Attack Surface Reduction to Minimize Private Data Loss from Breaches

George O. M. Yee
Computer Research Lab, Aptusinnova Inc., Ottawa, Canada
Dept. of Systems and Computer Engineering, Carleton University, Ottawa, Canada
e-mail: george@aptusinnova.com, gmyee@sce.carleton.ca

Abstract— **Organizations are increasingly being victimized by breaches of private data, resulting in heavy losses to both the organizations and the owners of the data. For organizations, these losses include large expenses to resume normal operation and damages to its reputation. For data owners, the losses may include financial loss and identity theft. To defend themselves from such data breaches, organizations install security controls (e.g., encryption) to secure their vulnerabilities. While such controls help, they are far from being fool proof. Reducing the attack surface is a sound core approach for protecting valuable data. This paper applies this reduction to minimize the data loss from e-commerce data breaches. The paper first examines the behaviour of Business-to-Consumer (B2C) e-commerce companies in terms of why they collect and store personal data. It then applies attack surface reduction by limiting the amount of private data that the company stores in its computer system, while preserving the company's ability to accomplish its purposes for collecting the private data. The paper illustrates the approach by applying it to different types of B2C e-commerce companies.**

Keywords-attack surface reduction; minimizing data loss; data breach; private data loss; B2C e-commerce.

I. INTRODUCTION

This work extends Yee [1] by a) relating the approach to attack surface reduction, b) improving the explanations throughout the paper, as well as updating the examples of breaches in Section I, c) showing mathematically how the approach reduces the risk of data loss, d) adding application examples, and e) increasing the number of references.

Data breaches of personal data or personal information are appearing more and more often in the news, devastating the victim organizations. The losses have serious negative consequences both to the consumer (e.g., financial loss, identity theft) and to the organization (e.g., loss of reputation, loss of trust). Breaches of private data held by companies and other types of organizations have been occurring at an alarming rate. Each year has been accompanied by its assortment of data breaches. Consider the following sampling of breaches in 2022 [2], the year of this work:

- August, 2022: Up to 20 Million Plex Users Compromised. Plex offers streaming services for movies, music, and games, and hosts user-produced audio and visual content. Plex informed its customers on August 24 that it suffered a data breach impacting most of its user accounts. The private data loss included usernames, email addresses, and passwords of approximately 20 million users.
- July, 2022: 69 Million Accounts Exposed in Neopets Breach. Neopets is a virtual pet website where users can own virtual pets and buy virtual items for them. On July 19, 2022, a hacker posted data on 69 million Neopets users for sale on an online forum. The private data loss included name, email address, date of birth, zip code, and more, as well as 460 MB of compressed source code for the Neopets website.
- June, 2022: Up to 2 Million People Compromised in Shields Health Care Group Breach. The Massachusetts-based Shields Health Care Group disclosed in June, 2022, that they had detected a breach in March, 2022. The loss of private data included names, social security numbers, medical records, and other sensitive personal information.

In response to these attacks, organizations attempt to identify the vulnerabilities in their computer systems and secure these vulnerabilities using security controls. Example security controls are firewalls, intrusion detection systems, encryption, two-factor authentication, and social engineering awareness training for employees. Unfortunately, securing vulnerabilities with security controls is far from being foolproof. One major weakness is that it is impossible to find all the vulnerabilities in a computer system. This means that it is highly likely that a determined attacker will find an attack path into the organization's system that has been overlooked and cause a data breach, even though the organization believes that it has done due diligence and secured all its vulnerabilities. Nevertheless, security controls do help to prevent breaches, and we are not advocating that they be eliminated. Rather, the approach in this work can be considered as an addition to the existing arsenal of security controls.

In this work, we propose an approach in which most of the private data collected by an organization is stored on the user's device. Thus, a smaller quantity of private data remains on the company's computer system, reducing the system's attack surface and minimizing the loss of private data should the company-stored data ever be breached. The approach also ensures that the needs of the company to carry out its

purposes for collecting the private data are satisfied. The user's device could be a desktop computer, a laptop, or a smart phone. The approach is intended for Business-to-Consumer (B2C) e-commerce companies, since B2C companies appear to collect large quantities of personal data and are often victimized by data breaches. Note that in this work when we write about data storage on or in the "company's computer system", we mean that the data is stored on company premises or in the cloud.

This paper is organized as follows. Section II looks at private data, attacks, and attack surface. Section III examines the behaviour of B2C companies in terms of why they collect and store personal information. It also looks at the nature of the collected information. Section IV presents the approach, including a mathematical description of how it reduces the risk of data loss. Section V gives examples of how the approach can fit with different types of e-commerce companies. Section VI describes related work. Section VII gives conclusions and future work.

II. Private Data, Attacks, and Attack Surface

This section explains private data, attacks, and attack surface.

A. Private Data, Attacks, and Attack Surface

Private data consists of information about a person that can identify or be linked to that person and is owned by that person [3]. Thus, private data is also "personal information", and consists of "personal data". For example, a person's height, weight, or credit card number can all be used to identify the person and are considered as personal information. There are other types of personal information, such as buying patterns and navigation habits (e.g., websites visited) [4]. An individual's privacy refers to his/her ability to control the collection (what private data and collected by which party), purpose of collection, retention, and disclosure of that data, as stated in the individual's privacy preferences [3]. In many countries, private data is protected by legislation in which the concept of "purpose" for collecting the personal information (how the collected information will be used) is important. Companies must disclose the purpose for collecting the personal information and cannot use the information for any other purpose. Private data needs protection and must not fall into the wrong hands.

DEFINITION 1: An *attack* is any action carried out against an organization's computer system that, if successful, results in the system being compromised.

This work focuses on attacks that compromise the private data (PD) held in the online systems of organizations. The attacker who launches an attack may be internal (inside attacker) or external (outside attacker) to the organization. An internal attacker usually has easier access to the targets of his/her attack and he/she may hide his/her attacks in the guise of normal duty. This work focuses on outside attackers. Reference [5] gives a good account of how to mitigate insider attacks.

Salter et al. [6] give an interesting insight into what enables a successful attack: "Any successful attack has three steps: One, diagnose the system to identify some attack. Two, gain the necessary access. And three, execute the attack. To protect a system, only one of these three steps needs to be blocked." Thus, an attack surface must contain a target that the attacker deems worthy of attack (suit his/her purpose for the attack) and that target must be accessible to the attacker. For this work, the target that is potentially worthy of attack is the PD that is accessible to attackers. In a computer system, this PD is either moving (travelling from one location to another), at rest (stored), or being used (by some process). This leads to the following definition of attack surface:

DEFINITION 2: The *attack surface* for private data, also called the *private data attack surface*, contained in an online computer system is the set of all locations in the system that contain attacker accessible PD in the clear, where the PD is moving, at rest, or being processed.

In Definition 2, "attacker accessible PD" means that the attacker is able to exfiltrate the PD using some agent of attack, such as malware against stored PD and PD being processed, or a man-in-the-middle attack against a link containing moving PD. Also, we assume that attackers would attack PD that is in the clear rather than PD that is encrypted. In the rest of this paper, by "attack surface" we mean the private data attack surface, unless otherwise indicated. Figure 1 shows an example private data attack surface.

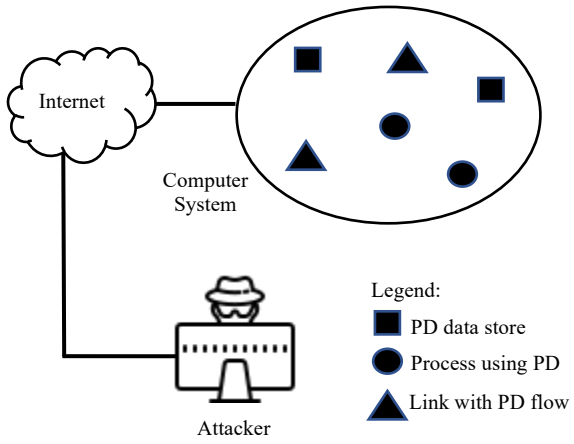


Figure 1. Example private data attack surface consisting of the set of all 6 attacker accessible locations in the system that contain PD in the clear.

An alternative definition of attack surface for PD contained in a computer system is the set of ways the attacker has to exfiltrate the PD. However, given the complexity of computer systems and the fact that the tools available to the attacker to use in his/her attacks are unknown to us, it is next to impossible to determine this set. On the other hand, locations that contain attacker accessible PD are easier to identify. Since an exfiltration must be from a location that contains PD, the set of such exfiltrations depends on the set of such locations. The larger the set of locations, the larger the set of exfiltrations. The smaller the set of locations, the smaller the set of exfiltrations. Therefore, Definition 2 in a

sense includes this alternative definition, but in addition, is more easily applied.

As mentioned above, in the first step of a successful attack, the attacker diagnoses the system to identify the attack [6]. A smaller attack surface will make this step more difficult for the attacker. Therefore, a smaller attack surface corresponds to higher security, which is why we wish to reduce the attack surface. Definition 2 also gives rise to this conclusion: a smaller attack surface means a smaller number of locations that contain PD, which in turn means fewer opportunities for exfiltration of the PD, or in other words, higher security.

Definition 2 is consistent with the intuitive understanding of an attack surface (the usual meaning), which is "the set of ways in which an adversary can enter the system and potentially cause damage" [7]. Each "way" corresponds to a location in Definition 2 that in turn corresponds to methods for exfiltrating PD from the location.

III. The Collection and Storage of Personal Information by B2C Companies

In this section we examine why B2C companies collect personal information and discuss the nature of this information.

A. Purposes for Collecting Personal Information

Companies engaged in B2C e-commerce, collect personal information for the following purposes:

- Transaction Requirements (self-evident): Personal information is needed and used in carrying out the transaction. For example, making an online purchase requires your name and address for goods delivery.
- Communication (self-evident): Personal contact information is needed to communicate with customers for resolving order issues or to answer product questions.
- To Secure Other Data: A personal biometric is needed for further authentication, e.g., a voice print, prior to allowing the customer to access more secure areas of his or her account [8]. The biometric may also be required for use in multi-factor authentication.
- Establishing Loyalty: A personal history of past transactions may be required to establish a customer's loyalty in order to reward the customer with certain benefits such as free shipping or product discounts [9].
- Targeted Advertising: A personal history of past transactions is needed to understand the type of products a particular customer has purchased in the past, and thereby create more appealing and effective ads directed at the customer [10].
- Market Research: The personal histories of past transactions for all customers are studied in order to understand what products appeal to customers in order to make decisions for stock purchases, or to provide a better customer experience in terms of app or website design [8].
- Sharing or Selling: Personal information collected is shared or sold to other organizations for a profit [8].

B. E-Commerce Data

In B2C e-commerce, online companies sell items and services to consumers. Example types of such companies include sellers of goods and services (e.g., Amazon.com), hotels (e.g., Mariott.com), travel agencies (e.g., Expedia.ca), financial services (e.g., CIBC.com), and the list goes on. All these companies share common data types. Each company offers products that customers purchase. Table 1 identifies the products for the e-commerce company types mentioned above.

Each customer has a set of personal identifying information, such as name, postal address, and phone number that identify the customer, and depending on the service provided by the company, include personal information such as credit card details, date of birth, amount of mortgage on house, and so on. We group all such personal identifying information under the heading Customer Personal Data (CPD). Each customer makes one or more product selections and effects payment for the product(s) selected. In addition, there is ancillary data, such as type of payment, date ordered, date shipped, date delivered (from delivery agent, e.g., courier), and so on. Table 2 shows these data types and whether they originate from the company or the customer.

TABLE 1. PRODUCTS ASSOCIATED WITH EACH COMPANY TYPE.

Company type	Products
Sellers of goods and services (e.g., Amazon.com)	Physical items such as pots, clothing, and electronics; services such as selling your items for you
Hotels (e.g., Mariott.com)	Rooms
Travel Agencies (e.g., Expedia.ca)	Travel bookings
Financial services (e.g., CIBC.com)	Fee-based banking accounts

TABLE 2. DATA TYPES AND WHERE THEY ORIGINATE.

Data type	Origin
Products	Company
CPD	Customer
Product selection	Customer
Amount paid	Company
Ancillary data	Company

We can see that each online customer order involves the data types shown in the left column of Table 2. Depending on the company, the instantiation of these data types will be different, with the possible exception of Amount paid. For example, the "Products" of Amazon.com would be different from the "Products" of eBay.com and the CPD for CIBC.com may be different from that for TD.com (another Canadian bank). Thus, each customer order may be represented by a data collection as shown in Figure 2. We wish to emphasize

that there is no implied ordering of the data types in Figure 2, i.e., Figure 2 does not state that the data types should be stored in any particular order one after the other. These data collections would be stored by the company in its own databases, which may be on company premises or on a cloud server. If the company were to suffer a data breach, this data (including CPD) would be exposed.

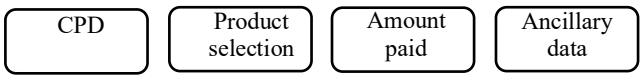


Figure 2. Data collection for a customer order.

IV. Approach

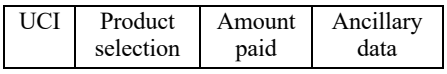
This section details our approach for minimizing the loss of PD from data breaches.

A. Strategy for Storing a Customer's Personal Data

The goal of this strategy is to reduce the storage of personal data on the company's computer system by storing the bulk of the personal data on customers' own devices, while allowing for all the purposes described in Section III-A to be carried out. The strategy consists of five parts, as follows:

1. Identification of data (Figure 2) to be stored on the customer's device: CPD.
2. Design for linking the data on the customer's device to the rest of the data stored on the company's computer system: Use a Unique Customer Identifier (UCI) that the company assigns to each customer. The UCI is the hash (e.g., SHA-3) of the customer's User ID and password for accessing the company. It will form part of the records shown in Figure 3 (shown as relational records without loss of generality since we could have shown them as other types of data structures, e.g., linked lists).

a) Record of personal data stored on customer's device.



b) Record of order data stored on company's system.

Figure 3. Data records corresponding to a customer order. Encrypted data types are shaded.

3. Design for enabling the company to carry out its communication purpose: Use the "Contact information" data record in Figure 4 to contact the customer, where "Contact information" consists of email address and telephone number. Figure 5 shows how the UCI links the three types of data records together.
4. Design to keep the CPD record should the customer a) use a new device with the company after using other devices, or b) loses a device used with the company. For a), the customer can register a new device with the company on its website after logging in. The company would then transfer the CPD record from a previously used device (on which the customer is also logged in) to the new device. For b), the customer may have used other devices with the company and wishes to replace the lost device, in which case the resolution for a) applies. If the lost device is the only device used with the company, the customer would need to re-enter his/her CPD. See also the third paragraph of Section IV-C below.

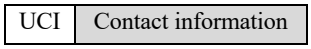


Figure 4. Data record for a customer's contact information. Encrypted data types are shaded

5. Enabling security: Use authenticated symmetric encryption (e.g., AES-GCM [11]) to encrypt the UCI and CPD in Figure 3(a), as well as the Contact information in Figure 4 (encrypted data types are shaded). The UCI in Figure 4 is not encrypted. The UCI and remaining data types in Figure 3 (b) are not encrypted, as it would be difficult for the attacker to use them alone to identify the customer, should the data be breached.

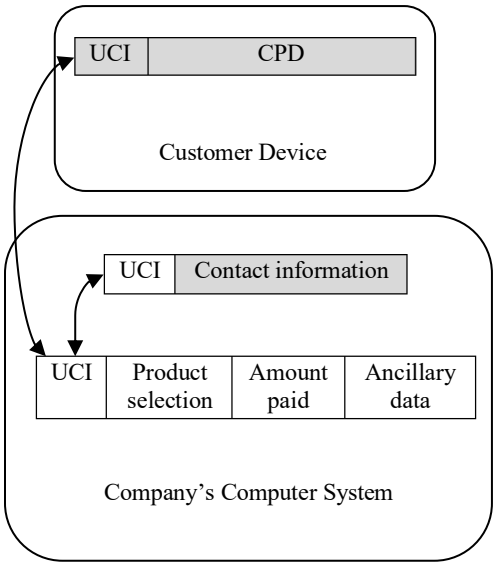


Figure 5. How the UCI links data records together.

B. Customer Walk-Through of the Strategy

1. The customer accesses the company using its website, running either on a desktop computer or on a mobile device such as a smart phone or tablet. In the following, all data transfers between the user's device and the company's system is done though a secure channel (e.g., TLS).
2. If it's the customers first use of the website on this device (detected by the absence of the CPD record), he/she will be asked if he/she has a different device that was used

with the website. If not, he/she will be prompted to enter his/her CPD. The company then generates the UCI, forms the record in Figure 3(a), encrypts it, and stores this encrypted record on the customer's device. The company then uses the unencrypted CPD entered by the customer for processing the current order. In addition, the company checks if the customer's Contact information is already in the system (possible if the customer's device was lost or stolen) and if not, creates and stores the record in Figure 4, after encrypting the Contact information (obtained from the CPD). If the customer has used the website before on a different device, he/she will be asked to also login using the other device, at which point the company stores the CPD record from the old device on the new device, decrypts the CPD record, and uses it for the current transaction.

If the customer has used the website before on this device (detected by the presence of the encrypted CPD record), the company automatically retrieves the encrypted CPD record (Figure 3(a)) from the customer's device and decrypts it for use in the current transaction.

Note that the only time the company retrieves the CPD record from a customer device is when the customer logs in to do a new transaction.

3. The customer proceeds with his/her shopping. Once the customer completes the shopping, the company creates and stores the customer's order data record as shown in Figure 3(b). Note that this record may have to be updated for some ancillary data (e.g., date delivered) once the data is available. This update process as out of scope for this work.

Figure 6 shows a message sequence diagram illustrating the case where the customer uses a device with the company's system for the first time and has not used any other device with the company in the past. Figure 7 presents a message sequence diagram for the case where the customer uses a device with the company that he/she has used before. Figure 8 gives a message sequence diagram depicting the case where the customer uses a device with the company for the first time and has used a different device with the company before.

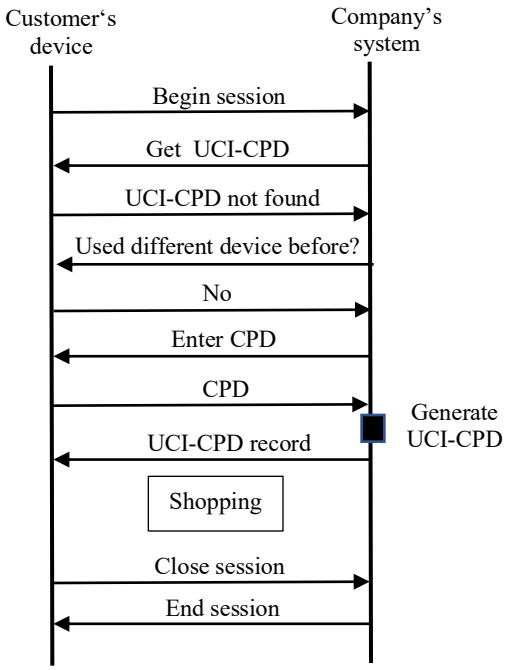


Figure 6. Customer uses a device with the company for the first time and has not used any other device before.

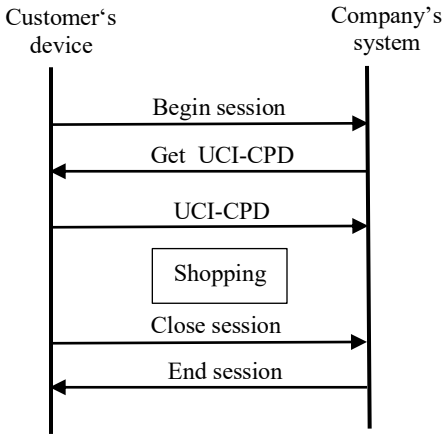


Figure 7. Customer uses a device with the company that he/she has used before.

C. Security Analysis

We first consider outside attacks against the company. Such attacks would result in breaching the company's data stores leading to the loss of the Contact information and the order data (Figure 5). This loss could be in the form of a copy taken of the data, deletion of the data from the company's data stores, modification of the data in the company's data stores, or certain combinations of these, namely copy followed by deletion, and copy followed by modification. However, the attacker fails to read the Contact information since it is encrypted. The attacker would be able to read the UCI from both the Contact information and the order data records but the UCI would appear as meaningless (hash). The attacker could also read the order data but would have a hard time identifying the customer using only this data. Further, deleting or modifying the data will also fail to damage the company provided that the company is aware of the attack and is able to re-populate the data stores using data back-ups. We assume that the company has implemented other security measures, including making data backups and having ways to detect attacks (e.g., intrusion detection system). Any modification of the encrypted Contact information would also be detected by a failure to decrypt the modified version, i.e., the modified encrypted data fails authentication. Note that for the rest of this paper, whenever we refer to failing to decrypt attacker-modified encrypted data, we mean that the modified encrypted data has failed authentication. In any case, the probability of being attacked after applying the approach is low, since the only attraction for attackers is

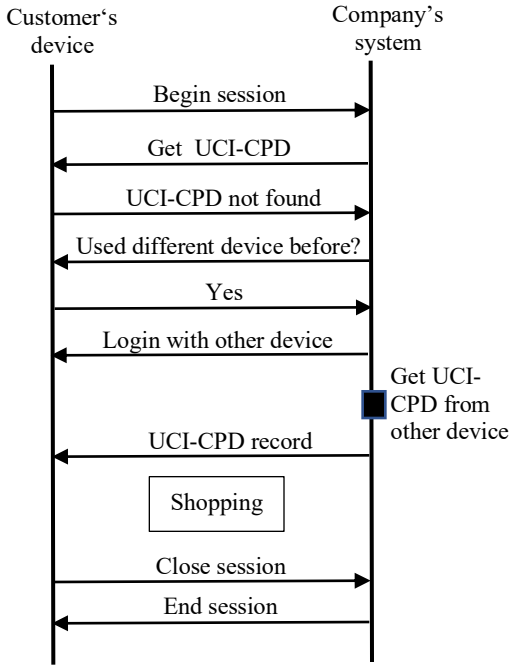


Figure 8. Customer uses a device with the company for the first time, having used a different device with the company before.

encrypted Contact information, consisting only of email address and telephone number. Attacks on the company side could also involve malware, that for example, exfiltrates the customer's CPD while in the clear. However, these attacks are not peculiar to the approach and can occur for any website that collects information from users. We assume that the company already has security measures for such attacks.

As for insider attacks against the company, we admit that our security scheme is vulnerable to such attacks. For example, an insider could simply access the CPD in its unencrypted form. Insider attacks are always among the harder ones to defend against and given their seriousness, we expect the company to have implemented other security measures (e.g., [5]) specifically against insider attacks. An exploration of these measures is outside the scope of this paper.

Attacks on the customer side with the device in the customer's possession or not (device lost or stolen) could also result in a copy taken of the customer's CPD record, deleting it, modifying it, or combinations thereof. Since the data is encrypted, the attacker would not be able to read the data if a copy is taken. Deletion or modification of the encrypted CPD record would be detected by the company's system when it fails to find it or fails to decrypt it, in which case the company's system would inform the customer that he/she needs to re-enter his/her CPD or have it transferred from another device (see Section IV-A, part 4).

The secure communication channel between the company's system and the customer device may also be attacked, but this is again not peculiar to the approach. Such attacks would be handled the same way as is done for the many other applications of secure communication channels.

D. Implementation Notes

The following are suggestions on how the above strategy should be implemented.

- On the company side, the implementation should include functionality to warn that its data stores have been compromised when it is unable to decrypt attacker-modified encrypted data, or when it finds its data stores empty. The implementation should also warn the customer that his/her device has been attacked when the encrypted CPD record was expected but is missing, or when it is unable to decrypt the attacker-modified record.
- If the customer changes or forgets his/her password for accessing the service (if forgotten, a conventional password reset procedure would be used), the company's computer system will need to generate a new UCI corresponding to the new user-ID/password combination. The company will have to create a new CPD record with the new UCI, and upload this new record to all customer devices via the website. The company will also have to update the UCI in the records of Figure 3(b) and Figure 4.
- The company's system needs to allow the customer to update his/her CPD and/or Contact information, and update the relevant records with the new information. For CPD, the system would need to upload an updated CPD record to all customer devices.

E. Verification of Purposes

We verify that the approach allows the company to carry out its purposes (Section III-A) for collecting private data.

- Transaction Requirements: The customer's CPD record is obtained from the customer's device for every transaction (either pre-existing or currently entered) and is available for carrying out the transaction.
- Communication: For contacting the customer, the customer's Contact information (Figure 4) can be obtained using the UCI link from the order data records since contacting is done for an order issue. The customer can contact the company by logging into the company's website. The company can determine the customer's UCI from the customer's User ID and password, and use it to access the contact information for the reply.
- To Secure Other Data: The personal biometric, once captured, can be stored as part of the customer's CPD record on the customer's device. Once the customer logs in for a new transaction, the CPD record is retrieved from the customer's device, at which point the personal biometric is available for use.
- Establishing Loyalty: The company has access to a customer's order history in the form of the order data records. These records (Figure 3(b)) are identified as belonging to a particular customer through the UCI link to the Contact information records. The company can thus establish the loyalty of a particular customer.

- Targeted Advertising: Understanding the type of products a customer has purchased in the past may be done by accessing the customer's order data records, as explained above for establishing loyalty.
- Market Research: The histories of past transactions for all customers can be studied by accessing the order data records, ignoring the UCI in each order record, since there is no need to identify the customers. We assume that market research is carried out without the CPD records, since the company probably does not have the customer's consent for such use of his/her CPD. If the company does require the CPD records, the company can always capture and store them, but would have to accept the risks of those records being breached and being sued for illegally using the CPD for market research.
- Sharing or Selling: There is nothing stopping the company from copying each customer's CPD record and sharing or selling the data. The company would have to accept the risks of the CPD records being breached and being sued for illegally sharing or selling the customer's CPD.

F. Strengths and Weaknesses of the Approach

The approach has the following strengths: a) it is straight-forward, which may make it easier to "sell" to upper management for approval, b) it is efficient in that attackers would have to breach the devices of all the company's customers, in order to breach the same quantity of personal data that are traditionally all stored in the company's system, c) it minimizes the risk of data loss (see subsection G below), d) it makes the company less attractive to attackers who intend to cause a data breach due to its efficiency as stated above and the fact that the only private data left on the company's system to be breached is the encrypted customer Contact information, and e) it should please customers who want more control over their private data, since most of it is stored only on their own devices.

The approach seems to have three weaknesses: a) the storage/retrieval of the CPD record may attract attacks on the secure transmission channel, b) there is additional overhead cost due to encryption / decryption operations, and c) it is vulnerable to insider attack. Weakness a) does not represent significant extra risk over conventional transactions since personal data is transmitted in conventional transactions as well. For weakness b), the extra overhead should not be significant. Finally, weakness c) is not exclusive to this approach, since it can arise wherever there are insiders. Potential remedies include the installation of specific security measures to defend against insider attacks [5].

G. Showing that the Approach Minimizes the Risk of Data Loss

Our approach of having most of a user's private data stored on his/her computing device rather than on the company's system minimizes data loss according to beliefs 1 and 2 as follows:

1. Much less private data is lost in the event of a system breach, because the storage of most of the private data has been relocated to user devices, and
2. There is a much-reduced risk of theft of the users' private data if that data is stored on user devices rather than stored in the company's system.

Belief 1 is self-evident. To verify belief 2, compare Case 1 where a portion of each users' private data is stored on the system, with Case 2 where the portions of private data in Case 1 are instead stored on user devices. Let D and D_i represent the private data in Cases 1 and 2 respectively, where D_i is the private data belonging to user i. Let E be the event that D is stolen in Case 1. Let E_i be the event that D_i is stolen from user i in Case 2. Let $P(E) = p$ where $P(E)$ is the probability of E. Finally, let $P(E_i) = q_i$. Figure 9 illustrates D and D_i . We postulate that for n users,

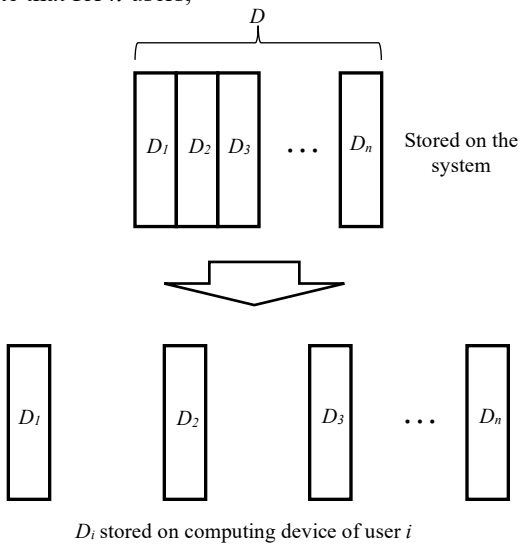


Figure 9. Moving the storage of private data from the server to user devices.

$$P(E_1 \cap E_2 \cap \ldots \cap E_n) \ll P(E) \quad (1)$$

meaning that the risk of theft of all the private data moved to user devices from the system (Case 2) is much lower than the risk of theft of that same data were it to remain on the system (Case 1), which is a statement of belief 2 above. Thus, to verify belief 2, we need to prove (1). To do this, let C be the event that an attacker chooses to attack the system. Let C_i be the event that an attacker chooses to attack the computing device of user i. Let S be the event that the attacker successfully defeats the security controls of the system. Let S_i be the event that the attacker successfully defeats the security controls of user i's device. We note that

$$P(E) = P(C)P(S|C) = p \quad (2)$$
$$P(E_i) = P(C_i)P(S_i|C_i) = q_i \quad (3)$$

What can we say about the conditional probabilities here? Since D has a lot more private information than D_i, an attacker would be more likely to choose D over D_i as his/her target. In other words, a company is a more attractive target than a user device. Thus, $P(C)>P(C_i)$ for all i. Further, the attacker would be more motivated to defeat the security controls of the company compared to the security controls of the user device, again due to the attractiveness of the company as a target. Thus, $P(S|C)>P(S_i|C_i)$ for all i. Equations (2) and (3) then give $p > q_i$ for all i. Now since the E_i are independent events, we have

$$P(E_1 \cap E_2 \cap \ldots \cap E_n) = \prod_1^n P(E_i) \quad (4)$$
$$= \prod_1^n q_i \quad (5)$$
$$< p^n \ll p = P(E) \quad (6)$$

proving (1) as we set out to do. Note that (5) follows from (4) due to $p > q_i$ and (6) follows from (5) due to the fact that p is a probability with $0 < p < 1$.

Another way to reason about (1) is simply to notice that the product in (4) decreases monotonically with increasing n due to the fact that the q_i are probabilities between 0 and 1. Thus, since $P(E)$ is fixed, (1) will be true for sufficiently large n. Since we are dealing with systems that have many users, it would not be difficult to achieve sufficiently large n.

We now have beliefs 1 and 2 both true, meaning that storing the private data on user devices instead of on the system does indeed minimize the risk of data loss.

V. Application Examples

We instantiate the data types in Figure 2 for four types of B2C companies, demonstrating that the approach can fit with different B2C companies.

Example 1: Seller of goods (e.g., Amazon.com). Table 3 shows the instantiation of the data types for this example.

TABLE 3. INSTANTIATION OF DATA TYPES FOR EXAMPLE 1.

CPD	Product selection	Amount paid	Ancillary data
Name	Camera	$159.00	Date ordered
Billing address	Hair clipper	$49.00	Date shipped
Default shipping address	Laser printer toner	$68.00	Date delivered
Alternate shipping address			Payment method
Email address			Product returned
Phone number			Reason for return
Credit card data			Refund status

Example 2: Travel Agency (e.g., Expedia.ca). Table 4 shows the instantiation of the data types for this example.

TABLE 4. INSTANTIATION OF DATA TYPES FOR EXAMPLE 2.

CPD	Product selection	Amount paid	Ancillary data
Name	Vacation package	$1059.00	Date ordered
Billing address	Trip insurance	$189.00	Date mailed
Default address			Date delivered
Alternate address			Payment method
Email address			Product returned
Phone number			Reason for return
Credit card data			Refund status

Example 3: Hotel (e.g., Mariott.com). Table 5 shows the instantiation of the data types for this example.

TABLE 5. INSTANTIATION OF DATA TYPES FOR EXAMPLE 3.

CPD	Product selection	Amount paid	Ancillary data
Name	Room - double	$200 / night	Date of reservation
Billing address			Arrival date
Home address			Departure date
Email address			Payment method
Phone number			Airport shuttle y/n
Credit card data			Daily laundry y/n
Loyalty ID number			Daily cleaning y/n
Country of origin			Wake-up call y/n
Passport country			Stay extended y/n
Passport number			
Room preferences			
Floor preference			

Example 4: Online Training (e.g., Udemy.com). Table 6 shows the instantiation of the data types for this example.

TABLE 6. INSTANTIATION OF DATA TYPES FOR EXAMPLE 4.

CPD	Product selection	Amount paid	Ancillary data
Name	Guitar	$30.00	Date of purchase
Billing address	Photography	$50.00	Date training started
Home address	Programming	$60.00	Date training ended
Email address			Certificate issued y/n
Phone number			Comprehension test taken y/n
Credit card data			Comprehension score
Training type preferred			Comprehension score issued y/n
Training length preferred			

We could have included other examples here, but the above examples suffice for demonstrating that the approach can be applied to different types of B2C companies.

VI. Related Work

Work that is most closely related to this work are as follows: Aggarwal et al. [12] propose that an organization outsource its data management to two untrusted servers to break associations of sensitive information. They show how the use of two servers, together with the use of encryption where needed, enables efficient data partitioning and guarantees that the contents of any one server does not violate data privacy. However, it is unclear if attackers can re-construct the sensitivity associations by breaching both servers. Ciriani et al. [13] present what they claim to be a solution that improves over Aggarwal et al. [12] by first splitting the information to be protected into different fragments so that sensitive associations represented by confidentiality constraints are broken, and minimizing the use of encryption. The resulting fragments may be stored at the same server or at different servers. Our work differs from Aggarwal et al. [12] and Ciriani et al. [13] as follows: a) the above two papers are solutions for securing databases, whereas our work is focused on reducing the loss of data in the event of a data breach by simply not storing some of the data in the company's computer system, b) we do not use data partitioning or fragmentation; rather, our data is distributed between the company and its customers from the point of data creation, c) we do not need to rely on breaking any sensitivity associations, d) our approach has been designed to satisfy the business needs of the organization, and e) our approach is more straightforward, and is therefore easier to apply.

Other work in the literature mostly deal with the prevention or risks of data breaches, the discovery of a data breach, and the aftermath of a data breach. Within these categories, the most closely related works have to do with preventing or evaluating the risks of data breaches. We describe some of these papers below, to give the reader a sense of this research. Note that these works all differ from this paper in that this paper aims to minimize the data lost if a breach were to happen, whereas the works described in the following are largely focused on preventing breaches from happening. Panou et al. [14] describe a framework for monitoring and describing insider behaviour anomalies that can potentially impact the risks of a data breach. The framework also enhances a company's understanding of cybersecurity and increases awareness of the threats and consequences related to breaches, and eventually enable faster recovery from a breach. Guha and Kandula [15] propose a data breach insurance mechanism together with risk assessment methodology to cover the risk from accidental data breaches and encourage best practices to prevent the breaches. They also present data supporting the feasibility of their approach. Zou and Schaub [16] interviewed consumers after the Equifax data breach and discovered that consumers' understanding of credit bureaus' data collection practices was incomplete. As such, consumers did not take sufficient protective actions to deal with the risks to their data. The authors describe the implications of their findings for the design of future security tools with the aim of empowering consumers to better manage their data and protect themselves from future breaches. Nicho and Fakhry [17] look at the application of system dynamics to cybersecurity, specifically to the Advanced Persistent Threat (APT) that can employ technical, as well as organizational factors to cause a data breach. They applied system dynamics to the APT that led to the Equinox breach and identified key independent variables contributing to the breach. Their work provides insights into the dynamics of the threat and suggests "what if" scenarios to minimize APT risks that could lead to a breach. Luh et al. [18] present an ontology for planning a defence against APTs that can lead to a data breach. The ontology is mapped to abstracted events and anomalies that can be detected by monitoring and helps with the understanding of how, why, and by whom certain resources are targeted. Other references in this category are readily available.

In terms of identifying and reducing the attack surface, this work is unique in reducing the attack surface of a company's system by storing private data on user devices. This author has published works [19][20][21] that deal with reducing the attack surface during software design, by identifying vulnerabilities using a model of the software system under development. A. Kurmus et al. [22] look at reducing the attack surface of commodity OS kernels by identifying code that is not used and removing it or preventing it from executing. T. Kroes et al. [23] investigate reducing the attack surface through dynamic binary lifting, removal of unnecessary features, and recompilation. M. Sherman [24] investigates attack surfaces for mobile devices. This author claims that mobile devices exhibit attack surfaces in capabilities, such as communication, computation, and sensors, that are generally not considered in current secure coding recommendations. C. Theisen et al. [25] propose the use of risk-based attack surface approximation (RASA) which uses crash dump stack traces to predict what code may contain attackable vulnerabilities. Their goal is to help software developers prioritize their security efforts by providing them with an attack surface approximation. It is worthwhile noting that some works propose to increase security through attack surface expansion rather than attack surface reduction. For cloud services, T. Al-Salah et al. [26] propose three attack surface expansion approaches that use decoy virtual machines co-existing with the real virtual machines in the same physical host. They claim that simulation shows that adding the decoy virtual machines can significantly reduce the attackers' success rate. For enterprise networks, K. Sun and S. Jajodia [27] propose a new mechanism that expands the attack surface, so that attackers have difficulty in identifying the real attack surface from the much larger expanded attack surface. Note that these two works do not contradict reducing the attack surface to improve security, since the attack surface is not really expanded but only appears to be expanded due to the addition of decoys.

VII. Conclusion and Future Work

We have presented an attack surface reduction approach, applicable to B2C e-commerce companies, that minimizes the loss of private data in the event of a data breach by storing most of a customer's private data in his/her own device rather than in the company's computer system. This redistribution of private data reduces the attack surface of the company's system, minimalizing the amount of data that would be lost in a breach. Not all of the private data is moved to the customer's device since we still allow some necessary personal data (customer contact information) to be stored on the company's system. We also verified that the approach allows the company to carry out its purposes for collecting private data, which is an important requirement of any company that may wish to implement this approach. Some readers may consider the approach overly simple, but if a simpler solution gets the job done, it should be preferred over a more complex solution. As well, a large contribution of this work is showing how the approach can be done securely. We look forward to readers' feedback and correcting any inadvertent omissions, if found, in a future paper.

In terms of future work, we would like to explore the application of the approach to other types of businesses and organizations, and adapt it where necessary. We would also like to have implementations of the approach in order to fine tune it, measure implementation effort, and check performance.

Acknowledgment

The author is grateful to Aptusinnova Inc. for financially supporting this work. He expresses his thanks to the reviewers of this paper for their insightful comments.

References

[1] G. Yee, "Towards Reducing the Impact of Data Breaches," Proc. Fourteenth International Conference on Emerging Security Information, Systems and Technologies (SECURWARE 2020), November 2020, pp. 75-81.

[2] M. Heiligenstein, "Top 10 Biggest Data Breaches of 2022 – So Far," Firewall Times, [retrieved: October, 2022] https://firewalltimes.com/biggest-data-breaches-2022/

[3] G. Yee, "Visualization and Prioritization of Privacy Risks in Software Systems," International Journal on Advances in Security, issn 1942-2636, vol. 10, no. 1&2, pp. 14-25, 2017, [retrieved: Dec., 2020] http://www.iariajournals.org/security/

[4] E. Aïmeur and M. Lafond, "The scourge of Internet personal data collection," Proc. 2013 International Conference on Availability, Reliability and Security (AReS 2013), Sept. 2013, pp. 821-828.

[5] CERT National Insider Threat Center, "Common Sense Guide to Mitigating Insider Threats, Sixth Edition," Technical Report CMU/SEI-2018-TR-010, Software Engineering Insititute, Carnegie Mellon University, December 2018.

[6] C. Salter, O. Sami Saydjari, B. Schneier, and J. Wallner, "Towards a Secure System Engineering Methodology," Proceedings of New Security Paradigms Workshop, Sept. 1998, pp. 2-10.

[7] P. K. Manadhata and J. M. Wing, "An Attack Surface Metric," IEEE Transactions on Software Engineering, vol. 37, no. 3, pp. 371-386, May/June, 2011.

[8] Business News Daily, "How businesses are collecting data (and what they're doing with it)," [retrieved: October, 2020] https://www.businessnewsdaily.com/10625-businesses-collecting-data.html

[9] R. Sarcar, "How to set up an ecommerce loyalty program to improve retention, build community and drive 5X in sales," [retrieved: October, 2020] https://www.bigcommerce.com/blog/online-customer-loyalty-programs/#how-to-create-and-implement-a-customer-loyalty-program

[10] PC, "How companies turn your data into money," [retrieved: October, 2020] https://www.pcmag.com/news/how-companies-turn-your-data-into-money

[11] M. Dworkin, "Recommendation for block cipher modes of operation: Galois/Counter Mode (GCM) and GMAC," NIST Special Publication 800-38D, November 2007.

[12] G. Aggarwal et al., "Two can keep a secret: a distributed architecture for secure database services," Proc. Second Biennial Conference on Innovative Data Systems Research (CIDR 2005), Jan. 2005, pp. 1-14.

[13] V. Ciriani et al., "Combining fragmentation and encryption to protect privacy in data storage," ACM Transactions on Information and System Security (TISSEC), Vol. 13, Issue 3, article 22, pp. 1-33, July 2010.

[14] A. Panou, C. Ntantogian, and C. Xenakis, "RiSKi: A framework for modeling cyber threats to estimate risk for data breach insurance," Proc. 21st Pan-Hellenic Conference on Informatics (PCI 2017), article no. 32, Sept. 2017, pp. 1-6.

[15] S. Guha and S. Kandula, "Act for affordable data care," Proc. 11th ACM Workshop on Hot Topics in Networks (HotNets-XI), Oct. 2012, pp. 103-108.

[16] Y. Zou and F. Schaub, "Concern but no action: consumers' reactions to the Equifax data breach," Extended Abstracts, 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18), paper no. LBW506, Apr. 2018, pp. 1-6.

[17] M. Nicho and H. Fakhry, "Applying system dynamics to model advanced persistent threats," Proc. 2019 International Communication Engineering and Cloud Computing Conference (CECCC 2019), Oct. 2019, pp 29-33.

[18] R. Luh, S. Schrittwieser, and S. Marschalek, "TAON: an ontology-based approach to mitigating targeted attacks," Proc. 18th International Conference on Information Integration and Web-based Applications and Services (iiWAS '16), Nov. 2016, pp. 303-312.

[19] G. Yee, "Reducing the Attack Surface for Sensitive Data," International Journal on Advances in Security, issn 1942-2636, vol. 13, no. 3&4, pp. 109-120, 2020, [retrieved: Oct., 2022] http://www.iariajournals.org/security/

[20] G. Yee, "Modeling and Reducing the Attack Surface in Software Systems," Proceedings, 11th Workshop on Modelling in Software Engineering (MiSE'2019), May 2019, pp. 55-62.

[21] G. Yee, "Attack Surface Identification and Reduction Model Applied in Scrum," Proceedings, 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), June 2019, pp. 1-8.

[22] A. Kurmus, A. Sorniotti, and R. Kapitza, "Attack Surface Reduction for Commodity OS Kernels: Trimmed Garden Plants May Attract Less Bugs," Proceedings of the Fourth European Workshop on System Security (EUROSEC '11), April 2011, article no. 6 (no page number available).

[23] T. Kroes, A. Altinay, J. Nash, Y. Na, and S. Volckaert, "BinRec: Attack Surface Reduction Through Dynamic Binary Recovery," Proceedings of the 2018 Workshop on Forming an Ecosystem Around Software Transformation (FEAST '18), October 2018, pp. 8-13.

[24] M. Sherman, "Attack Surfaces for Mobile Devices," Proceedings of the 2nd International Workshop on Software Development Lifecycle for Mobile (DeMobile 2014), November 2014, pp. 5-8.

[25] C. Theisen, B. Murphy, K. Herzig, and L. Williams, "Risk-Based Attack Surface Approximation: How Much Data is Enough?," Proceedings of the 39th International Conference on Software Engineering: Software Engineering in Practice Track (ICSE-SEIP '17), May 2017, pp. 273-282.

[26] T. Al-Salah, L. Hong, and S. Shetty, "Attack Surface Expansion Using Decoys to Protect Virtualized Infrastructure," Proceedings of the 2017 IEEE International Conference on Edge Computing (EDGE), June 2017, pp. 216-219.

[27] K. Sun and S. Jajodia, "Protecting Enterprise Networks through Attack Surface Expansion," Proceedings of the 2014 Workshop on Cyber Security Analytics, Intelligence and Automation (SafeConfig '14), November 2014, pp. 29-32.