UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computaçcão

Danilo Fabrino Favato

CONFIDENTIAL DATA LEAKAGE IN BRAZILIAN OFFICIAL FOREIGN TRADE STATISTICS

Belo Horizonte

2021

Danilo Fabrino Favato

CONFIDENTIAL DATA LEAKAGE IN BRAZILIAN OFFICIAL FOREIGN TRADE STATISTICS

Versão final

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientador: Gabriel de Morais Coutinho
Coorientador: Mário Sérgio Ferreira Alvim

Belo Horizonte
2021

Favato, Danilo Fabrino

F272c Confidential data leakage in Brazilian official foreign trade statistics [manuscrito] / Danilo Fabrino Favato – 2021.
62 f. il.

Orientador: Gabriel de Morais Coutinho.
Coorientador: Mário Sérgio Ferreira Alvim.
Dissertação (mestrado) — Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação
Referências: f. 61 – 62

1. Computação – Teses. 2. Fluxo quantitativo de informação – Teses. 3. Programação inteira – Teses. 4. Comércio exterior – Estatística – Teses. 5. Vazamento de informação – Teses. I. Coutinho, Gabriel de Morais. II. Alvim, Mário Sérgio Ferreira. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. IV. Título.

CDU 519.6*44 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Rezende Costa CRB 6/1510 – Instituto de Ciências Exatas da UFMG.

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

CONFIDENTIAL DATA LEAKAGE IN BRAZILIAN OFFICIAL FOREIGN TRADE STATISTICS

DANILO FABRINO FAVATO

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. GABRIEL DE MORAIS COUTINHO - Orientador
Departamento de Ciência da Computação - UFMG

Prof. Mário Sérgio Ferreira Alvim - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. CRISTIANO ARBEX VALLE
Departamento de Ciência da Computação - UFMG

PROF. JEROEN ANTONIUS MARIA VAN DE GRAAF
Departamento de Ciência da Computação - UFMG

JULIANO BRITO DA JUSTA NEVES
Tecnologia e Segurança da Informação - Receita Federal do Brasil

Dra. NATASHA FERNANDES
Ciência da Computação - Macquaire University

Belo Horizonte, 20 de Outubro de 2021.

Com gratidão, dedico este trabalho ao povo brasileiro que, através de seus impostos, me permitiu estudar nesta universidade pública.

Acknowledgments

I would like to thank the following people who have helped me undertake this challenge: My parents, Mário Nazareno Favato e Ana Maria Fabrino Favato, for their unconditional love and support throughout every single day of my life; my advisors, Gabriel Coutinho e Mário S. Alvim, for their extremely constructive criticism and guidance; and my friends whose company made this journey way longer than expected, but also more fulfilling.

"Writing laws is easy, but governing is difficult."
(Leo Tolstoy)

Resumo

Este trabalho apresenta um ataque, nunca antes documentado, de reconstrução de base dados que pode ser executado sobre estatísticas oficiais de comércio exterior Brasileiras de forma a revelar dados de empresas que são protegidos por leis de sigilo fiscal. Um algoritmo de otimização inteira é utilizado no cruzamento de dados inteiramente públicos e a quantidade de informação vazada é medida através da teoria de Fluxo Quantitativo de Informação. A análise inicial, desenvolvida aqui, sobre o provável alcance desse tipo de ataque indica que, apenas no mês Janeiro de 2021, mais de 348 importadores podem ter seus dados sigilosos vazados, o que representa mais de 137 milhões de dólares.

Palavras-chave: Fluxo quantitativo de informação, Programação de otimização inteira, Ataque de reconstrução de bases de dados, Estatísticas de comércio exterior.

Abstract

This work presents an undocumented database reconstruction attack that can be used over Brazilian official foreign trade statistics to reveal businesses' data which are subject to fiscal secrecy laws. An integer programming algorithm is applied over entirely public datasets and the information leakage is measured within the Quantitative Information Flow framework. The initial analysis, developed here, regarding the potential reach of this kind of attack shows that, accounting for only the month of January 2021, more than 348 importers might have their confidential data leaked which represents more than 137 million US dollars.

Palavras-chave: Quantitative information flow, Integer optimization programming, Database reconstruction attack, Foreign trade statistics.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

This study aims to present an undocumented attack that can be performed over Brazilian foreign trade statistics that enables the linkage between companies and their import transactions. The current Brazilian legislation not only forbids the data publishers[1] from disclosing such links but also regards the tax secrecy, that might be violated by this attack, as a highly protected right.

The database reconstruction attack is conducted through the usage of a custom-designed *Mixed Integer Programming* algorithm and the information leakage results are presented within the *Quantitative Information Flow* theory. Nonetheless, this text tries to be accessible for those who are unfamiliar with these knowledge areas as might be the case for some of the stakeholders of the datasets explored here.

1.1 Motivation

The current legal framework regarding governmental data usage has created two almost opposable objectives that public institutions must pursue: publicity and confidentiality. On the publicity side, there is the increasing advocacy for *Open Government Data*, which "promotes transparency, accountability and value creation by making government data available to all".[2] On the confidentiality side, the concern lies in the unauthorized disclosure of sensitive data about individuals or businesses, a worry that has sprouted many data protection laws around the globe.[3]

[1] Ministério da Economia and Receita Federal do Brasil (RFB)

[2] https://web.archive.org/web/20201120073847/https://www.oecd.org/gov/digital-government/open-government-data.htm

[3] General Data Protection Regulation in the European Union (GDPR 2016/679); Lei Geral de Proteção de Dados Pessoais in Brazil (LGPD 13.709/2018); Personal Information Protection and Electronic Documents Act in Canada (PIPEDA April 13th 2000); and The Data Privacy Act in the

Although these novel protection laws, as well as media and public attention, tend to focus on protecting data about individuals, this study is about the unwanted disclosure of commercial transactions between companies. Some people may be less concerned about businesses' privacy rights and disregard the importance of protecting them. However, as will be shown below, Brazilian laws are categorical about how businesses should be protected and their importance to a healthy business environment. Also, we invite these less sympathetic towards business rights to rethink their position in the light of past events, like when the NSA was accused of worldwide industrial espionage and the shock produced by these allegations.[4]

Particularly, here, we demonstrate how Brazilian foreign commerce statistics, published by government institutions, can be used to violate the legally protected tax secrecy of local businesses and, ultimately, cause economic harm to those companies.

To provide the evidence that supports this claim, this text will be organized into the following parts:

1. In Section (1.2) we briefly introduce the intuition of how the attack works.

2. In Chapter 2 we present the *Quantitative Information Flow* concepts that are used in the *information leakage* measurement and the *adversary's* chance of success.

3. In Chapter 3, we formally define the *Mixed Integer Programming* algorithm that is used in the attack.

4. In Chapter 4, we present the implementation details and foreign commerce statistics methodology aspects that are important to reproduce the findings of this study.

5. Finally, in Chapter 5 we discuss the attack's consequences and reach by

    a) estimating how many businesses might be in jeopardy and what might be the global economic damage.

    b) explaining, within the Brazilian current legal framework, why this identification might be violating the underlying businesses' tax secrecy rights

Philippines (2012), for example.

[4] https://web.archive.org/web/20210417032720if_/https://www.reuters.com/article/us-security-snowden-germany-idUSBREA0P0DE20140126

1.2 An intuition on how the attack works

This section provides the intuition behind the attack through a concrete example. Knowing this outline will help understand some of the theoretical concepts that will be presented later, as it will be easier to locate where those might fit in this general overview.

The terms in *italics* here might have everyday use but are formally defined in Chapter 2 to avoid confusion.

This example will showcase how it is possible to attribute to a specific and uniquely identifiable company[5] a transaction made in January 2021, which amounts to almost 5 million USD. To attribute a record in a *de-identified* database to a specific individual (a company in our case) is classified as a re-identification attack. [13]

1.2.1 Re-identification in our context

Figure 1.1 roughly depicts how foreign trade statistics in Brazil are produced. It starts with the companies which through their internal processes produce a set of **Import transactions**. This information is used internally by these companies and also sent to governmental agencies to fulfill legal requirements, like tax payment.

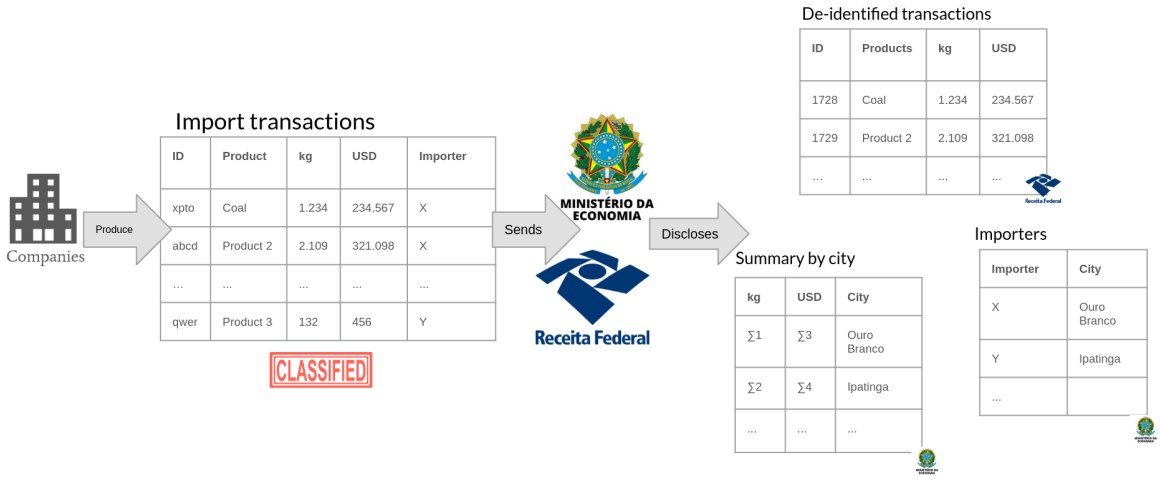


Figure 1.1. Overview of the information flow for foreign trade statistics

The raw data in **Import transactions** is private and it is not publicly available. However, the Brazilian Tax and Customs Administration (RFB[6]) and the Ministry of Economy are also obliged by transparency laws to publish data about foreign

[5]This company's name will be omitted in the text.
[6]Receita Federal do Brasil

commerce. To comply with these requirements both agencies apply *disclosure control* techniques, like *de-identification* and *generalization*. The RFB publishes a set of **De-identified transactions**, those are the same records seen in the **Import transactions** but the importers' *direct identifiers* are omitted. The Ministry of Economy publishes two other datasets: a **Summary by city** which aggregates the values (USD, kg) by the importers' city, and a registry of **Importers** that contains registration information about the companies that have done at least one foreign transaction. The *disclosure methods* applied by these agencies aim to break the connection between the goods imported and the companies behind such transactions. In this way the import transactions are de-identified.

A re-identification occurs when an adversary is able to reconstruct the link between the **De-identified transactions** and the **Importers**. How this adversary does such a feat depends on what is known. We will arbitrarily choose an adversary with very limited knowledge to demonstrate the attack. What we want to convey by doing so is how powerful the attack is. If this attack is performed by stronger adversaries, then we can only expect that their probability of success will be higher.

1.2.2 Prior knowledge

The adversaries' prior knowledge strictly defines what they know before starting the attack. Even though real-world adversaries have some common knowledge (like that steel mills are more likely to import coal than toy stores), we assume a very modest adversary, who does not know anything like that, they[7] only have access to two datasets:

1. Importers: A set of businesses that are known to have imported some good in January 2021. This is the one disclosed by the Ministry of Economy. This set has 18,430 records;

2. De-identified transactions: A set of all international commercial transactions, made in January 2021 disclosed by RFB. The size of this set is 817,468 transactions.

The sources and details about these datasets are in Section 4.1.1.

[7]The singular gender-neutral pronouns they/their will be used to refer to the adversary throughout the text.

Table 1.1. Arbitrarily chosen transaction that the adversary wants to *re-identify*

NR ORDEM	*Description*	*Qtt.*	*Value USD*
17280000100001	HULHA BETUMINOSA - CARVAO DE PEDRA, EM BRUTO, A GRANEL PARA PREPARO DE COQUE MARCA "WARRIOR BLUE CREEK LV". ESPECIFICACOES: UMIDADE (COMO RECEBIDA): 8,78%; MATERIA VOLATIL (BASE SECA): 20,60%; CINZA (BASE SECA): 10,03%; ENXOFRE (BASE SECA): 0,70%;VAL	39,783 tonnes	4,924,259.04

1.2.3 The goal

For the sake of this demonstration, the adversary's goal is to re-identify the arbitrarily chosen transaction presented in Table 1.1, *i.e.* to identify the company that made the import. Given the restrictions imposed by what is known a priori, the adversary will always try to maximize the chances of achieving this goal.

Throughout this example, this transaction will be referenced by the first 9 digits of its *NR ORDEM* (172800001).

1.2.4 Prior success

Given that the adversary already has some prior knowledge, they can try to achieve the goal above by making some type of guess. The way this guess is made will determine their chance of success before (prior to) the attack.

For instance, we can say that the adversary can randomly choose a company from the Importers dataset, and attribute to it the target transaction and their chance of success will be $1/18{,}430 \approx 0.005\%$. Although this feels intuitive care must be taken about the hidden assumptions here. The adversary is always trying to maximize their chance of getting their goal right, they will always choose the company with the highest probability of being the importer. By saying their prior success is $1/18{,}430$, we are assuming that the adversary views all companies as likely equal. Given that the adversary is very limited this is a reasonable assumption derived from the *maximum entropy principle*[8]

[8]This principle states that in the face of "partial information we must use the probability distribution which has the maximum entropy, subject to whatever is known". [10]. Here it means the uniform distribution.

Table 1.2. Summary by city possible cities of the selected transaction

City	*Value USD*	*kg*
SAO GONCALO DO AMARANTE	6,024,745	71,944,247
OURO BRANCO	9,002,981	71,738,400
SERRA	21,710,563	241,106,000
IPATINGA	8,333,965	77,109,186
Total	**45,072,254**	**461,897,833**

that we will adopt for lack of better information.

1.2.5 The attack

From what we have presented so far the adversary does not have much chance of being successful. However, during the attack, they will be able to access different datasets and gain new knowledge.

To increase their chance of success the adversary will break the task of linking the datasets they already have into two steps. In the first step, they will try to guess the city of the underlying business of the selected transaction in the De-identified transactions. In the second step, they will try to guess the underlying business' identity among those businesses within the chosen city (the business' city is already present in the Importers dataset). Of course, just breaking the task into two steps does not change the adversary's chance of success. However, during the attack, the adversary will have access to an auxiliary dataset that will increase their probability of getting it right. This auxiliary dataset is the Summary by city published by the Ministry of Economy.

In our example, the transaction presented in Table 1.1 amounts to more than 4.9 million USD and 39 thousand tonnes. So any city whose totals are less than this could be discarded and with it the companies based on their territories. Unfortunately, in our example, that is not the case. Table 1.2 shows that all[9] cities have totals that are above the values of the target transaction.

But the adversary can go even further because they have all[10] the transactions made (Table 1.3, the target is underlined) and they know that there is at least one valid partition of these transactions between the cities that also produces the same sums observed in Table 1.2 (notice that the totals on both tables are the same, ignoring

[9]Some readers might find it strange that Table 1.2 only presents 4 cities, and ask if this is a real example or a made up one. This is a real example with real data. There are only 4 cities because Summary by city dataset was filtered with a simple key lookup. How this filter is done will be explained on Chapter 4

[10]All transactions within a filter that will be better explained in Section 4.1

Table 1.3. All de-identified transactions (filtered)

NR ORDEM	*Value USD*	*kg*
172800001	4,924,259.04	39,783,000
172900001	4,078,722.43	31,955,400
228300001	2,272,290.64	29,293,252
229000001	3,752,453.61	42,650,995
256100001	3,496,134.03	49,503,000
256300001	6,508,740.07	77,000,000
257300001	3,183,628.58	31,660,000
257500001	2,319,861.17	22,443,000
316500001	4,660,922.94	44,017,851
317100001	3,673,042.49	33,091,335
94200001	6,202,199.17	60,500,000
Total	**45,072,254.17**	**461,897,833**

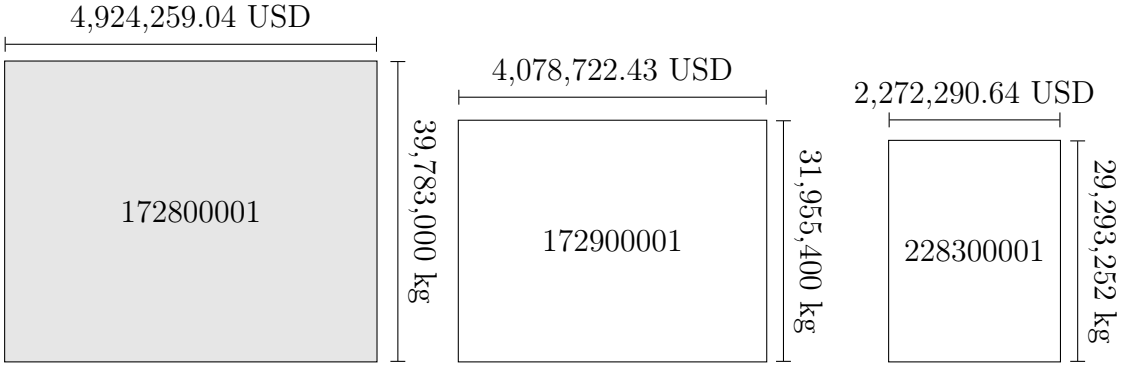
the decimal places). If there is just one valid partition then they can be sure that they have correctly identified all the transactions destinations. If so, this will greatly reduce the number of possible importers to only those importers whose city is the one that produces the valid partition.

The algorithm used to find these valid partitions is fully explained in Chapter 3, a graphical intuition is presented below.

1.2.5.1 Graphical intuition of the algorithm

The records in Table 1.3 will be represented as "packages" as shown in Figure 1.2. These packages have their width proportional to the transaction value, and height proportional to their weight. The gray one is the adversary's target.

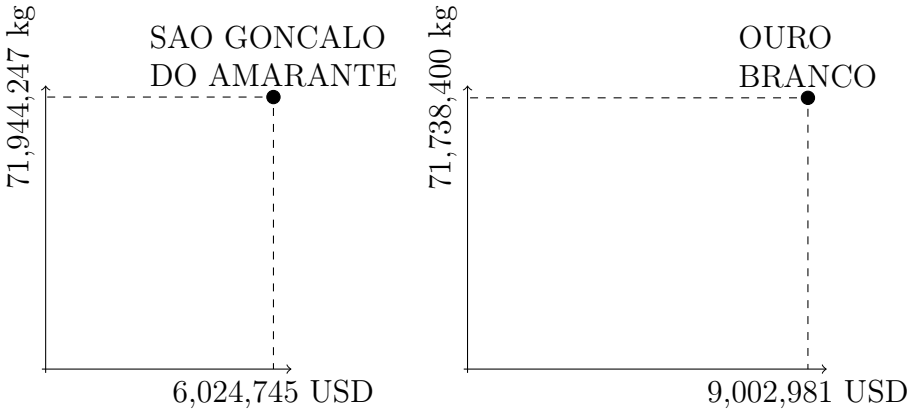
Figure 1.2. Graphical representations of some de-identified transactions



The totals by cities in Table 1.2 will be represented as target coordinates in the

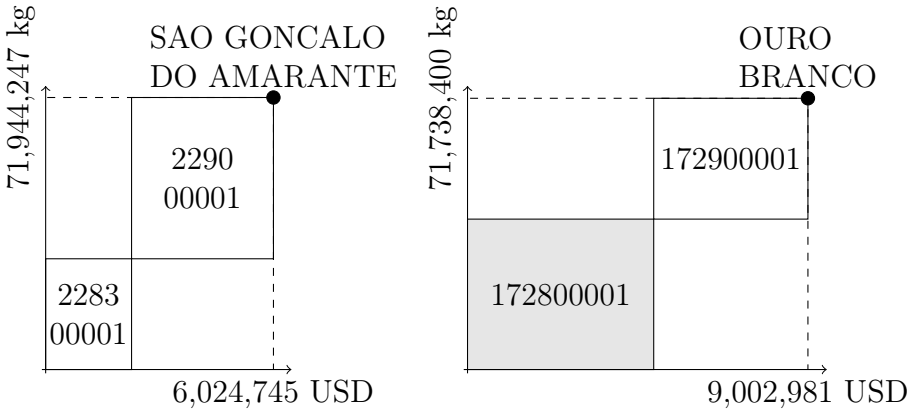
plane, as shown in Figure 1.3. The x coordinate of these targets is positioned at the total value USD, and the y coordinate is at the total weight.

Figure 1.3. Graphical representation of some summaries by cities



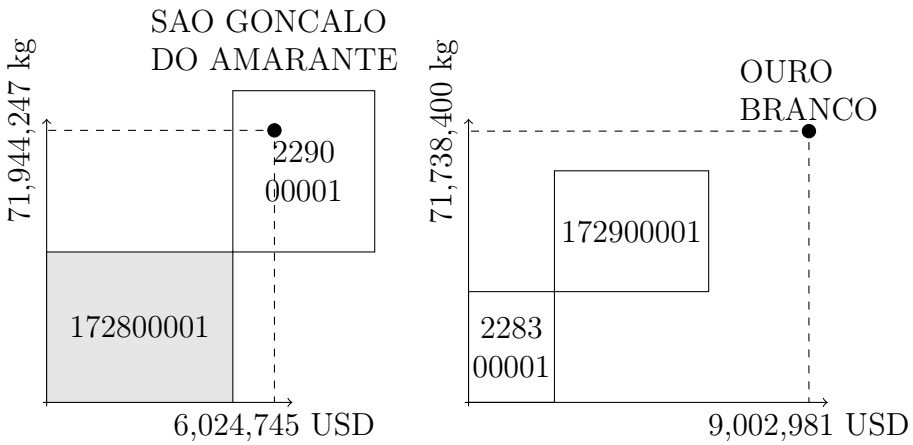
From this setup, the adversary will try to fit all the "packages" inside the planes without violating the boundaries shown by the dashed lines. The packages must be stacked by their corners as shown in Figures 1.4 and 1.5. This arrangement guarantees that in a valid solution the sum of packages' values will be equal to the total observed in the city represented by the dot. Figure 1.4 shows a valid solution and Figure 1.5 an invalid. It is easy to see that in the valid one, the sums of the packages 172800001 and 172900001 are equal to the totals for the city of Ouro Branco. If we swap the cities of packages 172800001 and 228900001 then their sum will be different from the city totals.

Figure 1.4. Graphical representation of a valid solution



Moreover, it is possible to test if there is any other valid allocation in which package 172800001 is placed in a city other than OURO BRANCO. In fact, for this

Figure 1.5. Graphical representation of an invalid solution



example, there is not. This is conclusive evidence that this package must have been purchased by a company based in OURO BRANCO.

1.2.6 Posterior success

The *posterior success* is calculated in the same way as the *prior*, but considering the new knowledge acquired during the attack: that the target package 172800001 went to a company based in OURO BRANCO.

By consulting the Importers dataset the adversary will find only one company in this city.[11] Hence, their chance of success, in this case, is 100%. We can say that the adversary's chance of success is deterministic and there's no plausible deniability, i.e. the importer cannot deny it was him. There isn't the slightest chance it was someone else, and for that we say that the attack was *deterministically* successful.

We can calculate the adversary's success chance in each phase of the attack as shown in Table 1.4. After having access to the Summary by city (phase 2) the adversary will know that the target transaction (172800001) could have gone to one of the four cities shown in Table 1.2. By looking in the Importers dataset they will know that there are 106 registered importers at SERRA, 8 at SAO GONCALO DO AMARANTE, 6 at IPATINGA, and 1 in OURO BRANCO. This reduces the number of possible importers from the initial 18,430 to 121 and increases the adversary's success chance by a factor of 152, this factor[12] is proportional to the *information leakage*.

Another way of quantifying the attack's success is by measuring the value, in US dollars, that is at risk of re-identification. For this example the transaction 172800001

[11] Whose name we will omit here.

[12] $\frac{1/121}{1/18430} \approx 152$

amounted to more than 4.9 million USD. Multiplying this value by the chance the adversary has in each of the phases we get an estimate of the *Value at risk* presented in Table 1.4.

Table 1.4. Adversary's of success by phase

	Phase	*Success odds and chance*	*Chance factor increase*	*Value at risk (USD)*
1	Prior	1 in 18,430 ($\approx 0.005\%$)	-	267
2	Summary by city	1 in 121 ($\approx 0.8\%$)	152x	40.7 k
3	Algorithm (Posterior)	1 in 1 (100%)	121x	4.9 mi

One last thing to note is that, the result obtained with the algorithm is qualitatively different from the phase 2. The 100% success chance is deterministic, it changes the knowledge from "possible" to "certain".

1.3 Fiscal secrecy violation

As has been demonstrated by the example above. After the attack, the adversary knows the following facts with certainty about the uniquely identified importer based in the city of OURO BRANCO:

1. It has purchased 39,783 tonnes of a specific type of coal.
2. That purchase cost 4,924,269.04 USD, so the average price is approximately 124 USD/ton.
3. The supplier was "Warrior Blue Creek LV".[13]

The Brazilian legal framework regarding the publication of foreign trade statistics will be presented more in-depth in Section 5.2. Nonetheless, for motivational purposes, it is sufficient to quote here Article 2$^{\text{nd}}$ of Ordinance 2,344 of March 24$^{\text{th}}$, 2011[14] which clearly states that

> Are subject to fiscal secrecy information regarding the economic or financial situation of the liable business (...) such as: (...) which can reveal (...) suppliers, (...) volumes or values of purchases and sales.[15]

[13] Mining company based in Alabama USA.
[14] Portaria RFB Nº 2.344, de 24 de março de 2011
[15] Free translation.

In our interpretation, the information disclosed by the RFB and Ministry of Economy can reveal suppliers, volumes and values purchased by the liable importers and as so should be subject to fiscal secrecy.

Chapter 2

Quantitative Information Flow

This chapter formally presents Quantitative Information Flow concepts used to describe the elements involved in the attack and quantify the information leakage. [2]

2.1 Basic Concepts

- ***Sensitive information***: what is worthy of protection [13], this may be law-enforced but it is not a necessity in terms of general privacy studies. In our concrete case, as will be discussed in section 5.2, the Brazilian legislation enforces the protection of data that can reveal the economic state of individual businesses such as volumes sold and purchased, commercial relationships, suppliers and customers.

- ***Adversary***: is "the person or entity from which *sensitive information* must be protected" [13]. Although the everyday use of this term can be associated with "bad" intentions, in our context they do not matter. For our use case the adversary could be a market analyst that is unknowingly causing harm by accessing the information or a hacker that sells private business data for companies that want to obtain illegal advantages in the market. All that matters is that an adversary is an entity not intended to infer sensitive information.

- ***Plausible deniability***: in our context is the ability to deny, with credibility, that some inferred company is not the underlying business behind a transaction. This ability is critical in negotiation. Imagine that one company X wants to renegotiate its contract with supplier Y. Now imagine that X knows that another company Z buys from Y at a price of 100 USD. In the negotiation the supplier Y could claim that they are unable to practice below 110 USD. However, if X

knows, without plausible deniability, that they in fact practice such prices in the market, then X will have an advantage in the negotiation by knowing that Y is bluffing.

2.2 Disclosure control

Disclosure control is the research area that aims to "guarantee that statistical patterns are revealed while the *sensitive information* ... is kept safe". [13]

Statistical publications are those in which the data gathered is only disclosed in summaries (like sum, count, mean) by groups. The idea that these kind of publications might not be sufficient to preserve individuals' privacy is not new. Data publishers usually rely on *disclosure control* techniques to avoid unintended private data leakage [1, 8].The effectiveness of these techniques can be measured within the *Quantitative Information Flow* theory. This measurement is made possible through (what can be thought as) a game in which an *adversary* will query the released datasets and try to reveal the *sensitive information*. How much knowledge about the *sensitive information* the adversary gains during the attack is its success measure.

A very effective way of limiting the adversary's chance of success is to simply not publish anything. Nonetheless, statistical publications are done with some legitimate goals in mind, and these publications' usefulness is related to how reachable these legitimate goals are. Not publishing the datasets can be very effective to protect the *sensitive information*, but it also heavily undermines its statistical usefulness.

Disclosure control methods try to balance this trade-off between usefulness and privacy. The ones implemented by RFB and the Ministry of Economy in the publication of foreign trade statistics are:

- ***De-identification***: by which the fields that could directly identify the importer, such as the company's name or id, are stripped from the transactions.
- ***Pseudonymization***: where each transaction is assigned a unique, artificially created identification code (NR ORDEM) for each one of the transactions. This number covers up the "official" code used internally by the government called DI (Import Declaration number).
- ***Generalization***: Totals by city are only disclosed using a more general class of the Harmonized System of products classification (HS4).[1]

[1]See the paragraph about NCM in Section 4.1.

- ***Suppression***: Whenever there are less than 4 importers behind the transactions in a given month in a given NCM[1] the RFB will suppress the information of all the transactions within that group.[2]

It is worth noting that there are disclosure control techniques within the *Differential Privacy* [7] framework that are known to be robust against the type of attack demonstrated here. They are becoming more and more used, for instance, the United States Census Bureau will adopt such techniques in the 2020 Census.[3]

2.3 Related studies

As has been demonstrated in the Netflix prize [12] and AOL[4] data leaks, techniques such as the ones used by the RFB and Ministry of Economy are weak against database reconstruction attacks. What happens is that, even when de-identification, pseudonymization and generalization are used, the data published usually enables the construction of a set of constraints which are sufficiently restrictive as to agree with just one arrangement of the microdata. [8]

An Australian report [6] shows how database reconstruction attacks can be used to re-identify individuals in the Australian health records public datasets. These Australian datasets contain billing information regarding procedures undertaken by individuals and paid by the government and their insurers. These billing records are de-identified and pseudonymized, nonetheless, the authors argue that around 900,000 individuals have unique sums regarding their paid expenditures and that the fact that they are unique can aid in the re-identification. Of course, those expenditure sums were calculated because the records were pseudonymized, *i.e.* there was a field used for grouping the transactions. We think that the algorithm presented here shows that even if the artificially created ids are stripped from the public datasets the same sums could be reconstructed.

There is also a Brazilian study [13] where the author presents how the national educational census database, released by Brazilian authorities, can be used to reveal if a student has a physical disability with 99.69% probability of success. The procedure we present here shows that even if the Brazilian authorities started to only release certain sensitive attributes in summaries, *i.e.* just disclose the total number of students

[2] §2º Portaria RFB nº 361, de 14 de março de 2016.

[3] https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance.html

[4] https://www.nytimes.com/2006/08/09/technology/09aol.html

with disabilities per school, re-identification would be possible, although the chance of success could be lower.

2.4 Adversary models

In the example of Section 1.2 the adversary adopted what can be called a *Journalist model* [13] which means that their goal was to find out if the re-identification could be done. In Chapter 5 we will use an adversary model with a *Marketer* approach, that is, the adversary will try to re-identify as many transactions as possible so we can estimate the attack's potential reach.

In both approaches, the adversary will be restricted to use just some minimal **public data** and will **lack any kind of expertise** in foreign exchange trades. This adversary might not reflect what is done by real world adversaries, but it provides a lower bound on the leakage that could be caused by even stronger ones. As we shall see, even such a modest attacker will cause unacceptable harm. This will allow us to highlight the following aspects for each one of these arbitrary decisions:

1. ***Public data***: by limiting the adversary knowledge to just public data we are finding the lower bound of their posterior chance of success. Public data is available to all and those who have access to any private data that aids this type of attack will only be more successful.

2. ***Lack of expertise***: to become an expert takes time and energy which limits the number of potential adversaries. By not requiring expertise we are maximizing the number of potential adversaries.[5] This also maximizes the number of potential targets as experts are limited to just one area of expertise (*i.e.* an expert in steel imports might not know anything about electronics commerce).

2.5 Information leakage

The following definitions are necessary to formally measure the information leakage derived from the attack. These definitions are based on the work by Alvim, Chatzikokolakis, McIver, Morgan, Palamidessi, and Smith [2], but are translated here to our concrete case.

[5]It could be argued that the adversary here requires expertise in computer science to implement the algorithm, however, the adversary just has to have access to the algorithm, that can be implemented once and then copied as many times as needed and is probably cheaper than a consultation with an expert.

Definition 1 (Secret)**.** Every transaction has the secret, called I (importer that made the transaction), the possible values for this secret are within the finite set $\mathcal{I}$ (possible importers). For a set of n transactions the secret is a n-tuple whose elements are within $\mathcal{I}$ and represent the importers behind each transaction.

Definition 2 (Prior knowledge)**.** The adversary prior knowledge about I is given by a probability distribution π on $\mathcal{I}$ that specifies the probability π_i of each possible value i of I.

In our specific case, the adversary's prior knowledge is given by the uniform probability distribution over all possible importers, $\mathcal{I}$ (which are obtained from the Importers dataset). Formally, this distribution is the mapping: $\pi : (i) \mapsto 1/|\mathcal{I}|^n$ where n is number of transactions being targeted at once.

Definition 3 (Measure of prior success)**.** The measure of prior success is a function that takes as input the prior knowledge probability distribution π and returns a real number that indicates the adversary's success.

We used two measures of prior success. 1) the maximum probability of correctly guessing the importers, *i.e.* chance of success ($\max(\pi) = 1/|\mathcal{I}|^n$) and; 2) the value at risk, which is the maximum probability of correctly guessing the importers times the target transactions' values ($\sum 1/|\mathcal{I}|^n \times$ value in USD).

Definition 4 (Channel)**.** The channel is a data release that gives the adversary more information about the secret. The channel models the attack itself and is defined by a triple ($\mathcal{I}$, $\mathcal{O}$, Ch) where $\mathcal{I}$ is the set of possible inputs, $\mathcal{O}$ is the set of possible outputs and Ch is a matrix of size $|\mathcal{I}| \times |\mathcal{O}|$. The elements of Ch are denoted as $Ch_{i,o}$ and represent a value between 0 to 1 which is the conditional probability of observing output o given the input i, the rows of Ch must sum to 1.

For a more concrete explanation: suppose there are only 3 packages as shown in Table 2.1 and the adversary wants to know the importer of package b.

Table 2.1. Example transactions

id	*Value USD*
a	1
b	2
c	3

Also, suppose that the set $\mathcal{I}$ has only 2 elements: *ob* which is a company from Ouro Branco and *sg* which is company from São Gonçalo do Amarante. In this case, the adversary's prior chance of success is $1/2$ and the value at risk is 1 USD.

If these transactions are all attributed to *ob* the input is $I = (ob, ob, ob)$. If we define the output as the total value in USD imported by the cities of Ouro Branco and São Gonçalo do Amarante, respectively, we can calculate it as $O = (6, 0)$ for this specific input, which means that the total for the city of Ouro Branco is 6 USD and 0 for the city of São Gonçalo do Amarante. If we do this for all possible inputs we get the matrix Ch below.

Ch	(6, 0)	(3, 3)	(4, 2)	(5, 1)	(1, 5)	(2, 4)	(0, 6)
ob, ob, ob	1	0	0	0	0	0	0
ob, ob, sg	0	1	0	0	0	0	0
ob, sg, ob	0	0	1	0	0	0	0
sg, ob, ob	0	0	0	1	0	0	0
ob, sg, sg	0	0	0	0	1	0	0
sg, ob, sg	0	0	0	0	0	1	0
sg, sg, ob	0	1	0	0	0	0	0
sg, sg, sg	0	0	0	0	0	0	1

This channel shows that the probability of observing the totals (3, 3) for Ouro Branco and São Gonçalo do Amarante is 1 when the transactions a and b are attributed to *ob* and transaction c is attributed to sg. This channel is deterministic, all its elements are either 0 or 1, each input determines the sum expected for the cities.

Definition 5 (Joint distribution)**.** The joint distribution on $\mathcal{I} \times \mathcal{O}$ is determined by the prior distribution π and the channel Ch and is defined as $\Pi_{i,o} = \pi_i Ch_{i,o} := p(i, o)$.

For the example above, this means multiplying each element of the matrix Ch by $1/2^3 = 1/8$ to produce the matrix Π. The row with the sum of each column of this matrix is the $\mathcal{O}$-marginal distribution p_O.

Π	(6, 0)	(3, 3)	(4, 2)	(5, 1)	(1, 5)	(2, 4)	(0, 6)
ob, ob, ob	$1/8$	0	0	0	0	0	0
ob, ob, sg	0	$1/8$	0	0	0	0	0
ob, sg, ob	0	0	$1/8$	0	0	0	0
sg, ob, ob	0	0	0	$1/8$	0	0	0
ob, sg, sg	0	0	0	0	$1/8$	0	0
sg, ob, sg	0	0	0	0	0	$1/8$	0
sg, sg, ob	0	$1/8$	0	0	0	0	0
sg, sg, sg	0	0	0	0	0	0	$1/8$
p_O	$1/8$	$2/8$	$1/8$	$1/8$	$1/8$	$1/8$	$1/8$

Lastly, if we divide each column by its sum (marginal probabilities) then each column will represent the posterior distributions, *i.e.* the distributions over inputs given the observed output denoted as $p_{I|o}$. For the example, $p_{I|(6,0)} = (1, 0, 0, 0, 0, 0, 0, 0)$ and $p_{I|(3,3)} = (0, 1/2, 0, 0, 0, 0, 1/2, 0)$. This whole result, from the prior to these posteriors distributions can be encapsulated in what is called *hyper-distributions*, which are distributions over distributions.

Definition 6 (Hyper-distributions)**.** Consider a prior π, an input space $\mathcal{I}$, an output space $\mathcal{O}$ and a channel Ch. This channel determines a joint distribution Π that has marginal distribution p_O and for each o a corresponding posterior distribution $p_{I|o}$ on the inputs. The hyper-distribution considers p_O to be a distribution over the posteriors (the normalized columns of Π) instead of the labels on the top.

In our example the hyper-distribution assigns the probability distribution ($1/8$, $2/8$, $1/8$, $1/8$, $1/8$, $1/8$, $1/8$, $1/8$) to the posteriors ($p_{I|(6,0)}$, $p_{I|(3,3)}$, $p_{I|(4,2)}$, $p_{I|(5,1)}$, $p_{I|(1,5)}$, $p_{I|(2,4)}$, $p_{I|(0,6)}$).

Definition 7 (Posterior knowledge)**.** The adversary's posterior knowledge is a Bayesian update of the prior knowledge. Once the output o is revealed, the adversary will know which posterior to get from the hyper-distribution defined by the channel.

For example, if the output revealed is (6, 0) then the adversary's posterior knowledge is the distribution $p_{I|(6,0)} = (1, 0, 0, 0, 0, 0, 0, 0)$, but if the output is (3,3) the posterior knowledge is $p_{I|(3,3)} = (0, 1/2, 0, 0, 0, 0, 1/2, 0)$.

Definition 8 (Measure of posterior success)**.** The measure of posterior success maps the posterior knowledge to a real number that indicates the adversary's success.

The same functions applied in the prior knowledge distribution (π) are applied over the posterior knowledge distribution ($p_{I|o}$), the chance of success and the value at risk. So if the observed output is (6, 0) then the adversary's chance of posterior success is $\max p_{I|(6,0)} = 1$ and the value at risk is 2 USD. If the output is instead (3, 3) than the chance of success is $\max p_{I|(3,3)} = 1/2$ and the value at risk is 1.

Definition 9 (Leakage)**.** Leakage is a comparison between posterior and prior success measures, indicating how much the attack execution increases the adversary's knowledge about the secret.

In this study two leakage measures are used:

1. **Chance of success increase**: $\frac{\text{Posterior chance}}{\text{Prior chance}}$.

2. **Value leaked**: (Posterior value at risk) - (Prior value at risk)

In our example here, if the output is (6, 0), the chance of success increase is $1/1/2 = 2\times$ (two times) and the value leaked is $1\times$ USD. If the output is (3, 3), the chance of success increase is 1 (no increase) and the value leaked is 0.

Chapter 3

The algorithm

This chapter presents the Database reconstruction attack used by the adversary to re-identify the importer behind a specific transaction. This algorithm will be called the ATTACK ALGORITHM. Before presenting the ATTACK ALGORITHM, we will present an optimization problem that is at its core. This optimization problem is closely related to the well-known GENERALIZED ASSIGNMENT, we are going to call our particular problem the PACKAGE ALLOCATION problem. Also, when written as a decision problem the PACKAGE ALLOCATION is closely related to the SUBSET SUM, NUMBER PARTITION, and BIN PACKING problems.

3.1 The PACKAGE ALLOCATION problem

In this problem there is a set of packages with different weights that must be allocated in bins with different capacities. There is also a specific package that is the tracked one. The package's weights sum is equal to the sum of the bin's capacities (within a rounding error) and every package has to be placed in one and only one bin, so each bin can have more than one package assigned to it. The problem is to decide if it is possible to construct a valid allocation of the packages (which does not overflow the bin's capacities, beyond the rounding error) and **includes the specific tracked package in the specified bin**.

Instance: $\mathfrak{I} = (P, W, B, C, p, b, \nu, \epsilon)$ is the input tuple for the PACKAGE ALLOCATION problem, where:

1. $P = \{p_1, p_2, ..., p_n\}$. Set of n packages.
2. $W = \{w_1, w_2, ..., w_n\}$. Set of n positive weights for each package in P.

3. $B = \{b_1, b_2, ..., b_m\}$. Set of m bins.
4. $C = \{c_1, c_2, ..., c_m\}$. Set of m positive capacities for each bin in B.
5. $p \in P$. The tracked package.
6. $b \in B$. The destination bin of the tracked package.
7. $\nu \in \mathbb{Q}^+$. A non-negative overflow tolerance.
8. $\epsilon \in \mathbb{Q}^+$. A non-negative maximum allowed global rounding error.

Properties: Every instance of this problem has the following properties.

1. The sum of packages weights is less than or equal to the sum of bins capacities, within a positive global rounding error $\sum(W) - \sum(C) \leq \epsilon$
2. The allocation matrix X maps each package to a bin. An allocation is valid if every package is placed in one and only one bin and there is no overflow beyond the rounding tolerance ν. Formally:
   a) $X : P \times B \to \{0, 1\}$ and $X(i, j) \mapsto x_{ij}$ for each $i \in P$ and $j \in B$, where $x_{ij} = 1$ if package i goes into bin j and $x_{ij} = 0$ otherwise.
   b) $\forall i \in P, \sum_{j \in B} x_{ij} = 1$, (every package is placed in one and only one bin).
   c) $\forall j \in B, \sum_{i=1}^{n} w_i x_{ij} - c_j \leq \nu$, (overflow is lower or equal than the tolerance)

3.2 Optimization problem

GENERALIZED ASSIGNMENT Problem The GENERALIZED ASSIGNMENT Problem (GAP) [5] is a combinatorial optimization problem which has as special cases the ASSIGNMENT and KNAPSACK problems.

In the GAP, there is a set of agents with different budgets and a set of tasks that can be assigned to any agent. For every agent, each task produces a cost and a profit that can vary among the agents. The problem is to find an assignment that does not exceed any of the agents' budgets and maximizes the profits.

GAP Instances:

1. Set of n tasks. $P = \{p_1, p_2, ..., p_n\}$.
2. Set of m agents. $B = \{b_1, b_2, ..., b_m\}$.

3. Budgets of each agent. $C = \{c_1, c_2, ..., c_m\}$.

4. for each agent $b_j \in B$, each task $p_i \in P$ has a positive profit r_{ij} and a positive cost w_{ij}

   The variable $x_{ij} = \{0, 1\}$ indicates if the task i is assigned to agent j.

GAP optimization:

$$\text{maximize} \sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij}x_{ij}. \tag{3.1}$$

$$\text{subject to} \sum_{i=1}^{n} w_{ij}x_{ij} \leq c_j \qquad j = 1, \ldots, m; \tag{3.2}$$

$$\sum_{j=1}^{m} x_{ij} = 1 \qquad i = 1, \ldots, n; \tag{3.3}$$

$$x_{ij} \in \{0, 1\} \qquad i = 1, \ldots, m \quad j = 1, \ldots, n; \tag{3.4}$$

The PACKAGE ALLOCATION can be modeled as a GENERALIZED ASSIGNMENT where the budgets have been already paid and there is no profit, so the goal is to consume as much as possible from the budgets without consuming beyond an overflow tolerance (ν). The cost of each task is the same across all agents and one specific task must be done by a designated agent.

PACKAGE ALLOCATION optimization problem:

$$\text{minimize } y. \tag{3.5}$$

$$\text{subject to} \sum_{i=1}^{n} w_i x_{ij} - c_j \leq y \qquad j = 1, \ldots, m; \tag{3.6}$$

$$\sum_{j=1}^{m} x_{ij} = 1 \qquad i = 1, \ldots, n; \tag{3.7}$$

$$x_{ij} \in \{0, 1\} \qquad i = 1, \ldots, m \quad j = 1, \ldots, n; \tag{3.8}$$

$$x_{pb} = 1 \tag{3.9}$$

$$y \leq \nu \tag{3.10}$$

$$y \geq 0 \tag{3.11}$$

If there is a feasible solution for this optimization problem then there is a valid allocation that contains package p in bin b. This specification is equivalent to the PACKAGE ALLOCATION decision problem that will be presented below.

For practical purposes, in the actual implementation of the attack we are going to use a variant of the above. The constraint (3.9) will be replaced by the set of constraints $\forall b \in \hat{B}, x_{pb} = 0$, where $\hat{B}$ is a subset of B.[1] This implementation of the PACKAGE ALLOCATION problem will have as input the tuple $\mathfrak{I} = (P, W, B, C, p, \hat{B}, \nu)$.

One last remark, in this description of the PACKAGE ALLOCATION problem the packages and bins have only one attribute: weight and capacity respectively. However, it is simple to modify the algorithm to any finite number of attributes.[2] In the example from Section 1.2 we used two attributes (value and weight), and in Phase of 2 Section 5.1 we used 5 attributes (value, weight, quantity, freight and insurance).

3.3 Decision problem and NP-Completeness

In this section we formalize the PACKAGE ALLOCATION as a decision problem to demonstrate that it is NP-Complete and also to better understand its complexity.

Decision problem: Given the instance $\mathfrak{I}$ is there a valid allocation matrix X that includes package p in bin b?

If PACKAGE ALLOCATION returns TRUE then we know that the package p fits in bin b. If we fix p and iterate over all $b \in B$ and end up with only one possible valid allocation for p, then we can be sure that package p can only be in bin b.

Theorem 1. PACKAGE ALLOCATION is in NP

Proof. X can be verified by:

1. For each bin: assert the overflow is lower than the tolerance ($\forall j \in B, \sum_{i=1}^{n} w_i x_{ij} - c_j \leq \nu$).
2. For each package: check that it is placed in one and only one bin ($\forall i \in P, \sum_{j \in B} x_{ij} = 1$).

Both checks can be done in polynomial time in the size of the input. This is enough to place the PACKAGE ALLOCATION in the **NP** class [3]. □

[1] Thus, the decision problem is: Is there a valid allocation where the package p is not placed in any $b \in \hat{B}$?

[2] Formally, replace the weight function $W : P \to \mathbb{Q}$ by an attribute function which maps P to $\mathbb{Q}^d$, where d is the number of attributes, and analogously for $C : B \to \mathbb{Q}^d$

3.3.1 Related problems

The PACKAGE ALLOCATION problem is closely related to some well known NP-complete and combinatorial optimization problems like the KNAPSACK, SUBSET SUM, BIN PACKING and GENERALIZED ASSIGNMENT[11, 9]. Using the SUBSET SUM it is possible to prove that our problem is NP-complete as follows.

SUBSET SUM problem

Instance:

1. A set of natural numbers: $N = \{n_1, n_2, ..., n_k\}$
2. The natural number: S

Decision problem: Does any subset of N sum to exactly S?

Theorem 2. PACKAGE ALLOCATION is NP-complete

Proof. We show a polynomial reduction of the SUBSET SUM problem to the PACKAGE ALLOCATION.

Given an instance of the SUBSET SUM problem:

1. For each $w_i \in W$, create the following instance $\mathfrak{I} = (P, W, B, C, p, b, \nu, \epsilon)$ of the PACKAGE ALLOCATION problem:
   a) $P = \{1, 2, ...k\}$. The set P contains the indexes of the k natural numbers in in the set N.
   b) $W = N$. The packages weights are the set of natural numbers.
   c) $B = \{1, 2\}$. Create 2 bins.
   d) $\epsilon = \nu = 0$. No rounding errors or overflows.
   e) $C = \{S, \sum(W) - S\}$. The first bin has a capacity equal to S. The second has capacity so as to satisfy $\sum(C) = \sum(W)$.
   f) $p = i$ and $b = 1$.
   g) Is there a valid allocation where the package with weight w_i is placed in the first bin?
2. If any of the created instances of the PACKAGE ALLOCATION problem returns TRUE, then return TRUE, otherwise, FALSE.

If any instance of the PACKAGE ALLOCATION returns TRUE, then there is a subset of W which sums up to exactly S.

If no instance of the PACKAGE ALLOCATION returns TRUE, that means that for all $w_i \in W$, no subset of W which includes w_i sums to S.

This reduction is polynomial in time and space and thus PACKAGE ALLOCATION is **NP-complete**. □

3.3.2 PACKAGE ALLOCATION complexity

The upper bound for the PACKAGE ALLOCATION problem computational complexity can be derived from a brute force approach. If we iterate over all possible allocations, that each package can be in any of the bins, this yields m^n (number of bins to the power of number of packages) possible allocations. Throughout this text we will use following "casual" definition of complexity for our algorithm: complexity$(n, m) = log_{10} m^n$.

Even small instances of the PACKAGE ALLOCATION problem with less than 15 packages and 30 bins can have as many as one sextillion (10^{21}) possible states. Nonetheless, integer programming solvers use a branch-and-bound (or branch-and-cut) strategy that enables the search for an optimal solution without requiring to test every possible allocation.

3.4 The algorithm used in the attack

In our attack, the adversary wants to know all the possible cities the target transaction (p) might fit. This is done by first getting one valid allocation without restricting the bin where the package p can go or the overflow tolerance ν.[3] After this first iteration, we retrieve the bin b where the package p ended and add it as a restriction ($x_{pb} = 0$) for future runs of the PACKAGE ALLOCATION, thus finding equivalent arrangements for the packages. The procedure then loops and adds these restrictions until it becomes infeasible. The set $R \subseteq B$ containing the possible bins for the package p is returned at the end. Algorithm 1 formally defines the attack. The auxiliary functions used are:

1. **PackageAllocation**: receives the tuple $\mathfrak{I} = (P, W, B, C, p, \hat{B}, \nu)$; solves the PACKAGE ALLOCATION optimization problem; returns the matrix X.

2. **GetBin**: receives p and X; returns b where $x_{pb} = 1$.

[3]The adversary knows that the summary statistics are produced from the micro data, so they can be sure that at least one valid allocation exists.

3. **GetMaxOverflow**: receives X, W and C; calculates the overflow for each bin ($\sum x_{ij}w_i - c_j$); returns the maximum overflow.

4. **IsFeasible**: asserts that the solution returned from the **PackageAllocation** was feasible, *i.e.* all constraints were met.

5. **AddBin**: adds the bin b to the set R.

Algorithm 1: ATTACK ALGORITHM

input : P, W, B, C, p, ϵ

output: R: a subset of B, that contains the bins where the allocation of package p is valid.

```
1  begin
2      R ← ∅;
3      if ∑(W) − ∑(C) > ϵ then return R ; // Invalid instance
4      ν = ∞;
5      X = PackageAllocation(P, W, B, C, p, R, ν);
6      b = GetBin(p, X);
7      ν = GetMaxOverflow(X, W, C);
8      while IsFeasible(X, P, W, B, C, p, R, ν) do
9          AddBin(b, R);
10         X = PackageAllocation(P, W, B, C, p, R, ν);
11         b = GetBin(p, X);
12     end
13     return R
14 end
```

This algorithm was implemented in Python 3.7 with the help of the Python optimization package Pyomo 6.0.1 and Pandas 1.3.1 analysis tool. The solver used was IBM® ILOG® CPLEX® version 12.9.0.0. The whole attack documented in Chapter 5 was run in a personal computer Intel® Core™ i7-8550U CPU @ 1.80GHz with 32GB of ram memory and it took almost 62 hours of active processing time to complete.

Chapter 4

Implementation details

This chapter presents the implementation details of the attack algorithm so the findings here can be checked and reproduced. The following sections will also provide details about foreign exchange trade statistics, the datasets used and the data manipulation that must be done before applying the algorithm described in Section 3.4.

4.1 Foreign trade statistics methodology

International trade statistics serve the needs of many users including supranational and international organizations. However, these statistics are collected, organized, and published by local governments in what, in the past, gave rise to many different methodologies. The United Nations has, since its creation, devoted efforts to unifying and standardizing these different methodologies for greater comparability between national statistics. [15] Brazil is one of the countries that follow the United Nations guidance and so its statistics are highly compatible with other countries. Below we briefly introduce some concepts used in international commerce statistics that are important to understand how the reconstruction attack is done.

DI - Declaration of Import is a document that every importer has to submit to the Siscomex (Brazilian International Commerce System) whenever an import transaction occurs.[1] The information submitted is divided into two groups: 1) regarding the import **transaction** and 2) regarding the **goods** being imported. The 1st group will be called the *DI header* throughout this text and the 2nd group will be called the *items details*.

[1] https://receita.economia.gov.br/orientacao/aduaneira/manuais/despacho-de-importacao/topicos-1/conceitos-e-definicoes/tipos-de-declaracao-de-importacao/declaracao-de-importacao-di

The data filled in this DI is the same that later will be present in the datasets discussed in Section 4.1.1. Almost all datasets will only present information from the DI's header, only the De-identified transactions dataset will also contain data from the items details.

NCM - Common Mercosur Nomenclature is a regional product categorization system used in the Mercosur Economic Region since 1995.[2] It was derived from the Harmonized Commodity Description and Coding System (HS), which was designed and is maintained by the World Customs Organization. The HS is comprised of more than 5 thousand commodity groups, that are identified by a hierarchical six-digit code.[3] These six-digit codes are arranged in a logical structure and supported by a well-defined set of rules that enable a uniform classification of any merchandise into one of those groups.

For example, HS code 080510 identifies *fresh oranges* as the imported good. One key aspect of HS codes is that they are hierarchical and can be used in aggregations. The hierarchy levels are commonly labeled HS2 and HS4, the HS2 is represented by the first 2 digits in the HS code and HS4 is represented by the first 4. In the example of fresh oranges, its HS2 code is 08 which identifies *Edible fruit & nuts* and its HS4 code is 0805 *Citrus Fruit, Fresh or Dried*.

The NCM is basically an extension of the HS system that uses 8 digit codes of which the first 6 digits agree with the HS. This enables the NCM to represent a greater granularity for products categorization. The NCM is not only used for product classification but also for taxation purposes as the import tax rates are based on the NCM the good is placed in.

Incoterms - International Commercial Terms is a set of predefined commercial clauses that is widely used in international trade contracts.[4] The goal of these terms is to provide universal clarity and predictability to business that engage into international commerce. The incoterms feature abbreviations like FOB (Free on Board), EXW (Ex Works) and CIP (Carriage and Insurance Paid To) to name a few, which all have a very precise meaning regarding the responsibilities of buyers and sellers in the overall sales process.

[2] https://receita.economia.gov.br/orientacao/aduaneira/classificacao-fiscal-de-mercadorias/ncm

[3] http://www.wcoomd.org/en/topics/nomenclature/overview/what-is-the-harmonized-system.aspx

[4] https://iccwbo.org/resources-for-business/incoterms-rules/incoterms-2020/

The import Brazilian statistics are currently disclosed using their FOB value, which is the merchandise value when it was placed on board in the port of origin in the seller's country.

4.1.1 The datasets

The reconstruction attack described in Section 5.1 uses 4 different datasets that for ease of explanation will be named De-identified transactions, Summary by city, Importers, and Summary by NCM. A brief description, the source, and available fields of each of these datasets are presented below.

De-identified transactions this dataset can be downloaded from the RFB website[5] and it contains detailed information about imported goods that entered the Brazilian borders. The information is not aggregated and it does not contain direct identifiers, such as the business name or address. The available fields are detailed in Table 4.1:

Table 4.1. Available fields in the De-identified transactions dataset

Field	*Description*	*Example*
NUMERO DE ORDEM	Sequential unique number that also aids in the identification of the Declaration of Imports (DI).	150320000100001
ANOMES	Year and month when this import operation was registered.	202011
COD.NCM	DI header imported good NCM code number.	42010090 (Saddlery and harness of other materials)
PAIS DE ORIGEM	DI header imported good country of origin.	CHINA
PAIS DE AQUISICAO	DI header imported good country of acquisition.	CHINA
UNIDADE DE MEDIDA	DI header measurement unit for statistical purposes.	NET KILOGRAM
UNIDADE COMERC.	DI item Measurement unit for commercial purposes.	PIECE

[5] https://siscori.receita.fazenda.gov.br/apoiosiscori/consulta.jsf

Table 4.1 – continued from previous page

Field	*Description*	*Example*
DESCRICAO DO PRODUTO	DI item detailed description as submitted by the importer (usually in portuguese).	Bandana para animal de estimação, composição 100% algodão, sem forro. MARCA: ACCESSORI - Artigo: 13556843 (P/N: 9.IMPOR.182.000003) (P/N Fab.: BANDANA ESTRELA)
QTDE ESTATISTICA	DI header statistical quantity.	477.3
PESO LIQUIDO	DI header net weight (in kg).	477.3
VMLE DOLAR	DI header FOB value in US dollars.	8,313.44
VL FRETE DOLAR	DI header international freight in US dollars.	245.23
VL SEGURO DOLAR	DI header insurance in US dollars.	0.00
VALOR UN. PROD.DOLAR	Calculated field = TOT.UN.PROD.DOLAR / QTD COMERCIAL	0.87
QTD COMERCIAL	DI item commercialized quantity.	4,773
TOT.UN. PROD.DOLAR	DI item value in US dollars.	4,152.51
UNIDADE DESEMBARQUE	DI header port of entrance.	NOT/INFORMED
UNIDADE DESEMBARACO	RFB unit responsible for the administrative treatment of this DI.	PORTO DE SANTOS
INCOTERM	International commercial terms of the DI.	FOB
NAT. INFORMACAO	DI headers nature of the operation.	EFFECTIVE

The field NAT. INFORMACAO is one of the key fields for the attack that is not self-explanatory. It is used for filtering out transactions that, although were processed in a customs facility, are not a usual definitive import.

For instance, what we usually think of as a "normal" import, buying something from another country, has the value for this field set to EFFECTIVE. However, this is not the case for all operations. Imagine the following scenario: your company needs a very specific equipment for a one-time task. This equipment is very expensive and is not available in your country. You found a foreign company that is willing to rent the equipment for a time period. If you agree the equipment will be shipped to your location. This scenario can be characterized in the Brazilian customs as a "temporary admission" and in such case, you would benefit from a special tax regime. This operation would have its NAT. INFORMACAO set to ADMINISTRATIVE or SPECIAL.[14]

The data disclosed by the Ministry of Economy, in most cases, only contains the data for EFFECTIVE transactions[6]. The totals from De-identified transactions, Summary by city, and Summary by NCM will only match if non-effective imports are filtered out from the first dataset.

Summary by city this dataset can be queried at the COMEX STAT website[7] or entirely downloaded from the Ministry of Economy website[8]. It contains aggregated data for value and weight detailed by the fields shown in Table 4.2.

Table 4.2. Available fields in the Summary by city dataset

Field	*Description*	*Example*
CO_ANO	Year when the import was registered.	2021
CO_MES	Month when the import was registered.	01

[6]In some edge cases ADMINISTRATIVE and SPECIAL can be also considered see Section 4.2.

[7]http://comexstat.mdic.gov.br/pt/geral

[8]https://www.gov.br/produtividade-e-comercio-exterior/pt-br/assuntos/comercio-exterior/estatisticas/base-de-dados-bruta

Table 4.2 – continued from previous page

Field	*Description*	*Example*
SH4	Imported goods HS4 code number.	3002 (Human blood; animal blood prepared for therapeutic, prophylactic or diagnostic uses; antisera and other blood fractions and modified immunological products, whether or not obtained by means of biotechnological processes; vaccines, toxins, cultures of micr)
CO_PAIS	Imported goods country of origin.	249 (USA)
SG_UF_MUN	Importer federation state.	SP
CO_MUN	Importer city.	3449904 (SAO JOSE DOS CAMPOS)
KG_LIQUIDO	Imported goods total net weight in kg.	326
VL_FOB	Imported goods total value in USD.	906,723

Importers this dataset can be downloaded from the Ministry of Economy webpage[9] and it contains registration data about businesses that have conducted at least one import operation in the current year. This data is updated monthly, by adding new businesses to the list. If the adversary wants they can pinpoint the exact month a given importer started importing by tracking the additions made to the list. The available fields in this dataset are:

[9]https://www.gov.br/produtividade-e-comercio-exterior/pt-br/assuntos/comercio-exterior/estatisticas/empresas-brasileiras-exportadoras-e-importadoras

Table 4.3. Available fields in the Importers dataset

Field	*Description*	*Example*
CNPJ	Business registration number in Brazil.	165.823.950/001-18
EMPRESA	Business name.	A. A. TELECHESKI
ENDEREÇO	Business address.	RUA INGLATERRA
NÚMERO	Business address number.	160
BAIRRO	Business address neighborhood	JARDIM EUROPA
CEP	Business zip code.	87111-090
MUNICÍPIO	Business address city.	SARANDI
UF	Business federation state.	PR
CNAE PRIMÁRIA	Economic nature / industry segment of the business	4649 - Wholesale commerce of equipment and personal articles for personal and domestic usage not previously specified.
NATUREZA JURÍDICA	Juridic nature of the business	213 - Individual merchant firm

Summary by NCM this dataset can be queried at the COMEX STAT website[10] or entirely downloaded from the Ministry of Economics website.[11] It contains aggregated data for value, weight and quantity detailed by the fields shown in Table 4.4. The figures shown in this dataset are what is considered to be the official figures regarding foreign commerce for the National Accounting System.[14]

This dataset was not used in the example in Section 1.2 but will be used in the attack shown in Section 5.1. This dataset is useful for identifying which NCMs were deliberately suppressed from the De-identified transactions as the RFB does not disclose any information for an NCM if in a given month less than 4 companies imported on it. If in a given month a NCM appears in Summary by NCM but not in De-identified transactions then we know this less-than-4-companies rule was applied.

[10] http://comexstat.mdic.gov.br/pt/geral

[11] https://www.gov.br/produtividade-e-comercio-exterior/pt-br/assuntos/comercio-exterior/estatisticas/base-de-dados-bruta

Table 4.4. Available fields in the Summary by NCM dataset

Field	*Description*	*Example*
CO_ANO	Year when the import was registered.	2020
CO_MES	Month when the import was registered.	10
CO_NCM	Imported goods NCM code number.	40161090 (Oth.works of vulcan. rubber alveol. n/harden.)
CO_PAIS	Imported goods country of origin.	249 (USA)
SG_UF_NCM	Importer federation state.	MG
CO_VIA	Imported goods entrance route.	04 (AERIAL)
CO_URF	Unit of the RFB responsible for administrative treatment of the imported good.	0617700 (BELO HORIZONTE)
CO_UNID	Unit of measurement.	10 (kg)
QT_ESTAT	Imported goods total quantity in the given unit (CO_UNID).	53
KG_LIQUIDO	Imported goods total net weight in kg.	53
VL_FOB	Imported goods total value in USD.	3,236
VL_FRETE	Total international freight in USD	20
VL_SEGURO	Total international insurance in USD	2

4.2 Divergences between the datasets

The foreign commerce statistics disclosed by the RFB and Ministry of Economy have the same information source. Nonetheless, there are some methodologies differences that can make the data published by them to diverge.

1. **Numeric precision**: The RFB report values with up to 2 decimal places for currency values and 5 decimal places for weight. The Ministry of Economy rounds every value to integers.

2. **Data suppression**: The RFB does not disclose certain NCMs if less than 4 importers operated in it, in an attempt to preserve privacy. Which NCMs were

omitted is easy to find by seeing the differences between De-identified transactions and Summary by NCM.

3. **Outliers treatment**: The Ministry of Economy publications are intended for Statistical purposes and international comparability. Transactions that are caught in their internal controls might be discarded from the official publications. These internal controls and discarded transactions are not publicly available.

4. **Transactions inclusion criteria**: On April 7th, 2021 the Ministry of Economy changed its foreign commerce statistics methodology and started including some imports classified as ADMINISTRATIVE or SPECIAL in their reports. The information used for the selection criteria is not present in the datasets [14] and we are unaware of publicly available datasets that have this information.

The divergences caused by rounding values are easier to deal (they do not make it much harder to track a transaction to its city)[12]. The divergences caused by the data suppression, outliers treatment, and the different inclusion criteria do have a detrimental effect on the algorithm.

Although just the data suppression technique, used by RFB, had confidentiality as a goal, all the other data treatments applied that create these divergences act as a kind of protection against database reconstruction attacks. Where the divergences are small the linking of the transaction to a city usually yields only one possible city. Where the divergences are larger there will be more than one possible solution, and if there are too many possible solutions the problem might not be feasible within a time constraint.

The effect these divergences produce in the efficacy of the database reconstruction attack is similar to what is expected from the implementation of Differential Privacy [7] techniques. Differential Privacy techniques rely on the introduction of random errors into the data. If applied to the datasets discussed here, that would also yield divergences between the summaries and the underlying micro-data (transactions). Nonetheless, Differential Privacy techniques have the advantage of providing theoretical limits for information leakage and defense against deterministic re-identifications.

By grouping the datasets De-identified transactions, Summary by city, and Summary by NCM by SH4 and NCM, and comparing the total FOB values we found the following for January/2021.

1. The total imports amount to 15.2 billion USD spread across 7,062 unique NCMs. These figures considers the Summary by NCM as the official source.

[12]They are the reason the variable y is included in the optimization problem from Section 3.4

2. There is no divergence between Summary by NCM and Summary by city when they are grouped by HS4, country of origin and federations state (UF).

Regarding the differences between De-identified transactions and Summary by NCM:

1. The RFB suppressed the information of all transactions among 2,277 NCMs. This amounts to 1.5 billion USD or 10% of the total imported.

2. For 1,672 NCMs which amount to 5.3 billion USD (35%) the differences for values are between 2 and -2 USD. This might be caused just by rounding or small transactions included or excluded by each publisher.

3. For 284 NCMs which amount to 1.8 billion USD (12%) the totals obtained from De-identified transactions values is greater than or equal to 2 USD, when compared to Summary by NCM. Those differences might be related to the outliers treatment done by the Ministry of Economy.

4. For the remaining 1,211 NCMs which amount to 6.4 billion USD (42%) the totals obtained from De-identified transactions values is lower than or equal to 2 USD. Those differences are mainly related to the inclusion of some ADMINISTRATIVE transactions in the Ministry of Economy releases.

4.3 Dealing with the datasets divergences

These different methodologies used by the RFB and the Ministry of Economy pose some challenges to the applicability of the ATTACK ALGORITHM. The main one is that the algorithm has the premise that each package must go to one bin, and one bin only. However, as discussed above, there are some cases where the Ministry of Economy deliberately discards some of the transactions from their summaries. In those cases, if we included those transactions in the algorithm we could end up with an allocation that does not reflect the reality. To avoid this kind of problem we need to identify these cases (where the Ministry of Economy has discarded some transaction) and either 1) use the ATTACK ALGORITHM to identify which transactions were discarded before trying to use the same algorithm to track the transaction importer; or 2) simply not solve these cases because of the extra complexity. Due to scope limitations we went on with option 2 to try to estimate the attack's reach in Chapter 5.

Even though this second option is less complex, we still need, in both cases, to correctly identify those cases where the Ministry of Economy might have discarded

some transactions when calculating the summaries. This task would be trivial[13] if it wasn't for the rounding errors differences. The challenge lies in how to differentiate divergences that are due to discarded transactions and which ones are due to the rounding methodology.

If the rounding mechanism is unknown[14] it is possible to model it as a random error with a uniform distribution. If we are interested only in the differences produced by rounding mechanisms, a function that rounds a number to an integer will produce similar results to adding a small random number to it.

For instance, assume n transactions with rational ($\mathbb{Q}$) values v_i. We can model those values as $v_i = k_i + \varepsilon_i$, where k_i is an integer and ε_i follows a uniform distribution from -0.5 to 0.5. We can write $\sum^n v_i - \sum^n k_i = \sum^n \varepsilon_i$. The maximum value that $\sum^n \varepsilon_i$ can assume is $n \times 0.5$. This could be used as a limit to tell apart differences due to rounding and transaction exclusion. However this limit is too extreme. As n grows it becomes increasingly unlikely for $\sum^n \varepsilon_i$ to be equal to $n \times 0.5$. In fact by the Central Theorem limit [4] the probability that $\sum^n \varepsilon_i > n \times {}^{2.33}/_{\sqrt{12n}}$ is lower than 1%. That is the limit we adopted for telling apart which differences in the figures reported were due to rounding and which were due to some exclusion criteria adopted by the Ministry of Economy.

For, example, if we have 3 transactions whose original values sum up to 5.7, but the reported sum is 4.5, the actual difference is 1.2 and the limit is 1.165. This means that (if our premises are right) there is less than 1% of chance that this difference is due to just rounding, more probably one of the transactions was not included in the total report. In these cases we did not tried to re-identify the transactions.

Data preprocessing Before starting the attack the adversary needs to execute a preprocessing routine over the De-identified transactions dataset. Only data from the DI header will be used. Records where NAT. INFORMACAO is not EFFECTIVE are filtered out. To eliminate duplicates[15] the first 17 digits of the NUMERO DE ORDEM field will be used as key.[16]

[13] Without rounding, if the totals in the summaries were less than the transactions' sum then some transaction were omitted from the summaries.

[14] We tried 1) rounding the transactions values to the closest integer, and then sum, and 2) sum the transactions and round the totals. In neither case we were able to replicate the rounding methodology used by the Ministry of Economy.

[15] As some DIs have more than one item the DI header information is duplicated for each one of the items.

[16] The relation between these 17 first digits and the fields that are from the DI header is obvious from what can be seen in the database.

Chapter 5

Attack reach and consequences

This chapter is divided into 3 sections. The first one (5.1) tries to estimate the attack reach by running the proposed algorithm over all possible instances for the foreign trade statistics of January 2021. In Section 5.2 the legal framework regarding the publication of the datasets and data protection laws is presented more in depth. Finally, Section 5.3 addresses what was not covered in this study and is subject to future works.

5.1 Estimating the reach

To estimate the reach of the attack we will model the adversary as a *Marketer* whose objective is to re-identify as many transactions as they can. As this adversary we are proposing does not know how to differentiate importers from the same city, they will try to track down the transactions back to the importers city and once they know that they will guess any importer of that city using a uniform distribution.

Beyond having their knowledge bounded by the information they can access, the adversary will also be bounded computationally. The adversary will not be able to run instances of the ATTACK ALGORITHM problem with complexity[1] above 24 and they will have a limit of 1 minute to solve each instance.[2]

Their prior knowledge will be the same as the one stated in the example. The attack development will be divided in 3 phases.

Phase 1 is similar to what was presented in the example from Section 1.2, they will try to reconstruct the link between Summary by city and De-identified transactions

[1] $\log_{10} m^n$, as discussed in Section 3.3.2

[2] This arbitrary limit of 24 was chosen based on the researchers difficulty with instantiating problems bigger than that as almost always that would take more than a minute.

and then select a fitting importer from Importers.

Phase 2 using the Summary by NCM the adversary will try to reduce the size of some instances of the ATTACK ALGORITHM that were too large to be solved within the constraints defined in phase 1. This will be done by using the Summary by NCM to narrow the possible federation states for each transaction.

Phase 3 by using the knowledge acquired in the previous phase the adversary will produce smaller instances of the ATTACK ALGORITHM by using the federation state of each transaction.

The following subsections will present the adversary *modus operandi*.

5.1.1 Prior

The adversary only has access to Importers and De-identified transactions datasets from January/2021. They want to re-identify as many transactions, from the 817,468 in total,[3] as they can. Their chance of achieving this for a randomly chosen transaction is $1/18{,}430 \approx 0.05\%$ and by doing this guessing game for all transactions they are expected to get around $1/18{,}430 \times 817{,}468 \approx 44$ of those right. These transactions amount to 13.1 billion USD, so the value at risk is $44/817{,}468 \times 13.1 \approx 712$ thousand USD.

5.1.2 The attack

During the attack development the adversary will have access to Summary by city and Summary by NCM datasets.

5.1.2.1 Phase 1

The goal of this phase is to create a direct link between Summary by city and De-identified transactions. This phase will be further divided into 3 steps.

Step 1 In this step the adversary will only perform a simple left join of De-identified transactions over Summary by city using the common fields of both datasets. Which are:

1. ANOMES $\leftrightarrow$ (CO_ANO, CO_MES) [MONTH key]
2. COD.NCM (first 4 digits) $\leftrightarrow$ HS4 [HS4 key]

[3] Considering only EFFECTIVE ones.

3. PAIS DE ORIGEM (code) ↔ CO_PAIS [ORIGIN key]

This join achieves 2 things. It reduces the number of possible cities a transaction might be linked to, and alongside the number of importers. It also creates the instances of the PACKAGE ALLOCATION problem. Each grouping defined by the MONTH, HS4 and ORIGIN key above is a instance for the ATTACK ALGORITHM where the packages are the transactions and the bins are the totals by city.

After this step the adversary chance of success for a randomly chosen transaction has increased to approximately 1 in 1,161 $\approx 0.1\%$ (16× factor increase). By random guessing they are expected to re-identify around 705 transactions and the value at risk is 46.6 million USD (45.8 million USD of leakage).

Step 2 In the second step the adversary will select the instances they will be able to use the PACKAGE ALLOCATION problem to further narrow the possible cities for each transaction. The inclusion criteria for the instances is the following:

1. The complexity ($log_{10}m^n$) is lower than or equal to 24 (this is the adversary's computational limit).
2. The global rounding error $\epsilon = \sum(W) - \sum(C)$ for the FOB value should be within what is expected from errors that are due only to rounding. The maximum value allowed is $n \times {}^{2.33}/\sqrt{12n}$, where n is the number of packages in the instance.

Only 127,346 (16%) transactions meet the criteria above. Their total value amount to 4.2 billion USD (28%). Table 5.1 details how many transactions met each criteria. The main exclusion factor is the complexity.

Table 5.1. Number of transactions that meet each criteria (Phase 1)

		Complexity		
		> 24	≤ 24	Total
Rounding error	$>$ limit	20,069	2,479	22,548
	$\leq$ limit	667,574	**127,346**	794,920
Total		687,643	129,825	817,468

Step 3 Here the adversary runs the bi-dimensional algorithm (value and weight) for each one of the instances selected in the previous step. The packages weights are extracted from the De-identified transactions field PESO LIQUIDO and the values

from VMLE DOLAR. The bins are extracted from the Summary by city dataset. The field KG_LIQUIDO is the bin capacity for weight and VL_FOB the capacity for value.

After this step the adversary chance of success for a randomly chosen transaction has increased to approximately 1 in 227 $\approx$ 0.44% (5$\times$ factor increase). By random guessing they are expected to re-identify around 3,615 transactions and the value at risk is 286.3 million USD (239.7 million USD of leakage) Also a total of 80,509 (9%) transactions were deterministically tracked back to their original city, of those, 980 went to cities with just one importer.

5.1.2.2 Phase 2

The goal of this phase is not to narrow down the number of possible importers, so the adversary will not update their chances in this phase. The goal of this phase is to reduce the size of the instances that were above the complexity threshold of 24 in the previous phase. This will be possible through the linking of De-identified transactions with Summary by NCM and the usage of the ATTACK ALGORITHM to narrow down the possible federation state of the transaction before retrying again to find its city. This phase is divided into 3 steps.

Step 1 In this step the adversary will only perform a simple left join of De-identified transactions over Summary by NCM using the common fields of both datasets, which are:

1. ANOMES $\leftrightarrow$ (CO_ANO, CO_MES) [MONTH key]
2. COD.NCM $\leftrightarrow$ CO_NCM [NCM key]
3. PAIS DE ORIGEM (code) $\leftrightarrow$ CO_PAIS [ORIGIN key]

Like in the previous phase, this join creates the instances of the ATTACK ALGORITHM. Now the key for each instance is based on the MONTH, NCM and ORIGIN key above. The packages are the transactions and the bins the totals by NCM.

Step 2 In a similar way as done in the previous phase the adversary now selects which instances of the ATTACK ALGORITHM they will try to solve. The criteria is the same. Table 5.2 shows the number of transactions that meet each criteria. The total for this table is lower than Table 5.1 because the transactions that have already been locked into a single city were excluded from this phase (the city is already known, there is no need to find out the federation State).

Table 5.2. Number of transactions that meet each criteria (Phase 2)

		Complexity		
		> 24	≤ 24	Total
Rounding error	$>$ limit	21,898	4,306	26,204
	$\leq$ limit	437,879	**272,876**	710,755
Total		459,777	277,182	736,959

Step 3 Likewise in the previous phase, in this step the adversary will run a five-dimensional (value, weight, quantity, freight and insurance) version of the ATTACK ALGORITHM. The goal here is to lock the transactions to a single federation state (field SG_UF_NCM of Summary by NCM). Table 5.3 shows the correspondence of the fields for each one of the datasets.

Table 5.3. Fields used in the five-dimensional PACKAGE ALLOCATION problem

Field	*Summary by NCM*	*De-identified transactions*
Value	VL_FOB	VMLE DOLAR
Weight	KG_LIQUIDO	PESO LIQUIDO
Quantity	QT_ESTAT	QTDE ESTATISTICA
Freight	VL_FRETE	VL FRETE DOLAR
Insurance	VL_SEGURO	VL SEGURO DOLAR

After running this step a total of 257,333 (31%) transactions were locked into a single federation state, of which 176,454 do not have their city already defined. Those are the transactions that will go to the next phase.

5.1.2.3 Phase 3

In this phase the adversary uses the information obtained in the previous one, regarding the transactions federation state, to build new instances of the PACKAGE ALLOCATION problem. This phase is also divided into 3 steps.

Step 1 in this step the adversary performs a left join of the transactions obtained in the previous phase (that are locked into a single federation) over the Summary by city using the following fields:

1. ANOMES $\leftrightarrow$ (CO_ANO, CO_MES) [MONTH key]
2. COD.NCM (first 4 digits) $\leftrightarrow$ HS4 [HS4 key]

3. PAIS DE ORIGEM (code) ↔ CO_PAIS [ORIGIN key]

4. SG_UF_NCM ↔ SG_UF_MUN [UF key]

Again this linkage provides the instances that will be used in the ATTACK ALGORITHM, just like in phase 1. The advantage is that by using the additional UF key some instances will be below the computational complexity set for the adversary.

After this step the adversary chance of success for a randomly chosen transaction has increased to approximately 1 in 153 ≈ 0.65% (1.5× factor increase). By random guessing they are expected to re-identify around 5,305 transactions and the value at risk is 391.2 million USD (104,9 million USD of leakage).

Step 2 This is the filtering step and the criteria is the same as before. Table 5.4 presents the the number of transactions that meet each criteria. The total number of transactions in this table represents the number of transactions that were locked into a single federation state in the previous phase.

Table 5.4. Number of transactions that meet each criteria (Phase 3)

		Complexity		
		> 24	≤ 24	Total
Rounding error	$>$ limit	0	760	760
	$\leq$ limit	48,871	**126,823**	175,694
Total		48,871	127,583	176,454

Step 3 Again the adversary runs the bi-dimensional algorithm (value and weight) like they did in phase 1.

After this last step the adversary chance of success for a randomly chosen transaction has increased to approximately 1 in 106 ≈ 0.94% (1.4× factor increase). By random guessing they are expected to re-identify around 7,716 transactions and the value at risk is 494.3 million USD (103.1 million USD of leakage).

5.1.3 Summary of the results

After going through all 3 phases, the adversary gained deterministic knowledge of the importers' city in 138,413 transactions, which amount to 6.1 billion dollars (40%). For 2,003 of such transactions, that amount to 137.3 million dollars (0.9%), there is only one importer in the city. That means that the adversary gained access to information

subject to fiscal secrecy of 348 companies (2% of the total) that imported something in the month of January 2021. Table 5.5 shows the top 5 cities (of this total of 348), by value, where there were more deterministically re-identified transactions. All these cities had just one importer in January 2021.

Table 5.5. Top 5 cities with most deterministically re-identified transactions

City	*# of re-identified transactions*	*FOB value USD*
OURO BRANCO (MG)	28	24.6 mi
IPAUSSU (SP)	3	10 mi
GUARANTA DO NORTE (MT)	5	9.9 mi
SAO GABRIEL (RS)	66	7.3 mi
COMODORO (MT)	3	4.1 mi

Table 5.6 shows the adversary deterministic and probabilistic success in each phase. The probabilistic chance is the average chance of re-identifying one random transaction, and the probabilistic value at risk (VAR) is the chance of re-identifying the transaction times its value. The deterministic success is for transactions that were traced back to just one importer, so the chance of re-identifying is 100%. Column "# of transactions" shows how many transactions were deterministically re-identified in each step.

Table 5.6. Success metrics and leakage by step

		Probabilistic				Deterministic	
		Chance	VAR (USD)	Chance Increase	Value (USD) Leaked	# trans.	VAR (USD)
Prior		0.05%	712k			0	0
Phase 1	Step 1	0.1%	47 mi	16 ×	46 mi	91	3 mi
	Step 3	0.44%	286 mi	5 ×	240 mi	980	99 mi
Phase 3	Step 1	0.65%	391 mi	1.5 ×	105 mi	1,463	110 mi
	Step 3	0.9%	494 mi	1.4 ×	103 mi	2,003	137 mi

Table 5.6 is the final result of all the work done here. The 0.9% chance shown in Phase 3 - Step 3 means that if we select some arbitrary transaction in the RFB data set the chance of the adversary correctly guessing that specific transaction underlying importer is 0.9%. At first glance this result might feel dismal, however, given that most transactions are huge in value the column VAR (Value-At-Risk) shows another perspective. After Phase 3 - Step 3, the adversary's chance of correctly guessing the importer times the value of each transaction shows that the economic harm can probably reach up to 494 million dollars (in this single month). The deterministic section of Table 5.6 is even more alarming from the legal stand point. It shows that for 2,003

transactions the adversary does not have a "chance of guessing correctly", he deterministically knows, without plausible deniability, the underlying importer. In fact, in these cases those importers have their data, subject to fiscal secrecy, disclosed in an unwanted manner. The total value of those 2,003 transactions is 137 million dollars.

5.1.3.1 Remarks about solving time

In total, after going through all the attack phases, 84,698 instances of the problem[4] were built of which 69,244 (82%) were solved, *i.e.* the algorithm returned a valid response within the 1 minute time constraint. The remainder 15,454 (18%) were either skipped due their high complexity, or failed to return a result within the time constraint. Table 5.7 shows that the average solving time (for the instances that were solved) was 1 second. In fact, more than 99% of the solved instances were solved under 28 seconds (around 68 thousand).

Table 5.7. Solved and non-solved instances statistics

Final status	*Count*	*Avg. Complexity*	*Avg. Solving time (s)*
Not solved	15,454	136	-
Solved	69,244	4	1
Total	84,698	28	-

Figure 5.1 shows how the average solving time (in seconds) increases with the instance complexity.

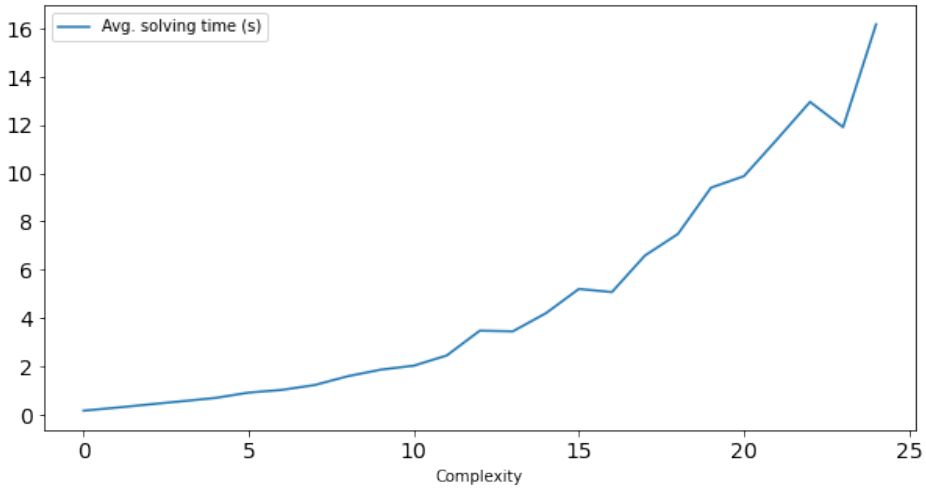


Figure 5.1. Average solving time by complexity ($\log_{10} m^n$)

[4] An instance ($\mathfrak{I}$) as defined in Section 3.1.

5.2 Brazilian foreign commerce statistics legal framework

In Brazil, the Foreign Trade Secretariat (COMEX)[5] is one of the government branches in charge of producing and disclosing statistical data about Brazilian international commerce exchange. The publication of this data aims to meet international recommendations about merchandise trade statistics and enable the integration and comparison of Brazilian foreign commerce statistics with other countries. This process is locally regulated by the ordinance 7,017 of March 11th, 2020[6] which limits what can and cannot be disclosed in such publications. In its 5th article this regulation explicitly mandates that the disclosure of the foreign trade statistics should respect the principles of publicity and confidentiality. Moreover, the 8th article states that data subject to secrecy laws should not be disclosed which includes any information that could reveal the economic state of individual business such as commercial relationships, suppliers, customers, values and volumes sold or purchased. This ordinance also allows the usage of anonymization techniques in order to meet these data protection requirements.

Other government institution that publishes data about Brazilian foreign trade is the Brazilian Tax and Customs Administration (RFB).[7] The publication of this data is regulated by the ordinance 361 of March 14th, 2016[8] which states that the purpose of such publications is to subside market studies, public policies, sector analysis and also to enable the identification of tax evasion, unfair competition and product forgery practices. Nonetheless, the limits of what can and cannot be made public by the RFB are stated in the ordinance 2,344 of March 24th, 2011.[9] This latter regulation states that any information disclosed by the institution should preserve the tax secrecy of the underlying liable businesses (the tax payers). Further details are given by its 2nd article, which agrees with the aforementioned ordinance 7,017, in fact, any data about the economic or financial state of the liable business, that was obtained by the RFB to meet tax payment ends, including customs, is subject to tax secrecy. The only exceptions allowed by this regulation are information about: registration, such as name, address and registration number; current tax situation (without revealing the amount due); and aggregated information that does not identify the underlying liable business.

[5]Secretaria Especial Substituta de Comércio Exterior e Assuntos Internacionas, do Ministério da Economia

[6]Portaria nº 7.017, de 11 de março de 2020, do Ministério da Economia

[7]Receita Federal do Brasil

[8]Portaria RFB nº 361, de 14 de março de 2016

[9]Portaria RFB nº 2344, de 24 de março de 2011

In summary, both ordinances (7,017 of March 11th, 2020 and 2,344 of March 24th, 2011) require the disclosure of foreign trade statistics while also recognizing as sensitive information **values** and **volumes** sold or purchased by individual companies and forbidding government entities from publishing this kind of information without the usage of *disclosure control* techniques that should protect the identity of the underlying businesses in each transaction.

In fact, the disclosure of all those datasets is legally required and each one, by itself, does not enables the violation of the fiscal secrecy of the underlying companies. However, these datasets in conjunction allow an adversary to have access to confidential information.

This study's findings were presented to the RFB at June 17th in a digital meeting where they exposed their concerns regarding the vulnerability as well as their willingness to address the problem. A follow-up meeting with some members of the companies alliance called PROCOMEX[10] happened on July 15th when, again the findings were presented. On December 16th 2021 the RFB published the ordinance 100 that took offline their website where foreign trade statistics were published. The official motive was not disclosed in the ordinance. Without the RFB database there is no way to implement the attack presented in this text.

5.3 Future work

This study presented an undocumented type of attack over official Brazilian foreign trade statistics and also tried to estimate its reach. Due to scope limitations many approaches were not developed. There are still some questions unanswered that could be the theme of future studies.

1. What is the effect of the computational limit on the adversary's success? How much more the adversary can do if its computational power increases?

2. If the ATTACK ALGORITHM was used, prior to the re-identification phases, to identify which transactions were excluded by the Ministry of Economy for statistical purposes,[11] would that affect the adversary's chance of success?

3. What are the chances of success of a slightly stronger adversary?

Regarding this last item, it is worth noting that: 1) a lot of information available in the datasets presented here was not explored in the adversary's advantage. This

[10] http://www.procomex.org.br/

[11] As discussed in Section 4.3.

includes the goods country of acquisition, port of entrance, importers industry segment and other minor details; 2) There is a lot more public information available in the internet, including the foreign trade statistics of other countries that can be used in the adversary's favor.

In complement, there are 7,590 importers in the month of January 2021 that were the only ones within a city and economic activity. This means that more than 41% of the Brazilian importers are at high risk of having their tax secrecy violated if a slightly stronger adversary uses the techniques presented here.

Lastly, and maybe more important, which disclosure controls techniques are more effective in preventing unwanted information leakages from occurring? The current research points to Differential Privacy techniques, however the balance between utility and data protection has to be discussed within each specific case and is not an easy choice. We hope that future works will be able help the RFB and the Ministry of Economy in their challenge regarding the publication of these statistics.

Bibliography

[1] Adam, N. R. and Worthmann, J. C. (1989). Security-control methods for statistical databases: A comparative study. *ACM Comput. Surv.*, 21(4):515–556. ISSN 0360-0300.

[2] Alvim, M., Chatzikokolakis, K., McIver, A., Morgan, C., Palamidessi, C., and Smith, G. S. (2020). *The Science of Quantitative Information Flow*. Springer.

[3] Arora, S. and Barak, B. (2007). *Computational Complexity*. Cambridge University Press.

[4] Bussab, W. and Morettin, P. (2002). *Estatística Básica*. Saraiva.

[5] Cohen, R., Katzir, L., and Raz, D. (2006). An efficient approximation for the generalized assignment problem. *Information Processing Letters*, 100(4):162 – 166. ISSN 0020-0190.

[6] Culnane, C., Rubinstein, B. I. P., and Teague, V. (2017). Health data in an open world. *CoRR*, abs/1712.05627. Available at http://arxiv.org/abs/1712.05627.

[7] Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407. ISSN 1551-305X.

[8] Garfinkel, S., Abowd, J. M., and Martindale, C. (2019). Understanding database reconstruction attacks on public data. *Commun. ACM*, 62(3):46–53. ISSN 0001-0782.

[9] Horowitz, E. and Sahni, S. (1974). Computing partitions with applications to the knapsack problem. *J. ACM*, 21(2):277–292. ISSN 0004-5411.

[10] Jaynes, E. T. (1957). Information theory and statistical mechanics. *The Physical Review*, 106(4):620–630.

[11] Karp, R. (1972). Reducibility among combinatorial problems. In Miller, R. and Thatcher, J., editors, *Complexity of Computer Computations*, pages 85--103. Plenum Press.

[12] Narayanan, A. and Shmatikov, V. (2006). How to break anonymity of the netflix prize dataset. *CoRR*, abs/cs/0610105. Available at http://arxiv.org/abs/cs/0610105.

[13] Nunes, G. (2021). A formal quantitative study on privacy in the publication of educational censuses in brazil. Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

[14] SECEX (2021). Revisão metodológica da contabilização dos fluxos de exportação e importação brasileira de bens. Available at https://www.gov.br/economia/pt-br/centrais-de-conteudo/publicacoes/notas-informativas/2021/nota-informativa-revisao-da-metodologia-da-balanca-comercial.pdf/view.

[15] United Nations (2011). *International Merchandise Trade Statistics*. United Nations. Available at https://unstats.un.org/unsd/trade/imts/eg-imts/IMTS