The inadvertent disclosure of personal health information through peer-to-peer file sharing programs

Khaled El Emam,1,2,3 Emilio Neri,2 Elizabeth Jonker,2 Marina Sokolova,2 Liam Peyton,3 Angelica Neisa,2 Teresa Scassa4

ABSTRACT
Objective There has been a consistent concern about the inadvertent disclosure of personal information through peer-to-peer file sharing applications, such as Limewire and Morpheus. Examples of personal health and financial information being exposed have been published. We wanted to estimate the extent to which personal health information (PHI) is being disclosed in this way, and compare that to the extent of disclosure of personal financial information (PFI).

Design After careful review and approval of our protocol by our institutional research ethics board, files were downloaded from peer-to-peer file sharing networks and manually analyzed for the presence of PHI and PFI. The geographic region of the IP addresses was determined, and classified as either USA or Canada.

Measurement We estimated the proportion of files that contain personal health and financial information for each region. We also estimated the proportion of search terms that return files with personal health and financial information. We ascertained and discuss the ethical issues related to this study.

Results Approximately 0.4% of Canadian IP addresses had PHI, as did 0.5% of US IP addresses. There was more disclosure of financial information, at 1.7% of Canadian IP addresses and 4.7% of US IP addresses. An analysis of search terms used in these file sharing networks showed that a small percentage of the terms would return PHI and PFI files (ie, there are people successfully searching for PHI and PFI on the peer-to-peer file sharing networks).

Conclusion There is a real risk of inadvertent disclosure of PHI through peer-to-peer file sharing networks, although the risk is not as large as for PFI. Anyone keeping PHI on their computers should avoid installing file sharing applications on their computers, or if they have to use such tools, actively manage the risks of inadvertent disclosure of their, their family’s, their clients’, or patients’ PHI.

INTRODUCTION
Between 15% and 17% of US adults have changed their behavior to protect the privacy of their personal health information (PHI), doing things such as: going to another doctor, paying out-of-pocket when insured to avoid disclosure, not seeking care to avoid disclosure to an employer, giving inaccurate or incomplete information on medical history, self-treating or self-medicating rather than seeing a provider, or asking a doctor not to write down the health problem or record a less serious or embarrassing condition.1–3 Privacy concerns have caused individuals to not be totally honest with their healthcare provider.4 In a survey of physicians in the USA, nearly 87% reported that a patient had asked that information be kept out of their record, and nearly 78% of physicians said that they had withheld information from a patient’s record due to privacy concerns.5 Public opinion surveys in Canada found that, over the prior year, between 5% and 5% of Canadians have withheld information from their provider because of privacy concerns, and 1–5% have decided not to seek care for the same reasons.6 Furthermore, between 11% and 13% of Canadians have at some point withheld information from a healthcare provider because of concerns over with whom the information might be shared, or how it might be used.7,8 with the highest regional percentage in Alberta at 20%.7 Similar results have been reported by the Canadian Medical Association.9 An estimated 735 000 Canadians decided not to see a healthcare provider because of concerns about the privacy of their information.9 Specific vulnerable populations have reported similar privacy protective behaviors, such as adolescents, people with HIV or at high risk for HIV, women undergoing genetic testing, mental health patients, and battered women.10

With the growing use of information technology in the provision of healthcare,11–20 patients and physicians are worried about unauthorized disclosure and use of PHI.1 4 21–26 In addition, a considerable amount of PHI is also disclosed with patient consent through compelled authorizations (eg, to obtain insurance, make an insurance claim, or seek employment).27 But, should the worst happen and there is a privacy breach affecting their health information, between 61% and 74% of Canadians want to be notified, as well as the oversight bodies.28 29

Breach notification can only happen if the data custodian knows that a breach has occurred. In some cases inadvertent disclosure of PHI may not even be known to the data custodian. For example, a recent study found that approximately 10% of personal computers bought on the secondhand market in Canada contained identifiable health information.30 This means that many Canadians (individuals and corporations) are selling or disposing of their computers unaware that they contain PHI.

Another potentially significant mechanism for inadvertent disclosure of PHI is not to be totally honest with their healthcare provider.4 In a survey of physicians in the USA, nearly 87% reported that a patient had asked that information be kept out of their record, and nearly 78% of physicians said that they had withheld information from a patient’s record due to privacy concerns.5 Public opinion surveys in Canada found that, over the prior year, between 5% and 5% of Canadians have withheld information from their provider because of privacy concerns, and 1–5% have decided not to seek care for the same reasons.6 Furthermore, between 11% and 13% of Canadians have at some point withheld information from a healthcare provider because of concerns over with whom the information might be shared, or how it might be used.7,8 with the highest regional percentage in Alberta at 20%.7 Similar results have been reported by the Canadian Medical Association.9 An estimated 735 000 Canadians decided not to see a healthcare provider because of concerns about the privacy of their information.9 Specific vulnerable populations have reported similar privacy protective behaviors, such as adolescents, people with HIV or at high risk for HIV, women undergoing genetic testing, mental health patients, and battered women.10

With the growing use of information technology in the provision of healthcare,11–20 patients and physicians are worried about unauthorized disclosure and use of PHI.1 4 21–26 In addition, a considerable amount of PHI is also disclosed with patient consent through compelled authorizations (eg, to obtain insurance, make an insurance claim, or seek employment).27 But, should the worst happen and there is a privacy breach affecting their health information, between 61% and 74% of Canadians want to be notified, as well as the oversight bodies.28 29

Breach notification can only happen if the data custodian knows that a breach has occurred. In some cases inadvertent disclosure of PHI may not even be known to the data custodian. For example, a recent study found that approximately 10% of personal computers bought on the secondhand market in Canada contained identifiable health information.30 This means that many Canadians (individuals and corporations) are selling or disposing of their computers unaware that they contain PHI.

Another potentially significant mechanism for inadvertent disclosure of PHI is not to be totally honest with their healthcare provider.4 In a survey of physicians in the USA, nearly 87% reported that a patient had asked that information be kept out of their record, and nearly 78% of physicians said that they had withheld information from a patient’s record due to privacy concerns.5 Public opinion surveys in Canada found that, over the prior year, between 5% and 5% of Canadians have withheld information from their provider because of privacy concerns, and 1–5% have decided not to seek care for the same reasons.6 Furthermore, between 11% and 13% of Canadians have at some point withheld information from a healthcare provider because of concerns over with whom the information might be shared, or how it might be used.7,8 with the highest regional percentage in Alberta at 20%.7 Similar results have been reported by the Canadian Medical Association.9 An estimated 735 000 Canadians decided not to see a healthcare provider because of concerns about the privacy of their information.9 Specific vulnerable populations have reported similar privacy protective behaviors, such as adolescents, people with HIV or at high risk for HIV, women undergoing genetic testing, mental health patients, and battered women.10

With the growing use of information technology in the provision of healthcare,11–20 patients and physicians are worried about unauthorized disclosure and use of PHI.1 4 21–26 In addition, a considerable amount of PHI is also disclosed with patient consent through compelled authorizations (eg, to obtain insurance, make an insurance claim, or seek employment).27 But, should the worst happen and there is a privacy breach affecting their health information, between 61% and 74% of Canadians want to be notified, as well as the oversight bodies.28 29

Breach notification can only happen if the data custodian knows that a breach has occurred. In some cases inadvertent disclosure of PHI may not even be known to the data custodian. For example, a recent study found that approximately 10% of personal computers bought on the secondhand market in Canada contained identifiable health information.30 This means that many Canadians (individuals and corporations) are selling or disposing of their computers unaware that they contain PHI.
Disclosure of PHI is through peer-to-peer file sharing applications. There are different peer-to-peer clients that can search and download files from various networks. Table 1 shows a mapping for the currently most popular clients and networks. The use of peer-to-peer file sharing applications is increasing, with one US study in 2003 estimating that 26 million adults share files online.

As summarized in box 1, many of these file sharing applications have features that facilitate (or even encourage) the inadvertent sharing of media files as well as documents, email, and database files. Other mechanisms whereby files may be inadvertently disclosed include users accidentally putting files in shared directories; peer-to-peer network incentives to share more files result in users sharing many directories (e.g., by scoring users based on how many files they share); and unawareness or forgetfulness about the contents of files that are being shared. This poses a significant privacy risk because users may be sharing a large amount of personal (health) information unknowingly by participating in these file sharing networks.

Thus far, there have been no systematic empirical estimates of the extent to which PHI is being disclosed through these file sharing applications, and whether anyone has successfully accessed such PHI. If users of these applications are unknowingly disclosing PHI, then this would constitute a serious privacy breach. In addition, there have been no comparisons of the extent of disclosure of PHI to personal financial information (PFI).

In this study we examine the extent to which PHI is being disclosed through peer-to-peer file sharing networks in Canada and the USA. We address two specific issues. First, we estimate the extent to which PHI is being disclosed in Canada, and compare that to the USA. We also compare the extent to which PHI is being disclosed to the extent to which PFI is being disclosed. However, if documents containing PHI are being made available, that does not necessarily mean that anyone is actually finding these documents. Therefore, the second issue we address is the proportion of searches on the file sharing networks that return documents containing PHI.

**Methods**

This research protocol was approved beforehand by the Research Ethics Board of the Children’s Hospital of Eastern Ontario Research Institute. Below we describe the methods that we followed for data collection and estimation. Our focus was on files that are likely to contain correspondence in various formats (e.g., word processing files, email files, PDF files, and spreadsheet files). As described in Appendix A, these are most likely to contain PHI.

<table>
<thead>
<tr>
<th>Networks</th>
<th>eDonkey</th>
<th>OverNet</th>
<th>Gnutella</th>
<th>WinMX</th>
<th>BitTorrent</th>
<th>FastTrack</th>
<th>Ares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clients</td>
<td>Ares</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ares</td>
</tr>
<tr>
<td></td>
<td>Bearshare</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BitTorrent</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>eMule</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kazaa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Limewire</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Morpheus</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shareaza</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>WinMX</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

While the method descriptions in this section are for PHI only, the exact same approaches were used for PFI.

**Estimating the proportion of IP addresses with PHI**

The primary objective is to determine the proportion of IP addresses that are exposing PHI. We also wanted to: (a) compare the proportion of IP addresses with PHI in Canada with those in the USA, and (b) compare the proportion of IP addresses with PFI with those with PHI.

We modified an open source peer-to-peer file sharing client to automatically search multiple peer-to-peer file sharing networks, and download and organize the files. This modified client performed a wild card search for all document files (Word documents, Outlook email files, PDF files, Access database files, and Excel spreadsheets). Whenever a match was found, the file was downloaded to a repository and its originating IP address recorded. The main networks that were targeted for search were FastTrack, Gnutella, and eDonkey. The specific tool we modified is called Shareaza. All documents that were downloaded were run through an anti-virus program to ensure that malicious documents were quarantined.

The wildcard search was run continuously for four months. The super-peers in the peer-to-peer network maintain an index of files in the nodes connected to them. They will also forward searches to other super-peers. Therefore, the longer the query runs the more indexes will be searched for the files. We continued searching for files until the IP address sample size (described below) was reached.

All documents were classified manually as containing PHI or not. The coding instructions are included in Appendix B. If there was at least one document that is classified as PHI, then the IP address was designated as containing PHI. The proportion of IP addresses with PHI gave us an estimate of the proportion of IP addresses on the network that have documents with PHI.

The IP addresses were geo-mapped so that we can determine the location of the machine. Each IP address was geo-mapped to a region: Canada or USA. The geographic location of IP addresses was obtained from IANA (http://www.iana.org). Geographic locations were determined based on the registration of each IP space and its assignment globally. Accuracy is almost 100% at the country level and becomes less precise at regional and municipal levels. To verify our approach, we selected a random subset of 75 Canadian and 75 US IP addresses from our results (according to the geolocations we determined) and sent only the IP addresses to Quova Inc. (one of the leading firms providing geolocation services). Their blinded country classifications matched ours 100% of the time.

Files that came from IP addresses outside the USA and Canada were discarded.
Examples of features which encourage inadvertent sharing can be divided into two high-level groups as described below. Not all peer-to-peer file sharing clients have all of these features, but they are still common.

**Inadvertently sharing downloaded files**
- **Redistribution feature**: This is often a default behavior such that when a user downloads a file from the network it is put in a directory that is also available for sharing. Therefore, all downloaded files are automatically available for sharing. If a user changes the download directory, then that is also automatically shared.
- **Coerced sharing feature**: The user interface makes it quite difficult to disable the sharing of the folder used to store downloaded files. In some cases, hidden functionality makes it quiet difficult to stop sharing. For example, in a recent version of LimeWire, a new “Individually Shared Files” feature was added, which allows the user to select which files can be shared individually rather than sharing whole directories. However, if the user un-shares the directory, that does not stop sharing the files inside it because they are also individually shared. Therefore, the user would also have to go in and un-share each individual file in the directory.

**Inadvertently sharing existing files**
- **Sub-folder sharing feature**: Whenever a folder is selected for sharing, then all of its sub-folders are also shared. The user interfaces often use the singular term “folder” when in fact all folder sharing is recursive.
- **Search wizard feature**: The search wizard will often be executed during the initial installation, or can be manually started after installation. The wizard will search all of the user’s machine and recommend directories to share. The recommendation is based on the existence of 'trigger' files, such as music or video files. If a user selects to share a directory, then all of its sub-directories are implicitly automatically shared.
- **Partial uninstall feature**: If a user uninstalls a file sharing client, it retains information about which directories were being shared. If later the user re-installs the same client or a new version of it, all of the previous sharing options are used. This makes it quite difficult to stop sharing certain directories, even if the user removes the program and starts again with a fresh installation.
- **Library share-folders feature**: If a user changes the directory that is being shared to another directory, say directory A to directory B, sharing from directory A does not actually stop. Changing the directory only adds directory B to be shared. Therefore, users can only share more directories but never less. Users are not explicitly made aware that by changing directories they are creating an incrementally expanding library of shared directories.

**Sample size**
Previous studies on the prevalence of personally identifying information (PII) in peer-to-peer file sharing networks reported rates between 49% and 61%. These were rates for documents, therefore they represent an upper bound for us since we were estimating the rates for IP addresses (because each IP address may have multiple documents). We make the most conservative assumption that PII will be prevalent at the same rate and use a 50% prevalence with a ±5% interval size. With a 95% CI for the estimate, 384 IP addresses would be needed. If the prevalence of PII is lower, this sample size will give us an interval at most of ±5%.

To compare the proportion of IP addresses with PII across regions, we tested the null hypothesis that the two regional proportions are equal. We performed a power analysis at a two-tailed $z$ level of 0.05 with 80% power. We assumed that the smallest inter-region difference is 0.1 (e.g., if the prevalence in Canada is 0.5 then the smallest detectable difference in prevalence for the USA would be 0.6 or 0.4). Under that smallest detectable difference assumption, we would need 392 IP addresses in each region.

To err on the conservative side, we planned to download the files from 800 IP addresses in each of the two regions. This would allow for descriptive estimates for each region and for comparisons of prevalence across regions.

**Estimating the proportion of actual searches that find PII**
We instrumented a super-peer in the eDonkey2000 network. This allowed us to capture the search/query terms that were being used over a two-month period. The node to obtain search terms was created using the Lugdunum eD2K node server. Modifications were made to the configuration to generate a log file containing all of the search terms which are passed to the node server. In an eDonkey2000 network the node servers act as index servers which index records contained on clients which connect to them. Search terms from clients are distributed along the node servers which then compare search terms to their indexed lists of files. When a match is found the details on the peer-to-peer connection are then passed back to the client. It is due to this mechanism that we were able to detect and log all search requests.

All search terms were then run against the files we downloaded in the first step of the study and classified as PII or not. A search term that matches a PII file was then considered PII-sensitive. We determined the proportion of search terms that were PII-sensitive. We retained the originating IP address for each search term so as not to double count the same searches coming from a single location.

**Reliability of ratings**
To determine the reliability of the manual classification of documents an inter-rater agreement analysis was performed. To decide how many documents need to be rated by a second rater we performed a power analysis for using the $\kappa$ statistic. To determine the expected $\kappa$ value for the power analysis we can rely on generally accepted benchmarks for $\kappa$ values. Hartman notes that $\kappa$ values should exceed 0.6. Landis and Koch provide a more general benchmark where values between 0.4 and 0.6 are considered moderate agreement. A similar benchmark is provided by Altman. Fleiss suggests that values between 0.4 and 0.75 represent intermediate to good agreement. To err on the conservative side we will assume that our value of $\kappa$ will be at least 0.5, which would be considered a moderate level of agreement according to the above benchmarks. At that level of agreement and 80% power to reject a null hypothesis comparing $\kappa$ to agreement by chance, the second rater needed to code 52 documents. A previous study found that the smallest $\kappa$ value when two observers coded PHI was 0.6 (and for a $\kappa$ that high we would need only 22 documents). To err on the conservative side we had two independent raters code 64 documents.
Two special protocols were put in place for this study and were overseen by the first author (KEE):

- If any illegal materials were discovered (e.g., child pornography or indications of incest), then that information would be passed on to the police.
- If there were cases of disclosure of particularly sensitive personal information or personal health information for a large number of individuals, then they would be reported to the appropriate (federal or provincial) privacy commissioner for follow-up.

Ethical considerations

In this study, PHI was collected from Canadian and American locations without consent. Common criteria used for deciding to waive the consent requirement for non-interventional research are:

1. The research involves no more than minimal risk to the participants.
2. The waiver is unlikely to adversely affect the rights and welfare of the subjects.
3. The research could not practically be carried out without the consent waiver.
4. Whenever possible or appropriate, the subjects will be provided with additional pertinent information after participation.

We will address these issues below.

About whom are we collecting personal information?

The PHI files do not necessarily pertain to the owners of the computers. They may contain health information about multiple family members, employees, or patients. Therefore, there are a variety of individuals who may be affected. We will make a distinction between the exposed individual about whom we have PHI and the sharers: the individuals who put the files on the file sharing network. An exposed individual may be a sharer, but not necessarily. A sharer may be an exposed individual, but not necessarily. Either of these may be the computer owner, but not necessarily.

Is the research of minimal risk to the participants?

Personal information is information about an identifiable individual. To enable us to classify files as either PHI or PFI, we are looking specifically for identifiable information. The study could not be performed if the information was de-identified. The main risk to the participants then would be if we revealed or inappropriately disclosed the identifiable personal information that we have collected.

All of the data gathered in this study were stored on encrypted computers for performing the manual and automated analysis reported in this paper. We will destroy the files after one year from the time of publication of the study (see the section below on “Data security and data destruction” for more details on our undertakings). No information regarding the IP addresses where any personal information was found was revealed in any publications or presentations, and no personal information was revealed in any publications. The collected data was only accessed by a subset of the research team (KEE, EN, EJ, MS); this subset is bound by the confidentiality clauses in their Children’s Hospital of Eastern Ontario employment contracts, and the data was not and will not be shared with other parties not participating directly in this study.

With these security measures in place, it can therefore be argued that there is minimal risk of harm to subjects.

Are files on a peer-to-peer file sharing network public information?

The information we collected is publicly available to anyone with access to the file sharing network—should this be considered public information?

The IRB Guideline notes that “some behavior that occurs in public places may not really be public behavior”, and that, for example, research involving covert recording of conversations in public parks raises invasion of privacy questions. Waskul frames it in these terms: if this recording happened to your conversations, “would you not feel that your privacy was grossly violated? Would you not be outraged?”

On the other hand, within a public place, subjects would be aware that their behavior and any information they share is available for observation by others. It has been argued that there is no reasonable expectation of privacy in a public setting, and it is assumed that most people would adjust their behavior accordingly to that knowledge. Some have further argued that we cannot reasonably expect to maintain privacy over that which another person could discover, overhear, or come to know without concerted effort on his/her part, such as talking in a normal voice in a public place, undressing before a window without shades, or dropping a personal letter on the sidewalk. Conversely, observation in private settings and/or private information is more restricted due to a greater expectation of privacy by subjects.

For internet interactions, there is debate about whether the web should be viewed as a public or private space. Much of this debate revolves around the perceived expectations of subjects within the specific context, that is, a public chat room versus personal email.

Whether a space can be deemed public depends on the expectations of privacy of its members. Kraut et al apply the measure of subjects’ expectations of privacy to show how many online contexts could be considered public.

Many online communication forums have unrestricted membership, allowing anyone who comes by to participate in conversation or observe it... In such cases, we believe that people who post in these groups should have no reasonable expectation of privacy, and researchers and IRBs should be able to treat online communication in them as public behavior.

Where communications are private—instant messages sent to a friend, for example—then researchers are obliged to request consent for access to such information.

Eisenbach and Till offer the following guidelines to judge the level of privacy expected by subjects within different contexts where interaction occurs on the web:

Firstly, if a subscription or some form of registration is required to gain access to a discussion group then most of the subscribers are likely to regard the group as a “private place” in cyberspace. Secondly, the number of (real or assumed) users of a community determines how “public” the space is perceived to be: a posting to a mailing list with 10 subscribers is different from a posting to a mailing list with 100 or 1000 subscribers. Thirdly, and perhaps most importantly, the perception of privacy depends on an individual group’s norms and codes, target audience, and aim, often laid down in the “frequently asked questions” or information files of an internet community.

Under the above definitions, a peer-to-peer file-sharing network could reasonably be viewed as a public space: the programs used to access the network are available to anyone who has internet access, many free of charge; the nature of a public file sharing network entails open sharing of information between all users; and there are millions of users. Therefore, observation of such
networks and the information found within these is arguably equivalent to observation of behavior in a public space.

Is there a reasonable expectation of privacy?
The sharers posted the files to a public space and thus the sharers volunteered the private information found within those files. An argument can be made that in this case, the sharers chose to decrease the relative amount of privacy for the information under their control by posting it on the network and failing to maintain its privacy. The reasonableness of any expectation that the privacy of this information will be observed is thereby decreased as a result of the sharers’ actions.57

In our study, the intentions of the sharers are unknown to us. They may have deliberately made the documents available, but they may not have known their contents or the sensitivity of their contents. For example, a sharer may not have realized that a particular set of files contain sensitive PHI about some other people when they posted them (eg, family members or employees in a parent’s workplace). Therefore, the sharing may have been deliberate, but not necessarily well informed. It is also possible that the sharers were aware of the information’s sensitivity but were inadvertently sharing the files (ie, unknowingly sharing all of the files in a particular directory—see box 1). Another possible scenario is that a sharer was both aware of the information’s sensitivity and still knowingly shared the files online. This scenario is plausible given that individuals post information, some personal, on the public internet (eg, newsgroups), expecting that it will be kept private or circulated only among a small group of similar-minded individuals.58 Such expectations are clearly misplaced.57

Without contacting each individual, we cannot be certain of the intentions of any given sharer. It should also be noted that there is a further tension between viewing a peer-to-peer file sharing network as a public space, and the assertion that sharers may not be fully aware that they are disclosing personal information within that public space. Sharers may know that they are in a public space and therefore have reduced expectations of privacy and may adjust their behaviors accordingly. However, because they believed that they had not disclosed any personal information, they may still have some expectations of privacy on that personal information.

Is the expectation of privacy of individuals a relevant consideration? In law, the concept of a reasonable expectation of privacy arises most often in the constitutional context where the issue is whether an individual has a reasonable expectation of privacy vis-à-vis the state.63 84 Thus the issue of reasonableness goes to balancing individual interests in privacy against the interest of the state in law enforcement, or national security. The concept of “reasonable expectation of privacy” also sometimes arises in the context of lawsuits for invasion of privacy. In these contexts courts consider whether an individual’s right of privacy was intentionally infringed by another; what was reasonable to expect in the circumstances is taken into account by the courts. In both tort law and in the US constitutional context, courts have found either no expectation of privacy or a diminished expectation of privacy in relation to activities taking place in public space.65—69

In the data protection context (where a set of rules govern the collection, use, or disclosure of personal information by public or private sector actors), the concept of reasonable expectation of privacy is less applicable because these regimes tend to be consent-based. Thus the issue is not what is generally considered reasonable in the circumstances; but rather, what the subject has or has not consented to, and whether express consent was required in the circumstances.

Our situation fits within the data protection context, as information is being collected by researchers for research purposes. There is no state actor that would attract constitutional attention. As for invasion of privacy, it would be difficult to bring the research being carried out within the tort framework. In any event, the tort in both Canada and the USA typically permits the use of information for fair comment or public interest purposes (see Privacy Act 199660; and in the USA, the First Amendment protection for free speech gives broad leeway for fair comment and media reporting). If indeed this is a question of data protection, then the central issues become whether there is consent to the collection and use of the data, or whether the collection and use of the data falls into an accepted exception to consent. Other issues will include whether appropriate safeguards for the protection of the data are in place.

Research ethics guidelines implicitly recognize that a component of research ethics includes basic data protection principles. Thus, compliance with research ethics norms is a key component of ensuring that the collection and use of the data is ethical. Research ethics guidelines address consent, collection or use without consent, and the control and safeguarding of data so as to protect subjects.

Is it necessary to obtain consent?
Within contemporary ethical guidelines for observational research, it is given that observational studies of subjects in public spaces and/or the collection of publically available data do not require consent from subjects as stipulated in national and professional ethics guidelines.49 60 70 Therefore, even if the information is considered to be private in nature, its presence in a public “space” obviates the need to obtain consent for research purposes.

Take the example of linguistics research offered by Herring.72 She points to a study by Zimmerman and West (1975) in which they secretly recorded conversations between couples in a public setting. “The conversations overheard by Zimmerman and West were never intended as public; they were private conversations between couples that happened to take place in public settings”.72 However, it was acceptable for the researchers to carry out their study without consent due to the fact that the information was discovered within a public context, and because the identities of the speakers were not revealed.72 Like in the Zimmerman study, any identifying information discovered in the present study would be kept confidential and would not be revealed in our analysis and reporting of the data. Herring also carries this argument into an online context when discussing her research on computer mediated communication (CMC). “Much CMC, such as that on Usenet newsgroups and on open-subscription listservs, resembles Zimmerman and West’s conversations in public places—researchers can easily ‘overhear’ it, although they may not have been the intended audience”.72

Another argument is to focus not solely on the “public space” as a marker of consent, but rather to place the emphasis on the research purposes and protocol under which the information, found in a public space, is used. In other words, the studies cited above can stand for the proposition that personal information which is disclosed into public spaces, regardless of the intention of the data subjects, may be used by researchers where: (a) their research expressly involves a study of these inadvertent public communications; and (b) appropriate ethical safeguards are in place to protect the personal information of these individuals within the research protocol. This approach recognizes that for
some kinds of research consent may not be practicable—a fact which is expressly recognized in some data protection statutes.\textsuperscript{79–75}

Is it practicable to obtain individual consent a priori?

We would not know in advance which computers have PHI on them, therefore it is not possible to seek targeted consent before actually examining the documents. Furthermore, we would not know before actually running the queries on the peer-to-peer network and downloading the files, which computers would have files with PHI (eg, not all computers are on all the time, therefore we may not be able to access the computers that have a match when we attempt to download the files). Practically we would not be able to obtain individual consent in advance.

There is another consideration with obtaining consent, even if it was practical: people change their behavior when they believe that they are being observed by another. This Hawthorne effect, or the principle of reactivity as it is also called, is characterized by people changing their behavior as the result of an awareness that they are being observed.\textsuperscript{76 77}

In a public setting, people act or do not act in certain ways according to what they deem to be acceptable in such a setting. Through the process of informing subjects of the purpose of a research project, even public behavior could be modified further by subjects in accordance with their awareness of the specific behaviors under study.\textsuperscript{56} For example, peer-to-peer file sharing users may be alerted to the possible sharing of sensitive information and then search and remove such documents from their computers, or discontinue file sharing altogether given information about research on the types of documents being shared. This is another reason why, it has been argued, such observational research ought to be exempt from consent requirements.\textsuperscript{49}

As Bakardjijuva points out, some types of research “are hard or impossible to reconcile with seeking informed consent” due to the fact that informing subjects of the goals of research would “have changed their behaviors substantively”.\textsuperscript{76} In the context of linguistics research: “the problem of how to collect authentic data without the collection process interfering with the phenomena observed” is ever present.\textsuperscript{72} Covert tape recording is often utilized in linguistics research to combat this problem. Herring goes so far as to argue that “covert tape recording may be justified even in private contexts, for example, if the knowledge that they are being recorded could make speakers self-conscious to the point of not producing the linguistic phenomena under investigation”.\textsuperscript{72}

Observing communication or interaction online in chat rooms, newsgroups, listservs, and other areas without participating is often called “lurking”.\textsuperscript{79} Many researchers have employed this method to unobtrusively observe interactions between people in an online setting. Herring, for example, has carried her linguistics research online to study computer mediated communication found on message boards,\textsuperscript{80} within weblog communities,\textsuperscript{81} and even in online gaming.\textsuperscript{82} Generally, for such observation she would not seek consent or inform subjects of the study beforehand as she is looking to uncover language that is “produced naturally (ie, by online discourse participants for their own purposes)”.\textsuperscript{83} According to the principle of reactivity, informing subjects that they are being observed would affect their behavior and would make studying natural language production impossible.

Another example of such unobtrusive observation, similar to our study in that it involved peer-to-peer networks, is a study by Mehta et al, examining pornographic videos found on peer-to-peer file sharing networks. This study was based on Mehta’s previous research concerning pornography found in online Usenet groups.\textsuperscript{83 84} His methods have been deemed to be lurking by others,\textsuperscript{79} in that he and his colleagues obtained the files under study without notifying the users within these groups that a study is taking place.\textsuperscript{85}

Some researchers have also chosen to interact with subjects online, while not revealing to the users that they are researchers. Such deception raises ethical concerns that are not presented with mere observation of online behavior, but these methods may be justifiable in certain circumstances when gaining consent would affect the responses of potential subjects and/or make the study impossible. For example, in a study by Glaser et al, researchers studied online white supremacist groups through public IRC chat rooms supported by such groups.\textsuperscript{66} The researcher entered the chat rooms as a visitor and engaged in conversation with the users, eventually conducting a semi-structured interview with an individual respondent whom the researcher was able to engage in conversation. The interview consisted of different scenarios which would presumably be threatening to a supremacist (eg, interracial marriage), presented in either a personal, local, or national context. The researchers did not obtain informed consent prior to the interviews as they believed that it would impede their ability “to gather candid responses without raising suspicion”.\textsuperscript{65} The Yale human participants committee, to whom the study protocol was submitted, agreed that consent would impede the study in that “respondents would have been very unlikely to participate, that those who did would not have been representative, and that responses would have been significantly biased”.\textsuperscript{65} Because this took place in a public forum, the subjects were not pressured into participating, the interview topics were common topics in this chat room, and the identities of the participants were protected, the Yale committee deemed that it was acceptable not to seek informed consent.\textsuperscript{66}

In our study, there is a tangible risk that if we inform the users of the purpose of the study prior to data collection, they could conceivably change their behavior on the basis of this information. In previous research on the inadvertent disclosure of personal information, informing the subjects resulted in the removal of that information off the internet.\textsuperscript{87} Furthermore, given the controversy over the high volume of copyrighted materials being shared in peer-to-peer file sharing networks, and the risk of users being charged with piracy and found financially liable,\textsuperscript{88–90} it would be almost certain that sharers in this context would be wary of any research taking place on sharing habits and alter or cease their participation in the network as a result. The purpose of our study, to assess the extent to which this data was being disclosed within file sharing networks, would most definitely be compromised as a result of any such changes in behavior.

Can we notify the affected individuals that we have downloaded their personal information?

One option is to notify the owner of the computer that we downloaded the information about the existing PHI. However, it is not obvious that we can determine the identity of the computer owner. One piece of information we have that can potentially be used to determine the identity of the computer owner is the IP address.

In some jurisdictions, IP addresses are regarded as personal information because it is argued that they can potentially identify an individual. For example, in the US Health Insurance Portability and Accountability Act Privacy Rule the removal of the IP address is required to claim that a data set is de-identified according to the Safe Harbor list.\textsuperscript{91} An opinion from the European Article 29 Working Party contends that IP addresses can be
Many externally visible IP addresses are re-mapped to
an internet service provider would know the identity of
a household from an IP address in use at a particular time.

- Often users at home are assigned a temporary IP address for
  a limited period of time through DHCP (Dynamic Host
  Configuration Protocol, which assigns IP addresses dynami-
  cally for specified periods of time). When that address expires,
  they are assigned another one. Therefore, in principal only
  the internet service provider would know the identity of
  the computer owners or subjects, nor to reliably inform them in
determinate violations of the rights of

- It has been argued that the IP address is identifying
  information because law enforcement or a private party can
  compel an ISP to reveal the physical address associated with
  an IP address through the courts. In our case we will not be
  seeking to identify the physical addresses that way,
  therefore the only reasonable mechanism for identification
  will not be used.

- Many externally visible IP addresses are re-mapped to
  machines in an internal network through NAT (Network
  Address Translation). For example, a company may have
  a single external IP address, but behind their firewall there are
  many machines on their network. To an external entity like
  us, all of those machines will appear as a single IP address.
  Therefore, it is not possible to determine from only the single
  external IP address which specific machine the files or searches
  came from.

Any attempt to communicate with an individual identified
through an IP address, given the above uncertainty, may result in
us revealing sensitive information to the wrong person. For
example, if we inform an individual that we had found their
mental health records on the peer-to-peer network, we might
realize later that this was not the individual’s computer and that
actually we had informed the individual’s parents who did not
know that their child had mental health records. In such a case
we would engage in a breach of privacy in an attempt to notify.

Furthermore, if we attempt to notify the exposed subjects
themselves, we may inadvertently cause a breach of privacy similar
to the one noted above. For example, if we have a
medical record about Mr Smith and are able to determine his
address, an attempt to contact him directly may inadvertently
reveal to other members of his household that he has a medical
record of which they were not aware. Or, Mr Smith may no
longer live at that address (it may be his parents’ address or his
ex-wife’s address) and notifying him at that address may inad-
vertently reveal his PHI to the current inhabitants.

However, through the publication of the study and making its
results broadly available, we hope to increase public awareness
of the risks of file sharing applications. Furthermore, in the
Discussion section we provide some suggestions on risk reduc-
tion and management.

Summary
We therefore contend that there is a minimal risk of harm for
this study, that it is not possible to reliably get consent from the
computer owners or subjects, nor to reliably inform them indi-
vidually after the fact. In such circumstances ethical obligations
are met through the measures put in place to safeguard any
personally identifiable information that is collected in the course
of the study, and to ensure that no subjects are individually
identifiable in the dissemination of the research results.

Data security and data destruction
After the study, all copies of files with personal information
collected for the purpose of this study were saved on an
encrypted DVD and stored in a locked safe at the Children’s
Hospital of Eastern Ontario Research Institute with access to
only two co-authors (KEE and EJ) and a member of the research
ethics board of the institution. After 12 months from the
publication of this paper, we undertake to destroy all of the data
with personal information, and all copies of it, that has been
collected for the purpose of this study. The data destruction will
be performed under the supervision of our institutional research
ethics board, which will have a delegate to witness that the data
destruction has occurred and a letter certifying that the data
destruction has occurred will be sent to the Journal.

RESULTS
Data files were downloaded from 807 Canadian IP addresses and
344 US IP addresses. Approximately 1% of all downloaded files
were viruses in Canada, as were 2.1% in the USA. All of the
viruses were Trojan horses opening back doors to the computer
and allowing an external entity to drop potentially malicious
files or control the machine.

The 2-rater reliability analysis on a subset of 64 files had a k
value of 0.63 (p<0.0001) for PII, 0.71 (p<0.0001) for PHI, and
0.857 (p<0.0001) for PFI.

We estimated that 10.9% (95% CI 9.3% to 13.6%) of the IP
addresses had PII in Canada, as did 7.1% (95% CI 5.6% to 8.9%)
in the USA. The proportions and CIs for PHI and PFI are shown
in figure 1. We can see that the proportion of IP addresses with
PHI was relatively low, at 0.5% of all addresses in Canada, and
0.4% in the USA. The regional difference was not statistically
significant. PFI was much more readily available at a prevalence
rate of almost 2% in Canada and 5% in the USA. The difference
between the USA and Canada on PFI was significant by a χ² test
(p=0.0007). Within regions, the difference between PHI and PFI
was significant in Canada (p=0.0329) and the USA (p<0.0001)
by a χ² test.

Examples of files containing PHI are medical authorization
forms for minors detailing their medical histories (eg, for chil-
dren going to camp) and personal health assessment forms.
There was also a statement under oath describing a knee injury
plus other medical conditions of a US soldier before deployment
to a specific overseas base of operation. Files containing PFI
included documents with banking details, such as account
numbers and passwords, credit card numbers, tax return docu-
ments, and documents related to personal bankruptcies.

During our data capture period, approximately 3.5 million
search terms were logged. We found evidence that there were
individuals looking for these files (see table 2 for counts of the
number of times these search terms were logged). Three search
terms were used that matched the PHI files, and eight terms
matched the PFI files. Despite their simplicity, the search terms
were quite effective in returning sensitive documents. Again, we
see that searches for PFI were more frequent than searches for
PHI. Out of all of the search terms, the proportion that returns
PHI and PFI is relatively small. Most search terms were for music
files and pornography.

DISCUSSION
Summary
Approximately 7–11% of the IP addresses were disclosing
documents with PII. Our prevalence rates for PII on file sharing
networks are smaller than the 49% and 61% previously

---


Research paper

---

considered, for practical purposes, as identifying information.92
The German data protection commissioner has advocated that IP
addresses be treated as identifying information.93 A similar view
prevails in Canada as established by the courts.94 However, for
the purpose of our study the IP address will not necessarily allow
us to determine the identity of computer owners for a number of
reasons:

- Often users at home are assigned a temporary IP address for a
  limited period of time through DHCP (Dynamic Host
  Configuration Protocol, which assigns IP addresses dynami-
  cally for specified periods of time). When that address expires,
  they are assigned another one. Therefore, in principal only
  the internet service provider would know the identity of
  the computer owners, we may inadvertently cause a breach of privacy
  in an attempt to notify.

- It has been argued that the IP address is identifying
  information because law enforcement or a private party can
  compel an ISP to reveal the physical address associated with
  an IP address through the courts. In our case we will not be
  seeking to identify the physical addresses that way,
  therefore the only reasonable mechanism for identification
  will not be used.

- Many externally visible IP addresses are re-mapped to
  machines in an internal network through NAT (Network
  Address Translation). For example, a company may have
  a single external IP address, but behind their firewall there are
  many machines on their network. To an external entity like
  us, all of those machines will appear as a single IP address.
  Therefore, it is not possible to determine from only the single
  external IP address which specific machine the files or searches
  came from.

Any attempt to communicate with an individual identified
through an IP address, given the above uncertainty, may result in
us revealing sensitive information to the wrong person. For
example, if we inform an individual that we had found their
mental health records on the peer-to-peer network, we might
realize later that this was not the individual’s computer and that
actually we had informed the individual’s parents who did not
know that their child had mental health records. In such a case
we would engage in a breach of privacy in an attempt to notify.

Furthermore, if we attempt to notify the exposed subjects
themselves, we may inadvertently cause a breach of privacy similar
to the one noted above. For example, if we have a
medical record about Mr Smith and are able to determine his
address, an attempt to contact him directly may inadvertently
reveal to other members of his household that he has a medical
record of which they were not aware. Or, Mr Smith may no
longer live at that address (it may be his parents’ address or his
ex-wife’s address) and notifying him at that address may inad-
vertently reveal his PHI to the current inhabitants.

However, through the publication of the study and making its
results broadly available, we hope to increase public awareness
of the risks of file sharing applications. Furthermore, in the
Discussion section we provide some suggestions on risk reduc-
tion and management.

Summary
We therefore contend that there is a minimal risk of harm for
this study, that it is not possible to reliably get consent from the
computer owners or subjects, nor to reliably inform them indi-
vidually after the fact. In such circumstances ethical obligations
are met through the measures put in place to safeguard any
personally identifiable information that is collected in the course
of the study, and to ensure that no subjects are individually
identifiable in the dissemination of the research results.
reported\textsuperscript{37–39} for three reasons: (a) we were estimating the proportion of IP addresses that were disclosing PII, whereas previous work reported the proportion of documents that contained PII; (b) earlier studies did not necessarily download the documents to examine them for PII but assumed that they do contain it based on type or name; or (c) earlier studies used targeted searches for documents about or belonging to specific organizations (eg, by using bank names in search queries) or specific domains (eg, medical). Each IP address may have multiple documents in them, therefore our reporting in terms of IP addresses, reason (a), will by definition be smaller than reporting in terms of documents. In case (b) the previously reported proportions are likely to be higher than ours because we did download and examine all documents and inevitably some documents that sound like they should contain PII do not. In case (c), we used a general or wildcard search, which would by definition result in a much lower prevalence rate than a targeted search because of the larger denominator, especially if the target is very likely to contain PII (eg, if bank and hospital documents are specifically targeted).

Our results indicate that a relatively small percentage of IP addresses disclose PHI. There were no significant regional differences in that rate. However, if we consider that tens of millions of people use peer-to-peer file sharing applications in North America, a rate as small as 0.5% indicates that overall there are tens of thousands of IP addresses disclosing PHI. Furthermore, our results suggest that there are searches being conducted on these networks that are specifically targeting PHI and that will successfully return PHI files (ie, they are effective searches). This clearly means that there are people actively looking for and finding documents containing PHI. It is not possible to tell whether these searchers would use the information for any malicious purposes.

Significantly more PFI is being disclosed from Canadian and American IP addresses. A 4.7% disclosure rate in the USA translates into hundreds of thousands of computers containing exposed PFI. Furthermore, there is clear evidence that there are searches being conducted on the peer-to-peer networks specifically targeting PFI and these searches are returning actual files with PFI. The disclosure rate of PFI is significantly lower in Canada.

The lower PHI disclosure rate is not surprising since individuals have more electronic financial information than health information. Therefore, if they are inadvertently exposing files on their computers they are more likely to expose PFI than PHI.

Table 2. Proportion of personal health information (PHI) and personal financial information (PFI) files that were matched using the search terms (note that search terms coming from the same IP, even if they matched multiple files, were only counted once)

<table>
<thead>
<tr>
<th>Search term</th>
<th>No. of matching files</th>
<th>Proportion</th>
<th>No. times search term was used</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tax return</td>
<td>18</td>
<td>0.33</td>
<td>528</td>
</tr>
<tr>
<td>Tax</td>
<td>30</td>
<td>0.56</td>
<td>113</td>
</tr>
<tr>
<td>Credit report</td>
<td>1</td>
<td>0.02</td>
<td>20</td>
</tr>
<tr>
<td>Credit card numbers</td>
<td>2</td>
<td>0.04</td>
<td>1</td>
</tr>
<tr>
<td>Credit card number</td>
<td>2</td>
<td>0.04</td>
<td>16</td>
</tr>
<tr>
<td>Credit card</td>
<td>5</td>
<td>0.09</td>
<td>1</td>
</tr>
<tr>
<td>Bank account</td>
<td>3</td>
<td>0.06</td>
<td>1</td>
</tr>
<tr>
<td>Amex</td>
<td>1</td>
<td>0.02</td>
<td>1</td>
</tr>
<tr>
<td>PHI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient file</td>
<td>1</td>
<td>0.14</td>
<td>1</td>
</tr>
<tr>
<td>Medical form</td>
<td>1</td>
<td>0.14</td>
<td>1</td>
</tr>
<tr>
<td>Medical</td>
<td>4</td>
<td>0.57</td>
<td>7</td>
</tr>
</tbody>
</table>

Figure 1. The proportion of IP addresses that exposed personal health information (PHI) and personal financial information (PFI) in Canada and the USA with 95% CIs.

Practical implications

As more PHI becomes electronic, more PHI would likely become inadvertently exposed. Therefore, there is a real need for peer-to-peer clients to make it easier for users to clearly know which files they are making available on the peer-to-peer networks and whether these files contain PHI or PFI. This would help individuals realize if they are exposing files that may contain sensitive information. The peer-to-peer software industry has produced voluntary guidelines\textsuperscript{97} and some peer-to-peer client tool developers have claimed they made improvements to reduce the opportunities for inadvertent disclosure of personal and sensitive information.\textsuperscript{98–99} However, doubts have been raised about how seriously the industry is implementing software features to reduce inadvertent disclosure of files.\textsuperscript{100–101} There are at least five possible reasons why software improvements may not have a large effect: (a) the recommended software improvements themselves are not
Box 2 Some recommendations for managing risks from inadvertent disclosure risks from peer-to-peer file sharing clients

The following are methods that can be used to protect yourself and/or your organization against inadvertent file sharing through peer-to-peer clients.

Avoid or block peer-to-peer clients
The following recommendations would help ensure that no peer-to-peer file sharing clients are installed and running on your machines and network(s):

- Educate your users about the risks from file sharing clients, how to recognize these programs, and about why they should not be installed.
- All non-administrative users on shared computers must have separate accounts without the ability to install software themselves. Do not have one shared account that everyone uses because then all users pay for any single user’s misdeeds. Separate accounts with minimal privileges ensures that users cannot install a peer-to-peer file sharing client at all, and if they somehow are able to, the underlying file system will not permit them to share other users’ files (e.g., a user would not be able to share another user’s “My Documents” folder).
- Some advanced anti-virus software can be set to detect the signature of files that are being installed or that are being executed on the machine, and can stop them. These would also alert an administrator if such programs are installed or executed. It is advisable to use such tools and set them to block peer-to-peer clients.
- Some peer-to-peer clients use fixed ports to communicate, and for these clients a basic firewall can be used to block traffic on the specific ports they use. However, many of the popular clients will scan ports until one is available to transfer control and data packets, will use standard ports (e.g., use port 80 which is normally used for HTTP traffic), and/or will use standard protocols (e.g., HTTP) to transfer data. In that case, more sophisticated intrusion detection and protection systems that can perform deep packet inspection, or analyze packet flow patterns, and detect the protocol and nature of the traffic are needed to block peer-to-peer file sharing packets.

Manage risks from peer-to-peer clients
If it is necessary to use peer-to-peer file sharing clients, then the following are recommendations for managing the risks:

- Do not put sensitive data on the computer with the peer-to-peer file sharing application on it, nor give it access to sensitive data on a shared network resource. This means that any machine with file sharing enabled on it would be treated as inherently insecure and untrusted.
- If it is absolutely necessary to use programs that share files on peer-to-peer networks, it is probably least risky to use one of the clients that have received significant congressional, researcher, and non-governmental organization scrutiny (e.g., Limewire) and that have been developed by US-based companies. The continuous attention will make it more difficult for these companies to continue having or adding features which encourage inadvertent file sharing.
- If file sharing among specified and known individuals or team members is required, then use an application such as Groove, which allow peer-to-peer file sharing but not on an open network. These provide tighter control on whom data is shared with, and all communication among peers is secured.

effective in reducing inadvertent sharing of files or are simply not implemented or retained in all subsequent releases of the software; (b) the features which encourage inadvertent sharing are replaced with equally damaging features; (c) many users already have older and more permissive versions of the peer-to-peer file sharing clients and will not upgrade to the newer, and possibly more restrictive, ones; (d) not all vendors have implemented the suggested improvements (it is estimated that there are 225 different peer-to-peer file sharing clients, and many still use aggressive tactics to promote sharing of information); and (e) it is possible that the main reasons for inadvertent sharing was not the usability of the tool but another behavioral mechanism (e.g., sharers not being aware of the content or the sensitivity of the content they were sharing).

Until additional convincing evidence emerges as to the effectiveness of the file sharing software improvements, it is advisable that healthcare providers not install any peer-to-peer file sharing applications on computers that contain sensitive personal information about their patients. Additional recommendations for managing the risks from such applications are provided in box 2.

Limitations
Our study did not consider BitTorrent clients. These represent a different protocol for peer-to-peer file sharing. Therefore, our results only represent PHI disclosure risks in one part, albeit a large one, of the file sharing universe.

CONCLUSIONS
The purpose of our study was to estimate the extent to which PHI was disclosed on peer-to-peer file sharing networks. We found that around 0.5% of IP addresses were disclosing PHI in the USA and Canada. This was significantly less than the amount of PFI that was being disclosed in both countries. However, given the number of users of such file sharing programs, 0.5% still represents tens of thousands of IP addresses exposing PHI in Canada and the USA.

Some of the files that were discovered included very sensitive medical and financial information about individuals. It is not likely that these individuals deliberately shared these files. It is more likely that they inadvertently made this information available through a misconfiguration of their file sharing client programs, a misunderstanding of how they work, not knowing the contents of the files, or misinformation/misunderstanding about the risks of sharing.

As more health information gets digitized, it is expected that the amount of health information available to individuals on their personal computers will increase. Therefore, it is most likely that the rates of PHI disclosure through peer-to-peer file sharing networks that we obtained will rise over time.

Acknowledgments
We wish to thank Bradley Malin (Vanderbilt University) and Fred Carter (Office of the Privacy Commissioner/Ontario) for giving us feedback on an earlier version of this paper.

Competing interests
None.

Ethics approval
This protocol was approved by the Research Ethics Board of the Children’s Hospital of Eastern Ontario Research Institute. The CHEO RI Research Ethics Board has a Federal Wide Assurance (FWA) certificate (FWA00003131) from the Department of Health and Human Services in the USA (see http://ohrp.nih.gov/search/ and search for FWA # 00003131). An FWA certificate formalizes the institution’s commitment to protect human subjects, including compliance with US laws, regulations, policies, and guidelines related to the conduct of research on human subjects (http://www.hhs.gov/ohrp/humansubjects/assurance/fisawert.html). This board also follows the Canadian Tri-Council Policy Statement on Ethical Conduct for Research
Involving Humans (http://pre.ethics.gc.ca/eng/policy-politique/tcp-epct/), which is produced by the three main research funding agencies in Canada: the Canadian Institutes for Health Research, the Natural Sciences and Engineering Research Council, and the Social Sciences and Humanities Research Council.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

11. Harristinteractive. Health information privacy (HIPAA) notices have improved public’s confidence that medical information is being handled properly. 2005.
59. U.S. v. Mayes. 507 F 3d 1013 (9th Cir. 2007).
63. Privacy Act (B.C.), S.B.C. 1996, p. c. 373, s. 2(6a).
92. Claburn C. European regulators mull protecting IP addresses information week; 2008.
102. Sydnor T, Knight J, Hollaar L. Filesharing programs and “Technological features to induce users to share”: United States patent and trademark office; 2007.
103. Sydnor T, Knight J, Hollaar L. Filesharing programs and “Technological features to induce users to share”: United States patent and trademark office; 2007.
104. Sydnor T. Inadvertent file-sharing re-invented: the dangerous design of Limewire 5: the progress & freedom foundation; 2009.