Danilo Fabrino Favato*, Gabriel Coutinho, Mário S. Alvim, and Natasha Fernandes

A novel reconstruction attack on foreign-trade official statistics, with a Brazilian case study

Abstract: In this paper we describe, formalize, implement, and experimentally evaluate a novel transaction re-identification attack against official foreign-trade statistics releases in Brazil. The attack's goal is to re-identify the importers of foreign-trade transactions (by revealing the identity of the company performing that transaction), which consequently violates those importers' fiscal secrecy (by revealing sensitive information: the value and volume of traded goods). We provide a mathematical formalization of this fiscal secrecy problem using principles from the framework of quantitative information flow (QIF), then carefully identify the main sources of imprecision in the official data releases used as auxiliary information in the attack, and model transaction re-construction as a linear optimization problem solvable through integer linear programming (ILP). We show that this problem is NP-complete, and provide a methodology to identify tractable instances. We exemplify the feasibility of our attack by performing 2,003 transaction re-identifications that in total amount to more than $137M, and affect 348 Brazilian companies. Further, since similar statistics are produced by other statistical agencies, our attack is of broader concern.

Keywords: Quantitative information flow, Integer optimization programming, Database reconstruction attack, Foreign trade statistics

*Corresponding Author: Danilo Fabrino Favato: Universidade Federal de Minas Gerais, Brazil, E-mail: dfavato@dcc.ufmg.br
Gabriel Coutinho: Universidade Federal de Minas Gerais, Brazil, E-mail: gabriel@dcc.ufmg.br
Mário S. Alvim: Universidade Federal de Minas Gerais, Brazil, E-mail: msalvim@dcc.ufmg.br
Natasha Fernandes: Macquarie University, Australia, E-mail: natasha.fernandes@mq.edu.au

1 Introduction

Foreign-trade statistics comprise national import and export summaries, and are instrumental both for national governments and for official supranational and international organizations. These statistics are used, for example, to produce a country's National Accounts, which is the official bookkeeping of its economic activity, as well as to provide accurate measures of that nation's Gross Domestic Product (GDP) [1].

The *utility* of accurate foreign-trade statistics, however, surpasses governments, as they can provide valuable information to society as a whole. The release to the public of accurate trade statistics enables many useful analyses including, for example, the identification of harmful economic practices such as forgery and predatory prices (dumping) [2], market and economic studies for new businesses planning, and assessment of the impact of governmental policies.

However, utility to society is not the government's only concern when releasing foreign-trade statistics. In order to foster a fair, free-market environment, several countries grant companies legal rights to some level of *fiscal secrecy* about their foreign-trade transactions.

In Brazil, federal legislation requires public disclosure of foreign-trade statistics,[1] but leaves the specific granularity of what is disclosed to each governmental agency's discretion. On the other hand, Brazilian fiscal secrecy legislation recognizes as sensitive information —and therefore worthy of protection— the values and volumes of goods traded by individual companies. It also forbids governmental entities from publishing this kind of information without the usage of disclosure control techniques protecting the identity of the importer in each transaction.[2] In order to balance the utility (to the general public) of foreign-trade statistics and the legally required fiscal secrecy (of companies), Brazilian governmental agencies have (since 2007) been publishing datasets of foreign-trade transactions anonymized

---

1 Item X of Article 92, Decree 9745 of April 8th, 2019 and Article 1, Item XII of the Brazilian Tax and Customs Administration internal regiment (Annex I Ordinance 284 of July 27th, 2020).
2 Article 198, National Tax Law 5.172, October 25th, 1966.

using techniques such as de-identification, generalization and suppression[3] (detailed in Section 2.2 ahead).

In this paper we describe, formalize, implement, and experimentally evaluate a novel transaction re-identification attack against official foreign-trade statistics releases in Brazil. The attack's goal is to re-identify the importers of foreign-trade transactions (by revealing the the identity of the company performing that transaction), which consequently violates those importers' fiscal secrecy (by revealing sensitive information: the value and volume of traded goods). After properly formalizing and implementing our attack, we demonstrate its effectiveness against Brazilian releases of official foreign-trade statistics by performing 2,003 transaction re-identifications that in total amount to over $100M in value, and affects 348 companies.

It is noteworthy that, although in the past the methodology for collection and organization of foreign-trade statistics varied significantly from country to country, nowadays a standardization provided by the United Nations (UN) [3] is adopted by many countries, including Brazil. This means that Brazilian foreign-trade statistics analyzed in this paper are highly compatible with those of other countries, and therefore our novel attack might also be of concern to those countries.

1.1 Contributions and plan of the paper

The main contributions of this paper are the following.
1. We identify and describe a novel transaction re-identification attack that violates the fiscal secrecy of companies by reconstructing foreign-trade transaction data from publicly available datasets in the context of Brazil. Since production of Brazilian foreign-trade statistics follows the UN's standard methodology, our attack may be of concern to other countries as well. (Sec. 2).
2. We provide a mathematical formalization of the fiscal secrecy problem of transaction re-identification using principles from the framework of *quantitative information flow* (QIF) [4], which has sound information- and decision-theoretic grounds to naturally model an adversary's knowledge, goals, and capabilities, as well as the leakage of information caused by an attack. (Sec. 3).

3. We carefully identify the main sources of imprecision in the official foreign-trade statistics used as side-information in our attack, including:
   (a) the splitting of transaction information across various datasets;
   (b) the sanitization techniques employed by the Brazilian government in the publishing of these statistics, such as de-identification, generalization, and suppression; and
   (c) rounding errors introduced (voluntarily or not) in the data publishing process.
   To cope with such sources of uncertainty, we model transaction re-construction as a linear optimization problem solvable through integer linear programming (ILP), and show that it is NP-complete. We then provide a methodology to identify tractable instances of the problem, as well as to simplify some intractable instances by finding further constraints in order to try to make them tractable (Sec. 4).
4. We exemplify the feasibility of our attack and its algorithmic solution through a concrete case study in which over 2,000 transactions are re-identified and the fiscal secrecy of hundreds of companies are violated in Brazil, using as computational resources only a modest personal computer (Sec. 5).

Finally, in Sec. 6 we discuss related work, and in Sec. 7 we present our conclusions and future prospects.

1.2 Ethical considerations

The results of this study have already been presented to the relevant government agencies —the Brazilian Customs and Tax administration— in order to give them time to act. On December 16th 2021 they removed public access to one of their main databases, which is sufficient to prevent the type of attack presented here. For ethical reasons we have also chosen not to disclose in this paper other points of access to the data used in the description of our attack. Throughout this text we omit the underlying business direct identifiers and, in particular, we do not provide identifiers for businesses whose privacy is affected by this attack.

Following the suggestions given by our reviewers, on May 4th we contacted the UN Statistical Commission and the Tax and Customs administration of a particular

---

3 Ordinance SRF 306, March 22nd of 2007.

country —whose name we omit here for precaution[4]. Neither had replied to our e-mail by the time we finished this document.

2 The problem, the datasets, and the attack

In this section we provide an overview of the problem tackled in this paper and motivate why it is a serious fiscal secrecy matter. We start with a description of the Brazilian environment for production, processing, and publication of foreign-trade information. We then describe the publicly available data used in our attack, and provide a concrete attack instance used to compromise fiscal secrecy. Since Brazilian foreign-trade statistics methodology is highly compatible with that of other countries, the attack we identify may be replicable in those countries as well. A proper formalization of the attack —including its general form— and full algorithmic solutions are deferred to Sections 3 and 4, respectively.

2.1 Fiscal secrecy vs. societal utility in foreign-trade statistics

The current legal framework for the collection, treatment, usage, and publication of individuals' and companies' data by the government has two conflicting goals. On one hand, there is an increasing desire for *Open Government Data*, which "promotes transparency, accountability and value creation by making government data available to all".[5] This transparency is extremely useful for civil society to, e. g., follow and evaluate the effectiveness of public policies, or detect fraud in the use of public money, or even malpractices by ill-intended business. In this context, it is desirable that the published data have high *utility*, in the sense of being as accurate, detailed, complete, and freely available as possible.

On the other hand, there are various concerns about the unauthorized disclosure of sensitive information about individuals or companies, a worry that has sprouted many data protection laws around the globe.[6] Although these novel protection laws, as well as media and public attention, tend to focus on protecting data about individuals (*privacy*), in this paper we focus on the unwanted disclosure of information about commercial transactions between companies (*fiscal secrecy*). A recent example of the risks associated with the latter is the case in which the American National Security Agency (NSA) was accused of worldwide industrial espionage and the shock produced by these allegations.[7]

For a more concrete example of the importance of fiscal secrecy, consider a small company A that, through its own research and merit, has established an agreement with a new foreign supplier which has given A a significant edge on the market. Now say that B, a very large (and ill-intentioned) competitor of A, is interested in finding out A's new supplier to approach it and either offer it an exclusivity contract (in which the supplier could only sell to B), or just start buying all the supplier's capacity in order to starve A. This could lead to A being pushed out of the market and to significant losses for its owners and employees.

In general, violations of fiscal secrecy can be harmful to the idea of free markets. Indeed, *symmetric information* is recognized as one the key requirements for a free competitive market [5]. Of course, in real world markets there is no such a thing as perfectly symmetric information. Each player will know something different about their market, and that is considered acceptable when such information was acquired through legal methods such as research, qualified personnel, etc. Nonetheless, when the information is acquired illegally it is usually tagged as "privileged" or "espionage", and seen as detrimental in fair free-market competition.

In this context, fiscal secrecy violations —whereby some companies may illegally obtain secret information about other companies— are of concrete concern. It is thus a challenge for governments to balance the utility (to the general public) of foreign-trade statistics and the fiscal secrecy (of companies).

---

4 This country was directly contacted as it is, to our knowledge, the one whose disclosure methodology most closely resembles the Brazilian, and thus for which our attack could be easily adapted.
5 https://web.archive.org/web/20201120073847/https://www.oecd.org/gov/digital-government/open-government-data.htm

6 Notable examples or privacy legislation include the General Data Protection Regulation in the European Union (GDPR 2016/679), the *Lei Geral de Proteção de Dados Pessoais* in Brazil (LGPD 13.709/2018), the Personal Information Protection and Electronic Documents Act in Canada (PIPEDA April 13th 2000), and The Data Privacy Act in the Philippines (2012).
7 https://web.archive.org/web/20210417032720if_/https://www.reuters.com/article/us-security-snowden-germany-idUSBREA0P0DE20140126

2.2 The data release methodology

Fig. 1 provides an overview of the process of production, compilation, sanitization, and publication of foreign-trade statistics in Brazil.
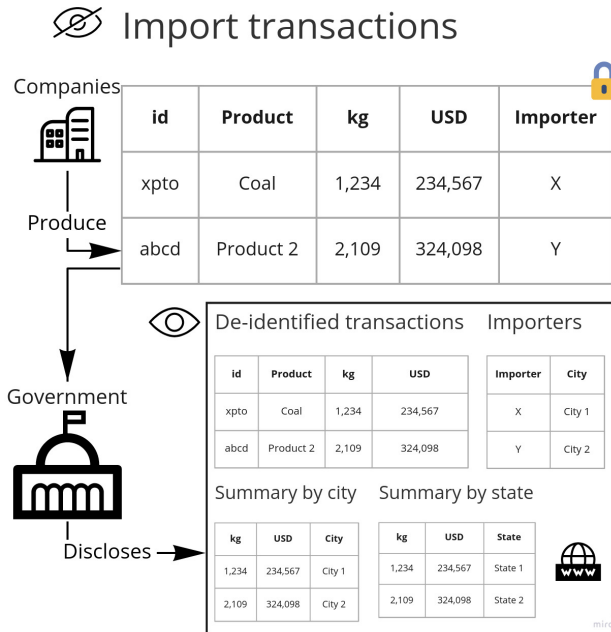
✑ Import transactions

Fig. 1. Overview of the generation of foreign-trade statistics.

The process begins with each company in the country producing and submitting to Siscomex (the Brazilian International Commerce System)[8] a legally required *Declaration of Imports* (DI) whenever an import transaction occurs.

The DI contains detailed information about each transaction performed by the company, including the goods imported, its quantity, from whom, and at what price and date. It is a document analogous to an invoice in the foreign-trade context, and it clearly states the seller and the importer of each transaction.

Upon receiving all DIs from all over the country, the government compiles them into a single collection of foreign-trade transactions. This raw-data collection is only used internally by the government, but a sanitized form of it is published after the application of the following sanitization techniques: *de-identification*, by which obvious individual identifiers (such as an im-

porter's name) are removed; *generalization*, by which data is provided in coarser granularity (e. g. when transaction values are grouped by city); and *suppression*, in which part of the data is removed from the collection. The goal of these disclosure methods is to break the connection between the goods imported and the companies behind such transactions.

The end result of this process are the 4 datasets described below, which have been publicly available in the period we conducted this study, and were updated monthly. For brevity, only relevant fields from each dataset are included here; full dataset details are provided in Appendix A. In addition, as explained in Sec. 1.2, we have chosen not to disclose public access urls for these datasets.

1. The DE-IDENTIFIED TRANSACTIONS dataset (Tbl. 1a) contains detailed records from the original compilation of all companies' DIs, with all importer's direct identifiers omitted.[9]
2. The IMPORTERS dataset (Tbl. 1b) contains data about businesses that have conducted at least one import operation in the current year. New businesses were added to the list each month, so anyone with access to the dataset could pinpoint the exact month in which a given importer started importing by tracking the additions made to the list.
3. The SUMMARY BY CITY dataset (Tbl. 1c) contains data of transactions' value and weight aggregated by city.
4. The SUMMARY BY STATE dataset, (Tbl. 1d) contains aggregated data for value, weight, freight, insurance and quantity of transactions, grouped by states. This dataset contains the official figures of foreign commerce for the National Accounting System, which are used to calculate the Brazilian GDP [6].

2.3 Overview of the attack, and the correlations it exploits

The crucial challenge in our transaction re-identification attack is how to reconstruct a row of the original raw-data dataset from the publicly released, sanitized datasets described in the previous section. Clearly, from the reconstructed transaction row the adversary immediately recovers the transaction's importer and its sen-

---

8 https://www.gov.br/receitafederal/pt-br/assuntos/aduana-e-comercio-exterior/manuais/despacho-de-importacao/topicos-1/conceitos-e-definicoes/despacho-de-importacao

9 During our study, and until December 16th 2021, the Tax and Customs Administration kept an up-to-date version of the dataset on its website.

Field	Description
NUMERO DE ORDEM	Transaction identification number.
ANOMES	Year and month of transaction registry.
COD.NCM	DI header: NCM of imported goods.
PAIS DE ORIGEM	DI header: country of origin.
PESO LIQUIDO	DI header: net weight in kg.
VMLE DOLAR	DI header: FOB value in USD.
NAT. INFORMACAO	DI header: nature of the operation.

(a) Relevant fields from the DE-IDENTIFIED TRANSACTIONS dataset.

Field	Description
CNPJ	Company's registration number.
EMPRESA	Company's name.
MUNICÍPIO	Company's city.
UF	Company's federation state.

(b) Relevant fields from the IMPORTERS dataset.

Field	Description
CO_ANO	Year of registry.
CO_MES	Month of registry.
SH4	Imported goods' SH4 code.
CO_PAIS	Imported goods' country of origin.
SG_UF_MUN	Importer's federation state.
CO_MUN	Importer's city.
KG_LIQUIDO	Imported goods' total net weight in kg.
VL_FOB	Imported goods' total value in USD.

(c) Relevant fields from the SUMMARY BY CITY dataset.

Field	Description
CO_ANO	Year of registry.
CO_MES	Month of registry.
CO_NCM	Imported goods NCM code number.
CO_PAIS	Imported goods' country of origin.
SG_UF_NCM	Importer' federation state.
KG_LIQUIDO	Imported goods' total net weight in kg.
VL_FOB	Imported goods' total value in USD.

(d) Relevant fields from the SUMMARY BY STATE dataset.

Table 1. Relevant fields of published datasets, whose full descriptions can be found in Appendix A.

sitive values. The attack exploits the following correlations among the datasets published by the government.

– **NCM fields.** The fields COD.NCM in DE-IDENTIFIED TRANSACTIONS, SH4 in SUMMARY BY CITY and CO_NCM in SUMMARY BY STATE are all related to the *Common Mercosur Nomenclature* (NCM), which is a regional product categorization system used in the Mercosur Economic Region since 1995.[10] It was derived from (and is highly compatible with) the *Harmonized Commodity Description and Coding System* (HS), which was designed and is maintained by the World Customs Organization.[11] NCM codes are 8-digits long and follow a hierarchical organization. Eg., the NCM code 08051000 identifies *Fresh oranges*. Its first four digits, 0805, represent the hierarchical level labeled SH4 *Citrus fruit, fresh or dried*. Whenever a NCM code appears in DE-IDENTIFIED TRANSACTIONS, it also appears in SUMMARY BY STATE, and its SH4 appears in SUMMARY BY CITY.

– **Final destination fields.** The fields MUNICÍPIO in IMPORTERS, CO_MUN in SUMMARY BY CITY and CO_URF in SUMMARY BY STATE are all related to the importer's official address. The first two fields refer to the importer's city, and the last refers to the importer's state. Whenever a city appears in SUMMARY BY CITY, it also appears in IMPORTERS, and its state appears in SUMMARY BY STATE.

– **Country of origin fields.** The fields PAIS DE ORIGEM in DE-IDENTIFIED TRANSACTIONS and CO_PAIS in both SUMMARY BY CITY and SUMMARY BY STATE are all related to the country of origin of the imported goods. If a country appears in DE-IDENTIFIED TRANSACTIONS, it also appears in SUMMARY BY CITY and in SUMMARY BY STATE.

In the absence of any anonymization technique, the solution to the problem of reconstructing transaction rows from the public datasets could be obtained via a system of linear equations, whose variables would be the quantities (values and weights) in each transaction, whose constraints would be the summary statistics, and whose solution space would be defined by the whole set of transactions. Even in this simpler situation, however, it would still be possible that multiple solutions existed, ie., multiple mappings of transactions to importers that satisfied all constraints. But since anonymization techniques such as suppression and generalization are employed on the published data, further uncertainty is added to the problem, increasing the number of possible solutions.

For the reasons above, we model transaction reconstruction as a linear optimization problem solvable through integer linear programming. In Sec. 3 we formalize the attack, and in Sec. 4 we detail its algorithmic solution, including how we cope with further uncertainties introduced by imperfect data. Before that, however, we present a concrete motivating example of our attack.

---

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

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

NR ORDEM	10653400001
Description	Cotton T-400 [...] Metros:1.013,70(09 rolos)
Quantity	420.69 kg
Value (USD)	3,388.41
NCM	52083900
Origin	China

Table 2. Target transaction for re-identification.

2.4 A motivating example

Here we provide a concrete instance of our transaction re-identification attack. which is formalized in Sec. 3.

In this attack, depicted in Fig. 2, an adversary with very limited prior knowledge applies integer programming techniques (the ATTACK ALGORITHM detailed in Sec. 4) to publicly available data. Her goal is to reconstruct the link between a transaction chosen as target in the DE-IDENTIFIED TRANSACTIONS dataset and a company in the IMPORTERS dataset, effectively re-identifying the target transaction's importer and inferring its sensitive attributes (value and volume).

In our concrete example we describe how the adversary is able to re-identify the importer of a real transaction valued at ∼$3400 USD, performed in January 2021. For ethical reasons, we omit the importer's identity.

Before the attack. The adversary starts with access to only the publicly available DE-IDENTIFIED TRANSACTIONS and IMPORTERS datasets (Tbls. 1a and 1b, respectively). She picks as her target for re-identification the transaction specified in Tbl. 2, which we refer to by its identifying NR ORDEM: 10653400001.

Although real-world adversaries typically have some common knowledge about foreign trade (eg., they know that a clothing manufacturer is more likely to buy 400 kg of cotton cloth than a toy store is), here we are conservative and assume a very modest adversary who lacks any such common knowledge. She just assumes, a priori, that every company is equally likely to be the importer of any given transaction. This means that our adversary's best strategy to re-identify the target transaction before the attack is performed is just to blindly guess which company among the 18,430 present in the IMPORTERS dataset is the actual importer. Therefore, the adversary's probability of a successful transaction re-identification at this point is $1/{18,430} \approx 0.005\%$.

Attack execution. During the attack the adversary gains access to the SUMMARY BY CITY and SUMMARY BY STATE datasets, and uses them as auxiliary information to re-identify the target transaction's importer. She breaks her task into two parts as follows.

– First, she tries to infer the city of the target transaction. This is done by using SUMMARY BY CITY in Phase 1 to try to link the target transaction directly to a city. If that direct step is computationally intractable (as we shall see in Sec. 4, the PACKAGE ALLOCATION algorithm used to link datasets is exponential), the adversary tries to find further constraints to simplify the problem. For that, she uses SUMMARY BY CITY in conjunction with SUMMARY BY STATE in Phase 2 to try to first link the target transaction to a state, and only then she moves to Phase 3 and tries to link the target transaction to a city, given that the state is already known.
– Second, if the adversary has successfully linked the target transaction to only one city, she guesses as the importer a company based on that city, found by querying the IMPORTERS dataset.

We now describe the attack phases in more detail.

Attack Phase 1: Trying to link the target transaction directly to a city. To try to directly link the target transaction (from Tbl. 2) to its city of occurrence, the adversary will perform two steps:

1. Filter SUMMARY BY CITY and DE-IDENTIFIED TRANSACTIONS by SH4 = 5208 and country of origin = China (same parameters as target transaction). This yields 19 cities and 99 transactions.
2. Try to run the PACKAGE ALLOCATION algorithm to link the target transaction to one of the cities.

Step 1 already reduces the number of possible importers from over 18,000 to fewer than 5,000 (just the ones linked to those 19 cities). However, for our computationally bounded adversary, Step 2 is intractable. Therefore she proceeds to Phase 2.

Attack Phase 2: Reducing the instance's size by first trying to link the target transaction to a state. Here the adversary tries to link the target transaction to its state as an intermediate step for finding the transaction's city. This is also done in two steps:

1. Filter SUMMARY BY STATE and DE-IDENTIFIED TRANSACTIONS by NCM = 52083900 and country of origin = China (same parameters as target transaction). The results of these queries are show in Tbls. 3b and 3a respectively.
2. Try to run the PACKAGE ALLOCATION algorithm to link the target transaction to one of the states.

The key idea behind this linkage algorithm is that the adversary knows that there must exist at least one valid partition of these transactions among states that pro-
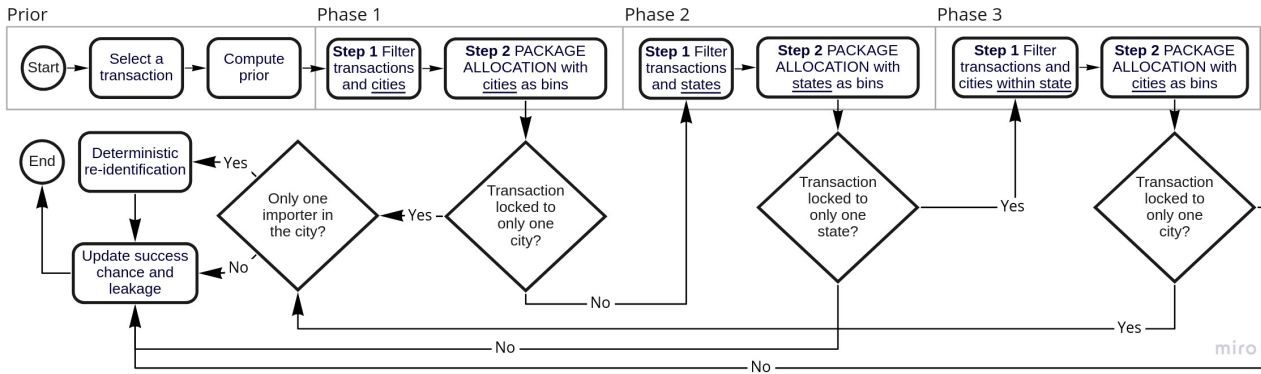
Fig. 2. Phases and steps of the attack algorithm.

duces the same sums observed in Tbl. 3b. (Notice that the totals on both tables are the same, ignoring the decimal places; we shall discuss rounding errors in Sec. 4.3).

If there is just one possible partition, or if in all possible partitions the target package would only fit in one destination, then the adversary can be sure that she has correctly identified the target transaction's destination state, and can proceed to Phase 3. Otherwise, the adversary will have more than one state and many companies left as possible importers, and she just makes a blind guess among those to infer the correct one.

In the latter case, the adversary will not be absolutely certain of a correct re-identification, but can still have some probabilistic confidence in the result.

The algorithm. The algorithm used to find these valid partitions that link a transaction to its city/state employs integer programming techniques, and is fully explained in Sec. 4. Here we provide an overview of it in the context of our motivating example.

For this algorithm, the records from DE-IDENTIFIED TRANSACTIONS (Tbl. 3a) are represented as "packages", as shown in Fig. 3a. Each package is represented by a rectangle whose width and height are the transaction's value and weight, respectively. The totals by state from SUMMARY BY STATE (Tbl. 3b) are represented as target coordinates in the plane, as shown in Fig. 3b. The x coordinate of these targets is positioned at the total value per state, and the y coordinate is at the total weight per state.

From this setup, the adversary tries to fit all the "packages" to the planes without violating the boundaries represented by the dashed lines. The packages must be stacked by their corners as shown in Figs. 3c and 3d. This arrangement guarantees that in a valid solution the sum of packages' values will be equal to the total observed in the state represented by the cor-

ner point. Figs. 3c and 3d depict valid and invalid solutions, respectively. It is easy to see that in the valid one, the sums of the packages with NO_ORDEM 12343400002, 10653400001 and 10652800002 match the total values of the state of MINAS GERAIS, whereas the package with NO_ORDEM 05566500003 alone matches the totals of the state of CEARA. If, however, we swap the packages with NO_ORDEM 12343400002 and 05566500003, we no longer have a valid solution for either state.

Moreover, it is possible to test if there is any other valid allocation in which the package with NO_ORDEM 10653400001 is placed in a state other than MINAS GERAIS. In fact, for this example there is not. This is conclusive evidence that this package must have been purchased by a company based in MINAS GERAIS.

With that, the adversary has successfully linked the target transaction to a single state, and she can move on to the next phase of the attack.

Attack Phase 3: Trying to link the target transaction to a city once the state is known. Now that the adversary knows that the target transaction was made by a company based in MINAS GERAIS she can return to a similar problem to the one she faced in Phase 1. Before knowing the state there were too many possible solutions and the problem was intractable. But by incorporating restrictions found in Phase 3, she is able to reduce the number of possible allocations and (potentially) make the problem tractable. The algorithm used to solve this new, reduced instance is the same as the one explained for the case of the states. If this reduced instance of the problem is tractable, the adversary may link the target transaction to a specific city. In our example, the adversary is able to identify the city of OURO BRANCO, in the state of MINAS GERAIS, as the location of the target transaction.
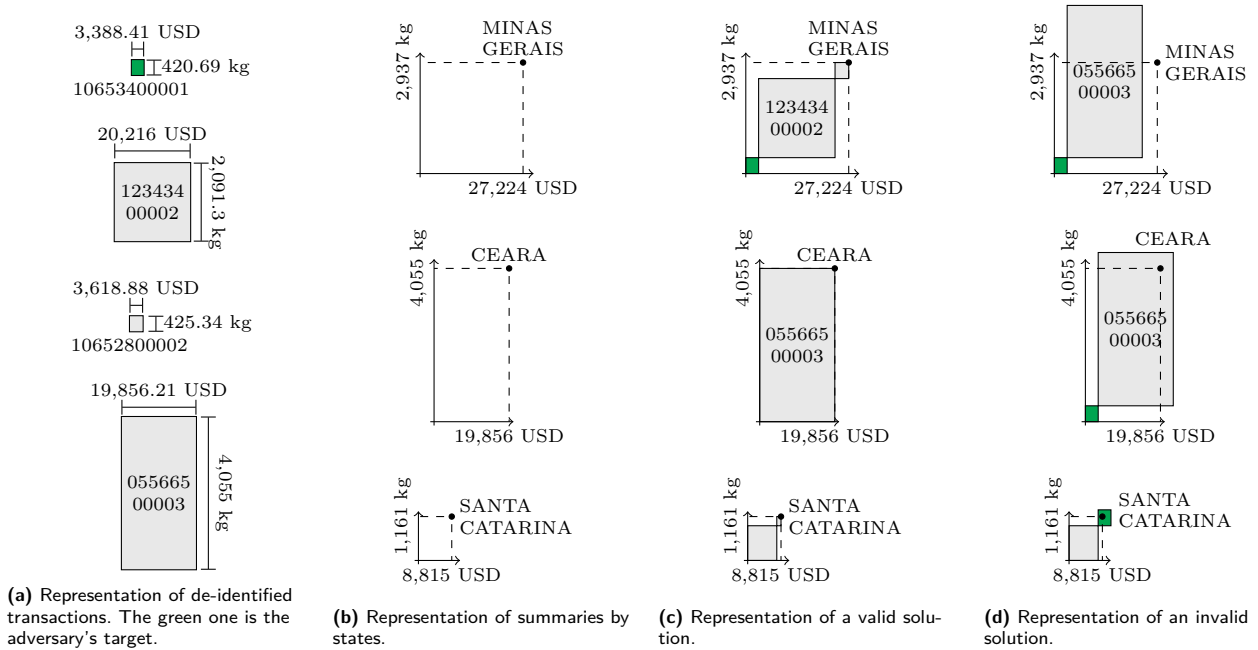
(a) Representation of de-identified transactions. The green one is the adversary's target.

(b) Representation of summaries by states.

(c) Representation of a valid solution.

(d) Representation of an invalid solution.

Fig. 3. Graphical representation of the package-allocation algorithm for Phase 2 of the attack, linking transactions to states.

After match: Inferring importer from city. After the city in which the transactions have occurred has been identified, the adversary proceeds to infer which company based on that city may be the importer. In our example, a simple inspection of the IMPORTERS dataset reveals to the adversary that there is only one company in the city of OURO BRANCO, which leads to an immediate re-identification of the importer. (For ethical reasons, we omit the affected company's name.)

Hence, in our example the adversary has a high confidence that she achieved a deterministic re-identification. This leads to the inference of the following sensitive information: (i) the importer purchased 420.69 kg of a specific type of cotton cloth; and (ii) that purchase cost 3,388.41 USD, so the average price is approximately 8.06 USD/ton. In our view, this is a clear violation of the importer's fiscal secrecy within Brazilian legal framework.[12]

In the following sections we formalize the general version of this attack, and explain how to compute its overall leakage of secret fiscal information.

---

3 Attack model

We now formalize our model, generalizing from the previous section's example, using principles from *quantitative information flow* (QIF) [4], which has sound information- and decision-theoretic grounds to model an adversary's knowledge, goals, and capabilities, as well as the leakage of information caused by an attack. This formalism is helpful for 1) measuring the attack effectiveness and reach and 2) to provide a baseline that future works can use to compare different defensive strategies against this attack.

3.1 Channels, secrets and adversaries

Using QIF we model adversaries who employ *Bayesian inference* on data release mechanisms modelled as noisy information-theoretic channels. The adversary updates her prior knowledge about sensitive values or "secrets" using Bayes' rule, thereby inducing a measure of "channel leakage", which incorporates the difference between her prior and posterior knowledge as well as her goals. We now make these ideas more precise.

A *secret* models the information sought by the adversary; the set of possible secrets is denoted by $\mathcal{X}$. A *prior* models the adversary's prior knowledge about the secrets (ie., before the attack is even performed), which can come from various sources, including other

NR ORDEM	Value USD	kg
<u>10653400001</u>	3,388.41	420.69
08986300003	111.91	23.1
02113100001	1,116.52	242.86
05566500003	19,856.21	4,055
12343400002	20,216.8	2,091.3
12319800004	1,346.64	136
12634700012	7,698.19	918.25
02898100002	57,194.45	14,986
13460400002	13,434.69	2,420
11554100002	2,400.31	182
11722800003	9,064.28	1,644
11717600002	22,297.1	1,850
11722600002	7,024.53	1,265
10652800002	3,618.88	425.34
Total	**168,768.92**	**30,659.54**

(a) Transactions in DE-IDENTIFIED TRANSACTIONS which are in the same NCM and country of origin as selected target-transaction.

State	Value (USD)	Weight (kg)
CEARA	19,856	4,055
ESPIRITO SANTO	112,762	22,483
MINAS GERAIS	27,224	2,937
SANTA CATARINA	8,815	1,161
SAO PAULO	112	23
Total	**168,769**	**30,659**

(b) States in the SUMMARY BY STATE dataset which are compatible with the selected target-transaction.

Table 3. Intermediate results used in the motivating example.

datasets, media, or simply common-sense about foreign trade. Prior knowledge is formalized as a probability distribution on possible secret values denoted $\pi \in \mathbb{D}\mathcal{X}$. (Here, $\mathbb{D}\mathcal{X}$ is the set of probability distributions over the elements of set $\mathcal{X}$.)

A *channel* C takes inputs (secrets) $x \in \mathcal{X}$ and produces an observation $y \in \mathcal{Y}$ according to some distribution in $\mathbb{D}\mathcal{Y}$. When $\mathcal{X}$ and $\mathcal{Y}$ are discrete we write C as a matrix whose element $C_{x,y}$ is the probability of producing $y \in \mathcal{Y}$ when the input is $x \in \mathcal{X}$. Rows in C are distributions over $\mathcal{Y}$ and hence are 1-summing; if C is deterministic then every entry in C is 0 or 1, with each row containing exactly one value 1.

Under the *g-leakage* framework [4], the adversary's goal is modelled with a *gain function*, and her strategy is to maximize her expected gain using her knowledge of the channel. In this work we adopt the "Bayes vulnerability" gain function which simply models an adversary whose goal is to guess the secret in a single attempt. We describe the attack in terms of the secret's "vulnerability" with respect to the adversary in question. Given the adversary's prior $\pi \in \mathbb{D}\mathcal{X}$, the prior vulnerability of the secret is given by

$$V_1(\pi) \;=\; \max_{x \in \mathcal{X}} \pi(x) \;,$$

and it measures how likely it is that the adversary can correctly guess the secret value in one try. Given access to the channel C, the posterior vulnerability (i. e., after the attack is performed) can be computed as[13]

$$V_1(\pi, C) \;=\; \sum_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}} \pi(x) \times C_{x,y} \;,$$

and it represents the expected probability of the adversary's correctly inferring the secret value in a single attempt after having observed the output of the channel.

Finally, we can compute the leakage of the channel (multiplicatively) by computing how much the channel increases the adversary's success, as

$$\mathcal{L}_1(\pi, C) \;=\; \frac{V_1(\pi, C)}{V_1(\pi)} \;. \tag{1}$$

3.2 Modelling the reconstruction

In our scenario, the adversary's goal is to identify the company behind a transaction chosen as the target, using the information from the data release. We model the *secrets* $\mathcal{X}$ as the set of possible importers and the adversary's *prior* as uniform on $\mathcal{X}$. That is, we assume a fairly weak adversary who lacks any common knowledge on trade transactions. We model the data release mechanism *for the target transaction* [14] as a channel [15] which maps importers (secrets) to cities (observations).

Before the attack, the channel is "noisy" from the adversary's perspective, as she lacks precise information about the correlation between the target transaction and the city of import. The attack's goal is to reconstruct (more precisely) this correlation, and therefore to improve her likelihood of guessing the correct importer, by exploiting the information released in the SUMMARY BY CITY and SUMMARY BY STATE datasets.

---

13 See [4, Ch 5 Thm 5.15] for details.

14 We would model the entire data release as a channel if the adversary's goal was to learn information about *any* transaction.

15 In fact what this encodes is the *correlation*, which is a joint distribution, but this can be factorized into a channel by taking appropriate marginals.

3.3 Measuring the leakage of information caused by the attack

We now return to the example of Sec. 2.4 to describe how the channel is constructed and to examine the overall leakage caused by the attack, which is quantified by comparing the adversary's measure of success before and after the attack was conducted, as per Eqn (1).

Tbl. 4 presents the adversary's probability of correctly re-identifying the importer of the target transaction (10653400001) at each phase of the attack. Before the attack, this probability is uniform over all importers, and thus the prior vulnerability is simply $1/18{,}430$, as our ill-informed adversary could only perform a blind guess over the all equally likely $18{,}430$ importers. At this point the *channel* representing the data release has as its inputs all $18{,}430$ importers and as its outputs all possible cities.

At the end of Phase 1, after having access to the SUMMARY BY CITY, the adversary knows that the target transaction could have gone to one of the 19 cities that imported from China in the SH4 5208. This reduces the number of possible importers from the initial $18{,}430$ to $4{,}923$. In our channel model, this has the effect of updating the adversary's *prior* to become uniform over the remaining $4{,}923$ importers (eliminating the remainder options). The (posterior) vulnerability after this update is $\sum_x \pi(x)C_{x,y} = 1/4{,}923$. The information leakage caused by this phase of the attack is $\mathcal{L}_1(\pi, C) = (1/4{,}923)/(1/18{,}430) \approx 3.7$, which means that the adversary's probability of correctly re-identifying the target transaction's importer has increased by a factor of 3.7 when compared to before the attack.

At the end of Phase 2, the adversary knows that the transaction went to one of the 3 cities of the state of MINAS GERAIS that imported in the SH4 from China. The IMPORTERS dataset shows that BELO HORIZONTE has 276 registered importers, EXTREMA has 49, and OURO BRANCO has 1. This reduces the number of possible importers from $4{,}923$ to 326. As before, the posterior vulnerability increases to $1/326$, and the information leakage caused by this phase is $(1/326)/(1/4{,}923) \approx 15$. Finally, at the end of Phase 3, the adversary knows that the transaction went to the city of OURO BRANCO, which has only one registered importer. Thus the posterior vulnerability is now 1 (i. e. the adversary's probability of guessing the importer is 1) and the leakage caused by the final phase is 326.

Phase	Probability of correct re-identification
Prior	$1/18{,}430 = 0.005\%$
Phase 1	$1/4{,}923 = 0.02\%$
Phase 2	$1/326 = 0.3\%$
Phase 3	$1/1 = 100\%$

Table 4. Measure of the adversary's success by phase. Notice that by the end of Phase 1 the adversary's probability of correct re-identification increases by a factor 3.7 over the prior. At the end of Phase 3, the increase factor is of 326 over the previous phase, and of $18{,}430$ over the prior.

4 The attack algorithm

In this section we introduce the mathematical modelling of the task of placing transactions (packages) into cities or states (bins) as an integer linear program (ILP). In this setting, we aim to introduce variables whose 1 or 0 value, after the execution of the algorithm, will indicate whether a package is or is not placed in a bin. The motivation to recourse to this framework is twofold.

First, a brute force attack that attempts to test all possible allocations and verify whether they satisfy validity constraints given by value and weight is intractable under current computational resources even for small examples. For instance, the example described in Sec. 2.4 yields more than six billion ways of distributing 14 transactions in 5 states.

In general, if one iterates over all possible allocations of n packages in m possible bins, this yields m^n possible allocations. A slightly larger instance than the one considered in Sec. 2.4, with $n = 40$ and $m = 12$, yields as many as roughly 10^{43} possibilities. Clearly, several of these possibilities do not need to be tested as they can be identified as sub-cases of already infeasible allocations. For instance, in Tbls. 3b and 3a, package 02898100002 can only fit in the state of ESPIRITO SANTO, thus all further tested attempts should be made with this allocation set, which eventually leads to the backtracking paradigm for testing exponentially many possibilities.

Upon modelling the problem as an integer linear program, standard techniques to solve such programs, implemented in what is typically referred to as an *ILP solver*, already employ the backtracking paradigm in conjunction with other geometric methods, yielding branch-and-bound or branch-and-cut strategies, for example. In principle, a suitable backtracking implementation should be marginally faster *if an exact allocation exists*. However, this may come at the cost of being much less flexible to modifications later observed in the data, and also of being harder to implement.

The second reason why we argue the ILP formulation is more suited than brute force, or the more sophisticated backtracking methods, is that the data typically does not allow for an exact allocation. As errors can be introduced in the data either involuntarily or by design (suppressing decimals, for instance), one should not expect the values in Tbl. 3a to sum up exactly to the ones seen in Tbl. 3b, for example. Therefore one will typically wish to find an allocation that minimizes error — as it happens, the branch-and-bound method to solve ILPs is essentially backtracking applied when a target linear function needs to be minimized.

Finally, note that it could still be the case that a clever approach could find a package allocation more quickly, upon exploring a possible specific combinatorial structure of the problem. We believe that is highly unlikely, however, as we can show that the formal problem underlying our modelling is NP-complete (Thm. 1 ahead).

4.1 Mathematical modelling and complexity

We now mathematically model the problem of allocating transactions (packages) into cities or states (bins). This shall be referred to as the PACKAGE ALLOCATION decision problem, whose instances are as follows. The mathematical formulation shown in Fig. 4 will be helpful for any future replication or expansion of this work, and Theorem 1 is important to base our decision of using ILP and not some other custom tailored algorithm.

Instance:
$$\mathfrak{I} = (P, k, w, B, c, \rho, \beta, \varepsilon) \tag{2}$$

is the input tuple for the PACKAGE ALLOCATION problem, where:

1. $P = \{p_1, p_2, ..., p_n\}$ is a set of n packages.
2. k is the number of attributes each package has. In principle, each attribute is a non-negative rational number. In the example shown in Sec. 2.4 there are two attributes: weight and value.
3. $w : P \to \mathbb{Q}_+^k$ is a function that determines for each package its attribute set. The ith attribute of package p will be denoted by $w(p; i)$.
4. $B = \{b_1, b_2, ..., b_m\}$ is a set of m bins.
5. $c : B \to \mathbb{Q}_+^k$ is a function that determines for each bin its capacity for each attribute.
6. $\rho \in P$ is a special, target package.

7. $\beta \in B$ is a special, target bin for the target package.
8. $\varepsilon : B \to \mathbb{Q}_+^k$ is a non negative tolerance for each bin and for each attribute.

An instance is a YES-instance if and only if there exists a function $X : P \times B \to \{0, 1\}$ (naturally identified with a $|P| \times |B|$ matrix with $0/1$ entries) satisfying:

(a) For each $p \in P$, $\sum_{b \in B} X_{pb} = 1$. (Each package is allocated to exactly one bin).
(b) For each $i \in \{1, ..., k\}$, and for each $b \in B$, $\sum_{p \in P} X_{pb} w(p; i) \leq c(b; i) + \varepsilon(b; i)$, and $\sum_{p \in P} X_{pb} w(p; i) \geq c(b; i) - \varepsilon(b; i)$. (Bins' capacities are filled up to tolerance for all attributes).
(c) $X_{\rho\beta} = 1$. (Package ρ is allocated at bin β).

It is immediate to verify that PACKAGE ALLOCATION is in NP, as any given function X can be verified in time polynomial to the size of the instance. More interestingly, we have the following result.

Theorem 1. *PACKAGE ALLOCATION is NP-complete.*

Proof. An instance of the SUBSET SUM problem is a set of natural numbers $N = \{n_1, ..., n_\ell\}$ and a natural number S. It is an YES-instance if and only if a subset of N sums to exactly S. This problem is well known to be NP-complete (see for instance [7, Theorem 8.23]). Given an instance of SUBSET SUM as above, generate ℓ instances of PACKAGE ALLOCATION, each of which contains ℓ packages (one for each number), only one attribute (the value of the natural number), two bins, one of capacity S and one of capacity $\sum n_i - S$, $\varepsilon \equiv 0$, fixing β to be the bin of capacity S and making ρ vary over all packages. It is immediate to verify that an instance of SUBSET SUM is YES if and only if there is at least one package ρ for which the corresponding instance of PACKAGE ALLOCATION is YES. As this is a polynomial reduction, the result follows. $\square$

Since a tailored efficient algorithm is likely nonexistent, we present in Fig. 4 an exact integer linear program formulation. We use Instance (2) and allow M to be a large enough number exceeding the sums of attributes.

It is immediate to verify that the optimum solution of the program above is equal to 0 if and only if the given instance is a YES instance. Moreover, the optimum solution, if not equal to 0, indicates which bins and attributes had to be violated in a best possible allocation, possibly helping to indicate where inconsistencies in the data occur, as we discuss in Sec. 4.3.

PACKAGE ALLOCATION ILP formulation:

minimize $y_1 + ... + y_k$.

subject to

(a) $\displaystyle\sum_{b \in B} X_{pb} = 1,$ for all $p \in P$;

(b) $\displaystyle\sum_{p \in P} X_{pb}w(p;i) \leq c(b;i) + \varepsilon(b;i) + y_iM,$

for all $b \in B, i \in [k]$;

(c) $\displaystyle\sum_{p \in P} X_{pb}w(p;i) \geq c(b;i) - \varepsilon(b;i) - y_iM,$

for all $b \in B, i \in [k]$;

(d) $X_{\rho\beta} = 1$

(e) $X_{pb} \in \{0,1\},$ for all $p \in P, b \in B$;

(f) $y_i \in \{0,1\},$ for all $i \in [k]$.

Fig. 4. Formulation of the PACKAGE ALLOCATION problem in ILP.

Finally, for the sake of intelligibility, below we discuss a modified model in which only one attribute is considered, rendering the formulation above significantly simpler. In this case, the objective function simply minimizes a real variable z, understood to model the maximum bin violation, thus restrictions of type (b) and (c) are replaced, respectively, for

$$\sum_{p \in P} X_{pb}w(p) \leq c(b)+z \quad \text{and} \quad \sum_{p \in P} X_{pb}w(p) \geq c(b)-z.$$

In this case, the instance is accepted if and only if the optimum z is below a given threshold, henceforth introduced as ν. One advantage of this model with only one attribute is that, upon removing the constraint for the special package and target bin, the solver will always return a feasible solution. This can be interpreted as a best possible allocation, and therefore allows for a good measure of the quality of the data. For example, data which has suffered artificial modifications will typically result in best possible allocations with large errors.

The ILP presented above can be readily implemented in any solver, as we discuss in Sec. 5.

4.2 Exact algorithm

In our attack model, we assume the adversary wants to know all the possible cities (or states) where the target transaction (ρ) might be placed.

There are two ways to achieve that. One of them requests the solver to print all feasible solutions whose objective value has achieved a given threshold. Not all solvers will handle this task particularly efficiently.

A second direct approach is as follows. First, request one valid allocation without restricting the bin where the package ρ can be placed, within the overflow tolerance ν.[16] After the first iteration, we retrieve the bin b where the package ρ was allocated and add as a restriction that $X_{\rho b} = 0$ for a next run of the PACKAGE ALLOCATION solver routine, thus finding possible new allocations for which package ρ is placed in a different bin. The procedure then loops, adding restrictions until it becomes infeasible. The set $R \subseteq B$ containing the possible bins for the package ρ is returned at the end. For later reference, we will call this our ATTACK ALGORITHM.

Notice how R contains all mathematically possible bins on which the package ρ would fit. If ρ is to be placed in a bin not present in R that would necessarily create a partition of packages in which at least one of the bins capacity would be greatly overflown (beyond the tolerance). The algorithm's exactness is translated to an assured re-identification given the assumption that DE-IDENTIFIED TRANSACTIONS contains the same transactions used to produce the total shown at SUMMARY BY CITY and SUMMARY BY STATE (which is ensured by the Brazilian government with high degree of certainty [6]).

4.3 Dealing with imperfect data

Although the foreign-trade statistics disclosed in the four datasets presented in Sec. 2.2 derive from the same source, some methodological differences produce divergences in the data.

1. **Numeric precision**: The DE-IDENTIFIED TRANSACTIONS reports values with up to 2 decimal places for currency and 5 decimal places for weight. SUMMARY BY CITY and SUMMARY BY STATE round all values to integers.

2. **Data suppression**: The DE-IDENTIFIED TRANSACTIONS does not disclose those NCMs in which less than 4 importers operated, in an attempt to preserve privacy. It is, however, easy to find which NCMs were omitted by observing the differences (in NCMs) between DE-IDENTIFIED TRANSACTIONS and SUMMARY BY STATE.

---

16 The adversary knows that the summary statistics are produced from the microdata, so they can be sure that at least one valid allocation exists.

3. **Outliers treatment**: Summary by city and Summary by state datasets are intended for statistical purposes and international comparability. Transactions that are caught in internal controls might be discarded from the official publications, and this process is opaque to external scrutiny.

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 [6] and we are unaware of publicly available datasets that have this information.

Divergences caused by rounding values are easy to deal with, as our tolerance parameters can be adjusted accordingly. Divergences caused by the data suppression, treatment of outliers, and the different inclusion criteria, however, do have a detrimental effect on our attack algorithm, even though only the data suppression technique has confidentiality as a goal.

Despite the fact that this kind of data manipulation may act as a kind of protection against database reconstruction attacks, it was not introduced with the explicit intent of increasing privacy, nor was it applied in a controlled manner. Importantly, it lacks the formal guarantees of techniques based on, e. g., differential privacy [8], and consequently, as we shall see, did not meaningfully protect against our attack.

For instance, by grouping the De-identified transactions, Summary by city, and Summary by state datasets by SH4 and NCM codes, and comparing the total USD values we found the following for January/2021:

1. Using the Summary by state dataset as the official source, the total imports amount to 15.2 billion USD spread across 7,062 unique NCMs.

2. There is no divergence between the Summary by state and Summary by city datasets when they are grouped by SH4, country of origin, and state.

As for differences between the De-identified transactions and Summary by state datasets in the month of January/2021, we can highlight:

1. De-identified transactions suppressed the information of all transactions among 2,277 NCMs. This amounts to 1.5 billion USD (10% of the total imported in the period of interest).

2. For 1,672 NCMs which amount to 5.3 billion USD (35% of the total imported) 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% of the total imported) the totals obtained from De-identified transactions values is greater than or equal to 2 USD, when compared to Summary by State. Those differences might be related to the outliers treatment.

4. For the remaining 1,211 NCMs which amount to 6.4 billion USD (42% of the total imported) 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 releases.

The different methodologies used pose some challenges to the applicability of the `ATTACK ALGORITHM`. The main one is that the algorithm operates under the premise that each package must go to exactly one bin. However, as discussed above, there are some cases in which the Ministry of Economy deliberately discards some of the transactions from their summaries. In such cases, if we included those transactions in the algorithm we would end up with an allocation that does not reflect reality. To avoid this kind of problem, we need to identify the cases in which the Ministry of Economy has discarded some transactions and either: (i) use the `ATTACK ALGORITHM` to identify which transactions were discarded before trying to use the same algorithm to track the transaction importer; or (ii) simply not solve these cases because of the extra complexity. In this work, we only considered the second alternative.

Although this second option is simpler, we still need, in both cases, to correctly identify those cases where the Ministry of Economy might have discarded some transactions from the summaries. This task would be trivial if we could model our problem without considering any rounding errors. The challenge therefore lies in how to distinguish divergences caused by discarded transactions from those caused by the rounding methodology.

If the rounding mechanism is unknown,[17] 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

---

[17] We tried 1) rounding the transactions' values to the closest integer, and then summing, and 2) adding up all transactions and rounding the totals. In neither case were we able to replicate the rounding methodology used by the Ministry of Economy.

a number to an integer will produce similar results to adding a small random number to it.

For instance, assume n transactions with rational values $v_i \in \mathbb{Q}$ were reported instead as their nearest integer. We can model those values as $v_i = k_i + \varepsilon_i$, where k_i is an integer and $-0.5 < \varepsilon_i \leq 0.5$. Assuming the decimals in the rational numbers follow a uniform distribution, then so do the variables ε_i. We can therefore write $\sum^n v_i - \sum^n k_i = \sum^n \varepsilon_i$, and the maximum value that $\sum^n \varepsilon_i$ can assume is $n \times 0.5$. This could be used as a threshold to distinguish differences due to rounding from those due to transaction exclusion.

Note, however, that this limit is already too loose. 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 Limit Theorem [9], the probability that $\sum^n \varepsilon_i > n \times 2.33/\sqrt{12n}$ is lower than 1%. That is the threshold 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 threshold is 1.165. Hence, if our premises are right, there is less than a 1% chance that this difference is due to just rounding, and the more probable hypothesis is that one of the transactions was not included in the total report. In such cases we did not try to re-identify the transactions.

5 Experimental evaluation

In this section we evaluate how our ATTACK ALGORITHM from Sec. 4.2 performs in a large-scale re-identification attack on the DE-IDENTIFIED TRANSACTIONS dataset from January/2021. This case study is conducted with the treatment of imperfect data discussed in Sec. 4.3, and the ILPs from Sec. 4.1 are used as sub-routines.

5.1 Experimental set-up and computational cost

Our ATTACK ALGORITHM was implemented in Python 3.7 and uses the optimization package Pyomo 6.0.1 (for the solver routine), as well as the Pandas 1.3.1 package. The solver used was IBM® ILOG® CPLEX® version 12.9.0.0. The whole experimental evaluation was run in a personal computer Intel® Core™ i7-8550U CPU @ 1.80GHz with 32GB of ram memory.

Status	Count	Avg. Complexity	Avg. Solving time (s)
Unsolved	15,454	136.4	-
Solved	69,244	3.73	1.39
Total	84,698	28	-

Table 5. Statistics about solved and non-solved instances. The Avg. complexity is the simple mean of $log_{10} m^n$ across all instances.
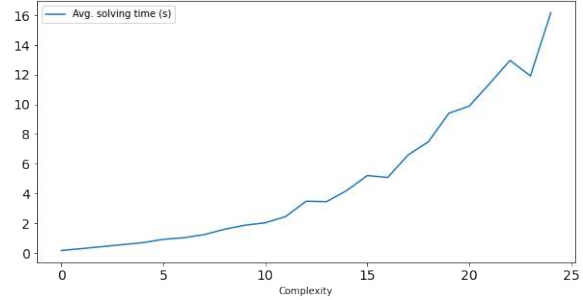

Fig. 5. Average solving time by complexity ($\log_{10} m^n$), where m is the number of bins and n the number of packages.

We set 1 minute as the threshold for considering an instance solved in reasonable time. Through all attack phases, 84,698 instances of the problem were built, and of, those, 69,244 (82%) were successfully solved within the time limit. The remaining 15,454 (18%) either failed to return a result within the time constraint or were ignored due to the errors produced by the rounding mechanism being over the threshold explained in Sec. 4.3. This took in total almost 62 hours of active processing.

Tbl. 5 shows that the average solving time (for the instances that were solved) was 1 second. In fact, more than 99% of the solved instances (around $68,000$) were solved in under 28 seconds. Finally, Fig. 5 shows how the average solving time (in seconds) increased with the instance complexity (defined here as $\log_{10} m^n$ where m is the number of bins and n the number of packages in the instance. This value gives a sense of the ILP search space, higher numbers represent bigger search spaces).

5.2 Results and analyses

After the complete execution of our attack (i. e., all 3 Phases described in Sec. 2.4), we were able to re-identify with certainty the importer's city for 138,413 transactions out of a total of 817,468 in the dataset (amounting to 6.1 billion dollars, or 40% of the total value imported in the period of interest). For 2,003 of such transactions (amounting to 137.3 million dollars, or 0.9% of the total value imported), there is only one importer in the

City	Transactions re-identified	Value (USD)
OURO BRANCO (MG)	28	24.6 mi
IPAUSSU (SP)	3	10 mi
G. DO NORTE (MT)	5	9.9 mi
SAO GABRIEL (RS)	66	7.3 mi
COMODORO (MT)	3	4.1 mi

(a) Cities with the largest values of transactions re-identified with certainty.

Attack Phase / Step	Prob. of re-identification	Mult. Leakage	Transactions re-identified	Value (USD)
Prior	0.05%	–	0	0
1/1	0.10%	16×	91	3 mi
1/2	0.44%	80×	980	99 mi
3/1	0.65%	120×	1,463	110 mi
3/2	0.90%	168×	2,003	137 mi

(b) Metrics of attack success by phase and step.

Table 6. Experimental results for the concrete case study..

recovered city. Thus, we were able to learn with high confidence sensitive details of 348 companies (2% of the total that imported something in the period of interest). Tbl. 6a shows the top 5 cities (from a total of 348) in which the total value of transactions re-identified with absolute certainty was largest.

Tbl. 6b summarizes the adversary's success in each attack phase and step. We provide the probability of re-identification of a randomly selected transaction, as well as the corresponding multiplicative leakage (i. e., the factor by which this probability increases wrt. the prior one). In particular, after the conclusion of Phase 3, Step 2, the adversary's probability of re-identification is 0.9%, corresponding to an increase of 168× over the prior risk. Although this probability appears to be small, observe that it corresponds to a significant number of transactions with a combined value of 137 mi USD. We also provide the number of transactions that can be re-identified with certainty, and these transaction values in USD. In our interpretation, it is clear that these are violations of fiscal secrecy.

6 Related work

As has been previously demonstrated in the Netflix prize [10] and AOL[18] data leaks, techniques such as

---

18 https://www.nytimes.com/2006/08/09/technology/09aol.html

the ones used by the Brazilian Government are weak against database reconstruction attacks. Even when de-identification, pseudonymization and generalization are used, the data release usually enables the construction of a set of constraints which are sufficiently restrictive as to agree with just one arrangement of the microdata. [11]

An Australian report [12] 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.

In the context of Brazil, a recent study has shown 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 [13]. 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 with disabilities per school, re-identification would be possible, although the chance of success could be lower.

7 Conclusion and prospects

In this paper we have presented a re-identification attack against a Brazilian dataset of foreign-trade statistics which successfully identified importers for transactions totalling 137 mi USD, thus violating those importers' rights to fiscal secrecy. The main real-world practical consequence of our work was to inform the Brazilian Tax and Customs Administration (RFB) of the vulnerabilities we observed, in June 17th, 2021. Later in 2021, RFB removed access to the dataset named DE-IDENTIFIED TRANSACTIONS (Ordinance 100 of December 16th of 2021). We cannot state this removal was a direct consequence of our work, but it was sufficient to avoid our attack. Also, the periodicity of the IMPORTERS dataset changed in February 7th, 2022

and a technical note[19] stated clearly that this was due to "the possible vulnerabilities to the fiscal secrecy regarding linking this dataset with other open data".

As mentioned, since Brazilian foreign-trade statics follow international standards, it is possible attacks similar to ours may still be of concern to other countries.

One immediate goal created by our work is the proposal of techniques within the realm of differential privacy and its variants to allow for the publication of data maintaining statistical utility while enforcing privacy. A primary goal should be to introduce enough uncertainty so that all importers are able to plausibly deny being responsible for any given transaction, therefore the total amount of published data cannot contain enough information so as to violate the privacy of any importer with full certainty. In the case of foreign-trade statistics, "plausible deniability" is a sufficient guarantee to deter malicious actors from making reliable inferences; the requirements for utility are yet to be fully investigated. We leave this goal as future work.

One important aspect we overlooked is the fact that any real-world attacker will make use of economic common-sense — a mining company is unlikely to import car parts, for example. Potential work in this area would be to incorporate machine learning methods for pre-processing data to provide a stronger (and perhaps more realistic) prior for the adversary against which new privacy defences could be evaluated.

In terms of the optimization modelling and attack algorithm, several research prospects unfold from this work. First, it would be interesting to implement database reconstruction attacks with tailored implementation of backtracking methods or dynamic programming, as these should be more scalable and less prone to numerical instability. The necessity to deal with errors might lead to imperatively complicated implementations. Second, the framework of integer linear programs more often than not leads to efficiently solvable approximation algorithms, which we intend to investigate both in the theoretical model of `Package Allocation` and also in the practical setting of our work. Third, and finally, the introduction of noise to the data deforms the feasibility region of the program. The inverse optimization problem of determining which direction in the feasible set is least affected by the noise can lead to very interesting considerations. Assuming noise

shall be eventually introduced to data with the intent of maximizing privacy protection and minimizing loss of utility, it might be possible to frame this task as a geometric modelling problem. We leave this investigation to future work.

Acknowledgements

References

[1] United Nations, European Commission, International Monetary Fund, Organisation for Economic Co-operation and Development, and World Bank, *System of national accounts 2008*. United Nations, 2009.

[2] W. T. Organization, "Agreement on implementation of article VI of the General Agreement on Tariffs and Trade 1994," 1994.

[3] United Nations, *International Merchandise Trade Statistics*. United Nations, 2011.

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

[5] K.-G. Lofgren, T. Persson, and J. W. Weibull, "Markets with asymmetric information: The contributions of George Akerlof, Michael Spence and Joseph Stiglitz," *The Scandinavian journal of economics*, vol. 104, no. 2, pp. 195–211, 2002.

[6] SECEX, "Revisão metodológica da contabilização dos fluxos de exportação e importação brasileira de bens," 2021.

[7] J. Kleinberg and E. Tardos, *Algorithm design*. Pearson, 2006.

[8] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, pp. 211–407, Aug. 2014.

[9] W. Bussab and P. Morettin, *Estatística Básica*. Saraiva, 2002.

[10] A. Narayanan and V. Shmatikov, "How to break anonymity of the netflix prize dataset," *CoRR*, vol. abs/cs/0610105, 2006.

[11] S. Garfinkel, J. M. Abowd, and C. Martindale, "Understanding database reconstruction attacks on public data," *Commun. ACM*, vol. 62, pp. 46–53, Feb. 2019.

[12] C. Culnane, B. I. P. Rubinstein, and V. Teague, "Health data in an open world," *CoRR*, vol. abs/1712.05627, 2017.

[13] G. Nunes, *A formal quantitative study on privacy in the publication of educational censuses in Brazil*. Master's thesis, UFMG, 2021.

---

19 https://balanca.economia.gov.br/balanca/metodologia/ Nota_publicacao_anual_da_lista_de_empresas_ exportadoras_e_importadoras.pdf

A Detailed description of the published datasets

Tables 7, 8, 9, and 10 below provide the detailed structure of of the published datasets used in our attacks, they were partially described in Sec. 2.2.

Field	Description	Example
NUMERO DE ORDEM	Sequential unique number aiding the identification of the Declaration of Imports (DI).	150320000100001
ANOMES	Year and month of transaction registry.	202011
COD.NCM	DI header: NCM code number of imported good.	42010090 (Saddlery and harness of other materials)
PAIS DE ORIGEM	DI header: country of origin.	CHINA
PAIS DE AQUISICAO	DI header: 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
DESCRICAO DO PRODUTO	DI item: detailed description of imported good, 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 USD.	8,313.44
VL FRETE DOLAR	DI header: international freight in USD.	245.23
VL SEGURO DOLAR	DI header: insurance in USD.	0.00
VALOR UN. PROD.DOLAR	Ratio TOT.UN.PROD.DOLAR / QTD COMERCIAL	0.87
QTD COMERCIAL	DI item: commercialized quantity.	4,773
TOT.UN. PROD.DOLAR	DI item: value in USD.	4,152.51
UNIDADE DESEMBARQUE	DI header: port of entrance.	NOT/INFORMED
UNIDADE DESEMBARACO	Tax and Customs Administration unit administratively treating this DI.	PORTO DE SANTOS
INCOTERM	International commercial terms of the DI.	FOB
NAT. INFORMACAO	DI header: nature of the operation.	EFFECTIVE

Table 7. Complete structure of the DE-IDENTIFIED TRANSACTIONS 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 Tax and Customs Administration 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

Table 8. Complete structure of the SUMMARY BY STATE 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

Table 9. Complete structure of the IMPORTERS dataset.

Field	Description	Example
CO_ANO	Year when the import was registered.	2021
CO_MES	Month when the import was registered.	01
SH4	Imported goods SH4 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

Table 10. Complete structure of the SUMMARY BY CITY dataset.