Visualization for Privacy Compliance[1]

George Yee
National Research Council Canada
1200 Montreal Road, Building M-50
Ottawa, Ontario, Canada K1A 0R6
1-613-990-4284
george.yee@nrc.ca

ABSTRACT
The growth of the Internet has been accompanied by the growth of e-services (e.g. e-commerce, e-health). This proliferation of e-services has put large quantities of consumer private information in the hands of the service providers, who in many cases have mishandled the information, either intentionally or unintentionally, to the detriment of consumer privacy. As a result, government bodies have put in place privacy legislation that spells out a consumer's privacy rights and how consumer private information is to be handled. Providers are required to comply with such privacy legislation. This paper proposes visualization as a tool that can be used by security or privacy analysts to understand how private information flows within and between provider organizations, as a way of identifying vulnerabilities that can lead to non-compliance. A model of private information flow and a graphical notation for visualizing this flow are proposed. An application example of using the notation to identify privacy vulnerabilities is given.

Categories and Subject Descriptors
C.2.0 [**General**]: Security and Protection (e.g., Firewalls)

General Terms
Security

Keywords
privacy, compliance, visualization, privacy legislation, e-services

1. INTRODUCTION
A large number of e-services targeting consumers has accompanied the rapid growth of the Internet. For example, e-services are available for banking, shopping, learning, healthcare, and Government Online. However, each service requires a consumer's private information in one form or another. This has led to a large amount of consumer private information being in the possession of service providers along with the accompanying concerns over potential loss of consumer privacy. Experience has shown that such concerns have not been unfounded, as providers have yielded to the temptation of, for example, employing the private consumer data for marketing purposes (e.g. selective advertising based on personal buying habits) and for additional financial gain (e.g. selling a consumer's contact information and buying habits to another provider).

In response to the above mentioned privacy concerns, various government jurisdictions have enacted privacy legislation that identify a consumer's privacy rights in terms of how providers must treat private information. For example, in the United States, the Health Insurance Portability and Accountability Act (HIPAA) [1] requires compliance by health care providers to "Standards for Privacy of Individually Identifiable Health Information" (the Privacy Rule). The Privacy Rule sets national standards for the protection of health information for health care providers who conduct health care transactions electronically. By the compliance date of April 14, 2003, such providers must have implemented measures to comply with the Privacy Rule. Failure to comply may trigger the imposition of civil or criminal penalties [2].

This paper describes preliminary research at using visualization techniques to understand how private information flows within and between provider organizations. The goal is to use visualization to more effectively identify privacy vulnerabilities, where an organization's compliance to privacy regulations may be compromised. To achieve this goal, privacy legislation is first examined to understand some of the requirements for compliance. Then a model of private information flow and a graphical notation for visualizing this flow are proposed. An example of using the notation to model private information flow and identify private information vulnerabilities that can lead to non-compliance is given.

Section 2 looks at privacy legislation to understand some of the requirements for compliance. Section 3 presents a graphic model for visualizing private information flow, and an example of using the model to identify privacy vulnerabilities. Section 4 presents an evaluation of the results of this paper. Section 5 discusses related works. Section 6 gives conclusions and future work.

2. PRIVACY LEGISLATION
To visualize private information flow in order to identify vulnerabilities that lead to non-compliance, one needs to understand which vulnerabilities pose risks of contravening privacy legislation. In other words, it is necessary to know what privacy legislation requires for preserving personal privacy. This may be a very daunting task, since legislation is in general framed in obscure legal terminology (obscure to non-lawyers anyway) and is usually spread out in many volumes of works. Fortunately, such legislation has been summarized in terms of privacy principles or summaries of rights that are easier to work with. Yee et al. [3] examined legislation-derived privacy principles or

VizSEC'06, November 3, 2006, Alexandria, Virginia, USA.

ACM 1-59593-549-5/06/0011.

rights from the United States (for HIPAA), Canada (national legislation), and the European Union (European Union legislation), and found that in terms of how a provider treats private information in its possession, all principles and rights boil down to the consumer's right to know and agree to:

- *Collector*: who collects the private information,
- *What*: what is the private information,
- *Purposes*: what the private information will be used for,
- *Retention Time*: how long the private information can be retained by the provider,
- *Disclose-To*: the party or parties to whom the provider can disclose the private information.

In addition, the provider has to perform certain maintenance level actions such as allowing the consumer to access and update her ("her" and "she" are used to stand for both sexes) information, be accountable for the information, accept challenges from the consumer regarding proper conduct, and very importantly, ensure that the information is kept safe from malicious attackers who want to compromise the information.

Key questions to ask at this point are: "How does the possibility of non-compliance or failing to uphold these rights and responsibilities manifested in terms of vulnerabilities?" and "Would visualizing private information flow be able to identify these vulnerabilities?" Of course, the answer to the second question depends on the graphic model used for visualization. However, it is useful to set reasonable limits on what is required of the graphic model by ruling out certain visualizations a priori. Table 1 attempts to answer these two questions.

Table 1. Vulnerability identification through visualization

Right or Responsibility	Vulnerability	Identify by Visualization
Collector	Potential to have a different collector	Yes
What	Potential to collect different information	Yes
Purposes	Potential to use the information for a different purpose	Yes
Retention Time	Potential to violate the retention time	Yes
Disclose-To	Potential to disclose to different party	Yes
Consumer Update	Lack of mechanism or procedure, possible failure of such mechanism or procedure	No
Accountability	Same as for "Consumer Update"	No
Consumer Challenge	Same as for "Consumer Update"	No
Security	Potential for malicious attack	Yes

In Table 1, "Identify by Visualization" means that one would be able to conclude by looking at the graphic that the vulnerability is present or absent. For example, in the cases of "Collector" and "Disclose-To" the graphic representing the flow could indicate (or not) that there is a potential for the flow to be redirected to the wrong collector. In the cases of "Consumer Update", "Accountability", and "Consumer Challenge", it is difficult to see how these could be more advantageously indicated graphically rather than simply by policy.

3. A MODEL FOR VISUALIZING PRIVATE INFORMATION FLOW

This section presents a model and notation for visualizing private information flow between the consumer and the provider, as well as within and between providers. The design of the model and notation will follow Table 1 in terms of what vulnerabilities can be visualized.

3.1 MODEL AND NOTATION

The model and notation, called a "Private Data Flow Chart (PDFC)", is based on Data Flow Diagrams that was popular in the 1970's and 1980's in the context of structured programming [4]. The following components are defined:

- *Regular Process*: a unit of private data processing; private data enters a regular process and is consumed or used to produce other data which leaves the process; represented by a circle with a single line border,
- *Composite Process*: construct for hiding a private data sub-chart to reduce visual complexity; may be replaced by the sub-chart it is hiding; represented by a circle with a double line border,
- *Computing Hardware*: computing platform; contains one or more regular processes that run on the platform; represented by a hexagon,
- *Private Data Flow*: flow of private data, e.g. between processes; represented by an arrow; a dashed arrow is used together with a "from" or "to" label where it is clear what the flow is from a previous chart,
- *Private Data Store*: a location where private data comes to rest; represented by two parallel lines,
- *Consumer*: customer or client of a provider; represented by an oval with a single line border,
- *Provider*: supplier of services to a consumer or to another provider; hides the PDFC of the provider; represented by an oval with a double line border.

The above covers the major components of the model. In addition, there are further corresponding elements describing the nature or type of the major components, as follows:

- *Description of Process*: the processing carried out in a regular process; indicated by annotations on the circle of a regular process,
- *Description of Composite Process*: the processing done in the process's hidden sub-chart; indicated by annotations on the circle of a composite process,
- *Type of Computing Hardware*: the type of computing platform, e.g. Windows PC; indicated by annotations on the hexagon representing computing hardware,
- *Description of Private Data Flow*: what the private data is, e.g. credit card number; indicated by annotations on the data flow arrow,
- *Type of Connection*: the type of connection carrying the private data flow, e.g. Ethernet; indicated by annotations enclosed in square brackets on the data flow arrow,

- *Type of Data Store*: the type of data store, e.g. flat file; indicated by annotations on the data store parallel lines,
- *Consumer ID*: the ID of the consumer, to distinguish her from other consumers; indicated by annotations on the consumer oval,
- *Provider ID*: the ID or service of the provider, to distinguish it from other providers; indicated by annotations on the provider oval.

In addition, to distinguish the PDFCs of different providers, a rectangle with a dashed border is used to enclose the PDFC of a provider. The name of the provider is placed on top of this rectangle. Figure 1 illustrates the application of this model. In Figure 1, a client supplies her private information to Books Online to purchase books. The PDFC of Books Online contains 3 processes and a data store running on 3 computing platforms. The Client Interaction process deposits the client's private information into the database, from which the Payment and Shipping processes each retrieve portions of the information needed for their individual functions.

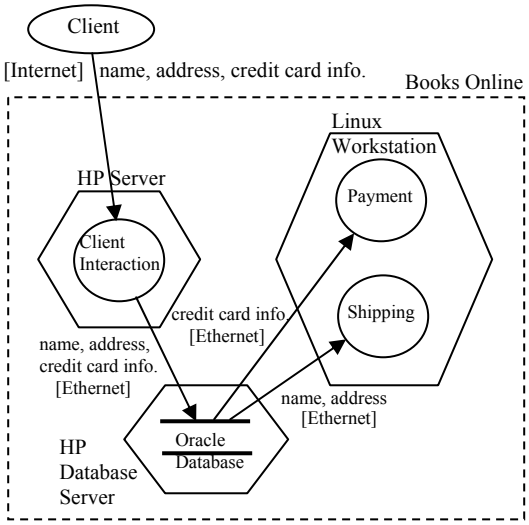

Figure 1. PDFC for Books Online

3.2 Application Example

Consider Drugs Online, an online pharmacy (see Figure 2) that uses the services of Pay All, a payment provider, and Ship Anywhere, a shipping provider, for its products, i.e. these providers take care of receiving payment from the pharmacy's client and shipping drugs to the pharmacy's client, respectively. In addition, the online pharmacy uses the services of Drugs Supply Wholesaler, a supplier of drugs, to replenish its stock. The top left chart in Figure 2 shows the entire system of providers for this example. Each provider in this chart is then decomposed into its PDFC in the remaining charts. Note that "account" as shown in the system chart and the PDFC for Drugs Online refers to the account of Drugs Online, for debit purposes by the providers that supply Drugs Online. Further, the "drug order" in the PDFC for Drugs Online is Drugs Online's private information, not the consumer's. Examining these charts, one can obtain the vulnerabilities shown in Table 2. For example, looking at the system chart (top left of Figure 2), one can see that the client information could flow to another provider masquerading as Drugs Online, leading to the phishing vulnerability. Looking at the Ethernet links within the PDFC for Drugs Online, one can see that they are vulnerable to internal attacks against data confidentiality and integrity.

Table 2. Some vulnerabilities of Drugs Online and its suppliers

Provider PDFC / Compliance Category	Vulnerability
Drugs Online	
Collector	Drugs Online could be spoofed (phishing).
What	The Client Interaction process could be compromised to request other information.
Purposes	This provider may decide to sell the client's information to another party for use in a different purpose. It may copy the data from the database without leaving any trace.
Retention Time	The client's information in the database could be retained after the agreed retention time.
Disclose-To	This could be violated by the information being sold or inadvertently disclosed by falling victim to phishing.
Security	a) phishing as noted for Collector and Disclose-To, b) loss of confidentiality and integrity of data for flows from the Internet, c) external attacks on the processes and platform operating systems since they are linked to the Internet, d) external attacks on the database, e) internal attacks on the processes and platform operating systems, f) internal attacks on the database, g) internal attacks on the Ethenet connections that carry the private information, and h) loss of confidentiality and integrity of data for the flows going out to the providers that supply Drugs Online .
Pay All	
Collector	Pay All could be spoofed (phishing).
What	Same as for Drugs Online.
Purposes	Same as for Drugs Online.
Retention Time	Same as for Drugs Online.
Disclose-To	Same as for Drugs Online.
Security	Same as for Drugs Online but without h).
Ship Anywhere	
Collector	Ship Anywhere could be spoofed (phishing).
What	Same as for Drugs Online.
Purposes	Same as for Drugs Online.
Retention Time	Same as for Drugs Online.
Disclose-To	Same as for Drugs Online.
Security	Same as for Pay All.
Drugs Supply Wholesaler	
Collector	Drugs Supply Wholesaler could be spoofed (phishing).
What	Same as for Drugs Online.
Purposes	Same as for Drugs Online.
Retention Time	Same as for Drugs Online.
Disclose-To	Same as for Drugs Online.
Security	Same as for Pay All.

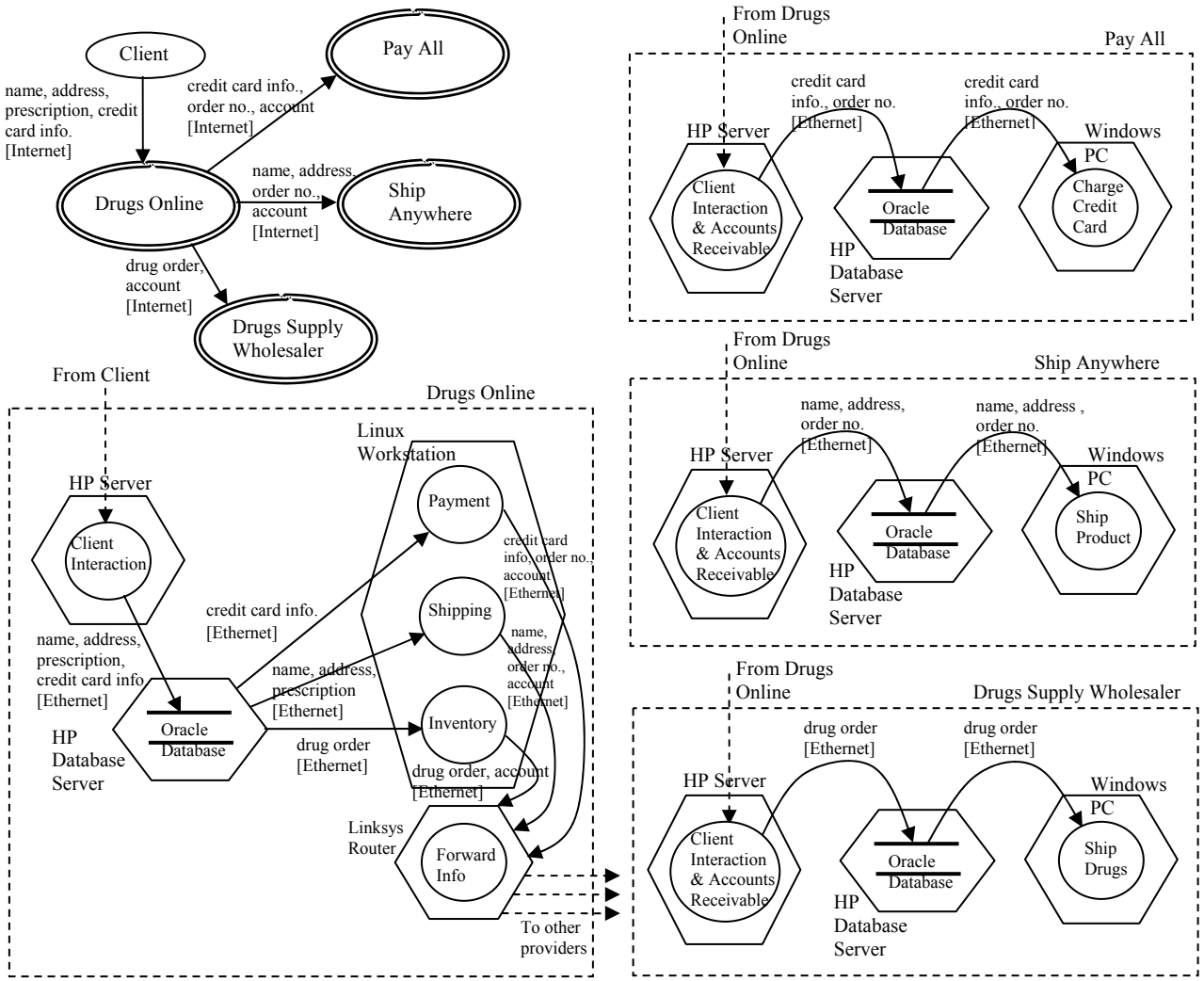

Figure 2. PDFCs for an online pharmacy and its suppliers

The vulnerabilities in Table 2 are somewhat repetitive due to the use of similar components in the PDFCs to keep the example simple. Drugs Online and its suppliers would need to install countermeasures against these vulnerabilities in order to ensure compliance with privacy legislation.

4. EVALUATION OF RESULTS

From personal experience to date, the proposed model and notation are easy to use, as are Data Flow Diagrams, upon which the model is based. It appears that the following characteristics hold for applications of the model and notation:

1. The proposed model and notation does facilitate the identification of private information vulnerabilities that can lead to non-compliance. This is probably due to the fact that the private information flows are mapped out graphically so that one can more easily see the dangerous locations.
2. Not all vulnerabilities can be found. This follows from the facts that a) finding vulnerabilities is a manual process done by error-prone humans, and b) new vulnerabilities appear in concert with the attacker's capabilities, which usually improve over time.
3. The more detailed the modeling, the greater the number of vulnerabilities that will be found. This makes sense since the more details shown, the greater the number of places seen where vulnerabilities may crop up. However, in practice, one may limit the amount of detail modeled due to the size of the system or the pressure of scheduling. Automated tools would be highly useful to speed up the process allowing the inclusion of greater detail.
4. The proposed model and notation appears to work better for identifying security vulnerabilities than for identifying other vulnerabilities. This can be seen in Table 2, where the number of vulnerabilities for the security compliance category is the largest among all the categories and even vulnerabilities from some of the other compliance categories such as Collector and What derive from security vulnerabilities (phishing, insecure software).
5. Finding vulnerabilities using the model and notation requires security/privacy knowledge and skill. This is not unexpected since it is a tool meant for the security/privacy analyst. It should be noted that finding vulnerabilities without this tool also requires such knowledge and skill. It is hoped that the

use of this tool will make finding vulnerabilities easier and more foolproof.

In addition to the above characteristics, there are some open questions that need to be investigated. One question is: "Is modeling the flow of only private information sufficient for identifying vulnerabilities, or would including some other non-private information flows help to identify vulnerabilities?" For example, in the PDFC for Drugs Online, a "synchronization message" from the Payment process to the Shipping process (to not ship before payment is made) was omitted. In this case, including it does not help identify more vulnerabilities to private information, but what about other cases? Another question is whether or not the above model needs to be explicitly linked with known vulnerabilities using, for example, tags and/or overlays. Such linking may facilitate "compiling" the PDFC to obtain an aggregate picture of privacy risks. A third question arises from the observation that workflow models also model the flow of information from one work process to the next. The question then is: "Could workflow models be adapted to model the flow of private information, and once adapted, would they be better suited for identifying private information vulnerabilities leading to non-compliance?"

Finally, one may object to the fact that the proposed approach has used one interpretation of a very particular set of privacy legislations, and therefore the approach is not general. To this objection there are two answers. Firstly, the approach can be "tuned" to any privacy legislation by analyzing the new legislation for compliance requirements. Secondly, the proposed model and notation have intrinsic value in facilitating the identification of threats to private information, whatever the privacy legislation.

5. RELATED WORK

As far as this author is aware, there is no other work that deals with visualization of private information flows in organizations in order to identify vulnerabilities that can lead to privacy legislation non-compliance. Related works follow. Korba et al. [5] describe private data flows between controller and processing components of a privacy rights management architecture whose purpose is to uphold an individual's rights to data privacy. Korba et al. do not, however, graphically depict the flows for any given system, with the purpose of analyzing the flows for privacy vulnerabilities. Their flows are part of a system whose purpose is to ensure compliance to privacy legislation. In addition, there are works making use of visualization for other purposes. Here are three samples. Pillat and Freitas [6] present a system for providing multiple coordinated views of multidimensional data. Users are allowed to set which visualizations they want to coordinate via a diagram showing the different visualizations. Walker et al. [7] describe a technique for visualizing the operation of an object oriented system using dynamic information collected as the system executes. They discuss preliminary qualitative studies into the technique's usefulness. Konyha et al. [8] present an interactive and intuitive 3D visualization framework for rigid body simulation data. The authors report that they have integrated their visualization technique into an application developed at a leading company in automotive engine design and simulation, for engine chain and belt driven timing drives. These three samples demonstrate the diverse applications of visualization. Of course, a related work is again the work on Data Flow Diagrams [4] upon which this work's proposed model and notation are based.

6. CONCLUSIONS AND FUTURE WORK

This work has proposed an effective model and graphical notation to facilitate the identification of private information vulnerabilities that can lead to privacy legislation non-compliance. An application example was also presented. Such identification and subsequent installation of effective countermeasures against the vulnerabilities are necessary to avoid heavy government imposed penalties for non-compliance. Perhaps more importantly, they are necessary to inspire public trust in e-service providers. Future work includes: a) validating and fine-tuning the model with more examples as well as applications using flow information from real provider organizations, and b) investigating the open questions mentioned in section 4.

7. REFERENCES

[1] U.S. Government Office for Civil Rights. HIPAA: Medical Privacy - National Standards to Protect the Privacy of Personal Health Information. Available as of Feb. 28, 2005 at: http://www.hhs.gov/ocr/hipaa/

[2] U.S. Government. General Overview of Standards for Privacy of Individually Identifiable Health Information. Available as of Feb. 28, 2005 at: http://www.hhs.gov/ocr/hipaa/guidelines/overview.pdf

[3] Yee, G., Korba, L., and Song, R. Legislative Bases for Personal Privacy Policy Specification. Chapter in *Privacy Protection for E-Services*, edited by G. Yee, Idea Group, Inc., 2006. NRC 48270. Available as of August 18, 2006 at: http://iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/index_e.html

[4] Wikipedia. Data Flow Diagram. Available at: http://en.wikipedia.org/wiki/Data_flow_diagram (visited July 1, 2006).

[5] Korba, L., Song, R., Yee, G., and Chen, Y.-C. Scenarios for Privacy Rights Management Using Digital Rights Management. *Proceedings of the 2005 Resource Management Association international conference (IRMA 2005)*, San Diego, CA, USA, 2005. NRC 47428. Available as of August 18, 2006 at: http://iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/index_e.html

[6] Pillat, R.M., Freitas, C.M.D.S. Coordinating views in the Infovis toolkit. *Proceedings of the working conference on advanced visual interfaces,* Venezia, Italy, 2006, 496-499.

[7] Walker, R.J., Murphy, G.C., Freeman-Benson, B., Wright, D., Swanson, D., Isaak, J. Visualizing dynamic software system information through high-level models. *Proceedings of the 13th ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications,* Vancouver, B.C., Canada, 1998, 271-283.

[8] Konyha, Z., Matkovic, K., Hauser, H. Interactive 3D Visualization of rigid body systems. *Proceedings of the 14th IEEE Visualization Conference (VIS'03)*, 2003, 539-546.

[1] NRC Paper Number: NRC 48772