KNOWLEDGE DISCOVERY IN CORPORATE EMAIL: THE COMPLIANCE BOT MEETS ENRON

by

K. Krasnow Waterman

Sloan Fellow

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
June, 2006

Certified by: _____
John Van Maanen
Erwin H. Schell Professor of Organization Studies
Thesis Advisor

Certified by: _____
Pat Bentley
Senior Lecturer
Thesis Reader

Certified by: _____
Stephen Sacca
Director
MIT Sloan Fellows Program in Innovation and Global Leadership

KNOWLEDGE DISCOVERY IN CORPORATE EMAIL:

THE COMPLIANCE BOT MEETS ENRON

by

K. Krasnow Waterman

Submitted to the Alfred P. Sloan School of Management on May 12, 2006
in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Management of Technology

ABSTRACT

I propose the creation of a real-time compliance "bot" – software to momentarily pause each employee's email at the moment of sending and to electronically assess whether that email is likely to create liability or unanticipated expense for the corporation. My thesis describes the confluence of historical events making such a product necessary and desirable – increase in corporate regulation, explosive growth of email, acceptance of email as evidence in litigation. The cautionary tale of Enron provides the backdrop for the thesis. The government released hundreds of thousands of Enron management emails and they have become research fodder for those interested in "Knowledge Discovery," a computer science discipline that gleans meaningful information from data otherwise indecipherable due to its sheer size. CEO's and other C-level corporate managers are my intended audience, so I have attempted to counter the weightiness of the technical topics by focusing on the search for readily understandable management headaches such as the loss of productivity due to high participation in the fantasy football pool or the potential for dirty jokes to become evidence in an employment law claim.

Thesis Supervisor: John Van Maanen, Ph.D.
Title: Erwin H. Schell Professor of Organization Studies

Thesis Reader: Pat Bentley, Ph.D.
Title: Senior Lecturer

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).

Acknowledgements

First and foremost, I thank John Van Maanen for being willing to advise a less-than-orthodox thesis. Having first enjoyed his classroom teaching, I had hopes that he could advise without changing the drummer I was following. Without his support and consent, this thesis would be an entirely less interesting work.

Second, Pat Bentley, my Reader, brought the wisdom of her years at Sapient and her eagle editing eyes to bear. Her enthusiasm for the subject and encouragement to use my own "voice" were invaluable. And, she helped me pass one of the toughest tests of all. I often told my law students to write as if a very bright twelve year old were the audience. Learning that Pat's exceptional ten year-old daughter volunteered to help review my thesis is the highest praise that the document is readable.

In the background, there are a host of people who contributed in so many ways and I thank them all, most notably:

- Sir Tim Berners-Lee for turning me away from a topic I did not know well enough;
- Doug Oard, Associate Professor, University of Maryland, and Sonia Sigler, General Counsel, Cataphora Corporation, for taking time in the middle of a National Science Foundation workshop to plant the seeds that became this thesis;
- Jeffrey Heer of the University of California, Berkeley, for the sheer joy that his work brings and for answering frantic emails;
- Robert Liscouski, former Undersecretary of the Department of Homeland Security, now President of Content Analyst, for making his company's technology available;
- Dharmesh Shah, a fellow Sloan Fellow and serial IT entrepreneur, for providing me a website where I posted early work and drew comments from the global IT community;
- Andy Brown, Chief Technology Architect, Merrill Lynch & Co., for expressing enthusiasm for the topic and sharing the "grenade" analogy; and
- Peter Weill, Director, MIT Center for Information Systems Research, for showing that technical topics can be offered in a way that is understandable to the business community.

Last, and most important, are the thanks due to my family. Many thanks are due my father, Arthur Krasnow, for letting me know that MIT exists and for always insisting that I be able to derive mathematical solutions for myself. Special thanks, to my mother, Pearl Krasnow, for being the embodiment of the ideal she always professed: you can do anything you set your mind to; she has changed the world in ways that most people will never know. And, to my husband, Matthew Waterman, who carried me when I was sick, encouraged my every dream, commuted across the country a thousand times; and who, I swear, stops people in the street just to tell them about me. Thanking them will take a lifetime…

TABLE OF CONTENTS

Chapter 1 – Introduction

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 a myriad of advances. Punch card systems became interactive systems; serial processors became multi-processors; the $1 Million (32 Megabit) mainframe computer became the more powerful $1,000 (2 Gigabyte) laptop; the 300 baud suction-cup modem became the wireless multi-Gigahertz modem card; programming advanced from machine language requiring the ability to convert things into 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.

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

---

[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).

introduction of the World Wide Web in the 1990's. As email gained dominance, business and personal communications migrated to this medium.

Email poses a tremendous challenge for organizational knowledge management. Business transactional data remains in corporate databases, but "soft" business discussions – planning, human resources, marketing, etc. – occur increasingly through email and out of formal organizational records. Each individual email account forms what is commonly called a "silo" of information, a negative connotation that the information is harder to access or apply because of its isolation. In the case of email, this is further exacerbated by the fact that the knowledge is generally lost altogether when employees leave the company.

Also, email fosters an informality that may reduce productivity or lead to corporate liability. Today, the CEO is ultimately responsible for every inappropriate employee act, whether that act involves violating government regulations, company policy, or the rights of others. A significant amount of that sort of inappropriate conduct takes place in, or is evidenced through, email. How, then, is the corporate manager to become aware of such conduct? Should he or she wait for one employee to turn in another employee? Should someone be assigned 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. Corporations need "Knowledge Discovery" tools 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. 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?

Some tools already have been built to analyze emails, either for spam-filtering or for the purpose of *retroactive* analysis: support for litigation, intelligence, or archival activities. I wanted to know if the same technologies could support business managers in *pro-active* management activities. In an effort to understand how Knowledge Discovery could be used on an organization's emails to support operations management, I surveyed existing research and performed some experiments of my own. The core of this thesis describes the research and my conclusions about how the technology can be applied to identify emails that could create costs, liability, or compliance issues for a corporation.

My research and my conclusions were aided by my prior professional experience. Based upon my Information Technology and broader operations management experience, I know that understanding the scope and volume of personal use of corporate email will provide a significant clue to sizing losses in systems costs and lost productivity. In addition to my general management experience, I have practiced law. From that work, I have some expertise in matters relating to employee misconduct and am aware of the sort of words, phrases, and documents that could lead to corporate liability.

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 also 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 his or her knowledge – finding other existing projects on the topic, other employees with similar interests, or obtain legacy knowledge – through email Knowledge Discovery.

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 uses the Enron email corpus to bring the concept of the "Compliance Bot" to life.

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

compliance issues. And, I have identified and used one tool that had not previously been used on the Enron data.

The latter part of the thesis describes what a bot could do: how it could intercept outgoing email and make instant decisions about whether emails are problematic, then block or reroute them to appropriate management personnel, and ultimately provide system-wide reporting on trends. I conclude that a compliance bot would be a useful tool for corporate management. Further, I believe I have shown that sufficient technology exists to build the first such bot.

Chapter 2 - 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. In 1971, the first inter-computer email was sent on ARPANET, a government-created precursor to the internet.[2] 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 equal 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.[3] In 1985 the Wang company, which sold word processing systems that were much less expensive than mainframes and accessible to smaller

[2] "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).
[3] "100 Years of IT," Frank Hayes, *Computer World* (April 5, 1999) (http://www.thocp.net/reference/info/100_years_of_it.htm).

companies, introduced Wang OFFICE which integrated internal company email with word processing.[4] By 1988, Wang recognized that companies would need to connect multiple email systems and offered gateways in the software 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).[5] Compuserve began offering service in 1989. At about the same time, Sir Tim Berners-Lee created the World Wide Web[6] and, in 1990, he posted the first website.[7] In 1993, AOL (America Online) began offering the sort of internet service we are still familiar with today.[8]

Email usage and storage became 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 limitation for an individual's email file on Microsoft's Outlook was 2 Gigabytes,[9] roughly equivalent to the storage needed for more than 16,000 20 page documents[10] or 642 copies of the e-book

---

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

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

[6] Many people mistakenly believe that the Internet and the World Wide Web are the same thing. The Internet is the network of networks that connects all the computers, while the World Wide Web is the means of accessing information on the Internet (through hyperlinks). *See, e.g.,*"Frequently Asked Questions" Sir Tim Berners-Lee (http://www.w3.org/People/Berners-Lee/FAQ.html).

[7] *Weaving the Web*, Sir Tim Berners-Lee with Mark Fischetti, pp. 28-30 (Harper Business, 2000).

[8] *See,* Crocker, above at n.5.

[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).

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 20 Gigabytes.[13]

By 2005, one market study determined that corporate users were averaging 133 email messages (sent and received) per day, adding a storage requirement of 294 Megabytes (MB) per user 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 five 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 archiving, this equates to $113.29[17] per average user per year for each year's worth of

---

[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 releasing software that would keep users from reaching the maximum file size) (http://www.windowsitpro.com/Article/ArticleID/22509/22509.html).

[13] *See*, "The .pst file," above at n. 9.

[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) (costs not adjusted for inflation) (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/aiim2003/2003TCOSummary).

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

[17] Calculated by adding all numbers in the series 1 through 12 (representing the aggregation of twelve months' data) and dividing by twelve (to determine the average monthly storage requirement) and multiplying by the average one month cost of $17.43. See also, "Linux e-mail set-up slashes costs to £8 per user," Cliff Saran, Computer Weekly.com (May 6, 2003) (finding £8017 per user for MS exchange email

email for data storage costs alone. In a company with a five-year retention period, the storage cost is $566.45 per user.

Chapter Summary: Email is a phenomenon with its roots in the 1960's. Its primary growth driver was the creation of the World Wide Web in the mid-1990's. Power users now maintain email files greater than the equivalent of 320,000 pages of text. It is estimated that corporations with five-year email retention policies are spending approximately $566.45 per employee to store emails.

---

services in 2003) (http://www.computerweekly.com/Articles/2003/05/06/194340/Linuxe-mailset-upslashescoststo%0c2%0a38peruser.htm).

Chapter 3 - Email Challenges for Corporate 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.[19] With the passage of the Civil Rights Act of 1964,[20] 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.[21]

Since 1964, Congress and the States have passed a flurry of laws regulating employer/employee relationships. The law now prohibits discrimination based upon race,

---

[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] *See, e.g.,* Fair Labor Standards Act, 29 U.S.C. § 201, *et seq.* (enacted June 25, 1938) (setting overtime pay requirements) (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) and National Labor Relations Act, 29 U.S.C. § 151, *et seq.* (enacted July 5, 1935) (establishing employees' rights to collective bargaining and unions) (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).

[20] 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).

[21] *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.

national origin, gender, religion[22] and, in some circumstances, age,[23] disability, [24] pregnancy,[25] familial status,[26] or sexual orientation.[27] The law requires employers of a particular size to grant employees leave to handle serious family matters.[28] There are laws detailing the manner in which benefits, pensions, and insurance[29] can be provided. And, there are laws regulating employment contracts, background investigations, termination procedures, payment of salary, and many of other topics.[30]

In addition to all of these laws that regulate how an employing organization should treat its employees, there is quite a bit of law allocating responsibility to the employer for the conduct of its employees. Since the 1850's, stockholders have been able to bring lawsuits against companies for management conduct which inappropriately diminishes the value of

---

[22] The Civil Rights Act of 1964 outlaws discrimination based upon race, national origin, religion, and gender. 42 U.S.C. § 2000e-2(a) (http://www.law.cornell.edu/uscode/html/uscode42/usc_sec_42_00002000---e002-.html).

[23] Age Discrimination in Employment Act outlawed discrimination against people over the age of 40 (29 U.S.C., Chapter 14, §§ 631 *et seq.* (1967) (http://www.law.cornell.edu/uscode/html/uscode29/usc_sup_01_29_10_14.html).

[24] Americans with Disabilities Act severely limits the circumstances under which disability may be considered in an employment decision (42 U.S.C., Chapter 126, §§ 12101 *et seq.*) (1990) (http://www.law.cornell.edu/uscode/html/uscode42/usc_sup_01_42_10_126.html).

[25] Pregnancy Discrimination Act (42 U.S.C. § 2000e(k)) (1978) (http://www4.law.cornell.edu/uscode/html/uscode42/usc_sec_42_00002000---e000-.html).

[26] *See, e.g.,* Md. Ann. Code art. 49B § 16 (including "marital status" and "sexual orientation") (http://mlis.state.md.us/cgi-win/web_statutes.exe); 10 New Jersey Statutes Annotated 5-4 (New Jersey Law Against Discrimination includes "marital status," "familial status," and "affectional or sexual orientation") (http://lis.njleg.state.nj.us/cgi-bin/om_isapi.dll?clientID=133006&Depth=2&depth=2&expandheadings=on&headingswithhits=on&hitsperheading=on&infobase=statutes.nfo&record={34F6}&softpage=Doc_Frame_PG42); CA Govt Code § 12940 (including "marital status" and "sexual orientation") (http://www.leginfo.ca.gov/cgi-bin/waisgate?WAISdocID=1662057699+0+0+0&WAISaction=retrieve).

[27] *Id.*

[28] Family Medical Leave Act (29 U.S.C. §§ 2601, *et seq.* (1993) (http://www4.law.cornell.edu/uscode/html/uscode29/usc_sec_29_00002601----000-.html).

[29] The Employee Retirement Income Security Act (ERISA) regulates all three. 29 U.S.C. §§ 1001, *et seq.* (1974) (http://www.law.cornell.edu/uscode/html/uscode29/usc_sec_29_00001001----000-.html).

[30] *See, e.g.,* 23 Arizona Revised Statutes §§ 201, *et seq.*

the corporation.[31] And, corporations can be sued for negligent hiring – for failing to perform the pre-employment investigation that would have revealed the likelihood that an individual would cause harm.[32] Some states will hold the corporation liable if a supervisor attempts to coerce an employee to break the law and then causes the termination of the employee for being unwilling to do so – for example, refusing to lie to a legislative committee[33] or refusing an instruction to perform medical work for which one is unqualified.[34]

Because email is used so widely and so frequently, it is a statistical certainty that an entire corporation's repository will contain 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 for employee misconduct, 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 seems 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. Apparently, these individuals do not realize that "deleted" emails do not really disappear; they remain in digital storage and can be discovered later, often many years later, when someone asserts or searches for misconduct.

---

[31] *Ross v. Bernhard*, 396 U.S. 531, 534-35 (1970) (describing the history of derivative actions and listing *Dodge v. Woolsey*, 18 How. 331 (1856) as establishing this principle). (http://caselaw.lp.findlaw.com/cgi-bin/getcase.pl?court=us&vol=396&invol=531).

[32] *See,*e.g., *Proctor v. Wackenhut Corrections Corp.*, 232 F.Supp.2d 709, 2002 WL 31528482 (N.D. Tex. 2002 ; Garcia v Duffy 492 So.2d 435 (1986).

[33] *Peterman v International Brotherhood of Teamsters, Local 396,* 174 Cal. App. 2d 184, 344 P.2d 25 (1959) (http://online.ceb.com/calcases/CA2/174CA2d184.htm).

[34] *Winkleman v Beloit Memorial Hospital,* 483 N.W.2d 211, (Wisconsin 1992).

A 2005 survey of 1,000 people found that 68% of employees have sent or received an email through a work-based account that could place the company at risk.[35] A 2004 study of 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.[36] The number could be significantly higher, as another quarter of the survey respondents did not know the answer to the question.

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 protecting their emails that parallels the Constitution's protections against warrantless seizures. Generally, these people are mistaken. At present, there is no single federal law that addresses the question of privacy for workplace email.

The Electronic Communications Privacy Act of 1986 (ECPA) makes it illegal to intercept electronic communications between two people.[37] Some privacy advocates believed that this would protect employees from employer interceptions of their emails. However, the

---

[35] "Risky Business: New Survey Shows Almost 70 Per Cent of Email-Using Employees Have Sent or Received Email that May Pose a Threat to Businesses," PR Newswire (November 15, 2005) (referring to 2204-2005 Harris Interactive survey commissioned by Fortiva) (http://www.prnewswire.com/cgi-bin/stories.pl?ACCT=104&STORY=/www/story/11-15-2005/0004216193&EDATE=)

[36] "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).

[37] 18 U.S.C. § 2511(1) (http://www.law.cornell.edu/uscode/html/uscode18/usc_sec_18_00002511----000-.html).

law 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.[38] 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 in ECPA for consent.[39] If the employee consents to the employer looking at his or her email, the employee has no claim to privacy for the email. And, an employer can require a potential employee to waive most rights as a condition of employment. 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 they agree never to talk about these things 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.[40] In 2005, more than half were monitoring employee emails.[41]

---

[38] 18 U.S.C. § 2511(2)(a)(i) (http://www.law.cornell.edu/uscode/html/uscode18/usc_sec_18_00002511----000-.html).
[39] *Id.*, at § 2511(2)(d).
[40] *See,* "2004 Workplace E-mail and Instant Messaging Survey," above at n. 36, pp.2 & 4..

With so much monitoring going on, I expect there to be new issues raised regarding the inappropriateness of certain monitoring (i.e., when does monitoring become stalking?) It is very likely that there will be additional lawsuits, court decisions, and new laws to balance the employer's need to manage the business and the employee's desire for privacy.

In addition to voluntary access to employee emails, employers are also subject to involuntary searches. A 2004 survey showed that 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.[42] The existence of email retention requirements implies that those saved emails may be audited or reviewed by others in order to determine compliance. This too is a potential to have persons other than senders and recipients reading email – the emails are not "private."

Another form of involuntary access to employee emails is litigation related disclosure. The 2004 survey indicates that more than 1 in 5 corporations have had employee email subpoenaed by a court.[43] 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 the American Bar Association adopted rules for electronic discovery in August 2004;[44] the rules explicitly list email as a form of data that parties and courts should consider when

---

[41] 2005 AMA survey.
[42] *See,* "2004 Workplace E-mail and Instant Messaging Survey Summary," above at n. 36, p. 3.
[43] *Id.*, at p. 1.
[44] "Final Revised Standards," subsection of Report 103B - Amendments to the Civil Discovery Standards (revised as of 6/04), Electronic Discovery Task Force, Section of Litigation, American Bar Association (http://www.abanet.org/litigation/taskforces/electronic/ and http://www.fjc.gov/public/pdf.nsf/lookup/ElecDi12.pdf/$file/ElecDi12.pdf).

preparing a list of materials required to be preserved or produced. By 2005, the proposed amendments had been modified and submitted to the US Supreme Court for adoption throughout the federal system. The submitted rules imposed even greater burdens, adding electronically stored data to essentially every rule that addresses discoverable material and specifically requiring relevant electronically stored information to be pro-actively disclosed at the beginning of litigation.[45] Thus, the reasons and methods for obtaining access to employee emails continue to grow.

Obligation to Proactively Search for Violations

We know that employers *may* look at employee emails and sometimes do. *Must* employers 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.

In 1998, the United States Supreme Court issued a decision[46] 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

---

[45] Report of the Committee on Rules of Practice and Procedure, and at Pp.12-13 (September 2005) (recommending amendment to Rule 26) (http://www.lexisnexis.com/applieddiscovery/lawlibrary/Excerpt_CV_Report_072505.pdf).

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

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?

The Enron, WorldCom, Tyco, Arthur Andersen, and other scandals propelled Congress[47] to pass the Sarbanes-Oxley Act in 2002.[48] In the simplest terms, the law requires publicly traded companies to report on their accounting controls.[49] However, corporations are struggling with what tasks they need to undertake to show that they have effective financial controls in place. A variety of IT strategies have been undertaken, not only to further secure and account for access to the accounting system, but also to find any indications of financial manipulation.[50]

---

[47] "Reforming the Boardroom: One Year Later, the Impact of Sarbanes Oxley," Allison Fass, Forbes.com (July 22, 2003) (http://www.forbes.com/technology/corpgov/2003/07/22/cz_af_0722sarbanes.html).

[48] Sarbanes-Oxley Act of 2002, Pub. L. No. 107-204, 116 Stat. 745 (July 30, 2002).(http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=107_cong_public_laws&docid=f:publ204.107).

[49] *Id.* (stating the purpose is "To protect investors by improving the accuracy and reliability of corporate disclosures made pursuant to the securities laws...").

[50] *See, e.g.,* "More Companies Tap IT for Sarbanes-Oxley," Thomas Hoffman, *Computerworld* (Oct. 17, 2005) (stating that 75% of respondents to a survey expected to spent significantly on IT as part of the

What to Look for in Emails?

Assuming that corporate management wants or is obligated to look at employee emails, what exactly should it seek? There are several obvious target categories. As described above, the corporation should look for emails that indicate current corporate malfeasance or employment misconduct. As one senior corporate IT manager puts it,[51] he's looking for any live "grenades" and wants to know when the pin is pulled! And, there is another category of email to hunt, the casual personal email - the siphon of corporate resources.

Criminal and Regulatory Malfeasance

Emails may contain the trail of a variety of crimes. This can include emails which provide the proof of insider trading (i.e., emails that send privileged information about a publicly-traded company to someone outside the scope of the privilege); emails that directly, or obliquely, reference inappropriate changes to books and records in an attempt to create the false appearance of financial health; or emails that indicate illegal pass-through of personal expenses as corporate expenses. Such emails can provide evidence of violations of securities, internal revenue, and other laws.

---

methodology to comply with Sarbanes-Oxley.)
(http://www.computerworld.com/governmenttopics/government/legislation/story/0,10801,105463,00.html).

[51] Conversation with Andy Brown, Chief Technology Architect, Merrill Lynch & Co.(March 1,2006).

Personal Use of Corporate Resources

While I worked on Wall Street in the 1980's, corporations realized that employees' use of company long-distance telephone services had grown to the point that it was affecting the bottom line. The companies were not only saddled with high toll charges, but they had expanded infrastructure to support the call volume. Throughout the region, consultants were offering services to reduce these overall costs to corporations. I remember one client that discovered an employee, who worked as an operator, was patching her siblings through to their home country *on a daily basis.* Throughout the country, companies began to scrutinize their phone bills, implement control policies and audit controls on their long-distance usage. In larger corporations, millions of dollars of savings were realized. At the same time, the corporations reaped the dual benefit of gaining back what had been lost in employee productivity. The general consensus was that the value of the savings was far greater than the cost of the consultants.

In a 2004 survey conducted by the American Management Association, nearly all employees claimed that they engage in personal use of corporate email less than 10% of the time.[52] In another survey, nearly 10% of employees admitted to sending their resume to a potential employer from their work email account.[53] However, one company that mines corporate email for litigation estimates that non-work-related emails make up approximately 1/3 of email traffic.[54] Employees use corporate email to talk about and gamble on sports, make social plans, disseminate jokes and inspirational stories, exchange

[52] *See,* "2004 Workplace E-mail and Instant Messaging Survey Summary," above at n. 36, p. 6.
[53] *See,* "Risky Business" above at n. 35.
[54] Tel. call with CTO of litigation discovery company (Feb. 10, 2006) (anonymity requested).

pornography, and handle household chores. This correlates to an annual cost of $188 per user, if the company has a standard five-year retention policy.

Evidence of Discrimination

Emails can provide evidence to support claims of discrimination. In the most direct cases, emails actually state the specific intention to prefer someone of one "class" (a race, ethnicity, gender, etc.) to person(s) of a different class. Imagine a series of emails by the partners in a law firm about hiring "the busty blonde."[55] Or, consider a hypothetical email that says, "I think we should give the promotion to [*male name*]. I know [*female name*] is probably better qualified, but her husband makes a good living, so they don't need the money." In other, more glaring cases, an email can contain such outrageous slurs against a person (or people) of a particular group, that discriminatory animus cannot be denied; these would include emails with terms such as "nigger," "kike," and "raghead." In most organizations, the discovery of such statements will result in swift action by management. Disciplinary action would be instigated against the sender and management would act to mitigate the effects on the recipients.

Another sort of discriminatory animus is the "hostile environment." In those situations, there is a pervasive attitude toward a particular class, an agglomeration of words and acts

[55] One email in the Enron corpus 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 SDOC_No 805666; " Offensive E-Mail http://news.bbc.co.uk/hi/english/sci/tech/newsid_1530000/1530458.stm ."]

that make it clear that a particular group is not welcome, or is considered lesser. This might be evidenced, for example, through a litany of distributed jokes about Polish people. In a recent survey, 48% of the respondents had sent or received emails of questionable tone or content that might be implicated here.[56] And, of course, sexual harassment by a manager of a subordinate is a form of hostile environment. I remember a story from my trial days about a boss who made all his female employees sit on his lap at the company Christmas party each year. Imagine the emails that are sure to have circulated about this!

Other Issues – Management, Liability, Risk

Emails also hide the tell-tale signs of other problems that may ultimately represent costs to the company. I remember another company Christmas party, at which I was present, where a junior employee got so drunk that she lost consciousness in the restroom and paramedics had to be called to resuscitate her. Had she suffered brain damage or died, the following days' emails would likely be evidence in the lawsuits filed against both the firm and the restaurant for continuing to serve someone so inebriated. Or, perhaps, they would be relevant in a human resources decision to have her evaluated and/or treated for alcoholism.

We all can think of circumstances in which someone has shared his or her user-ID and password for a computer system. In a recent Harris survey, 22% of respondents admitted

[56] *See,* "Risky Business" above at n. 35 (respondents admitted to sending /receiving joke emails, funny pictures/movies, funny stories of a questionable tone (e.g., inc racy/sexual content, politically incorrect) (http://www.prnewswire.com/cgi-bin/stories.pl?ACCT=104&STORY=/www/story/11-15-2005/0004216193&EDATE).

to such conduct.[57] Email repositories contain evidence of such violations of corporate policy and more than one person has been fired for sending a "joke" email from someone else's account.

More than just violations of policy (or good conduct) in and of themselves, password sharing can cause another, more significant problem. Individual email stores are generally not secured to the same extent as repositories specifically identified as containing high value content. Thus, it is more likely that a user-ID and password can be stolen from an email than from a system administrator. In a hacker's hands, the user-ID and password can be carte blanche to damage a system or steal its information. This is the domain of the corporate risk manager.

When to Look in Emails?

In 2004, nearly 80% of corporations surveyed had email content policies and more than half provided email policy training to their employees.[58] Nonetheless, it is clear that employees regularly violate those and other policies. In each of the aforementioned cases, if the emails are in the corporate store, they are "grenades" whose pins have already been pulled. In a perfect world, management would have the capacity to analyze emails in real-time and to block the transmission of those that are problematic; they would stop employees from pulling the pins.

---

[57] *Id.*

[58] *See,* "2004 Workplace E-mail and Instant Messaging Survey Summary," above at n. 36, pp. 2 & 4.

Chapter Summary: During the same time period that email has grown exponentially, the law defining corporate responsibility and liability has grown tremendously as well. Laws and regulations now assess corporate liability for management acts, or failures to act, on topics as diverse as sexual harassment and accounting fraud. Part of the method for fulfilling compliance obligations in these areas is to have better transparency of activity occurring through email. Despite common mythology to the contrary, employees do not generally have a right to privacy in their workplace emails. And, there is at least some trend towards the idea that management should pro-actively search emails for signs of crime, discrimination, regulatory non-compliance, as well as violations of policies for use of corporate resources. The current approach to this obligation is to search stored emails after they have been sent – described by one corporate manager as akin to looking for "grenades" after the pins have been pulled. Instead, I propose that management be given the ability to identify such problematic emails before they are transmitted and to stop them from being transmitted.

Chapter 4 - Knowledge Discovery: Meaning from Chaos

There has been a fortuitous confluence of events in modern business history. Just at the time email usage began exploding and the regulation of corporate conduct began to mature, a third topic also began to evolve. The field of "Knowledge Discovery," also known as "data mining," began to take shape.

What is "Knowledge Discovery"?

Consider that a corporation must handle, on average, one million emails per year for every 28 employees.[59] Obviously, no manager or team of compliance employees is going to read this volume of traffic one email at a time. It is too expensive to hire the number of people required and they could not possibly retain sufficient concentration or follow the thread of multiple emails between parties. So, how can corporations figure out if there are issues to be addressed?

[59] Calculated as 133 emails per person per day times 22 business days per month x 12 months per year (based upon email count from "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).

"Knowledge Discovery" is the answer. Knowledge Discovery addresses the issue "how does one understand and use one's data"[60] 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."[61] Knowledge Discovery is a cross-disciplinary field that draws from "statistics, databases, pattern recognition and learning, data visualization, uncertainty modeling, data warehousing, [On Line Analytical Processing], optimization, and high performance computing."[62] Most simply, it is described as the ability to convert "data" to "knowledge."[63] In this case, it is the means for getting valuable information about millions of emails without reading each one.

The Association for Computing Machinery (ACM), the first computing society, founded in 1947 and currently sustaining over 80,000 members,[64] is one of the world's premier professional computing organizations. The term "Knowledge Discovery" was coined at a 1989 ACM workshop,[65] the same year that Compuserve began offering internet-based

---

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

[61] 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/).

[62] 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).

[63] "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) .

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

[65] "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*

email. By 1995, interest in the topic had spread throughout the world, into governmental, commercial, and academic communities.[66] The first journal on the subject, *Data Mining and Knowledge Discovery Journal,* began publication in 1997.[67] In 1998, the year the Supreme Court focused on the proactive obligations of employers towards sexual harassment, the understanding of Knowledge Discovery was still nascent – described as approximately fifteen years behind the understanding of databases.[68]

By 2003, the "Business Analytics" market was estimated at $13 Billion (US).[69] A recent study revealed that companies using such tools reported a median Return on Investment of 112%, while a significant number saw a return of 1,000% or more.[70] The mean payback period was a swift 1.6 years, with the average project costing $4.5 Million.[71] "Business intelligence," which is largely Knowledge Discovery/data mining, is estimated to reach $3.3 Billion in 2006.[72] The Data Mining market is expected to continue to grow at 10% to 20% per year.[73] This bodes well for being able to produce an email data mining system at a price point that would be acceptable to consumer corporations.

---

for Information Science, Vol. 27, No. 1 (October/November 2000) (http://www.asis.org/Bulletin/Oct-00/jurisica.html).

[66] 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/).

[67] 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).

[68] *Id.*

[69] "Eye on Information," Alan Joch, *Oracle Technology Network* website (http://www.oracle.com/technology/oramag/oracle/05-jan/o15eye.html)

[70] "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).

[71] *Id.*

[72] *Id.*

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

How Can Knowledge Discovery Help?

This section explains, in layman's terms, the general mechanisms by which Knowledge Discovery works. Much like the way you don't absolutely need to know how a car is built to drive a car, you don't absolutely need to know what the magic is inside Knowledge Discovery to understand the rest of this thesis. However, my father – the engineer – wouldn't let any of his children drive a car without that understanding and, in the long run, that served me well. I could make judgments about sounds and smells from a car, making good decisions about when to pull over immediately and when to keep driving. As important, it gave me the ability to talk to auto mechanics and make fast judgments about whether to trust their work and their price. This section is offered in much the same spirit. The business manager who has a basic understanding of the underlying mechanisms of Knowledge Discovery may be better able to recognize a fatal flaw or to describe a problem to his/her "mechanic" – the programmer building or adapting a compliance bot.

Knowledge Discovery generally refers to three steps: pre-processing, processing, and visualization. Pre-processing is the work necessary to make data useable. Processing is the automated finding of patterns in data. Visualization 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. I do not, and discuss all three phases in this thesis.

Pre-processing

More than forty years ago, the phrase "garbage in garbage out" came into common usage[74] 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, in a 2003 poll 89% of respondents reported that at least 40% of data mining project time was spent on pre-processing and nearly two-thirds of respondents indicated that they spent more than 60% of their time on pre-processing.[75]

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,[76] it may not be possible to manipulate the data with the very software that created it. Even a novice user has had the experience of receiving an email or an email attachment that wouldn't open at all or opened but was unreadable. Also, I have seen instances in which data entry personnel typed the right information into the wrong fields, guaranteeing that searching the database by field would not yield the best possible results. It has been estimated that field error rates are at least

---

[74] "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.

[75] "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)).

[76] "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).

5%.[77] These are the sorts of problems that are addressed by data "cleansing." The following items are sometimes included within the broad umbrella of "cleansing."

- Integration: Data collected or created in one data platform – a program, or a vendor's software – 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 migration 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 (e.g., matching "Gina" and "Regina" or "Connolly" and "Conelly").

[77] *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").

- 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.[78] Everyone has had the experience of receiving two of the same catalog in the mail and discovering some slight difference in his or her name on the label (ie., 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 who share the identical name. Generally, disambiguating tools attempt to find other data (e.g., address, birthdate, height) associated with each record that will answer the question conclusively.

- De-duplicating: 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.[79] For example, as described more fully in my discussion of the Enron email processing, de-duplicating can result in under-counting the size or underestimating the impact of stored information.

---

[78] *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).

[79] *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).

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.[80] Others – Bayesian, neural networks, and genetic algorithms[81] – 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 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.[82]

- 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

---

[80] "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).

[81] *Id.*

[82] "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).

computer automatically determine rules that better-than-market performing stocks have in common.[83]

- 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?).[84] And, although 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.[85]

- Deviation analysis looks for outliers – data which falls outside normal patterns – and then attempts to discover the cause for the variation.[86] A classic example is credit card fraud detection.[87] 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

---

[83] "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).
[84] "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).
[85] *See,* "Knowledge Discovery in Databases" above at n. 80.
[86] "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/).
[87] *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).

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 or she wants. As the game continues, the player keeps track of all cards he or she 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.[88] 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.[89]

---

[88] "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).
[89] *Id.*, at 6.3.2.

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, an application could be handled in 1 second, using 250K of processor memory.[90] At that efficiency, any business quality personal computer could likely handle more than a thousand at once.[91] This is welcome news for the business manager wondering how to keep pace with the millions of emails moving through the corporate system.

Visualization

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 analysis.[92] 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 five years. In reality, that one small graph is presenting the knowledge of about 126,252 data points,[93] but it is easy to absorb the essence of that information. The difference between such a graph and a great

---

[90] *Id.*

[91] 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.

[92] "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).

[93] Calculated as (52 weeks * 5 days a week) minus 8 holidays per year)) times (500 stocks + 1 calculated average each day). 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).

Knowledge Discovery visualization tool is that the great tool will allow you to zoom in and see the details underlying the simple image.[94]

Chapter Summary: About twenty-five years after the creation of email and the enactment of the Civil Rights Act, and just a few years before the creation of the World Wide Web, a new field of "Knowledge Discovery" was begun. Knowledge Discovery uses a variety of automated strategies to make it possible to find meaningful information in volumes of data that are too large for people to manually comprehend. These tools use mathematics and statistics to analyze the data. Generally, the methodology involves three parts. The first part is pre-processing, getting the data into a format that can readily be analyzed. The second part is processing, the analyzing process. And, the last part is visualization, providing results in a manner that can readily be assimilated – a picture is worth a hundred thousand numbers.

[94] *See,* "Crossing the Information Visualization Chasm" above at n. 92, slide 13.

Chapter 5 - Enron Emails: The Practice Set

A significant challenge for Knowledge Discovery researchers has been the lack of availability of real emails for study.[95] A major research opportunity unfolded when the Federal Energy Regulatory Commission (FERC) released a large set of emails from the Enron Corporation's repository in March 2003.[96] Enron was a very high profile,[97] seemingly extraordinarily successful[98] energy company in Houston, Texas that was ultimately revealed to have engaged in systematic accounting fraud. The company filed a Chapter 11 bankruptcy in 2001 when the fraud was revealed, and operated as a reorganized company for some time, though it is now liquidating all remaining assets.[99] Criminal trials began in January 2006.[100] FERC released the emails (on an Aspen Corp. website) as part of its investigation into the manipulation of oil and gas prices by a number of firms.

---

[95] "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) .

[96] "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).

[97] The company was called "America's Most Innovative Company" for six consecutive years by *Fortune* magazine. *See, e.g.,*"The Rise and Fall of an Energy Giant," BBC News World Edition (Nov. 28, 2001) (http://newswww.bbc.net.uk/2/hi/business/1681758.stm).

[98] *Id.* (At its peak it claimed more than $100 billion in revenues.)

[99] See, Voluntary Petition of Enron Corp., electing Chapter 11 protection (dated 12/2/01) http://files.findlaw.com;news.findlaw.com/docs/enron/enronchp11pt120201.pdf); the Enron Corporation's website page entitled "Confirmation Order (Including Debtors' Supplemental Modified Fifth Amended Chapter 11 Plan) and Related Documents") (The company's Plan of Reorganization was confirmed in July 2004 and the reorganized debtor had been in operation since that time) (http://www.enron.com/corp/por/); In re: Enron Corporation, 01-16034- AJG (SDNY) (substantial legal proceedings have continued regarding the bankruptcy estate; approximately 10,000 legal pleadings have been filed in the case since that time) (docket at https://ecf.nysb.uscourts.gov/cgi-bin/login.pl?376956217176112-l 826 0-1); Enron webpage (announcing in April 2006 that remaining assets are being liquidated and distributed) (http://www.enron.com/corp/)

[100] *See, e.g.,* "Top Enron Officials' Trial Begins Today," Ben White and Carrie Johnson, *The Washington Post* (Jan. 29, 2006) (http://www.washingtonpost.com/wp-dyn/content/article/2006/01/29/AR2006012900864.html).

Email Statistics

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.[101] The emails quickly became notorious for the variety of non-business content (including spam, jokes, and pornography) as well as the evidence of inappropriate business conduct.[102] 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.[103] 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.[104] This resulted in a reduction of the database by approximately 8%.[105] 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.[106] Viewing the official site, it appears

---

[101] "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).

[102] 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/).

[103] 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/2004/03/09/2004030916163910 42.doc).

[104] *Id.*

[105] *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).

[106] *See,* "Third Order On Re-Release Of Data" above at n. 104.

that there are approximately 1.4 million emails.[107] A closer examination of the data quickly reveals that some have no message[108] and others are duplicates.[109] Also, there is very little obvious spam in the collection, so it is assumed that these were the emails actually received, after spam-filtering.

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.[110] Also, most users received far more emails than they sent;[111] most employees had sent 500 or fewer emails, with a significant number sending up to 1,000, but only 8 users had sent more than 2,000.[112] 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.[113]

---

[107] 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).

[108] See, e.g., S_DOC Nos. 21, 22, 25, 27 by continuing from the steps in n. 108 above., and sequentially reviewing documents.

[109] 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).

[110] *See,* "The Enron Email Dataset" above at n. 95, p. 4 & Figure 2.

[111] *Id.*

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

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

The Simple Boolean Search – Preliminary Knowledge

To appreciate what Knowledge Discovery can do for a corporation, it helps to understand what one would know without Knowledge Discovery. Any corporation does have the ability to do a bit more than just random searching in the data; it has the ability to perform Boolean searches. Think of the data like a pile of playing cards. Random searching correlates to "pick a card, any card." Boolean searching offers a sophisticated game of "go fish" – "is there an email with the word 'football'?" or "is there an email with the word 'blonde' and the word 'joke'?" Boolean logic permits search questions using the three terms "and" "or" and "not."[114] I used Boolean search tools offered by FERC/Aspen and the University of California, Berkeley[115] to get a peek into the dataset. I searched for evidence of some of the concerns for the corporate manager. This manual search provides a baseline to compare against the results of Knowledge Discovery work described later in the thesis.

Discrimination/Hostile Environment

First, I searched for emails containing the word "blonde" and looked at the first one hundred closely. Even in this small group, it was clear that a corporate manager would need to subdivide them further to identify emails of concern. For example, within the

---

[114] "Boolean Searching for the Web," Joe Barker, University of California, Berkeley, The Teaching Library (2002) (http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/Boolean.pdf).

[115] During the time of my research, Berkeley had made its web-based searched tool available over the internet (*see*, reference to the tool at: http://bailando.sims.berkeley.edu/), however access (through a link which was at: http://bailando.sims.berkeley.edu/ENRON_email.html) has recently been withdrawn (and reference removed from the page).

group there were emails that related to corporate social events and emails that related to purely personal social events. The compliance officer is unlikely to be concerned with emails announcing corporate social activities.

Within the group, there was a broad range of jokes. Using a rough approximation of the often-described but not released movie rating system,[116] I could see jokes that I would rate

- G: those meeting none of the following criteria;
- PG: one or two uses 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)

Whether the corporation wants to permit transmission of jokes at the G or PG level is more a question of personnel policy – a question of the mood and tone the company wants to set. The transmission of R and X rated content raises the specter of the "grenade" with the pin already pulled. So, too, do the emails I discovered in this subset with content derogatory to women, derogatory to men, derogatory to gay people, and derogatory to various religions. Boolean search alone doesn't provide the details of whether these are being mailed by a supervisor to a subordinate, or among management personnel. To determine this, the

---

[116] *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).

compliance officer would need to hand-match the discovered emails against a corporate organization chart. If these are the senders/recipients, the corporation could be facing liability related to sexual harassment or hostile environment.

Just out of curiosity, I tried looking for a few other items. Searching for "see you tonight" produced a few instances of people who were in intimate relationships. Searching for a variety of offensive slang terms for female anatomy rapidly produced instances of pornography. A little bit of directed surfing produced many copies of the "booty call agreement," an email that has been circulating on the internet since 1999, containing a "contract" with the social rules for casual sexual relationships.

Personal Business

Looking for straight-forward examples of personal business was relatively easy. I searched the FERC/Aspen (F/A) database for "doctor" and it produced more than 2,500 results. Eighty percent of the first 50 were personal: mostly about doctor's appointments and discussions of doctors; drugs for sale; and jokes. Ten of the 50 were news stories or the bio of Ken Lay, described as receiving an honorary "doctor" of laws degree. I searched for "plumber" and received only 95 hits, but nearly all were emails about plumbers' appointments at home or copies of an inspirational email that happened to mention "plumber" in a list of service people. And a search for "babysitter" produced 181 hits that were overwhelmingly about finding a babysitter, having a babysitter, feeling like a babysitter, and multiple copies of an obscene joke that mentions a babysitter.

Financial Misconduct

It quickly becomes clear that Boolean search is not an effective means for finding the financial scandal that was brewing at Enron. I searched for "books" in the F/A dataset and the result was more than 12,000 hits; I reviewed the first hundred and found articles circulated after the news of the Enron scandal broke and a few emails from outside vendors selling various books. A search for "restate" also produces emails about the scandal and the requirements placed upon Enron after the fact. The simple Boolean search method did not readily produce anything that would provide evidence of accounting impropriety.

Boolean searching does provide some ability to find emails about issues of concern, but it is quite limited. Like the "go fish" analogy, Boolean search allows you to find only a card you can describe exactly. It differs from "go fish" because you don't know how many cards are in the deck or how many of that particular card are in there.

Chapter Summary: In March 2003, the federal government released hundreds of thousands of emails from the senior managers of Enron. The emails have provided the first significant repository for researchers and have received significant media attention for the large amount of non-business mails (including spam, jokes, and pornography). In order to have some baseline understanding, I performed Boolean searches, looking for evidence of inappropriate sexual content, personal chores, and financial improprieties. I was able to find some of the first two items but none of the third.

Chapter 6 - Pre-processing: The Case Against "Cleansed" Data

Cleansing data, a process described in Chapter 4 under "Pre-processing," is usually thought of as a helpful tool in the analysis of large data sets. For the purposes of compliance analysis of corporate email, this may not be the case. Cleansing can obscure a variety of issues.

The original dataset released by FERC was over a million items. There are a variety of versions of the dataset in use. MIT acquired a copy of the data and discovered a variety of integrity problems.[117] SRI, International attempted to cleanse the data as a part of its CALO (Cognitive Assistant that Learns and Organizes) Project;[118] That version of the data, which is available for research, contains 517,431 emails from 151 users.[119] The CALO version has removed all attachments from the emails; attachments remain available in the FERC data. Multiple researchers determined that this dataset also contained emails they considered appropriate for cleansing; duplicates and error messages. USC researchers further cleansed the data and reduced the total to 252,759 emails (48.84%).[120] Carnegie Mellon researchers created a dataset of 619,446 from 158 users that they reduced to

---

[117] "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/).

[118] *Id.*

[119] *Id.* And, *see,* "The Enron Email Dataset" above at n. 95.

[120] *See,* "The Enron Email Dataset" above at n. 95 (indicating a dataset of "252,759 messages from 151 employees distributed in around 3000 user defined folders").

200,399 (32.35%) from 158 users.[121] Cleansing techniques affect results, as two other research groups identified 149[122] and 161[123] users (without 100% overlap). Each cleansing activity created different numbers of users and different volumes of data. It appears that there are at least four versions of the dataset: FERC/Aspen, USC, CALO (used by Carnegie Mellon and University of California at Berkeley), and Queens University.

I wanted to understand how a cleansed dataset might differ from an original and what impact that might have on compliance analysis, so I structured a test around the word "blonde." Based upon prior work experience, and an unscientific review of the data, I expected the emails containing that word to be personal in nature and mostly to contain jokes. 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.[124] I knew that the original FERC/Aspen dataset had nearly three times the number of emails, and wanted to know what the two-thirds were that were eliminated in the cleansing process..

I manually reviewed the resulting emails and categorized one hundred of them in an Excel spreadsheet (attached as Appendix 1). To understand the cleansing process, I tracked the following elements:

---

[121] "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).

[122] "Enron Email Dataset Research" Andres Corrada-Emmanuel, University of Massachusetts, Center for Intelligent Information Retrieval, Department of Computer Science (mapping file identified in "MD5 Digest to Relative Filepath Mapping")(http://ciir.cs.umass.edu/~corrada/enron/index.html).

[123] "Enron Dataset" Jafar Adibi and Jitesh Shetty (a link to an Excel spreadsheet with the list of 161) (http://www.isi.edu/~adibi/Enron/Enron.htm; http://www.isi.edu/~adibi/Enron/Enron_Employee_Status.xls)

[124] In later research, I discovered that the Queen's University research shows 88 occurrences of "blonde" despite a much larger cleansed set of 289,695 emails.

- **FERC/Aspen (F/A) Sdoc_No**
  the unique numeric identifier added by F/A
- **Dup**
  the unique number identifier for an F/A stored email that was a duplicate of another email already tracked
- **UCB DatabaseID**
  the unique numeric identifier added by UC Berkeley (UCB)
- **Date**
  the date the email was sent
- **Topic**
  a short description of the content of the email
- **What F/A recorded**
  - From an Enron email account?
  - To
    - How many Enron email accounts?
    - How many non-Enron email accounts?
  - Folder
    - Sender or Recipient
    - Location
- **What UCB recorded**
  - From an Enron email account?
  - To
    - How many Enron email accounts?
    - How many non-Enron email accounts?

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 their online search tool and found the same email. The two copies appeared not to have any identifying number in common.[125]

[125] F/A showed an "SDOC_NO" 31046 while UCB showed a "DatabaseID" 18295.

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. 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 also will be important to understand the traffic flows between official corporate email accounts and personal email accounts. For example, in this subset, there were two occasions on which a person received something relatively obscene (a dirty joke[126] and a pornography subscription[127]) and then forwarded it to an account that appeared on its face to be his own personal (non-business) email account (i.e., samename@yahoo.com or samename@hotmail.com). If the person next forwarded the entire email to others from his personal account, the email would still contain a reference to Enron (listed as the original user@enron.com email recipient) and the company would have no notice of how many times or places it traveled. This should be of tremendous concern to the company both because of the unknown cost to reputation and the complete inability to mitigate any circumstances in which a subsequent forwarding constitutes sexual harassment. The same

[126] F/A SDOC_No 46619.
[127] F/A SDOC_No 757975.

issue will be of even greater concern if an employee is emailing corporate financial information, legal advice, or insider secrets to his or her personal account.

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."[128] A search of the UCB data revealed two emails[129] with the same date and subject from the same employee but neither of them 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[130] is the one with a timestamp of 08:29 PDT. This timestamp conversion occasionally results in a different date (e.g., converting a timestamp from 7/31/01 02:01:40 GMT to 7/30/01 19:01 PDT).[131] It appears that the majority of the emails were sent or received in Texas at the Enron headquarters city. From the perspective of compliance, the local time for the email would be most useful, as personal emails may be read differently in the context of daytime and nighttime. Imagine how differently a female employee might read an email about her appearance or clothing from a male coworker if it arrives in the middle of the workday or arrives at 11 pm when they are the

---

[128] F/A SDOC_No 793655.
[129] UCB DatabaseID 207169 and 207170.
[130] UCB DatabaseID 207169.
[131] See, e.g., F/A SDOC_No 806665 stamped 7/31/01 02:01:40 GMT and matching UCB DatabaseId 7/30/01 19:01 PDT.

only two people left in the building – it might be the difference in perception between inappropriate and stalking .

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[132] is from the employee's "all documents" folder and three of the copies[133] are from the employee's "inbox" folder. Interestingly, there are other similar duplications involving the same user. Three more[134] are the responsive emails 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 at least four "sent" folder copies of an email[135] from the employee about car trouble and his possible interest in being fixed up with a "tall blonde." It is unknown whether these errors existed in the Enron database or were the result of the FERC/Aspen recovery process.

De-duplication and the Loss of Location Data

[132] F/A SDOC_No 160741.
[133] F/A SDOC_No 162270, 166329, and 171762.
[134] F/A SDOC_No 160755, 166343, and 173325.
[135] F/A SDOC_No 155940, 163584, and 173175.

The F/A set includes the details of where the email was found but the UCB search result does not include that data. For example, F/A data reveals whether 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 goals of the party analyzing the emails. 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: did the recipient of an X-rated joke put it in the "Deleted" folder? Save it to a personally-created folder called "Fun Emails"? Or, perhaps to one called "Harassment" or "Evidence"?

Summary Statistics

Here are the relevant statistics for the comparison of the first 100 emails returned by the F/A system's search for "blonde" and the search for matching emails in the UCB cleansed dataset:

- Within the F/A's 100 emails:
  - 47 (47%) are unique emails
    - This correlates closely with Berkeley's overall result of producing a cleansed set that is 48.8% of the size of the complete F/A set.
    - 45 of the 47 (≈96%) are personal emails

    - 45 of the 100 (45%) were additional copies of the unique emails
    - 8 of 100 (8%) were blank
        - correlating exactly with the 8% removal by FERC in response to privacy requests

- Comparing the F/A's 47 unique emails with the UCB cleansed emails:
    - 46 of the 47 (≈98%) are in the UCB set
        - 1 email is not there, an approximately 2% loss rate
    - 12 of UCB's cleansed copies of the 46 emails in common (26%) identify the sender or recipient email addresses differently
        - Relative agreement on number of emails "sent" by Enron employees
        - 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

It is important to recognize that the parties who cleansed the dataset were doing so for other analytic purposes. No criticism is intended; as described later, their work is fascinating and advances the state of research overall. The small changes to the data are irrelevant for their purposes. For example, if one is analyzing the text of the messages, the time or user-ID is of no consequence.

"Cleansing," though, may not be the best tool for compliance monitoring. In the case of this particular cleansing mechanism, a compliance manager could easily argue that not reaching 2% of the emails and having IDs and times changed will have a significant impact on overall effectiveness. Each cleansing tool will have a different impact on the data and it may be difficult to determine what ancillary issues are created. Since cleansing is normally performed on a stored copy of a dataset, this indicates that there is no strong reason to work on a stored copy as opposed to the "live" data. This conclusion supports my earlier stated suggestion of performing real-time analysis on emails before they are

transmitted. Based upon these results, and all of the foregoing information, I would recommend against cleansing data before processing it for compliance.

Chapter Summary: In order to understand what "cleansing" might do to an email dataset, I compared the results of searching for "blonde" in the full government-released dataset and in a cleansed dataset. In both cases, all the emails are personal and most are jokes. The most significant difference between the sets was that more than a quarter of my sample of the cleansed emails reflected different sender or recipient email addresses. Also, the cleansing process altered the times of the emails. These changes appear to have been irrelevant for the cleanser's purposes, but would be important in a compliance context because they affect the perception of the parties to a communication and the timing of those communications. At a minimum, someone using cleansed data must know exactly what changes the cleansing is causing. Overall, I assert that this is another reason to support real-time analysis of emails in transmission over analysis of emails in storage.

Chapter 7 - Processing: Gathering the Details about Enron

A number of Knowledge Discovery research activities have already centered on the Enron emails. This chapter describes the work performed and, where possible, how it would contribute to the creation of a compliance bot.

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.[136]

Deception Analysis

[136] "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).

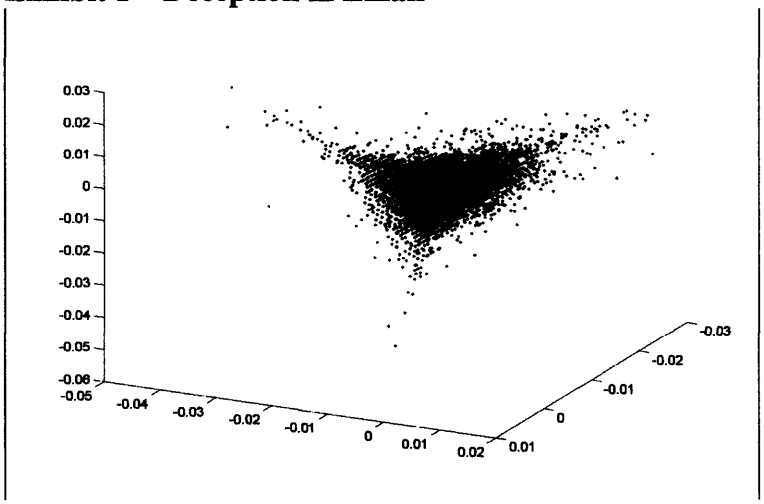
This group presented a paper in October 2005, explaining how they used the word counts in the application of "deception theory," which asserts that certain word choices are more common in deceptive writing.[137] Specifically, they looked for less than normal usage of "first person pronouns (I, me, my, 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.)."[138] 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 – a technique that reveals the components that underlie a matrix.[139]

The resulting plot is roughly a downward pointing triangle shape with elongated points:[140]

---

[137] "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).
[138] *Id.*, at p. 4.
[139] "Using the Singular Value Decomposition," by Emmett J. Ientilucci, Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, p.1 (May 29, 2003) (http://www.cis.rit.edu/~ejipci/Reports/svd.pdf).
[140] *Id.*, at p. 6, Figure 2.

Exhibit 1 – Deception in Email[141]



The upper left point represents high usage of exclusive words and is described as "emotionally charged" emails to co-workers, family, and friends.[142] 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.[143] 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.[144]

The Queens research team notes the value of this success. A corporate manager could select emails of interest without engaging in the labor intensive task of reading them all.

---

[141] *Id.*
[142] *Id.*, at p. 4
[143] *Id.*, at p. 5.
[144] *Id.*, at p. 8 and p.9, Figure 5.

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 email at a farthest extreme could be chosen to be read.

I believe this research provides additional valuable information for the compliance 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, on the upper left, might reveal discussions of other employees' misconduct.

And, while the emails seemed relatively evenly distributed, this perception was dispelled when the researchers color-coded the data points to reflect the authors of the emails.[145] Based upon the color-coding, Enron senior executives appear most often in the personal pronoun and action verb points.[146] 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[147] would suggest) or deception (as the current indictments[148] suggest), another explanation is

---

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

[146] 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.

[147] "The Decline and Fall of the Enron Empire, Tim Grieve, Salon.com (Oct. 14, 2003) (http://dir.salon.com/story/news/feature/2003/10/14/enron/index_np.html) .

[148] "Former Enron Chairman and Chief Executive Officer Kenneth L. Lay Charged with Conspiracy, Fraud, False Statements," Press Release of the United States Department of Justice (July 8, 2004) ("This indictment alleges that every member of Enron's senior management participated in a criminal conspiracy to commit one of the largest corporate frauds in American history") (http://www.usdoj.gov/opa/pr/2004/July/04_crm_470.htm).

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 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 more frequently angry, deceptive, or focused on outside recreation than the people outside the company with whom they exchanged communications. Again, further analysis should be performed: in this case, to determine if the total number of emails from inside or outside the corporation could skew the data.

Pure Word Counts

The Queen's University count contains 160,203 words[149] drawn from its own cleansed version of the data containing 289,695 emails.[150] Clearly, a business manager cannot

[149] "Other Forms of the Enron Data" webpage, Professor David Skillcorn, "Word list in decreasing frequency order"(http://www.cs.queensu.ca/home/skill/otherforms.html and http://www.cs.queensu.ca/home/skill/unique_n3.txt).
[150] *See,* "Detecting Unusual and Deceptive Communication" above at n. 138.

regularly review a list that's more than one hundred thousand items long. However, that doesn't make the list unusable. 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
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

A fast scan of this list of highest usage words could satisfy such a manager that the majority of the references seem reasonably related to official business.

Hostile Environment

A human resources manager (or attorney) might look at the occurrence list for words associated with potential employment law issues such as the previously described "hostile environment" claim. For example, emails containing words and slang describing parts of a woman's anatomy are potential evidence of a hostile work environment for women. I

searched for such words, leaving out words for which I could quickly identify another possible connotation (i.e., 379 occurrences of "breast" 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 and another 172 occurrences, related to the word "sex", related to the concept of sex, or that likely demarcated a pornographic website (e.g., "sexxx" and "SexyWhiteThang18"). Thus, in about two hours, I had identified 556 occurrences that might lead to liability for the company.

It is important to note that the significance of such a finding is not how many occurrences were found, but that any occurrences were found. Depending upon the circumstances, even a few examples could support an employee's hostile environment claim. Hundreds of occurrences of crass references to female anatomy and pornography could reflect many managers whose attitudes would be considered "hostile" to women and, therefore, discriminatory. Since all the emails in our sample are from management employees, a compliance manager receiving this report would be seeking additional information (e.g., how many different managers were involved, whether highest level managers were involved, and how many of the emails were sent to female subordinates) to determine whether these are individual incidents or a widespread trend. The manager would need to determine whether the emails represented potential liability for individual employee claims

or whether they might support a claim that the corporation as a whole tolerated or fostered a hostile environment for women, presaging a more expensive class-action liability situation.

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, management will immediately terminate the employment of the author. With such a small number of results, the company could easily address the issue.

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 nine terms.

3230[151]	doctor	1108
10756	mechanic	161
11549	plumbing	143
13179	dentist	113
20143	babysitter	53
21546	plumber	47
36329	babysitting	19
55281	repairman	9
57445	babysit	8
	TOTAL	1,661

[151] The numbers in the first column indicate where the word sits in the list of occurrences in order of usage. "Enron" was number 1 with over 370,000 occurrences and "forThanksgiving" was number 160203 with one occurrence.

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 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: The football search, in particular, will be relevant to later discussions of more revealing Knowledge Discovery technology.

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.

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 earlier) 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" (e.g., female or Catholic) 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.

Automated Categorization

One approach to processing is to reorganize emails into categories. At least two groups have taken subsets 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 annotated the emails with the categories.[152] The activity was a class exercise that did not result in statistics or visualizations, but the labeled emails have been made available for review or use.

As of March 2006, a Masters student at the University of Minnesota, Duluth, Department of Computer Science, under the direction of Associate Professor Ted Pedersen[153] reported the manual annotation of 3,000 emails from the University of Massachusetts, Amherst collection.[154] She will use the manual annotations as a benchmark against which to

---

[152] "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).

[153] *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).

[154] "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) (reported the annotation of 1,000 emails) (www.d.umn.edu/~tpederse/Group05/ap-slides-nov4.ppt): the information was subsequently updated via an

compare the results of automated clustering.[155] Unlike the Amherst work, though, multiple users' emails were categorized into a single set of common categories and subcategories.[156] So far, they have calculated the following distribution: Business – 45.25%; Personal – 26.22%; Human Resources – 14.2%; General Announcements – 10.82%; Enron Online – 2.98%; and Chain Mails – 0.53%.[157] Combining the 26.22% Personal, the 0.53% Chain letters, and the 8% personal emails that FERC removed, this indicates a total of nearly 35% personal email in the Enron corpus, a lot of time and money spent by the corporation's employees on activities that did not benefit the corporation.

At least one group has attempted to categorize the emails using an automated method. 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.[158] 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.[159] For each person, it took only the emails he or she 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

---

email from Apurva Padye to K. Krasnow Waterman (March 23, 2006) (based upon having seen a draft copy of this thesis posted online) (directing me to http://www.d.umn.edu/~tpederse/enron.html).

[155] 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).

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

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

[158] *See,* "Automatic Categorization of Email into Folders" above at n. 156.

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

few emails to allow meaningful study.[160] The topical emails were then removed from their folders and re-sorted into chronological order, [161] essentially re-creating the stream of emails delivered over time.

A series of tests were performed in which a tool first was given a set of emails and the name of the folder where each had been filed by the user and then the tool was instructed to file a similar number of emails that had arrived next in time. [162] 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[163] 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.[164] Winnow ran substantially faster than the other applications and Wide Margin Winnow was appreciably more accurate than Winnow.[165] 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 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.

---

[160] *Id.*, at pp. 4-5 and see p. 11, Table 1 (showing that the Enron sample set was approximately 19,500).
[161] *Id.* at p. 5 ("...after sorting the messages according to their time-stamp, we train the classifier...").
[162] *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).
[163] *Id.*, at p. 11.
[164] *See, Id.*, pp. 12-13 and Tables 3 & 4 (providing and discussing accuracy results per user per application).
[165] *Id.*, at pp. 14-15.

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.[166] 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 from the body of the text.[167]

If this task could be done successfully, there would be several benefits for general 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." And, in theory, the compliance manager

[166] *Id.*, at p. 14.
[167] *Id.*, at p. 6.

could sort the entire set into "questionable accounting," "credit card misconduct," or "password sharing."

Thread search

We are all used to seeing topical discussions on online bulletin boards, web pages where someone posts an initial item and then other people respond and, sometimes, the original poster writes again. Those websites are built to collect and organize information by "threads." Producing a similar presentation from a corporate email corpus involves a much greater technical challenge because the emails were not collected or stored in that way. This requires a technology that will detect topics in emails, find related emails, and put them in chronological order. The Joint Institute for Knowledge Discovery, at the University of Maryland, is working on a tool that makes it possible to follow a topic thread through the Enron corpus.[168] Unfortunately, the technology was between iterations during the time of my work so I didn't get to try it or see it in operation.

This technology could be very useful to a business manager. If the manager finds an email with inappropriate "hostile environment" content, he can follow the trail to determine everyone who received it, forwarded it, reacted to it, etc. The same holds true if the manager finds an email that indicates a compliance irregularity.

---

[168] *See,* e.g., "JIKD Email/Speech Update" Doug Oard, University of Maryland, Joint Institute for Knowledge Discovery, slide 3 (October 26, 2005) (zaphod.mindlab.umd.edu:16080/JIKD/Presentations/05Oct2005/jikdupdateoct05.ppt).

Latent Semantic Indexing

Another technology, Latent Semantic Indexing (LSI) categorizes the words in emails into context categories and then uses the contexts as the means of comparison. Like SVD described in the section about Deception Analysis, it also uses "vector space" techniques.[169] In simple terms, this means that it should be able to match an email about "dogs" to one about "canines" even if each never uses the other word. This is an interesting concept because it means that one need not identify every possible word choice in order to capture all of the relevant emails; if it works, it represents a significant improvement over the Boolean search technique.

Content Analyst is a commercially available software tool that provides LSI. The provider company, of the same name, provided me with research access to the tool. And, they loaded the CMU cleansed version of the dataset containing approximately 200,000 emails. I was the first person to use the tool against the Enron emails.

Before beginning any formal experiments, I played with the tools a bit. I quickly came up with the sense that clustering emails about "things" would be easier than clustering emails that represented concepts. For example, I thought it might be easier to successfully cluster emails about football than to cluster jokes or emails that would support a "hostile environment" claim.

---

[169] "Abstract: Large-Scale Information Retrieval with Latent Semantic Indexing" Todd A. Letsche and Michael W. Berry, University of Tennessee, Department of Computer Science (1996) (http://www.cs.utk.edu/~berry/lsi++/node1.html#SECTION00010000000000000000).

Personal emails

My first experiment, then, was to attempt to cluster all emails about football. The "query" function returns two results: (1) a list of terms that cluster most closely with your search terms and phrases and (2) a list of emails starting with the highest correlation first. The system is designed to work against substantial content input, but I tried giving the system only three words ("colts steelers rams") to see if that was enough. Even though three words should have been insufficient for maximum effect, using only those three terms yielded surprisingly good results as described below. I asked the system to show me the 100 most frequently associated terms. (*See,* the table on the next page.) Almost all of the 100 terms are the names of teams, coaches, and players. There are a few references to game, injury, websites, and other football-related terms.

The first ten thousand emails (5% of the total corpus) the system returned had correlation scores from 7431 down to 854 (out of a possible 10,000). Initially, the results also seemed to strongly support the benefits of LSI. The very first email is not about the Rams, Steelers, or Colts, but it is about football. The first email is an article about running back Jason Brookins returning to the Ravens after previous service to them, the Raiders, and the Jaguars. And the 10,000th email was a forward about watching two wide receivers at a football practice at the University of Texas.

Query Results
Found 100 Terms: *(Best Score = 10000)*

	Term	Score
1.	colts	9642
2.	denver's	9560
3.	steelers	9538
4.	rams	9521
5.	bengals	9505
6.	panthers	9442
7.	browns	9410
8.	broncos	9395
9.	edgerrin	9302
10.	jets	9194
11.	packers	9191
12.	cardinals	8873
13.	toomer	8839
14.	redskins	8589
15.	hairline	8568
16.	canidate	8565
17.	amani	8513
18.	lions	8465
19.	proehl	8461
20.	biakabutuka's	8430
21.	correll	8396
22.	grbac	8304
23.	saints	8271
24.	tendonitis	8264
25.	sharpe	8259
26.	bledsoe's	8254
27.	ravens	8207
28.	champing	8121
29.	hambrick	8114
30.	haslett	8113
31.	vikings	8073
32.	trung	8044
33.	buckhalter	8040
34.	faulk	8029
35.	tiki	8014
36.	bucs	8008
37.	glenn's	8007
38.	navies	7970
39.	frerotte	7942
40.	donnalley	7880
41.	ayanbadejo	7878
42.	obafemi	7878
43.	flashes	7877
44.	chrebet	7867
45.	vick	7845
46.	workhorse	7811
47.	samari	7781
48.	quincy	7774
49.	biakabutuka	7697
50.	riemersma	7665
51.	seahawks	7646
52.	dayne	7641
53.	olindo	7615
54.	kitna	7598
55.	falcons	7561
56.	tshimanga	7556
57.	pinkston	7555
58.	deuce	7550
59.	mcnabb's	7517
60.	longshot	7500
61.	roaf	7426
62.	canidate's	7397
63.	laveranues	7378
64.	footballguys	7369
65.	heiden	7362
66.	cheatsheets.net	7354
67.	cheatsheets-unsubscribe	7345
68.	mare	7340
69.	fourth-round	7329
70.	biakabatuka	7317
71.	martz	7311
72.	j.j	7307
73.	mariucci	7296
74.	cheatsheets	7285
75.	freddie	7284
76.	gruden	7277
77.	flexor	7276
78.	rolle	7276
79.	eagles	7252
80.	green's	7243
81.	chrebet's	7238
82.	kitna's	7237
83.	pittsburgh	7222
84.	duce	7204
85.	kennison	7167
86.	peerless	7137
87.	soreness	7131
88.	dugans	7117
89.	bears	7110
90.	roster	7107
91.	kevan	7073
92.	www.footballguy stalk.com	7072
93.	serwanga	7063
94.	germane	7043
95.	fabiano	7033
96.	crowell	7032
97.	footballguy	7024
98.	jacksonville's	7021
99.	pathon's	7014
100.	gladiators	7010

A sampling of the emails in between shows that 75% of them contain discussions or mentions of football, and 90% of them are personal (non-business). One of the two business related emails mentions a torn ACL (anterior cruciate ligament), a not-uncommon

football related injury, which probably caused the correlation. Considering that this is the result of providing only three terms, this seems to be a remarkable result.

Hit #	Score	Topic	Message-Id
00001	7431	Ravens running back	28789944.1075853090240
00500	4993	Fantasy football league	9613285.1075862002855
01000	3324	Offering a fantasy football trade	5158749.1075854607972
01500	2503	Football pool	27582376.1075855041175
02000	2103	College football schedule change	29756389.1075851681049
02500	1853	Golf fantasy league	19842248.1075852742677
03000	1661	Personal chat (mentions going to a game)	1419452.1075845686836
03500	1519	"Tonight's game"	26623748.1075854772836
04000	1420	Personal chat (mentions Denver & Miami)	5538512.1075849750025
04500	1330	Personal chat (mentions watching a football game on tv)	13133092.1075856166090
05000	1260	Offering a fantasy football trade	24208671.1075854685347
05500	1186	Fantasy football league	29979611.1075861652567
06000	1124	Playing hockey	6580033.1075857667857
06500	1072	Personal chat (mentions "game" and "basketball")	23198712.1075852548403
07000	1027	Post-bankruptcy office expenses	7353626.1075855436505
07500	988	Personal note (mentions traffic jam that will be caused by UT game)	21050168.1075845613562
08000	954	Energy related; shift change notes (mentions a torn ACL)	10531170.1075860742966
08500	924	Personal chat (mentions Notre Dame/Texas A&M football game)	7815633.1075852761469
09000	898	Fantasy football scoring	24221903.1075852727233
09500	873	Arizona Public Service Company	10183741.1075842267887
10000	854	Wide receivers at UT	7669758.1075857326256

I tried to tighten the cluster by adding more search terms. I ran a search using the thirty names of all the NFL teams. The list of 100 terms it produced had 75% of the same terms as the list generated by just three terms. And the correlation scores were remarkably similar to the first result. The individual emails were different, though. All of the top scoring items were about fantasy football leagues and a smaller majority mentioned football. And, one more was business related.

Hit #	Score	Topic	Message-Id
00001	8050	Fantasy football	31646619.1075855032431.
00500	6245	Fantasy football	12581365.1075861207496
01000	4364	Fantasy football	2840495.1075855170577
01500	2854	Fantasy football	3049284.1075854606472
02000	2327	Fantasy football	2783567.1075855353882
02500	2011	Sports picks	19009393.1075843002132
03000	1794	Sports betting website	26044337.1075840978359
03500	1612	"Game tonight"	26623748.1075854772836
04000	1493	Summary of 2000 (mentions "bears" in market context)	15669054.1075845453895
04500	1395	Kids baseball	17614808.1075861090841
05000	1309	American Express Gold Card ticket services (mentions multiple NFL teams)	17383711.1075852632849
05500	1234	Dallas Cowboy jokes	9973812.1075841191591
06000	1169	Joke about drinking	14020261.1075849789277
06500	1109	"Game at 6"	3698825.1075854153868
07000	1061	Personal chat (mentions children's sports)	9412053.1075845986756
07500	1020	Sports betting website	27750143.1075840013695
08000	981	Anti-Osama rhetoric	14369101.1075862597950
08500	944	Freddie Mac exec on office space	3478362.1075840249370
09000	913	Ford Expedition	22221803.1075858195180
09500	887	SW Air specials	2957087.1075840736489
10000	864	Job applicant thank you note	5565080.1075856352837

My theory is that fantasy football rises to the top because emails on that topic mention a higher number of teams per email.

To really understand this result, I went back to the FERC/Aspen Boolean search tool. Entering "colts steelers rams" produced zero results! This is presumably because it couldn't find the three words in a row. Entering "colts or steelers or rams" produced 971 results. If I really cared about those three teams alone, the Boolean search would give me the better result, because it would give me only emails containing exactly the terms I requested. But, as a manager concerned with use of corporate resources, I'm much more interested to know that there are more than 10,000 emails generally about football, which the LSI tool seems to prove. And, considering that the Queens University word counts indicated high occurrences of words related to other personal topics such as other sports

and drinking, I predict that the LSI tool will rapidly find me thousands more emails of a non-business nature.

Discrimination/Hostile Environment

My second set of experiments with LSI were focused on finding emails that could lead to liability if a female employee filed a hostile work environment claim based upon gender. First, I ran the "query" function using five x-rated terms for female genitalia. After the impressive football results, I was surprised to see that the 100 cluster terms were not all closely correlated to sex. Among the top ten terms were "fabled," "ironic," "vestiges," "chrome," "spies," and "blueberry."[170] I reviewed the emails and discovered that they were not correlated to sex either; there appeared to be a significant mix of personal topics – ranging from dirty jokes to inspirational stories.

Second, I ran a query using only the word "sex." This produced much more relevant cluster terms, but still produced a broad spectrum of personal messages, still producing many which were unrelated to sex or a hostile environment.

It had been explained to me that LSI was not intended to produce results on such a small set of words. So, I next attempted a "find like" search. This function permitted me to identify a number of specific x-rated emails and ask for similar documents. I selected five emails that were both x-rated and derogatory towards women. The results did seem to

[170] The other terms were "impatience," "excitable," "adoring," and "trims" which were at least potentially sex-related.

have a higher proportion of x-rated content, but were still quite varied, including many emails unrelated to sex or hostile environment.

Next, I tried another LSI tool. The software has a function that automatically categorizes results. It creates a taxonomy, a multi-level hierarchy based upon the terms with the highest correlation factor. The result is a series of folders whose names are the few highest correlation terms. My next experiment was to give the "Taxonomist" the body of a single, long email explaining, in graphic detail, the sexual activities associated with a series of slang terms (e.g., "teabagging," "Houdini").[171] This actually produced the least successful result. The 1,602 emails retrieved were completely varied, including a significant number of strictly business emails. I suspected this was due to the large proportion of general vocabulary terms and edited the text down to just the graphic phrases and ran the tool again. This returned over 2,500 emails, but still without any obvious connection to the original pornographic concept and in categorized folders that didn't offer any correlation terms related to the concept of pornography.

I had some success when I adjusted the Taxonomist's setting for the similarity of retrieved documents. Using just the sex related words from the same x-rated email, I ran the Taxonomist at the default setting for document similarity – 30 out of a possible 100. This produced 1024 emails, again including many unrelated to the topic. Moving the setting to 70 produced 0 emails. I then worked down incrementally to 45, which produced 28

[171] While these terms can be readily found through any internet search engine, the author cautions that the resulting pages will be extremely pornographic.

emails, 21 of which (75%) were pornographic jokes, including 7 variants (25%) of the original email from which the group of terms were excerpted..

I tried many other variations and studied the results. It seems that there are a few factors that affect how tightly a result correlates to the topic of interest. As I had suspected, clearly defined topics are easier to find. For example, I reran the Taxonomist with sports terms (football, basketball, etc.) and the results were overwhelming about sports. But, pornography has always been hard to define. Even a Supreme Court Justice ruling on obscenity was ultimately reduced to saying "I know it when I see it."[172] From an analytic standpoint, the first challenge is that the majority of words in pornographic emails are not sexual in nature and, conversely, many non-pornographic emails use potentially sexual terms in non-sexual ways (e.g., "he's a boob," "she's an ass").

Another factor that seemed to have significant impact on results was the fact that this test included email header data (date, time, multiple mentions of a sender's username, recipient's username, etc.). Often the "topic titles" generated by the Taxonomist for the folders were usernames, ISP names, etc. This implies a significant weight being given to those items. I believe that if the tool had the ability to ignore the header data it would produce better results and would be quite useful for some compliance topics.

Despite these challenges, I think that LSI shows promise as a component of the compliance manager's toolkit. The fact that I could create a test for such a vague concept as pornography and produce results increasingly closely tied to the topic is encouraging.

[172] *Jacobellis v. Ohio*, 378 U.S. 184, 197 (1964).

Since I know, from reviewing the emails generally, that there are many more pornographic emails, the work going forward will be to figure out how to generate a larger set of accurate results. And, of course, the work will need to expand to other important but conceptual topics of interest such as accounting fraud and insider trading.

Social Network Analysis

Traditionally, Social Network Analysis has been the category of analytics that determines the *existence* of a relationship between people or the intensity of contact (i.e., volume of interaction) between the parties.[173] A group at the University of Massachusetts, Amherst, extended the reach of the concept by experimenting with technology that attempts to determine the *nature* of the relationship between two people based upon the content of their emails to each other.[174] They ran an experiment trying to determine the likelihood of the nature of the relationship between people based upon fifty topics. The results appear to show some promise, with the highest results occurring in (a) the topic of "sports pool" between the organizer of the pool and a person who I recognize as arising frequently in my various searches of the topic and (b) the topic of "government relations" between an

[173] *See,* e.g, "Social Network Data," Robert A. Hanneman and Mark Ridlle, University of California, Riverside, Department of Sociology, Chapter 1 of *Introduction to Social Network Methods* (2005) (http://www.faculty.ucr.edu/~hanneman/nettext/).

[174] "The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks, with Application to Enron and Academic Email," Andrew McCallum, Andr´es Corrada-Emmanuel, Xuerui Wang, University of Massachusetts, Amherst, Department of Computer Science, Technical Report UM-CS-2004-096 (2004) (http://www.cs.queensu.ca/~skill/proceedings/mccallum.pdf).

executive responsible for Government Relations and the Vice President of Regulatory Affairs.[175]

If this technology could be successfully deployed, it could assist in finding people engaged in a number of activities of concern to a corporation. In theory, it could distinguish among the emails between two people; sorting out those that are strictly work-related and those that are personal. It could probably find management personnel engaged in sexual relationships with subordinates. It might be able to find employees who were agreeing on how to "cook the books"

Chapter Summary: I reviewed the research of many others testing Knowledge Discovery tools over the Enron emails and I experimented with a Latent Semantic Indexing tool as well. I conclude that most of these techniques provide a significant advance from Boolean searching; they make it possible to move from searching for keywords (e.g., names of sports teams) to searching for concepts (e.g., football generally or pornography). Since keywords often have multiple meanings, including ones unrelated to compliance or liability, these Knowledge Discovery tools can help a manager get better tailored results. Since concept searching allows the system to find words the manager didn't think to supply, these Knowledge Discovery tools can help a manager get much more complete results.

[175] *Id.*, p. 6, Table 1.

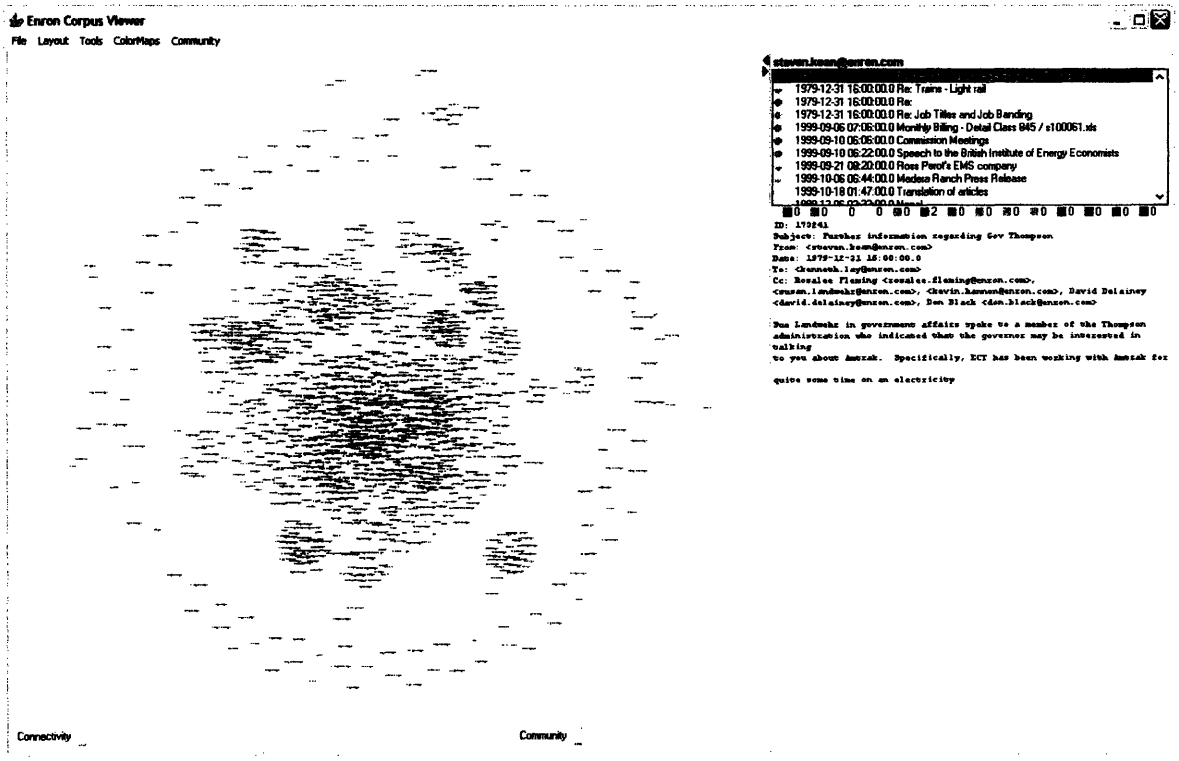
Chapter 8 - Visualization: Seeing the Relationships of Enron

Jeffrey Heer, a graduate student at University of California, Berkeley, provided insights into Enron while working on an Information Visualization ("infovis") toolkit he calls "prefuse."[176] 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, [177] 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

[176] "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).
[177] "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/).

Exhibit 2 – enronic (zoomed out)[178]

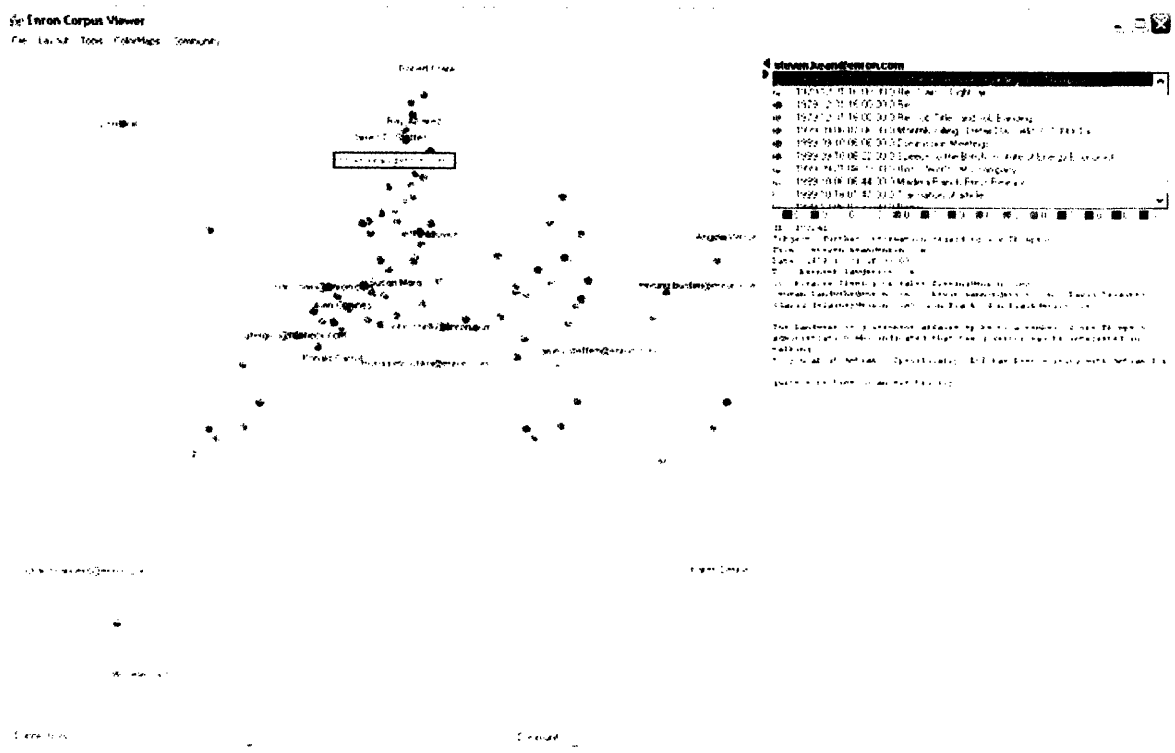


linking line whose thickness grows for each additional email. The picture is supplemented by a pie chart on each line, with each color reflecting a type of email categorized by Professor Hearst's class. For example, a thick line between two people with a dot that's half red and half a particular shade 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,[179] allowing the user to move the nodes, highlight segments, zoom in and out, etc. He added some extremely useful features. The first is a

[178] *Id.*

Exhibit 3 – enronic (zoomed in)[180]



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. The third is another 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

[179] *E.g.*, i2 (http://www.i2inc.com/) and Visual Analytics (http://www.visualanalytics.com/).
[180] *See*, "Exploring Enron: Visualizing ANLP Results [Version 1: white]" above at n. 179.

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 as well, but never responded to any of them. Without knowing the details of the Enron case, Heer 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. So, his work does not show personal relationships. However, if the tool were applied to the hand-categorized data from Minnesota, it would. This would provide another interesting step along the road to producing a compliance bot.

Chapter Summary: I found one advanced visualization tool that had been run against the Enron data. The Berkeley doctoral student who built it didn't know the facts of the scandal, but the tool quickly highlighted an anomaly that identified one of key suspects in the scandal. Generally, this tool makes it possible to sub-select segments from the vast amounts of data, allowing a manager to focus in on the email traffic in a single relationship or group or on the email traffic in a single topic area.

Chapter 9 – Conclusion: Putting it All Together to Build a "Compliance Bot"

What would be the best way to apply Knowledge Discovery to email in order to meet compliance management needs? Some companies use Knowledge Discovery tools to review stored emails and search for problems. But, as the manager I mentioned[181] says, that's looking for "grenades" after the pins have been pulled; that is, it's looking for the problem after potential liability has attached. Wouldn't it be better to try to prevent the problem before it occurs, or at least stop it in its tracks? I propose the building of a "compliance bot," software that automates what a human would see and do at the time the inappropriate act is begun.

The closest corollary to this is a spam filter. Spam, the "junk mail" of the email realm, creates a different sort of cost for the corporation. One research study released in 2003 estimated that employers lost $874 per employee per year due to lost productivity dealing with spam.[182] 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 – estimates that 50-60% of all incoming email, before

[181] "What to Look for in Emails" subsection of Chapter 3, above.
[182] "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).

filtering, is spam.[183] 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.[184] About half of the respondents in a 2004 survey of corporations claimed that spam was less than 10%,[185] but this may be due to corporate filtering before email is delivered to individual users.

E-mining: the Bot that Hunts Email "Grenades"

Technologists have devised software that runs in real-time, grabbing spam at the moment it comes across the communications line. As everyone knows, spam-filtering is not a perfect technology. Spam still slips into people's email boxes and, occasionally, valuable emails end up in the spam file. From the broad brush perspective, though, spam-filtering is an invaluable tool.

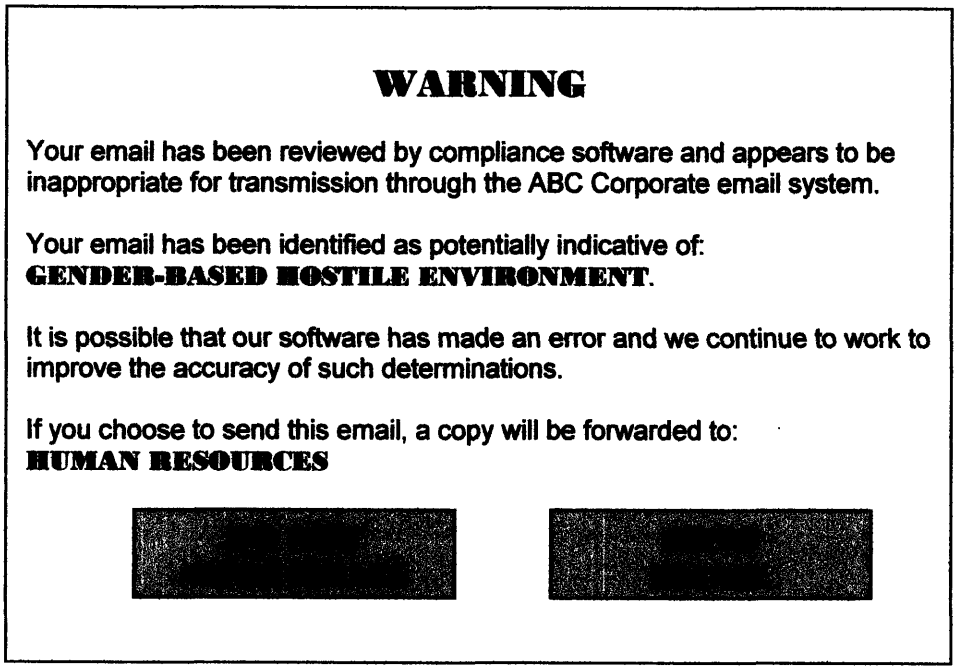
In the context of creating a "compliance bot," then, we already know it is possible to build software that checks each email in transmission without degrading system performance to the point of unacceptability. And, we know from spam-filtering that it is possible to reroute emails based upon a sender's user-ID or keywords. Compliance raises some additional challenges.

---

[183] *See*, VeriSign ROI Calculator (http://www.verisign.com/products-services/security-services/messaging-security-and-compliance/email-security/ROI_Calculator/).
[184] *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).
[185] *See*, "2004 Workplace E-mail and Instant Messaging Survey Summary," above at n. 36, p. 6.

I propose a bot that serves as a compliance monitor *before* an email is transmitted. When a user hits "send," the bot will quickly determine if the email appears to raise any concerns. If the email raises concerns, the bot will display a pop-up window that says something like:

Exhibit 4 – Sample Pop-up Window: Compliance Issue Identified



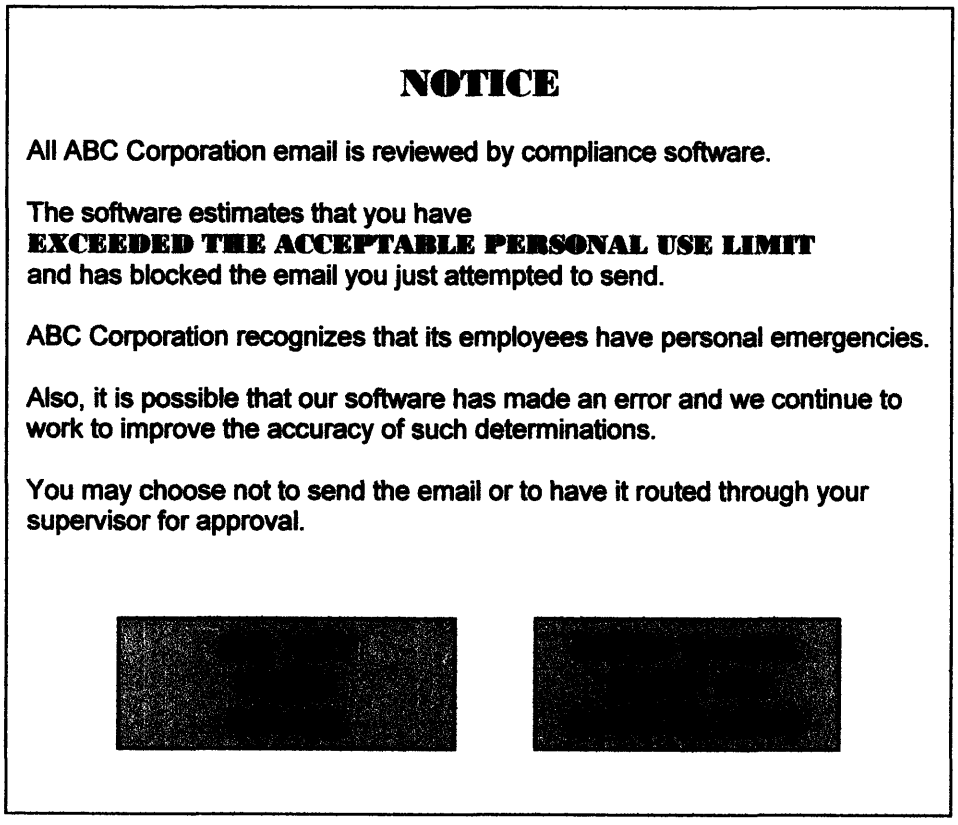
The idea is that the bot will not only determine that something is wrong, but also the likely general category of problem. In the above window the phrase "GENDER-BASED HOSTILE ENVIRONMENT" would be inserted by the bot from a list of options. Each item on the problem list would be paired with one or more interested compliance officers, permitting the second item, "HUMAN RESOURCES," to be extracted from the list and inserted into the pop-up as well.

This sort of functionality is intended to have multiple results. First, for most employees, the mere discovery that the system has such capabilities will result in behavior modification. It is expected that employees will significantly reduce the number of

inappropriate emails that they send. And, it is anticipated that the viral word-of-mouth spread about such functionality will cause a diminution of incoming inappropriate emails from friends and associates; this is not expected to have any impact on incoming spam. Second, by routing copies of high risk emails to the most-likely-appropriate compliance office, the corporation gains a tool for rapid containment of liability.

The bot could also address personal emails in a similar manner, generating a pop-up window like this:

Exhibit 5 – Sample Pop-up Window: Personal Use Limit Exceeded



This, too, is very likely to quickly reduce the total number of personal emails sent (or received) by employees. As the language of the sample pop-up window makes clear, ***I am not suggesting that a corporation should ban all personal emails through its corporate system.*** For corporations that focus on being a highly-rated place to work, that would be public relations suicide. The acceptable level of personal traffic will vary significantly depending upon the nature of the business and the management model of each company.

E-mining: The Senior Management Perspective

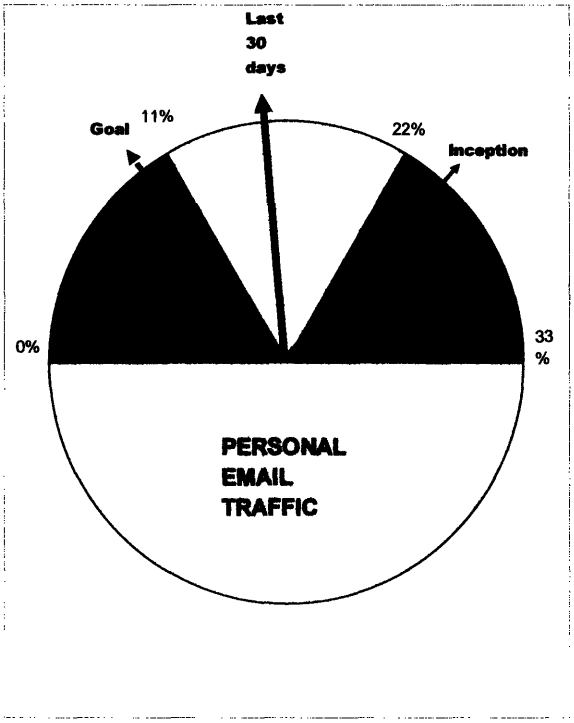
For most of my life, employees have been referred to as "staff." Today, the operative phrase is "human capital," clearly expressing the asset value that employees bring to the corporate bottom line. Email, too, is a corporate asset that should be tracked and managed. The compliance bot will provide this function.

Corporate senior managers could be able to check the status of email usage as quickly as they absorb the daily Dow Jones average. The compliance bot could produce a variety of outputs such as dashboards, pie charts, line or bar graphs. In moments, management could have an update on:

- Liability monitoring
    - the number of emails identified by the bot as potentially creating liability
    - the percentage of emails not sent after the pop-up warning appeared
    - the percentage of sent emails determined not to be of concern by compliance officers
    - reduction in percentage of compliance-issue emails since inception
    - breakdown by compliance category
    - estimated savings (avoidance of legal fees, court awards, and fines)

- Personal use monitoring
    - number of items per day
    - percent of total traffic
    - identities of people with (consistently) higher than average numbers
    - reduction in personal traffic since inception
    - savings in email expenses per user

Exhibit 5: Sample Dashboard Component: Personal Email Statistics



- System use auditing
    - topics being collected
        - persons receiving topic details

This last category will allow corporate senior managers to ensure that the system is not collecting information they consider inappropriate. And, considering the potentially sensitive nature of the emails being retrieved, the audit data also will permit senior managers to ensure that the data is only being transmitted to duly authorized persons.

The bot software could contain a feature that a technical professional I know refers to as "dial-up/dial-down" functionality. This would allow an authorized manager to raise or lower the tolerated level of personal use email. For example, if the bot consistently shows the company's email system has 33% personal emails, the manager could begin to drive the number down by setting the acceptable level at 29%, then later at 25%, and so on. My guess would be that most companies will settle on an acceptable level between 5% and 12%. This could vary by employment category as well. A line worker in a factory may send and receive very few emails, resulting in the number of personal emails being a high percentage. On the other hand, a busy professional can receive 200 emails a day; 10% would be a lot of personal traffic.

Besides being able to drive down personal usage as a long-term cost saving strategy, the dial-up/dial-down function would allow companies to expand and contract the tolerated level depending upon where they are in their business cycle – tight controls in the busy season, laxer controls in the off-season. And, dial-up/dial-down functionality would provide a needed role in crisis management. In the face of natural disasters or terrorist events, the dial-up/dial-down feature will give the company the fast path to compassion, allowing employees to quickly get word to whomever necessary.

As the best Knowledge Discovery tools do, the bot software must "learn." It must be able to adjust its future determinations based upon the accuracy of prior determinations. The compliance officer who receives copied potential-liability emails must be able to press a button or check a box that tells the system whether the item was correctly or incorrectly

chosen. And the supervisor receiving personal emails must be able to check/uncheck a box indicating that the item is not personal. Over time, as the bot adjusts its opinions, system performance and employee observance of rules should improve, driving down the number of emails referred to compliance officers and supervisors. This could result in some other valuable metrics:

- Compliance monitoring:
  - Number of issues addressed
  - Method of addressing issue
    - Employees disciplined/terminated
    - Financial controls changed
  - Time spent by compliance officers and supervisors

The use of the bot and the collection of these metrics would be very helpful in establishing the good faith of senior management in pursuing their fiduciary duties and compliance duties on behalf of the shareholders and the public.

Bots of the Future

My thesis supports the value of using Knowledge Discovery mechanisms on a corporation's email. It shows the vast amount of personal traffic currently riding through corporate systems and the associated costs. In the short term, available Knowledge Discovery tools will allow corporations to drive down personal email traffic to an acceptable level. This should improve the corporate bottom line through immediate incremental improvements to employee productivity and savings on data storage and related costs.

Next, available and in-development tools described herein, will be combined in ways that can aid corporations in meeting their increasing burdens to serve as watchdog over their employees. The thesis provides ample support for corporate obligations to do so and the growing trend towards requiring proactive rather than reactive strategies. This technical advance will have a larger impact on corporate profitability as it will significantly reduce the numbers of claims against the corporation which will translate to reduced legal fees and related costs. It also should vastly improve the ability to provide evidence of mitigating actions, resulting in earlier dismissal of lawsuits or smaller liability determinations, both reducing costs again.

In the longer term, corporations will reap the greatest benefit. The technologies described here will mature until a person can find all relevant information, projects, and people no matter where they reside in the corporation's digital stores. That is the point at which a corporation will most efficiently recognize the value of the work of all employees, present and past.

Appendix A

Appendix A – Comparative Analysis of Cleansed Data

FA Sdoc	dup	UCB dbID	date	topic	from E? (F/A)	To E?	to not E?	location sender recip	location folder	from E? (UCB)	To E?	To not E?	
4040		6224	01/10/02	notice re: band concert	0	1	u	r	inbox				
31046		18295	01/14/02	joke - cosmetic surgery	0	1	5	r	deleted items				
31133		18367	12/18/01	joke - casino	0	1	4	r	deleted items				
41828		35082	05/28/02	joke - 710 cap	1	7	0	r	deleted items				
46619		42180	12/20/01	joke - Santa	1	0	1	r	sent items				Note: to own personal email address
47071		43525	11/29/01	joke - T-G-I-F	0	1	14	r	deleted items				
0													
0													
0													
54113			01/21/02	joke subscription	0	1	0	r	deleted items				
70878		68956	01/22/02	hair dyed	1	1	0	s	deleted items				
70879		68959	01/22/02	hair dyed	1	0	0	r	deleted items				
70881		68957	01/22/02	hair dyed	1	1	0	s	deleted items				
70883		68961	01/22/02	hair dyed	1	0	0	r	deleted items				
77180		74891	02/03/02	columbia house ("Legally Blonde")	0	1	u	r	deleted items				
81411		80131	01/27/02	columbia house ("Legally Blonde")	0	1	u	r	deleted items				
84045		82465	01/12/02	porn sites	0	u	u	r	deleted items				
94082		90609	03/13/01	note between friends	1	0	1	s	all documents				
0													
0													
103050	94082							s	sent				
0													
0													
130367		101908	04/12/01	confirming receipt of blonde jokeSP	1	1	0	s	all documents				

130427		101957	04/18/01	joke - priests on vacation	1	1	3	r	all documents				
130476		101996	04/25/01	joke - ice fishing	1	1	1	s	all documents				
130479		101998	04/25/01	joke - ice fishing	1	1	1	s	all documents				
130627		102122	05/25/01	joke - kidnapper	1	3	0	s	all documents				
131395	130467							s	miscellaneous				
131402	130427							s	miscellaneous				
132914	130367							s	sent				
132251	130476							s	sent				
132328	130479							s	sent				
159940		107314	06/07/00	weekend chit-char + possible fix-up	1	1	0	s	sent	1	1	0	
160741		107962	09/12/00	naked blonde sliding	0	1	0	r	all documents	0	1	0	
160755		107974	09/12/00	naked blonde response	1	0	1	s	all documents	1	0	1	
162270	160741							r	inbox				
163584	159940							s	sent				
163774	160755							s	sent				
165532	159940							s	all documents				
166329	160741							r	inbox				
166343	160755							s	all documents				
171762	160741							r	inbox				
173135	159940							s	sent				
173325	160755							s	sent				
178945		11582	09/25/00	bull joke	0	1	4	r	all documents	0	1	4	
183350	178945							r	inbox				
244608		141150	02/14/00	That's nice website announced	0	1	9	r	all documents	0	1	9	
244610		141152	02/14/00	forwarding that's nice to personal	1	0	1	r	all documents	1	0	1	Note: to own personal email address
252725	244608							r	personal				
253746	244610							r	sent				
258344	244608							r	all documents				
258346	244610							r	all documents				
267546	244608							r	personal				
268573	244610							r	sent				

301883		157523	04/16/01	Blind man on a bar stool	1	0	1	s	sent	1	1	0	
301884		157524	04/16/01	defending blind man email	1	0	1	s	sent	1	1	0	
301885		157525	04/16/01	further defending blind mail	1	0	1	s	sent	1	1	0	
302369		65057	05/30/01	The Daily Reckoning	0	1	0	r	deleted	0	1	0	
310934		160211	05/25/01	Spark Life	0	1	0	r	deleted	0	1	0	
311338		160663	06/01/01	Black eye joke	1	1	0	r	inbox	1	1	0	
427921		179971	07/20/00	Comments re: women & dating	1	1	0	r	all documents	1	1	0	
428140		180174	02/11/00	What's the Difference - gender jokes	1	1	0	r	all documents	1	1	0	
428289	427921							r	correspondence				
428402	428140							r	discussion threads				
428524	427921							r	discussion threads				
429233	428140							r	fun emails				
441422		185003	09/08/00	fw'ing baby announcement	1	1	0	r	all documents				
446061	441422							r	inbox				
577458	301883							s	all documents				
577459	301884							s	all documents				
577460	301885							s	all documents				
577877	301883							s	discussion threads				
578769	301883							s	sent				
578770	301884							s	sent				
578771	301885							s	sent				
720713		161344	01/30/01	Offensive to everyone	0	1	1	r	all documents	0	1	1	
720937	720713							r	discussion threads				
721139	720713							r	saved-01				
733887		195212	12/21/00	email list Xmas	0	1	50+	r	all documents				
734542	733887							r	discussion threads				
735135	733887							r	personal				
753074		3657	10/24/01	Q about meeting a friend	1	0	1	s	deleted	1	0	1	
757040		41996	08/18/01	Blonde in a boat	0	1	7	r	inbox	0	6	0	
757975		83169	07/03/01	Daily pornographic email subscription	1	0	1	s	sent	1	0	1	Note: to own personal email address
761291		110564	10/24/01	Tech products 2	0	1	0	r	deleted	0	1	0	

784901		193953	10/18/01	Tech products for sale	0	1	0	r	deleted	0	1	0	
785018		194006	10/24/01	Tech products 2	0	1	0	r	deleted	0	1	0	
793655		207169	10/04/01	7 Degrees of Blonde	1	4	11	s	sent	1	14	0	
793656		207170	10/04/01	FW: resp to 7 Degrees of Blonde	1	0	1	s	sent	1	0	1	
805666		228756	09/19/01	Labor & Employment alert	0	u	1	u	research	0	2	0	
806665		229344	07/31/01	between friends + sports tix	1	0	1	s	sent	1	0	1	
0													
818766	130367							r	all documents				
818832	130427							r	all documents				
818880	130476							r	all documents				
818883	130479							r	all documents				
819022	130627							r	all documents				
819479	130476							r	miscellaneous				
819485	130427							r	miscellaneous				

BIBLIOGRAPHY

BIBLIOGRAPHY

[ABA] "Final Revised Standards," subsection of Report 103B - Amendments to the Civil Discovery Standards (revised as of 6/04), Electronic Discovery Task Force, Section of Litigation, American Bar Association (http://www.abanet.org/litigation/taskforces/electronic/ and http://www.fjc.gov/public/pdf.nsf/lookup/ElecDi12.pdf/$file/ElecDi12.pdf).

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

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

[ACM SIGKDD] Charter of ACM SIGKDD (http://www.acm.org/sigs/sigkdd/charter.php).

[ADA] Americans with Disability Act, 42 U.S.C., Chapter 126, §§ 12101 *et seq.* (1990) (http://www.law.cornell.edu/uscode/html/uscode42/usc_sup_01_42_10_126.html).

[ADEA] Age Discrimination in Employment Act outlawed discrimination against people over the age of 40 (29 U.S.C., Chapter 14, §§ 631 *et seq.* (1967) (http://www.law.cornell.edu/uscode/html/uscode29/usc_sup_01_29_10_14.html).

[Adibi] "Enron Dataset" Jafar Adibi and Jitesh Shetty (http://www.isi.edu/~adibi/Enron/Enron.htm; http://www.isi.edu/~adibi/Enron/Enron_Employee_Status.xls)

[AMA] "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).

[AZ] 23 Arizona Revised Statutes §§ 201, *et seq.*

[Barker] "Boolean Searching for the Web," Joe Barker, University of California, Berkeley, The Teaching Library (2002) (http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/Boolean.pdf).

[BBC] "The Rise and Fall of an Energy Giant," BBC News World Edition (Nov. 28, 2001) (http://newswww.bbc.net.uk/2/hi/business/1681758.stm).

[Berkeley] "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)

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

[Berners-Lee] "Frequently Asked Questions" Sir Tim Berners-Lee (http://www.w3.org/People/Berners-Lee/FAQ.html).

[Berners-Lee 2] *Weaving the Web*, Sir Tim Berners-Lee with Mark Fischetti (Harper Business, 2000).

[Bethke] "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) (http://www.uscharterschools.org/gb/personnel/intro.htm).

[Bhattacharya] "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).

[Blackburn], "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).

[Braue] "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).

[Burney] "EDD: Demystifying Deduplication," Brett Burney, *Law Technology News* (April 2005) (http://www.law.com/jsp/ltn/pubArticleLTN.jsp?id=1113901507580).

[CA] CA Govt Code § 12940 (including "marital status" and "sexual orientation") (http://www.leginfo.ca.gov/cgi-bin/waisgate?WAISdocID=1662057699+0+0+0&WAISaction=retrieve).

[Civil Rights] 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).

[Civ P] Report of the Committee on Rules of Practice and Procedure, and at Pp.12-13 (September 2005) (recommending amendment to Rule 26) (http://www.lexisnexis.com/applieddiscovery/lawlibrary/Excerpt_CV_Report_072505.pdf) .

[Cohen] "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/).

[Corrada] "Enron Email Dataset Research" Andres Corrada-Emmanuel, University of Massachusetts, Center for Intelligent Information Retrieval, Department of Computer Science (mapping file identified in "MD5 Digest to Relative Filepath Mapping") (http://ciir.cs.umass.edu/~corrada/enron/index.html).

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

[DOJ] "Former Enron Chairman and Chief Executive Officer Kenneth L. Lay Charged with Conspiracy, Fraud, False Statements," Press Release of the United States Department of Justice (July 8, 2004) (http://www.usdoj.gov/opa/pr/2004/July/04_crm_470.htm).

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

[Enron] Voluntary Petition of Enron Corp., electing Chapter 11 protection (dated 12/2/01) ((http://files.findlaw.com;news.findlaw.com/docs/enron/enronchp11pt120201.pdf).

[Enron 2] "Confirmation Order (Including Debtors' Supplemental Modified Fifth Amended Chapter 11 Plan) and Related Documents") (http://www.enron.com/corp/por/).

[Enron 3] *In re: Enron Corporation*, 01-16034- AJG (SDNY) (docket at https://ecf.nysb.uscourts.gov/cgi-bin/login.pl?376956217176112-l 826 0-1).

[Equal Rights] 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).

[ERISA] The Employee Retirement Income Security Act (ERISA) regulates all three. 29 U.S.C. §§ 1001, *et seq.* (1974) (http://www.law.cornell.edu/uscode/html/uscode29/usc_sec_29_00001001----000-.html).

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

[Fass] "Reforming the Boardroom: One Year Later, the Impact of Sarbanes Oxley," Allison Fass, Forbes.com (July 22, 2003) http://www.forbes.com/technology/corpgov/2003/07/22/cz_af_0722sarbanes.html).

[Fayyad] "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).

[FERC] 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/2004/03/09/2004030916163910 42.doc).

[FERC 2] "Addressing the Western Energy Crisis: Information Released in Enron Investigation," Federal Energy Regulatory Commission Website (April 28, 2005) (http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp

[FLSA] Fair Labor Standards Act, 29 U.S.C. § 201, *et seq.* (enacted June 25, 1938) (setting overtime pay requirements) (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) .

[FMLA] Family Medical Leave Act (29 U.S.C. §§ 2601, *et seq.* (1993) (http://www4.law.cornell.edu/uscode/html/uscode29/usc_sec_29_00002601----000-.html).

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

[Fortiva] "Risky Business: New Survey Shows Almost 70 Per Cent of Email-Using Employees Have Sent or Received Email that May Pose a Threat to Businesses," PR Newswire (November 15, 2005) (referring to 2204-2005 Harris Interactive survey commissioned by Fortiva) (http://www.prnewswire.com/cgi-bin/stories.pl?ACCT=104&STORY=/www/story/11-15-2005/0004216193&EDATE=).

[Garcia] *Garcia v Duffy,* 492 So.2d 435((1986).

[Goebel] "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) .

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

[Griffiths] "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).

[Hanneman] "Social Network Data," Robert A. Hanneman and Mark Ridlle, University of California, Riverside, Department of Sociology, Chapter 1 of *Introduction to Social Network Methods* (2005) (http://www.faculty.ucr.edu/~hanneman/nettext/).

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

[Hearst] 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).

[Heer] "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).

[Heer 2] "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/).

[Hoffman] "More Companies Tap IT for Sarbanes-Oxley," Thomas Hoffman, *Computerworld* (Oct. 17, 2005) (stating that 75% of respondents to a survey expected to spent significantly on IT as part of the methodology to comply with Sarbanes-Oxley.) (http://www.computerworld.com/governmenttopics/government/legislation/story/0,10801,105463,00.html).

[Ientilucci] "Using the Singular Value Decomposition," by Emmett J. Ientilucci, Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, p.1 (May 29, 2003) (http://www.cis.rit.edu/~ejipci/Reports/svd.pdf).

[Jacobellis] *Jacobellis v. Ohio*, 378 U.S. 184, 197 (1964).

[Joch] "Eye on Information," Alan Joch, *Oracle Technology Network* website (http://www.oracle.com/technology/oramag/oracle/05-jan/o15eye.html)

[John] "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).

[Jurisica] "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).

[Kalashnikov] "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)).

[KDD-95] Program Committee List, The First International Conference on Knowledge Discovery and Data Mining, KDD-95, at Montreal, Canada (Aug. 20-21, 1995) (http://www-aig.jpl.nasa.gov/public/kdd95/).

[Keila] "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).

[Klimt] "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).

[Korab] "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).

[Kuchar] "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).

[Letsche] "Abstract: Large-Scale Information Retrieval with Latent Semantic Indexing" Todd A. Letsche and Michael W. Berry, University of Tennessee, Department of Computer Science (1996) (http://www.cs.utk.edu/~berry/lsi++/node1.html#SECTION00010000000000000000).

[Maletic] "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).

[McCallum] "The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks, with Application to Enron and Academic Email," Andrew McCallum, Andr´es Corrada-Emmanuel, Xuerui Wang, University of Massachusetts, Amherst, Department of Computer Science, Technical Report UM-CS-2004-096 (2004) (http://www.cs.queensu.ca/~skill/proceedings/mccallum.pdf).

[MD] Md. Ann. Code art. 49B § 16 (http://mlis.state.md.us/cgi-win/web_statutes.exe).

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

[Microsoft] "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).

[Mosher] "Outlook 2002 Hotfix Addresses 2GB Size Limit," Sue Mosher, Contributing Editor, *Windows ITPro Magazine* (Sept. 13, 2001) (http://www.windowsitpro.com/Article/ArticleID/22509/22509.html).

[MPAA] "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).

[NJ] 10 New Jersey Statutes Annotated 5-4 (New Jersey Law Against Discrimination includes "marital status," "familial status," and "affectional or sexual orientation") (http://lis.njleg.state.nj.us/cgi-bin/om_isapi.dll?clientID=133006&Depth=2&depth=2&expandheadings=on&headingswithhits=on&hitsperheading=on&infobase=statutes.nfo&record={34F6}&softpage=Doc_Frame_PG42).

[NLRA] National Labor Relations Act, 29 U.S.C. § 151, *et seq.* (enacted July 5, 1935) (establishing employees' rights to collective bargaining and unions) (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).

[NYSE] "Holidays and Hours" webpage of the NYSE (http://www.nyse.com/Frameset.html?displayPage=/about/1022963613686.html)).

[Oard] "JIKD Email/Speech Update" Doug Oard, University of Maryland, Joint Institute for Knowledge Discovery (October 26, 2005) (zaphod.mindlab.umd.edu:16080/JIKD/Presentations/05Oct2005/jikdupdateoct05.ppt).

[Padhye] "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 (November 4, 2005) (www.d.umn.edu/~tpederse/Group05/ap-slides-nov4.ppt: http://www.d.umn.edu/~tpederse/enron.html).

[PDA] Pregnancy Discrimination Act (42 U.S.C. § 2000e(k)) (1978) (http://www4.law.cornell.edu/uscode/html/uscode42/usc_sec_42_00002000---e000-.html).

[Pedersen] Webpages of Associate Professor Ted Pedersen, University of Minnesota, Duluth, Department of Computer Science (http://www.d.umn.edu/~tpederse/; http://www.d.umn.edu/~tpederse/research.html; and http://www.d.umn.edu/~tpederse/students.html).

[Peterman] *Peterman v International Brotherhood of Teamsters, Local 396,* 174 Cal. App. 2d 184, 344 P.2d 25 (1959) (http://online.ceb.com/calcases/CA2/174CA2d184.htm).

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

[Proctor] *Proctor v. Wackenhut Corrections Corp.*, 232 F.Supp.2d 709, 2002 WL 31528482 (N.D. Tex. 2002.

[Quinion] "Garbage In Garbage Out," Michael Quinion, *World Wide Words* (Oct. 29, 2005) http://www.worldwidewords.org/qa/qa-gar1.htm; http://www.penguin.co.uk/nf/Author/AuthorPage/0,,0_1000065494,00.html.

[Radicati] "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).

[Radicati 2] "Messaging Total Cost of Ownership," by Sarah Radicati & Laura Venutura, The Radicati Group, Inc., p. 4 (1998) (costs not adjusted for inflation) (www.terracetech.com/jp/data/Messaging%20Total%20Cost%20of%20Ownership.pdf)

[Radicati 3] "Messaging Total Cost of Ownership -2003: in Enterprise and Service Provider Environments," The Radicati Group, Inc. (2003) www.sun.com/aboutsun/media/presskits/aiim2003/2003TCOSummary).

[Redman] "The Impact of Poor Data Quality on the Typical Enterprise," Redman, T., *CACM*, vol. 41, no. 2, February 1998, pp. 79-82).

[Roberts] "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).

[Ross] *Ross v. Bernhard*, 396 U.S. 531, 534-35 (1970). (http://caselaw.lp.findlaw.com/cgi-bin/getcase.pl?court=us&vol=396&invol=531).

[Saran] "Linux e-mail set-up slashes costs to £8 per user," Cliff Saran, Computer Weekly.com (May 6, 2003) (http://www.computerweekly.com/Articles/2003/05/06/194340/Linuxe-mailset-upslashescoststo%c2%a38peruser.htm).

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

[Seward] "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 (Feb. 2004) (http://www.e-evidence.info/seward1.pdf).

[Shetty] "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) .

[Skillcorn] "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).

[Slowinski] 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/).

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

[Snyder] "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).

[SOX] Sarbanes-Oxley Act of 2002, Pub. L. No. 107-204, 116 Stat. 745 (July 30, 2002) (http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=107_cong_public_laws&docid=f:publ204.107).

[Springer] 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).

[Stergiou] "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).

[Texas] "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* (http://www.tsl.state.tx.us/ld/pubs/wireless/chapter5.html).

[UNH] "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).

[White] "Top Enron Officials' Trial Begins Today," Ben White and Carrie Johnson, *The Washington Post* (Jan. 29, 2006) (http://www.washingtonpost.com/wp-dyn/content/article/2006/01/29/AR2006012900864.html).

[Wikipedia] "The Rating Process" section of the Wikipedia entry for "MPAA Film Rating System" (http://en.wikipedia.org/wiki/MPAA_film_rating_system).

[Winkleman] *Winkleman v Beloit Memorial Hospital,* 483 N.W.2d 211, (Wisconsin 1992).

[Wright] "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).

[Zaiane] "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/).