Towards Reducing the Impact of Data 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, i.e., the people described by 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. This paper examines the behaviour of Business-to-Consumer (B2C) e-commerce companies, in terms of why they collect and store personal data. It then proposes an approach that reduces the impact of a data breach 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— reducing impact; data breach; private data loss; B2C e-commerce.

I. INTRODUCTION

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). Recently in the news is a report that MGM Resorts has suffered a data breach in which the personal details of over 10.6 million guests who stayed at its hotels have appeared on a hacking forum [1]. MGM Resorts is being sued for this breach [2]. In the first half of 2019, there were 3,800 publicly disclosed breaches worldwide, exposing 4.1 billion records. This represents an increase of 54% in the number of reported breaches compared to the first half of 2018 [3].

In response to attacks that result in data breaches, 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 is stored on the user’s device. Thus, a smaller quantity of private data remains on the company’s computer system, reducing the impact or 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 smartphone. 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 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 III presents the approach. Section IV contains examples of how the approach can fit with different types of e-commerce companies. Section V describes related work. Section VI gives conclusions and future work.

II. 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. Private Data

Private information consists of data about a person that can identify or be linked to that person and is owned by that person. Thus, private information 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]. In many countries, personal information 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.

B. Purposes for Collecting Personal Information

Companies engaged in B2C e-commerce, collect personal information for the following purposes (the first two purposes are self-evident):

- Transaction Requirements: 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: 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 [5]. 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 [6].
- 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 [7].
- 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 [5].
- Sharing or Selling: Personal information collected is shared or sold to other organizations for a profit [5].

C. 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., Marriott.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>
<thead>
<tr>
<th>Company type</th>
<th>Products</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sellers of goods and services (e.g., Amazon.com)</td>
<td>Physical items such as pots, clothing, and electronics; services such as selling your items for you</td>
</tr>
<tr>
<td>Hotels (e.g., Mariott.com)</td>
<td>Rooms</td>
</tr>
<tr>
<td>Travel Agencies (e.g., Expedia.ca)</td>
<td>Travel bookings</td>
</tr>
<tr>
<td>Financial services (e.g., CIBC.com)</td>
<td>Fee-based banking accounts</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data type</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Products</td>
<td>Company</td>
</tr>
<tr>
<td>CPD</td>
<td>Customer</td>
</tr>
<tr>
<td>Product selection</td>
<td>Customer</td>
</tr>
<tr>
<td>Amount paid</td>
<td>Company</td>
</tr>
<tr>
<td>Ancillary data</td>
<td>Company</td>
</tr>
</tbody>
</table>

Figure 1. Data collection for a customer order.

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 1. We wish to emphasize that there is no implied ordering of the data types in Figure 1, i.e., Figure 1 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.
III. Approach

This section details our approach for reducing the impact of 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 II-B to be carried out. The strategy consists of five parts, as follows:

1. Identification of data (Figure 1) 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 2 (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.

   | UCI | Product selection | Amount paid | Ancillary data |

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

   Figure 2. 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 3 to contact the customer, where “Contact information” consists of email address and telephone number. Figure 4 shows how the UCI links the three types of data records together.

   | UCI | Contact information |

   Figure 3. Data record for a customer’s contact information. Encrypted data types are shaded.

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 III-C.

5. Enabling security: Use authenticated symmetric encryption (e.g., AES-GCM [8]) to encrypt the UCI and CPD in Figure 2(a), as well as the Contact information in Figure 3 (encrypted data types are shaded). The UCI in Figure 3 is not encrypted. The UCI and remaining data types in Figure 2 (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.

   UCI  CPD
   Customer Device

   UCI  Contact information
   Company’s Computer System

   UCI  Product selection  Amount paid  Ancillary data

   Figure 4. 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 through 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 2(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 3, 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 2(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 2(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.

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 4). 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
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
encrypted 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., [9]) 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 III-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 2(b) and Figure 3.

• 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 II-B) for collecting personal 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 3) 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 2(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 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 [9].

IV. APPLICATION EXAMPLES

We instantiate the data types in Figure 1 for two 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>
<thead>
<tr>
<th>CPD</th>
<th>Product selection</th>
<th>Amount paid</th>
<th>Ancillary data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Camera</td>
<td>$159.00</td>
<td>Date ordered</td>
</tr>
<tr>
<td>Billing address</td>
<td>Hair clipper</td>
<td>$49.00</td>
<td>Date shipped</td>
</tr>
<tr>
<td>Default shipping address</td>
<td>Laser printer toner</td>
<td>$68.00</td>
<td>Date delivered</td>
</tr>
<tr>
<td>Alternate shipping address</td>
<td></td>
<td></td>
<td>Payment method</td>
</tr>
<tr>
<td>Email address</td>
<td></td>
<td></td>
<td>Product returned</td>
</tr>
<tr>
<td>Phone number</td>
<td></td>
<td></td>
<td>Reason for return</td>
</tr>
<tr>
<td>Credit card data</td>
<td></td>
<td></td>
<td>Refund status</td>
</tr>
</tbody>
</table>

Example 2: Hotel (e.g., Mariott.com). Table 4 shows the instantiation of the data types for this example.

<table>
<thead>
<tr>
<th>CPD</th>
<th>Product selection</th>
<th>Amount paid</th>
<th>Ancillary data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Room - double</td>
<td>$200 / night</td>
<td>Date of reservation</td>
</tr>
<tr>
<td>Billing address</td>
<td></td>
<td></td>
<td>Arrival date</td>
</tr>
<tr>
<td>Home address</td>
<td></td>
<td></td>
<td>Departure date</td>
</tr>
<tr>
<td>Email address</td>
<td></td>
<td></td>
<td>Payment method</td>
</tr>
<tr>
<td>Phone number</td>
<td></td>
<td></td>
<td>Airport shuttle y/n</td>
</tr>
<tr>
<td>Credit card data</td>
<td></td>
<td></td>
<td>Daily laundry y/n</td>
</tr>
<tr>
<td>Loyalty ID number</td>
<td></td>
<td></td>
<td>Daily cleaning y/n</td>
</tr>
<tr>
<td>Country of origin</td>
<td></td>
<td></td>
<td>Wake-up call y/n</td>
</tr>
<tr>
<td>Passport country</td>
<td></td>
<td></td>
<td>Stay extended y/n</td>
</tr>
<tr>
<td>Passport number</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Room preferences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Floor preference</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We could have included other examples here, but the above examples suffice for our purposes.

V. RELATED WORK

Work that is most closely related to this work are as follows: Aggarwal et al. [10] 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 reconstruct the sensitivity associations by breaching both servers. Ciriani et al. [11] present what they claim to be a solution that improves over Aggarwal et al. [10] 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. [10] and Ciriani et al. [11] 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 reduce the impact or 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. [12] 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 [13] 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 [14] 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 [15] 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. [16] 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. There are other references in this category which have not been included here due to lack of space.

VI. CONCLUSIONS AND FUTURE WORK

We have presented an approach, applicable to B2C e-commerce companies, that reduces the impact 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. The word “most” is key, 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. Some readers may consider the approach overly simple, but if a simple solution gets the job done, it should be preferred over a complex solution. As well, we do not claim that the approach as presented is complete, as there may be details that we have overlooked. We look forward to receiving feedback and correcting this 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 conference reviewers of this paper for their detailed and useful comments.

REFERENCES


