Data anonymization and privacy

SPISE Camp

Miti Mazmudar

Learning Outcomes

- Security planner exercise: Continued from "Securing yourself and others" session. Identify security tools that you can use in your daily life.

Learning Outcomes

- Security planner exercise: Continued from "Securing yourself and others" session. Identify security tools that you can use in your daily life.
- Canada Computers exercise: Continued from "FIPs and Fibs" session. Complete a privacy risk assessment questionnaire.

Learning Outcomes

- Security planner exercise: Continued from "Securing yourself and others" session. Identify security tools that you can use in your daily life.
- Canada Computers exercise: Continued from "FIPs and Fibs" session. Complete a privacy risk assessment questionnaire.
- Define quasi-identifiers.

Learning Outcomes

- Security planner exercise: Continued from "Securing yourself and others" session. Identify security tools that you can use in your daily life.
- Canada Computers exercise: Continued from "FIPs and Fibs" session. Complete a privacy risk assessment questionnaire.
- Define quasi-identifiers.
- Describe how quasi-identifiers can be used to link records across datasets.

Learning Outcomes

- Security planner exercise: Continued from "Securing yourself and others" session. Identify security tools that you can use in your daily life.
- Canada Computers exercise: Continued from "FIPs and Fibs" session. Complete a privacy risk assessment questionnaire.
- Define quasi-identifiers.
- Describe how quasi-identifiers can be used to link records across datasets.
- Define k-anonymity.

Learning Outcomes

- Security planner exercise: Continued from "Securing yourself and others" session. Identify security tools that you can use in your daily life.
- Canada Computers exercise: Continued from "FIPs and Fibs" session. Complete a privacy risk assessment questionnaire.
- Define quasi-identifiers.
- Describe how quasi-identifiers can be used to link records across datasets.
- Define k-anonymity.
- Determine whether a data release has the k-anonymity property.

Learning Outcomes

- Security planner exercise: Continued from "Securing yourself and others" session. Identify security tools that you can use in your daily life.
- Canada Computers exercise: Continued from "FIPs and Fibs" session. Complete a privacy risk assessment questionnaire.
- Define quasi-identifiers.
- Describe how quasi-identifiers can be used to link records across datasets.
- Define k-anonymity.
- Determine whether a data release has the k-anonymity property.
- Outline two attacks against k-anonymity: homogenity & background knowledge attacks.

Security Planner

Security Planner activity

- **1 min** In pairs, go to https://securityplanner.consumerreports.org/

Security Planner activity

- **1 min** In pairs, go to https://securityplanner.consumerreports.org/
- **2 mins** Follow through each step of the planner.

Security Planner activity

- **1 min** In pairs, go to https://securityplanner.consumerreports.org/
- **2 mins** Follow through each step of the planner.
- **5 mins** Read all suggestions.

Security Planner activity

- **1 min** In pairs, go to https://securityplanner.consumerreports.org/
- **2 mins** Follow through each step of the planner.
- **5 mins** Read all suggestions.
- **1 min** Pick one suggestion that you don't understand.

Security Planner activity

- **1 min** In pairs, go to https://securityplanner.consumerreports.org/
- **2 mins** Follow through each step of the planner.
- **5 mins** Read all suggestions.
- **1 min** Pick one suggestion that you don't understand.
- **1 min** Come up with one question (about that suggestion) and write it down.

Security Planner activity

- **1 min** In pairs, go to https://securityplanner.consumerreports.org/
- **2 mins** Follow through each step of the planner.
- **5 mins** Read all suggestions.
- **1 min** Pick one suggestion that you don't understand.
- **1 min** Come up with one question (about that suggestion) and write it down.
- **10th min** Each pair will ask their question to me at the end of the activity.

Privacy breach risk self-assessment activity

Example: Canada Computers breach

r/bapccanada • 5d ago
Extension-Fly1044

Canada Computers online card skimmer

If you have made a purchase recently on Canada Computers' online store, you should immediately freeze or cancel the card you used.

I found a card skimmer on Canada Computers' online checkout page. This malware steals any information you enter on the page and sends it to the attacker's website.

The malware is a Magecart-style script that listens to any input on the payment form fields, validates them, and steals them. It's obfuscated and loads from CodePen through a disguised Google Analytics script (something a real payment processor would never do). The malware captures credit card number, CVV, expiration date, first name, last name, billing address, billing city, billing province, billing postal code, phone number, email address and the Canada Computers account you're logged into.

I found this on January 18th when buying something on the website with DevTools open. I saw a suspicious WebSocket connection to `rozenfeld[.]xyz`. This domain isn't related to Canada Computers or any payment processor in any way. It looks similar to `rozenfeld[.]ca`, which I believe is a legitimate e-commerce related company. This could be an attempt from the attackers to seem legitimate.

Keep in mind I'm just a person who does web development as a hobby, I'm not a cybersecurity expert. I have opened two support tickets with them via email to try and tell them about this privately and they have closed both with no response. I'm assuming this is because they thought it was a scam or prank. I'm posting this publicly because they're closing my support requests and because the skimmer is still on the website, stealing data.

I have frozen my card that was stolen and have reported this to the Canadian Anti-Fraud Centre.

Example: Canada Computers breach

UPDATE (Jan 22, 4:54 PM EST):

The skimmer seems to have been removed from the live site. As of 4:54 PM EST, the checkout page no longer contains the malicious script or connections to `rozenfeld[.]xyz`. However, there is archived proof of this on Archive.org from December 31st 2025 that confirms the skimmer was on the checkout page.

Archive link: https://web.archive.org/web/20251231195438/https://www.canadacomputers.com/en/

Archive timestamp: Wed, 31 Dec 2025 19:54:38 GMT

This means the skimmer was active for at least 3 weeks.

Canada Computers has yet to acknowledge this breach or notify customers at all.

The latest snapshot I found on Archive.org that didn't have the skimmer was made on December 8 2025. If you bought anything on their online store between Dec 8 and Jan 22, your card info has been stolen and you should take the precautions I recommended at the top of the post (cancel/freeze). Even if you bought something before December 8 on the online store, I'd watch my bank statements very closely since their website has a history of data breaches and bad practices.

[2]

[2] https://www.reddit.com/r/bapccanada/comments/1qk4axy/canada_computers_online_card_skimmer/

Takeaways: Responsible Disclosure

What does this breach report get right?

- Described the attack vector.

Takeaways: Responsible Disclosure

What does this breach report get right?

- Described the attack vector.
- Mentioned when the attack was discovered.

Takeaways: Responsible Disclosure

What does this breach report get right?

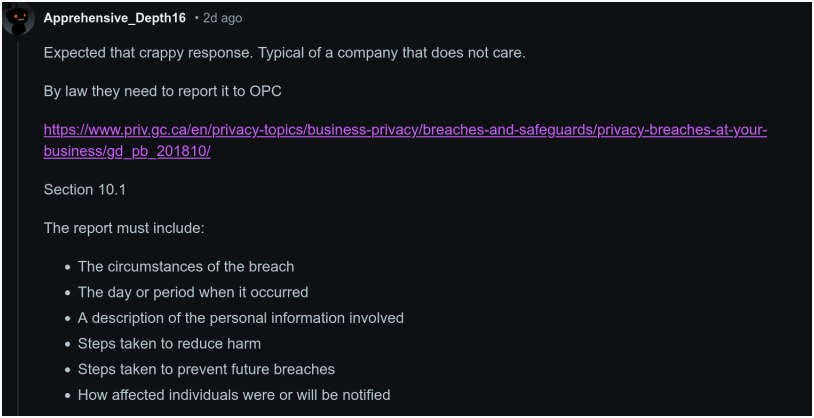
- Described the attack vector.
- Mentioned when the attack was discovered.
- Estimated when the attack might have occurred.

Takeaways: Responsible Disclosure

What does this breach report get right?

- Described the attack vector.
- Mentioned when the attack was discovered.
- Estimated when the attack might have occurred.
- Informed the business about the attack.

Example: Reactions to Canada Computers response



Apprehensive_Depth16 • 2d ago

Expected that crappy response. Typical of a company that does not care.

By law they need to report it to OPC

https://www.priv.gc.ca/en/privacy-topics/business-privacy/breaches-and-safeguards/privacy-breaches-at-your-business/gd_pb_201810/

Section 10.1

The report must include:

- The circumstances of the breach
- The day or period when it occurred
- A description of the personal information involved
- Steps taken to reduce harm
- Steps taken to prevent future breaches
- How affected individuals were or will be notified

Worksheet task

- An organization is only legally required to report breaches to the OPC if there is a "real risk of significant harm (RROSH)".

Worksheet task

- An organization is only legally required to report breaches to the OPC if there is a "real risk of significant harm (RROSH)".
- The risk of significant harm depends on the sensitivity of the information breached and the likelihood of its misuse.

Worksheet task

- An organization is only legally required to report breaches to the OPC if there is a "real risk of significant harm (RROSH)".
- The risk of significant harm depends on the sensitivity of the information breached and the likelihood of its misuse.
- The OPC has released a privacy breach risk self-assessment guide here:

Worksheet task

- An organization is only legally required to report breaches to the OPC if there is a "real risk of significant harm (RROSH)".
- The risk of significant harm depends on the sensitivity of the information breached and the likelihood of its misuse.
- The OPC has released a privacy breach risk self-assessment guide here:
- https://services.priv.gc.ca/rrpg-rrosh.

Worksheet task

- An organization is only legally required to report breaches to the OPC if there is a "real risk of significant harm (RROSH)".
- The risk of significant harm depends on the sensitivity of the information breached and the likelihood of its misuse.
- The OPC has released a privacy breach risk self-assessment guide here:
- https://services.priv.gc.ca/rrpg-rrosh.
- This tool takes as input the sensitivity of the information breached, the incident response steps taken to handle the breach, and estimates the likelihood of its misuse.

Worksheet task

- An organization is only legally required to report breaches to the OPC if there is a "real risk of significant harm (RROSH)".
- The risk of significant harm depends on the sensitivity of the information breached and the likelihood of its misuse.
- The OPC has released a privacy breach risk self-assessment guide here:
- https://services.priv.gc.ca/rrpg-rrosh.
- This tool takes as input the sensitivity of the information breached, the incident response steps taken to handle the breach, and estimates the likelihood of its misuse.
- The tool outputs the potential risks that the breach presents to an impacted person & whether Canada Computers should report the breach to OPC.

Data releases

- Publishing data by a one-time upload.

Data releases

- Publishing data by a one-time upload.
- Analysts can do computations on the published data.

Data releases

- Publishing data by a one-time upload.
- Analysts can do computations on the published data.
- Data aggregation and linking: Data released from different sources can be combined together.

Data aggregation

- Use publicly available data, e.g. census data, regional records.

Data aggregation

- Use publicly available data, e.g. census data, regional records.
- Buy data records from a data broker at a low price.

Data aggregation

- Use publicly available data, e.g. census data, regional records.
- Buy data records from a data broker at a low price.
- Governments can also do steps a, b and share their dossiers with each other.

Data aggregation

- Use publicly available data, e.g. census data, regional records.
- Buy data records from a data broker at a low price.
- Governments can also do steps a, b and share their dossiers with each other.
- Large companies may aggregate information about their customers.

Anonymity fails

Anonymity fail: Massachusetts Insurance Health Records.

- Massachusetts Group Insurance Commission released “anonymized” health records

Anonymity fail: Massachusetts Insurance Health Records.

- Massachusetts Group Insurance Commission released "anonymized" health records
- Record format: [ZIP codes, gender, date of birth, health information] (No names).

Anonymity fail: Massachusetts Insurance Health Records.

- Massachusetts Group Insurance Commission released "anonymized" health records
- Record format: [ZIP codes, gender, date of birth, health information] (No names).
- Massachusetts voter registration lists contain these three identifiers plus individuals' names.

Anonymity fail: Massachusetts Insurance Health Records.

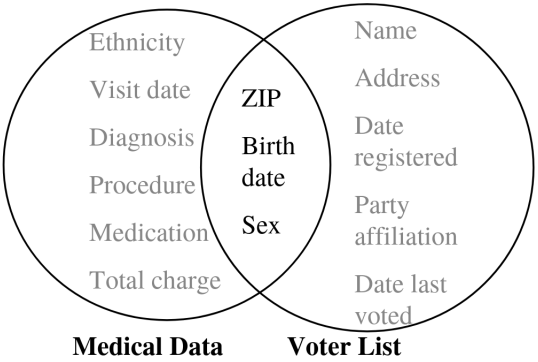


Figure 1 Linking to re-identify data

Anonymity fail: Massachusetts Insurance Health Records.

- Massachusetts Group Insurance Commission released "anonymized" health records
- Record format: [ZIP codes, gender, date of birth, health information] (No names).
- Massachusetts voter registration lists contain these three identifiers plus individuals' names.

Anonymity fail: Massachusetts Insurance Health Records.

- Massachusetts Group Insurance Commission released "anonymized" health records
- Record format: [ZIP codes, gender, date of birth, health information] (No names).
- Massachusetts voter registration lists contain these three identifiers plus individuals' names.
- We can aggregate voter registration lists and these health records, and re-identify patients by linking, using these three identifiers.

Anonymity fail: Massachusetts Insurance Health Records.

- Massachusetts Group Insurance Commission released "anonymized" health records
- Record format: [ZIP codes, gender, date of birth, health information] (No names).
- Massachusetts voter registration lists contain these three identifiers plus individuals' names.
- We can aggregate voter registration lists and these health records, and re-identify patients by linking, using these three identifiers.
- 87% of U.S. population can be uniquely identified using these identifiers.

Lessons from anonymity fails in private data releases

1. Datasets included data that was useful for research (primary data), as well as some identifiers ("quasi-identifiers").

Lessons from anonymity fails in private data releases

1. Datasets included data that was useful for research (primary data), as well as some identifiers ("quasi-identifiers").
2. *"Quasi-identifiers"* can be used to link data across different datasets (Massachusetts case).

Prevent anonymity fails & privacy/utility tradeoff

- Quasi-identifiers:

Prevent anonymity fails & privacy/utility tradeoff

- Quasi-identifiers:
  - Reduce granularity to *deter* linking: e.g. year instead of DOB, only first couple digits of zip code. Increases anonymity set.

Prevent anonymity fails & privacy/utility tradeoff

- Quasi-identifiers:
  - Reduce granularity to *deter* linking: e.g. year instead of DOB, only first couple digits of zip code. Increases anonymity set.
  - Remove attribute(s) to *prevent* linking altogether: e.g. no random number in AOL dataset or no medallion/license number in NYC taxi dataset. Will reduce utility of the dataset.

k-anonymity

1. Pre-processing quasi-identifiers to deter data linking.
    - Remove gender altogether.
    - Reduce granularity of ZIP code and date of birth.

k-anonymity

1. Pre-processing quasi-identifiers to deter data linking.
   - Remove gender altogether.
   - Reduce granularity of ZIP code and date of birth.
2. k-anonymity: Ensure that for each published record, we publish at least $k - 1$ other records with the same quasi-identifier (where $k \geq 2$).

k-anonymity

1. Pre-processing quasi-identifiers to deter data linking.
   - Remove gender altogether.
   - Reduce granularity of ZIP code and date of birth.
2. k-anonymity: Ensure that for each published record, we publish at least $k - 1$ other records with the same quasi-identifier (where $k \geq 2$).
3. This constraint ensures that any person who satisfies a given quasi-identifier could correspond to any one of the k records with the same quasi-identifier.

k-anonymity example

A 3-anonymized table

ZIP	DOB	Party affiliation
N1C**	196*-*-*	Green Party
N1C**	196*-*-*	Green Party
N1C**	196*-*-*	Green Party
G0A*	197*-*-*	Liberal Party
G0A*	197*-*-*	Green Party
G0A*	197*-*-*	Conservative Party
N2J*	199*-*-*	Conservative Party
N2J*	199*-*-*	New Democratic Party
N2J*	199*-*-*	Liberal Party

k-anonymity — homogeneity attack

- Homogeneity attack: For a given quasi-identifier value, all other data values are identical.

k-anonymity — homogeneity attack

ZIP	DOB	Party affiliation
N1C**	196*-*-*	Green Party
N1C**	196*-*-*	Green Party
N1C**	196*-*-*	Green Party
G0A*	197*-*-*	Liberal Party
G0A*	197*-*-*	Green Party
G0A*	197*-*-*	Conservative Party
N2J*	199*-*-*	Conservative Party
N2J*	199*-*-*	New Democratic Party
N2J*	199*-*-*	Liberal Party

- Homogeneity attack: For a given quasi-identifier value, all other data values are identical.
  - If you know Alice (N1C** , 196*-*-*) is in the table, then Alice has voted for the Green Party.

k-anonymity — background knowledge attack

- Background knowledge attack: For a given quasi-identifier value, other data values are distinct, but you know background info to rule out many of them.

k-anonymity — background knowledge attack

ZIP	DOB	Party affiliation
N1C**	196*-*-*	Green Party
N1C**	196*-*-*	Green Party
N1C**	196*-*-*	Green Party
G0A*	197*-*-*	Liberal Party
G0A*	197*-*-*	Green Party
G0A*	197*-*-*	Conservative Party
N2J*	199*-*-*	Conservative Party
N2J*	199*-*-*	New Democratic Party
N2J*	199*-*-*	Liberal Party

- Background knowledge attack: For a given quasi-identifier value, other data values are distinct, but you know background info to rule out many of them.
  - If you know Bob (G0A*, 197*-*-*) is in the table, and that he definitely doesn't like the LP or GP, then Bob has voted for the CP.