference on Knowledge Discovery and Data Mining, KDD-95, at Montreal, Canada (Aug. 20-21, 1995) (listing 30 members from 12 universities, 7 corporations, 4 government research centers, and representing 8 countries) (http://www-aig.jpl.nasa.gov/public/kdd95/).
[57] Charter of ACM SIGKDD (identifying the inception of the *Journal* as one of the supporting factors for creating an ACM SIG) (http://www.acm.org/sigs/sigkdd/charter.php).
[58] See, e.g., ACM SIGs home page (http://www.acm.org/sigs/) and ACM SIGs Guide (http://www.acm.org/sigs/guide98.html).
[59] ACM SIGs home page (http://www.acm.org/sigs).
[60] ACM SIGs Guide (http://www.acm.org/sigs/guide98.html); ACM SIGKDD home page (http://www.acm.org/sigs/sigkdd/).
[61] ACM SIGKDD Charter (noting that the "first year budget" and bylaws were approved in 1998) (http://www.acm.org/sigs/sigkdd/charter.php).
[62] *Id.*
[63] "Eye on Information," Alan Joch, *Oracle Technology Network* website (http://www.oracle.com/technology/oramag/oracle/05-jan/o15eye.html) **[look for non-commercial source – or non-vendor source]**
[64] "There's Gold in Them Thar Databases," David Braue, *Business & Technology Magazine*, (Aug. 7, 2003) (http://www.zdnet.com.au/insight/0,39023731,20275647,00.htm).
[65] *Id.*

$3.3billion in 2006.[66] The Data Mining market is expected to continue to grow at 10% to 20% per year.[67]

D. Knowledge Discovery Tools

Knowledge Discovery generally refers to three steps: pre-processing, processing, and presentation. Pre-processing is the work necessary to make data useable. Processing is the automated finding of patterns in data. Reporting is the means of making the discoveries understandable. Some people use the term "Knowledge Discovery" only to refer to the middle step – the act of finding patterns in data.

1. Pre-processing

More than forty years ago, the phrase "garbage in garbage out" came into common usage[68] to describe the historical fact that a computer could not tell if it was being given bad information. While the field of Artificial Intelligence has not progressed sufficiently to make the phrase obsolete, its impact is being eroded by the development of an array of pre-processing tools. Nonetheless, a 2003 poll reported that at least 40% of data mining project time was spent on pre-processing by 89% of respondents and nearly two-thirds of respondents indicated that they spent more than 60% of their time on pre-processing.[69]

When first acquired, data may have internal integrity issues. For example, if bits are lost in transmission or data is saved in the wrong format,[70] it may not be possible to manipulate the data with the very software that created it. Even the most novice user has had the experience of receiving a word processing or spreadsheet file that wouldn't open at all or opened but was unreadable. Also, I have seen instances in which data entry personnel typed the wrong information into the wrong fields, guaranteeing that databases searches by field would not yield the best possible results. It has been estimated that field error rates

[66] *Id.* **[look for underlying study – this number seems low]**

[67] "Data Mining Tools: METASpectrum^{SM} Evaluation," METASpectrum^{SM} Market Suvey (2004) (http://www.oracle.com/technology/products/bi/odm/pdf/odm_metaspectrum_1004.pdf).

[68] "Garbage In Garbage Out," Michael Quinion, *World Wide Words* (Oct. 29, 2005) (renowned etymologist and advisor to the Oxford English Dictionary cites a syndicated newspaper article about IRS computerization from April 1, 1963 as predating the OED first reference of 1964, but notes that the 1963 article indicated that the term was already long-standing) http://www.worldwidewords.org/qa/qa-gar1.htm; http://www.penguin.co.uk/nf/Author/AuthorPage/0,,0_1000065494,00.html.

[69] "Data Preparation Part in Data Mining Projects," KDnuggets: Polls, (Sept. 30 – Oct. 12, 2003) (slight rounding skew; reported total is 101%) http://www.kdnuggets.com/polls/2003/data_preparation.htm (cited in "Exploiting Relationships for domain-independent data cleaning," Dmitri V. Kalashnikov & Sharad Mehrotra, University of California Irvine, Computer Science Department, *TR-RESCUE-04-20* (Sept. 22, 2004) (http://www.ics.uci.edu/~dvk/RelDC/TR/TR-RESCUE-04-20.pdf)).

[70] "Data Cleansing: Beyond Integrity Analysis," Jonathan I. Maletic and Andrian Marcus, Software Division of Computer Science, Department of Mathematical Sciences, University of Memphis, Proceedings of the Conference on Information Quality at MIT, pp. 200-209 (Oct. 20-22, 2000) (http://www.sdml.info/papers/IQ2000.pdf).

are at least 5%.[71] These are the sorts of problems that addressed by data "cleansing." The following items are sometimes included within the broad umbrella of "cleansing."

Data collected or created in one data platform is not inherently readable by other software. At one time, tremendous programmer effort was required to move any data to any other system. Today, more vendors are offering the ability to automatically load data from other major platforms or to load data from lesser systems if certain information about the data structure (usually the "data dictionary") can be provided. However, there are still tremendous numbers of legacy systems for which no fast path exists.

- Fuzzy Matching: Data within and between systems is often not represented in the same way. Simple things such as dates and addresses can appear in a variety of forms. Typographical errors are common and names in foreign alphabets are often transliterated differently from day to day. One approach to this problem is to translate all data into the same representation (e.g., changing "January 31, 2001"; "31 Jan. 2001"; and "1/31/01" to 01312001) before any processing is done. Using this method, processing simply matches like data. However, a second approach also is now being used. That approach skips harmonization in the pre-processing stage; it leaves data in its existing form and seeks to accomplish matching through "fuzzy" logic which allows for some variation in representation.

- Disambiguation: In large data collections, there are often different items with the same name. The most common issue is two data entries with the same or nearly the same name. The challenge is to figure out whether this refers to one person or two people.[72] Everyone has had the experience of receiving two of the same catalog in the mail and discovering some slight difference in their name on the label (e.g., one with and one without a middle initial). With common names in large data collections, however, it is also likely to have two or more people with same name. Generally, disambiguating tools attempt to find other data (e.g., address, birthdate, height) associated with each record that will answer the question conclusively.

- Deduplicating: It is also common to find duplicate copies of records in data. Usually, removing duplicates is part of the pre-processing activity. However, it is important to understand the goal of the project before taking this step.[73] For example, as described more fully in my discussion of the Enron email processing, de-duplicating can result in under-counting the size or impact of stored information.

[71] *Id.*, at "Introduction" (with citations to "Orr, K., 'Data Quality and Systems Theory,' *CACM*, vol. 41, no. 2, February 1998, pp. 66-71" and "Redman, T., 'The Impact of Poor Data Quality on the Typical Enterprise,' *CACM*, vol. 41, no. 2, February 1998, pp. 79-82").

[72] *See,* "Deduplication and Group Detection Using Links," Indrajit Bhattacharya & Lise Getoor, University of Maryland, Department of Computer Science KDD Workshop on Link Analysis and Group Detection, Seattle, WA (Aug. 2004) (http://www.cs.umd.edu/~getoor/Publications/linkKDD04.pdf).

[73] *Cf.*, "EDD: Demystifying Deduplication," Brett Burney, *Law Technology News* (April 2005) (explaining impact of deduplication and reduplication on electronic discovery disputes in litigation) (http://www.law.com/jsp/ltn/pubArticleLTN.jsp?id=1113901507580).

2. Processing

The processing stage is the one that performs analysis on the data. Developing methods for conducting the analysis is a burgeoning field. A business manager is likely to have at least a visceral understanding of many of the techniques - probabilistic, case based reasoning, statistical, classification (including decision tree and pattern discovery); deviation; and trend.[74] Others though – Bayesian, neural networks, and genetic algorithms[75] – call up visions of programmer/sorcerers toiling over frothy pots of numbers indecipherable to mere mortals. For the business person, the important thing to know is that these methods focus on trying to determine which items are related or form a pattern.

- Probabilistic analysis determines a probability for each [piece?/cluster?] of data and is used in applications such as diagnosis and planning. For example, probabilistic analysis can be used to determine the likelihood that an airplane alarm system will be effective under particular weather or hazard conditions.[76]

- Statistical analysis, or rule induction, automatically creates rules from patterns. This is one method for attempting to beat the stock market – trying to have a computer automatically determine rules that better-than-market performing stocks have in common.[77]

- Classification sorts data according to similarities. Decision trees are one common method of classification. A decision tree subdivides data into progressively smaller categories, such as the way a lender makes a credit decision (e.g., is the loan applicant employed? ever had credit before? ever paid late?).[78] And, although

[74] "Knowledge Discovery in Databases: Tools and Techniques," Peggy Wright, *Crossroads: The Student Journal of the Association of Computing Machinery*, Networks & Distributed Systems, 5.2 (Winter 1998) (http://www.acm.org/crossroads/xrds5-2/kdd.html) and "A Survey of Data Mining and Knowledge Discovery Software Tools," Michal Goebel, University of Auckland, Department of Computer Science and Le Gruenwald, University of Oklahoma, School of Computer Science, *ACM SIGKDD Explorations Newsletter*, Vol. 1, No. 1 (June 1999) http://portal.acm.org/citation.cfm?id=846172&coll=portal&dl=ACM&CFID=61582900&CFTOKEN=98899665)/

[75] "Knowledge Discovery in Databases: Tools and Techniques," Peggy Wright, *Crossroads: The Student Journal of the Association of Computing Machinery*, Networks & Distributed Systems, 5.2 (Winter 1998) (http://www.acm.org/crossroads/xrds5-2/kdd.html) and "A Survey of Data Mining and Knowledge Discovery Software Tools," Michal Goebel, University of Auckland, Department of Computer Science and Le Gruenwald, University of Oklahoma, School of Computer Science, *ACM SIGKDD Explorations Newsletter*, Vol. 1, No. 1 (June 1999) http://portal.acm.org/citation.cfm?id=846172&coll=portal&dl=ACM&CFID=61582900&CFTOKEN=98899665)/

[76] "Probabilistic Analysis of Hazard Situations," J.K. Kuchar & R.J. Hansman, Massachusetts Institute of Technology, Aeronautical Systems Laboratory (Aug. 1996) (http://web.mit.edu/aeroastro/www/labs/ASL/probability/prob_hazard.html).

[77] "Stock Selection Using Rule Induction," George H. John, Peter Miller, & Randy Kerber, *IEEE Intelligent Systems*, Vol. 11, No. 5 (Oct. 1996) (abstract at http://doi.ieeecomputersociety.org/10.1109/64.539017).

[78] "Rule Induction: Decision Trees and Rules," Holly Korab, *Access Online* (publication of the National Center for Supercomputing Applications at University of Illinois, Urbana-Champaign) (Aug. 1997) (http://access.ncsa.uiuc.edu/Stories/97Stories/KUFRIN.html).

discussed in the previous section as a pre-processing technique, some refer to data cleansing as a pattern discovery technique because patterns may be readily evident in a smaller dataset.[79]

- Deviation analysis looks for outliers – data which falls outside normal patterns – and then attempts to discover the cause for the variation.[80] A classic example is credit card fraud detection.[81] A system might compute that a particular customer does 95% of her purchasing in Los Angeles; the other 5% is spent on online purchases. Multiple purchases arrive from Romania. The system identifies a deviation. A more sophisticated system might also look at how often a customer makes purchases, the value of an average purchase, and the historical maximum; in this case the system would note deviations because the prices were outside of normal range and were being made at a much faster pace than normal. To find the cause of this Romanian variation, the system might check for previously charged airplane tickets or hotel deposits in Romania.

- Bayes theorem determines probability where a fact is known. For example, a classical "card counter" at a Black Jack table is engaging in Bayesian analysis. In the first round after the cards are shuffled, the "counter" combines the knowledge of how many decks of cards are in play (total number of cards) and all of the cards that are face up on the table to determine the probability of being dealt a card he wants. As the game continues, the player keeps track of all cards he has seen in all hands played since the shuffle and adjusts the probability accordingly.

- Neural networks are intended to replicate brain function. They "learn" by being provided a large number of input patterns and resulting output patterns.[82] One example of a practical application of this technology is the processing of mortgage applications. As early as 1996, there was a reported case in which a system was trained to reach mortgage loan decisions and was able to do so with results that matched humans 84%-97% of the time.[83]

One of the major benefits of these techniques is the pace at which they can perform. In the case of the mortgage application processing technique in the last paragraph, even in 1996,

[79] "Knowledge Discovery in Databases: Tools and Techniques," Peggy Wright, *Crossroads: The Student Journal of the Association of Computing Machinery*, Networks & Distributed Systems, 5.2 (Winter 1998) (http://www.acm.org/crossroads/xrds5-2/kdd.html)

[80] "Chapter 1: Introduction to Data Mining," Osmar A. Zaiane, University of Alberta, Department of Computing Science, *Principles of Knowledge Discovery in Databases* (Fall 1999) (http://www.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/).

[81] *See, e.g.,* "Microsoft Technical Roadshow 2005: Business Intelligence in SQL Server 2005: Technical Overview," Peter Blackburn, *Microsoft TechNet*, slide 21 (2005) (http://download.microsoft.com/documents/uk/resources/techroadshow/it-professional-track/10_Business_Intelligence_in_SQL_Server_2005_Technical_Overview.ppt).

[82] "Neural Networks," Christos Stergiou and Dimitrios Siganos, Imperial College London, Faculty of Engineering, Department of Computing, *Surveys and Presentations in Information Systems Engineering (SURPRISE)*, vol. 4, 1.1 (1996) (http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html).

[83] *Id.*, at 6.3.2.

an application could be handled in 1 second, using 250K of processor memory.[84] At that efficiency, any business quality personal computer could likely handle more than a thousand at once.[85] For the business person, this means that some of these analytic processes can keep up with the pace at which new data arrives.

3. Presentation

Knowledge discovery results are most often provided in a format known as "visualization", referring to a methodology of providing images to represent the results of complex data.[86] Again, the goal is to make a large amount of data understandable quickly. We've all seen a graph showing a single trend line of stock performance over time. Consider a graph of S&P500 performance for 5 years. In reality, that one small graph is presenting the knowledge of about 126,252 data points – ((52 weeks * 5 days a week) minus 8 holidays per year) [87] times (500 stocks + 1 calculated average each day) – but it is easy to absorb the essence of that information. The difference between such a graph and a great knowledge discovery visualization tool, is that the great tool will allow you to zoom in and see the details underlying the simple image.[88]

IV. Knowledge Discovery Attacks Email Problems

Until recently, most Knowledge Discovery work involving emails was focused on spam filtering and historical analysis. [**HAVEN'T STARTED HERE YET. spam filter section + historical discussion from NSF Expedition Workshop (Cites to Oard, Baron, Underwood, etc..]**

Spam –

One research study released in 2003 estimated that employers lost $874 per employee per year due to lost productivity dealing with spam[89]. That was based upon an employee receiving 13.3 spam messages per day. VeriSign, Inc. – an infrastructure services company best known to the general public for its online payment processing products –

[84] *Id*.

[85] This is a rough assumption based upon 1,000 calculations using 250K absorbing 250MB of a 1GB RAM and assuming the remaining 75% of RAM is used to support the multi-processing and the underlying operating system.

[86] "Crossing the Information Visualization Chasm," Ben Schneiderman, University of Maryland, Human-Computer Interaction Laboratory, Public Presentation, slide 11 (Oct. 1999) (http://www.cs.umd.edu/hcil/pubs/presentations/info-viz-chasmslides/sld001.htm).

[87] The New York stock exchange is open Monday to Friday all year, except for eight specific holidays. *See,* "Holidays and Hours" webpage of the NYSE (http://www.nyse.com/Frameset.html?displayPage=/about/1022963613686.html)).

[88] *See, supra,* n. 40, at slide 13. **[Schneiderman]**

[89] "Report: Spam Costs $874 Per Employee Per Year," Paul Roberts, *InfoWorld*, Special Reports (July 1, 2003) (http://www.infoworld.com/article/03/07/01/HNspamcost_1.html).

estimates that 50-60% of all incoming email, before filtering, is spam.[90] And, Postini, a message management company, claims that 70-80% of incoming mail is spam, with small businesses receiving the high average figure of 50 per user per day.[91] About half of the respondents in a 2004 survey of corporations claimed that spam was less than 10%,[92] but this may be due to corporate filtering before email is delivered to individual users.

Talk about how Rules of Evidence are being changed to address discovery of email.

Talk about companies starting to offer compliance s/w tools for Sarbanes-Oxley

V. Boon for Researchers: Enron Emails Made Available

A significant challenge for knowledge discovery researchers has been the lack of availability of real datasets for study.[93] A major research opportunity unfolded when the Federal Energy Regulatory Commission (FERC) released Enron's email repository in March 2003.[94] **[Very brief description of Enron and its history]**

A. The Number of Available Emails

The exact number of emails is somewhat unclear. *The Wall Street Journal* reported that FERC had released 1.6 million emails and other documents, generally from the period 2000 to 2002.[95] The emails quickly became notorious for the variety of non-business content (including spam, jokes, and pornography) as well as the evidence of inappropriate

[90] *See*, VeriSign ROI Calculator (http://www.verisign.com/products-services/security-services/messaging-security-and-compliance/email-security/ROI_Calculator/).

[91] *See*, summary of Postini's Annual Message Management and Threat Report (posted Jan. 30, 2006) (http://www.postini.com/news_events/pr/pr013006_tr.php).

[92] "2004 Workplace E-mail and Instant Messaging Survey Summary," American Management Association, p.6 (2004) (http://www.amanet.org/research/pdfs/IM_2004_Summary.pdf

[93] "The Enron Email Dataset Database Schema and Brief Statistical Report," Jitesh Shetty, University of Southern California, and Jafar Adibi, USC Information Sciences Institute (http://www.isi.edu/~adibi/Enron/Enron_Dataset_Report.pdf).

[94] "E-sleuthing and the Art of Electronic Data Retrieval. Uncovering Hidden Assets in the Digital Age: Part 1," Jack Seward and Daniel A. Austin, McGuire Woods LLP, *American Bankruptcy Institute Journal*, Vol. 23: 1, fn. 7 (Feb. 2004) (http://www.e-evidence.info/seward1.pdf).

[95] "Online Laundry: Government Posts Enron's Emails," Dennis K. Berman, *The Wall Street Journal* (October 6, 2003) (copy available at: http://flatrock.org.nz/topics/info_and_tech/it_is_for_your_own_good.htm).

business conduct.[96] Employees complained about the invasion of their privacy and, although Enron had missed prior deadlines for requesting removal of specific emails, FERC ultimately agreed to remove and review 141,379 emails identified by Enron.[97] Those emails were described as ones which appeared to create a high risk of identity theft – those containing social security numbers, credit card numbers, birthdates, etc. – or extremely personal matters involving divorce or children.[98] This resulted in a reduction of the database by approximately 8%.[99] By September 2003, FERC had reviewed over 17,000 of the questioned emails and decided that less than a third were entitled to removal; FERC ordered approximately 12,000 re-released.[100] Viewing the official site, it appears that there are approximately 1.4 million emails.[101] A closer examination of the data quickly reveals that some have no message[102] and others are duplicates.[103]

MIT acquired a copy of the data and discovered these and other integrity problems;[104] SRI, International attempted to cleanse the data as a part of its CALO (Cognitive Assistant that Learns and Analyzes) Project.[105] That version of the data, which is available for research, contains 517,431 emails from 151 users.[106] The CALO version, however, has removed all attachments from the emails; attachments remain available in the FERC data. Multiple researchers determined that this dataset still contained duplicates and error messages. USC researchers further cleansed the data and reduced the total to 252, 759 emails (48.84%).[107]

[96] See, e.g., "The Decline and Fall of the Enron Empire," Tim Grieve, *Salon* (Oct. 14, 2003) (http://www.salon.com/news/feature/2003/10/14/enron/).

[97] Third Order On Re-Release Of Data Removed From Public Accessibility On April 7, 2003, Fact Finding Investigation of Potential Manipulation of Electric and Natural Gas Prices, 106 FERC ¶ 61,239, Docket No. PA02-2-000 (Issued March 8, 2004) (www.caiso.com/docs/2004/03/09/200403091616391042.doc).

[98] *Id.*

[99] *Id.* and "Addressing the Western Energy Crisis: Information Released in Enron Investigation," Federal Energy Regulatory Commission Website (http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp (page updated April 28, 2005)) ("Contents" description of "Enron email" as "92% of Enron's staff emails).

[100] Third Order On Re-Release Of Data Removed From Public Accessibility On April 7, 2003, Fact Finding Investigation of Potential Manipulation of Electric and Natural Gas Prices, 106 FERC ¶ 61,239, Docket No. PA02-2-000 (Issued March 8, 2004) (www.caiso.com/docs/2004/03/09/200403091616391042.doc).

[101] FERC's official site (http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp) directs one to the Aspen Corporation's iConnect 24/7 site (http://fercic.aspensys.com/members/manager.asp), which provides four versions of the Enron email. Selecting the .pst file which is not a re-release, and choosing document database view and the notification that this "You are viewing Document 1(1) of 1,368,775." (http://fercic.aspensys.com/iconect247/iconect247.exe).

[102] See, e.g., S_DOC Nos. 21, 22, 25, 27 by continuing from the steps in fn.**[immed preceding]** *supro.,* and sequentially reviewing documents.

[103] See, e.g., S_DOC Nos. 49010 and 50078 (same email from Kimberly Kirkwood to Mark Guzman, Subject "Fwd: Fw: THIS IS SCARY!!! DO IT!!" dated 12/12/2000, 18:24:00 GMT).

[104] "Enron Email Dataset," by William W. Cohen, Carnegie Mellon University, Center for Automated Learning & Discovery (Webpage last modified: April 4, 2005, 10:55:50 EDT) (http://www.cs.cmu.edu/~enron/).

[105] *Id.*

[106] *Id.* and "The Enron Email Dataset Database Schema and Brief Statistical Report," Jitesh Shetty, University of Southern California, and Jafar Adibi, USC Information Sciences Institute (http://www.isi.edu/~adibi/Enron/Enron_Dataset_Report.pdf).

[107] "The Enron Email Dataset Database Schema and Brief Statistical Report," Jitesh Shetty, University of Southern California, and Jafar Adibi, USC Information Sciences Institute (indicating a dataset of "252,759

Carnegie Mellon researchers created a dataset of 619,446 from 158 users that they reduced to 200,399 (a much smaller 32.35%) from 158 users.[108] Cleansing techniques affect results, as two other research groups identified 149[109] and 161[110] users (without 100% overlap).[111]

Work done at the University of Southern California by Jitesh Shetty and Jafar Adibi provided significant understanding of the basic statistics for the data. Consistent with anecdotal evidence and expectations, they determined that most users had saved a small number of emails and a small number had saved a large number – the majority of the employees had 1,000 to 5,000 emails while a small number had 5,000 to 10,000 emails.[112] Also, most users received far more emails than they sent;[113] most employees had sent 500 or less emails, with a significant number sending up to 1,000, but only 8 users had sent more than 2,000.[114]

Perhaps most important for the purpose of this thesis is that the emails were not distributed equally over time. There are no emails from 1998, progressively more through 1999, 2000, and 2001, and then less again in 2002.[115] [**Research whether this is the result of the company's retention policies or the limit of the release ordered by FERC.**]

B. Understanding the Cleansed Set

I wanted to understand how the cleansed datasets differed from the original, so I structured a small test. First, I searched for the word "blonde" in the FERC/Aspen dataset and was returned 309 emails; in the Berkeley set the result was 112.[116]

I manually reviewed them and categorized one hundred of them in an Excel spreadsheet. To understand the cleansing process, the elements tracked in the spreadsheet were:

- FA Sdoc_No
- UCB DatabaseID
- Date

messages from 151 employees distributed in around 3000 user defined folders") (http://www.isi.edu/~adibi/Enron/Enron_Dataset_Report.pdf).

[108] "Introducing the Enron Corpus," Bryan Klimt & Yiming Yang, Carnegie Mellon University, Language Technology Institute, p. 1 (2004) (presented at First Conference on Email and Anti-Spam (CEAS), Mountain View, CA)) (http://www.ceas.cc/papers-2004/index.html & http://www.ceas.cc/papers-2004/168.pdf).

[109] **Andres Corrada-Emmanuel cite here.**

[110] **Insert Shetty/Adibi cite (to the excel spreadsheet listing their 161)** http://www.isi.edu/~adibi/Enron/Enron.htm ("Ex-Employee Status Report")

[111] **Identify (by id number?) how many were different**

[112] Shetty & Adibi, p. 4 & Figure 2.

[113] *Id.*

[114] *Id.*, at p. 5 & Figure 3.

[115] *Id.*, at p. 7, Figures 5 & 6.

[116] **NOTE: The Queen's University shows only 88 occurrences of "blonde."**

- Topic – short description of content
- From an Enron email account?
- To
    - How many Enron email accounts?
    - How many non-Enron email accounts?
- Folder Location

1. Unique record identifiers

I wanted to know if the datasets used the same unique identifiers for the emails, which would make comparison simplest. The second email returned by the FERC/Aspen ("F/A") tool was a January 14, 2002 email containing a joke with the subject header "FW: Cosmetic Surgery." I searched for the same subject header in the Berkeley ("UCB") data using the online search tool and found the same email. They appeared not to have any identifying number in common: F/A showed an "SDOC_NO" 31046 while UCB showed a "DatabaseID" 18295.

2. Changes to Email Addresses

My review quickly uncovered that something in the UCB set had been altered. The UCB version of this particular email showed all six recipients as having email addresses at enron.com. The original F/A document showed only one recipient having an email address at enron.com; the other five were at swbell.net; burypartners.com; kochind.com; hotmail.com; and tmh.tmc.edu. For my purposes, this is a significant change. For the purpose of compliance analysis, it will be important to know if employees are exchanging inappropriate material with people outside the company. It likely also will be important to understand the traffic flows between official corporate email accounts and personal email accounts.

3. Conversion of Time Stamps

A curious difference between the F/A and UCB datasets is the conversion of the timestamps. The F/A dataset mostly provides time as Greenwich Mean Time (GMT). The UCB dataset converted all timestamps to Pacific Time (PDT or PST). For example, the F/A dataset has a 10/04/01 email from an Enron employee with the subject: "7 Degrees of Blonde" (SDOC_No 793655). A search of the UCB data revealed two emails (DatabaseID 207169 and 207170) with the same date and subject from the same employee. Neither of the two UCB emails matched the timestamp of the F/A email, 15:29:00 GMT. By reviewing the contents it was possible to determine that the matching UCB email (DatabaseID 207169) is the one with a timestamp of 08:29 PDT. This timestamp conversion occasionally results in a different date (see, e.g., SDOC_No 806665 stamped 7/31/01 02:01:40 GMT and matching DatabaseId 7/30/01 19:01 PDT). It appears that the majority of the emails were sent or received in Texas at the Enron headquarters city. From the perspective of the compliance analyst, the local time for the email would be most useful, as personal emails may be read differently in the context of daytime and nighttime.

4. Duplicates in the Original Dataset

Not surprisingly, the F/A database had its own errors. For example, there are four identical copies of an email from a non-employee to an employee about a naked blonde woman at a party and her near sexual encounter with a mutual acquaintance. All four have the same date and time stamp; although one copy (SDOC_No 160741) is from the employee's "all documents" folder, three of the copies (SDOC_No 162270, 166329, and 171762) are from the employee's "inbox" folder. Interestingly, there are other similar duplications involving the same user. SDOC_No 160755, 166343, and 173325 are the responsive email expressing regret for missing the party, but explaining that he had "[h]ooked up with a chick" on vacation in "Cabo." In another set, there are multiple "sent" folder copies of an email (SDOC_No 155940, 163584, and 173175) from the employee about car trouble and his possible interest in being fixed up a "tall blonde." It is unknown whether these errors existed in the Enron database or were the result of the FERC/Aspen recovery process.

5. De-duplication and the Loss of Location Data

The F/A set include the details of where the email was found but the UCB search result does not include that data. For example, F/A data reveals if an email was found in the sender's "Sent" folder or the recipient's "Inbox" folder. This is an excellent example of the importance of understanding the user's goals. UCB intentionally removed duplicate copies. Typically, upon sending an email, the sender will have a copy in his "Sent" folder and his "All Documents" folder and the recipient will have a copy in her "Inbox" folder. If all three copies were retained in the database, UCB's social network analysis tool likely would have incorrectly counted them as three distinct communications. So, for UCB's purpose, deleting duplicates provides a more accurate result. Eliminating duplicates effectively means eliminating at least two of the locations. While the location folder wasn't important for the particular type of social analysis that UCB was performing, it might be informative for a compliance analysis. If an X-rated joke is sent, did the recipient put it in the "Deleted" folder? Save it to a personally-created folder called "Fun Emails"? Or, perhaps to one called "Harassment" or "Evidence"?

6. Summary Statistics

[insert text]

- 46 of the 100 (46%) are unique emails (correlating closely with Berkeley's overall rate of 48.8%)
  - 12 of the 46 emails identify the sender or recipient email addresses differently
  - Resulting in the following
    - Relative agreement on number of emails sent by Enron employees (26 or 27)
    - Drastically different statistics on number of emails received by Enron employees

        - F/A indicates that recipients were 44 employees and 81 non-employees
        - UCB indicates that recipients were 87 employees and 36 non-employees
    - 44 of the 46 are personal emails
- 8 of 100 (8%) were blank (correlating exactly with the 8% removal by FERC in response to privacy requests)
- 45 of the 100 (45%) were additional copies of the unique emails
- 1 of the 100 (1%) was a unique email that does not appear in the UCB dataset

C. ENRON Knowledge Discovery Work to Date

A number of Knowledge Discovery research activities have already centered on the Enron emails.

1. Occurrence Counts

Word counts are often performed as a pre-processing activity, a precursor to a more sophisticated analysis. In this pre-processing activity, software identifies every unique word (or character string) and counts the number of occurrences of that word. Traditionally, these counts will drop out pronouns (he, she, me, I), prepositions (under, over, on, etc.) and other words that are not likely to provide clues to meaning. Queen's University in Canada performed this task on the Enron emails, sorted both by descending order of occurrence and alphabetically.[117]

This group presented a paper in October 2005, explaining how they used the word counts in the application of "deception theory," which asserts that certain word choices are more common in deceptive writing.[118] Specifically, they looked for less than normal usage of "first person pronouns (I, me, my, etc.)" and "exclusive words (but, except, without, etc.)" and higher than normal usage of "negative emotion words (hate, anger, greed, etc.)" and "action verbs (go, carry, run, etc.)."[119] For each email in their cleansed set, they counted words that fell into these four categories and then plotted the results using a Singular Value Decomposition (SVD) matrix. **[produce lay explanation of SVD]** .

The resulting plot is roughly a downward pointing triangle shape with elongated points.[120] The upper left point represents high usage of exclusive words and is described as

[117] "Other Forms of the Enron Data," Web-page posted by Professor David Skillcorn, Queen's University (Canada),School of Computing, data prepared by his former graduate student Nikhil Vats (http://www.cs.queensu.ca/home/skill/otherforms.html).

[118] "Detecting Unusual and Deceptive Communication in Email," P.S. Keila and D.B. Skillcorn, Queen's University, School of Computing, presented at CASCON 2005 (Oct. 20, 2005) (http://www.cs.queensu.ca/TechReports/Reports/2005-498.pdf).

[119] *Id.*, at p. 4.

[120] *Id.*, at p. 6, Figure 2.

"emotionally charged" emails to co-workers, family, and friends.[121] The upper right point represents high usage of personal pronouns and correlates strongly with non-business recreational activity. The bottom point contains high usage of action verbs.[122] Since the authors are searching for deception, they focus on the lower portion of the triangle.

Since the authors are searching for deception, they focused on the confluence of the four factors. Based upon learning during the research activity work (e.g., that use of personal pronouns is lower than normal throughout the dataset), they make some adjustments to the values and produce another matrix. In this one, they successfully create two clusters of deceptive emails; the clusters are differentiated based upon whether they do or do not contain negative emotional words as well.[123]

The research team notes the value of this success. A management or compliance officer could select emails of interest without engaging in the labor intensive task of reading them all. The identities of employees need not be revealed unless or until email of interest is identified. Also, the authors show that the emails of any individual employee could be evaluated using this technique and the one at the farthest point be read.

I believe this research provides additional valuable information for the human resources or operations manager. The person searching for personal use of corporate email might choose to focus on the upper right, which reflects high usage of mail to discuss personal recreation. And, further analysis of the "emotionally charged" emails might reveal discussions of other employees' misconduct.

And, while the mails seemed relatively evenly distributed, this perception was dispelled when the researchers color-coded the data points to reflect the authors of the email.[124] Based upon the color-coding it also appears that Enron senior executives most often in the personal pronoun and action verb points.[125] While 20/20 hindsight would make it easy to make a quick assessment that these senior managers were more heavily engaged in their own recreation (as the oft-cited emails about the wedding planning of Ken Lay's daughter[126] would suggest) or deception (as the current indictments[127] suggest), another explanation is possible. It is certainly possible that people who are in senior executive positions refer to themselves and to action verbs more frequently because they are the ultimate decision-makers. Further study should be done in this area.

The most interesting observation from the color-coded plot is that the Enron employees in the dataset generally were writing emails at the edges of the triangle (meaning, the employee emails had large numbers of words in one or more of the three categories) and

[121] *Id.*, at p. 4

[122] *Id.*, at p. 5.

[123] *Id.*, at p. 8 and p.9, Figure 5.

[124] *Id.*, at p. 7, Figure 3.

[125] While this would seem to imply that the senior executives spent their time writing about personal recreation or writing deceptively, further research might be useful to determine if it is the nature of senior executives to talk more frequently about themselves and to talk in active terms.

[126] **Insert cites**

[127] **Insert cite re: percentage of senior execs indicted?**

that non-employees were most heavily represented in the moderate range. The fact that employees generally were outside of the normative pool seems to provide an insight into the mood of Enron. It's important to remember that these emails belong to the managers of Enron. Based upon this analysis, its management employees appear to have been much more frequently angry, deceptive, or focused on outside recreation than the people outside the company with whom they exchanged communications.

2. Automated Categorization According to Personal Preference

In the summer of 2004, a group at University of Massachusetts, Amherst reported on their study of the accuracy of multiple software applications that sought to "learn" a person's strategy for sorting emails into folders.[128] The project essentially recognizes that people have different mental models for organization and, therefore, make different choices about how to file their records. The research used the emails associated with Enron's seven heaviest email users as one of its study datasets.[129] For each person, it took only the emails s/he had sorted into topic-related files (ignoring files such as in-box, all_documents, discussion_threads, etc.) and then also removed those files with too few emails to allow meaningful study.[130] The topical emails were then removed from their folders and re-sorted into chronological order, [131] representing the stream of emails delivered over time.

A series of tests were performed in which a tool first was given a set of emails with information about where the user had filed each one and then the tool was instructed to file a similar number of emails that had arrived next in time. [132] Because topics change over time and the tools could end up with nothing in their learning sample to assist in classification, the test allowed the system to learn its mistakes before moving on to the next set. The group benchmarked four classifying tools: Maximum Entropy; Naïve Bayes; Support Vector Machine (SVM); and Winnow[133] and concluded, primarily that Naïve Bayes was the weakest for this task, with accuracy results generally 10% to 20% lower than the next most accurate application.[134] Winnow ran substantially faster than the other applications and Wide Margin Winnow was appreciably more accurate than Winnow.[135] Using the Enron datasets, it appeared that the other three methods showed promise, with accuracy scores ranging from around 50% to over 90%, and that SVM was the most

[128] "Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora," Ron Bekkerman, Andrew McCallum, & Gary Huang, University of Massachusetts, Amhearst, Computer Science Department, Center for Intelligent Information Retrieval, Technical Report No. IR 418 2004 (http://www.cs.umass.edu/~ronb/papers/email.pdf).

[129] *Id.*, at p. 7.

[130] *Id.*, at pp. 4-5 and see p. 11, Table 1 (showing that the Enron sample set was approximately 19,500).

[131] *Id.* at p. 5 ("…after sorting the messages according to their time-stamp, we train the classifier…")

[132] *Id.*, at p. 5 (rejecting a methodology of learning from the first half and testing on the second half that had been used for spam filtering and rejecting a methodology of re-training after the filing of each single email as too resource intensive for a functioning organization.)

[133] *Id.*, at p. 11.

[134] *See, Id.*, pp. 12-13 and Tables 3 & 4 (providing and discussing accuracy results per user per application).

[135] *Id.*, at pp. 14-15.

accurate. However, the same tests were run on a second non-Enron sample set with significantly lower results – more than half the tests resulted in less than 50% accuracy – and that there was little differentiation between the accuracy of the three applications.

Significant observations arising from this study were 1) if a user had a small number of dominant folders, the accuracy rate was significantly higher and 2) accuracy rates fell at times when folders were created, moved, or abandoned.[136] It is also important to note that each email was treated as a "bag of words"; the protocol simply identified and removed the 100 most common words in a person's aggregated email collection and any word that appeared only once. The researchers suggest that accuracy might be improved by applying tools that that would weight or emphasize the information in fields such as Subject, To, and Signature and tools that extract entity names.[137]

If this task could be done successfully, there would be several benefits for business value. Users could file and retrieve emails more quickly, thus increasing efficiency and creating beneficial cost reductions for their employers. Users could more often find the information they are seeking, thus increasing productivity – another bottom-line benefit to an employer.

Theoretically, the method could be extended beyond an individual's files to an organization's files, dynamically reorganizing all information into a custom structure for each person that was his/her most effective map for assimilating information. For example, theoretically, the human resources manager might have the system sort the entire organization's emails into folders called "hostile environment" (with sub-folders for race, gender, nationality, etc.), "sexual misconduct," "drinking," "drugs," "office gambling," "other personal traffic," and "ordinary business."

3. Categorization According to Single Standard

At least two groups have taken sub-sets of the Enron corpus and attempted to hand sort the messages into categories.

- In November 2004, Associate Professor Marti Hearst at the University of California, Berkeley, School of Information and Management Systems and her students in an Applied Natural Language Processing class created categories for annotating a series of emails; chose approximately 1,700 emails that were focused on business topics (intentionally avoiding jokes and "very personal" messages); and then annotating the mails with the categories.[138] The activity

[136] *Id.*, at p. 14.
[137] *Id.*, at p. 6.
[138] "UC Berkeley Enron Email Analysis," a webpage posted by the University of California, Berkeley, BAILANDO ("Better Access to Information using Language Analysis and New Displays and Organizations") project (http://bailando.sims.berkeley.edu/enron_email.html) and Syllabus of SIMS 290–2, Applied Natural Language Processing Class, Professor Marti Hearst, University of California, Berkeley, School of Information and Management Systems (Fall 2004) (Class Assignments for November 1 & 3) (http://www.sims.berkeley.edu/courses/is290-2/f04/sched.html).

was a class exercise that did not result in statistics or visualizations, but the labeled emails have been made available for review or use.

- In November 2005, a Masters student at the University of Minnesota, Duluth, Department of Computer Science, under the direction of Associate Professor Ted Pedersen[139] reported the manual annotation of approximately 1,000 emails from the University of Massachusetts, Amhearst collection.[140] Unlike the Amhearst work, thought, multiple users' emails were categorized into a single set of common categories and subcategories.[141] The emails were divided as follows: Business – 60.2%; Personal – 14.3%; Human Resources – 14.2%; General Announcements – 8.5%; Enron Online – 2.0%; and Chain Mails – 0.7%.[142] The manual annotations will be used as a benchmark against which to compare the results of automated clustering.[143]

In both cases, the activity is applying one sort standard to all of the email. If it could be accomplished in an automated fashion (perhaps through clustering technology), it might be utilized in an interesting manner to help employees improve their compliance with corporate expectations. I would be interested in a tool that could run over an employee's draft email and produce a message such as:

"Our Compliance Scan tool has reviewed the contents of your draft email. It has determined that the email will be referred to the Human Resources Department as potentially indicative of a hostile work environment. This determination was made by an automated tool which could be in error. Do you wish to send the email anyway?"

4. Social Network Analysis

Insert text

a) "enronic"

Jeffrey Heer, a graduate student at University of California, Berkeley, provided insights into Enron while working on an Information Visualization ("infovis") toolkit he calls

[139] *See,* Webpages of Associate Professor Ted Pedersen, University of Minnesota, Duluth, Department of Computer Science (identifying himself, his research, and the students he supervises including Apurva Padhye) (http://www.d.umn.edu/~tpederse/; http://www.d.umn.edu/~tpederse/research.html; and http://www.d.umn.edu/~tpederse/students.html).

[140] "Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora." Apurva Padhye, Masters Student, University of Minnesota, Duluth, Department of Computer Science, Powerpoint Slide 13 (November 4, 2005) (www.d.umn.edu/~tpederse/Group05/ap-slides-nov4.ppt).

[141] *Id.*, at Slide 15.

[142] *Id.*, at Slide 16.

[143] Research Page of Apurva Padhye, Masters Student, University of Minnesota, Duluth, Department of Computer Science (http://www.d.umn.edu/~padhy005/ and http://www.d.umn.edu/%7Epadhy005/research.html).

"prefuse."[144] The premise of the toolkit is to provide programmers a fast path to multiple visualizations of the results of whatever analysis technique they are applying. In a white paper describing his early work on Enron, [145] he discussed the risks of automated tools that don't permit the user to ensure the accuracy of the results and explained that his tool would allow the user to see the underlying data.

Heer produced a social network analysis of the people communicating via email, treating each email as a link between the sender and recipient. He took the emails labeled by Professor Hearst's class and produced a visualization. In it, people (or their email addresses) are treated as nodes and the emails between each pair are represented by a linking line whose thickness grows for each additional email. The picture is supplemented by a pie chart on each line, with each color reflecting a type of email categorized by Professor Hearst's class. For example, a thick line between two people with a pie that's half red and half a particular shade of green indicates a lot of emails between the two and that the topics were evenly split between "company business, strategy, etc." (red) and "political influence/contributions/contacts" (green). Another function uses an existing algorithm to identify "communities."

His tool offers a variety of practical interactive features that I had seen before on successful commercial products,[146] allowing the user to move the nodes, highlight segments, zoom in and out, etc. He added some extremely useful features that I could not recall having seen before. The first is a reading pane that lets the user see a list of the emails being represented (either by person or between two people), with a pie graph for each email, plus the text of any email highlighted on the list. This directly addresses his concerns about checking automated tools for accuracy; the user can see if the emails really are between the right people and about the topic identified in the tag. The second is a slider bar that removes links that represent fewer connections. Since social network analysis often produces graphics that look like massive spider webs, this feature allows the user a clean view of the weightiest links. And, there's a second slider bar that makes it possible to watch the algorithm identify communities. This permits the user to see smaller sub-communities that are merged by the algorithm. This might, for example, make it possible to see alliances within a group of people sharing a project.

Heer used social network analysis to study the emails about the California energy crisis. When he reviewed the emails, he discovered the unusual pattern of one person reporting on all Congressional meetings to a person who never responded. Further investigation revealed that the individual received legal reports from other people in the company, but

[144] "prefuse: a toolkit for interactive information visualization," Jeffrey Heer, University of California, Berkeley, Computer Science Division; Stuart K. Card, Palo Alto Research Center; and James A. Landay, University of Washington, Computer Science & Engineering, presented at the Conference on Human Factors in Computing Systems (CHI) (April 2005) and Email Archive Visualization Workshop, University of Maryland (June 2, 2005) (http://www.chi2005.org/program/prog_papers.html; http://www.cs.umd.edu/hcil/emailviz/workshop/; and http://guir.berkeley.edu/pubs/chi2005/prefuse.pdf).

[145] "exploring enron: Visualizing ANLP Results [Version 1: white]" Jeffrey Heer, University of California, Berkeley (Fall 2004) (lower case in title in the original) (http://jheer.org/enron/v1/).

[146] E.g., i2 (http://www.i2inc.com/) and Visual Analytics (http://www.visualanalytics.com/).

never responded to any of them. Without knowing the details of the Enron case, he had identified the first person indicted.

Heer intentionally focused his work on the emails tagged with business topics and avoided those with personal or social tags.

It is also interesting to note that Heer suggests another kind of social network analysis. He notes that one could infer social networks of people and organizations from the text of email.

- **[Discuss work of Corrada-Emmanuel[147]]**

VI. Thesis Experimentation

…what I could achieve through the use of automated tools. In that respect, I was primarily interested in comparing two things: level of effort and quality of result.

A. Occurrence Counts

The Queen's University count contains 160,203 words drawn from its own cleansed version of the data containing 289,695 emails.[148] Clearly, a business manager cannot regularly review a list that's more than one hundred thousand items long. However, that doesn't make the list unusable. A fast scan of the beginning of the usage list could satisfy such a manager that the majority of the references seem reasonably related to official business. For example, the list below shows the most used words and the frequency of their use:

1	Enron	371971
2	ENERGY	244838
3	Power	243465
4	Company	151112
5	information	135604

[147] http://www.cnlp.org/presentations/slides/Corrada_Enron.pdf

[148] "Detecting Unusual and Deceptive Communication in Email," P.S. Keila and D.B. Skillcorn, Queen's University, School of Computing,

6	market	121906
7	time	120978
8	California	114828
9	business	111153
10	Thanks	101483
11	state	94524
12	Price	87119
13	Houston	82886
14	trading	76493
15	electricity	75423
16	Week	72083
17	Need	70652
18	email	70642
19	Agreement	69970
20	know	68601
21	year	68500
22	group	68085
23	services	67840
24	contact	65947
25	Call	64730

1. Hostile Environment

A human resources manager (or attorney) might look at the occurrence list for words associated with potential employment law issues. For example, emails containing words and slang describing parts of a woman’s anatomy (i.e., “boobs”) are potential evidence of a hostile work environment for women. I left out words for which I could quickly identify another possible connotation, such as “breast” (with 379 occurrences) because of the likelihood of emails relating to breast cancer fundraising and health awareness programs. In about an hour, I could identify twelve such terms – not all suitable for a PG-rated thesis – that totaled 384 occurrences.

In approximately another hour, I was able to identify another 17 terms, related to the word “sex”, the concept of sex, or that likely demarcated a pornographic website (e.g., “sexxx” and “SexyWhiteThang18”), with another 172 occurrences. Thus, in about two hours, I had identified 556 occurrences that might lead to liability for the company.

There were far fewer racial or ethnic slurs that could readily be identified. In part, this is due to the fact that many words used as derogatory terms have non-derogatory meanings in other contexts (e.g., “chink” or “spic”). There were 4 occurrences of “nigger” and 2 of “raghead.” In many organizations, the human resources manager will immediately terminate the employment of the author. With such a small number, the human resources manager could immediately address the issue.

2. Personal Use of Corporate Resource

Today, companies are increasingly concerned about the personal use of corporate resources. I looked for words that might signal use of the corporate email system for personal business. First, I looked at home related activities and discovered more than 1,500 occurrences for 9 terms.

20143	babysitter	53
36329	babysitting	19
57445	babysit	8
11549	plumbing	143
21546	plumber	47
10756	mechanic	161
55281	repairman	9
3230	doctor	1108
13179	dentist	113
	TOTAL	1,661

Having seen many references to parties, I searched for drinking related terms. From 16 terms, I discovered nearly 10,500 occurrences.

1441	wine	3534
1889	beer	2452
2609	drinks	1563
2690	drink	1489
4077	drinking	782
7821	liquor	278
13580	drunk	107
15992	martini	80
17382	whiskey	68
29165	drinkin	27
48494	drunks	11
57420	drinker	8
95256	drunkest	3
104823	nondrinkers	3
104957	drunkards	3
147752	drunkenness	1
	TOTAL	10,409

Then, I searched for words related to sports. I looked only for things related to the names of sports. I did not search for the names of teams or athletes. From just 19 terms, I uncovered nearly 17,000 uses.

899	football	6208
1157	golf	4701
2589	basketball	1578
2717	baseball	1470
4172	tennis	754
5231	soccer	523
6418	softball	384
6535	hockey	376
10108	golfing	181
10896	golfers	158
1127	rugby	150
15753	golfer	82
19292	footballguy	57

29598	footballs	27
40244	footballers	16
58038	baseballs	8
74331	softballs	5
87287	footballer	4
88047	arenafootball	4
99226	basketballer	3
	TOTAL	16,689

Searching the names of NFL teams (excluding “bills” as too common a term), produced more than 15,000 more hits:

2971	giants	1259
3288	Bears	1082
3308	jets	1073
3827	Texans	859
4301	broncos	719
4559	cowboys	658
4675	chiefs	629
4734	lions	616
4819	Raiders	600
5230	saints	524
5250	Eagles	520
5355	patriots	503
5678	ravens	462
5875	dolphins	438
5953	chargers	432
6000	rams	426
6023	packers	424
6183	Titans	405
6398	seahawks	386
6507	panthers	377
6641	Redskins	366
6984	colts	337
7041	jaguars	331
7258	Bengals	315
8111	falcons	260
8186	vikings	257
8240	steelers	253
8497	Buccaneers	240
8891	cardinals	223
17728	Niners	66
	TOTAL	15,040

In about a day, I had identified nearly 44,000 word occurrences that are likely evidence of personal use of the corporate email resource.

3. Limitations

The word count methodology has clear limitations. Most notably, it doesn't tell you *who* is using these terms.

The count does not provide indications of when a word is being used for the meaning sought and when it is not. For example, there are more than 12,000 occurrences of the word "bills" but there is no way to determine when the reference is to the "Buffalo Bills" and when it is to "utility bills." The count also provides no indication of when a word with a meaning in English is used as a word in another language or as a proper noun. While reading the "blonde" emails, I had seen a reference to "Tatas" as a reference to a woman's breasts. The word count shows 55 uses of this word. A Boolean search (discussed later) reveals that this is also the name of a power plant in India.

Many problematic emails cannot be identified by a single word. For example, many of the blonde jokes, which are derogatory to a particular legally "protected class" of individual do not contain any of the words I searched.

Because the list of words is much too long for regular review, the ability to find any information is limited to the creativity of the reviewer in choosing words to research. For example, while looking for words related to drinking, I missed "hangover" (#15706), "margarita" (#11069), and "margaritas" (#15793) with 82, 154, and 67 occurrences respectively. Undoubtedly, I missed other terms as well.

B. Boolean search

1. Evidence of a Hostile Environment

I returned to the "blonde" emails I had searched before, and began to expand the spreadsheet to include information that someone focused on compliance issues might wish to know:

- Category – social/Enron, social/personal, business
- Rating – this is a rough approximation of the often-described but not released movie rating system[149]:
    - G
    - PG: one or two use of a "harsher sexually derived word" as an expletive (not in a sexual context);
    - R: more than two uses of such words; discussion of sex; visual display of total female nudity;
    - X: "an accumulation of sexually oriented language," explicit sex scenes; visual display of male genitalia (except if in a non-sexual context)

[149] *See, e.g.,* "Questions & Answers: Everything You Always Wanted to Know about the Movie Rating System," from the official website of the Classification and Ratings Administration (http://www.filmratings.com/questions.htm); "F-bombs catch a break: MPAA lets 'Palace' push profanity limits," Gabriel Snyder and Ian Mohr, *Variety* (Feb. 25, 2005) (http://www.variety.com/article/VR1117918509?categoryid=1236&cs=1); and "The Rating Process" section of the Wikipedia entry for "MPAA Film Rating System" (http://en.wikipedia.org/wiki/MPAA_film_rating_system).

- Issue: Derogatory to: Women, Men, Gay, Religion, All or Neutral
- From
  - Enron employee?
  - Gender
  - Management employee?
- To
  - Enron employee?
  - Gender
  - Management employee?
- Folder Location of the
  - Filename
  - Enron employee?
  - Gender
  - Management employee?

C. Thread search

[Tool (in development) provided by Tamer El Sayed at U Md.

Offers search by thread + ….]

D. Latent Semantic Indexing

[Data available for research as of 2/10/06]

E. [InBox?]

[Available via internet]

x[150]

[150] [Saving in case there is an appropriate spot: SDOC_No 805666 is a long labor law email newsletter from the prestigious law firm of Baker & McKenzie. The term "blonde" appeared in a brief statement about a London law firm employee asserting "sex and race discrimination after she read offensive emails sent by a

partner in the firm and another solicitor suggesting that they choose as her successor a "busty ***blonde***." See " Offensive E-Mail <http://news.bbc.co.uk/hi/english/sci/tech/newsid_1530000/1530458.stm> ."]