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

M. ERIC JOHNSON

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. He holds a B.S. in Engineering, a 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. 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, the Department of Justice, the Department of Homeland Security, and the National Science Foundation, he is studying how information security and trust affect supply-chain relationships. He has testified before the U.S. 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 academic journals such 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, Dr. Johnson taught at the Owen Graduate School of Management, Vanderbilt University. He was previously employed by Hewlett-Packard and Systems Modeling. 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. Dr. Johnson recently served as president of the Production and Operations Management Society (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.*

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.

Journal of Management Information Systems / Fall 2008, Vol. 25, No. 2, pp. 97–123.

0742–1222 / 2008 $9.50 + 0.00.
DOI 10.2753/MIS0742-1222250205

KEY WORDS AND PHRASES: data breaches, file-sharing, information security, inadvertent disclosure, intellectual property leaks, peer-to-peer networks, risk management.

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 [31]. However, it is often not clear what security initiatives offer firms the greatest improvement [11]. A close look at the headlines reveals a bewildering set of information breaches. While hackers regularly penetrate poorly secured networks [25] and devices [1], many of the large recent security breaches were not technical break-ins, but rather inadvertent disclosures. For example, 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 [5, 19].

Organizations have mistakenly posted on the Web many different types of sensitive information, from legal to medical to financial (e.g., [17] or [30]). Even technology firms such as Google and AOL have suffered the embarrassment of inadvertent Web posting of sensitive information [2, 22]—in their cases, customer information. Still other firms have seen their internal information and intellectual property appear on blogs, YouTube, and MySpace [29]. 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 (P2P) file-sharing networks.

Despite significant efforts of the music industry, P2P file sharing has become mainstream among large segments of the Internet population. With estimates of 10 million simultaneous users [20] sharing music, video, software, and photos, file-sharing clients have joined the suite of standard PC applications for many users. Unrecognized by many of these users though is the serious security threat participation in these networks poses to both corporate and individual security [10].

In earlier research, Johnson et al. [12] showed why P2P file sharing represents a growing security risk. 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., Recording Industry Association of America 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 [27]. 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 U.S. 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 that the firms experiencing greater leakage also experience increased threat. Finally, we discuss managerial implications and propose a simple benchmarking technique to compare leakages. 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 P2P NETWORKS enables 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 [21], and a string of legal challenges by the music and video content industry (e.g., the Recording Industry Association of America and the Motion Picture Association of America). 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, eDonkey, 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 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, e-mail 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 [15].

Today, file sharing continues to grow, with usage doubling from less than 4 million in 2003 to nearly 10 million simultaneous users in 2006 [20]. 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 such as 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 whereas others, such as 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. Moreover, P2P file-sharing systems often provide incentives for users to share files via 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 [7] and features of P2P clients that facilitate inadvertent sharing [28]. 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 honeypot experiments, that there are significant threats from individuals actively seeking this information to commit theft [12]. In that paper, it was argued 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. They include:

- *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 [14, 23]. Viruses that exist in e-mail and other programs also have variants that exist in P2P networks [9].

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 file-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 held by customers, suppliers, contractors, and other business partners, and they may be leaking documents, too. The nature of information flows within the extended enterprise significantly increases the challenge of preventing leaks.

Methodology and Data

TO CHARACTERIZE THE RISKS 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. As noted earlier, we focused our analysis on the supply chains of the *Forbes* top 30 U.S.-based banks [3]. Those institutions collectively employ more than 1 million people, manage more than $7 trillion, and comprise a wide range of sizes as shown in Table 1.

With the help of Tiversa Inc., which monitors global P2P file-sharing networks, we gathered and categorized P2P searches and shared files related to these institutions over a seven-week period (December 27 to 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 eDonkey (e.g., eMule, eDonkey2000). 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 faction of the population at any moment. As mentioned earlier, recent estimates place the P2P population at nearly 10 million simultaneous users

Table 1. Summary Statistics on Institutions in Data Set ($N = 30$)

	Average	Standard deviation	Maximum	Minimum
Employees	47,406	68,020	307,000	2,202
Number of branches	1,567	1,919	7,237	41
Sales (billions of dollars)	17.94	28.84	120.32	1.28
Assets (billions of dollars)	248.42	395.25	1,494.04	26.28
Market value (billions of dollars)	40.34	56.95	230.93	4.49

Sources: Forbes and Hoovers.

[20]. The networks themselves are dynamic, with members constantly joining (and sharing files) and leaving. Thus, over a period of a day, some estimate that as many as 20 million users issue upwards of 800 million searches. Using Tiversa's systems, we participated in those networks globally and collected a 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 such as 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, and Pavilion);
- subsidiaries, divisional names (e.g., HP Shopping and Home Products Division); and
- suppliers, contractors, vendors (e.g., Celestica and 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 to 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

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"

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 four-point threat scale: high (3), medium (2), low (1), and public (0). Although 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 that 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 [4]. 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, and so on. 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 expose vulnerabilities that could still represent a threat to the institution and its customers.

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

Medium-threat searches, such as those for bank names, are more generic. Low-threat searches such as "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 Citizens Bank. See other examples in Table 4.

Inadvertently Disclosed Files: The Vulnerability

During this same period, we also collected files that we observed being shared on the networks. We focused on business-related files, particularly those from Microsoft

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	soverignbank	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

Table 4. Examples of Digital Wind

Institution affected	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 programs
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)

Office Suite (including file extensions doc, xls, ppt, mdb, along with rft, pdf, and txt). Using the digital footprint, file names with any related terms were captured. In some P2P networks, files are also indexed by their associated meta-data (such as the name of the firm to which a word processor is registered). Thus, we captured those documents as well. Using this approach, we collected more than 114,000 files totaling more than 15GB of data over the seven-week period. Tiversa’s systems allowed us to limit the files harvested to unique Internet protocol (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, thereby reducing the data set from all files found over all seven weeks to those found during the last week of collection. Files were manually evaluated on multiple dimensions [24]. For each file examined, we noted if the file was flagged to reduce distribution; for example, if it was marked “Confidential,” “Restricted,” “Internal Use,” and so on. We recorded the file’s age by examining both the file’s meta-data (e.g., creation and editing dates) and dates inside the document. We also assessed the source of the leak (customers, suppliers, internal) by examining

IP addresses and clues within the document. After examining the document, it was classified based on its type and on a four-point scale of its sensitivity as reflected by the potential consequence if exploited. (These methods are further described in Appendix Tables A1 and A2.) Like the search classification scheme, the scale included a high (3), medium (2), and low (1) along with (0) for public documents. *Public documents* are ones that the firm would want widely distributed (although they may be surprised to know these document 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 [8], which has observed that small accidents often precede much larger ones.

Results

THIS SECTION PROVIDES AN OVERVIEW OF 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, bank names are not included, and bank numbers shown in the figure are randomly assigned. They do not represent the *Forbes* ranking numbers. 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 brands seemed to experience the most search activity. It is hypothesized 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 [16]. Firms 2 and 6 represent banks in this category and experienced a large number of highly threatening searches. Bank 20 represents the case of a bank experiencing significant digital wind. That bank does not 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 Figure 1, many of those small institutions still experienced targeted searches. The figure clearly demonstrates the threat faced by these institutions.

To test the hypothesis that search activity is correlated with bank brand visibility, we performed a least squares regression on a linear model of *searches* (Y). Brand strength in marketing [16] is often measured on positive brand attributes (e.g., quality, value, trustworthiness, reliability). However, we were more interested in the notoriety of the

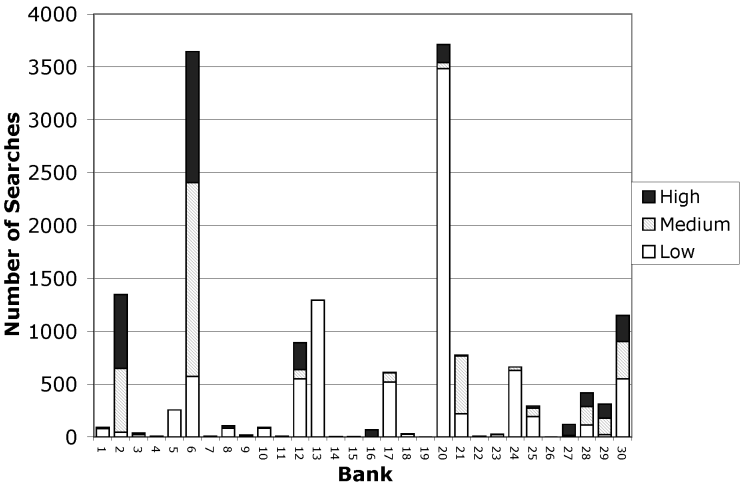

Figure 1. Search Threat Categorization for Top 30 Financial Institutions over Seven-Week Time Period (Sequence Does Not Correlate with Rank)

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, which 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.

Because low-threat searches (1) were driven by other phenomena unrelated to the bank, such as popularity of a 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 percent of the variation of search activity between banks. A regression limiting *Y* to high-threat searches (3) yielded even stronger support (R^2 of 0.86 with significant coefficients at 0.01).

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

Table 5. Support for the Relationship Between Brand Visibility (Measured by Employees) and Searches (7,194 Searches)

Regression statistics	
Multiple R	0.89
R^2	0.80
Adjusted R^2	0.79
Standard error	276.01
Observations	30

Analysis of variance

	Degrees of freedom	Sum of squares	Mean square	F	Significance F
Regression	1	8,330,109.8	8,330,109.8	109.3	3.56E-11
Residual	28	2,133,029.0	76,179.6		
Total	29	10,463,138.8			

	Coefficients	Standard error	t-statistic	p-value	Lower 95 percent	Upper 95 percent
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

to determine bank relevance and sensitivity. We chose this approach based on our hypothesis that documents found 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 [18].

In the end, we found limited support for this hypothesis from the data. Given the vast sea of files floating in the P2P environment 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 for those found early or later in our data collection.

The 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 percent were duplicates). Duplicate documents are 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 sources indicated a breakdown as shown in Figure 2.

The source was determined by an analysis of the content of the file, its meta-data, and the disclosing IP address, categorizing them into three groups: individuals 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, and also included consultants, information technology (IT) suppliers, processors, and so on. 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.

We found files of nearly every type (see Figure 3), but personally identifiable information (PII) documents were the most numerous, accounting for 49 percent of all unique documents. Many of these documents contained enough information to easily commit fraud or identity theft. (See Appendix Table A1 for group definitions.)

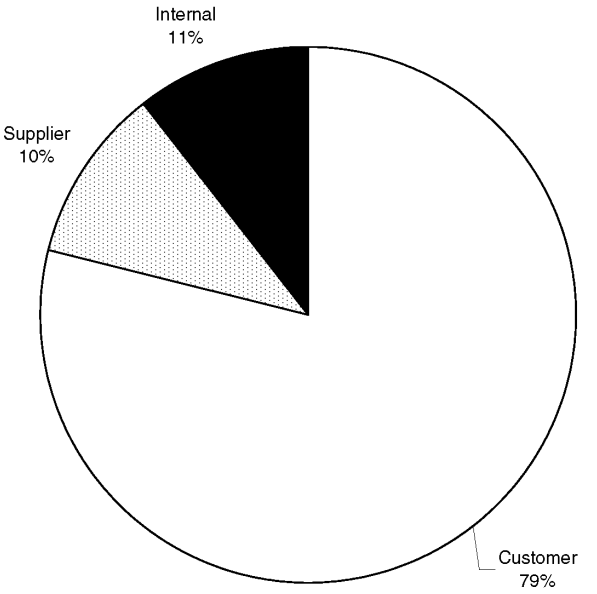

Figure 2. Document Source

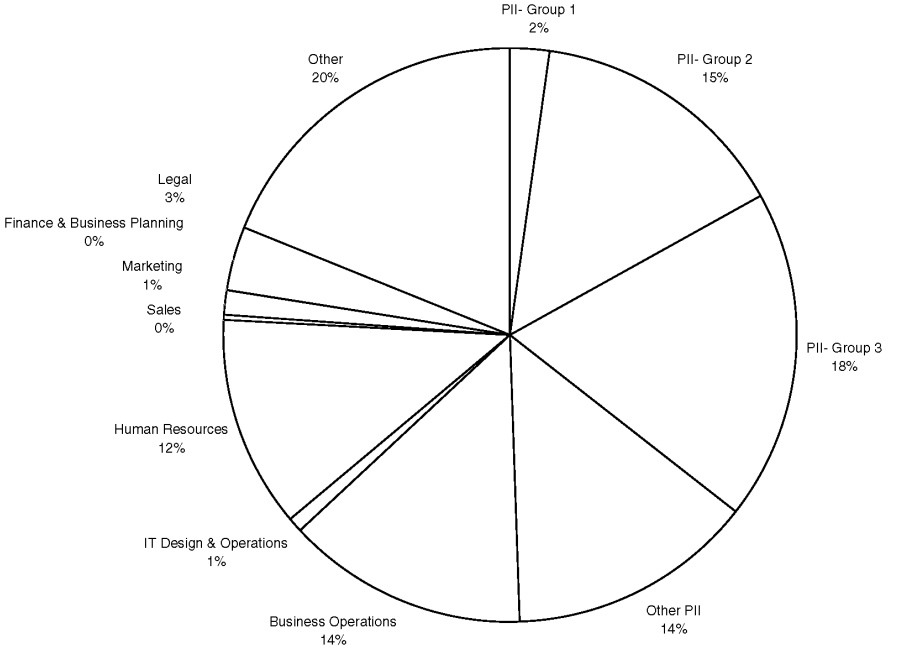

Figure 3. Document Type (Among All 1,708 Unique, Relevant Documents)

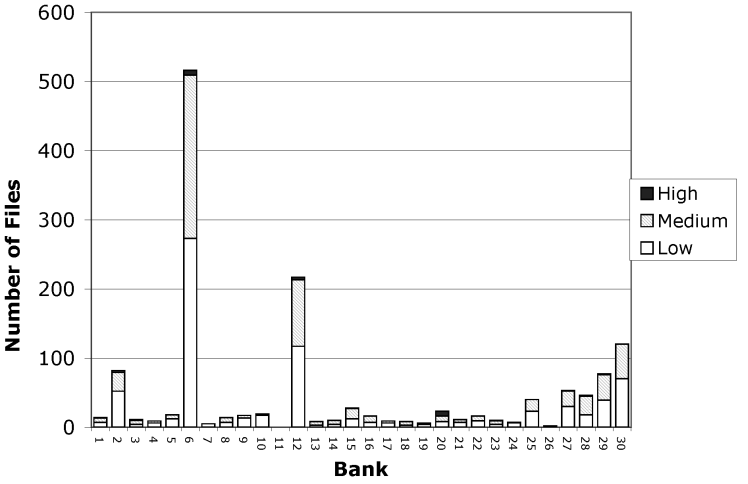

Figure 4. File Disclosure Categorization (Risk Rated as High, Medium, or Low) for Top 30 Institutions (Sequence Does Not Correlate with Bank Rank; see the Appendix for Rating Details)

The next largest category was the "other" category, which included bank addresses, charity requests, instructions, articles, fax cover sheets, and blank (public) forms. Business operations documents included employee training manuals, internal policies and procedures, and work plans. Many others originated from suppliers with regard 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 resumes, 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.

A graphic summary of the sensitivity of the 1,708 unique, relevant documents is shown in Figure 4. Again, to protect specific institutions, bank names are not included. The 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, it is argued that the number of employees is directed related to internal leak sources 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-sensitivity documents. This represented 1,412 files.

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 percent of the variation of document activity between banks. A regression limiting Y to medium- plus high-sensitivity files (levels 2 and 3) yielded a similar result (R^2 of 0.81 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, online retail activity, digital practices of the banks, outsourcing activity, international presence, and so on.

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 Federal Deposit Insurance Corporation (FDIC) collects and publishes data on retail accounts, which are defined so that the dollar value of deposits accounts for $100,000 or less [26]. Using that data, it can be seen in Table 7 that the overall model provides a strong fit with statistically significant coefficients. However, surprisingly, the coefficient on the number of accounts is negative. First, note that one might expect multicollinearity between employees and accounts: banks with many accounts will likely have many employees. Using an auxiliary regression [13] between accounts and employees, we detected the presence of some multicollinearity ($R^2 = 0.56$), but not a near-exact linear dependence. Of course, regressing just accounts against files does produce a positive and significant coefficient but with a low $R^2 = 0.19$. Reflecting on the negative coefficient in the multiple 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 the banks' focus: those with a strong retail focus may be taking better steps to educate and protect customers from leakage.

The Link Between Threat and Vulnerability

With a picture of the vulnerability, we return to further examine the related threat. While bank visibility is certainly linked to the threat, there are other factors that

Table 6. Support for the Relationship Between Employment Base and Files Found (1,412 Files)

Regression statistics	
Multiple R	0.92
R^2	0.84
Adjusted R^2	0.83
Standard error	40.54
Observations	30.00

Analysis of variance

	Degrees of freedom	Sum of squares	Mean square	F	Significance F
Regression	1	238,490.3	238,490.3	145.1	1.35828E-12
Residual	28	46,011.6	1,643.3		
Total	29	284,501.9			

	Coefficients	Standard error	t-statistic	p-value	Lower 95 percent	Upper 95 percent
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

Table 7. Support for the Relationship Between Employment Base and Number of Retail Accounts and Files Found (1,412 Files)

Regression statistics	
Multiple *R*	0.99
R^2	0.97
Adjusted R^2	0.97
Standard error	17.06
Observations	30

Analysis of variance

	Degrees of freedom	Sum of squares	Mean square	*F*	Significance *F*
Regression	2	276,643.0	138,321.5	475.2	9.06E-22
Residual	27	7,858.9	291.1		
Total	29	284,501.9			

	Coefficients	Standard error	*t*-statistic	*p*-value	Lower 95 percent	Upper 95 percent
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

security professionals point toward. 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 increase 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 increase the threat of discovery and exploitation of inadvertent disclosures.*

To test this modification of Hypothesis 1a, 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 H1b with good model fit and statistically significant coefficients at the 0.05 and 0.01 levels. Again, one might be concerned with multicollinearity between employee and files, since our earlier result showed a significant positive relationship. We note that a model with searches alone also produces strong fit ($R^2 = 0.82$) with a significant, positive coefficient.

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 P2P file-sharing networks, have created many new security risks for organizations. In this paper, we 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 meta-data 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

Table 8. Support for the Relationship Between Brand Visibility (Measured by Employees) and Vulnerability (Measured by Leaked Documents) and Searches

Regression statistics	
Multiple R	0.92
R^2	0.84
Adjusted R^2	0.83
Standard error	246.03
Observations	30

Analysis of variance

	Degrees of freedom	Sum of squares	Mean square	F	Significance F
Regression	2	8,828,752.6	4,414,376.3	72.9	1.3E-11
Residual	27	1,634,386.2	60,532.8		
Total	29	10,463,138.8			

	Coefficients	Standard error	t-statistic	p-value	Lower 95 percent	Upper 95 percent
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

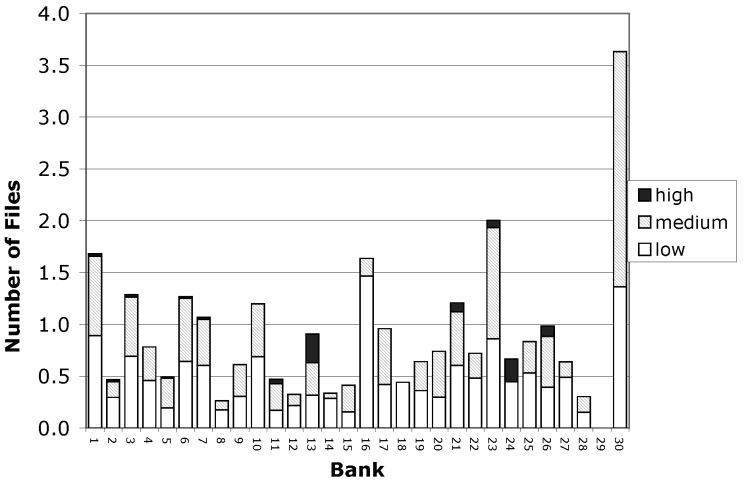

Figure 5. File Disclosure Categorization per 1,000 Employees for Top 30 Institutions (Banks Reordered to Disguise Identity; Risk Rated as High, Medium, or Low)

developing effective strategies to help individuals in the extended enterprise make better information risk decisions [6]. For Web applications such as file sharing, the benefits to the individual sometimes outweigh the perceived risks because users do not 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 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.

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

Acknowledgments: The author gratefully acknowledges helpful comments from seminar audiences at Emory University, University of Maryland, Carnegie Melon (WEIS), the Forty-First Annual Hawaii International Conference on System Sciences, 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, which 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 and do not necessarily represent the official position of the U.S. Department of Homeland Security, the Science and Technology Directorate, I3P, or Dartmouth College.

References

1. Bank, D. Stores blame checkout software for security breaches. *Wall Street Journal* (April 27, 2005) (available at http://online.wsj.com/public/article/SB111455367943717582-khRhpogsLZXJxrYrn0YAfx9bTvA_20050529.html).

2. Claburn, T. Minor Google security lapse obscures ongoing online data risk. *Information Week* (January 22, 2007) (available at www.informationweek.com/news/internet/showArticle.jhtml?articleID=196902585).

3. DeCarlo, S. The world's largest public companies. *Forbes.com* (March 31, 2006) (available at www.forbes.com/2006/03/30/largest-public-companies_06f2k_cz_sk_0331forbes2000intro.html).

4. DeVellis, R.F. *Scale Development: Theory and Applications,* 2d ed. London: Sage, 2003.

5. Francis, T. Towers Perrin laptops, client data stolen. *Wall Street Journal* (January 9, 2006), B2.

6. Goetz, E., and Johnson, M.E. Security through information risk management. Technical report, Institute for Information Infrastructure Protection, Saginaw, MI, 2007 (available at http://mba.tuck.dartmouth.edu/digital/Programs/CorporateEvents/CISO2007/Overview.pdf).

7. Good, N.S., and Krekelberg, A. Usability and privacy: A study of Kazaa P2P file-sharing. In G. Cockton and P. Korhnen (eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* New York: ACM Press, 2003, pp. 137–144.

8. Hammer, W., and Price, D. *Occupational Safety Management and Engineering,* 5th ed. New York: Prentice Hall, 2000.

9. Ingram, M. 66,000 names and personal details leak on P2P. *Slyck News* (April 29, 2006) (available at www.slyck.com/news.php?story=1169).

10. Johnson, M.E. Inadvertent file-sharing over peer-to-peer networks. Testimony before the Committee on Oversight and Government Reform, United States House of Representatives, Washington, DC, July 24, 2007 (available at http://oversight.house.gov/documents/20070724103956.pdf).

11. Johnson, M.E., and Goetz, E. Embedding information security risk management into the extended enterprise. *IEEE Security and Privacy, 5,* 3 (May–June 2007), 16–24.

12. Johnson, M.E.; McGuire, D.; and Willey, N.D. Why file-sharing networks are dangerous. *Communications of the ACM,* 2008, forthcoming.

13. Judge, G.G.; Hill, R.C.; Griffiths, W.E.; Lutkepohl, H.; and Lee, T. *Introduction to the Theory and Practice of Econometrics.* New York: John Wiley and Sons, 1982.

14. Kalafut, A., Acharya, A., and Gupta, M. A study of malware in peer-to-peer networks. In *Proceedings of the Internet Measurement Conference.* New York: ACM Press, 2006, pp. 327–332.

15. Karagiannis., T.; Broido, A.; Brownlee, N.; Claffy, K.; and Faloutsos, M. File-sharing in the Internet: A characterization of P2P traffic in the backbone. Technical report, Computer Science, University of California, Riverside, 2003.

16. Keller, K.L. *Strategic Brand Management,* 2d ed. Upper Saddle River, NJ: Prentice Hall, 2003.

17. Kenworthy, T. Bryant's accuser files civil suit. *USA Today* (August 10, 2004) (available at www.usatoday.com/sports/basketball/nba/2004-08-10-bryant-accuser-civil-suit_x.htm).

18. Law, A.M., and Kelton, W.D. *Simulation Modeling and Analysis,* 3d ed. New York: McGraw-Hill, 2000.

19. Levitz, J., and Hechinger, J. Laptops prove weakest link in data security. *Wall Street Journal* (March 26, 2006), B1.

20. Mennecke, T. P2P population continues climb. *Slyck News* (June 14, 2006) (available at www.slyck.com/news.php?story=1220).

21. Oberhozer-Gee, F., and Strumpf, K. The effect of file-sharing on record sales: An empirical analysis. *Journal of Political Economy, 115,* 1 (February 2007), 1–42.

22. Olson, P. AOL shoots itself in the foot. *Forbes* (August 8, 2006) (available at www.forbes.com/technology/2006/08/08/aol-internet-update-cx_po_0808privacy.html).

23. Shin, S.; Jung, J.; and Balakrishnan, H. Malware prevalence in the KaZaA file-sharing network. In *Proceedings of the Internet Measurement Conference.* New York: ACM Press, 2006, pp. 333–338.

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

25. Sidel, R. Giant retailer reveals customer data breach. *Wall Street Journal* (January 18, 2007), D1.

26. Statistics on depository institutions. Federal Deposit Insurance Corporation, Washington, DC, 2008 (available at www2.fdic.gov/sdi/index.asp).

27. Sun, L.; Srivastava, R.P.; and Mock, T.J. An information systems security risk assessment model under the Dempster–Shafer theory of belief functions. *Journal of Management Information Systems, 22,* 4 (Spring 2006), 109–242.

28. Sydnor, T.D., II; Knight, J.; and Hollaar, L.A. Filesharing programs and "technological features to induce users to share." Report to the United States Patent and Trademark Office from the Office of International Relations, United States Patent and Trademark Office, Washington, DC, November 2006 (available at www.uspto.gov/web/offices/dcom/olia/copyright/oir_report_on_inadvertent_sharing_v1012.pdf).

29. Totty, M. Security: How to protect your private information. *Wall Street Journal* (January 29, 2007), R1.

30. Twedt, S. UPMC patients' personal data left on Web. *Pittsburgh Post-Gazette* (April 12, 2007) (available at www.post-gazette.com/pg/07102/777281-114.stm).

31. Yue, W.T., and Çakanyildirim, M. Intrusion prevention in information systems: Reactive and proactive responses. *Journal of Management Information Systems, 24,* 1 (Summer 2007), 329–353.

Appendix Table A1. Disclosure Classification Scheme (Document Type)

	Major category	Definition	File categories
A	Personally identifiable information (PII)—group 1	Files that contain information that can uniquely identify a person to enable fraud or identity theft. Files contain at least three of: • Social Security number • Credit card number • User account number • User ID and password • Full address • Signature	1. Dispute letters 2. Wire transfer authorizations 3. Credit reporting agency records (e.g., Equifax) 4. User ID/password list records and account records 5. Tax returns 6. Customer service correspondence 7. Account closure 8. Statements/payment receipts 9. Other
B	PII—group 2	Files that contain information that can uniquely identify a person to enable fraud or identity theft. Files contain at least two of: • Social Security number • Credit card number • User account number • User ID and password • Full address • Signature	1. Dispute letters 2. Wire transfer authorizations 3. Credit reporting agency records (e.g., Equifax) 4. User ID/password list records and account records 5. Tax returns 6. Customer service correspondence 7. Account closure 8. Statements/payment receipts 9. Other
C	PII—group 3	Files that contain information that can uniquely identify a person to enable fraud or identity theft? Files contain at least one of: • Social Security number • Credit card number • User account number • User ID and password • Full address • Signature	1. Dispute letters 2. Wire transfer authorizations 3. Credit reporting agency records (e.g., Equifax) 4. User ID/password list records and account records 5. Tax returns 6. Customer service correspondence 7. Account closure 8. Statements/payment receipts 9. Other

(continues)

Appendix Table A1. Continued

	Major category	Definition	File categories
D	Other PII	PII that does not meet the criteria in A, B, or C	1. Dispute letters 2. Wire transfer authorizations 3. Credit reporting agency records (e.g., Equifax) 4. User ID/password list records and account records 5. Tax returns 6. Customer service correspondence 7. Account closure 8. Statements/payment receipts 9. Other
E	Business operations		1. Internally released PII 2. Internal organizational phone/e-mail lists 3. Customer lists 4. Employee training materials 5. Internal policies and 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 and operations		1. Network and 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 road maps 9. Other
G	Human resources	1. Employee pay or bonus records 2. Existing employee reviews and 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	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	1. Current press releases in markup 2. Past press releases 3. Focus group study results 4. Public relations plans 5. Other

(continues)

Appendix Table A1. Continued

	Major category	Definition	File categories
J	Finance and business planning		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		1. Confidentiality agreements 2. Supplier contracts 3. Customer contracts 4. Blank legal contracts or templates 5. Presubmission SEC filings 6. Submitted SEC filings 7. Litigation documents 8. Leases 9. Other
L	Other		1. Blank public application 2. Case study 3. Other (bank address, fax cover sheets, Web pages for the bank, charity requests, general firm info, instructions, etc.)
M	R&D		1. Product/service road maps 2. Nonpublic R&D results 3. Preapplication patent records 4. Other
Z	Not banking related		1. Other

Appendix A2. Document Sensitivity Rating Scale

Level	Definition
High (3)	Any file marked "CONFIDENTIAL," "PRIVATE," "RESTRICTED," "SECRET," "SENSITIVE"
	Documents that commonly require signing a nondisclosure agreement or private background check: examples include information relating to contracts, financial information, policies, internal memos, mergers, acquisitions, R&D results, and so on.
	Public disclosure could materially 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)	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, Social Security numbers, or other government IDs.
	Public disclosure will (1) negatively affect the safety, career, reputation, or lifestyle of an employee, customer, agent, or supplier; (2) lead to crimes such as identity theft or fraud; (3) subject organization to civil remedies or criminal penalties for noncompliance in record keeping; (4) cause significant public relations damage and loss of brand equity.
Low (1)	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)	Designed for public consumption.
	Public disclosure can do no harm to organization, its customers, or its suppliers.