Formal Privacy Analyses for Open Banking

Luigi D. C. Soares(✉)[1,2][0000−0002−9579−8427],
Mário S. Alvim[1][0000−0002−4196−7467], Di Bu[2][0000−0002−8714−6808], Natasha Fernandes[2][0000−0002−9212−7839], and Yin Liao[2][0000−0002−5343−0579]

[1] UFMG, Belo Horizonte, Brazil
contact@luigidcsoares.com, msalvim@dcc.ufmg.br
[2] Macquarie University, Sydney, Australia
{di.bu,natasha.fernandes,yin.liao}@mq.edu.au

Abstract. The term "Open Banking" describes a series of global initiatives to allow the sharing of customer data between financial companies to facilitate competition within their sector. In this paper, we formalise in the rigorous framework of quantitative information flow (QIF) relevant privacy risks in a concrete Open Banking scenario, namely: (i) transaction-history recovery and (ii) collateral attribute-inferences using external correlations. We provide extensive analyses of these risks in real-world data from Open Banking, supplied by a fintech in Australia. We show that the Open Banking system studied presents considerable privacy risks with respect to transactions, both in the presence and in the absence of demographic data. Finally, we exemplify potential real-world collateral attribute-inference attacks, in which we show how an attacker might leverage scientific correlations to infer individuals' level of neuroticism and self-control from their transaction history. We hope that this work may: (i) help financial customers in Australia make better-informed decisions about what kind of information, and how much of it, to share via Open Banking; (ii) raise awareness about the potential privacy risks of Open Banking in other countries; and (iii) foster the development of privacy regulation in digital finance and the open data economy.

Keywords: Open Banking, Privacy-Risk Analysis, Quantitative Information Flow

1 Introduction

The term "Open Banking" describes a series of global initiatives to allow the sharing of customer data between financial companies, such as banks or fintechs (i.e., financial technology companies), in order to facilitate competition within their sector. Under Open Banking, financial institutions provide access to customer banking information via an API.[3] Customers then provide consent for 3rd parties to access all or some of their banking information. The advantage to the

[3] See https://standards.openbanking.org.uk/api-specifications/ or https://consumerdatastandardsaustralia.github.io/standards/#banking-apis

customer is that it (ideally) gives them control over the sharing of their financial data, facilitating access to new financial products or services. The advantage to the financial sector is that it encourages new business models and provides opportunities for smaller fintechs to evaluate customers without requiring negotiation with other banks or relying on customer-provided information.

In this work, we provide a thorough formal analysis of privacy risks associated with sharing de-identified data via Open Banking. This involves (i) *transaction data*, which is de-identified data released via Open Banking and includes details of individual transactions such as amount spent, vendor, location, and date, and (ii) in some cases, *demographic data*, which is de-identified personal data including details such as age, gender, zip code, and job. More precisely, we assess the sensitive information that can be inferred from the collected datasets by any entity with access to them (e.g., the financial institution that collected the data in the first place or any other 3rd party with whom the data is shared).

We formalise our attack models in the rigorous mathematical framework of *quantitative information flow (QIF)* [3]. A crucial advantage of the use of QIF is that it allows for great flexibility in privacy analyses, since variations of practical scenarios of interest can be seamlessly captured in the framework. We notice, however, that QIF is not a privacy guarantee that a system may satisfy (such as differential privacy or k-anonymity, for some proper choice of parameters), but it is rather a framework for quantifying the privacy provided by the system in terms of its resistance to inference attacks. This is a crucial advantage of QIF: its guarantees are presented in terms of threats, which are easier to interpret for data managers and consumers. In terms of scalability, QIF has been put to test in a thorough formal privacy analysis of the Official Educational Censuses in Brazil, covering over a decade of microdata for more than 65 million individuals [4], which led to concrete changes in data-release policies by the Brazilian government.[4] Moreover, QIF has been successfully applied to a wide variety of scenarios that can be modelled as some form of information flow, including privacy and security [17,13,5,1,10], machine learning [23,27], and fairness [6].

Contributions. We rigorously formalise and provide extensive analyses of privacy risks in real-world data from Open Banking, supplied by a fintech in Australia:

- **Transaction-history recovery:** Success in recovering all transactions provided by a customer is clearly a damaging scenario, since it may lead the adversary to infer sensitive attributes, such as customers' buying habits, stores visited, or income. The availability of demographic data can facilitate this process. Nevertheless, as we shall see, this attack remains possible even without any sort of personal information at the adversary's disposal.
- **Collateral attribute-inference:** A crucial novelty of our work is the assessment of a more subtle privacy risk: *collateral attribute-inference* (a.k.a., *Dalenius attacks* in QIF [2,3]). In this case, the goal is not to infer individuals' attributes that are explicitly present in the dataset, but to exploit *known*

[4] In Portuguese: https://www.gov.br/inep/pt-br/assuntos/noticias/institucional/nota-de-esclarecimento-divulgacao-dos-microdados.

correlations between *attributes in the dataset* with *sensitive information not explicitly present in the dataset.* Such correlations have been used in, e.g., the Cambridge Analytica scandal to sway voters during elections, with dire consequences.[5] We exemplify a possible real-world collateral-inference attack, where we demonstrate how to leverage external, known correlations to infer individuals' personality traits from their financial activities.

Moreover, we set up a rigorous framework that can be used to model other Open Banking scenarios in the future, and analyze the corresponding privacy threats.

Related Work. Early privacy concerns about Open Banking focused on phishing attacks designed to lure consumers into handing over consent to their data [14]. The potential privacy risks of demographic data have been well-known since at least Sweeney's seminal work with the US Census [25]. Detailed transactions have also been shown to be problematic: four data points in credit card transaction data are enough to re-identify 90% individuals out of 1.1 million people [18]. However, in contrast to our work, these privacy analyses are deterministic.

More recently, studies have found correlations between people's spending behaviours and personal traits, which have been exploited to predict such traits. Examples of correlations range from people's lifestyles and preferences [12] to (possibly) harmful correlations such as individuals' psychological conditions [15,26] — knowing someone's personality traits might, for instance, bear influence on hiring decisions, either for good or bad [7,16,19,9]. Nevertheless, to the best of our knowledge, no previous work has formally quantitatively evaluated collateral-inference types of information leaks from financial data due to known correlations. Our analyses are in line with those of Alvim et al. [4] on educational data, but we incorporate collateral-inferences and consider Open Banking data.

Ethical Disclosure. The demographic and transaction datasets used in this work have been provided to us by a financial institution for the purposes of this research and are not publicly available. Consent was provided by customers for use of this data, and our use of the data has been ethically reviewed by our organisation. No real information about individuals has been revealed in this paper, as our examples use dummy names and values in place of true data. Moreover, no real re-identification or inference was performed in our analyses, since we, as researchers, lack the auxiliary knowledge that we assume the adversary has about customers to complete the attacks.

Plan of the Paper. The remainder of this paper is organised as follows. Section 2 provides necessary background on quantitative information flow (QIF). In Section 3, we formalise our attack models in QIF. In Section 4, we analyse the first kind of privacy risk, transaction-history recovery, both in the presence and absence of demographic data. In Section 5, we consider the more refined scenario in which the adversary explores a correlation to infer sensitive information not immediately available in the database. Finally, Section 6 concludes this work.

[5] Information obtained from https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie.

2 Preliminaries: QIF

In this section, we review fundamental concepts from quantitative information flow (QIF) that we use to formalise privacy risks. Our model assumes a Bayesian adversary who makes an optimal guess by combining their prior knowledge of the data-sharing mechanism with some auxiliary information.

Secrets, Prior Knowledge, and Prior Vulnerability. A *secret* X models the information sought by the adversary; the set of possible secrets is denoted by $\mathcal{X}$. The adversary's prior knowledge about the secrets can be modelled as a distribution $\pi \in \mathbb{D}\mathcal{X}$, where $\mathbb{D}\mathcal{X}$ denotes the set of all probability distributions over the values in $\mathcal{X}$. We write π_x for the probability assigned by π to secret value x.

In this work, we adopt as a privacy measure the *Bayes vulnerability*, which is closely related to Rényi min-entropy and Bayes risk [21,24,8]. It represents the adversary's probability of guessing the secret value correctly in one try. (See Appendix A for a deeper discussion on privacy measures in QIF.) The *prior Bayes vulnerability* can be computed as

$$V(\pi) = \max_{x \in \mathcal{X}} \pi_x. \tag{1}$$

Channels, Posterior Knowledge, and Posterior Vulnerability. The secret value is fed into and processed by a system (modelled as a channel) that produces some observable behaviour that the adversary can use to launch an attack. Formally, an *information-theoretic channel* $C : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ takes an input (secret) $x \in \mathcal{X}$ and produces an output (observation) $y \in \mathcal{Y}$ according to some distribution in $\mathbb{D}\mathcal{Y}$. When $\mathcal{X}$ and $\mathcal{Y}$ are discrete, we write C as a matrix whose element $C_{x,y}$ is the probability of producing output $y \in \mathcal{Y}$ when the input is $x \in \mathcal{X}$. Rows in C are distributions over $\mathcal{Y}$; if C is deterministic, every entry in C is 0 or 1, and each row contains exactly one value 1. By combining the prior $\pi : \mathbb{D}\mathcal{X}$ on secrets with knowledge of the channel $C : \mathcal{X} \to \mathbb{D}Y$ representing how the system works, the adversary can compute a joint distribution $\pi \triangleright C : \mathbb{D}(\mathcal{X} \times \mathcal{Y})$:

$$(\pi \triangleright C)_{x,y} = \pi_x C_{x,y}, \text{ for every } x \in \mathcal{X} \text{ and } y \in \mathcal{Y}. \tag{2}$$

Now, from the joint $\pi \triangleright C$, the adversary can perform Bayesian reasoning to update their knowledge about the secret from the prior π to a revised knowledge consisting of: (i) a marginal distribution on possible values of y obtained as

$$p(y) \stackrel{\text{def}}{=} \sum_{x \in \mathcal{X}} (\pi \triangleright C)_{x,y}, \tag{3}$$

and (ii) for each observation y, a posterior distribution δ^y on set $\mathcal{X}$ obtained as the conditional probability $p(x \mid y)$, i.e.,

$$\delta^y_x \stackrel{\text{def}}{=} \frac{(\pi \triangleright C)_{x,y}}{p(y)}, \text{ for each } x \in \mathcal{X}. \tag{4}$$

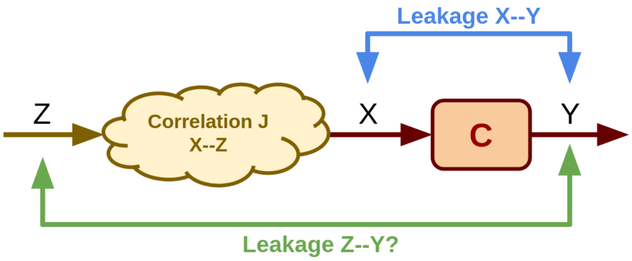


Fig. 1: Collateral-inference leakage: a channel C maps X to Y, but the secret of interest is another value Z correlated with X. Observations of the output Y of C can leak information about Z, given the correlation between X and Z.

The *adversary's posterior knowledge* is denoted by $[\pi \triangleright C]$, and consists of the distribution on the output values $y \in \mathcal{Y}$ together with each corresponding posterior distribution δ^y on secret values (that is, each δ^y_x is the updated probability of secret value x given that output value y was observed). We consider two possible definitions of *posterior Bayes vulnerability* in this paper:

- *Dynamic* posterior Bayes vulnerability, which corresponds to the adversary's maximum probability of guessing the secret value correctly for a fixed observation y. It is defined as

$$V^y[\pi \triangleright C] \overset{\text{def}}{=} V(\delta^y). \tag{5}$$

- *(Expected/static)* posterior Bayes vulnerability, which corresponds to the expected maximum probability of the adversary guessing the secret correctly, weighted over all possible values y that the observation can take. It is

$$V[\pi \triangleright C] \overset{\text{def}}{=} \sum_{y \in \mathcal{Y}} p(y) V^y[\pi \triangleright C] = \sum_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}} (\pi \triangleright C)_{x,y}. \tag{6}$$

Information Leakage. Finally, we can compute *channel leakage* as the ratio between posterior and prior vulnerabilities. This corresponds to the multiplicative factor by which the observation of the system increases the adversary's maximum probability of guessing the secret value correctly. In the context of this paper, this captures to how much the customers' privacy risk has increased.

Collateral-Inference Leakage (a.k.a., Dalenius-Scenarios Leakage). Up to this point, we have computed the leakage of secret X caused by a channel C. But now assume that there is another secret Z from a set $\mathcal{Z}$ that: (i) apparently has nothing to do with C, but (ii) is correlated with X via some joint distribution $\mathit{\Pi} : \mathbb{D}(\mathcal{Z} \times \mathcal{X})$ known to the adversary. In this case, we can quantify how much C (surprisingly) teaches the adversary about Z, as in Figure 1.

Notice that the joint $\mathit{\Pi} : \mathbb{D}(\mathcal{Z} \times \mathcal{X})$ must induce the marginal distribution π on $\mathcal{X}$ that is the prior to channel C. $\mathit{\Pi}$ also induces a prior $\rho : \mathbb{D}\mathcal{Z}$ representing

Table 1: Relevant fields from each dataset in our attacks. Fields marked as "QID" were considered as the adversary's auxiliary knowledge, whereas fields tagged as "Secret" are the sensitive information that adversaries seek to learn.

	Field	**Description**	**Role**
Demographic	Age	Customer's exact age	QID
	Employment	Part/full-time, unemployed, student, etc	QID
	Gender	Customer's gender (male or female)	QID
	Zip Code	Location where the customer lives	QID
	User ID	Customer's identification number	Secret
Transaction	Amount	Total amount paid or received	QID
	Category	Transaction's category (e.g., groceries)	QID
	Date	Year, month and day that the transaction took place	QID
	Payee	To whom the transaction was paid (e.g., Aldi)	QID
	Description	Detailed description of the transaction	Secret
	User ID	Customer's identification number	Secret

the adversary's knowledge about Z. Moreover, we can express the joint Π as the result of combining the prior $\rho : \mathbb{D}\mathcal{Z}$ with a channel $B : \mathcal{Z} \rightarrow \mathbb{D}\mathcal{X}$. More precisely: (i) $\rho_z B_{z,x} = \Pi_{z,x}$ for all $z \in \mathcal{Z}$ and $x \in \mathcal{X}$, and (ii) each row $B_{z,-}$ of B is found by normalising row $\Pi_{z,-}$. Now, it can be shown [2,3] that the conditional probability $p(y \mid z)$ for each $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$ from Figure 1 is equal to $(BC)_{z,y}$, where BC is just the ordinary matrix multiplication of matrices B and C, i.e., $(BC)_{z,y} = p(y \mid z)$. This leads to the following definitions:

- The *prior* collateral Bayes vulnerability of Z is

$$\mathcal{D}V(\Pi) \overset{\text{def}}{=} V(\rho). \tag{7}$$

- The dynamic and expected collateral Bayes vulnerabilities are, respectively,

$$\mathcal{D}V^y[\Pi \triangleright C] \overset{\text{def}}{=} V^y[\rho \triangleright BC] \tag{8}$$

$$\mathcal{D}V[\Pi \triangleright C] \overset{\text{def}}{=} V[\rho \triangleright BC] \tag{9}$$

3 Formalisation of Attacks Under Open Banking

Table 1 describes the relevant information from the demographic and the transaction datasets used in our attacks. Tables 2a and 2b provide examples of such datasets. Each individual has a unique, artificially created ID value, which is attached to all of their transactions. Figure 2 provides an overview of our model for Open Banking attacks, whose components we describe in detail below.

In our model, the secret X could be, for instance, an individual's employment status in the demographic dataset or the maximum amount spent by a

Table 2: Example of demographic and transaction data for five individuals.

(a) Demographic dataset.

User ID	Age	Gender
1	46	M
2	21	M
3	46	Female
4	23	Female
5	23	Female

(b) Transaction dataset.

User ID	Payee	Description
1	Red Rooster	Chicken Burger
1	Clinic	Fertility Treatment
2	Aldi	Groceries
3	Uber	23 minutes trip
3	Lakeside Hotel	One night
3	Clinic	Skin-Cancer Treat.
4	Uber	13 minutes trip
5	Uber	25 minutes trip

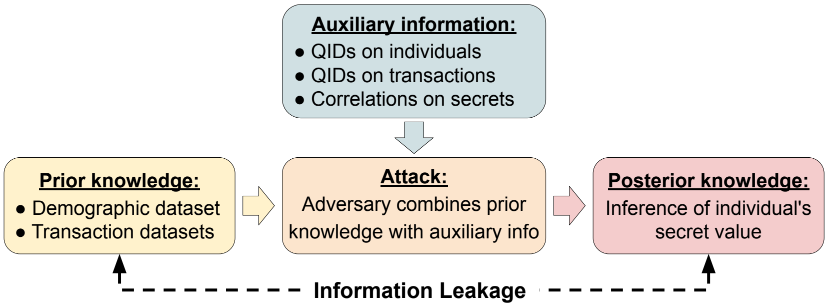


Fig. 2: Overview of attack model.

given individual in the transaction dataset. It could even be information not explicitly present in the datasets, such as a person's personality trait (which may be correlated to buying habits), in the case of collateral-inference attacks.

We consider that the adversary has a particular target and has access to the database D of interest (be it the demographic dataset, the transaction dataset, or a combination of both). This gives them a prior π over transaction histories, and their goal is to identify which transaction history belongs to their target — thus recovering all of their target's transactions — or infer some sensitive information correlated to their target's financial activities (i.e., a collateral inference).

In all of our attacks, a crucial element is the use of *quasi-identifiers (QIDs)*, which are attributes that, although not unique in themselves, may be combined to (almost) uniquely identify a record in a dataset [11,22,25,20]. For instance, Sweeney has famously shown that 87% of the population in the USA Census of 1 990 could be uniquely re-identified using only three QIDs: date of birth, gender, and zip code [25]. In this work, we exemplify how QIDs can be used in attacks to Open Banking systems in which we are interested.

Finally, the system C is composed of the demographic and/or transaction datasets, associating secret values X with observable QID values Y. For instance, in Example 1 (Section 4.1), the observation Y corresponds to the combination of two QIDs: a customer's age and gender. With access to the database D, the adversary can compute the channel $C : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ mapping secrets to QIDs as

$$C_{x,y} = \frac{\#\textit{rows with secret } x \textit{ and QIDs } y \text{ in database } D}{\#\textit{rows with secret } x \text{ in database } D}. \tag{10}$$

The channel C encodes the probability of the adversary learning a particular information $y \in \mathcal{Y}$ under the assumption that the secret is $x \in \mathcal{X}$, i.e., $p(y \mid x)$. Then, once an observation $y \in \mathcal{Y}$ is made from C — or, in other words, upon learning some QIDs y — the adversary can update their knowledge about the secret X via Bayesian reasoning to compute a posterior distribution, as per (4).

4 Quantification of Transaction-History Recovery Risks

In this section, we explore the first kind of privacy risk described in Section 1. We start by analysing in Section 4.1 the scenario in which the adversary has access to both the demographic and the transaction datasets. Then, in Section 4.2 we assess the privacy risks that remain even when the access to demographic data is removed. Each risk is presented with a simple and concrete example of attack, which is then modelled using the formalism from Section 3. We then provide experimental analyses of these attacks in real Open Banking data.

We consider as adversary any entity with access to the collected data, be it the original institution or any 3rd party with which the data is shared. We personify such an adversary as Charlize, a data analyst working for a fintech `F`, whose job is to assess customers' transactions to help them find suitable investment plans. We also assume that Charlize knows that her target is a client of fintech `F`.

4.1 Transaction-History Recovery via Demographic QIDs

Consider the case where the fintech `F` collects both demographic and financial data from customers, and let Alex be a client of `F`. In this scenario, Charlize has access to the demographic dataset from Table 2a as well as to the transaction dataset from Table 2b, and her goal is to recover Alex's financial data in the Open Banking system. Given that no two customers have the exact same transaction history in Table 2b, Charlize's objective reduces to determining Alex's ID.

Example 1. Assume that Charlize learns some (perhaps seemingly innocuous) QIDs about Alex: that she is a 46-year-old woman. Using this knowledge, Charlize can query the demographic dataset of Table 2a and discover that there is only one person with such QIDs in the dataset, thus learning that Alex's ID in the Open Banking system must be 3. Then, Charlize can retrieve Alex's financial history from the transaction dataset and learn all of Alex's transactions details, which includes a transaction to a skin-cancer clinic!

Instantiating the Example Using QIF. In the attack above, the secret X is Alex's transaction history, which, for this example, can be seen as Alex's user ID. Therefore, the secret X takes values in the set $\mathcal{X} = \{1, 2, 3, 4, 5\}$ of five possible IDs. Although Charlize knows that Alex is in the dataset, before the attack she does not have any reason to believe that any ID is more likely than any other to be Alex's. Hence, Charlize's prior knowledge π on Alex's ID is a uniform distribution $\pi = (1/5, 1/5, 1/5, 1/5, 1/5)$ over $\mathcal{X}$. A rational strategy allows her to guess any secret as the correct one, so the corresponding prior Bayes vulnerability is $V(\pi) = 1/5$, meaning that her probability of correctly re-identifying Alex in the Open Banking system assuming no auxiliary information is 20%.

But, since Charlize has access to the demographic dataset, she can employ (10) to build the channel $C^{\text{Re-id}}$ below representing the mapping from user IDs to QIDs in the Open Banking system. Then, she can combine her uniform prior π on IDs with channel $C^{\text{Re-id}}$ to obtain a joint $\pi \triangleright C^{\text{Re-id}}$, as per (2):

$$\begin{array}{c|c} & \pi \\ \hline 1 & \frac{1}{5} \\ 2 & \frac{1}{5} \\ 3 & \frac{1}{5} \\ 4 & \frac{1}{5} \\ 5 & \frac{1}{5} \end{array} \;\triangleright\; \begin{array}{c|cccc} C^{\text{Re-id}} & (46,\text{M}) & (21,\text{M}) & (46,\text{F}) & (23,\text{F}) \\ \hline 1 & 1 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 \\ 3 & 0 & 0 & 1 & 0 \\ 4 & 0 & 0 & 0 & 1 \\ 5 & 0 & 0 & 0 & 1 \end{array} \;=\; \begin{array}{c|cccc} \pi \triangleright C^{\text{Re-id}} & (46,\text{M}) & (21,\text{M}) & (46,\text{F}) & (23,\text{F}) \\ \hline 1 & \frac{1}{5} & 0 & 0 & 0 \\ 2 & 0 & \frac{1}{5} & 0 & 0 \\ 3 & 0 & 0 & \frac{1}{5} & 0 \\ 4 & 0 & 0 & 0 & \frac{1}{5} \\ 5 & 0 & 0 & 0 & \frac{1}{5} \end{array}$$

From $\pi \triangleright C^{\text{Re-id}}$, Charlize can perform Bayesian reasoning to update her prior knowledge about Alex's ID to some posterior, revised knowledge $[\pi \triangleright C^{\text{Re-id}}]$:

$$\begin{array}{c|cccc} [\pi \triangleright C^{\text{Re-id}}] & p(y_1) = \frac{1}{5} & p(y_2) = \frac{1}{5} & p(y_3) = \frac{1}{5} & p(y_4) = \frac{2}{5} \\ & (46,\text{M}) & (23,\text{M}) & (46,\text{F}) & (23,\text{F}) \\ \hline 1 & 1 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 \\ 3 & 0 & 0 & 1 & 0 \\ 4 & 0 & 0 & 0 & \frac{1}{2} \\ 5 & 0 & 0 & 0 & \frac{1}{2} \end{array}$$

The first row of $[\pi \triangleright C^{\text{Re-id}}]$ represents Charlize's updated knowledge about the probability of QIDs. Each column under a combination of values for QIDs represents the conditional probability of the corresponding ID being the right one, given these QIDs values. For instance, the column under QIDs $y_4 = (23, \text{F})$ indicates that, given these QIDs, there is a 50% probability that the individual in question has ID $x = 4$ and a 50% probability that the individual has ID $x = 5$.

Now, the attack itself starts when Charlize learns some auxiliary information about Alex: her age (46) and gender (Female). Using these two QIDs, Charlize can filter the demographic dataset and find that there is a unique record matching the criteria: that of user ID 3. Formally, Charlize's posterior knowledge (after learning the QIDs) is updated from a uniform prior on all IDs to a posterior distribution assigning all probability to the ID $x = 3$, as indicated in $[\pi \triangleright C^{\text{Re-id}}]$ above. Now, Charlize's rational strategy is to guess that Alex's ID is

3, and she would be right with probability 1. Hence, the *dynamic* posterior Bayes vulnerability with respect to the observation $y_3 = (46, \mathrm{F})$ is $V^{y_3}[\pi \triangleright C^{\text{Re-id}}] = 1$, as per (5), which is five times higher than the prior Bayes vulnerability.

Notice that, although Alex's QIDs were unique enough to single her out with certainty in the demographic dataset, this may not always be the case. As the column under QIDs $y_4 = (23, \mathrm{F})$ in $[\pi \triangleright C^{\text{Re-id}}]$ indicates, if Charlize's target is a 23-year-old female, the re-identification can happen with only 50% accuracy. Therefore, it may be relevant to assess the expected leakage *over all possible individuals that can be chosen as targets*. In the above scenario, the *expected* posterior Bayes vulnerability, according to (6), is $V[\pi \triangleright C] = {}^1\!/\!_5 + {}^1\!/\!_5 + {}^1\!/\!_5 + {}^1\!/\!_5 = {}^4\!/\!_5$. This, in turn, means that the expected leakage is $({}^4\!/\!_5)/({}^1\!/\!_5) = 4$, indicating that Charlize's expected chance of correctly re-identifying Alex upon learning her age and gender is four times higher than Charlize's initial chance of success.

Results. To evaluate the transaction-history recovery risks when demographic data is available, we considered a database with 17 206 customers, who together made 10 223 473 transactions. Table 3 summarises the privacy risks for the four combinations of demographic QIDs that yielded the highest Bayes vulnerabilities, along with the highest outcome when zip code is unknown. Since the demographic dataset contains 17 206 users, the prior Bayes vulnerability is $5.8 \cdot 10^{-5}$. Assuming that the adversary knows their targets' age, gender, zip code, and employment status, the adversary's expected probability of recovering their target's financial activities is 90%. In the absence of the target's employment status, the adversary's expected chance of success of re-identification decreases to 83%. Notice that lack of knowledge about the target's zip code brings the adversary's posterior success rate down to only 3%.

4.2 Transaction-history Recovery via Transaction QIDs

The risks analysed in Section 4.1 consider an adversary that has access not only to the transaction dataset, but also to the demographic dataset, which contains revealing QIDs (such as zip code and age) that can be exploited in attacks. In this section, we consider privacy risks that remain even when demographic data is not available. Thus, in the examples below we still consider Charlize, an employee of fintech `F`, as the adversary, but we assume that the fintech does not maintain demographic data anymore. Charlize's target is now her acquaintance Bob. Furthermore, to investigate how transaction QIDs can be composed as transaction histories grow, we consider Charlize acting in a longitudinal scenario in which the transaction dataset is updated monthly with the corresponding novel activities performed by customers. We start with a simple motivating example:

Example 2. Charlize learned through some social media platform that Bob has recently eaten at Red Rooster. Using this information, she queries the transaction dataset (Table 2b) and finds that the only compatible transaction belongs to user ID 1. Consequently, Charlize can recover Bob's whole transaction history, which includes a transaction related to a fertility treatment!

Table 3: Risks of transaction-history recovery using the demographic dataset. Results are rounded to three decimal places. The first row corresponds to the case of an adversary whose auxiliary information does not include the target's zip code. The remaining rows correspond to the four combinations of demographic QIDs that resulted in the largest Bayes vulnerabilities, sorted in ascending order. Recall that $V(\pi)$ represents the prior Bayes vulnerability of the secret (i.e, the adversary's probability of guessing the secret correctly before observing the output of the system) and $V[\pi\triangleright C]$ represents the expected posterior vulnerability of the secret (after observing the output of the system). The prior is the same for all cases, and the leakage of information in each case is given by the ratio between posterior and prior vulnerabilities.

QIDs	$\boldsymbol{V(\pi)}$	$\boldsymbol{V[\pi\triangleright C]}$
Age, Gender, Employment	$5.8 \cdot 10^{-5}$	0.034
Gender, Zip Code, Employment		0.368
Age, Gender, Zip Code		0.835
Age, Zip Code, Employment		0.851
Age, Gender, Zip Code, Employment		0.905

Instantiating the Example Using QIF. Here, again, the secret X is Bob's transaction history, which can be seen as the set $\mathcal{X} = \{1, 2, 3, 4, 5\}$ of possible IDs. Before the attack is performed, Charlize has no reason to believe that any value of X is more likely than any other, so she considers a uniform prior π on $\mathcal{X}$. Hence, the prior Bayes vulnerability of the secret is $V(\pi) = {}^1\!/_5$.

Charlize can build, from the transaction dataset, the channel $C^{\text{Hist.}}$ mapping transaction histories to QIDs, as per (10). Then, by combining the uniform prior $\pi = ({}^1\!/_5, {}^1\!/_5, {}^1\!/_5, {}^1\!/_5, {}^1\!/_5)$ on IDs with channel $C^{\text{Hist.}}$, Charlize can obtain the joint $\pi\triangleright C^{\text{Hist.}}$, as per (2), and, from that, compute the posterior knowledge $[\pi\triangleright C^{\text{Hist.}}]$:

$$\begin{array}{c|ccccc|}
[\pi\triangleright C^{\text{Hist.}}] & p(y_1)=\frac{1}{10} & p(y_2)=\frac{1}{6} & p(y_3)=\frac{1}{5} & p(y_4)=\frac{7}{15} & p(y_4)=\frac{1}{15} \\
 & \text{Red Rooster} & \text{Clinic} & \text{Aldi} & \text{Uber} & \text{Lakeside} \\
1 & 1 & \frac{3}{5} & 0 & 0 & 0 \\
2 & 0 & 0 & 1 & 0 & 0 \\
3 & 0 & \frac{2}{5} & 0 & \frac{1}{7} & 1 \\
4 & 0 & 0 & 0 & \frac{3}{7} & 0 \\
5 & 0 & 0 & 0 & \frac{2}{7} & 0
\end{array}$$

The attack is actually executed when Charlize uses Bob's transaction payee (y_1 = Red Rooster) as a QID to identify that the corresponding posterior distribution in $[\pi\triangleright C^{\text{Hist.}}]$ (the column below y_1) assigns all probability to $x = 1$. Hence, the posterior *dynamic* Bayes vulnerability is $V^{y_1}[\pi\triangleright C^{\text{Hist.}}] = 1$, which means that, by using as QID the transaction's payee, Charlize's probability of recovering Bob's whole transaction history increased five-fold. The correspond-

Table 4: Transaction dataset (second month).

User ID	1	2	3	3	4	5
Payee	Uber	Transfer	Uber	Red Rooster	Uber	Transfer

ing expected posterior Bayes vulnerability, as per (6), is $V[\pi{\triangleright}C^{\text{Hist.}}] = {}^1\!/_{10} + {}^1\!/_{10} + {}^1\!/_5 + {}^1\!/_5 + {}^1\!/_{15} = {}^2\!/_3$. Consequently, the expected leakage is ${}^{({}^2\!/_3)}\!/_{({}^1\!/_5)} = {}^{10}\!/_3$.

Now, suppose that Charlize gains access to a second month of transaction data, summarised in Table 4. Charlize can construct a channel $D^{\text{Hist.}} : \mathcal{X} \to \mathbb{D}\mathcal{Y}'$ similar to $C^{\text{Hist.}}$, but for the second month of data. Then, due to independence of the observations, she can compose the two channels into the following channel $C^{\text{Hist.}} \parallel D^{\text{Hist.}} : \mathcal{X} \to \mathbb{D}(\mathcal{Y} \times \mathcal{Y}')$, where $\left(C^{\text{Hist.}} \parallel D^{\text{Hist.}}\right)_{x,(y,y')} = C^{\text{Hist.}}_{x,y} \cdot D^{\text{Hist.}}_{x,y'}$:

$$\begin{array}{c|ccccc|}
C^{\text{Hist.}} \parallel D^{\text{Hist.}} & \text{(RR, Uber)} & \text{(Clinic, Uber)} & \text{(Clinic, RR)} & \text{(Aldi, Transfer)} & \cdots \\
1 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & \cdots \\
2 & 0 & 0 & 0 & 1 & \cdots \\
3 & 0 & \frac{1}{6} & \frac{1}{6} & 0 & \cdots \\
4 & 0 & 0 & 0 & 0 & \cdots \\
5 & 0 & 0 & 0 & 0 & \cdots
\end{array}$$

Each output $(y, y') \in \mathbb{D}(\mathcal{Y} \times \mathcal{Y}')$ corresponds to one observation y in the first month and another y' in the second. In this case, the expected posterior Bayes vulnerability reaches ${}^{14}\!/_{15}$, and the corresponding expected leakage is ${}^{14}\!/_3$. Overall, the addition of only one extra month of data increased leakage by 40%.

Results. Table 5 shows the risks of transaction-history discovery for the four worst combinations of transaction QIDs (i.e., the four combinations that resulted in the largest Bayes vulnerabilities), in conjunction with the case when the exact amount is unknown to the adversary (i.e., the adversary knows only the transaction's date, payee, and category). We constructed the channel from the transaction dataset, as per (10), and considered a uniform prior on transaction histories. For this experiment, we considered four (consecutive) months of transactions. The database holds a total of 14 998 194 transactions (ranging from 3 037 108 to 5 121 799 per month) distributed into 42 073 financial histories. Consequently, the prior Bayes vulnerability is ${}^1\!/_{42073} = 2.37 \cdot 10^{-5}$.

Assuming that the adversary learns as QIDs the date, payee, amount, and category of one of their target's transactions, the expected chance of recovering their target's whole transaction history is 54%. Notice that the risk, although much smaller than the 90% found in Section 4.1 when the demographic dataset was also available, remains significant. It is worth noting that much of this risk comes from knowing the exact value of one transaction, given that the removal of the transaction's amount would significantly reduce the vulnerability, down to a 5% expected success rate. Nevertheless, as customers share more data over

Table 5: Risks of re-identification and of transaction-history recovery without using the demographic dataset. Results are rounded to three decimal places. The first row corresponds to the case of an adversary whose auxiliary information does not include the transaction's exact amount. The remaining rows correspond to the four combinations of transaction QIDs that resulted in the largest Bayes vulnerabilities, sorted in ascending order accordingly to the last column, i.e., $n = 4$, where n is the number of months considered in the longitudinal setup. Recall that $V(\pi)$ represents the prior Bayes vulnerability of the secret (i.e, the adversary's probability of guessing the secret correctly before observing the output of the system) and $V[\pi \triangleright C_1 \parallel \cdots \parallel C_n]$ represents the expected posterior vulnerability of the secret (after observing the output of the system for n months). The prior is the same for all cases, and the leakage of information in each case is given by the ratio between posterior and prior vulnerabilities.

	$\boldsymbol{V(\pi)}$	$\boldsymbol{V[\pi \triangleright C_1 \parallel \cdots \parallel C_n]}$			
QIDS		$\boldsymbol{n=1}$	$\boldsymbol{n=2}$	$\boldsymbol{n=3}$	$\boldsymbol{n=4}$
Date, Payee, Category	$2.73 \cdot 10^{-5}$	0.053	0.292	0.609	0.836
Date, Amount		0.266	0.807	0.975	0.998
Date, Payee, Amount		0.521	0.911	0.992	0.999
Date, Amount, Category		0.437	0.917	0.994	1.000
Date, Payee, Amount, Category		0.544	0.936	0.996	1.000

time, this risk increases considerably. A second month's worth of data raises the adversary's success rate to 93%. After a third month, the adversary can recover virtually every transaction history in the dataset, with 99.6% accuracy. And, even with no knowledge about the transaction value, a longitudinal attack over four months of financial activities can raise the adversary's success rate to 83%.

5 Quantification of Collateral Attribute-Inference Risks

In the previous section, we considered the risk of transaction-history recovery, which constitutes a privacy breach for individuals and can reveal very sensitive attributes present in the Open Banking database. In this section, we explore the more refined collateral attribute-inference attacks, where the adversary's goal is not to infer individuals' attributes *that are explicitly present in the Open Banking database*, but to exploit known correlations to infer sensitive information *not explicitly present in the Open Banking database.* To illustrate, consider the following correlation Π between number of transactions and neuroticism level: $\Pi = \{(\text{Low}, 1) : {}^2\!/_5, (\text{Mid}, 1) : {}^1\!/_5, (\text{Mid}, 2) : {}^1\!/_5, (\text{High}, 3) : {}^1\!/_5\}$. (Neuroticism is scored from 2 to 14, but for simplicity in this example we grouped the scores.)

Example 3. While examining Bob's transactions, Charlize recalls a scientific study that uncovers a strong correlation (the joint Π above) between buying habits and psychological traits. She then decides to apply this study to try to infer Bob's level of neuroticism. Charlize knows that Bob's transaction history is composed of two transactions. By linking this information with the correlation between number of transactions and neuroticism score, she can infer that Bob might have an intermediate level of neuroticism.

Instantiating the Example Using QIF. We want to quantify how much information channel $C^{\#\mathrm{Tr.}} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ mapping customers' transaction counts X to transaction QIDs Y (indirectly) leaks about the individuals' neuroticism level Z, which takes values in $\mathcal{Z} = \{\mathrm{Low}, \mathrm{Mid}, \mathrm{High}\}$. $C^{\#\mathrm{Tr.}}$ can be constructed from an extension of the transaction dataset, which incorporates the number of transactions made by each individual. Then, we multiply $C^{\#\mathrm{Tr}}$ by channel B obtained from the joint $\mathit{\Pi}$. The result is the channel $BC^{\#\mathrm{Tr.}}$ below:

$$\begin{array}{c|ccc} B & 1 & 2 & 3 \\ \hline \mathrm{Low} & 1 & 0 & 0 \\ \mathrm{Mid} & \frac{1}{2} & \frac{1}{2} & 0 \\ \mathrm{High} & 0 & 0 & 1 \end{array} \quad \begin{array}{c|ccccc} C^{\#\mathrm{Tr.}} & \mathrm{RR} & \mathrm{Clinic} & \mathrm{Aldi} & \mathrm{Uber} & \mathrm{Lakeside} \\ \hline 1 & 0 & 0 & \frac{1}{3} & \frac{2}{3} & 0 \\ 2 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 3 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \end{array} = \begin{array}{c|ccccc} BC^{\#\mathrm{Tr.}} & \mathrm{RR} & \mathrm{Clinic} & \mathrm{Aldi} & \mathrm{Uber} & \mathrm{Lakeside} \\ \hline \mathrm{Low} & 0 & 0 & \frac{1}{3} & \frac{2}{3} & 0 \\ \mathrm{Mid} & \frac{1}{4} & \frac{1}{4} & \frac{1}{6} & \frac{1}{3} & 0 \\ \mathrm{High} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \end{array}$$

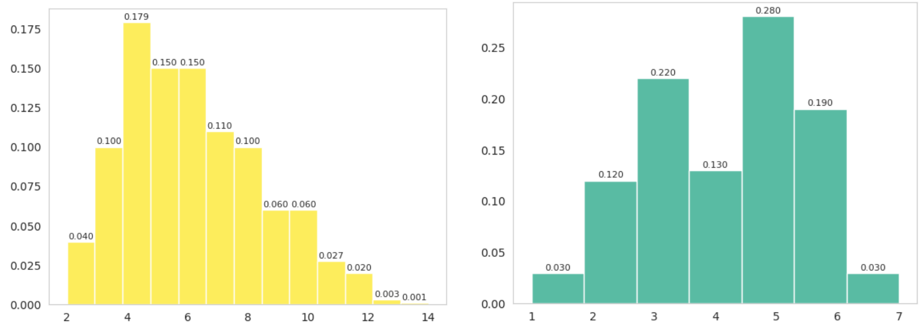
By combining channel $BC^{\#\mathrm{Tr.}}$ with a prior distribution $\rho = ({}^2\!/\!{}_5, {}^2\!/\!{}_5, {}^1\!/\!{}_5)$ on neuroticism levels (also extracted from the correlation $\mathit{\Pi}$) and employing Bayesian reasoning, Charlize obtains the posterior knowledge

$$\begin{array}{c|ccccc} [\pi \triangleright BC^{\#\mathrm{Tr.}}] & \begin{array}{c} p(y_1) = \frac{1}{10} \\ \text{Red Rooster} \end{array} & \begin{array}{c} p(y_2) = \frac{1}{6} \\ \text{Clinic} \end{array} & \begin{array}{c} p(y_3) = \frac{1}{5} \\ \text{Aldi} \end{array} & \begin{array}{c} p(y_4) = \frac{7}{15} \\ \text{Uber} \end{array} & \begin{array}{c} p(y_4) = \frac{1}{15} \\ \text{Lakeside} \end{array} \\ \hline \mathrm{Low} & 0 & 0 & \frac{2}{3} & \frac{4}{7} & 0 \\ \mathrm{Mid} & 1 & \frac{3}{5} & \frac{1}{3} & \frac{2}{7} & 0 \\ \mathrm{High} & 0 & \frac{2}{5} & 0 & \frac{1}{7} & 1 \end{array}$$

Then, knowing that Bob went to Red Rooster, Charlize updates her knowledge to the posterior under column y_1. Hence, in this scenario she can be confident that Bob has an intermediate level of neuroticism. In the expected case, the adversary's chance of discovering someone's neuroticism level is $\mathcal{D}V[\mathit{\Pi} \triangleright C^{\#\mathrm{Tr.}}] = {}^1\!/\!{}_{10} + {}^1\!/\!{}_{10} + {}^2\!/\!{}_{15} + {}^4\!/\!{}_{15} + {}^1\!/\!{}_{15} = {}^2\!/\!{}_3$, as per (9), which is ${}^5\!/\!{}_3$ times higher than their prior chance of success $V(\rho) = {}^2\!/\!{}_5$ (i.e., before learning any transaction QIDs).

Results. We now assess the privacy risks of collateral-inference contexts in a real-world dataset. The experimental setup is similar to that adopted in Section 4.1. We assess the information leaked from the Open Banking system about someone's psychological trait — neuroticism, scored from 2 to 14, and self-control, scored from 1 to 7 — due to a correlation with that person's spending behaviour. To characterise spending behaviour, we chose three of the metrics analysed by Tovanich et al. [26]: the total number of transactions (n_{tot}), the total amount (a_{tot}), and the average amount (a_{avg}) spent by each customer. We considered only the integer parts of a_{tot} and a_{avg}.

To construct the collateral-inference scenarios, we require a joint distribution between the psychological traits that the adversary seeks and the spending metrics. Tovanich et al. [26] did not provide the joint distributions that correspond



(a) Prior ρ^{neuro} on neuroticism

(b) Prior ρ^{self} on self-control

Fig. 3: Prior distributions on psychological traits.

to the data they used in their experiments. However, they provided some useful Pearson correlations, together with plots of the marginal distributions on the psychological traits scores. Using these, we determined the marginal distributions from the plots for neuroticism and self-control.

The derived prior (marginal) distributions ρ^{neuro} and ρ^{self} for, respectively, neuroticism and self-control are depicted in Figure 3. In the absence of auxiliary knowledge, from the prior distribution ρ^{neuro} we conclude that the adversary's optimal guess for a random individual's neuroticism score is 4, the secret value that maximises prior Bayes vulnerability: $V(\rho^{neuro}) = 0.179$. Similarly, the a priori optimal guess for self-control is 5, with a corresponding prior Bayes vulnerability of $V(\rho^{self}) = 0.280$.

For the attack itself, we constructed synthetic datasets correlating the spending metrics and the personality trait scores. For that, we first generate scores for each user in the transaction dataset, following the prior distributions in Figure 3. Then, we shuffle the trait scores and the metric values. After that, we iteratively randomly choose users (indices), sort the trait scores in ascending order if the target Pearson correlation is non-negative or in descending order otherwise, and sort the metric values in ascending order, until achieving a Pearson correlation as close as possible to the target. (See Appendix B for the pseudocode that was used in the construction of the joints and for the Pearson correlations.)[6]

Table 6 shows the posterior collateral Bayes vulnerability and the corresponding information leakage in our experiments. Knowledge of date, payee, amount, and category of one of their target's transactions and access to the customers' transaction count n_{tot} increases the adversary's probability of guessing their tar-

[6] We reinforce that the use of synthetic joints was due to the fact that we only had access to correlations in the form of Pearson correlations; the datasets from which these correlations were computed are not publicly available. Nevertheless, we do believe the synthetic data is illustrative of the kind of concrete threat we are considering.

Table 6: Information leakage about individuals' psychological traits, given correlations with their spending behaviour. Results correpond to the 95% confidence interval of 30 randomly constructed joints.

Trait	**Metric**	$\boldsymbol{V(\rho)}$	$\boldsymbol{\mathcal{D}V[\Pi \triangleright C]}$	**Information Leakage**
Neuroticism	n_{tot}	0.179	$[0.198, 0.200]$	$[1.114, 1.119]$
	a_{tot}		$[0.600, 0.601]$	$[3.362, 3.364]$
	a_{avg}		$[0.597, 0.598]$	$[3.346, 3.348]$
Self-control	n_{tot}	0.280	$[0.290, 0.292]$	$[1.038, 1.041]$
	a_{tot}		$[0.616, 0.620]$	$[2.202, 2.214]$
	a_{avg}		$[0.641, 0.643]$	$[2.292, 2.294]$

get's neuroticism and self-control score by a factor of, respectively, 1.114 and 1.038 (reaching a posterior success of about 19% and 29%, respectively).

In constrast, knowledge of the total or average amount spent by customers incur on a much higher privacy risk. Access to the total amount spent by customers (a_{tot}) boosts the adversary's expected success by a factor of 3.362 for neuroticism and a factor of 2.202 for self-control. Finally, if the adversary knows the correlation between personality traits and the average amount spent by customers, their chance of identifying their target's trait correctly increases by a factor of 3.346 for neuroticism and by a factor of 2.292 for self-control. In all four scenarios, the posterior Bayes vulnerability reached around 60%.

6 Discussion and Conclusion

Our study has formalised and highlighted the risks of data sharing for consumers, for both transaction-recovery and indirect attribute-inference risks (via collateral-inference scenarios) in a real Open Banking system in Australia. Although our analysis focuses on the risks associated with customers who have provided consent to use their data, we argue that the onus should be on regulators to ensure that privacy-risk mitigation is inherent in the design of data-sharing protocols. Furthermore, we highlight the issue of accountability for privacy breaches and with whom responsibility lies. Moves to increase consumer controls over data amplify the chances that consumers will unwittingly expose themselves to privacy risks. It appears that consumer-controlled sharing exonerates institutions from obligations regarding private data sharing and forces consumers to absorb that risk. Our view is that it should be up to organisations and regulators to provide guidelines to assist consumers in making informed decisions.

On the mechanisation of our approach. The QIF framework provides formulas according to which the leakage of sensitive information is measured (in terms of prior and posterior vulnerabilities) taking as input: (i) a prior distribution representing the adversary's prior knowledge and (ii) a channel representing the system's behaviour. To compute such formulas in practice, we need to first model the prior and the channel in each case of interest (often writing personalised code to extract the prior and the channel from the data), and then these parameters can be passed onto a QIF library that can compute leakage.

Analyses of other scenarios. We highlight that, although this work focused on the implementation of Open Banking in Australia, by analysing one particular dataset provided by a fintech company, the framework described can be used to quantify the privacy risks of any other similar Open Banking model. However, this requires access to other real-world datasets that reflect the specification of the Open Banking system that one wishes to analyse. Consequently, the results may vary according to the data.

Key Recommendations. Our research shows that the release of arbitrary text fields in transaction data leads to unnecessary privacy risks to customers, and privacy liabilities to small fintechs. To our knowledge, these fields are only used to identify transaction categories, and could be redacted to disclose far less information than they do at present. The transactions' payee field may also carry sensitive information, and thus should (at the very least) be treated before being shared. We also recommend to customers that no more than one month's worth of data is released to prevent longitudinal attacks, which are much more damaging with the release of additional data. However, it is unclear how much utility is lost to the fintech in this case (i.e., how much information they would require in order to make a reasonable financial judgement). We leave this investigation to future work. We also believe that the release of demographic data leads to unnecessary privacy invasions above what utility this information might provide, and we recommend that demographic data on individual users is not collected, or, at the least, not directly associated with the provided transaction data.

Future Work. We plan to extend our analysis to other collateral-inference scenarios (e.g., health issues, political preferences) and use the theory of QIF to limit the damage caused by such attacks over all possible correlations and gain functions using collateral (Dalenius) capacity [3]. We also want to study the trade-off between privacy and utility if different privacy mechanisms are employed.

Acknowledgments. Luigi D. C. Soares was supported by the Brazilian National Council of Scientific and Technological Development (CNPq), under grant number 140359/2022-2, and by Macquarie University, via the International Cotutelle Macquarie University Research Excellence Scholarship (Cotutelle "iMQRES").

References

1. Alvim, M.S., Andrés, M.E., Chatzikokolakis, K., Degano, P., Palamidessi, C.: On the information leakage of differentially-private mechanisms. Journal Computer Security **23**(4), 427–469 (2015). https://doi.org/10.3233/JCS-150528, https://doi.org/10.3233/JCS-150528

2. Alvim, M.S., Chatzikokolakis, K., McIver, A., Morgan, C., Palamidessi, C., Smith, G.: Additive and Multiplicative Notions of Leakage, and Their Capacities. In: Proc. of CSF. pp. 308–322. IEEE (2014). https://doi.org/10.1109/CSF.2014.29
3. Alvim, M.S., Chatzikokolakis, K., McIver, A., Morgan, C., Palamidessi, C., Smith, G.: The Science of Quantitative Information Flow. Springer (2020)
4. Alvim, M.S., Fernandes, N., McIver, A., Morgan, C., Nunes, G.H.: Flexible and scalable privacy assessment for very large datasets, with an application to official governmental microdata. Proc. Priv. Enhancing Technol. **2022**(4), 378–399 (2022). https://doi.org/10.56553/popets-2022-0114, https://doi.org/10.56553/popets-2022-0114
5. Alvim, M.S., Fernandes, N., McIver, A., Morgan, C., Nunes, G.H.: A novel analysis of utility in privacy pipelines, using Kronecker products and quantitative information flow. In: Proc. SIGSAC (2023). https://doi.org/10.1145/3576915.362308
6. Alvim, M.S., Fernandes, N., Nogueira, B.D., Palamidessi, C., Silva, T.V.A.: on the duality of privacy and fairness (extended abstract). In: Proc. CADE (2023). https://doi.org/10.1049/icp.2023.2563
7. Behling, O.: Employee selection: Will intelligence and conscientiousness do the job? Academy of Management Perspectives **12**(1), 77–86 (1998)
8. Braun, C., Chatzikokolakis, K., Palamidessi, C.: Quantitative notions of leakage for one-try attacks. Electronic Notes in Theoretical Computer Science **249**, 75–91 (2009)
9. Butz, N.T., Stratton, R., Trzebiatowski, M.E., Hillery, T.P.: Inside the hiring process: how managers assess employability based on grit, the big five, and other factors. INTERNATIONAL JOURNAL OF BUSINESS ENVIRONMENT **10**(4), 306–328 (2019)
10. Chatzikokolakis, K., Fernandes, N., Palamidessi, C.: Comparing systems: Max-case refinement orders and application to differential privacy. In: Proc. CSF (2019). https://doi.org/10.1109/CSF.2019.00037
11. Dalenius, T.: Finding a needle in a haystack or identifying anonymous census records. Journal of official statistics **2**(3), 329 (1986)
12. Di Clemente, R., Luengo-Oroz, M., Travizano, M., Xu, S., Vaitla, B., González, M.C.: Sequences of purchases in credit card data reveal lifestyles in urban populations. Nature communications **9**(1), 3330 (2018)
13. Fernandes, N., Dras, M., McIver, A.: Processing text for privacy: an information flow perspective. In: Proc. FM (2018). https://doi.org/10.1007/978-3-319-95582-7_1
14. Freebairn, P.: Response to the farrell report into open banking. Policy (2018)
15. Gladstone, J.J., Matz, S.C., Lemaire, A.: Can psychological traits be inferred from spending? evidence from transaction data. Psychological Science **30**(7), 1087–1096 (2019). https://doi.org/10.1177/0956797619849435, https://doi.org/10.1177/0956797619849435, pMID: 31166847
16. Judge, T.A., Ilies, R.: Relationship of personality to performance motivation: a meta-analytic review. Journal of applied psychology **87**(4), 797 (2002)
17. Jurado, M., Palamidessi, C., Smith, G.: A formal information-theoretic leakage analysis of order-revealing encryption. In: Proc. CSF (2021). https://doi.org/10.1109/CSF51468.2021.00046
18. de Montjoye, Y.A., Radaelli, L., Singh, V.K., Pentland, A.S.: Unique in the shopping mall: On the reidentifiability of credit card metadata. Science **347**(6221), 536–539 (2015). https://doi.org/10.1126/science.1256297, https://www.science.org/doi/abs/10.1126/science.1256297

19. Moy, J., Lam, K.: Selection criteria and the impact of personality on getting hired. PERSONNEL REVIEW **33**(5-6), 521–535 (2004). https://doi.org/10.1108/00483480410550134
20. Narayanan, A., Shmatikov, V.: Robust De-anonymization of Large Sparse Datasets. In: Proc. of S&P. pp. 111–125 (2008). https://doi.org/10.1109/SP.2008.33
21. Rényi, A.: On measures of entropy and information. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. vol. 4, pp. 547–562. University of California Press (1961)
22. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression (1998)
23. Silva, R.M., Gomes, G.C.M., Alvim, M.S., Gonçalves, M.A.: How to build high quality L2R training data: Unsupervised compression-based selective sampling for learning to rank. Information Sciences **601** (2022). https://doi.org/10.1016/j.ins.2022.04.012
24. Smith, G.: On the Foundations of Quantitative Information Flow. In: FOSSACS. LNCS, vol. 5504. Springer (2009)
25. Sweeney, L.: Simple demographics often identify people uniquely. Health (San Francisco) **671**(2000), 1–34 (2000)
26. Tovanich, N., Centellegher, S., Bennacer Seghouani, N., Gladstone, J., Matz, S., Lepri, B.: Inferring psychological traits from spending categories and dynamic consumption patterns. EPJ Data Science **10**(1), 24 (May 2021). https://doi.org/10.1140/epjds/s13688-021-00281-y, https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-021-00281-y
27. Viegas, F., Alvim, M.S., Canuto, S.D., Rosa, T., Gonçalves, M.A., Rocha, L.: Exploiting semantic relationships for unsupervised expansion of sentiment lexicons. Information Systems **94** (2020). https://doi.org/10.1016/j.is.2020.101606

A More Details on Quantitative Information Flow

The leakage measures used in Quantitative Information Flow (QIF) were developed by the security community in line with the principle that leakages should correspond with an adversarial attack, and if the leakage of a system increases, this increase should be *justifiable* by demonstrating an adversary who is able to learn more from the leakier system. However, leakages can be difficult to interpret in practice, and so we give here some toy examples to explain the leakage measures used in this paper and how to use them.

Let us say that we are given a dataset of 100 users in which one user is identifiable with certainty (using their QIDs) and every other user is identifiable with at most 50% probability. This could be depicted by the following channel:

$$\begin{array}{c|cccccc|}
C & Q_1 & Q_2 & Q_3 & \cdots & Q_{n-1} & Q_n \\
u_1 & 1 & 0 & 0 & \cdots & 0 & 0 \\
u_2 & 0 & 1 & 0 & \cdots & 0 & 0 \\
u_3 & 0 & 1 & 0 & \cdots & 0 & 0 \\
u_4 & 0 & 1 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
u_{99} & 0 & 0 & 0 & \cdots & 0 & 1 \\
u_{100} & 0 & 0 & 0 & \cdots & 0 & 1
\end{array}$$

Now, the risk to user u_1 of having their data breached is *not* the same as the leakage of the system. The risk to a user (with respect to an attack) is given by the posterior vulnerability; this measure focuses on the posterior knowledge of the adversary, i.e., that *after the attack*. The leakage of the system tells us how much the system itself contributes to the attacker's *gain in knowledge* with respect to the prior state, i.e., that *before the attack was performed*.

Let us denote the dynamic Bayes leakage of a system C, assuming a prior π and given an observation y, by $\mathcal{L}^y(\pi, C)$. In the above case, the dynamic posterior vulnerability for the individual user (using a uniform prior υ) is $V^{Q_1}[\upsilon \triangleright C] = 1$, meaning that their probability of being re-identified given observation Q_1 is 1. The dynamic leakage of this system for user u_1 is $\mathcal{L}^{Q_1}(\upsilon, C) = {}^1/_{({}^1/_{100})} = 100$. Hence, the adversary's knowledge has increased 100-fold.

Now consider a dataset of 10 000 users in which one is identifiable with 10% probability. This could be depicted by the channel

$$\begin{array}{c|cccccc|}
D & Q_1 & Q_2 & Q_3 & \cdots & Q_{n-1} & Q_n \\
u_1 & 1 & 0 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
u_{10} & 1 & 0 & 0 & \cdots & 0 & 0 \\
u_{11} & 0 & 1 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
u_{99} & 0 & 0 & 0 & \cdots & 0 & 1 \\
u_{100} & 0 & 0 & 0 & \cdots & 0 & 1
\end{array}$$

Here the first 10 users share the same QIDs Q_1. Therefore, the probability of an adversary guessing user u_1 given the observation Q_1 (under a uniform prior υ) is $V^{Q_1}[\upsilon{\triangleright}D] = {}^1\!/\!_{10}$, as expected, and the dynamic leakage of this system for user u_1 and the observation Q_1 is $\mathcal{L}^{Q_1}(\upsilon, D) = ({}^1\!/\!_{10})/({}^1\!/\!_{10000}) = 1000$. Notice that this is higher than what we computed for channel C, which might be interpreted as indicating that this channel D is less safe for user u_1 than is channel C. However, the leakage between two systems operating on different secret spaces cannot be compared in this way, since the adversary has different priors on the different spaces (thus, we are not comparing apples with apples). If we were just considering user u_1, we would just compare posterior vulnerabilities, which implies that channel D is safer than channel C (for that user).

In terms of when to rely on leakages, let us consider now the channel C_2 below, which has the same secret space as C. However, this time no user is vulnerable to attack with certainty.

$$\begin{array}{c|cccccc|}
C_2 & Q_1 & Q_2 & Q_3 & \cdots & Q_{n-1} & Q_n \\
u_1 & 1 & 0 & 0 & \cdots & 0 & 0 \\
u_2 & 1 & 0 & 0 & \cdots & 0 & 0 \\
u_3 & 0 & 1 & 0 & \cdots & 0 & 0 \\
u_4 & 0 & 1 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
u_{99} & 0 & 0 & 0 & \cdots & 0 & 1 \\
u_{100} & 0 & 0 & 0 & \cdots & 0 & 1
\end{array}$$

Now we can compute the dynamic leakage of this system for user u_1 and observation Q_1 and compare it with the corresponding dynamic leakage of C. For C_2, we have $\mathcal{L}^{Q_1}(\upsilon, C_2) = ({}^1\!/\!_2)/({}^1\!/\!_{100}) = 50$, meaning that the adversary's knowledge has increased 50-fold on C_2, compared with their 100-fold increase on C, indicating that C_2 is safer than C for user u_1.

Alternatively, we can compare the multiplicative leakages for the more general cases (expected and maximum gain of the adversary, where the max-case takes into account the "best" observation — in the adversary's perspective — irregardless of its probability):

$$\begin{array}{rclcrcl}
\mathcal{L}(\upsilon, C) & = & n & \qquad & \mathcal{L}(\upsilon, C_2) & = & n \\
\mathcal{L}^{max}(\upsilon, C) & = & 100 & & \mathcal{L}^{max}(\upsilon, C_2) & = & 50
\end{array}$$

This shows that channel C is not better than C_2 for any leakage measure, and is worse for the worst-case measures that we use.

B Construction of Joints from Pearson Correlations

As explained in Section 2, a collateral attribute-inference attack requires a joint Π representing a known correlation between secrets Z and X. However, as mentioned in Section 5, we only had access to correlations in the form of *Pearson*

Table 7: Comparison between the Pearson correlations given in [26] and the correlations we obtained for the synthetically built datasets, when demographic information is available. Numbers correspond to the 95% confidence interval of 1 000 randomly constructed joints. The interval's lower and upper limits are rounded down and up to four decimal places.

Trait	**Metric**	**Pearson [26]**	**Pearson achieved**
Neuroticism	n_{tot}	-0.0735	$[-0.0735, -0.0734]$
	a_{tot}	-0.1644	$[-0.1308, -0.1307]$*
	a_{avg}	-0.1496	$[-0.1496, -0.1495]$
Self-control	n_{tot}	-0.0717	$[-0.0717, -0.0716]$
	a_{tot}	$+0.0976$	$[+0.0975, +0.0976]$
	a_{avg}	$+0.1524$	$[+0.1523, +0.1524]$

* Target correlation is not reachable with the data available.

correlations; the datasets from which these correlations were computed are not publicly available. In view of this, we opted to construct synthetic joints from the Pearson correlations, using the transaction dataset we have available and the Pearson correlations and marginal distributions on the personality traits given in [26]. Algorithm 1 shows the pseudocode for the construction of such joints.

Table 7 shows the Pearson correlations from [26] between neuroticism/self-control and each of the chosen spending metrics, side by side with the Pearson correlations achieved in the synthetic joints that we constructed.

Algorithm 1 Pseudocode for generating synthetic joints, given two lists, Z and X, and a Pearson correlation between Z and X. The input Z here would be, e.g., a list of neuroticism scores, distributed according to a marginal distribution on neuroticism. And, X would be a list of people's total number of transactions, computed from the transaction dataset. Assume that *pearson* is an existing function that computes the Pearson correlation between two lists.

```
function GENERATE_JOINT($Z$, $X$, target_pcorr, $\epsilon$)
    // First, update target if original target is unreachable
    if target_pcorr $\geq 0$ then
        upper_limit $\leftarrow$ pearson_upper_limit($Z, X$)
        target_pcorr $\leftarrow$ min(target_pcorr, upper_limit)
    else
        lower_limit $\leftarrow$ pearson_lower_limit($Z, X$)
        target_pcorr $\leftarrow$ max(target_pcorr, lower_limit)
    end if
    // Then, shuffle data to get an initial joint
    $Z \leftarrow$ shuffle $Z$
    $X \leftarrow$ shuffle $X$
    curr_pcorr $\leftarrow$ pearson($Z, X$)
    $n \leftarrow |Z| * 10/100$
    // Finally, iteratively sort data at random
    while | curr_pcorr - target_pcorr | $> \epsilon$ do
        indices $\leftarrow$ choose $n$ indices at random
        if target_pcorr $\geq 0$ then
            $Z' \leftarrow Z$, with $Z$[indices] sorted in ascending order
        else
            $Z' \leftarrow Z$, with $Z$[indices] sorted in descending order
        end if
        $X' \leftarrow X$, with $X$[indices] sorted in ascending order
        new_pcorr $\leftarrow$ pearson($Z', X'$)
        // Due to sorting, pcorr always goes in one direction
        if | new_pcorr | $<$ | target_pcorr | $+ \epsilon$ then
            $Z \leftarrow Z'$
            $X \leftarrow X'$
            curr_pcorr $\leftarrow$ new_pcorr
        else
            $n \leftarrow$ int($n/2$)
        end if
    end while
    return $(Z, X)$
end function
```