Defining a Framework for Teaching Privacy in Information Assurance Curriculum

Svetlana Peltsverger and Guangzhi Zheng, *Southern Polytechnic* State *University*

Abstract – Many Information Assurance courses include privacy topics. However, many of them do not address privacy issues systematically and comprehensively. Those courses do not offer students a complete picture of privacy from both data providers' and data collectors' perspectives. A coherent and consistent curriculum framework on teaching privacy needs to be defined. Moreover, students learn about possible invasion of privacy as a result of poor information system security, not about privacy as an essential principle in information systems. This paper discusses the importance of defining a consistent framework for teaching privacy in IA curriculum. Authors propose key learning outcomes and content modules, as well as two options to implement the framework. The framework can be used as a guide to design privacy courses and learning modules.

Index terms: Information Assurance, Privacy, Curriculum

I. INTRODUCTION

Information Assurance (IA) has become a broadly recognized priority for both government and private industry. Many universities offer degrees and certificates in IA, in which they teach students how to develop secure software and use the best practices for system configuration. In many of these programs, the focus of the education is on confidentiality, a part of both sets of core security principals: CIA triad (confidentiality, integrity, availability) and the newer Parkerian Hexad (confidentiality, possession or control, integrity, authenticity, availability, utility). However, confidentiality does not guarantee privacy. IA specialists must consider issues such as privacy as they design and manage information systems that protect private information from misuse. The authors believe that universities are still not producing enough graduates that will satisfy the requirements of industry in this area.

II. BACKGROUND

According to Oxford English Dictionary, privacy is the state or condition of being alone, undisturbed, or free from public attention, as a matter of choice or right; seclusion; freedom from interference or intrusion. Since it first use in *The Burgh Court book of Selkirk* in 1534, the definition of privacy has not changed as much as our society. The attack surface on privacy continues to expand. The change started in mid-1990s when the Internet introduced a nearly instantaneous way to access data via web browsers, emails and later instant messages, forums, blogs, social networks, e-commerce sites, global positioning systems, etc. People expose their personal information when they use computers, smart phones and other devices in everyday business and personal life. Every time a digital content is retrieved, a digital footprint that can be used to trace the action to a particular individual or device is created. The explosion of online content and the growth of online services like online banking and electronic health records expanded the surface of attacks on personal privacy. According to Cisco, global IP traffic is expected to grow to 44 exabytes per month by 2012.

In recent years, most companies have posted their privacy policies on websites, every year customers around the world receive letters and emails about changes in privacy policies. Privacy is increasingly popular topic in the press and courts worldwide. Contributing factors for growing concerns about privacy are:

- Decreasing storage cost.
- Decreasing cost of computing power.
- Decreasing cost of transmitting data.
- Increasing popularity of business intelligence that is ranked number five on the list of the top 10 technology priorities for chief information officers (CIOs) in 2011, according to Gartner's annual global CIO survey [1].

In 2001 due to the USA Patriot Act banks started performing due diligence on customers such as verifying identity and indefinitely saving account activity. It triggered collection of personal data such as credit card records, phone records, health care records, e-mails, mobile devices location.

Computing professionals are the ones who must design and implement information system to prevent private information disclosure. Privacy violations are not always the result of poor security, so students must learn about privacy as an essential principle in information systems, instead of learning it as consequences of poor information system security. The disclosure of personal information will eventually compromise security, as it happened with Sarah Palin's account when a published answer to a security question allowed a password change [2].

ISBN 1-933510-95-1/$15.00

Despite such importance of privacy, many educational programs do not cover the subject systematically and comprehensively. Authors studied several curriculum models including ACM curriculum requirements, Committee on National Security Systems (CNSS) Training requirements, and several university programs, and found that coverage of privacy is not adequate. Clearly an educational framework is missing. In this paper, we are proposing a framework that can be used as a guide for privacy curriculum design.

In the rest of the paper, we survey existing programs, describe the proposed framework, and suggest ways to implement those changes.

III. SURVEY OF THE CURRENT EDUCATIONAL PROGRAMS

Many universities base their computing curriculum on the Association for Computing Machinery's (ACM) curriculum model. ACM Information Technology Curricula IT 2008 [3] calls for one core hour in Social and Professional Issues (23 core hours) for *SP. Privacy and Civil Liberties IT426*, which exclusively discuss privacy from a social, ethical, and legal perspective. "Reasonable expectation of privacy" is covered under "Information Assurance and Security", as a learning outcome in Forensics. Privacy is also mentioned under "Web Technologies Social Software" when talking about the use of cookies. The coverage of privacy is clearly scattered.

ACM Computer Science Curricula 2008 (CS2008) [4] extends the coverage of privacy to information management, computer crime, and digital library. However, most of the learning objectives only stay on the understanding level instead of the application level. Words like "describe" and "clarify" are used. Computer science/software engineering students are future software developers and must learn how to incorporate privacy protection in their software designs. They must be able not only to describe, but also to design and implement privacy protection in information systems. In addition, both curriculum models do not provide details for the learning objective and topics. Therefore, the interpretation and implementation of learning outcomes largely depends on individual instructors.

In 2011 there were 145 institutions designated as Centers of Academic Excellence in Information Assurance. Before institution can be designated as CAE/IAE the courseware must be certified under the IA Courseware Evaluation Program as meeting the Committee on National Security Systems (CNSS). Certification for CNSS Training Standard 4011 is required.

NIST 4011 [5] calls for coverage of privacy in Policies and Procedures area (f) Auditing and Monitoring. Education, training and awareness section mentions "Which information requires protection is often debated in government circles. One historic problem is the clash of society's right to know and an individual's right to privacy." There is no recommendation on topics to cover.

We also examined a number of courses, programs, and popular textbooks. Privacy is often covered from different and narrow perspectives, either as a compliance or ethical issue, or as a pure computing issue. For example, CS 7301 Introduction to Data Privacy (University of Texas Dallas[1]) is a graduate computer science course focus on privacy techniques (algorithms). Computer Science 105 Privacy and Technology (Harvard[2]) is another course covering several rather independent topics including biometrics, surveillance, and data mining. In CIST 1601 Introduction to Security Fundamentals (Chattahoochee Technical College), privacy topics are focused on privacy awareness and ethics, where impacts and arising issues related to technology use are discussed.

Many programs have a dedicated course that covers Professional Practices and Ethics. One of the popular textbooks is the *Gift of Fire* by Sara Baase [6]. In one chapter dedicated to privacy the author covers Fourth Amendment, Privacy Regulations in US and Europe, Technology to Protect Privacy, Expectation of Privacy, Surveillance Technologies, Wiretapping and E-mail Protection, Designing Communications Systems for Interception, and ten more topics.

The above analysis reveals that many existing courses and programs cover only selected independent topics in different areas, depending on the perspective and focus of the course. The authors believe that a consistent learning framework needs to be established and used in design of IA courses and programs. In the next section, we present a preliminary framework that systematically addresses the information privacy education.

IV. A PRIVACY EDUCATION FRAMEWORK

The proposed learning framework is based on the following competencies:

1. Knowledge of major privacy issues in common domains (scenarios) involving data collection, tracking, usage, and sharing.
2. In each domain, from a data collector's (service provider's) perspective:
   a. Know the legal, social, and business issues.
   b. Analyze threats and risks.
   c. Propose, design and implement technical solutions to protect customers or clients.

[1] http://www.utdallas.edu/~mxk055100/courses/privacy08f.html
[2] http://isites.harvard.edu/icb/icb.do?keyword=k75223&pageid=icb.page379623

3. In each domain, from a data provider's perspective:
   a. Know the legal, social, and technical issues.
   b. Know types of data that need protection from privacy invasions.
   c. Configure systems and implement technical solutions to protect a data provider.

Domains of privacy issues

Students should be aware and understand major scenarios where privacy may be of concern. Moreover, they also should learn how to deal with them with caution. The following list of major scenarios involving privacy need to be considered in a course:

Explicit Data Collection

Explicit data collection: privacy has to be addressed in many activities that explicitly collect data. These include surveys, questionnaires, focus groups, or experiments conducted to specifically target data collection and analysis. Very often these techniques are part of commercial activities, such as market survey, consumer survey, industry survey, satisfaction survey, or public review and opinion, but they are also often used in non-profit and academic research.

Implicit Data Collection

Implicit data collection: privacy issues are often neglected in these areas because people may not be aware about the data collection process. This has become evident in the Internet advertising where user activity is being tracked by cookies, spyware, and action history. For example, log files contain a record of user activities and are subject to privacy concerns. It is also an issue of the Internet where Customer Relationship Management (CRM) has widely been implemented to record customers' data. Many websites require users to create a free account to track user activity. Collected data then can be used to improve or expand service offerings, e.g. Google launched a social service called Buzz in 2010 and exposed the personal contacts of its email users.

Data Management

Privacy and security issues are often treated together in industries where the management of these data is the key to the business, such as healthcare, education, financial, hiring and recruiting, background checking, etc. Information brokers also share information with partners. Every privacy policy includes language similar to "... may share data with trusted partners to help us perform statistical analysis, ...". As of February 2012 according to Google search, there were more than two million websites with such policies.

Online Privacy, Social Networking and Web Mining

This technology when used on data of personal nature might cause concerns. Especially in the proliferation of the social activities and contents these days. People may not be aware that the information they leave publicly and separately on different sites can be compiled into a more comprehensive profile.

Other Technology Specific Areas

Privacy issues exist in many tools or technology specific areas such as Peer–to-Peer networks where intermediate nodes must be considered untrusted parties (Androutsellis and Spinellis, 2004 [7]), emails, GPS, wireless communication, RFID, cloud, bio-metrics, etc. According to Maxmind.com you can find geo location of an IP address. For example, Hostname 74.125.65.100 belongs to Google and located Miami, FL 33144 Latitude 25.7660 Longitude 80.3112. Researchers in the paper "I Know Where You are and What You are Sharing" [8] used Skype API and Maxmind to track user mobility even if they were behind NAT. Authors suggested several solutions including not revealing the callee's IP address until the callee accepts the call. Privacy issues exist in non-computing environments as well. For example, hard drives and memory modules in copy machines retain a soft copy of scanned or copied document and are subjects to privacy concerns too.

Physical Environment

Students should know how to deal with video surveillance, wiretapping, etc. Facebook was forced to implement a more secure authentication method to protect users from widely publicized FireSheep wireless networking attack. Google now uses https as a default protocol to deliver query results.

From a Data Collector or Service Provider's Perspective:

First, students must be aware of the laws and regulations that govern the data collection and sharing process. Such privacy laws should at least include Family Educational Rights and Privacy Act (FERPA), Gramm Leach Bliley Act (GLBA), and Health Insurance Portability and Accountability Act (HIPAA).

Second, students should understand common approaches of data collection and tracking, and learn to manage common threats and risks in these processes.

Privacy policy is an important document that often becomes the center of a dispute. Students must learn not only how to write those policies, but how to enforce them. In reviewed courses and programs, enforcement is often an overlooked aspect. Another missing part is developing technical/automatic procedures for privacy policy enforcement.

In the online advertising industry, DoubleClick can be protected from legal actions if all data they sell is protected from deanonymization. The company tracks the

individual Internet users. As soon as the first ad is displayed on a computer, a unique number is generated and saved in a cookie file on the user's computer. When the user visits another website with DoubleClick ads, DoubleClick reads the cookie and can customize what add will be displayed. Participated businesses need to fully understand the threats and risks involved in DoubleClick's technologies.

Third, students should also be introduced to basic approaches, technologies, and systems for privacy preservation. Not only should they know and evaluate these technologies, but also know how to develop and configure systems to prevent privacy violations. Privacy preservation concepts and techniques are not covered in many Professional Practices and Ethics course. Examples of such technical areas are:

Anonymization

Anonymity is a result of not having identifying characteristics (such as a name or description of physical appearance) disclosed. A survey of these techniques and software should be covered, such as I2P (anonymous network), Tor (anonymity network), Phantom, Bitblinder, packet trace anonymization [9]. Student should also understand the threats to common anonymization techniques as anonymization does not always guarantee privacy [10][11].

Record Level Privacy vs. Source Level Privacy

Students who will design information systems must know how to achieve record level privacy, besides source level privacy. An example of record level privacy attack is Record linkage. Fung et al [12] used publicly available anonymized data from Table 1 and knowledge of a gender and age of a new patient to show that a 32 year old male patient was admitted to a clinic has HIV.

Table 1. 3-anonymous patient table

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[30-35)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

These solutions focus on distributed vertical partitioned datasets mapping, without identity disclosure, using non-identifying information induction. Common techniques used are statistical databases, privacy-preserving join computation, and privacy-preserving top-K queries as in Clifton et al. [13]. These projects mostly focus on record level anonymity, proposing solutions to prevent inducing association from a collection of non-identifying information. Clifton et al. [13] lay out a privacy framework for the area of "privacy preserving data integration" by introducing private views and private policies. Private views are used to limit access to information when data is considered to be private if appears together (e.g. address and date of birth). Privacy data policy defines access rights to private views. It can be defined based on the owner's explicit consent or more complex privacy policies.

Vaidya et al [14] in the paper on "distributed data mining" use perturbation – "modifying the data so that it no longer represents real individuals." They use the US Census Bureau's Public Use Microdata Sets were data values swapped between records that preserve statistical properties, but destroy real values.

Commercial solutions are also available and can used in courses, such as the DB2 Anonymous Resolution from IBM to anonymize data sources[3].

Table Linkage Aware Data Protection

Students should learn how to protect data before sharing, to prevent possible use of linkage with publicly available data and reveal the identity of the records. Most statistical agencies publicize anonymization process they use which provides valuable information for an attacker. Example by Nin et al [15] shows how to link U.S. Census and U.S. Energy Information data sets. They demonstrate that when only half of attributes are known more than 70% of the records can be linked.

Protocols

Protocols should also be covered as major privacy techniques. These include the lightweight protocols such as the secure sum technique [13], and many heavy weight protocols that usually make intensive use of encryption and routing. Examples include those P2P communication protocols such as Hordes [16], Onion Routing [17], "group anonymity" [18] and ad-hoc P2P collaboration [19]. for privacy preserving data mining. There are also non-encryption based techniques. One example is the Trusted Query Network [20], which is a policy based and automated query system that is aimed to make individual data unidentifiable and untraceable. The TQN protocol is shown in Figure 1.

Reputation Based Techniques

In reputation based techniques sharing decisions are made based on reputation certificates rather than real identification [21].

[3] http://www-01.ibm.com/software/data/infosphere/anonymous-resolution/

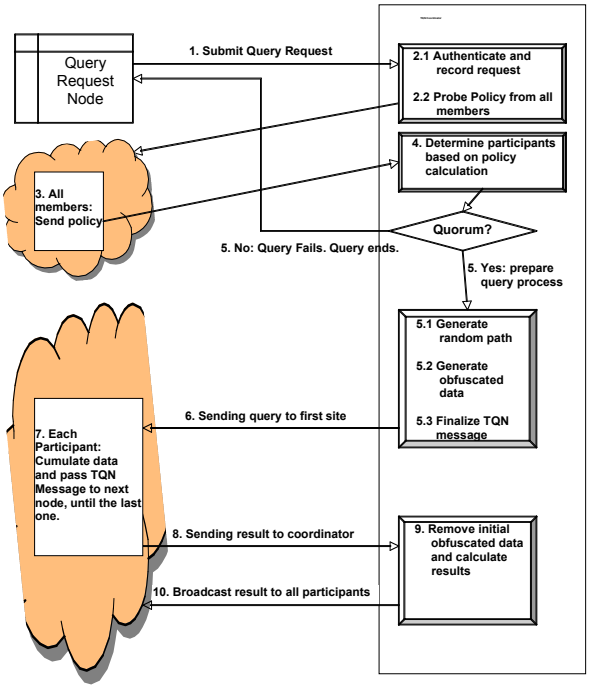

Figure 1: TQN Protocol [20]

Protecting Privacy from a Data Provider's Perspective:

In this part, students should be able to understand the privacy issues involved and know the tools to protect their own privacy when releasing their own information, either from an individual perspective or from an organization's perspective. This aspect of privacy is most often covered in existing IA curriculum.

Know the Rights

The legal issues involved in data collection and sharing should be covered: regulations, ethics, compliances, policies, etc. This is consistent with the legal issues covered from a data collector's perspective.

Students should also learn types of sensitive data (e.g. GEO location, personal information, health records, browsing history, etc.), be aware of what information can be used to identify oneself, and know how to analyze the risks. Such awareness is important to protect oneself. For example:

- Apple was under fire after experts found that every iPhone and iPad has been keeping track of gadget's GEO location without user consent.
- All businesses collect customer data to provide better services, beat competitors and increase profits. This information often includes but not limited to names, demographic data and transaction history. The NIST Guide to Protecting the Confidentiality of Personally Identifiable Information (SP 800-122) [22] ensures that businesses handle this information properly: "Information which can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc."
- Many businesses sell or exchange collected "non-personally identifiable information". In case of online businesses, non-personally identifiable information may also include user IP addresses, browser types, domain names, etc. Companies like InfoUSA.com sells data for direct marketing. The data can be searched by geography, age, income, home value and other great selections. InfoUSA.com claim that they process more than 1.5 billion records per year using United States Postal Service's (USPS) National Change of Address (NCOA) database. The data is certified using the USPS's Coding Accuracy Support Systems (CASS).

From a Technical Standpoint

Students must know common data collection and tracking technologies, so they can configure systems to protect their own privacy. Google's CEO Eric Schmidt in his interview with the Wall Street Journal [23] said: "every young person one day will be entitled automatically to change his or her name on reaching adulthood in order to disown youthful hijinks stored on their friends' social media sites."

V. How to Deliver This Content?

There are two options to deliver the privacy content defined in this framework in a current information security curriculum. The first option is to develop specific courses concentrating on privacy comprehensively. The courses' main learning objectives and topics may follow the structure proposed in the framework. The second option is to enhance current security related courses with more coverage on privacy, with each focusing on a particular issue listed in the framework. For example, the online privacy technologies can be covered in a web security course or e-commerce course, and anonymity techniques can be discussed in a web mining or data mining class.

The first option is an ideal choice as it delivers a coherent and consistent learning framework on privacy. However, it requires more preparation, longer planning time, and additional education resources. The second choice can take immediate enhancements and be implemented at the same time by different instructors. The current resources can be effectively utilized. But these different pieces may not contribute to a comprehensive goal as different courses have different objectives and schedules. If the institution is setting up a new program then the first option can be chosen; if current security courses have

been set up, then the second option can be chosen in the short term, and the first option may be considered in the long run.

VI. CONCLUSION

In this paper, we present a coherent and consistent framework for teaching privacy in IA programs. This framework allows students to learn major privacy issues in common domains from perspective of both data providers and data collectors. Though much work remains to be done, authors believe that such a framework will help universities design well-structured privacy courses and programs, and produce graduates that satisfy requirements of the industry, government, and society.

VII. REFERENCES

[1] Gartner's annual global CIO survey, url: http://www.gartner.com/it/page.jsp?id=1526414, visited February 2012
[2] Why security questions are important http://www.securitymatters.iu.edu/security-questions-important-2/, visited February 2012
[3] ACM Information Technology Curricula IT 2008, url: http://www.acm.org//education/curricula/IT2008%20Curriculum.pdf
[4] ACM Computer science Curricula 2008 CS2008 http://www.acm.org//education/curricula/ComputerScience2008.pdf
[5] NIST 4011 http://www.cnss.gov/Assets/pdf/nstissi_4011.pdf
[6] Baase, S (2007), "A Gift of Fire: Social, Legal, and Ethical Issues in Computing", Prentice Hall, New Jersey.
[7] Androutsellis, S. and Spinellis (2004) "A Survey of Peer-to-Peer Content Distribution Technologies," *ACM Computing Surveys* (36) 4, pp. 335-371.
[8] Blond, S., Zhang, C., Legout, A., Ross, K., Dabbous, W. (2011) "I Know Where You are and What You are Sharing, " In Proceedings of ACM SIGCOMM Internet Measurement Conference, November 2-4, 2011
[9] Pan, Ruoming, Mark Allman, Vern Paxson, Jason Lee, "The Devil and Packet Trace Anonymization", ACM SIGCOMM Computer Communication Review Archive Volume 36 , Issue 1 January 2006
[10] Narayanan, Arvind and Vitaly Shmatikov, "Robust De-anonymization of Large Datasets", 2008, http://arxiv.org/abs/cs/0610105v2, visited Feb 20, 2012.
[11] Backstorm et al. "Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography", Proceedings of WWW '07 the 16th international conference on World Wide Web.
[12] B. Fung, K. Wang, R. Chen, P. Yu., "Privacy-preserving data publishing: a survey on recent developments", ACM Computing Surveys, Vol. 42, Issue No 4, December 2010
[13] Clifton, C., A. Doan, A. Elmagarmid, M. Kantarcıoglu et al. (2004) Privacy Preserving Data Integration and Sharing. *DMKD'04*, Paris, France, 2004.
[14] Vaidya, J. and C. Clifton (2004) "Privacy-Preserving Data Mining: Why, How, and When," IEEE Security & Privacy pp. 19-27.
[15] J. Nin, J. Herranz and V. Torra, On method-specific record linkage for risk assessment, Proceedings of Joint UNECE/Eurostat work session on statistical data confidentiality, 2007
[16] Shields, C. and B. N. Levine. (2000) A protocol for Anonymous Communication over the Internet. seventh ACM Conference on Computer and Communication Security, (ACM CCS 2000), 2000.
[17] Reed, M. G., P. F. Syverson, and D. M. Goldschlag (1998) "Anonymous connections and onion routing," IEEE Journal on Selected Areas in Communications (16) 4, pp. 482-494.
[18] Krishnan, S., Jeffrey Uhlmann, (2004) "The design of an anonymous file sharing system based on group anonymity," Information and Software Technology (46) pp. 273-278.
[19] Abhichandani, T., K. Kosaka, and S. Chatterjee. (2006) Designing a P2P Ad-hoc Secured Collaboration Tool. First International Conference on Design Science Research in Information systems and Technology, Claremot, CA., 2006.
[20] Vaishnavi, Vijay, Art Vandenberg, Richard Baskerville, Guangzhi Zheng, "TQN: A Novel Approach to Generating Information Security Data", Proceedings of the 16th Workshop on Information Technologies and Systems (WITS), Milwaukee, Wisconsin, 2006
[21] Ooi, B. C., C. Liau, and K.L.Tan. (2003) Managing trust in peer-to-peer systems using reputation-based techniques. International Conference on Web Age Information Management (WAIM 2003), 2003, pp. 2-12.
[22] NIST Guide to Protecting the Confidentiality of Personally Identifiable Information (SP 800-122) http:// csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf viewed February 2012
[23] Google and the Search for the Future, The World Street Journal, August 14, 2010