Develop and Disseminate Hands-on Lab Materials of Privacy Concepts and Technologies to Educators

Na Li
Department of Computer Science
Prairie View A&M University
Prairie View, TX, USA
nali@pvamu.edu
0000-0001-9296-1260

Lin Li
Department of Computer Science
Prairie View A&M University
Prairie View, TX, USA
lilin@pvamu.edu
0000-0002-9652-8111

Mengjun Xie
Department of Computer Science and Engineering
University of Tennessee at Chattanooga
Chattanooga, TN, USA
mengjun-xie@utc.edu
0000-0001-5089-9614

Bugrahan Yalvac
Department of Teaching, Learning and Culture
Texas A&M University
College Station, TX, USA
yalvac@tamu.edu
0000-0002-3675-1936

***Abstract*—In the era of digitalization, a massive amount of data has been generated from people's online activities or use of portable/wearable devices. The data often carries rich information about people. Therefore, privacy technologies are needed, from data generation to usage and from transmission to storage, to protect people's sensitive information. Although the research community is making great progress in addressing advanced privacy protection technologies, very few educational materials have been developed to incorporate the latest research results and engage students in learning privacy technologies, especially for younger generations. In this paper, we present our newly designed educational materials on privacy technologies, which can be used for training high quality cybersecurity professionals to meet the ever-increasing demand. The developed learning modules not only incorporate the latest research results in privacy technologies but also include effective hand-on lab activities. To help other institutions effectively teach privacy technologies, we organized a faculty training workshop in summer 2022. Twenty-nine faculty from twenty institutions nationwide participated in the training. Survey results show that the participants gained a better understanding of privacy issues and demonstrated strong interest in teaching privacy technologies after attending the workshop.**

Keywords—Data Privacy, De-anonymization, Relationship Privacy, Image Privacy, Location Privacy, Web Tracking, IoT Security and Privacy

I. INTRODUCTION

With the fast development of the networking services such as social networks, Internet of Things, and mobile applications, cybersecurity has never been more challenging than today [1], [2]. Cybersecurity education has been attached with great importance in the newly released computer science curricula. Similarly, industries favor qualified workers with security technologies and the demand is higher than ever [3], [4]. A student possessing this skill set will have a distinct competitive advantage in the market. As an integral part of cybersecurity, privacy protection has gained more and more attention from both industry and academia due to serious privacy breaches in recent years. For example, Facebook was sued over the Cambridge Analytica data scandal [5] in which 87 million users' profiles were harvested; the Federal Communications Commission fined AT&T $25 million for failing to protect the clients' personal information in 2015 [6]; and the University of Mississippi Medical Center was hit with a $2.75 million fine in 2016 by the Department of Health and Human Services over a health data breach [7]. According to Ponemon Institute, the average total cost of a data breach increased from $3.86 million in 2020 to $4.24 million in 2021 [8]. The occurrence of the violations caused people to panic [9], especially considering their daily online activities which generate a massive amount of data containing personal information.

Research on privacy has been intensively conducted in the scientific community. Despite the critical societal importance, privacy education has not been well integrated into the undergraduate computer science curricula. Privacy issues are often treated as optional topics instead of key fundamental concepts in security learning. A main obstacle is the serious lack of effective learning materials which can enable students to understand critical privacy concepts and gain hands-on skills. To address these problems and better prepare qualified graduates for future U.S. workforce, the researchers from AAA University and BBB University collaborated in developing innovative learning materials on privacy protection.

The rest of this paper is organized as follows: Section II presents a background of current privacy education and our project objectives. Section III briefly introduces seven lab modules that we developed. Section IV gives two examples of the hands-on lab activities of the learning modules: *Data Anonymization and Web Tracking*. Section V analyzes the feedback of the faculty who attended our training workshop on using the developed lab modules. Section VI concludes the paper.

II. BACKGROUND AND PROJECT OBJECTIVES

Recently, a team of cross-disciplinary members, including computer scientists, educators, and social scientists, at the International Computer Science Institute (ICSI) and UC Berkeley, developed an online privacy curriculum which targets younger students [10]. This team designed and implemented ten principles with the purpose of spreading the awareness of protecting privacy among younger students and helping them better understand what happens to personal information when it goes online, how it might be used to negatively affect users, and how they can defend their privacy by limiting what they share. They focused on online privacy in general, which doesn't cover as broad topics of privacy protection as we do in this project. Additionally, they are more knowledge based instead of hands-on learning based.

In this project, we focused on developing effective and hands-on learning modules on privacy protection. Through the engaging lab activities, we expect to motivate students' interests in privacy technologies and deepen their understanding of privacy issues. The specific objectives of the project include:

- Design self-contained privacy learning modules by encapsulating the hands-on labs and related lecture contents, which can be infused into teaching different security and privacy subjects and be easily adapted by other institutions.
- Develop effective hands-on labs on privacy breach and protection on various topics, including cutting-edge fields such as the Internet of Things (IoT) and social media, with a special effort on developing engaging lab setting/labware which enables students to gain first-hand experience.
- Evaluate the effectiveness of the experiential learning approach on students' learning outcomes, experience, motivation and attitudes towards privacy study.

III. LAB MODULE OVERVIEW

So far, we have developed seven privacy learning modules on different topics, including Data Privacy, De-anonymization, Relationship Privacy in Online Social Networks, Image Privacy, Location Privacy, Web Tracking, and IoT Security & Privacy. We created a VirtualBox image for each lab in the modules, which can easily be adopted by educators from other institutions. Next, we briefly introduce the lab modules. Two detailed lab examples will be described in Section IV.

A. Data Privacy

Prevalent data collection, both offline and online, by governments and private entities has gradually become a new "norm". On the one hand, new technologies such as artificial intelligence and data mining heavily rely on gigantic volumes of data. On the other hand, privacy concerns about omnipresent data collections have been growing especially when many different datasets can be obtained and crossed by the same entity. A well-known linking attack against people's data privacy was revealed by Dr. Sweeney [11]. This module aims to highlight the importance and challenges of data privacy and provide hands-on experience for students to understand data anonymization. The lab introduces basic concepts of data anonymization, linkage attack, and the k-anonymity privacy model. This lab provides an opportunity for students to practice using ARX [12], a data anonymization tool, to apply k-anonymity.

Learning Outcomes: Students will be able to (1) explain the importance and necessity of data privacy, (2) explain k-anonymity and its weaknesses, (3) analyze the utility of anonymized data, and (4) apply ARX to anonymize sensitive personal data.

Lab Design and Implementation: This lab consists of three tasks. The first task is about linkage attack. In this task, a brief explanation of linkage attack is provided and then an exercise is given for students to apply the linkage attack on a small dataset. The second task is on k-anonymity. In this task, the relevant concepts of k-anonymity protection model are first provided with small examples; Then, how to apply an open-source software, ARX, to data anonymization using the k-anonymity model is introduced. Hands-on exercises are provided for both parts. The reason that ARX was picked for this lab is two-fold: First, ARX is free, open-source, functionally rich, and well maintained. Second, ARX can be installed on Windows, Linux, and macOS. The third task is on attacking k-anonymity in which two attack methods, homogeneity attack and background knowledge attack, are introduced and practiced.

Challenges: Students are required to apply a linkage attack and explain the attack result in the first task. In the second task, students need to apply the k-anonymity model to a given table and perform required operations using ARX to anonymize a given dataset. Students are asked to use k-anonymity attack methods to perform attacks on a large anonymized dataset in the third task.

B. De-anonymization

De-anonymization is referred to as re-identifying target people from anonymized data with extra knowledge. The anonymized dataset is called Target Set (TS) and the dataset with extra knowledge is called Auxiliary Set (AS). De-anonymizing social media data can use either descriptive information, such as users' hobbies, membership groups, location information or behavioral patterns online [13]–[16], or structural information, such as centrality and neighborhood topology [17]–[21], or both [22]. De-anonymization can be implemented with seed based and signature based attacks. The seed based attacks [23], [24] start with a small number of seeds which are identifiable users, and attempt to identify their neighbors, and then their neighbors' neighbors, and so forth. The signature based attacks do not assume the availability of any seeds; instead, they rely on node signatures [22], [25], [26], which are uniquely generated from the nodes' descriptive or/and structural information, and then match node signatures between TS and AS to re-identify users.

Learning Outcomes: students will be able to (1) master the definition of de-anonymization, (2) know some de-anonymization technologies, (3) understand how the target set and the auxiliary set are prepared, (4) understand the implementation of seed based de-anonymization, (5) understand how to use profile and topological attributes in de-anonymization and (6) analyze experiment results to observe what could impact the de-anonymization accuracy.

Lab Design and Implementation: we developed a web application using Angular JS, Node.js, Spring, and MongoDB. A dataset [27] collected from Weibo (i.e., a Chinese Twitter) was used to generate TS and AS. The system consists of three primary components: TS and AS data generation, de-anonymization, and experiment analysis. The data generation component allows users to configure the size of each set, and the percentage of their common nodes as well as how to anonymize TS, like injecting randomness into gender or year of birth values. The de-anonymization component runs a seed-based algorithm in the backend and visualizes results in a graph implemented with Cytoscape.js [28]. It allows users to specify the number of initial seeds and decide what information to leverage for de-anonymization, either profile attributes (i.e., Year of Birth and Gender) or structural attributes (i.e., degree and centrality). The summary of the de-anonymization result presents with the number of pairs of nodes matched between TS and AS and the number of correct matches. Lastly, the analysis component provides a user-friendly interface to analyze the experiment results by plotting different charts.

Challenges: Students are required to run several groups of experiments with different configurations, from data generation to de-anonymization. Then they need to observe and analyze what could affect the de-anonymization accuracy of TS (e.g., how much of TS is anonymized) and how.

C. *Relationship Privacy*

Among the massive amount of data generated by social media platforms, relationship data has been focused more on protection in recent years [29], [30]. Researchers noticed that users' relationship privacy might be compromised by others using friend search engines [31]. A friend search engine is an application which can retrieve friend lists of individual users. In designing a friend search engine, OSN operators tend to display the entire friend list in response to each query in order to increase sociability of the site. However, some users may not feel comfortable to displaying their full friend lists. In [31], the author proposed a privacy-aware friend search engine which handles the trade-off between privacy protection and the sociability of OSNs by setting a k value to control the number of friends displayed and selecting which friends should be displayed.

Learning Outcomes: Students are able to (1) be aware that their behaviors through OSN applications may compromise other users' privacy, (2) observe the results from different display strategies of a friend search engine, (3) compare and evaluate different display strategies in terms of privacy preservation and impact on sociability, and (4) understand the trade-off between preserving users' privacy and enhancing the sociability of OSNs.

Labware Design and Implementation: We developed a web application interacting with the Twitter API "GET followers/list" [32] to query users' followers. The system implements three display strategies Random K (randomly selecting K friends), Rank K (choosing the most influential ones to the sociability of the OSN) and Top K [31], where $K = 4$. A user can make queries using any display strategies implemented and visualize the results in a graph implemented with Cytoscape Web [28]. In reality, each user's impact on the sociability of an OSN should be measured by multiple factors, including but not limited to his online activities and social connections. However, due to the restriction on the use of Twitter APIs, only a certain number of queries can be made in a time window through Twitter APIs, therefore, a real-time evaluation was not doable. Instead, the system randomly selects node weights in [0, 1000) as impact indicators. The total weight of nodes visible in the result graph represents those nodes' impact on the site sociability and it increases with more queries. A line chart is plotted with the number of users whose relations are compromised with queries.

Challenges: Students are required to run several queries with three different display strategies and observe their privacy violation levels. Also, they need to compare the impact of the strategies on the sociability of OSNs to evaluate the trade-off between the preservation of relationship privacy and the retention of sociability.

D. *Image Privacy*

Photo sharing has become a popular activity for online social network users. Semantically rich photos often contain not only the information that the uploaders want to share but also the information that is sensitive to others. However, most of the current OSNs do not have well-defined mechanisms for user privacy protection. This labware, which we named Facelock was developed for teaching photo privacy. The goal is to increase students' awareness of privacy protection while sharing photos in OSN. Through the hands-on activities, students will gain understanding of photo privacy and the essential concepts of face recognition.

Learning Outcomes: Students will (1) be aware of the privacy issues related to photo sharing on OSNs, (2) gain basic understanding of face recognition and object detection through deep learning technologies, (3) know how to use blurring techniques to maintain a trade-off between privacy protection and utility loss, and (4) understand the access control mechanism of OSN users for image sharing and privacy protection.

Labware Design and Implementation: The photo privacy labware consists of three components: (1) a SQLite database, (2) a web server, and (3) a deep learning based face recognition library. The SQLite database is used to store the information of the registered users, including their post records, profiles, friend lists, etc. For face recognition purposes, each registered user must upload a ”standard” picture as part of his profile. The picture will be used for user detection in the photos through a face recognition library. The Facelock web server hosts a Facebook-like social network environment where users can post messages and pictures, search people, and join friend circles, etc. Besides, Facelock reinforces photo privacy protection. When a user attempts to post a photo, the web server will use the face recognition library to tag users in the photo. Each tagged user will receive an alert and then take action to edit the photo to meet his privacy preference. The post will not be made publicly available until all tagged users respond with their privacy protection choices. The face recognition library plays a key role in user tagging. There have been different deep learning models developed for face recognition. Considering the model complexity and the speed required, the one adopted by Facelock is an open source application available on GitHub [33]. This library is light weight and highly accurate. Requiring only one profile picture for each OSN user, it can achieve the recognition accuracy of 99.38%.

Challenges: Students are required to play different roles in the social network: post uploader, user(s) involved in the photos, and normal post readers. With different blurring schemes, students can observe their privacy violation level.

E. Location Privacy in Location Based Services

Location Based Services (LBS) have been applied in many web and mobile applications. With LBS-enabled services, individuals can share their real-time and historical geographic location information online to facilitate social interactions or events. However, alongside the benefits, mobile LBS capabilities also cause users' privacy concerns. Therefore, a large group of researchers have been devoted to designing secure and practical mechanisms to protect users' location information in LBS [34]–[36].

Learning Outcomes: Students will be able to (1) increase their awareness of location privacy protection, (2) know basic anonymization algorithms to protect their location information, (3) understand how to track users based on their requests from the perspective of a malicious LBS, (4) understand the trade-off between location anonymization and its cost from the viewpoint of a privacy analyst, and (5) understand different users' privacy preferences.

Labware Design and Implementation: The labware we developed consists of three components: an Android client, a LBS server, and an analytic server. First, the Android client provides users with an interface to request the nearest landmark searched by individual requestors with Google Places APIs. The Android client implements three anonymization approaches, *no anonymization* (using real location), *shift based anonymization* (using a dummy location for query) and *area based anonymization* (instead of sending a single location point, an area is included in the query). Second, the LBS server is assumed to be semi-trustable, therefore, it has only whatever is included in users' queries, instead of their real location information. The server can filter queries and plot traces of individual users based on the queries they have made. Lastly, the analytic server gathers data from both the Android client and the LBS server in order to analyze the trade-off between preserving users' location privacy and its cost which is referred to as the distance between the real nearest landmark requested and the one returned from the LBS. The analytic server can also filter queries according to the time period of data collection, user id, or anonymization approach. Additionally, the server can analyze users' privacy preferences, showing the distributions of uses using different anonymization options. For individual users, the server can plot charts to show how a particular user's privacy preference changes over time.

Challenges: Students are required to use our Android client to request landmarks with different anonymization options. Then, they need to go to the LSB web site to filter requests and view his trace on the map. Last, students are required to visit the analytic service to analyze the trade-off of location protection and its cost (i.e., the accuracy loss of LBS responses).

F. Web Tracking

Web tracking happens when visitors browse the Internet. The websites or the third parties collect, store, and share information about visitors' activities. By analyzing users' behaviours, the websites may infer their preferences and provide content that attract the visitors in order to maximize the commercial benefits. In general, web tracking technologies can be categorized into two groups: *stated tracking* and *stateless tracking*. The former is usually done through cookies, and the latter is often conducted by browser fingerprinting. This lab was developed for teaching both concepts.

Learning Outcomes: through the hands-on activities, students will (1) be aware of web tracking and its commercial importance and potential threats to user privacy, (2) be able to explain the cookie mechanism and how to prevent web tracking through cookies, (3) be able to explain the browser fingerprint mechanism and how to prevent web tracking through browser fingerprint, and (4) be able to analyze web scripts and effectively prevent web tracking.

Labware Design and Implementation: The web tracking lab consists of three components: (1) several E-commerce websites, (2) an OSN website, and (3) the advertisement server. The E-commerce websites host different product information (e.g., appliances, shoes, phones, etc.) and each web page is embedded with a script from the advertisement server. When a visitor browses the products from page to page, the script will track the browsing record and transmit the information, including the visitor computer's ID (i.e., cookie or fingerprint) and the product ID, to the advertisement server. The OSN website plays a third-party role in which the visitor browses for other purposes (e.g., social networking). Similarly, a script from the advertisement server is embedded into the website. After a visitor browses some products of the E-commerce websites and then comes to visit the OSN website, the script can retrieve the cookie or fingerprint information and compare it with the historic records stored at the advertisement server. The user's behaviors will be analyzed. A result will be displayed at the OSN website for advertising purposes. The advertisement server is the hub for user information collection, storage, and analysis. Since its scripts are embedded in a hidden mode in the E-commerce and OSN websites, it is invisible to the visitors unless they know how to examine the website source code and how to analyze the network traffic.

Challenges: Students are required to browse different commercial websites, study scripts embedded into each web page, observe the change of database records upon each product browsing activity, and analyze how the advertisement server displays the visitors' browser history information on the third-party website. Students are also required to manipulate browser settings to understand different web tracking and protection mechanisms.

G. *IoT Security & Privacy*

Recent years have witnessed the exponential growth of Internet of Things (IoT) technologies as well as the soaring increase of attacks against IoT devices. Those attacks often exploit weak security protection exposed on many IoT technologies. Once IoT devices are compromised, they often become "bots" and are remotely controlled by attackers. Those IoT bots can be used to launch a variety of attacks that not only can damage system security, such as distributed denial-of-service (DDoS) attacks against legitimate servers, but also can breach data privacy, such as data theft or espionage through compromised wireless routers and Internet cameras. The famous Mirai botnet is such an example [37]. This module was designed to help students better understand how IoT security and privacy are attacked in practice by adapting the Mirai source code.

Learning Outcomes: Students will be able to (1) describe key facts about the Mirai malware, (2) explain Mirai operates in both the infecting and the attacking phases, (3) explain how to prevent the spread of Mirai, and (4) practice infecting a simulated Internet of Things (IoT) device with provided Mirai emulation executables.

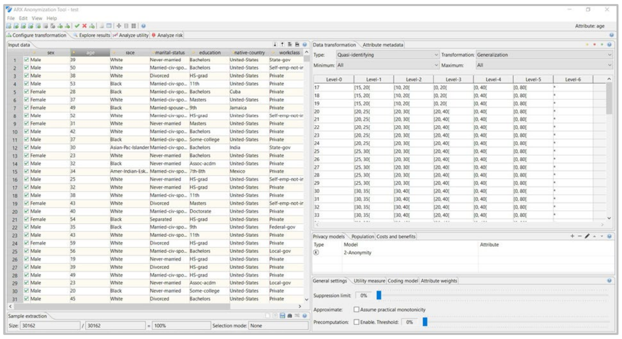
Lab Design and Implementation: This lab was built using modified Mirai source code, which was carefully designed to preserve essential Mirai operations including infection and spreading without the worry of security breach on the local area network (LAN) or even Internet. The lab environment consists of four virtual machines (VM): a command and control (C&C) server VM, a loader VM, a router VM, and a LAMP (Linux, Apache, MySQL, PHP/Perl/Python) server VM. The C&C server VM is to control the execution of the Mirai "botnet", which is the router VM in our case. The router VM is the machine to be infected and used to launch a DoS attack on the LAMP server VM, which is a local web server accessible to all other VMs. The loader VM is used to load Mirai onto the router VM. All the four VMs are placed in an internal network within VirtualBox to contain the spread of Mirai. The lab implemented important functions for operating a Mirai botnet including network scanning for victims, loading appropriate code to attack a victim, controlling a bot remotely, and launching an attack from a bot.

Challenges: Students are required to follow instructions to perform a sequence of command line operations on different VMs to practice how an IoT (emulated) device can be compromised by Mirai and later used for spreading malware or launching an attack.

IV. EXAMPLES OF HANDS-ON LAB ACTIVITIES

A. *Data Anonymization*

Our data privacy lab asks students to apply the knowledge of data anonymization and k-anonymity and skills of using ARX anonymization tool to anonymize a reasonably large dataset that contains 30,162 records. During the hands-on practice, the instructor will first provide the necessary background of k-anonymity and then introduce ARX and demonstrate its basic operations. After that, students can launch ARX from the provided VirtualBox VM instance. They need to follow the instructions to perform the following operations: (1) start a new project; (2) import the provided dataset using the File Import function; (3) mark quasi-identifying attributes in the input data section; (4) create a hierarchy for each quasi-identifying attribute without a predefined hierarchy or import a hierarchy file for its corresponding attribute; (5) create a k-anonymity privacy model with k set to 2 once all the hierarchies are set; (6) customize certain configuration attributes in the general settings; (7) perform anonymization; (8) visualize and review anonymization results; (9) analyze utility; and (10) generate a certificate file. A screenshot of the ARX interface after step 5 is depicted in Fig. 1.



		sex	age	race	marital-status	education	native-country	workclass
1	☑	Male	39	White	Never-married	Bachelors	United-States	State-gov
2	☑	Male	50	White	Married-civ-spo...	Bachelors	United-States	Self-emp-not-ir
3	☑	Male	38	White	Divorced	HS-grad	United-States	Private
4	☑	Male	53	Black	Married-civ-spo...	11th	United-States	Private
5	☑	Female	28	Black	Married-civ-spo...	Bachelors	Cuba	Private
6	☑	Female	37	White	Married-civ-spo...	Masters	United-States	Private
7	☑	Female	49	Black	Married-spouse-...	9th	Jamaica	Private
8	☑	Male	52	White	Married-civ-spo...	HS-grad	United-States	Self-emp-not-ir
9	☑	Female	31	White	Never-married	Masters	United-States	Private
10	☑	Male	42	White	Married-civ-spo...	Bachelors	United-States	Private
11	☑	Male	37	Black	Married-civ-spo...	Some-college	United-States	Private
12	☑	Male	30	Asian-Pac-Islander	Married-civ-spo...	Bachelors	India	State-gov
13	☑	Female	23	White	Never-married	Bachelors	United-States	Private
14	☑	Male	32	Black	Never-married	Assoc-acdm	United-States	Private
15	☑	Male	34	Amer-Indian-Esk...	Married-civ-spo...	7th-8th	Mexico	Private
16	☑	Male	25	White	Never-married	HS-grad	United-States	Self-emp-not-ir
17	☑	Male	32	White	Never-married	HS-grad	United-States	Private
18	☑	Male	38	White	Married-civ-spo...	11th	United-States	Private
19	☑	Female	43	White	Divorced	Masters	United-States	Self-emp-not-ir
20	☑	Male	40	White	Married-civ-spo...	Doctorate	United-States	Private
21	☑	Female	54	Black	Separated	HS-grad	United-States	Private
22	☑	Male	35	Black	Married-civ-spo...	9th	United-States	Federal-gov
23	☑	Male	43	White	Married-civ-spo...	11th	United-States	Private
24	☑	Female	59	White	Divorced	HS-grad	United-States	Private
25	☑	Male	56	White	Married-civ-spo...	Bachelors	United-States	Local-gov
26	☑	Male	19	White	Never-married	HS-grad	United-States	Private
27	☑	Male	39	White	Divorced	HS-grad	United-States	Private
28	☑	Male	49	White	Married-civ-spo...	HS-grad	United-States	Private
29	☑	Male	23	White	Never-married	Assoc-acdm	United-States	Local-gov
30	☑	Male	20	Black	Never-married	Some-college	United-States	Private
31	☑	Male	45	White	Divorced	Bachelors	United-States	Private

Size: 30162 / 30162 = 100% Selection mode: None

Type: Quasi-identifying Transformation: Generalization
Minimum: All Maximum: All

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5	Level-6
17	[15, 20[	[10, 20[	[0, 20[	[0, 40[	[0, 80[	*
18	[15, 20[	[10, 20[	[0, 20[	[0, 40[	[0, 80[	*
19	[15, 20[	[10, 20[	[0, 20[	[0, 40[	[0, 80[	*
20	[20, 25[	[20, 30[	[20, 40[	[0, 40[	[0, 80[	*
21	[20, 25[	[20, 30[	[20, 40[	[0, 40[	[0, 80[	*
22	[20, 25[	[20, 30[	[20, 40[	[0, 40[	[0, 80[	*
23	[20, 25[	[20, 30[	[20, 40[	[0, 40[	[0, 80[	*
24	[20, 25[	[20, 30[	[20, 40[	[0, 40[	[0, 80[	*
25	[25, 30[	[20, 30[	[20, 40[	[0, 40[	[0, 80[	*
26	[25, 30[	[20, 30[	[20, 40[	[0, 40[	[0, 80[	*
27	[25, 30[	[20, 30[	[20, 40[	[0, 40[	[0, 80[	*
28	[25, 30[	[20, 30[	[20, 40[	[0, 40[	[0, 80[	*
29	[25, 30[	[20, 30[	[20, 40[	[0, 40[	[0, 80[	*
30	[30, 35[	[30, 40[	[20, 40[	[0, 40[	[0, 80[	*
31	[30, 35[	[30, 40[	[20, 40[	[0, 40[	[0, 80[	*
32	[30, 35[	[30, 40[	[20, 40[	[0, 40[	[0, 80[	*
33	[30, 35[	[30, 40[	[20, 40[	[0, 40[	[0, 80[	*

Type	Model	Attribute
ⓚ	2-Anonymity	

Suppression limit: 0%
Approximate: □ Assume practical monotonicity
Precomputation: □ Enable. Threshold: 0%

Fig. 1. A Screenshot of Using ARX for Data Anonymization

B. *Web Tracking by Browser Fingerprinting*

Our web tracking lab, depicted in Fig. 2, was developed on top of Dr. Wenliang Du's SEED lab "Web Tracking" [38] which only introduces web cookies. Therefore, we focus on the introduction of our browser fingerprinting lab and skip the cookie part. During the hands-on activities, students are required to take the following steps to understand the concepts, observe the results, and study the scripts: (1) visit E-commerce websites and browse the products; (2) log into the MySQL database of the advertisement server, select the database (named "revive adserver"), and open the record table (named "bt FingerprintLog"). After selecting each product, they will observe the record change in the table; (3) visit the E-commerce site ("www.wtlabelgg.com") and observe the product displayed in the banner area; (4) revisit an E-commerce website and open the source code of it to examine the JavaScript code on how the browser fingerprint is generated and how the product information is embedded; (5) open each product page and examine the source code, study how the web uses PHP script to pass the browser

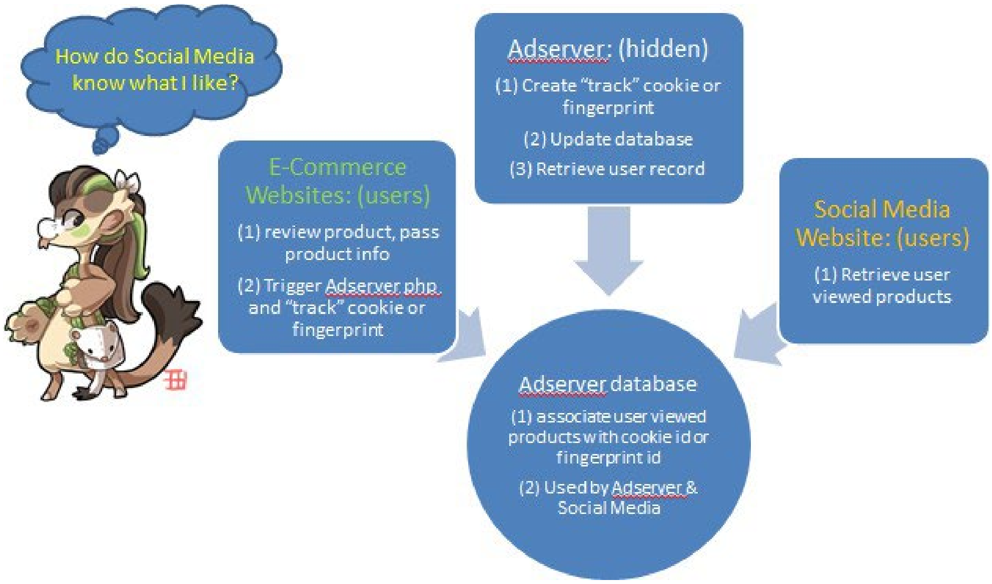

How do Social Media know what I like?
Adserver: (hidden)
(1) Create "track" cookie or fingerprint
(2) Update database
(3) Retrieve user record
E-Commerce Websites: (users)
(1) review product, pass product info
(2) Trigger Adserver.php and "track" cookie or fingerprint
Social Media Website: (users)
(1) Retrieve user viewed products
Adserver database
(1) associate user viewed products with cookie id or fingerprint id
(2) Used by Adserver & Social Media

Fig. 2. Illustration of Web Tracking

fingerprint and product ID to a tracking script at the advertisement server side; (6) open the source code of the tracking script to examine how the browser fingerprint and product information are inserted into the database; (7) revisit the social network and open its source code to study how the script retrieve the historic records from the advertisement server and display the result based on an analysis of the visitor behaviors. Finally, students will be asked to manipulate the scripts to display different private information of the user.

V. EVALUATION

Some of the developed lab modules had been tested among students through some security courses that we taught in the past few years [XX, YY, ZZ]. To disseminate the project outcomes and widely evaluate the lab modules, we organized a faculty training workshop in summer 2022. We believe that the well trained faculty can broaden the project impact and engage more students into the fields of privacy and cybersecurity. In this paper, we focus on the survey analysis of the workshop participants.

A total of 29 faculty from twenty institutions nationwide participated in the workshop, including 20 men (69%) and 9 women (31%). There were 10 Full Professors (34%), 8 Assistant Professor (28%), 6 Associate Professor (21%), 4 lecturers (14%), and 1 part-time instructor (3%). 25 participants completed both the pre-workshop and post-workshop surveys. Majority of the respondents were at the Computer Science Department in their institutions. Computer science and computer security were the two mostly reported course titles among the courses participants have taught. 17 of the participants (68%) reported their ethnicity as Asian, 5 of them (20%) as African American, and 3 of them (12%) as Caucasian.

Before the workshop, participants were not well aware of the session topics. The means of their responses to a 5 point scale was around 3.51, which meant they knew only a few words about the session topics. After the workshop was completed, participants' response mean increased to 4.66, which meant that they knew the basic terms and could apply the concepts. Comparing the participants' responses to the pre- and post-workshop surveys, we found that participants statistically significantly improved their awareness of all the seven session topics: Data Privacy (**DP**), Relationship Privacy (**RP**), Image Privacy (**IP**), De-anonymization (**DA**), IoT Security & Privacy (**IoT**), Location Privacy (**LP**), and Web Tracking (**WT**). The means and standard deviations of the participants' awareness of the workshop session topics before and after the workshop, the paired sampled student t-test results, and cohen's *d* effect sizes are presented in Table I. The comparisons of the pre versus post awareness found statistically significant difference at $p = 0.01$ level. This means that participants greatly improved their awareness of the session topics. Cohen's *d* effect sizes close to 1 or bigger than 1 indicate a large group mean difference.

TABLE I. CHANGES IN PARTICIPANTS' AWARENESS OF THE TOPICS

Lab	**Pre**: μ (σ)	**Post**: μ (σ)	**t**	**df**	**p-value**	*d*
DP	4.08 (0.79)	4.84 (0.36)	5.467	24	< 0.001 *	1.072
RP	3.44 (1.06)	4.72 (0.79)	6.799	24	< 0.001 *	1.333
IP	3.32 (0.88)	4.56 (0.63)	6.972	24	< 0.001 *	1.367
DA	3.04 (1.25)	4.52 (0.81)	6.268	24	< 0.001 *	1.229

Lab	**Pre**: μ (σ)	**Post**: μ (σ)	**t**	**df**	**p-value**	*d*
IoT	3.04 (1.25)	4.52 (0.81)	5.892	24	< 0.001 *	1.155
LP	3.44 (1.13)	4.56 (0.75)	4.594	24	< 0.001 *	0.901
WT	3.56 (1.09)	4.64 (0.55)	5.417	24	< 0.001 *	1.062

Similarly, before the workshop, participants were moderately interested in teaching the session topics. The means of their responses to a 5 point scale was around 3.95, which meant they were "somewhat" to "a lot" interested in teaching the topics. After the workshop, participants' response mean increased to 4.25, which meant that they were "a lot" to "a great deal" interested in teaching the topics to their students. The pre and post surveys show that participants statistically significantly improved their interest in teaching all the session topics, however the group mean differences were not statistically significant. Table II presents the results about the participants' interest change before and after the workshop. For the Data Privacy and Location Privacy, participants' pre and post responses show statistically significant difference at $p = 0.05$ level. This means that participants' interest in teaching these two topics greatly improved. In other session topics, participants also increased their interests to teach the topics. But the group mean differences were not statistically significant. Cohen's *d* effect sizes close to 0.5 indicate a medium group mean difference. Effect sizes close to 0.1 indicate a very small group mean difference.

TABLE II. CHANGES IN PARTICIPANTS' INTEREST IN TEACHING THE TOPICS

Lab	**Pre**: μ (σ)	**Post**: μ (σ)	**t**	**df**	**p-value**	*d*
DP	4.16 (0.96)	4.56 (0.69)	2.28	24	0.031 *	0.447
RP	3.72 (1.11)	4.04 (1.11)	1.557	24	0.132	0.305
IP	3.76 (0.99)	3.96 (1.14)	0.787	24	0.439	0.154
DA	3.72 (1.18)	4.04 (1.07)	1.504	24	0.145	0.295
IoT	4.24 (0.91)	4.48 (0.85)	1.604	24	0.121	0.314
LP	3.92 (0.97)	4.28 (0.91)	2.178	24	0.039 *	0.427
WT	4.12 (0.81)	4.36 (0.79)	1.883	24	0.071	0.369

After each session, the participants were surveyed about the learning materials and workshop organization in terms of the following seven statements: (1) This lab increased my knowledge and skills in [session topic]. (2) I learned how to teach [session topic] more effectively. (3) The session was well organized. (4) The session objectives were stated clearly and met. (5) The information provided and/or skills presented were relevant and useful. (6) The presenter(s) provided adequate time for Q&A. (7) The session materials provided were useful.

Table III presents participants' responses to the session surveys on a 6 point Likert-Scale (1 = Strongly disagree, 2 = Disagree, 3 = Slightly disagree, 4 = Slightly Agree, 5 = Agree, 6 = Strongly Agree). Results show that almost all participants agreed that the sessions increased their knowledge and skills in the privacy technologies and they learned how to teach the topics effectively. Participants reported that the sessions were well organized, objectives were stated clearly and met, the information provided and/or skills presented were relevant and useful, presenter(s) provided adequate time for questions and answers, and the session materials were useful.

TABLE III. PARTICIPANTS' RESPONSES TO THE SEVEN STATEMENTS.

#	**DP** (22)	**RP** (22)	**IP** (19)	**DP** (20)	**IoT** (18)	**LP** (18)	**WT** (21)
1	5.27	5.61	5.42	5.35	5.67	5.33	5.43
2	5.14	5.3	5.11	5.05	5.44	5.11	5.24
3	5.41	5.70	5.58	5.40	5.72	5.50	5.29
4	5.45	5.57	5.63	5.35	5.72	5.72	5.19
5	5.55	5.65	5.58	5.30	5.67	5.61	5.38
6	4.95	5.70	5.74	5.60	5.50	5.67	5.48
7	5.36	5.65	5.53	5.40	5.67	5.44	5.43
The # within the Parenthesis Represents the Surveys Returned							

VI. CONCLUSION

Privacy education is critical for training younger generations to become future cybersecurity professionals. The emergence of new communication methods (e.g., OSN and IoT) and networking services (e.g., Location Service) has greatly improved people's life quality. While people benefit from these new technologies, they also expose more personal information to the Internet which brings serious concerns about privacy protection. Thus, we designed multiple privacy labs to help students learn specific privacy issues. We organized a faculty training workshop to disseminate our project outcomes and enable them to teach privacy topics effectively. Our study shows that the faculty trainees were very satisfied with the lab materials in terms of concept learning and hands-on experience.

ACKNOWLEDGEMENT

This project is supported by the National Science Foundation (NSF) under grant DUE-1712496.

REFERENCES

[1] P. Thomas, "Cybersecurity: A look at performance and growing opportunities," November 09, 2015. http://marketrealist.com/2015/11/ise-cyber-security-etf-perform-last-week/.

[2] W. House, "Fact sheet: Biden administration and private sector leaders announce ambitious initiatives to bolster the nation's cybersecurity," August 25, 2021. https://www.whitehouse.gov/briefing-room/statements-releases/2021/08/25/fact-sheet-biden-administration-and-private-sector-leaders-announce-ambitious-initiatives-to-bolster-the-nations-cybersecurity/.

[3] U. B. of Labor Statistics, "Occupational outlook handbook, information security analysts," last accessed, August 12, 2022. https://www.bls.gov/ooh/computer-and-information-technology/information-security-analysts.htm.

[4] Cybint, "15 alarming cyber security facts and stats," December 23, 2020. https://www.cybintsolutions.com/cyber-security-facts-stats/.

[5] Wikipedia, "Facebook–cambridge analytica data scandal," last accessed August 8, 2022. https://en.wikipedia.org/wiki/Facebook-Cambridge_Analytica_data_scandal.

[6] R. R. Ruiz, "F.c.c. fines AT&T $25 million for privacy breach," last accessed August 8, 2022. https://bits.blogs.nytimes.com/2015/04/08/f-c-c-fines-att-25-million-for-privacy-breach/?r%20=%200.

[7] L. Clason, "Mississippi medical center hit with $2.75m fine for privacy breach," last accessed August 8, 2022. https://thehill.com/policy/healthcare/289131-mississippi-medical-center-hit-with-275m-fine-for-privacy-breach/

[8] IBM, "Cost of a data breach report 2021," last accessed, August 8, 2022. https://www.dataendure.com/wp-content/uploads/2021_Cost_of_a_Data_Breach_-2.pdf.

[9] J. Fitzgerald, "The 10 biggest data breaches of 2022 (so far)," July 01, 2022. https://www.crn.com/news/security/the-10-biggest-data-breaches-of-2022-so-far-.

[10] I. C. S. Institute and U. of California-Berkeley, "Teaching privacy," last accessed August 15, 2022. https://teachingprivacy.org/.

[11] L. Sweeney, "k-anonymity: A model for protecting privacy," *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.

[12] F. Prasser, F. Kohlmayer, R. Lautenschläger, and K. A. Kuhn, "Arx - a comprehensive tool for anonymizing biomedical data," in *AMIA Annual Symposium Proceedings*, pp. 984–993, American Medical Informatics Association, 2014.

[13] R. Zafarani and H. Liu, "Connecting users across social media sites: a behavioral-modeling approach," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, SIGKDD'08, pp. 41–49, 2013.

[14] T. Okuno, M. Ichino, T. Kuboyama, and H. Yoshiura, "Content-based de-anonymization of tweets," in *Proceedings of the Seventh International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 53–56, oct 2011.

[15] T. Okuno, M. Ichino, T. Kuboyama, and H. Yoshiura, "Re-identification of anonymized cdr datasets using social network data," in *Proceedings of the Third IEEE International Workshop on the Impact of Human Mobility in Pervasive Systems and Applications*, pp. 237–242, mar 2014.

[16] J. Qian, X.-Y. Li, C. Zhang, L. Chen, T. Jung, and J. Han, "Social network de-anonymization and privacy inference with knowledge graph model," *IEEE Transactions on Dependable and Secure Computing*, 2007.

[17] W. Peng, F. Li, X. Zou, , and J. Wu, "A two-stage deanonymization attack against anonymized social networks," *IEEE Transactions on Computers*, vol. 63, no. 2, 2014.

[18] S. Nilizadeh, A. Kapadia, and Y.-Y. Ahn, "Community-enhanced de-anonymization of online social networks," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 537–548, 2014.

[19] S. Ji, W. Li, M. Srivatsa, and R. Beyah, "Structural data de-anonymization: Theory and practice," *IEEE/ACM Transactions on Networking*, vol. 24, no. 6, 2016.

[20] K. Sharad and G. Danezis, "An automated social graph de-anonymization technique," in *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, pp. 47–58, 2014.

[21] S. Ji, W. Li, M. Srivatsa, J. S. He, and R. Beyah, "General graph data de-anonymization: From mobility traces to social networks," *ACM Transactions on Information and System Security*, vol. 18, no. 4, 2016.

[22] H. Fu, A. Zhang, and X. Xie, "Effective social graph deanonymization based on graph structure and descriptive information," *ACM Transactions on Intelligent Systems and Technology*, vol. 6, no. 4, 2015.

[23] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in Proceedings of the 2009 *IEEE Symposium on Security and Privacy*, pp. 173–187, 2009.

[24] A. Narayanan, E. Shi, and B. I. P. Rubinstein, "Link prediction by de-anonymization: How we won the kaggle social network challenge," in *Proceedings of the 2011 International Joint Conference on Neural Networks*, pp. 1825–1834, 2011.

[25] K. Henderson, B. Gallagher, L. Li, L. Akoglu, T. Eliassi-Rad, H. Tong, and C. Faloutsos, "It's who you know: Graph mining using recursive structural features," in *Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining*, pp. 663–671, 2011.

[26] M. Korayem and D. J. Crandall, "De-anonymizing users across heterogeneous social computing platforms," in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, 2013.

[27] "Kdd cup 2012, track 1, predict which users (or information sources) one user might follow in tencent weibo.," 2012.

[28] M. Franz, C. T. Lopes, G. Huck, Y. Dong, O. Sumer, and G. D. Bader, "Cytoscape.js: a graph theory library for visualisation and analysis," *Bioinformatics*, vol. 32, no. 2, 2016.

[29] N. Li, N. Zhang, and S. Das, "Preserving relation privacy in online social network data," *IEEE Internet Computing*, vol. 15, no. 3, pp. 35– 42, 2011.

[30] Y. Liu and N. Li, "Retrieving hidden friends: A collusion privacy attack against online friend search engine," *IEEE Transactions on Information Forensics and Security*, 2018.

[31] N. Li, "Privacy-aware display strategy in friend search," in *Proceedings of IEEE International Conference on Communications (ICC), Communication and Information Systems Security Symposium*, pp. 951–956, 2014.

[32] "Get followers/list," n.d.

[33] A. Geitgey, "Face recognition," last accessed, July 20, 2022. https://github.com/ageitgey/face_recognition.

[34] H. Kido, Y. Yanagisawa, and T. Satoh, "An anonymous communication technique using dummies for location-based services," in *Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops*, pp. 88–97, 2005.

[35] A. R. Beresford and F. Stajano, "Mix zones: User privacy in location-aware services," in *Proceedings of International Conference on International Conference on Pervasive Services*, pp. 127–131, 2004.

[36] B. Gedik and L. Liu, "Location privacy in mobile systems: A personalized anonymization model," in *Proceedings of 25th IEEE*

International Conference on Distributed Computing Systems, pp. 620–629, 2005.

[37] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas, and Y. Zhou, "Understanding the mirai botnet," in *26th USENIX Security Symposium (USENIX Security 17)*, (Vancouver, BC), pp. 1093–1110, USENIX Association, Aug. 2017.

[38] W. Du, "Seed," last accessed August 8, 2022. https://seedsecuritylabs.org/.