Bank of England

An approach to cleaning MiFID II corporate bond transaction reports

Staff Working Paper No. 1,071
April 2024

Simon Jurkatis

Staff Working Papers describe research in progress by the author(s) and are published to elicit comments and to further debate. Any views expressed are solely those of the author(s) and so cannot be taken to represent those of the Bank of England or to state Bank of England policy. This paper should therefore not be reported as representing the views of the Bank of England or members of the Monetary Policy Committee, Financial Policy Committee or Prudential Regulation Committee.

Bank of England

Staff Working Paper No. 1,071

An approach to cleaning MiFID II corporate bond transaction reports

Simon Jurkatis[1]

Abstract

Since 2018, EU and UK financial markets regulators have been in receipt of data on transactions in debt instruments, such as corporate bonds, reported under the Markets in Financial Instrument Regulation. The data gives regulators a more detailed and broader view of trading in these instruments than previously. Reports submitted under this framework, however, come with a number of unique challenges that require careful consideration. Among those challenges are that reports are not submitted in a completely standardised way, that prices and quantities can be reported in different units, and that reports may be submitted by both counterparties of a transaction. This paper describes an approach for handling these issues for transaction reports on corporate bonds, with the aim of helping to enhance the data quality and supporting robust research into this market.

Key words: Corporate bonds, data cleaning, deduplication, outlier detection.

JEL classification: C55, C58, C81, G10.

[1] Bank of England. Email: simon.jurkatis@bankofengland.co.uk

The views expressed in this paper are those of the author, and not necessarily those of the Bank of England or its committees. I am grateful to Marcus Buckmann, Chiranjit Chakraborty, Gabor Pinter, Nick Vause and Jack Worlidge for their help, comments and stimulating discussions.

The Bank's working paper series can be found at www.bankofengland.co.uk/working-paper/staff-working-papers

Bank of England, Threadneedle Street, London, EC2R 8AH
Email: enquiries@bankofengland.co.uk

ISSN 1749-9135 (on-line)

1 Introduction

This manual outlines an approach to cleaning corporate bond transaction data reported under the Markets in Financial Instruments Regulation (EU) No 600/2014 as on-shored and set out in Article 26 of UK MiFIR, and made available to the Bank of England (BoE) by the Financial Conduct Authority (FCA). General information on the regulatory framework and reporting requirements can be found on the website of the FCA and in the reporting guidelines on transaction reporting by the European Securities and Markets Authority (ESMA).

This document is intended to serve three purposes. Firstly, to provide additional detailed information on the cleaning approach that underlies the analysis in a recent working paper by Jurkatis et al. (2023). This should help others to understand, reproduce and build on our results. Secondly, I hope that this document will be useful to other regulators in other jurisdictions with access to MiFID II transaction reports and that it may provide a basis for discussions on further improvements to the handling of this complex and incredibly rich data set. Finally, while the data discussed in this paper is confidential regulatory data whose access is restricted to EU and UK central bankers and financial market supervisors, both the EU and UK are in discussions to introduce a consolidated tape for fixed income instruments with transaction-level information on prices, volumes and transaction time.[1] I therefore hope that this document may provide some helpful guidance on how to work with that data, once it becomes available — even though it can be expected that the reporting requirements and the layout of the data will differ.

Caveats. Before proceeding to the details of the approach, the reader should take note of two caveats. First, the cleaning approach is tailored to suit corporate bond transaction reports. The corporate bond market is characterised by infrequent trading. This makes the identification of price and quantity outliers particularly challenging. Most bonds will have only few observations available to make a statistical assessment of what an unreasonably high or low value may be based on a bond's price and quantity distribution. At the same time, fewer observations

[1] https://www.fca.org.uk/publications/consultation-papers/cp23-33-payments-data-providers-drsp-policy-statement-framework-consolidated-tape-cp23-15.

make the identification of duplicates (reports on the same transaction reported by its different counterparties) easier as the smaller sample size per bond allows for more sophisticated, but computationally more demanding methods. This contrasts with data on, for example, government bonds, which trade more frequently, so the identification of duplicates is more demanding, but the identification of price and quantity outliers is easier and more reliable. Hence, different techniques for different data sets are more likely to achieve optimal results, and the approach presented here should not be blindly applied to other asset classes.

Second, the approach has been designed to accommodate the ongoing growth of the data since the implementation of the MiFID II framework in Jan 2018. That is, the approach is applied to batches of the data as they come into the Bank of England, independent of previous batches. This means that the identification of price and quantity outliers makes only limited use of historical data.

The main steps of the cleaning process include normalizing reports to have a common format; identifying price and quantity outliers; identifying duplicate reports; and correcting entries, including identified outliers, where possible. The rest of the paper is organised along those lines. Section 2 describes the selection of the data and steps to normalize the reports. The normalization is particularly helpful for the identification of duplicates. Section 3 presents the approach to identifying price outliers and describes steps to correct those outliers where possible. Section 4, in turn, presents steps to identify quantity outliers. Section 5 outlines the procedure to link together duplicate reports and describes the procedure to correct erroneous entries based on the information in both duplicate reports. Finally, Section 6 offers some concluding remarks.

2 Data preparation

2.1 Selecting relevant reports

While the cleaning approach is designed to handle corporate bond transactions, MiFID II transaction reports do not directly contain an identifier for this security class. The reported instrument is identified by its International Securities Identification Number (ISIN) and, in addition, is classified by its CFI (Classification of Financial Instruments) code.

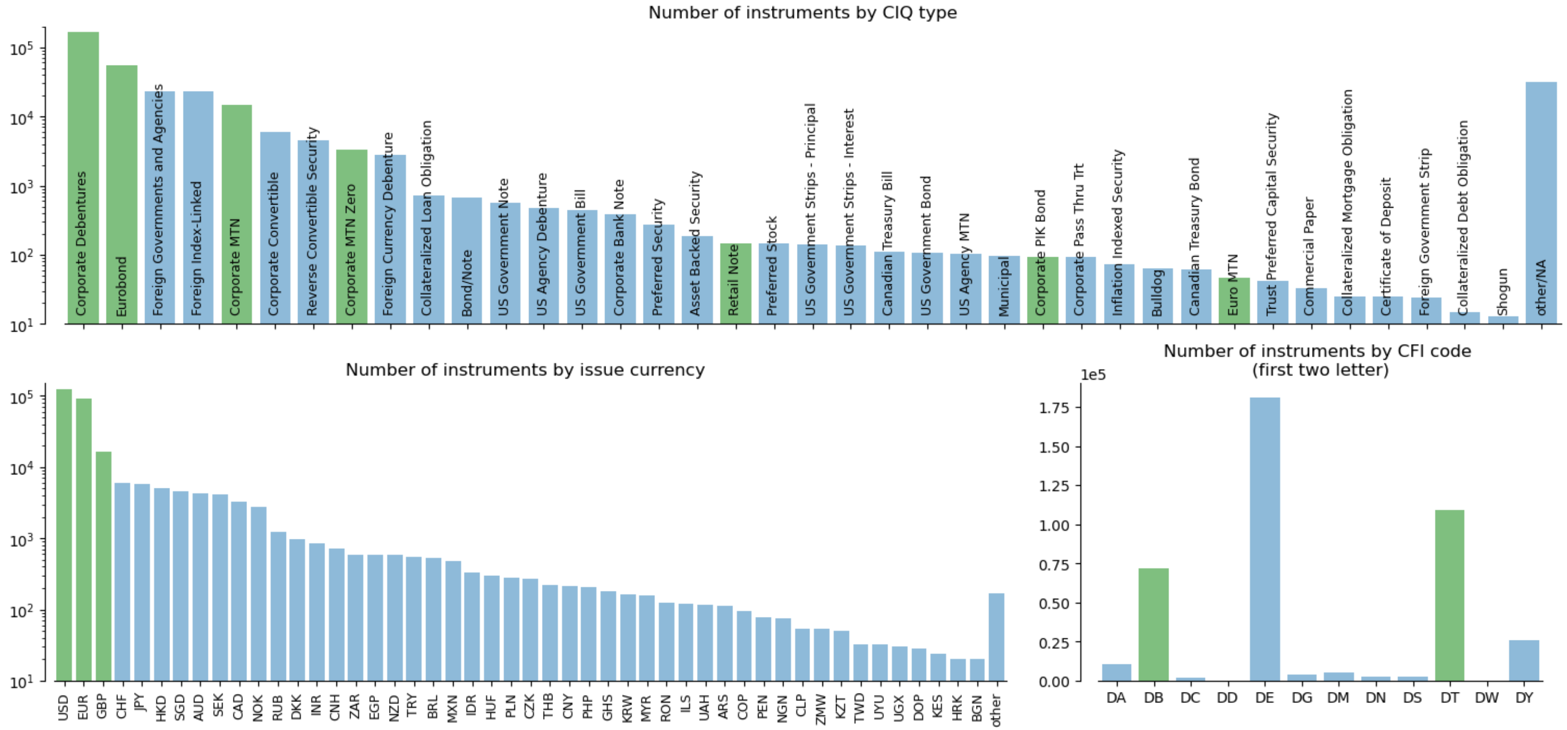
The first two letters of the CFI code could in principle be used to identify bonds (CFI code starting with DB, where D denotes debt debt instrument and B denotes bond), but this is insufficient to identify instruments that would generally be considered to be corporate bonds. I therefore use S&P Capital IQ (CIQ) data in conjunction with the CFI code to filter out reports on corporate bonds from the overall data. Figure 1 shows the distribution of CIQ security types, CFI codes and currencies across the full sample of instruments.

Among all instruments with a CFI code starting with 'D', I select those that are classified as 'Corporate Debentures', 'Corporate PIK Bond', 'Retail Note' or 'Corporate Zero' according to CIQ. In addition, I select all instruments that are classified as 'Eurobond', 'Corporate MTN', 'Corporate MTN Zero' or 'Euro MTN' if their CFI code starts with 'DB' or 'DT' (where T denotes medium-term note).

I further restrict the sample to bonds issued in Sterling (GBP), US Dollar (USD) or Euro (EUR), which are the three most prevalent currencies in the sample (as of 1 Oct 2023), and to bonds of issuers that are either public or private companies according to CIQ.

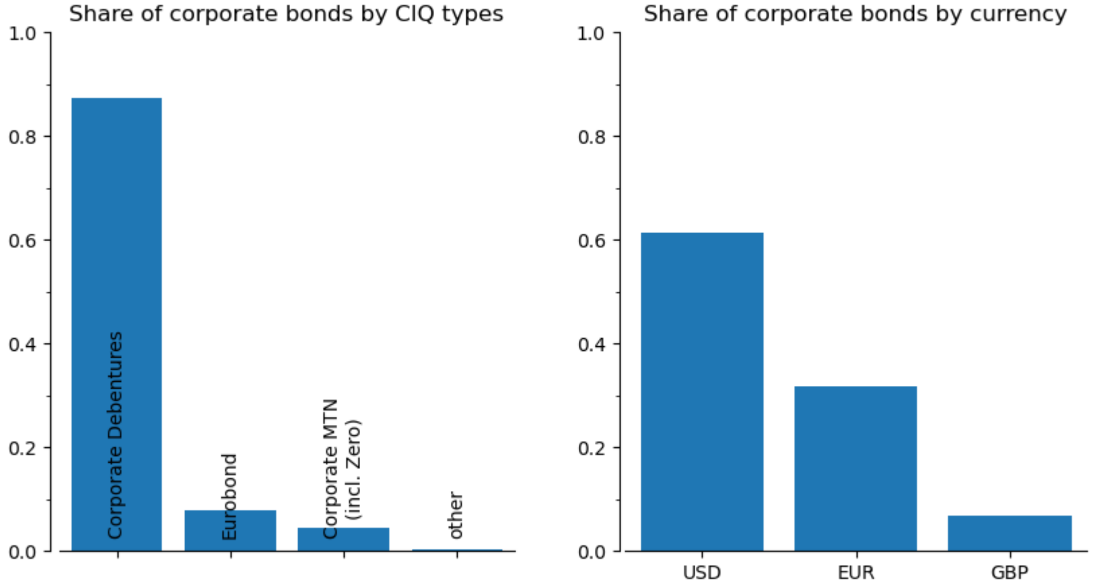
Figure 2 shows that the set of selected corporate bonds consists largely of Corporate Debentures (87%) and Eurobonds (8%), which are also the most frequent types in the overall sample. Most corporate bonds are issued in USD (61%), followed by EUR (32%), with only a small fraction issued in GBP (7%). The total number of reports that satisfy the selection criteria is 59,861,537 as of 1 Oct 2023, distributed across 144,962 different bonds.

Figure 1: Instrument types and issued currency across the full sample



Notes: The top graph shows the number of different instrument types according to S&P Capital IQ (CIQ) across the sample of transaction reports in debt instruments (according to their CFI code) for the period 3 Jan 2018 to 1 Oct 2023. Only types with at least 10 instruments are shown. The bottom-left graph shows the number of instruments by issued currency across the same sample. Only currencies with at least 20 instruments are shown. The bottom-left Panel shows the number of instruments according to the first two letters of their CFI code. Green bars show the selected types.

Figure 2: Distribution of instrument types and issued currency for the selected sample of bonds



Notes: The left graph shows the share of different S&P Capital IQ (CIQ) instrument types for the selected sample of corporate bonds over the period 3 Jan 2018 to 1 Oct 2023. The right graph shows the share of instruments by issued currency across the same sample.

2.2 Correcting internal account flags

When a firm fills several client orders with one or more market-side transactions on a riskless agency basis, i.e. without taking any inventory risk, reports on these transactions have to be linked with an "internal account" flag (see Section 5.23 in the ESMA Guidlines ESMA/2016/1452 and examples therein). This flag is especially useful for identifying such transactions when they are spread over time. For instance, the internal account flag is used to identify riskless agency trades in Jurkatis et al. (2023) (as part of 'matched trades', which are described in Section 5).

Given the usefulness of the flag, I check that it has been reported correctly. This is possible because transactions by an executing entity that are linked with the internal account flag must net to zero at the end of the day. That is, the amount bought into and sold out of the internal account by an executing entity must be the same at the end of the day for every bond where the internal account flag has been used.

For every executing-entity-bond-day, I therefore filter the cases where the ac-

Table 1: Correcting internal account reports

	`executing`	`buyer`	`seller`	`quantity`	`time` ···
original	E	E-INTC	C1	100	14:43:22 ···
⇐ original	E	C2	C1	100	14:43:22 ···
⇒ corrected	E	C2	E-INTC	100	14:43:22 ···

Notes: This table shows a generic example where an executing entity uses the internal account flag incorrectly. The seller in the second report should have been reported as the executing entity trading out of its internal account.

count flag does not net zero and check if the internal account flag has accidentally not been used on one of the sides (buy or sell). I take this to be the case when the executing entity reports an internal account on one side, e.g. buying into the internal account, but then reports having matched the same client to another client at the same quantity and time, as shown in the top two rows of Table 1.[2]

In all such cases I insert the internal account flag accordingly (for instance, by substituting the bottom row of Table 1 for the middle row) and verify that the corrections have led to internal account trades to net to zero at day end. In all cases where the internal account flags do not net to zero, all flags are removed from the transaction reports of the corresponding executing entity for the respective bond-day.

Around 12% of reports contain an internal account flag and in around 90% of them the internal account seems to have been applied correctly. Of the 10% erroneous reports 4% could be corrected.

2.3 Report normalization

To enhance the identification of duplicates, I normalize reports to a common standard. Normalization of reports is applied in two ways: splitting single reports into two, and aggregating two or more reports into one.

[2] Note that while the table displays the internal account flag as a suffix to the reported counterparty identifiers, we actually store the information in a separate column. This is necessary for the identification of duplicates discussed later.

Table 2: Report normalization - splitting matched reports

	`executing`	`buyer`	`seller`	···
original	E	C1	C2	···
⇒ normalized {	E	E	C2	···
	E	C1	E	···

Notes: This table shows a generic example of a report where the executing entity matches two clients directly. Such reports are split in two with the executing entity as the buyer in one report and as the seller in the other.

Splitting reports. Transactions where an executing entity matches two counterparties can be reported in two ways: either as two reports where the executing entity buys from one client in one report and sells to the other client in the other report, or as a single transaction where the executing entity reports one client as the buyer and the other as the seller, as shown in the first row of Table 2. Around 15% of all reports in corporate bonds are of the latter type.

Reports set out in this way complicate the identification of duplicate reports (if such a duplicate exists). To see this, imagine a report where executing entity E matches clients C1 and C2 and reports the transaction as shown in row one of Table 2. If C1 also reports the trade, not being aware that it was matched to C2, it would report its counterparty E as the seller. With diverging information in the seller field across the reports submitted by E and C1, it would be more complicated to identify those reports as duplicates of the same transaction. Imagine further that both clients, C1 and C2, report the transaction. In this case the report by the executing entity E would be a duplicate of two reports, and care would be needed to eliminate multiple duplicates for subsequent analyses.

Therefore, to increase the probability of finding duplicate reports and to ensure that each report has at most one duplicate, I split reports in which an executing entity matches two counterparties into two as shown in row 2 and 3 of Table 2.

Aggregating reports. Firms occasionally report to have traded with the same counterparty multiple times at the exact same time and price in the same bond. Around 9% of all reports fall into this category. Such reports are aggregated into

Table 3: Report normalization - aggregating reports

	executing	buyer	seller	quantity	price	time	···
original	E	E	C	100	98.95	09:45:3.100	···
original	E	E	C	50	98.95	09:45:3.100	···
⇒ normalized	E	E	C	150	98.95	09:45:3.100	···

Notes: This table shows a generic example where the executing entity buys the same bond from the same counterparty at the same price at the same time. Such reports are aggregated into one report by adding up the quantity across the multiple reports.

a single report by summing up their quantities as shown in Table 3. Doing so helps to identify duplicate reports where one counterparty reports a number of smaller transactions while the other counterparty reports a single transaction with the total transacted volume.

3 Price outliers

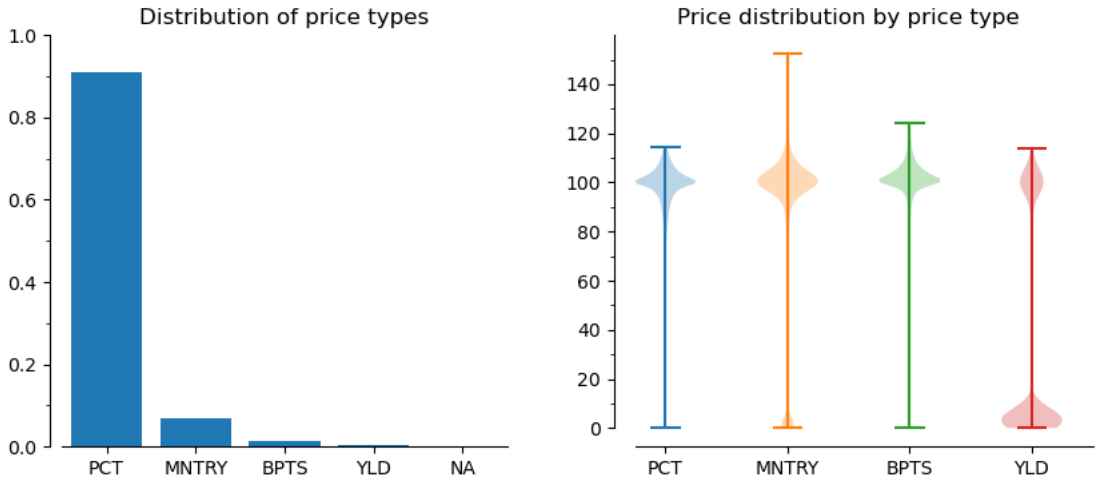
3.1 Reporting standard

The reporting requirements allow for prices to be reported in different types: percentage of par (PCT), yield (YLD), monetary (MNTRY) or basis points (BPTS).[3] As most prices are reported in PCT (and a large number of prices appear to be in PCT even when they are reported to be of another type), I treat prices generally to be reported as PCT and do not attempt to convert them when they are reported to be of a different type. Moreover, prices have to be reported as *clean*, meaning without the accrued interest.

The left panel of Figure 3 shows that around 91% of prices are reported to be PCT, followed by MNTRY with 7%. Prices reported in BPTS are just above 1% of the sample and less than that are reported in YLD. A very small fraction of reports (0.1%) do not have a price type. These cases refer to reports where prices are missing, usually because they are reported as not applicable or pending.

[3] The monetary price is value in the respective currency per contract exchanged, i.e. percentage-of-par price times the par value of the bond.

Figure 3: Price types and price distribution



Notes: The left panel shows the distribution of price types across reports in corporate bonds between 3 Jan 2018 and 1 Oct 2023. The 'NA' values refers reports where prices are missing, typically because they are pending or not applicable. The right panel shows the distribution of prices across reports with the same price type over the same sample period. Prices below zero and above the 95th quantile have been dropped form the illustration.

The right panel of Figure 3 shows the distribution of prices by price type. It shows that for transactions with reported price types other than PCT, the reported price is often more similar to a PCT value than a value that would be expected for the other price types.

3.2 Identifying price outliers

The identification of price outliers proceeds in two steps. First, all prices that are below 20 or above 500 are labelled as outliers. The reason is that values outside these bounds are unreasonable if quoted in percentage of par and indistinguishable from possible values of other price type such as yield (if below 20) or monetary (if above 500). The percentage of all reports flagged in this way is 1.6%, with the largest percentage concentrated in YLD (67%) and MNTRY (14%). Prices outside the range of 20 to 500 are less than 1% for prices reported in PCT and BPTS.

Prices that are within the bounds of 20 to 500 are then separately processed in a second step to determine whether they should be flagged as outliers.

For any given bond-day, I determine lower l_{bt} and upper u_{bt} as

$$l_{bt}, u_{bt} = m_{bt} \pm \begin{cases} \lambda & \text{if } 2 \leq n_{bt} < 10 \\ \max\{IQR_{mb} \times 5, \lambda\} & \text{if } n_{bt} \geq 10 \end{cases} \tag{1}$$

where m_{bt} is the median price of bond b on day t (among all prices within the fixed bounds of 20 and 500), IQR_{mb} is the corresponding inter-quartile range, and n_{bt} is the number of unique entity-price observations.[4] Prices that are outside these thresholds are labelled as outliers.

The parameter $\lambda \in [5, 15]$ is calibrated to achieve an outlier percentage of 0.5% over the given sample. The target outlier percentage is set relatively low as the percentage applies only to the observations within the strict bounds of 20 and 500 and comes on top of the outliers already flagged in the first step.[5]

For bond-days with more than one but less than ten observations, thresholds are discarded and all observations are treated as outliers if the number of outliers was more than 4 or 50% of the observations. This is because a high number of outliers may prevent reliable estimation of the median in the first place.

3.3 Correcting price outliers

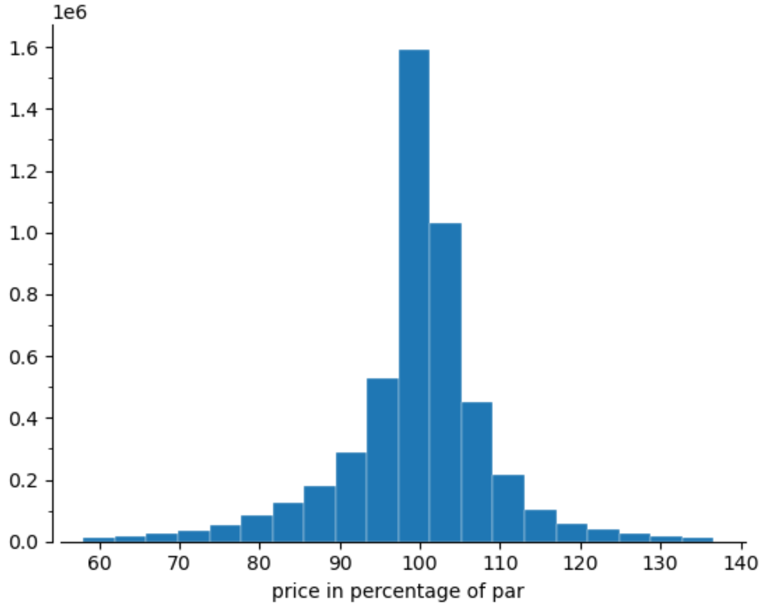
All prices flagged as outliers based on the thresholds l_{bt} and u_{bt} are checked to see if they can be corrected. This is done by moving the decimal point of the outlier price such that the number of digits before the decimal point (excluding leading zeros) corresponds to the number of digits before the decimal point of both thresholds. If the new price resulting from that manipulation falls within the thresholds, the new price is kept and the outlier label removed.

For example, consider the thresholds 99.75 and 101.05 and a reported price of 0.101, which is flagged as an outlier. The reported price would first be multiplied by 100 such that the number of digits before the decimal point corresponds to those of the lower bound. As the transformed price of 10.1 would not fall within

[4] Prices that are the same for a given bond-day and reported by the same executing entity are counted only once.

[5] The data is obtained by the Bank of England on a weekly basis (with the exception daily data updates if required). The data is therefore processed week by week, including the calibration of λ. In 86% of cleaning runs, λ takes its lower bound of 5.

Figure 4: Inlier price distribution



Notes: The figure shows the distribution of prices flagged as inliers across a random selection of 5,000,000 reports in corporate bonds between 3 Jan 2018 and 1 Oct 2023. The distribution is truncated at its 1st and 99th quantiles.

the thresholds the transformed price is ignored. Next, the original price would be multiplied by 1000 such that the number of digits before the decimal corresponds to those of the upper bound. As the transformed price of 101 falls within the thresholds, the reported price would be replaced with the new one and the outlier flag dropped.

Figure 4 displays the distribution of inliers for a random sample of 5,000,000 reports between 3 Jan 2018 and 1 Oct 2023. The distribution is truncated at the 1st and 99th quantile reflecting common practice in the finance literature. The figure shows that prices are distributed as a bell-curve centered at 100 and ranging from roughly 60 to 140.

4 Quantity outliers

4.1 Reporting standard

Similar to prices, quantities can be reported in different types: units or nominal/monetary. Since most quantities are reported in nominal amounts and because values reported in units could in principle also be valid nominal values, I treat all reported values as nominal. Inspection of the distribution of quantities reported in units suggests that many of them are indeed nominal quantities. Multiplying values reported in units by the bonds' par values to convert them to nominal often leads to unreasonably large numbers.

Figure 5 shows that 85% of quantities are reported as nominal (NMNL), 13% as monetary (MNTRY), and 2% as units (UNIT). The distribution of quantities in NMNL and MNTRY are very much the same suggesting that these labels are used inter-changeably. Also the distribution of quantities reported as UNIT has a large overlap with the other distributions, despite having a longer tail of smaller values, suggesting that these values are also often reported as NMNL/MNTRY.

One way to check the validity of the reported quantity would be to compare it to the reported 'net-amount' which is defined as the nominal quantity times the dirty price (i.e. clean price price plus accrued interest). However, if the comparison fails it may be because any one of the three variables is reported erroneously. In fact, it seems that the net-amount is frequently misreported and I therefore do not use the triplet of net-amount, price and quantity to evaluate the validity of any of its components.

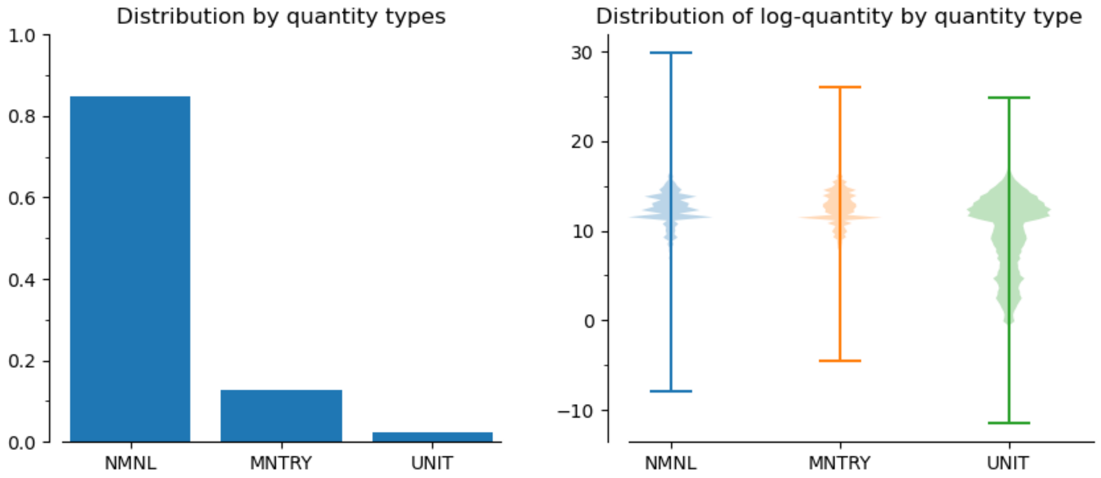
4.2 Identifying quantity outliers

To identify quantity outliers, I merge the transaction reports with information from CIQ or, if not available there, from ESMA and FCA reference files. Specifically, I obtain data on each bond's issuance date, issuance size (i.e. the nominal amount issued), and the currency in which the bond was issued.

Identification of the outliers then proceeds in different steps depending on the time of the trade relative to the issuance date. This is because bonds are typically more heavily traded and more liquid close to the issuance date, so the trade-size distribution may have a larger average and heavier tails around those dates.[6]

[6] Note that I shift weekend observations to the previous Friday as weekend observations are typically too few to base the outlier assessment on those observations alone.

Figure 5: Quantity type and quantity distribution by type



Notes: The left panel shows the distribution of quantity types across reports in corporate bonds between 3 Jan 2018 and 1 Oct 2023. Possible quantity types are nominal (NMNL), monetary (MNTRY), and unit (UNIT). The right panel shows the distribution of the logarithm of quantities across reports with the same quantity type. The distributions are displayed across all observations for types MNTRY and UNIT. The NMNL observations are down-sampled (randomly) to the number of MNTRY observation to reduce the computational burden in producing this graph.

Quantities smaller or equal to 0. First, any quantity smaller or equal to zero is flagged as an outlier. So far, no quantity has been reported that is smaller/equal zero.

Before the issuance date. (≈3% of reports.) The reporting requirements state that quantities exchanged as part of the issuing process have to be reported (see Question 8 of the ESMA Q&A on MiFIR data reporting). It is therefore possible to observe reports that have a trade execution timestamp that dates before the date the bond was issued. Across those reports, quantities that are larger than the amount issued are flagged as outliers.

Issuance date +1. (≈1% of reports) In this step, I focus on reports with timestamps that indicate that the trade was executed on the day of the issuance or one day after. First, all quantities across those reports are divided by the issued amount, $q := Q/\text{amount-issued}$, where Q is the reported quantity and amount-issued is the nominal issuance size of the bond. Next, I determine the 99.5th quantile of the distribution of q across all reports with $0 < q < 1$. All reports with a q greater or equal to that quantile are flagged as outliers.

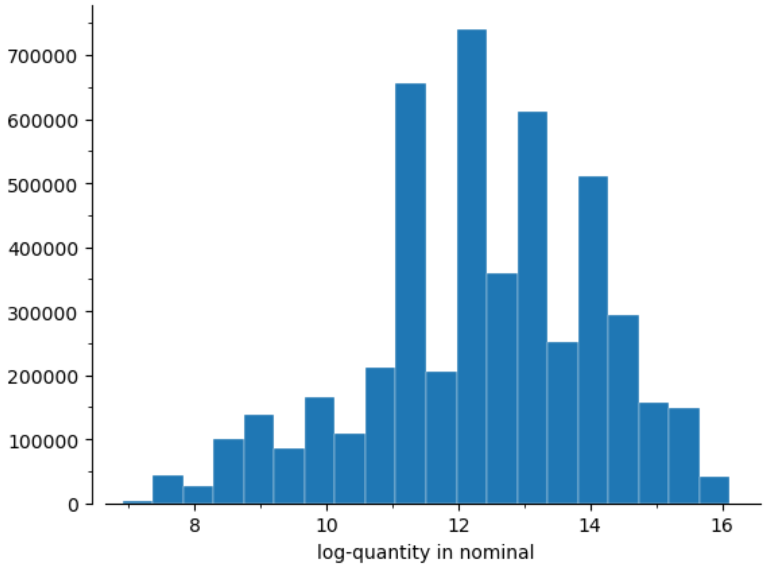
After issuance date +1 or no issuance date information. (≈94% and 2% of reports respectively) In this step, I focus on all trades executed at least two days after the date of issuance and all trades for which I had no information on the issuance date of the bond. Again, for each such report I compute the ratio of reported quantity over amount issued, q. I then determine the 99.5th quantile of the q distribution across all reports with $0 < q < 1$ for each date separately. All reports with a q greater or equal to the 99.5th quantile for the given day or with a q greater than 70% are flagged as outliers.

No issuance size information or traded in a currency other than the issued currency. (<1% of reports) The data contain a field on the currency of the traded quantity. In most cases the field is empty in which case it is assumed that the quantity was traded in the currency in which the bond was issued. Occasionally, however, the field is populated in which case the currency may deviate from the issued currency. For these that trade in a different currency to that of the original issuance or with no issuance size information, the handling process again has different steps for (a) before the issuance date, (b) on the issuance date plus the next day, and (c) at least two days after the issuance and reports where the information on the issuance date of the corresponding bond was not available.

For category (a), I determine the 99.5th quantile across all quantities reported in this category as well as all quantities previously processed under the "before the issuance date" category that were not flagged as outliers. This is done for each currency separately. Quantities that are greater or equal to their currency-specific 99.5th quantile are flagged as outliers. The process for category (b) is equivalent, using the 99.5th currency-specific quantile across reported quantities in this category as well as all non-outlier quantities previous processed under the "issuance date +1" category. For category (c), the process is the same as in (b) with the only difference that the 99.5th quantile is currency-*date* specific.

In total, only around 0.6% of quantities are flagged as outliers. The distribution of inliers for a random sample of 5,000,000 reports is displayed in Figure 6. The distribution is truncated at the 1st and 99th quantile. The figure shows a jump in the distribution close to 100,000 (around 11.5 in log). This is indeed a common trade-size occurring in around 11% of reports. Other values around which the

Figure 6: Inlier quantity distribution



Notes: The figure shows the distribution of the logarithm of nominal quantities flagged as inliers across a random selection of 5,000,000 reports in corporate bonds between 3 Jan 2018 and 1 Oct 2023. The distribution is truncated at its 1st and 99th quantiles.

distribution clusters are 200,000 (12.2 in log), 500,000 (13.1 in log) and 1m (13.8 in log).

5 Duplicate reports

5.1 Identifying duplicates

For a number of transactions, though not all, both counterparties have to report the trade. In many types of analysis, it is important to have such duplicate reports linked or removed, for example, to not over-estimate overall transaction volume.

Ideally, such duplicate reports would match on the buyer, seller, date-time, instrument-code, price, quantity and venue fields. Unfortunately, many reports that are very likely to be duplicates disagree on one or more of those fields. For example, one side may erroneously report the counterparty as the parent company of the counterparty or another child of the same parent company, or one counterparty may erroneously report the local time instead of the Coordinated Universal Time (UTC).

To help identify such disagreements across duplicate records, I follow a probabilistic record linkage approach (Fellegi and Sunter, 1969; Grannis et al., 2003). Consider the set of all reports in a given bond on a given day, $\mathcal{R}$. First, I compare every report on that bond-day to every other report of the same bond-day, excluding reports by the same executing entity, to obtain a vector of agreement for each report pair:

$$x_r = \left(x_r^{(1)}, \cdots, x_r^{(K)}\right)' \tag{2}$$

$$\text{with} \quad x_r^{(k)} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ agree on field } k \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

$$\text{and} \quad r := (i,j) \in \mathcal{R}^2 = \mathcal{R} \times \mathcal{R}\, s.t.\, e_i \neq e_j, \tag{4}$$

where e_i is the executing firm in report i and e_j the executing firm in report j.

The vector shows to which extent two reports agree on buyer, seller, time, price, quantity and venue. For the buyer, seller, price and quantity fields, the comparison is strict, meaning, $x_r^{(k)}$ takes the value 1 if and only if the values reported by the two counterparties are exactly the same. The reported trade time is floored to the second and allowed to deviate by one hour across the two reports and still be considered in agreement.[7] The venue field contains the market identifier code (MIC) if the transaction was executed on a Regulated Market, a Multilateral Trading Facility, an Organized Trading Facility or a Systematic Internalizer. Otherwise, the field is populated with 'XOFF'. The venue values of a given pair of reports are considered to be in agreement when the Levenshtein distance (which is a popular metric for comparing the similarity of strings) between them is at least 0.95. Here, the Levenshtein distance is normalized to be between 0 and 1, taking the value one if two strings match exactly.

The agreement vector x_r is then treated as a multinominal variable with different probabilities of 'success' in any of its k categories depending on whether the

[7] Timestamps are floored to the second since some reports have to provide the timestamp only to the second, while others have to provide millisecond precision. I allow for a tolerance of one hour because time should be reported in UTC which, however, is not always adhered to and British summer time differs from UTC time by one hour.

report pair is truly a pair of duplicates or not:

$$P(x_r|d_r = 1) = \prod_k m_k^{x_r^{(k)}}(1 - m_k)^{(1-x_r^{(k)})}$$
$$P(x_r|d_r = 0) = \prod_k u_k^{x_r^{(k)}}(1 - u_k)^{(1-x_r^{(k)})}$$

where

$$d_r = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are true duplicates} \\ 0 & \text{otherwise} \end{cases} \qquad \text{(unobserved)}$$

with priors

$$P(d_r = 1) = p$$
$$P(d_r = 0) = 1 - p.$$

The likelihood of the data $\{x_r\}$ given parameters $\theta = (m_1, u_1 \ldots, m_K, u_K, p)$ is then given by

$$\begin{aligned}\mathcal{L}(\theta) :=& \log P(\{x_r\}|\theta) \\ =& \sum_r \{d_r[\log P(x_r|d_r = 1) + \log p] \\ & +(1 - d_r)[\log P(x_r|d_r = 0) + \log(1 - p)]\}\,.\end{aligned}$$

Since $\{d_r\}$ are unobserved, the maximization of the likelihood has a missing data problem as encountered, for example, in the estimation of Gaussian mixture-models, which can be solved using the Expectation-Maximization algorithm (Dempster et al., 1977). By applying this algorithm, I obtain maximum-likelihood estimates of the parameters and, for every report pair, the posterior probability that reports i and j are duplicates given their agreement vector, i.e. $P(d_r|\hat{\theta}, x_r)$.

However, a report can be the duplicate of at most one other report. To reduce duplicate assignments across reports to a one-to-one mapping I use a method from graph theory, maximum-weight-matching (Galil, 1986). The assignment of reports to each other can be represented as an undirected graph, where each node (i.e. report) links to every other node except where both reports reference the same executing entity. The links (or edges) are weighted according to the estimated probabilities that the reports are duplicates. The maximum-weight-matching algorithm finds the one-to-one assignment that maximizes the sum of weights across the edges. Before the application of the algorithm, I remove all the edges with weights of less than 0.5 (i.e. effectively setting the duplicate probabilities of less than 0.5 to zero) to reduce the computational burden of the exercise.

Figure 7 (left panel) shows that 45% of reports are identified as duplicates. The middle and right bars in this chart suggest we should not expect every report to be assigned a duplicate. Around 5% of reports are the only report for a given bond on a given day, and around 41% of all bond-days see only a single executing entity reporting. In all these cases, there are no other reports that could be assigned as a duplicate.

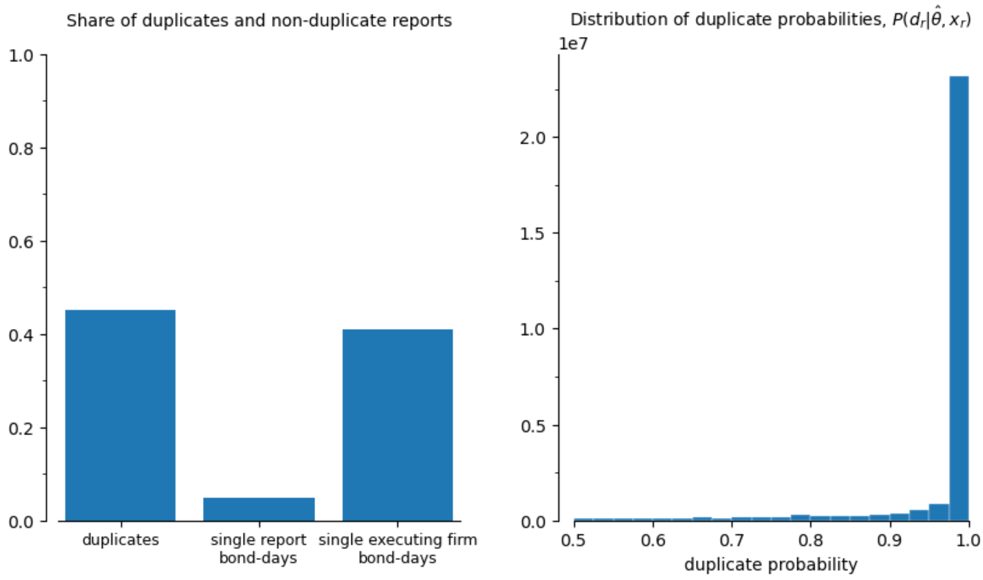
The right panel of Figure 7 shows the distribution of estimated probabilities. The estimated probabilities are generally very high, close to 1, meaning the assigned duplicates generally match well across fields and that we can take confidence in the assignments.

The assignment of duplicates and their corresponding probabilities are then added to the data. Each duplicate pair is assigned a unique ID. Depending on the analysis, one may not want to drop duplicate reports or drop a specific report (since reports contain a number of fields that are specific to the executing firm). Moreover, duplicate reports can be used to correct previously flagged outliers and to correct other erroneous entries, as outlined in the next section.

5.2 Correcting entries based on duplicates

Given that the method to identify duplicates allows for disagreement between the content of two matched reports, I can use duplicates to correct entries in the case of such disagreements. This includes misreported counterparties, price and quantity outliers flagged in one report but not the other, and verification of price and quantity outliers (a value that is flagged as an outlier but is consistent across two reports by different executing entities may not be an outlier after all.)

Figure 7: Identified duplicates and duplicate probabilities



Notes: The left panel shows the share of reports identified as duplicates of another report (left column), the share of reports where the report is the only report for a given bond on a given day (middle column), and the share of bond-days with only one executing entity being reported. The right panel shows the distribution of estimated duplicate probabilities. The sample in both panels spans the period 3 Jan 2018 to 1 Oct 2023 for all corporate bond reports.

The process of correcting outliers and resolving disagreements across duplicates is described below.

Correcting prices. Prices that are flagged as an outlier in one report but not the other are replaced with the non-outlier value. If prices are flagged as outliers in both reports, but both prices are the same, the outlier flag is removed from both reports.

Correcting counterparty information. Consider two reports identified as duplicates but with diverging information in the buyer and seller field as shown in Table 4. In Example 1, firm X is the buyer and reports W to be the seller. At the same time, firm Y is the seller and reports firm Z as the buyer. Since these reports are linked as duplicates with a sufficiently high probability, it must be that they matched closely or exactly on the other fields, including the venue and time of the execution, the price and quantity. This provides us with some confidence that the executing firms simply misreported the counterparties. Typically, such inconsis-

Table 4: Duplicate reports with inconsistent counterparty information

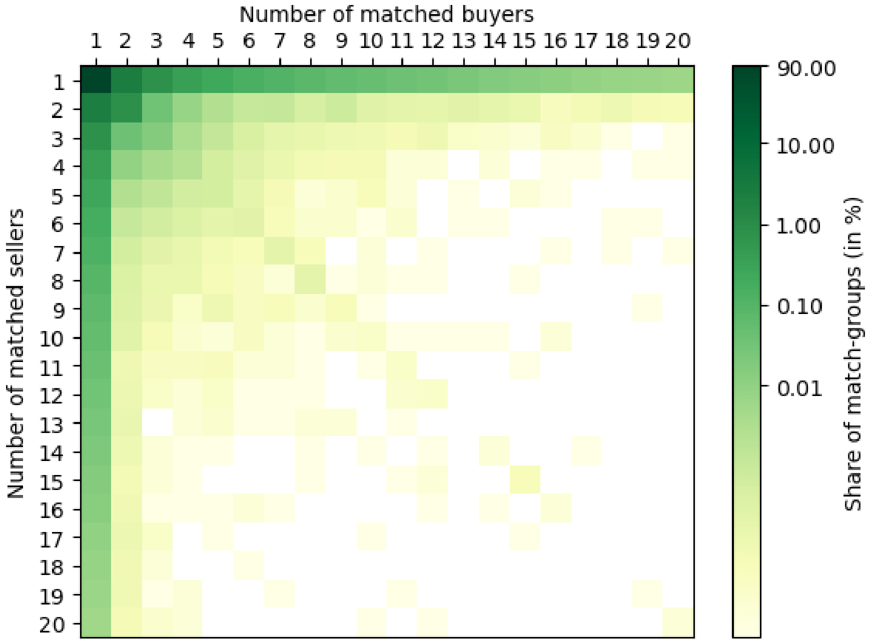
	`executing`	`buyer`	`seller`	`dupid`	⋯
Example 1	X	X	W (Y)	1	⋯
	Y	Z (X)	Y	1	⋯
Example 2	X	X	Y	2	⋯
	Y	Z (X)	Y	2	⋯
Example 3	X	X	W	3 (NaN)	⋯
	Y	W	Y	3 (NaN)	⋯
Example 4	X	X	Z	4 (NaN)	⋯
	Y	Y	Z	4 (NaN)	⋯

Notes: This table shows some generic examples of reports identified as duplicates with inconsistent counterparty information. The content in parentheses displays the correction applied to such report pairs. In Examples 1 and 2, inconsistent counterparty information is corrected. In Examples 3 and 4, the duplicate assignments are removed.

tencies happen when one firm reports the parent or another related company of its actual counterparty. Such misreporting is not always as severe as in Example 1. More often, it would only be one firm misreporting its counterparty, as shown in Example 2. Where such inconsistencies arise, I replace the counterparty identifier in the field that is not the same as the executing entity with the identifier of the executing firm of the other report. That is, W in Example 1 is replaced by Y, and Z is replaced by X.

It is possible that reports are linked as duplicates with inconsistent counterparty information which, on close inspection, do not appear to be duplicates after all, even though they matched well on the other fields. Such instances are shown in Examples 3 and 4 of Table 4. In Example 3, both X and W report the same counterparty, W, which indicates that W probably traded with X and Y at a similar time on similar terms, effectively intermediating between them. Since W's reports are missing (possibly because W does not have any reporting obligations), the reports by X and Y were linked by mistake. Similarly, in Example 4, it seems more likely that Z traded with both X and Y, rather than Z being misreported by those firms. In such cases, the counterparty information remains unchanged and

Figure 8: Distribution of number of matched counterparties



Notes: This figure shows the distribution of the number of matched sellers (y-axis) and matched buyers (x-axis) across all matched groups of corporate bond reports over the period 3 Jan 2018 to 1 Oct 2023. The figure is truncated to show a maximum of 40 matched counterparties.

the duplicate ID is removed from the pair of reports.

Correcting quantities. Reported quantities across duplicates may disagree. Where this is the case, one counterparty typically reports a quantity that is a factor larger than that reported by the other counterparty (often 2, 100 or 1000).[8] If one of the duplicate reports is flagged as an outlier and its quantity is a factor larger than the quantity in the other reports, the outlier quantity is divided by that factor and the outlier flag removed. If both reports are marked as outliers, but have the same value, they are no longer considered as outliers.

This correction procedure is applied to all trades that were previously identified as belonging to the same match-group:

Definition (Matched trades)**.** *A match-group is defined to be a set of reports in which the executing entity is buying and selling the same instrument at the same time, as shown in first three rows of Table 5, or trades linked by an internal account*

[8] The factor 2 can arise from the aggregation of reports when one counterparty mistakenly reports the same transaction twice.

Table 5: Match-group, duplicates and quantity outliers

instr	time	exe	buyer	seller	quantity	outl	matchid	dupid
A	9:45:03	X	X	Z	100 (10)	True	1	1
A	9:45:03	X	W	X	60 (6)	False	1	
A	9:45:03	X	Y	X	40 (4)	False	1	
A	9:45:03	Z	X	Z	10	False		1

Notes: This table shows a generic example for correcting a quantity outlier that is part of a match group. All quantities of a match group are corrected by the same factor, even when those reports are themselves not flagged as outliers in order to preserve the balance of buys and sells within the match group. The correction factor is determined by comparing the outlier quantity to the quantity of its duplicate report. The correction is shown in parentheses.

identifier as discussed in Section 2. In these cases, the executing entity effectively acts as a broker, bringing together different clients without taking a position on its own balance sheet. I link such reports by proving them an ID that is unique to their group. A total of 48% of reports are part of a matched trade and in 98% of these cases the quantity that the executing entity buys and sells nets to zero.[9] *Figure 8 shows that in 90% of matched trades the executing entity matches one buyer with one seller, and in a further 5% it matches either one buyer or one seller with two other counterparties.*

If one report (or more) in a match-group is flagged as a quantity outlier, the above correction procedure is applied, correcting all quantities of the match-group by the same factor, even those that are not flagged as an outlier. This is done because the quantities that an executing entity buys and sells in a match-group typically net to zero. Therefore, to preserve the balance of buys and sells in match-groups where one report is subject to a correction, the correction is expanded to all reports of the same group. For example, in Table 5, the quantities in the first three rows would be divided by 10 as the quantity in the duplicate report by Z is a factor 10 smaller than the outlier quantity reported by X.

Note, however, that the correction is only applied if the correcting factor is the

[9] The remaining 2% of matched trades have median absolute difference of 200,000 between the amount bought and sold.

same across all duplicates of the same match-group. If, for example, there was another duplicate report by W stating a quantity of 60, the reports by X would remain unchanged as there would now be disagreement across duplicates of the match-group about how or whether to correct the reports.

6 Conclusion

This paper presents an approach for cleaning transaction reports submitted under the Markets in Financial Instruments Regulation as on-shored in the UK. The approach focuses on transactions in corporate bonds. Transaction reports for another asset class, even if reported under the same regulatory framework, would likely benefit from an approach more-tailored to that class.

I do not consider the approach outlined in this paper to be the final word. For example, more could be done to flag and correct erroneous reports by using the time-series dimension of the data. Nevertheless, I hope that the detailed cleaning methodology and various post-cleaning summary statistics presented above provide a useful guideline for future research on this data set and also informs the cleaning process of similar data sources.

References

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: series B (methodological) 39*(1), 1–22.

Fellegi, I. P. and A. B. Sunter (1969). A theory for record linkage. *Journal of the American Statistical Association 64*(328), 1183–1210.

Galil, Z. (1986). Efficient algorithms for finding maximum matching in graphs. *ACM Computing Surveys (CSUR) 18*(1), 23–38.

Grannis, S. J., J. M. Overhage, S. Hui, and C. J. McDonald (2003). Analysis of a probabilistic record linkage technique without human review. In *AMIA Annual*

Symposium Proceedings, Volume 2003, pp. 259. American Medical Informatics Association.

Jurkatis, S., A. Schrimpf, K. Todorov, and N. Vause (2023). Relationship discounts in corporate trading. *Bank of England Working Paper Nov*(1049), 1–45.