

Information Risk of Inadvertent Disclosure – An Analysis of File-Sharing Risk in the Financial Supply Chain

M. Eric Johnson¹
Center for Digital Strategies
Tuck School of Business
Dartmouth College, Hanover NH
M.Eric.Johnson@dartmouth.edu

Abstract: Firms face many different types of information security risk. Inadvertent disclosure of sensitive business information represents one of the largest classes of recent security breaches. We examine a specific instance of this problem – inadvertent disclosures through peer-to-peer file-sharing networks. We characterize the extent of the security risk for a group of large financial institutions using a direct analysis of leaked documents. We also characterize the threat of loss by examining search patterns in peer-to-peer networks. Our analysis demonstrates both a substantial threat and vulnerability for large financial firms. We find a statistically significant link between leakage and leak sources including the firm employment base and the number of retail accounts. We also find a link between firm visibility and threat activity. Finally, we find that firms with more leaks also experience increased threat.

Key Words: inadvertent disclosure, intellectual property leaks, data breaches, security, risk management, peer-to-peer networks, file sharing.

¹The author gratefully acknowledges helpful comments from seminar audiences at Emory University, University of Maryland, Carnegie Mellon (WEIS), HICSS 2008, Northwestern University, and Vanderbilt University. This work would not be possible without the assistance of Tiversa Inc., Scott Dynes, Christopher Graves, Daniel McGuire, and Nicholas Willey. Experiments described in this paper were conducted in collaboration with Tiversa who has developed a patent-pending technology that, in real-time, monitors global P2P file sharing networks.

This research was supported by award number 2003-TK-TX-0003 from the U.S. Department of Homeland Security, Science and Technology Directorate under the auspices of the Institute for Information Infrastructure Protection (I3P). Points of view in this document are those of the author(s) and do not necessarily represent the official position of the U.S. Department of Homeland Security, the Science and Technology Directorate, the I3P, or Dartmouth College.

Bio: M. Eric Johnson is Director of Tuck's Glassmeyer/McNamee Center for Digital Strategies and Professor of Operations Management at the Tuck School of Business, Dartmouth College. His teaching and research focuses on the impact of information technology on supply chain management. Through grants from the National Institute of Standards and Technology, Department of Justice, the Department of Homeland Security, and the National Science Foundation, he is studying how information security and trust effect supply chain relationships. He has testified before the US Congress on information security and collaboration and published many related articles in the *Financial Times*, *Sloan Management Review*, *Harvard Business Review*, and *CIO Magazine*. His research articles have appeared in such academic journals as *Production and Operations Management*, *Management Science*, *Operations Research*, *IEEE Security and Privacy*, *Communications of the ACM*, *IEEE Transactions on Engineering Management*, *Manufacturing and Service Operations Management*, and *Transportation Science*.

Before joining Tuck, Johnson taught at the Owen Graduate School of Management, Vanderbilt University. He was previously employed by Hewlett-Packard Co. and Systems Modeling Corp. He has consulted for diverse companies such as Sprint-Nextel, Lucent, Mattel, Nokia, Hewlett-Packard, Accenture, Pepsi, DHL, The Parthenon Group, Campbell-Hausfeld, Fleetguard, and Kulic&Soffa. Johnson recently served as president of the POMS College of Supply Chain Management and has served on numerous editorial boards including *Production and Operations Management*, *Management Science*, *Interfaces*, *Operations Research*, *International Journal of Logistics Management*, and *Manufacturing and Service Operations Management*. He holds a B.S. in Engineering, B.S. in Economics, an M.S. in Engineering and Operations Research from Penn State University, and a Ph.D. in Engineering from Stanford University.

Introduction

As firms become ever more dependent on information, new risks to that information arise from unexpected sources. Information security breaches have become a steady feature of the business press. With each new story, firms come under increased pressure to harden their networks and take a more aggressive security posture (Yue and Çakanyildirim 2007). However, it is often not clear what security initiatives offer firms the greatest improvement (Johnson and Goetz 2007). A close look at the headlines reveals a bewildering set of information breaches. While hackers regularly penetrate poorly secured networks (Sidel 2007) and devices (Bank 2005), many of the large recent security breaches were not technical break-ins, but rather inadvertent disclosures. For example, in the last few months alone, laptops at Towers Perrin, Boeing, Fidelity, and the U.S. Department of Veterans administration were lost or stolen – in each case inadvertently disclosing personal and business information (Francis 2007, Levitz and Hechinger 2006). Organizations have mistakenly posted on the web many different types of sensitive information, from legal to medical to financial (e.g., Twedt, 2007 or Kenworthy 2004). Even technology firms like Google and AOL have suffered the embarrassment of inadvertent web posting of sensitive information (Claburn 2007, Olson 2006) – in their cases, customer information. Still other firms have seen their internal information and intellectual property appear on blogs, YouTube, and MySpace (Totty 2007). In each case, the result was the same: sensitive information inadvertently leaked creating embarrassment, vulnerabilities, and financial losses for the firm, its investors, and customers. In this paper we examine a common, but widely misunderstood source of inadvertent disclosure: peer-to-peer file-sharing networks.

Despite significant efforts of the music industry, peer-to-peer (P2P) file sharing has become mainstream among large segments of the internet population. With estimates of 10 million simultaneous users (Mennecke 2006) sharing music, video, software, and photos, file-

sharing clients have joined the suite of standard PC applications for many users. Unrecognized to many of these users is the serious security threat participation in these networks poses to both corporate and individual security (Johnson 2007).

In our earlier research, we showed why P2P file sharing represents a growing security risk (Johnson et al. 2008). The evolution of these networks has done little but increase the risk. Efforts by internet service providers (ISPs), worried firms and organizations, and copyright holders to limit P2P both technically (e.g., site blocking, traffic filtering and content poisoning) and legally (e.g., RIAA prosecution of individual users and file-sharing firms) have prompted P2P developers to create decentralized, encrypted, anonymous networks that can find their way through corporate and residential firewalls. These networks are almost impossible to track, are designed to accommodate large numbers of clients, and are capable of transferring vast amounts of data.

Understanding security risk in management information systems is an important and rapidly evolving topic (Sun et al. 2007). In this paper, we analyze the information risk posed by file sharing. We show that confidential and potentially damaging documents have made their way onto these networks. We also show that attackers actively search P2P networks hoping to find information that they can exploit. First we describe the P2P security issues, establishing the vulnerabilities these software clients represent. Then we examine the vulnerability, threat and potential consequences through an analysis of documents we found circulating on these networks. Focusing on the top 30 US banks, we analyze a set of leaked documents collected throughout the supply chain including suppliers, customers, and the banks themselves. We also analyze user-issued search information on these same institutions, finding an astonishing number of searches targeted to uncover sensitive documents and data. For our sample of banks, we analyze tens of thousands of relevant searches and documents. We characterize the nature of these searches and files and the

underlying drivers of file leakage and movement. We find statistically significant links between leakage, firm employment base and the number of retail accounts. We also find a link between firm visibility and threat activity. More importantly, we find the firms experiencing greater leakage also experience increased threat. Finally, we discuss managerial implications and propose a simple benchmarking technique to compare leakage. Our analysis clearly reveals a significant information risk firms and individuals face from P2P file-sharing networks.

File Sharing in Peer-to-Peer Networks

File sharing on peer-to-peer networks enable users to publish and distribute any file from music to video to spreadsheets. Napster brought the concept of file sharing into the mainstream with its wildly popular music-sharing service. While only operating for two years before its court-ordered closure in 2001, Napster enabled tens of millions of users to share MP3-formatted song files. In its place many other file-sharing systems have emerged, driving an endless debate over the impact of music sharing (Oberhozer-Gee and Strumpf 2007) and a string of legal challenges by the music and video content industry (RIAA/MPAA). Yet none of these efforts seem to reduce file sharing. Rather the industry's legal and communication pressures have pushed users onto new clients and networks that pose new and more challenging security issues. In fact, some argue that Napster's success and failure simply spurred innovation, paving the way for many new P2P file-sharing networks such as Gnutella, FastTrack, e-donkey, and Bittorrent, with related software clients such as Limewire, KaZaA, Morpheus, eMule, and BearShare.

There have been many attempts to thwart file sharing. Firms, universities, and internet service providers (ISPs) block or throttle traffic associated with P2P systems using approaches such as port filtering. Client developers responded by using ports associated with

other services (web traffic, email traffic, etc.) to exchange data, blending file-sharing traffic with other data streams. Recent traffic studies suggest that P2P connections are now distributed across all ports (Karagiannis et al. 2003).

Today file sharing continues to grow, with usage doubling in the past three years, from less than four million in 2003 to nearly ten million simultaneous users in 2006 (Mennecke 2006). Many more files are shared in hard-to-track private networks, sometimes called dark networks (or darknets), accessed through invitations from other users. Even faster sharing growth is occurring in Bittorrent, which is one of the most popular applications for very large files like video. Users of these systems readily adapt and change to new networks based on legal pressure, features, and popularity. For example, the FastTrack network (used by KaZaA) has seen declines over the past three years while others like the Gnutella network, and the popular Limewire client, have grown. These rapid shifts suggest low barriers to entry for new sharing technologies, supported by a well-informed user base that is willing to explore new alternatives.

Inadvertent Disclosure in File Sharing

File-sharing clients typically allow users to share data in a particular folder and often direct users to move media files they wish to share into that folder. In normal operation, the client simply writes files to disk as it downloads them and reads files from disk as it uploads them. There are several routes for confidential data to get on to the network: a user accidentally shares folders containing the information; a user stores music and other data in the same folder that is shared; a user downloads malware that, when executed, exposes files; or the client software has bugs that result in unintentional sharing of file directories. Of course it is not necessary for a worm or virus to expose personal or sensitive documents because many users will unknowingly expose these documents for many reasons. For example, some users

mistakenly point to My Documents and end up sharing all of their files. In some cases the client interface design makes it difficult to see what is being shared. Additionally, P2P file-sharing systems often provide incentives for users to share files, such as faster downloads or broader searches. The clients typically come with wizards that are designed to find all media files and share the directories where media files are located. So a single MP3 file in My Documents can lead to sharing everything in My Documents. Moreover, the clients often share all subdirectories of a shared directory.

Many of these reasons point to the interface design (Good and Krekelberg 2003) and features of P2P clients that facilitate inadvertent sharing (Sydnor et al. 2006). In our earlier research, we illustrated the problem by uncovering a wide range of private personal information including passports, birth certificates, and tax returns. We also showed, through honey pot experiments, that there are significant threats from individuals actively seeking this information to commit theft (Johnson et al. 2008). In that paper, we argue that, while we believe that many information leaks are the result of accidentally shared data rather than the result of malicious outsiders, there are many other trends that are driving more security concerns.

- *Growing usage and network heterogeneity means more leaks* – With many networks and clients, users are not likely to grasp the security issues and P2P developers will likely not focus on security.
- *“Set and forget” increases losses* – P2P clients tend to be “set and forget” applications that run in the background while the user is not at the computer. This suggests that the user is not carefully tracking the activities of the P2P client, increasing the opportunity for abuse. Further, even benign file-sharing programs

consume significant processor time and network bandwidth, conditioning the P2P user to tolerate sluggish performance that, for others, might be a first sign that a system has been compromised.

- *No borders result in global losses* – Geography is largely irrelevant in P2P networks, meaning no particular country or region is safer than another. A computer logging on in Bombay or Brussels becomes part of the same network as a computer in Pittsburgh.
- *Malware* – While the overwhelming majority of traffic on P2P networks is entertainment content (games, movies, music, etc.), also lurking on P2P networks are files that pose severe security risks (Kalafut et al 2006, Shin et al 2006). Viruses that exist in email and other programs also have variants that exist in P2P networks (Ingram 2006).

Firms often mistakenly believe that they are immune from P2P disclosure problems because they protect the perimeter of their networks with firewalls and even use software to block corporate users from accessing files-sharing networks. However, even the best perimeter systems fail when corporate users connect to the web on public networks while traveling or at home. More importantly, sensitive corporate information is also held by customers, suppliers, contractors, and other business partners who also may be leaking documents. The nature of information flows within the extended enterprise significantly increases the challenge of preventing leaks.

Methodology and Data

To characterize the risk facing large financial institutions, their partners (suppliers, contractors), and their customers, we examined both the vulnerability and resulting consequences of leaked files and the threat posed by those searching to exploit the vulnerability. We focused our analysis on the supply chains of the Forbes top 30 U.S.-based banks (Forbes 2006). Those institutions collectively employ over a million people, manage over seven trillion dollars, and comprise a wide range of sizes as show in Table 1.

Table 1. Summary statistics on institutions in data set ($N=30$).

	Employees	Number of Branches	Sales (\$bil)	Assets (\$bil)	Market Value (\$bil)
Average	47,406	1,567	\$ 17.94	\$ 248.42	\$ 40.34
<i>Standard Deviation</i>	<i>68,020</i>	<i>1,919</i>	<i>28.84</i>	<i>395.25</i>	<i>56.95</i>
Max	307,000	7,237	\$ 120.3	\$ 1,494.0	\$ 230.9
Min	2,202	41	\$ 1.3	\$ 26.3	\$ 4.5

Primary Data Sources: Forbes and Hoovers.

With the help of Tiversa Inc.², we gathered and categorized P2P searches and shared files related to these institutions over a 7-week period (December 27-February 13, 2006). Tiversa's servers and software allowed us to monitor and to participate in the three most popular networks (each of which supports the most popular clients) including Gnutella (e.g., Limewire, BearShare), FastTrack (e.g., KaZaA, Grokster), and e-donkey (e.g., eMule, EDonkey2K). Given the nature of P2P networks, it is difficult to make statements regarding the exact population size in aggregate or at any particular moment or our ability to observe some fraction of the population at any moment. As mentioned earlier, recent estimates place the P2P population at nearly 10 million simultaneous users (Mennecke 2006). The networks themselves are dynamic, with members constantly joining (and sharing files) and leaving. Thus, over a period of a day some estimate as many as 20 million users issue upwards of 800

² Tiversa Inc. monitors global P2P file sharing networks.

million searches. Using Tiversa's systems, we participated in those networks globally and collected a very large sample of this activity.

To gather relevant searches and files, we developed a *digital footprint* for each financial institution. A digital footprint comprises terms that would quickly lead you back to the host firm or important trading partners (suppliers, contractors, vendors). These terms, if Googled, would often (but not always) lead you directly back to the host firms. For example, for a firm like Hewlett-Packard they would include:

- Firm names, abbreviations, nicknames, ticker symbol (e.g., Hewlett-Packard, Hewlett, HP, HPQ); If the organization is the merger of two or more companies, each one could be active (Compaq);
- Key brands and subbrands (e.g., Compaq, Inkjet, Pavilion...);
- Subsidiaries, divisional names (e.g., HP Shopping, Home Products Division);
- Supplier, contractors, vendors(e.g., Celestica, Accenture);

Searches or files containing any one or combination of these terms were captured. Of course, increasing the number of terms included in the digital footprint increases the number of search and file matches found, but also increases false positives – searches and files captured that have nothing to do with the institution in question. In practice, we developed a footprint and then tuned it to eliminate terms that seemed less useful and added ones that were. Our goal was to cast a large initial net with 20-30 terms and then further refine the footprint to eliminate unrelated items, reducing the collected searches and files that must be manually analyzed.

P2P User-Issued Searches – the threat

Using this approach, we collected over 437,800 searches issued by P2P users looking for terms that matched our digital footprints including 41,700 unique strings. Those searches were evaluated and reduced to nearly 16,000 searches with good fit for the banking institutions. The resulting searches were then manually analyzed to assess their intent. Our goal was to categorize the searches by a measure of their threat. After studying thousands of searches, we developed a three-point threat scale: High (3), Medium (2) and Low (1). While a five- or seven-point scale would allow for greater discrimination, in practice, we found we could not further distinguish between the searches. Thus we concluded a more detailed scale would increase the scale’s variance through the induction of random noise rather than a systematic variance attributable to the underlying threat phenomenon (DeVellies 2003). As shown in Table 2, those categorized as high threat (i.e., 3) were searches directed for specific documents or data that could fuel malicious activity. Medium threat searches were ones targeted generically against the firm. Such searches would uncover sensitive files along with music, video, etc. Low threat searches were ones searching for music, picture, or video files related to the bank’s footprint. While these searches could be seen as benign, they would also uncover sensitive files and thus the expose vulnerabilities that could still represent a threat to the institution and its customers.

Table 2. Three-point search threat scale with example.

Threat Level	Search Group Type	Example Search
High – 3	Fraud / ID Theft Intent	“Citibank August Statement”
	Internal File Search	“Citibank Hotel RFP.doc”
Medium- 2	Company Search	“Citibank”
Low -1	Public File or Media Search	“Citibank Commercial”
	Partial Match Term	“Jimmy Buffet Wachovia”

Table 3 shows examples of searches we observed in each of the three categories. Directed searches for databases, account user information, passwords, routing, and pin numbers represent clear threats.

Table 3. Examples of searches observed in each category.

High - 3	Medium - 2	Low - 1
bank pnc checking account for	bank of new york	wachovia center
wachovia bank online user id	regions bank	state street cutie
clientauthorization wachovia	union planters	deep in the music suntrust
wells fargo.*pdf	first horizon	a day in the life pnc
suntrust letter	m&t bank	wells fargo music man
citi bank balance transfer	huntington bank	first national city band march
bank of america database	wachovia bank	bank of america tower
washington mutual statement	golden west	Girls Of The Golden West
GlobalStrategy-Citigroup.pdf	sovereignbank	paul mccartney tour wachovia
us bank check register to end	banco popular	new orleans rap pnc hotboy
mellonbank creditreport	national bank of america	chase away morgan
pin bank of america	amsouth	the chase fleetwood mac

Medium searches, like those for bank names, are more generic. Low threat searches like “bank of america tower” or “wells fargo music man” may seem innocent, but keep in mind that these are searches on P2P file-sharing networks, not Google. Each of these searches would uncover other bank-related files.

For many firms, coincidental association with a popular song or brand represents another problem we call “digital wind.” Millions of searches for that song increase the likelihood of exposing a sensitive bank document. Either by mistake or by curiosity, when these documents are exposed, they are sometimes downloaded to other clients, thus spreading the file and making it more likely to fall into the hands of someone who will try to exploit its information. For example, the popular song “Citizen Cope” creates digital wind for the Citizens Bank (see other examples in Table 4).

Table 4. Examples of digital wind

Institution Effected	Digital Wind
Citizens Bank	Citizen Cope (song)
Fifth Third	HP printer driver (for the model 5300)
Golden West (Wachovia)	Songs with “golden” or “west” in the title
Keycorp	People looking for key generators for various program
National City	The National (music group) with City Middle (a song)
PNC	Music rappers (PNC and P-Money)
State Street	“State Street Residential” (song by Death Cab for Cutie)

Inadvertently Disclosed Files – the vulnerability

During this same period, we also collected files we observed being shared on the networks. We focused on business-related files – particularly those from the Microsoft Office Suite (including file extensions doc, xls, ppt, mdb, along with rft, pdf, txt). Using the digital footprint, file names with any related terms were captured. In some P2P networks, files are also indexed by their associated metadata (like the name of the firm to which a word processor is registered). Thus we captured those documents as well. Using this approach, we collected over 114,000 files totaling more than 15GB of data over the 7-week period. Tiversa’s systems allowed us to limit the files harvested to unique IP addresses, thus reducing the number of duplicate files collected.

With the vast sample of files, we conducted a cross-sectional analysis of files for all banks found in a single week – therefore reducing our data set from all files found over all 7-weeks to those found in the last week of our collection. Files were manually evaluated on multiple dimensions (Shye 1985). For each file examined, we noted if the file was flagged to reduce distribution – for example if it was marked, “Confidential,” “Restricted,” “Internal Use,” etc. We recorded the file’s age by either examining both the file’s metadata (e.g., creation and

editing dates) and dates inside the document itself. We also assessed the source of the leak (customers, suppliers, internal) by examining IP addresses and clues within the document. After examining the document, we classified the document based on its type and on a four-point scale of its sensitivity as reflected by the potential consequence if exploited (method further described in Appendices A and B). Like the search classification scheme, the scale included a High (3), Medium (2), and Low (1) along with the addition of (0) for public documents. Public documents are ones that the firm would want widely distributed (although they may be surprised to know these documents are circulating in music-sharing networks). Keep in mind that while leaking a low sensitivity document (like a 0) may seem harmless, if that document is leaked from a source with access to other more sensitive documents, it is likely a matter of time before that source leaks a more damaging document. This outcome is analogous to the safety literature (Hammer and Price 2000), which has observed that small accidents often precede much larger ones.

Results

We overview some of the key observations from this extensive data set of searches and disclosed files.

Searches – the threat

A graphic summary of the 15,989 searches with good fit for the banking institutions is shown in Figure 1. To protect specific institutions, we have not included bank names and bank numbers shown in the figure are randomly assigned (they do not represent the Forbes ranking number). As might be expected, there is wide dispersion of search interest in the banks. From an initial examination of the data, we observed that the largest firms with strong global

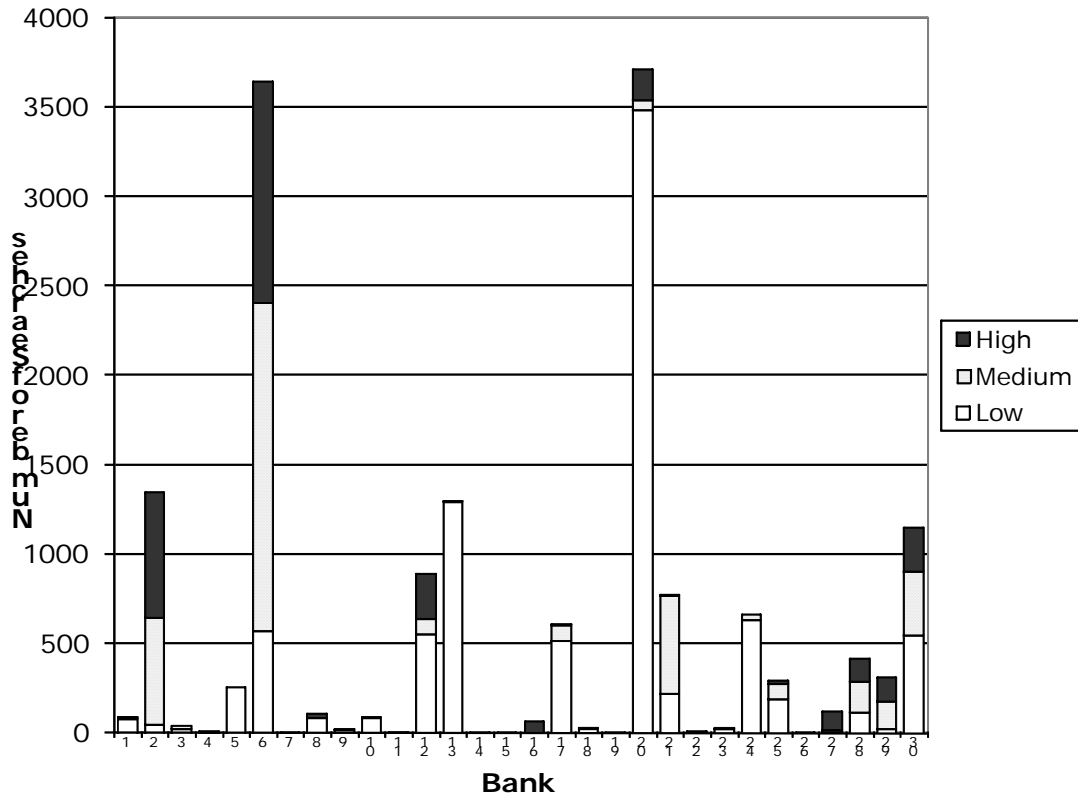


Figure 1. Search threat categorization for top 30 financial institutions over 7-week time period (sequence does not correlate with rank).

brands seemed to experience the most search activity. We hypothesize that firm visibility is a key driver of search activity. Formally, we propose:

Hypothesis 1a (Firm Visibility Increases Threat). *Firm visibility increases the threat of discovery and exploitation of inadvertent disclosures.*

Of course, marketing theory would link brand strength to consumer awareness (Keller 2003). Firms 2 and 6 represent banks in this category and experienced a large number of highly threatening searches. Bank 20 represents the case of bank experiencing significant digital wind. That bank doesn't have a well-known global brand, but its name and associated products have names that unfortunately share common elements with a popular music group. Many of the smaller banks experienced far less search activity, either by luck (less digital

wind) or by obscurity. Yet, as can be seen in the figure, many of those small institutions still experienced targeted searches. Figure 1 clearly demonstrates the threat faced by these institutions.

To test our hypothesis that search activity is correlated to bank brand visibility, we performed a least squares regression on a linear model of searches (Y). Brand strength in marketing (Keller 2003) is often measured on positive brand attributes (e.g., quality, value, trustworthiness, reliability). However, we were more interested in the notoriety of the brand, which is not limited to positive elements. So, as a simple measure of brand visibility, we chose the number of firm employees (X).

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Banks with a large employment base typically have a large retail customer base (rather than business customers) and many visible branch offices that are open to the public. We note that this is simply one of many possible surrogates for visibility – others include the number of retail accounts, total assets, or the number of bank branches. We argue that number of employees is a good measure of the visibility of the bank – better than revenues or assets since those may be driven by large business customers who provide less public visibility for the bank. Likewise the number of locations might not capture the impact of urban and rural markets.

Since low-threat searches (1) were driven by other phenomenon unrelated to the bank, such as popularity of song that coincidentally shared digital footprint elements, we limited searches (Y) to include medium and high threat searches (which accounted for 7,194 searches).

Table 5 shows that the visibility of the bank explains much of the variation in P2P search activity. This parsimonious model explains nearly 80% of the variation of search activity between banks. A regression limiting Y to high threat searches (3) yielded even stronger support (R square of .86 with significant coefficients at .01).

Table 5. Support for the relationship between brand visibility (measured by employees) and searches (7,194 searches)

<i>Regression Statistics</i>						
Multiple R	0.89					
R Square	0.80					
Adjusted R Square	0.79					
Standard Error	276.01					
Observations	30					
<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	8330109.8	8330109.8	109.3	3.56E-11	
Residual	28	2133029.0	76179.6			
Total	29	10463138.8				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
β_0	-133.73	61.77	-2.17	0.04	-260.26	-7.21
β_1	0.00788	0.00075	10.45697	0.00000	0.00634	0.00942

Inadvertently Disclosed Files – the vulnerability

With a massive collection of documents, we conducted a cross-sectional analysis (files for all banks found in a single week). We chose to focus on the last week of collection. This week included 12,706 documents that required largely manual analysis to determine bank relevance and sensitivity. We chose this approach based on our hypothesis that documents found very early in the collection process would likely include many public ones available on many clients or ones that had been circulating for some period while ones found later would more likely represent recent leaks. Keep in mind the nature of P2P networks where some users are constantly sharing files while others periodically join the network as they 1) turn on their computers; 2) launch a P2P client to find music or other files; or 3) download a P2P client

and begin sharing files as a new network member. We hypothesized that our collected documents would thus experience an initial transient phenomenon often seen in simulation analysis of complex systems (Law and Kelton 2000).

In the end, we found limited support for this hypothesis from the data. Given the vast sea of files floating in the P2P and the transient nature of users, the file discoveries (particularly of relevant, unique files) varied significantly from day to day. While our daily finds fluctuated based on many factors, we did not observe a noticeable drop-off in the number of files from week to week nor did we find a statistically significant difference in document age of those found early or later in our collection.

Our last week contained 12,706 documents, many of which were not related to any of the banks in question. After hundreds of hours of manual analysis, we categorized 2,432 documents as relevant to the banks of which 1,708 were unique (30% were duplicate). Duplicate documents are themselves interesting as they show the spread of certain files. Given the nature of P2P networks, duplicates increase the likelihood of threatening searches successfully finding a document. An analysis of unique document source indicated a breakdown as shown in Figure 2. The source was determined by an analysis of content of file itself, its metadata, and the disclosing IP address, categorizing them into three groups: individual not involved in the banking operation (customers), another company working with or for the bank (suppliers), or by someone within the bank (internal). As one would expect, the majority of documents came from the most numerous demographic, customers. Customer computers often double as both office and entertainment machines and many have multiple users, therefore users may be unaware of what someone else in the household has stored on the computer. Similarly, the documents originating from suppliers were often from smaller firms and contractors whose computers would likely be used for both home and business purposes. These were often painters, landscapers, electricians, and building contractors but

also included consultants, IT suppliers, processors, etc. However, we also found documents from major professional service providers such as auditors and consultants. Internal documents were about as numerous as documents coming from suppliers. Many of these seemed to come from individuals more likely to work in the field than in an office environment.

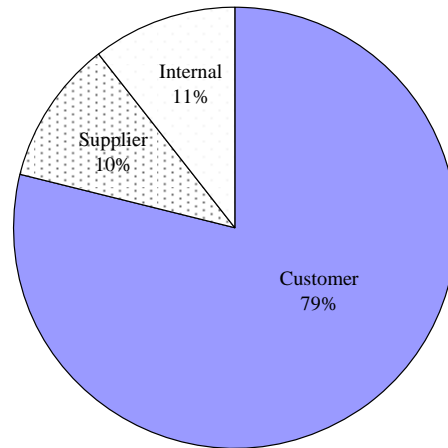


Figure 2. Document source.

We found files of nearly every type, but Personally Identifiable Information (PII) documents were the most numerous, accounting for 49% of all unique documents. Many of these documents contained enough information to easily commit fraud or identity theft (see appendix A for group definitions). The next largest category was the other category including bank addresses, charity requests, instructions, articles, fax coversheets, and blank (public) forms. Business Operations documents included employee training manuals, internal policies and procedure, and work plans. Many others originated from suppliers in regards to work that had been or would be completed for the bank (invoices, proposals, and estimates). Also numerous in this category were various internal forms (both complete and incomplete). The human resources category was also well represented with employee résumés, job descriptions, employee performance reviews, and employee lists. Along with many public

and low sensitivity documents we found some (apparently) sensitive documents including IT documentation, auditing evaluations conducted by third parties, and many sensitive customer documents. For one bank, we found a spreadsheet with 23,000 business accounts including their contact names and addresses, account numbers, company positions, and relationship managers at the bank. Clearly such a data trove would be very useful for a competing bank not to mention the fraud potential. Ironically, for one bank, we found a detailed manual of their security review process!

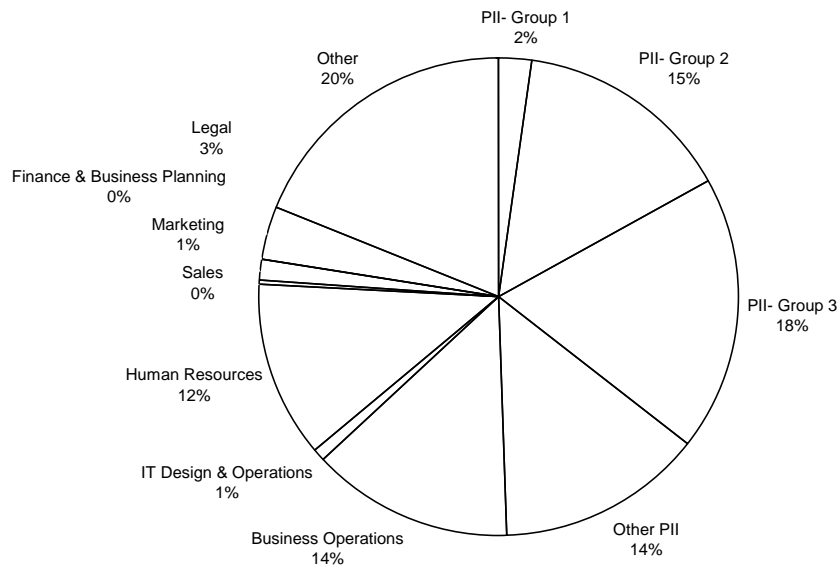


Figure 3. Document type (among all 1,708 unique, relevant documents).

A graphic summary of the sensitivity of the 1,708 unique, relevant documents is shown in Figure 4. Again, to protect specific institutions, we have not included bank names and bank numbers shown in the figure are randomly assigned (they do not represent the Forbes ranking number). Like searches, there was wide dispersion of document disclosures among banks. The largest firms again seemed to have the most documents. We hypothesize that the number of leaked documents is linked to number of leak sources.

Hypothesis 2 (Leak Sources Drive Vulnerability). *Firm leak sources increase the vulnerability of inadvertent disclosure.*

In this case, we argue that number of employees is directly related to leak sources (internal) and that firms with a large employment base also have many customers and suppliers (each representing classes of leak sources). Thus we tested the link to bank size as represented by the number of employees, using a least squares linear model of documents (Y). We ignored all public documents, limiting files (Y) to include low, medium and high sensitive documents (which represented 1,412 files).

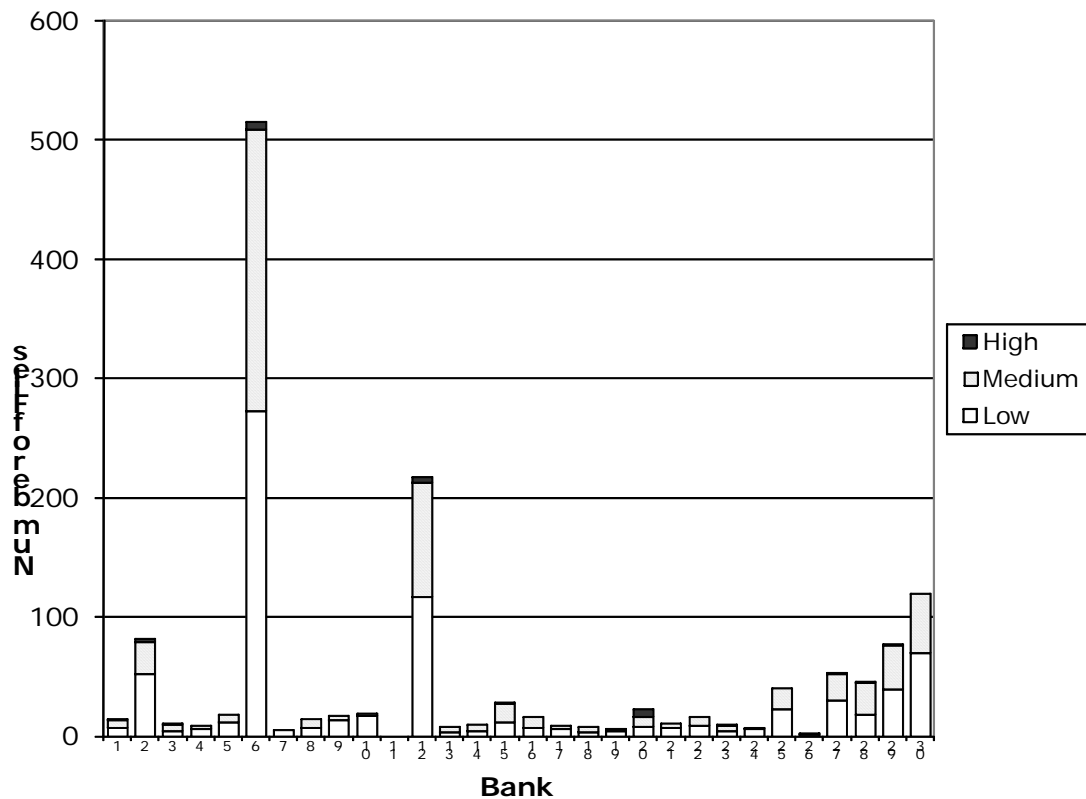


Figure 4. File disclosure categorization (risk rated as high, medium, and low) for top 30 institutions (sequence does not correlate with bank rank, see appendix for rating details).

Table 6 shows that the employment base of the bank explains much of the variation between banks in the number of sensitive files found. Again, this parsimonious model

explains nearly 84% of the variation of document activity between banks. A regression limiting Y to medium plus high sensitive files (levels 2 and 3) yielded a similar result (R square of .81 with significant coefficients). Of course, this model could be further instrumented to account for other factors such as the number of retail accounts, number of suppliers, on-line retail activity, digital practices of the banks, outsourcing activity, international presence, etc.

Table 6. Support for the relationship between employment base and files found (1,412 files).

<i>Regression Statistics</i>						
Multiple R	0.92					
R Square	0.84					
Adjusted R Square	0.83					
Standard Error	40.54					
Observations	30.00					
<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	238490.3	238490.3	145.1	1.35828E-12	
Residual	28	46011.6	1643.3			
Total	29	284501.9				
	<i>Coefficient</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
β_0	-16.14	9.07	-1.78	0.09	-34.72	2.45
β_1	0.0013	0.0001	12.0470	0.0000	0.0011	0.0016

Of these, certainly the number of retail accounts is the most interesting addition. Given the number of leaks we found flowing from customers, it is likely that the number of retail customers is a significant factor. Suppliers, on the other hand, had few leaks and are much more difficult to characterize. Given these observations, we further instrumented the model to include customers. We performed a least squares multiple regression on a linear model of searches (Y) where X_1 is the number of firm employees and X_2 is the number of retail accounts.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Each year in June, the FDIC collects and published data on retail accounts which is defined as the number of deposit accounts of \$100,000 or less (FDIC 2008). Using that data, we see in Table 7 that the overall model provides strong fit with statistically significant coefficients. However, surprisingly, the coefficient on number of accounts is negative! First, note that one might expect multicollinearity between employees and accounts – banks with many accounts will likely have many employees. Using an auxiliary regression (Judge et al 1985) between accounts and employees we detect the presence of some multicollinearity ($R^2 = .56$), but not a near-exact linear dependence. Of course, regressing just accounts against files, does produce a positive coefficient (significant) but with low $R^2 = .19$. Reflecting on the negative coefficient in the multi regression, one might argue that employment base best captures the size difference between banks, but efficient banks with larger retail customer bases (or accounts per employee) do better than those with lower retail account bases. This could be driven by the nature of bank focus – those with a very strong retail focus may be taking better steps to educate and protect customers from leakage.

Table 7. Support for the relationship between employment base and number of retail accounts and files found (1,412 files).

<i>Regression Statistics</i>						
Multiple R	0.99					
R Square	0.97					
Adjusted R Square	0.97					
Standard Error	17.06					
Observations	30					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	2	276643.0	138321.5	475.2	9.06E-22	
Residual	27	7858.9	291.1			
Total	29	284501.9				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
β_0	-8.15	3.88	-2.10	0.05	-16.11	-0.19
β_1	-4.75E-06	4.15E-07	-1.14E+01	7.22E-12	-5.60E-06	-3.90E-06
β_2	0.00194	0.00007	27.54837	0.00000	0.00179	0.00208

The link between threat and vulnerability

With a picture of the vulnerability, we return to further examine the threat. While bank visibility is certainly linked to the threat, there are other factors that security professionals point towards. One of the most interesting is the role of visible vulnerability. Past security failures (highlighted in the media) and the existence of visible vulnerabilities often increases the criminal activity and threat. We argue that in the case of inadvertent disclosure, the existence and magnitude of leaks may very well drive search activity. Individuals who have successfully found leaked documents are encouraged to increase their search activity.

Hypothesis 1b (Firm Visibility and Leak Propensity Increase Threat). *Firm visibility and leak propensity increases the threat of discovery and exploitation of inadvertent disclosures.*

To test this modification of hypothesis 1, we performed a multiple regression of searches (Y) where X_1 is the number of firm employees and X_2 is the number of sensitive files found (low, medium and high).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Table 8 shows support for this hypothesis with good model fit and statistically significant (at the .05 and .01 level respectively) coefficients. Again, one might be concerned with multicollinearity between employee and files, since our earlier result shows a significant positive relationship. We note that a model with searches alone also produces strong fit ($R^2 = .82$) with a significant, positive coefficient.

Table 8. Support for the relationship between brand visibility (measured by employees) and vulnerability (measured by leaked documents) and searches.

<i>Regression Statistics</i>						
Multiple R		0.92				
R Square		0.84				
Adjusted R Square		0.83				
Standard Error		246.03				
Observations		30				
<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	2	8828752.6	4414376.3	72.9	1.3E-11	
Residual	27	1634386.2	60532.8			
Total	29	10463138.8				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
β_0	-80.61	58.09	-1.39	0.18	-199.80	38.58
β_1	0.003	0.002	2.090	0.046	6.342E-05	0.007
β_2	3.292	1.147	2.870	0.008	0.939	5.645

Conclusions and Managerial Implications

Inadvertent disclosure of sensitive business information represents a major information risk facing firms. The popularity of many Web 2.0 applications, including collaboration tools and peer-to-peer (P2P) file-sharing networks, have created many new security risks for organizations. In this paper, we have illustrated the threat and vulnerability of firms to leaks in P2P networks, characterizing the extent of the problem for large financial institutions. We found that both the vulnerability and threat are well explained by institution visibility and the number of leak sources. We also found that banks leaking information experience greater search threat. Thus reducing the leaks not only reduces the vulnerability but may also reduce the threat activity of those looking to exploit the leaks.

Faced with this P2P threat and vulnerability, executives can take many actions to improve their information security. While brand strength and recognition are certainly desirable attributes, firms should consider branding in light of the digital wind created by other media. Such considerations would also be helpful in making their brands more likely to stand out in traditional internet searches via Google or Yahoo. Firms could also introduce file naming

conventions and policies to reduce the metadata footprint of their documents. These types of initiatives reduce the threat of documents being found and spread.

On the other hand, many other initiatives can be taken to reduce the leaks. Key among them is employee, contractor, supplier, and customer education on the dangers of P2P file sharing. One of the security challenges many organizations face, is developing effective strategies to help individuals in the extended enterprise make better information-risk decisions (Goetz and Johnson 2007). For web applications like file sharing, the benefits to the individual sometimes outweigh the perceived risks since users don't always bear the cost of a security failure. Technical steps to block P2P participation on firm equipment help address this issue along with policies for home machine use and supplier security qualification.

Periodic P2P monitoring and benchmarking is also useful in gauging progress and comparing firm performance with peers. Based on our statistical analysis, we propose that firms measure document leaks in terms of documents per employee per unit time (holding the document search and collection effort constant). Such a measure provides a useful benchmarking tool for security executives. As shown in Figure 5, summarizing file disclosures this way provides a very different picture of bank security performance. In our case, over the week we analyzed, firms with less than 0.5 documents per 1,000 employees appear to be the leaders. Of course, document sensitivity must be likewise considered. Moreover, it is important to realize that even a single high-sensitivity document can be very damaging.

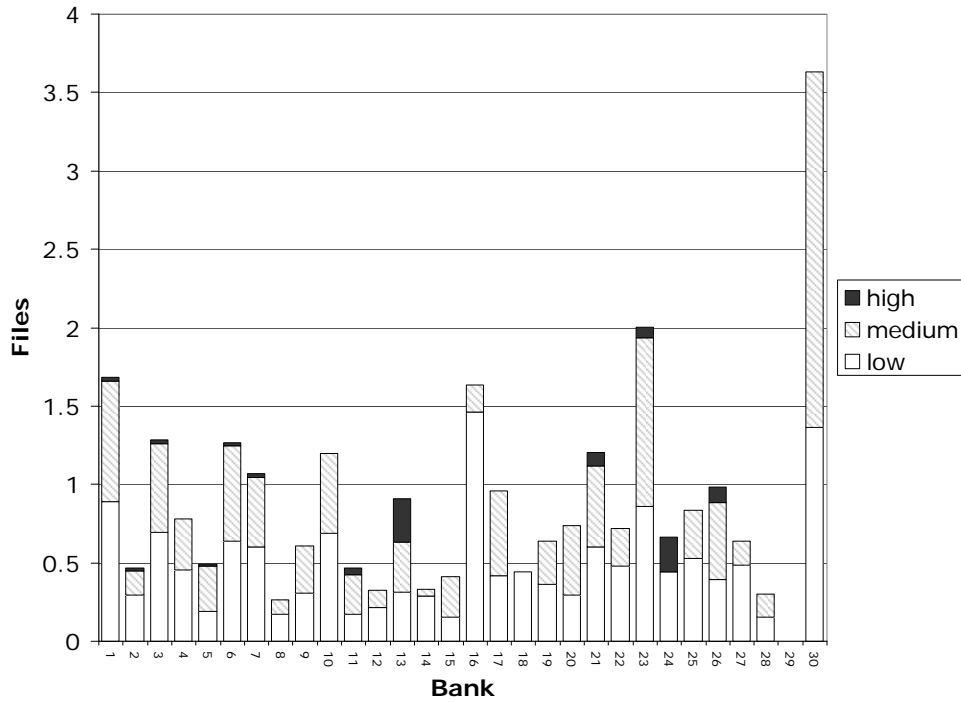


Figure 5. File disclosure categorization per 1000 employees for top 30 institutions (banks reordered to disguise identity).

We see many of the current P2P trends further increasing the problem. In ongoing work, we are further analyzing the data we gathered to provide managers and developers with clues on how to best control these inadvertent disclosures.

References

Bank, D. (2005), "Stores Blame Checkout Software For Security Breaches,"

Wall Street Journal, April 27.

Claburn, T. (2007), "Minor Google Security Lapse Obscures Ongoing Online Data Risk,"

Information Week, January 22.

DeVellis, R. F. (2003), **Scale Development: Theory and Applications**, Second Edition,

Sage Publications, London.

FDIC (2008), Statistics on Depository Institutions. <http://www2.fdic.gov/sdi/index.asp>

DeCarlo, S (2006), "The World's Largest Public Companies: Forbes Global 2000,"

Forbes, 03.31.06,

http://www.forbes.com/2006/03/30/largest-public-companies_06f2k_cz_sk_0331forbes2000intro.html.

Francis, T. (2007), "Towers Perrin Laptops, Client Data Stolen," *Wall Street Journal*, January

9. B2.

Gerber, A. J. Houle, H. Nguyen, M. Roughan, and S. Sen. (2003), "The Gorilla in the Cable,"

in National Cable & Telecommunications Association (NCTA) 2003 National Show,

Chicago, IL, June 8-11, 2003.

Goetz, E. and M. E. Johnson (2007), "Security through Information Risk

Management," *I3P Technical Report*.

<http://mba.tuck.dartmouth.edu/digital/Programs/CorporateEvents/CISO2007/Overview.pdf>

Good, N.S. and A. Krekelberg (2003), "Usability and privacy: a study of Kazaa P2P file-sharing," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Ft. Lauderdale, Florida, April 05-10. ACM Press, 137 – 144.

Ingram, M. (2006), "66,000 Names and Personal Details Leak on P2P," April 29, <http://www.slyck.com/news.php?story=1169>

Hammer, W. and D Price (2000), **Occupational Safety Management and Engineering** (5th Edition), Prentice-Hall International Series in Industrial and Systems Engineering, NY, NY.

Johnson, M.E. and E. Goetz (2007), "Embedding Information Security Risk Management into the Extended Enterprise," *IEEE Security and Privacy*, May-June, 24-32.

Johnson, M. E. (2007), "Inadvertent File Sharing Over Peer-to-Peer Networks," Testimony before the Committee on Oversight and Government Reform, United States House of Representatives, July 24.

Johnson, M.E., D. McGuire, N. D. Willey (2008), "Why File-Sharing Networks Are Dangerous." forthcoming in *Communications of the ACM*.

Judge, G.G., R.C. Hill, W.E. Griffiths, H. Lutkepohl, and T. Lee (1982), **Introduction to the Theory and Practice of Econometrics**, Wiley, NY, NY.

Kalafut, A., A. Acharya, M. Gupta (2006), “A Study of Malware in Peer-to-peer Networks,” *Proceedings of the Internet Measurement Conference*, ACM 2006, 327 – 332.

Karagiannis, T, A. Broido, N. Brownlee, K. Claffy, M. Faloutsos (2003) “File sharing in the Internet: A characterization of P2P traffic in the backbone,” Technical Report, UC Riverside.

Karagiannis, T. A. Broido, M. Faloutsos, and K. Claffy (2004), “Transport Layer Identification of P2P Traffic” *Proceedings of the 4th ACM SIGCOMM conference on Internet Measurement*. Taormina, Sicily, Italy, 121-134.

Keller, K. L. (2003), *Strategic Brand Management*, 2nd edition, Upper Saddle River, NJ: Prentice-Hall.

Kenworthy, T. (2004) “Bryant's accuser files civil suit,” *USA Today*, August 10.

Law, A.M. and W.D. Kelton (2000), **Simulation Modeling and Analysis**, Third Edition, McGraw Hill, New York, NY.

Levitz, J. and J. Hechinger (2006), “Laptops Prove Weakest Link in Data Security,” *Wall Street Journal*, March 26, B1.

Mennecke, T. (2006), “Slyck News – P2P Population Continues Climb” June 14,

<http://www.slyck.com/news.php?story=1220> .

Oberhozer-Gee, F. and K. Strumpf (2007), “The Effect of File Sharing on Record Sales: An Empirical Analysis,” *Journal of Political Economy*, Vol. 115, No. 1, 1-42.

Olson, P. (2006), "AOL Shoots Itself in the Foot," *Forbes*, August 8.

Pew Internet Activities and Trends Report – June 05. Survey Question: "Ever share files from your own computer such as music, video, or picture files, or computer games with others online?" Pew (2003), Pew Internet Project Data Memo, http://www.pewinternet.org/pdfs/PIP_Copyright_Memo.pdf, July.

Sidel, R. (2007), "Giant Retailer Reveals Customer Data Breach," *Wall Street Journal*, January 18, D1.

Shin, S.J. Jung, H. Balakrishnan (2006), "Malware Prevalence in the KaZaA File-Sharing Network," *Proceedings of the Internet Measurement Conference*, ACM 2006, 333 – 338.

Shye, S. (1985), **Multiple Scaling: The Theory and Application of Partial Order Scalogram Analysis**, North-Holland, Amsterdam.

Sun, L., Srivastava, R. P. and M. Theodore J (2006), "An Information Systems Security Risk Assessment Model Under the Dempster--Shafer Theory of Belief Functions," *Journal of Management Information Systems*, Vol. 22 No. 4, 109 – 142.

Sydnor II, T. D., J Knight, and L.A. Hollaar (2006), "File Sharing Programs and Technological Features to Induce Users to Share," A Report to the United States Patent and Trademark Office from the Office of International Relations, November.

Symantec (2006), "W32.Antinny.Q," http://www.symantec.com/security_response/writeup.jsp?docid=2004-053016-5101-99&tabid=2

Totty, M. (2007) "Security: How to Protect Your Private Information," *Wall Street Journal*,
January 29. R1.

Twedt, S. (2007), "UPMC patients' personal data left on Web," *Pittsburgh Post-Gazette*,
April 12.

Yue, W. T. and M. Çakanyildirim (2007), "Intrusion Prevention in Information Systems:
Reactive and Proactive Responses," *Journal of Management Information Systems*
Vol. 24 No. 1, 329 – 353.

Appendix A – Disclosure Classification Scheme (Document Type)

	Major Category	Definition	File Categories
A	Personally Identifiable Information (PII) – <u>Group 1</u>	Are files that contain information that can uniquely identify a person to enable fraud or identity theft? Files contain at least three of: <ul style="list-style-type: none"> • SS# • Credit Card Number • User Account Number • User ID and password • Full Address • Signature 	<ol style="list-style-type: none"> 1. Dispute letters 2. Wire Transfer Authorizations 3. Credit Reporting Agency Records (e.g. Equifax) 4. User ID / Password List Records & Account Records 5. Tax Returns 6. Customer Service Correspondence 7. Account Closure 8. Statements/Payment Receipts 9. Other
B	Personally Identifiable Information (PII) – <u>Group 2</u>	Are files that contain information that can uniquely identify a person to enable fraud or identity theft? Files contain at least two of: <ul style="list-style-type: none"> • SS# • Credit Card Number • User Account Number • User ID and password • Full Address • Signature 	<ol style="list-style-type: none"> 1. Dispute letters 2. Wire Transfer Authorizations 3. Credit Reporting Agency Records (e.g. Equifax) 4. User ID / Password List Records & Account Records 5. Tax Returns 6. Customer Service Correspondence 7. Account Closure 8. Statements/Payment Receipts 9. Other
C	Personally Identifiable Information (PII) – <u>Group 3</u>	Are files that contain information that can uniquely identify a person to enable fraud or identity theft? Files contain at least one of: <ul style="list-style-type: none"> • SS# • Credit Card Number • User Account Number • User ID and password • Full Address • Signature 	<ol style="list-style-type: none"> 1. Dispute letters 2. Wire Transfer Authorizations 3. Credit Reporting Agency Records (e.g. Equifax) 4. User ID / Password List Records & Account Records 5. Tax Returns 6. Customer Service Correspondence 7. Account Closure 8. Statements/Payment Receipts 9. Other
D	Other PII	PII that does not meet the criteria in A,B, or C	<ol style="list-style-type: none"> 1. Dispute letters 2. Wire Transfer Authorizations 3. Credit Reporting Agency Records (e.g. Equifax)

			<ol style="list-style-type: none"> 4. User ID / Password List Records & Account Records 5. Tax Returns 6. Customer Service Correspondence 7. Account Closure 8. Statements./Payment Receipts 9. Other
E	Business Operations		<ol style="list-style-type: none"> 1. Internally released PII 2. Internal Organizational Phone / e-mail lists 3. Customer lists 4. Employee Training Materials 5. Internal Policies & Procedures 6. Supplier proposals 7. Project work plans (non-IT) 8. Supplier Portal Access Records 9. Purchase Orders 10. Invoices 11. Completed Internal Forms 12. Internal Forms 13. Charitable activities records 14. Mortgage appraisals 15. Supplier correspondence 16. Supplier / Contractor / Consultant work product or deliverable 17. Other
F	IT Design & Operations		<ol style="list-style-type: none"> 1. Network & Systems operations documents 2. Disaster Recovery Plans 3. Network design 4. Organizational access codes 5. Functional / Software Specifications 6. IT project work plans 7. Acceptable use policies 8. Internal IT roadmaps 9. Other
G	HR		<ol style="list-style-type: none"> 1. Employee pay or bonus records 2. Existing employee reviews & performance appraisals 3. Employee medical records 4. New hire candidate interview records (Hire / Pass) 5. Promotion / Termination records 6. Resumes/Cover Letters 7. Resignation letters 8. Job Descriptions 9. Employee lists 10. Individual employee benefits records 11. Other

H	Sales		<ol style="list-style-type: none"> 1. Sales group (region, product line, etc.) projections 2. Sales presentations 3. Territory or Account Plans 4. Target prospect lists 5. Competitive analysis 6. Client proposals 7. Price quotes 8. Internal price and discount lists 9. Other
I	Marketing		<ol style="list-style-type: none"> 1. Current press-releases in mark-up 2. Past press releases 3. Focus group study results 4. PR plans 5. Other
J	Finance & Business Planning		<ol style="list-style-type: none"> 1. Revenue projections / corporate level sales projections 2. Business plans 3. Internal budget records 4. Merger or acquisition records 5. Investor relations records 6. Other
K	Legal		<ol style="list-style-type: none"> 1. Confidentiality Agreements 2. Supplier Contracts 3. Customer Contracts 4. Blank legal contracts or templates 5. Pre-submission SEC filings 6. Submitted SEC filing 7. Litigation documents 8. Leases 9. Other
L	Other		<ol style="list-style-type: none"> 1. Blank Public Application 2. Case Study 3. Other (bank address, fax coversheets, unintelligible, web pages for the bank, charity requests, general firm info, instructions)
M	R&D		<ol style="list-style-type: none"> 1. Product / Service Roadmaps 2. Non-public R&D Results 3. Pre-application patent records 4. Other
Z	Not Banking Related		<ol style="list-style-type: none"> 1. Other

Appendix B – Document Sensitivity Rating Scale

Level	Definition
High 3	<ul style="list-style-type: none"> • Any file marked “CONFIDENTIAL”, “PRIVATE”, “RESTRICTED”, “SECRET”, “SENSITIVE” • Documents that commonly require signing a Non Disclosure Agreement or private background check: Examples include information relating to contracts, financial information, policies, internal memos, mergers, acquisitions, R&D results, etc. • Public disclosure could <i>materially</i> damage business operations, market position (patentability, competitive position, brand equity), equity price, or damage a large number of customers or suppliers of organization. • Trade secrets (e.g., as described in the "Economic Espionage Act of 1996 (18 USC 1831-39)")
Medium 2	<ul style="list-style-type: none"> • Information that is either protected by privacy laws or must be kept private for other reasons. Human resources data is one example of data that can be classified as medium risk. Also, identifying information such as credit card or other financial information, SSN or other government IDs • Public disclosure will (a) negatively affect the safety, career, reputation or lifestyle of an employee, customer, agent, or supplier; (b) lead to crimes such as identity theft or fraud; (c) subject organization to civil remedies and/or criminal penalties for non-compliance in record keeping (d) cause significant PR damage and loss of brand equity
Low 1	<ul style="list-style-type: none"> • Information that is commonly shared with others in course of business but not with the general public (and is therefore quasi-public). • Examples include resumes, cover letters, forms, sales presentations. • Public disclosure might breach privacy or pose some business risk
Public 0	<ul style="list-style-type: none"> • Designed for public consumption. • Public disclosure can do no harm to organization, its customers, or its suppliers.