Who’s telling the truth? Statistical techniques for error detection in double-sided reporting of money market transactions

Gemma Agostoni, European Central Bank, [gemma.agostoni@ecb.int](mailto:gemma.agostoni@ecb.int)

Steven Cassimon, European Central Bank, [steven.cassimon@ecb.int](mailto:steven.cassimon@ecb.int)

Sébastien Pérez-Duarte, European Central Bank, [sebastien.perez\_duarte@ecb.int](mailto:sebastien.perez_duarte@ecb.int)

**Abstract**

*In 2016 the European Central Bank started collecting statistical data on money market transactions based on the Money Market Statistical Reporting (MMSR) Regulation. This granular dataset covers four segments of the euro money markets, namely unsecured, secured, foreign exchange swaps, and euro overnight index swaps. The detailed trade data to be provided comprise amongst others the volume, rate, and counterparty or collateral type information together with the time when the transaction was conducted. On average 45,000 transactions are being received every day from the largest 52 banking institutions in the euro area. Different procedures to ensure high quality have been adopted, including using ISO 20022 XML standard for the exchange of information. Particularly relevant are the methods aiming to reconcile the data reported by the different institutions. Both sides of a transaction (borrowing and lending) are reported by the parties involved. The lack of a unique transaction identifier poses significant challenges to the identification of the two sides of a single transaction.*

*This paper presents different techniques, applied to MMSR, for pairing and matching the two sides of a transaction based on incomplete and partially incorrect information. Errors can be of several kinds: over-reporting and under-reporting of transactions, and misreporting. Absolute, partial and fuzzy matching techniques for pairing and matching individual transactions are developed, which allow automatically identifying out-of-scope or missing transactions, and misreported values, which would have not been detected otherwise. Inconsistencies detected are then referred to the specific reporting agents. These techniques have been critical in enhancing the quality of the MMSR, and may find applications in other datasets.*

**Keywords:** Transaction-level data, double-sided reporting, pairing and matching techniques, under- and over- reporting, misreporting

# Introduction

The increasing availability of very granular data expands the possibilities of error detection and quality assurance. One such area is the collection of transaction-level data, where a large number of transactions between different economic agents are reported. Recent European regulations (e.g. European Market Infrastructure Regulation [EMIR], Securities Financing Transaction Regulation, Money Market Statistical Regulation [MMSR]) impose such form of reporting and mandate that both counterparties of such a transaction (when applicable) have the obligation to report, implying that transactions between such parties are reported twice. This double-sided reporting allows new techniques to assess the quality of the data and detect possible errors, as long as the two sides of the same transaction can be identified. In the context of the EMIR work, the process is known as “pairing” (to identify the two sides of each transaction) and “matching” (to reconcile the differences between these two sides). For practical reasons however, the identification of the two “legs” is often challenging (e.g. CPMI-IOSCO 2017, McLellan and Rugarber 2014) and, as long as the full unique identification of transactions is not implemented, alternative statistical techniques are needed. Even with a unique identification, full reconciliation is not assured (as is unfortunately the case in EMIR). Under the MMSR collection, there is no UTI and we need to rely on the other observed characteristics of the transactions to achieve the pairing and matching.

There is so far little scientific literature on the pairing and matching. The closest is the literature on clustering, and in particular the definition of similarity measures, starting from Goodman and Kruskal (1954), Gower (1971), Gower and Legendre (1986), with more recent contributions by e.g. Morlini and Zani (2012). We borrow from the literature in the construction of the pairing distance below.

The paper is organised as follows. The second section describes the procedure to identify the two legs of each transaction under reporting error, while the third section briefly describes the data and the application of the pairing and matching procedure to the MMSR data. A short conclusion outlines the future research directions.

# Pairing and matching under double-sided reporting without unique identification

In this section we consider individual trades between two counterparties in the sample. These trades are to be reported by each of the counterparties and should be present twice in the data, in the absence of reporting error. We assume no unique identification of the trade and hence no immediate and foolproof way to link the two legs of the trade.

Each leg has some characteristics (or variables) that are trade-specific (i.e. they should in principle be identical for both legs of the same trade), noted and the other characteristics that are counterparty-specific (i.e. they differ between the two legs), noted .[[1]](#footnote-1)

## Grouping of trades

We consider a subset and of the trade- and counterparty-specific variables.

**Definition**: Two legs and are **grouped**, if the variables in and match, i.e. and .

The objective is to have all trades that have comparable characteristics together. By increasing the number of variables in and the number of grouping classes increases, and the number of legs in each class will decrease. In the extreme case, each leg might be in its own class.

## Pairing of trades

For counterparty-specific elements we introduce , which is formed from the variables of by exchanging the counterparty-specific information. For example, the side of the transaction “borrowing” would become “lending”.

**Definition:** two legs and are **paired**, if the variables match in the following way: and . The pairing relation is symmetric.

It can be seen that if one leg is paired to another, then all legs in the grouping class of the first leg are paired to all legs in the grouping class of the other, and that one class can only pair to at most one other class.

## Pairing distance

We now consider pairing within two grouping classes and . We select a subset of characteristics and . The pairing distance between two legs and is based on the distance between the variables of and , and is finite, positive, and symmetric.

For variables the distance should take into account the reversal of the roles (e.g. borrowing associated with lending). For continuous variables the distance could allow for some fuzziness (i.e. the notional amounts do not need to match exactly). The distance could also take into account that some variables are more crucial than others in determining whether two legs are indeed from the same trade or not (expert knowledge). Finally, the relative frequency or rarity of categories could be used in the calculation of the distance. We will explore these considerations in follow-up work. In this paper we consider a simple binary distance on the characteristics, allowing for some fuzziness in the numeric variables.

## Matching of legs

We can compute the pairing distance between all legs of paired to all legs of .

**Definition:** two legs are **matched** if they are paired, are the unique best pairing of the other, and the pairing distance is below the **matching threshold**.

In other words, two legs as matched when they are paired (implying that the legs have the same grouping variables, allowing for the counterparty-specific variables) and when each leg is the best pairing of the other one based on the distance.

If the best pairing is not unique, the legs are considered **ambiguous**.

If is unpaired (i.e., there is no class that pairs to ) then all legs of are also **unpaired**, and thus also **unmatched**.

## Summary

A leg can therefore be:

* ***Unmatched***, if there is no leg that can be paired or matched to it
* ***Ambiguous***, if there are several legs that pairs to it and that are both the best matches of the other
* ***Disputable***, if there is a leg that pairs to it, and both are the unique best pairing, but the pairing distance is above the matching threshold
* ***Matched***, if there is a leg that pairs to it, that both are the unique best pairing of the other, and the pairing distance is below the matching threshold.[[2]](#footnote-2)

Within the matched category, we can further distinguish the ***perfect*** matches (with a pairing distance of 0) and the ***misreported*** matches (with a strictly positive pairing distance, but still below the matching threshold.

# Application to MMSR data

## Description of the data

In 2016 the European Central Bank started collecting statistical data on money market transactions based on the Money Market Statistical Reporting (MMSR) Regulation (ECB/2014/48). This granular dataset covers four segments of the euro money markets, namely unsecured, secured, foreign exchange swaps (FX Swap), and euro overnight index swaps (OIS). On average 45,000 transactions are being received every day from the largest 52 banking institutions in the euro area. The counterparties are identified by their LEI in case of credit institutions or supranational authority, or the transaction is conducted via a central counter party (CCP).

A particular case is when both the reporting agent and the counterparty are in the MMSR reporting population and therefore the dataset includes both sides of such transactions. However, transactions reported are not identified through a unique identifier, posing significant challenges to the identification of the two sides of a single transaction.

A transaction is described both by numerical and categorical variables, for instance the volume and the rate, counterparty or collateral type information together with the dates when the transaction was conducted, when it will be settled, and when it matures. The majority of the categorical variables are reported according to code lists and therefore no arbitrary value is allowed. We consider here the transactions between the banks in the MMSR reporting population that have taken place from 1 April 2017 to 31 March 2018. The percentage of transactions between MMSR reporting agents is not homogenous in all market segments. On average on a daily basis, 14% of transactions in the secured and FX Swap segments is between MMSR reporting agents, while in the unsecured and OIS segments this figure is respectively less than 1% and 3%. Therefore, in the paper we focus on the secured and FX Swap market segments only, for which on average we have daily 4,500 and 500 transactions respectively. The transactions are characterised by common variables and specific ones at segment level.

## Results of the pairing and matching exercise

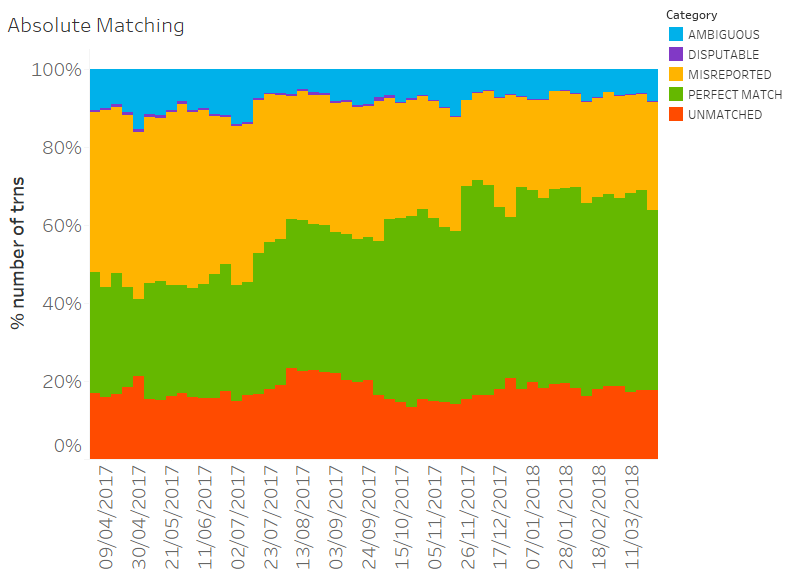
The first step is the selection of the variables for grouping the trades. The counterparty-specific variables and the trade date are always considered among the variables used for the grouping of trades in order to ensure well defined grouping classes. The selection of additional grouping variables for each segment was derived from the analysis of the distribution of errors of the single-matched transactions. Furthermore, in the selection of variables we considered the trade-off between improving the comparability of grouping classes and potentially increasing the number of unmatched transactions. For both segments, we select as additional variable the nominal amount rounded to 3 significant digits. This allows increasing the percentage of transactions matched while having a limited increase in the number of unmatched transactions.

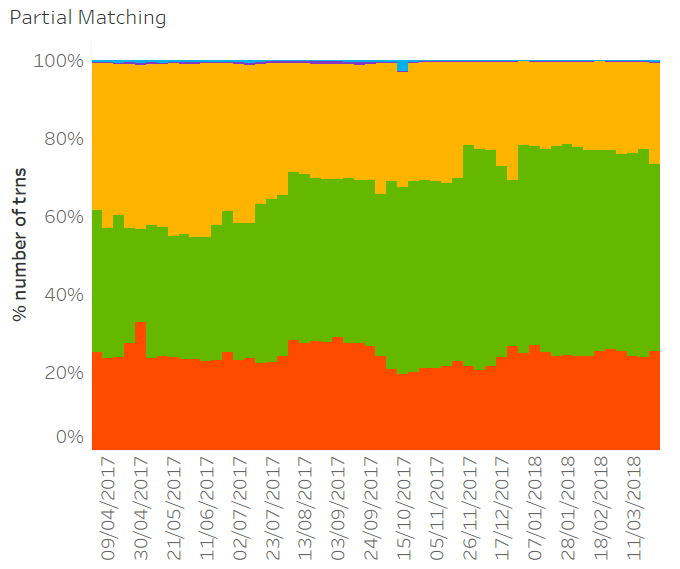
We consider two procedures for the pairing and matching. With “absolute matching”, we limit the number of grouping variables but enforce strict equality between the matching variables. With “partial matching”, we include more grouping variables but allow for small deviations in the numerical values (deal rate, forward points). A binary distance is used throughout. Partial matching allows focusing on the most severe cases of misreporting and to exclude not significant cases, such as the ones due to rounding. Transactions classified as misreported in the absolute matching may then be classified into perfect matches under the partial matching.

From Chart 1 and Chart 2, respectively for secured and FX Swap segment, we observe that the transactions classified as ambiguous in the absolute matching are the ones that are distributed between the categories matched and unmatched in the partial matching. This effect can be explained by the additional grouping variable used in the partial matching method.

The remaining impact on the classification of cases is limited for the secured segment, while for the FX Swap segment it is significant. In the secured segment, there are very little rounding issues for the numerical values, while in the FX Swap segment rounding differences are more common, in particular for the variable FX Forward points.

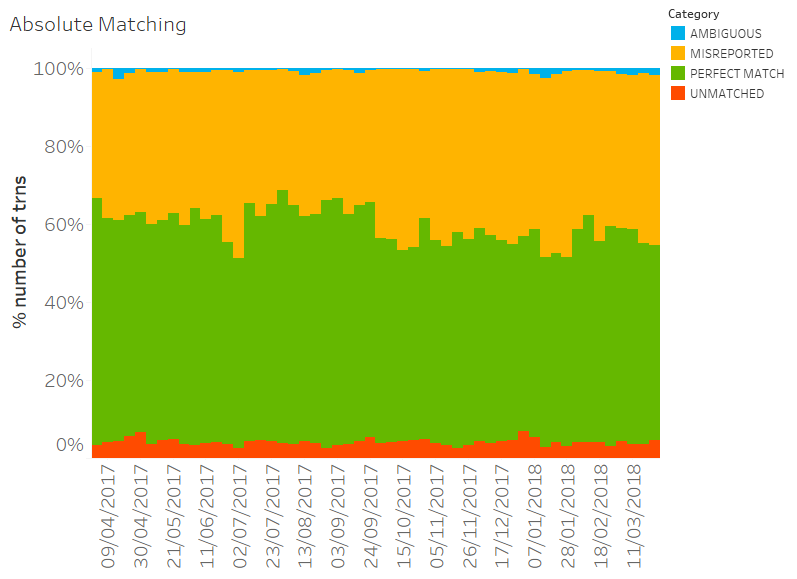
**Chart 1: Secured segment**

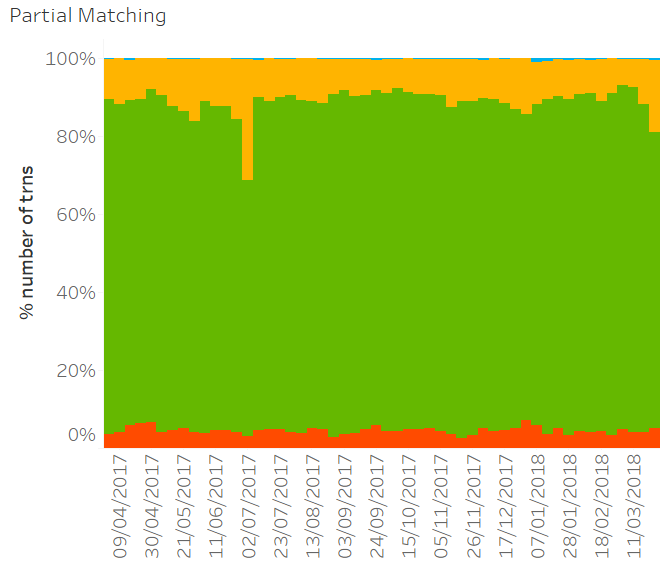




Source: MMSR data, authors’ calculations

Chart 2: FX Swap segment





Source: MMSR data, authors’ calculations

Apart from the comparison of the absolute and partial matching technique, Chart 1 and 2 also show the evolution over time of the different categories for both segments. For the secured segment the percentage of perfectly matched transactions increases from 34% at the beginning of April 2017 to 50% at the end of March 2018, a sizeable increase in the quality of the data, following extensive data quality monitoring, feedback to reporting agents, detailed reporting instructions and questions and answers, and including a thorough matching exercise conducted in 2017.

The FX Swap segment shows very high data quality with 80% of the transactions matching perfectly consistently over time, while the remaining transactions are mainly concentrated in the misreported. This can be explained to a large extend by the fact that the FX Swap segment has a simpler reporting schema than the secured segment. Furthermore, FX Swap covers only a class of transactions, i.e. swaps, while the secured segment covers a broader scope of financial products. An example of complex products are the ones with open maturity that require daily reporting following precise rules for the date structure (i.e. settlement and maturity date).

Looking into the source of errors, the exercise reveals that the unmatched transactions are caused by either reporting of transactions that are not in the scope of MMSR, misidentification of the counterparty, or transactions with foreign branches not covered by the regulation. Misreporting affects numeric and categorical variables, with overall low impact on most of the numeric variables. The most frequently misreported variables differ in the secured and FX Swap segments. For the secured, several errors are due to the different reporting of the date structure and collateral information, such as the haircut or the collateral nominal amount. While in the FX Swap segment the main differences are caused by erroneous calculation of FX forward points. Unmatched and misreported transactions are then followed-up with the respective reporting agents, and corrected data is requested as necessary.

# Conclusion

This paper develops a procedure to pair double-sided transactions in the MMSR data in the absence of a unique trade identifier, from the reported characteristics of the transactions. The procedure groups transactions sharing some common characteristics, pairs the two legs of the transaction, and matches to the possible unique best pair, based on a distance that allows fuzziness. Trades are then determined to be perfectly matched, misreported, unmatched, or undetermined (ambiguous). The exercise reveals 1) the difference in quality between the secured and FX swap segments, due to the complicated structure of the secured transactions; 2) the overall high quality of both segments, with 50% to 80% of the trades perfectly reported in the two segments; 3) the increase in quality in the secured segment over time.

Although the process of pairing and matching as applied to MMSR data has given very positive results in terms of improvement of quality, it is still based on somewhat heuristic decisions in terms of grouping and pairing variables, and the choice of the pairing distance. Future work will extend the procedure to a more data-driven selection of the relevant parameters, which will also allow the process to be more easily applied to other datasets.

# References

Committee on Payments and Market Infrastructures and Board of the International Organization of Securities Commissions (2017), Technical Guidance: Harmonisation of the Unique Transaction Identifier, *Bureau of International Settlements*.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American statistical association*, 49(268), 732-764.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857-871.

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of classification*, *3*(1), 5-48.

Morlini, I. & Zani, S. (2012), A New Class of Weighted Similarity Indices Using Polytomous Variables, *Journal of Classification*, 29(2), pp. 199-226.

McLellan, S. and Rugarber, L. A. (2014), Pairing/Matching – the last hurdle in the race to reporting transparently? *Sapient Global Markets blog*. <https://www.sapientglobalmarkets.com/blog/pairingmatching-the-last-hurdle-in-the-race-to-reporting-transparently>

1. Examples of trade-specific variables would be the dates (trade, settlement, maturity), instrument type, reference rate, collateral, currencies, notional amount. Examples of counterparty-specific variables would be the identification of the reporting entity and the counterparty (their roles being interchanged between the two legs) and the transaction type (lending-borrowing, selling-buying). [↑](#footnote-ref-1)
2. This category could be split further, as s*ingle match*, if and are the only members of their respective grouping class, and as b*est match*, if the grouping classes of and/or have more than 1 member. [↑](#footnote-ref-2)