Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

Ninth IFC Conference on "Are post-crisis statistical initiatives completed?"

Basel, 30-31 August 2018

Two is company, three's a crowd: automated pairing and matching of two-sided reporting in EMIR derivatives' data[1]

Sébastien Pérez-Duarte and Grzegorz Skrzypczynski,
European Central Bank

[1] This paper was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

Two is company, three's a crowd

Automated pairing and matching of two-sided reporting in EMIR derivatives' data

Sébastien Pérez-Duarte, Grzegorz Skrzypczynski

Abstract

The European Market Infrastructure Regulation (EMIR), which was the European response to the G20 commitment to reform OTC derivatives markets, mandates EU counterparties to report extensive details of their derivative transactions to trade repositories. One of the features of EMIR is the double-sided reporting obligation, which means that details of a trade between two EU entities will be reported separately by each of the counterparties. According to the regulation, the two counterparties have to agree on a unique trade identifier and on the characteristics of the trade itself (so-called common data) before submitting the report. However, after 4 years of EMIR reporting the regulators are still faced with a significant number (up to 55%) of trades that cannot be reconciled.

The existence of both paired and non-paired trades in the EMIR dataset offers an outstanding opportunity to analyse patterns of regulatory misreporting. This paper studies the set of paired trades to understand the pitfalls of double-sided reporting and proposes different measures to assess the level of consistency between two sides of the same trade. It discusses also ways to choose automatically one set of data, if two counterparties provide conflicting information on the trade. Those insights are further used to design an algorithm to pair the trades in the set of non-paired transactions, i.e. transactions that could not be paired using the unique trade identifier.

The proper detection of these misreporting errors allows supervisory authorities to better address the issue with reporting counterparties, provides statisticians with correct aggregates without double counting the same transactions, and ensures that policy makers have a more accurate view of the distribution of risks.

Keywords: pairing, matching, derivatives, two-sided reporting

JEL classification: C18

Table of contents

1. Introduction

In response to the financial crisis of 2008, the G20 leaders committed in 2009 in Pittsburgh to implement a set of reforms that would strengthen the international financial regulatory system. One of the objectives set by G20 was improving over-the-counter derivatives market and one of the measures to achieve this was reporting of the OTC derivatives contracts to trade repositories.[1] In Europe, this obligation was imposed by the Regulation (EU) 648/2012 of the European Parliament and of the Council of 4 July 2012 on OTC derivatives, central counterparties and trade repositories, or European Market Infrastructure Regulation (EMIR).[2]

EMIR provided European authorities with access to information on both over-the-counter (OTC) and exchange-traded (ETD) derivative contracts of unprecedented granularity and size. The ECB receives detailed information about over 30 million open contracts of euro-area entities every day.[3] This rich dataset offers unique opportunities, but presents also significant challenges due to its size and complexity.

One of the particular challenges, which the researchers working with the data have to face, is the issue of double-sided reporting. According to the EMIR regulation, both counterparties to the contract have to report the trade to the trade repository (assuming that they are both located in the EU). This means that both legs of the trade are reported as separate observations, which has to be taken into account in the process of data analysis. While this brings considerable benefits in terms of data quality management, as the regulators can compare the information provided by two counterparties, it leads also to certain issues in data aggregation and analysis.

As envisaged by EMIR regulatory and implementing technical standards, the counterparties are obliged to mutually agree on a unique trade identifier before reporting the trade. This requirement, however, turned out to be very difficult to comply with by the reporting agents, in particular in the first years of reporting.[4] Since then there have been significant improvement in the legal framework, as well as in the industry's ability to exchange the trade identifier before the reporting deadline. Additionally, the trade repositories, together with ESMA, have put a lot of effort into implementing the reconciliation process, where information is exchanged on a daily basis between the TRs, in order to assess the completeness and consistency of reporting, and ensure the pairing of the two legs of the transactions. As reported by ESMA, "*average pairing rates in November 2017 rose to 87%, from 55% in November 2016*".[5] However, in the set of all outstanding trades, which includes both new and old trades, the number of non-paired observations is much

[1] For more information on post-crisis reforms of derivative markets see ECB (2016)

[2] See: https://www.esma.europa.eu/regulation/post-trading

[3] Apart from the report on outstanding contracts ("trade state" report), the authorities receive also daily updates on the new transactions, valuation updates and other lifecycle events ("trade activity" report).

[4] See Maxwell (2104)

[5] See ESMA (2017), p. 40

higher, reaching around half of the reported dataset. This poses significant challenges to researchers using the data and hinders its meaningful aggregation.

The goal of this paper is to provide insights into the nature of the non-paired trades and to attempt to apply an automated procedure to find the corresponding legs in the non-paired sample. For this purpose, we draw on the method applied by Agostoni, et al. (2018) to the dataset collected under the ECB Money Market Statistical Regulation (MMSR), with some modification to account for differences between the two reporting frameworks and operational challenges related to the size of the EMIR dataset.

The paper is organized as follows. Chapter 2 briefly outlines the method used. Chapter 3 describes the choice of grouping variables and other parameters of the matching procedure. Chapter 4 presents the outcome of the quantitative analysis of the unpaired EMIR sub-sample, while Chapter 5 concludes.

2. Method

For the purpose of analysis we have adopted a modified method from Agostoni, et al. (2018). Similarly to MMSR dataset EMIR reports consist of counterparty-specific variables, noted Y_i,[6] and trade-specific variables, noted X_i.[7] The counterparties are obliged to agree on the values of trade-specific variables before reporting the trade. While this cannot be safely assumed in the cases when counterparties failed to agree on the trade ID, we can still expect that the characteristics of the trade reported by the two counterparties will not differ significantly.

It is not expected that Y_i will be consistent between two legs of the same trade. However some variables of Y_i may contain information that the reporting entity provides on its counterparty,[8] thus they could be cross-compared with the information included in the other leg. The identifiers of the counterparties are a good example of such relationship: for a paired trade *id_of_the_reporting_counterparty*$_{t1}$ should be equal to *id_of_the_other_counterparty*$_{t2}$, and vice versa. We denote $\tilde{Y}_i$ as a vector formed by switching corresponding variables of Y_i.

In terms of data type, we can distinguish the following types of variables:

- categorical (discrete) variables
- dates

[6] Those include, identifiers of counterparties to the trade and other agents involved in the transactions, sector of the reporting counterparty, and information on valuation and collateral

[7] Those include information on the contract, like asset class, product and underlying ID, information on notional, various timestamps related to the transaction, details on clearing, and asset-class specific variables.

[8] The contract value is one exception to this rule. It is included in the counterparty-specific variables, as, by construction, two counterparties observe the contract value with opposite sign. Additionally, there may arise differences due to different valuation methodologies, time of valuation, etc. For the purpose of our pairing exercise, however, the absolute value of this variable proves to be useful, and was treated in a similar way to trade-specific variables.

- timestamps
- numerical variables

2.1 Clustering and grouping of trades

We denote X^g and Y^g a subset of X and Y, respectively. Those will be called "grouping variables" below.

Definition: two legs u and v are **clustered** if the variables match in the following way: $X_u^g = X_v^g$ and $Y_u^g = Y_v^g$. The clustering relationship is symmetric, reflexive and transitive, and allows partitioning the set of all trades into disjoint clusters.

Definition: two legs u and v are **grouped**, if the variables match in the following way: $X_u^g = X_v^g$ and $Y_u^g = \tilde{Y}_v^g$. The grouping relation is symmetric.

If two legs u and v are grouped, then all trades of the cluster of u are grouped with all trades of the cluster of v. The grouping relation splits the dataset into sub-groups, from which potential paired trades can be drawn. To illustrate with an example: if Y^g consists of ID's of reporting and other counterparty, and X^g contains only asset class, then each trade in asset class Z, reported by the entity A, with counterparty B, will be grouped with every trade in asset class Z, reported by the entity B, with counterparty A.

2.2 Matching distance

We denote X^m a subset of X, further described as the set of "matching variables". X^g and X^m do not contain the same variables.[9]

For the purpose of determining the optimal candidates for paired trades in the grouped dataset, we calculate a matching distance between each pair of trades in a group:

$$d_m(u,v) = \sum_{x \in X^m} w_x \, f_{T(x)}(u,v,\tau_x)$$

where:

- w_x is the weight associated to variable x, indicating the importance of the variable in the decision to pair two trades
- $f_{T(x)}(u,v,\tau_x)$ is a distance function between the values of the variable x between legs u and v, taking parameter τ_x
- $T(x)$ is the type of variable x (categorical, date, timestamp, or numerical)

For simplicity we have considered the discrete distance function depending on the type of the underlying variables. The distance functions' representations are taken in this paper to be binary, but other choices are possible; the selection is summarized in Table 1.

[9] By definition, for all trades in the group $X_u^g = X_m^g$. Thus, using any variable from X^g in the matching process does not bring any additional benefit.

Distance functions Table 1

Data type	$f_{T(x)}(u, v, \tau_x)$
Categorical	If $x_u = x_v$ then $f_{T(x)}(u, v, \tau_x) = 0$ else $f_{T(x)}(u, v, \tau_x) = 1$
Date	If the absolute distance between x_u and x_v is less or equal to τ_x days then $f_{T(x)}(u, v, \tau_x) = 0$ else $f_{T(x)}(u, v, \tau_x) = 1$
Timestamp	If the absolute distance between x_u and x_v is less or equal to τ_x seconds then $f_{T(x)}(u, v, \tau_x) = 0$ else $f_{T(x)}(u, v, \tau_x) = 1$
Numerical	If the relative difference[10] between x_u and x_v is less or equal τ_x% then $f_{T(x)}(u, v, \tau_x) = 0$ else $f_{T(x)}(u, v, \tau_x) = 1$
All variables	If x_u is NULL and x_v is NULL then $f_{T(x)}(u, v, \tau_x) = 0$ If x_u is NULL and x_v is not NULL, or vice versa then $f_{T(x)}(u, v, \tau_x) = 0.5$

The methodology could be further extended to allow for continuous output of the distance function for continuous variables, introduce string metrics to better measure the distance between categorical variables, account for correlation between variables, or take into account common misreporting patterns (e.g. reversing the legs of interest rate derivatives). We leave these considerations for future work.

2.3 Classification of trades

The calculation of the matching distance between grouped trades allows us to classify the trades along the conditions described in the table below. For this purpose we define the *best match* of a trade as the trade from its group, to which it has the lowest matching distance (if this condition is satisfied by multiple trades, then *best match* is not determined).

Trade classification Table 2

Trade classification	Definition
Perfect match	The trade has a best match, and the relation is reciprocal. Additionally, the matching distance is equal to 0, i.e. the trades are identical within the bounds of the matching conditions.
Imperfect match	The trade has a best match, and the relation is reciprocal. The matching distance is higher than 0, which may indicate misreporting of some characteristics of the trade.
No match	There exists no trade, with which the trade has a grouping relation.
Perfect matching group	There exist multiple trades, with distance 0 to each other. The trades can be considered perfect matches, but it is not possible to determine the exact legs of the particular trades.
Ambiguous	All other cases – the grouped trades could not be unambiguously matched.

[10] The relative difference is $\mathrm{rd}(x, y) = 2\frac{|x-y|}{|x|+|y|}$ if both $x \neq 0$ and $y \neq 0$, 0 otherwise.

3. Determination of the grouping and matching variables

One of the differences between the EMIR and MMSR datasets is the fact that a significant sub-sample of the EMIR trades can be unambiguously paired by using the trade identifier reported by the counterparties. This UTI-paired sub-sample can be used to determine the optimal parameters of our matching procedure, in particular X^g, w_x, and p_x. For this purpose we have grouped together trades by their counterparties' identifiers and trade identifier, and calculated a variety of statistics on a variable-by-variable basis to check the consistency of data reported in two distinct legs. This allows selecting the variables that offer the highest degree of similarity between the two legs as grouping variables, other variables as matching variables, and the observed patterns between the matching variables as elements in the selection of the thresholds.

All calculations in this paper were carried out on data received from five trade repositories[11] for the reference date of 5 July 2018, with a total of 32,780,000 trade state reports.

3.1 Grouping variables

The identifiers of both counterparties were by construction included in the set of grouping variables, as it allows restricting the size of the groups to the level, which facilitates calculating matching distance between all members of the group. This approach has some limitations, as it does not allow addressing the potential issue of ID misreporting. [12] This phenomenon will be further explored in follow-up work.

We have considered variables with following characteristics as potential candidates for grouping variables set X^g:

- the variables that are very well-populated, i.e. less than 10% of paired trades have missing information in both legs of the trade,
- the variables that exhibit high matching of the non-empty values, i.e. the information provided in the two legs is equal for more than 98% observations.

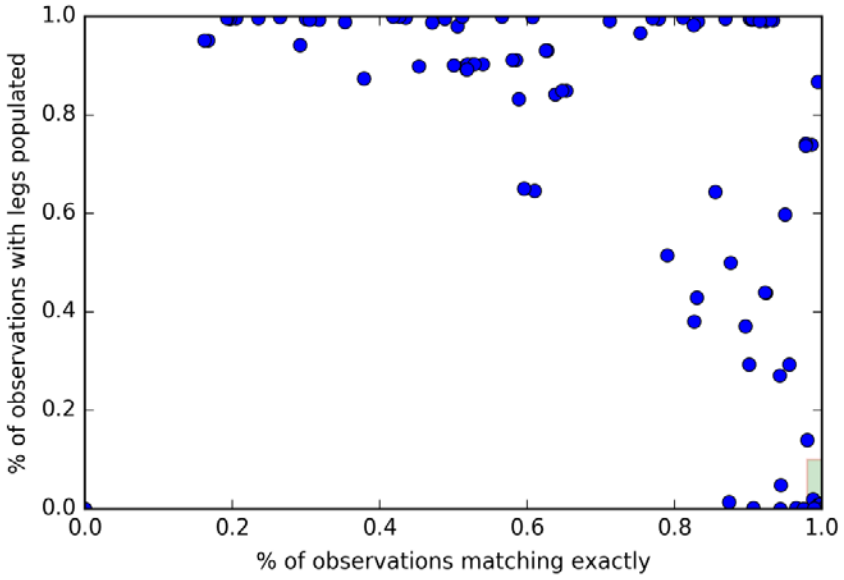
Figure 1 shows the distributions of two abovementioned metrics across different variables. The green rectangle in bottom right of the chart represents the criteria for the grouping variables.

[11] DTCC Derivatives Repository Ltd. (DDRL), Krajowy Depozyt Papierów Wartosciowych S.A. (KDPW), Regis-TR S.A., UnaVista Ltd, and ICE Trade Vault Europe Ltd.

[12] The problem of ID misreporting if further discussed in chapter 4.

Determination of grouping variables

Scatterplot between share reported and share matching exactly Figure 1



Source: EMIR data, UTI-paired sample, ECB calculations. The green rectangle in bottom right of the chart represents the criteria for the grouping variables.

Based on this analysis, combined with expert knowledge, the following variables were chosen as grouping variables:

- ID of the reporting counterparty (LEI)
- ID of the other counterparty (LEI)
- Asset class (interest rate, currency, equity, commodity, credit)
- Contract type (swap, option, forward, etc.)
- Clearing status (cleared, not cleared)
- Execution date (extracted from the execution timestamp)

3.2 Matching distance weights

The weights introduced in section 2.2 serve the purpose of assigning more importance to variables that signal a higher probability of two legs being the correct match. The weight can be derived from the likelihood that a noisy observation is in fact equal to the target.

We describe the case of a categorical variable X; we observe only $\hat{X}$, a perturbed version of X. We assume that with a probability p the variable is not perturbed, and when the variable is perturbed it is given randomly one of the values of the variable, with assignment proportional to the existing distribution of values. The probability p is the **fidelity** of the variable (the higher the fidelity, the more we can trust the match) and the share of the category is the inverse of the specificity (the higher the **specificity**, the lower the chance that a match is random).

We have a known value of x and want to determine the probability that when we observe the perturbed value $\hat{X}$ the underlying true value of X is k. This corresponds to $P(X = k|\hat{X} = k)$. We note q_k the share of category k of the variable X. Then $P(X = k|\hat{X} = k) = p + (1-p)q_k$. Similarly, $P(X = k|\hat{X} \neq k) = (1-p)q_k$. We bound by below the first probability by p and bound by above the second one by $(1-p)q$, where q is the share of the most common category of the variable X. Then the contribution of variable X in the match can be estimated in a worst-case way by

$$P(X = \hat{X}|\hat{X}) = p\left(\frac{1-p}{p}q\right)^{\mathbb{1}_{\hat{X}\neq X}}$$

The product of several such variables is the probability of the good match, and the contribution of each variable to the opposite of the log-likelihood is thus the term in the matching distance in section 2.2, and we can thus set

$$w_x = \log\left(\frac{p_x}{q_x(1-p_x)}\right).$$

The higher the specificity $1/q_x$ and the higher the fidelity p_x, the higher the weight (as long as $p_x \geq 1/2$, which we assume in what follows).

3.3 Parameters of distance function

The threshold parameters τ_x of the distance function where chosen on the basis of percentiles of (absolute and relative) differences between values of the legs of the trades in the paired sample. The parameters were selected so that $f_{T(x)}(u, v, \tau_x)$ accepts 98% observations in the paired sample.

3.4 Verification of matching parameters

The selected grouping variables, weights, and parameters were verified by running the procedure on the UTI-paired sample, shown in Table 3. Due to the complexity of the measures and the size of the data, the calculations were carried out on a random representative 1% sample of trades.

Classification of trades in the UTI-paired sub-sample — Table 3

Classification	Number of trades	Percentage
Perfect match	91,861	48.74%
Imperfect match	53,505	28.39%
Perfect matching group	24,484	12.99%
Ambiguous	10,431	5.53%
No match	8,191	4.35%
Total	188,472	100.00%

Source: ECB calculations, based on data received from trade repositories (1% sample of trades).

The results obtained confirm the robustness of the method. For 77% trades the procedure was able to find the matching leg. For another 13% the procedure determined the existence of the perfect matching group, although the trades were too similar to distinguish the individual trades.

4. Analysis of the non-UTI-paired trades

The procedure applied above was applied to the non-UTI-paired sub-sample with the following results:

Classification of trades in the non-UTI-paired sub-sample — Table 4

Classification	Number of trades	Percentage
Perfect match	70,175	0.48%
Imperfect match	421,613	2.88%
Perfect matching group	11,233	0.08%
Ambiguous	700,686	4.78%
No match	13,459,158	91.79%
Total	14,662,865	100,00%

Source: ECB calculations, based on data received from trade repositories[13]

As shown in Table 4, the results of the pairing exercise in the non-UTI-paired sample indicate that most of the trades cannot be paired. 92% of the trades are not in a grouping relationship with any other trades (those trades are further defined as the **unpaired sample)**. Put differently, it means that counterparty A reports a trade with counterparty B, but there is no corresponding trade reported by counterparty B with counterparty A within the same group [same asset class and contract type; same clearing status and execution date]. Potential reasons are discussed and quantified in the following sections.

4.1 Trades with entities from non-EU jurisdictions

EMIR applies to entities resident in the EU, thus the trades concluded with counterparties outside the EU are expected to appear only once in the dataset. In order to assess the extent of this phenomenon we have used the GLEIF dataset,[14] combined with information reported within EMIR, to identify the trades that were carried out with counterparties from other jurisdictions. As shown in Figure 2 the share of those trades in the unpaired sample amounts to over 35%, with significant contribution of trades with US (18.5%)[15] and Swiss (6%) counterparties.[16]

[13] Around 140,000 of the trades (1% of the non-UTI-paired sample), forming one particularly large grouping set, could not be classified by the procedure due to the size of the set. The inclusion of those trades, however, would not materially change the results.

[14] GLEIF is Global Legal Entity Identifier Foundation, which is responsible i.a. for managing reference database of LEI codes. The LEI is an ISO standard for identification of legal entities. For more details visit https://www.gleif.org

[15] All the percentages refer to the share in the unpaired sample.

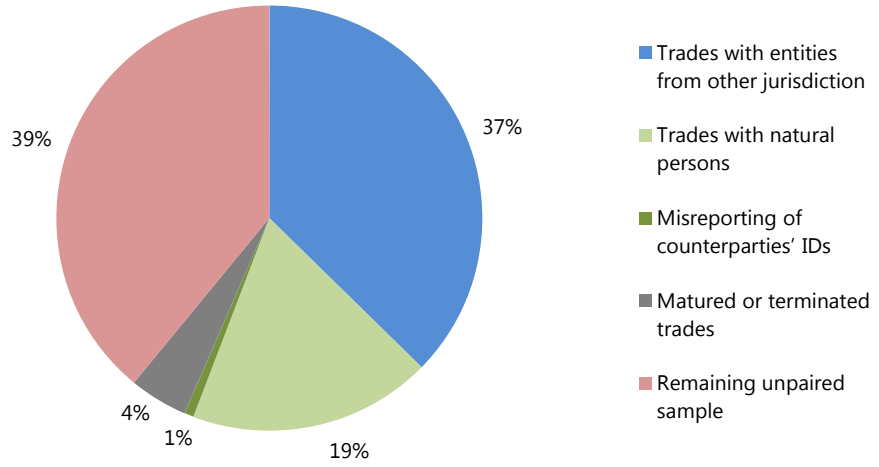
[16] We have also identified a non-negligible amount of trades (around 9%), where information about the country of the other counterparty is missing. Those trades are kept in the unpaired sample.

4.2 Trades with natural persons

The EMIR regulation does not impose the reporting obligation on natural persons, hence for trades carried out by private individuals we will see only the leg reported by the counterparty of the trade. To identify such trades we used the type of the identifier, assuming that the counterparties identified by the "client code", instead of LEI are not legal entities.[17] As shown in Figure 2, this filtering allows us to restrict the size of the unexplained unpaired sample by a further 18%.

Breakdown of the unpaired sample

Figure 2



Source: EMIR data, ECB calculations. Full sample. Non-exclusive categories removed from the sample in the order shown in the chart, e.g. the Trades with natural persons does not include any Trade with entities from other jurisdictions, although the latter can include Trades with natural persons.

4.3 Misreporting of counterparties' IDs

In case the reporting entity misreports its ID or the ID of its counterparty, then the two legs of the trade will have different counterparty pairs, and they will not be grouped together for the matching procedure. To assess the degree to which the dataset could be affected by that issue, we checked whether the IDs reported in the unpaired sample could be found in the GLEIF reference database. As shown in Figure 2, the extent to which this effect could explain the unpaired sample seems to be limited.

Another type of ID misreporting can occur, when the reporting entity submits a valid LEI code as the ID of the other entity, but this is not the LEI with which the other counterparty identifies itself. This could happen, for instance, when the reporting entity reports the ID of the parent company of its counterparty. To assess this phenomenon, we carried out a separate exercise, in which we replaced the IDs of the entities with the LEIs of their ultimate parents from the GLEIF relationship

[17] While this approach follows EMIR guidelines on reporting, it cannot be excluded that some counterparties incorrectly assign a client code to a legal entity, which should be identified by LEI. This is, however, outside the scope of this paper.

database,[18] and then re-run the grouping procedure. The result was the reduction of the unpaired sample by less than 0.5% observations. Thus, it was concluded that information from GLEIF relationship database is not useful in improving the outcome of the pairing exercise.[19]

4.4 Matured or terminated trades

If a trade is terminated early or compressed away, the entity is expected to send this information to the trade repository, which should remove the trade from the trade state report. Furthermore, the trade repository should remove all the trades that reached their maturity date. If any of these obligations is not met, we may see in the trade state report trades that do not exist anymore. To assess this we:

- checked the remaining unpaired sample for existence of trades, for which the maturity date or termination date lies in the past,
- cross-checked the trades with information reported on trade activity reports from the preceding two months. If there was any indication that the trade has been terminated or compressed, we flagged it as an expired trade.

By following the above steps, we have identified further 4% of the trades, which could be deducted from the unpaired sample.

4.5 Understanding the unpaired sample

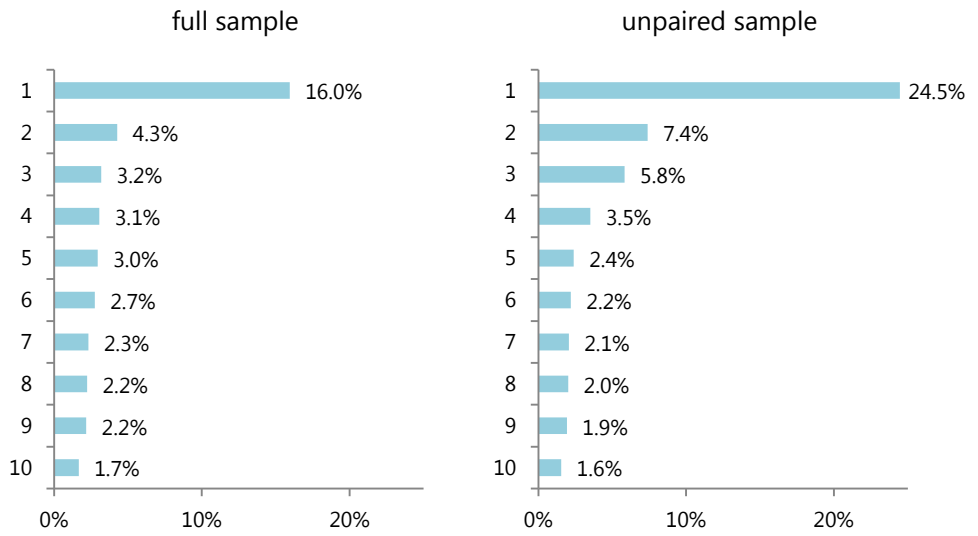
After taking the steps described above, we are left with around 5 million trades, constituting 40% of the original unpaired sample. The remaining observations may constitute a case of underreporting (i.e. one of the counterparties failed to meet its reporting obligation), or may be an incorrect reports. Figure 3 presents some descriptive statistics of the set of remaining trades.

[18] See https://www.gleif.org/en/lei-data/access-and-use-lei-data/level-2-data-who-owns-whom

[19] It may be possible that the identification issue lies in the entity managing the fund rather than the fund itself. Other sources of data, also covering other relation types (e.g. fund–management company) could be more successful in improving the pairing outcomes. We leave these considerations for future work.

Share of largest reporters in the unpaired and the full sample

Figure 3



Source: EMIR data, ECB calculations.

In particular the trades in the unpaired sample seem to be significantly more concentrated that in the full sample of trades. This may hint at the existence of systematic issues in reporting by some large reporters of EMIR data.

We investigated further the explanatory factors behind the possibility of pairing with a simple logistic regression (Table 4) on a randomly selected 1% sample. All coefficients are significant at the 1% level and t-statistics are not displayed. Among the most salient results, and other things being equal, cleared trades, intragroup trades, and trades within the euro area are more likely to be paired, while trades that have no contract value reported (as in Abad et al. (2016)) and trades executed before 2018 or close to the reporting date are less likely to be paired. With regards to other breakdowns, like asset class or contract type, the trades with the lower pairing are those that are classified as "Other".

Logit regression

Odds ratios, probability of being paired

Table 4

Variable	Odds ratio	Variable	Odds ratio
Asset class		Location of other counterparty	
Commodity	1	Euro area	1
Credit	0.547	Other EU	0.405
Currency	0.508	RoW	0.00676
Equity	0.619	Nature of reporting party	
Interest rate	0.745	CCP	1
Other	0.324	Financial	1.378
Missing	1.371	Non-financial	4.740
Contract type		Other	2.600
Contracts for difference	1	Missing	5.220
Forwards	1.906	Execution date	
Forward rate agreements	1.919	<= 2013	0.247
Futures	0.879	2014-2017	0.403
Option	1.197	2017	0.645
Other	0.420	2018 Q1	1.178
Swap	1.770	2018 Apr-May	1.197
Swaption	1.899	2018 Jun	1
Missing	0.173	2018 Jul	0.599
Clearing Status		> Aug 2018	0.608
No	1	Contract value missing	
Yes	1.720	No	1
Missing	0.439	Yes	0.590
Intra group		Notional amount (log)	1.036
No	1		
Yes	4.228		
Missing	0.669		
Observations	295,721		

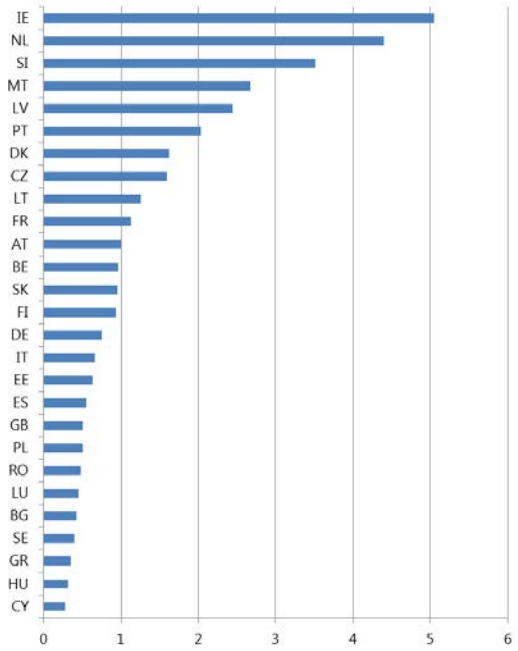
Source: ECB calculations, based on data received from trade repositories for 5 July 2018. 1% sample.

While it is clear that non-EU trades will not be double-reported in the context of EMIR, it is reasonably surprising that within the EU, the pairing success varies substantially by country. This could imply that counterparties from those countries fail to meet the reporting obligation more often, or that there exist some particular difficulties in agreeing on the content of the report by entities from those countries. Another reason could be, however, misreporting or overreporting of the reporting counterparties, and further case-by-case analysis would be needed to fully understand underlying reasons for country disparities.

Pairing success by country of the other counterparty

Figure 4

European Union countries



Source: EMIR data, ECB calculations.

5. Conclusions

The paper applies an automated pairing procedure to the EMIR dataset on derivatives, by grouping the similar trades together, and then classifying them according to the matching distance between each member of the group. Although the robustness of the procedure is successfully verified on the UTI paired sub-sample of the EMIR dataset, and contrary to the similar work on the MMSR dataset, the procedure fails to produce significant improvements for understanding the unpaired dataset. The paper further analyses the set of trades that could not be successfully paired.

In terms of data aggregation, the paper indicates that there is no single optimal approach to the treatment of the non-paired sample. While a limited set of the observations represents trades that could indeed be paired, the rest of the unpaired sample has to be treated with caution. While some of the trades can be considered unpairable (trades with counterparties outside the EU, or with natural persons), the others may be the effect of underreporting, or may be incorrectly reported transactions, which should be then removed from the dataset.

The paper finally offers also some insights regarding further improvement of the quality of data reported under EMIR. We have observed clear patterns between some characteristics of the trades and the probability of being paired. Furthermore, the concentration of entities in the unpaired sample is higher than in the full EMIR dataset. This suggests that focusing on the few most important contract types and/or entities may bring significant benefits in terms of data quality. These results

may be of benefit to national competent authorities, supporting their efforts in improving the quality of EMIR reporting.

The follow-up work may include further refinement of the pairing procedure, and incorporating the time dimension into the analysis. A particularly interesting area of interest could be addressing the issue of potential counterparties' ID misreporting. To this end, alternative reference datasets could be added to understand the links between entities, or the pairing procedure could be applied to the dataset without grouping the trades by the counterparties (to allow matching of the trades with non-identical counterparty pairs. Those paths will be explored in future research.

References

Abad, J. et al (2016). Shedding light on dark markets: First insights from the new EU-wide OTC derivatives dataset. *ESRB Occasional Paper Series*. No 11, September 2016.

Agostoni G., Cassimon S., Pérez-Duarte S. (2018). Who's telling the truth? Statistical techniques for error detection in double-sided reporting of money market transactions. *2018 European Conference on Quality in Official Statistics*.

Ascolese, M., A. Molino, G. Skrzypczynski, S. Pérez-Duarte (2017). Euro-area derivatives markets: structure, dynamics and challenges. *IFC-National Bank of Belgium Workshop on "Data needs and Statistics compilation for macroprudential analysis"*, May 2017

CPMI-IOSCO (2017). Harmonisation of the Unique Transactions Identifier. https://www.bis.org/cpmi/publ/d158.pdf

ECB (2016). Looking back at OTC derivative reforms - objectives, progress and gaps, Economic Bulletin Issue 8, 2016.

ESMA (2017). Annual report 2017. https://www.esma.europa.eu/sites/default/files/library/esma20-95-916_2017_annual_report_0.pdf

Maxwell F. (2014). Majority of EMIR derivatives reports cannot be matched, say repositories. https://www.risk.net/regulation/emir/2335669/majority-of-emir-derivatives-reports-cannot-be-matched-say-repositories

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

Ninth IFC Conference on "Are post-crisis statistical initiatives completed?"

Basel, 30-31 August 2018

Two is company, three's a crowd: automated pairing and matching of two-sided reporting in EMIR derivatives' data[1]

Sébastien Pérez-Duarte and Grzegorz Skrzypczynski,
European Central Bank

[1] This presentation was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

Grzegorz Skrzypczynski

Sébastien Pérez-Duarte

DG-Statistics, European Central Bank

Two is company, three's a crowd:

Automated pairing and matching of two-sided reporting in EMIR derivatives' data

IFC conference

Are post-crisis statistical initiatives completed?

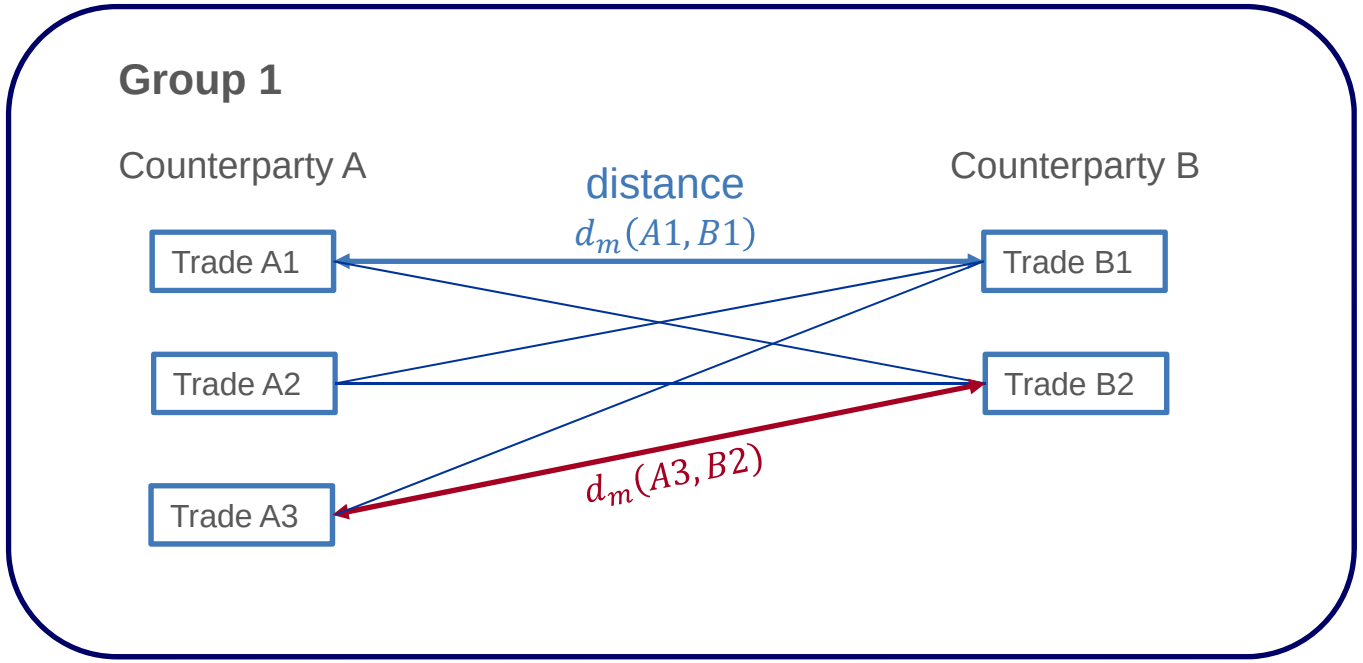
30-31 August 2018, Basel

Motivation: double-sided reporting in EMIR

- EU counterparties report **derivative transactions** to trade repositories
- Separately by both counterparties (**double-sided reporting**)
  - Improves data quality monitoring
  - ⓘ Risk of double-counting when analysing and aggregating the data
- **UTI (Unique Transaction Identifier)** to link trades, agreed between counterparties
  - Challenges in implementation (not unique, different UTIs for the same trade)
  - Work on improving pairing and matching (inter-TR reconciliation process)
  - Global initiatives to harmonise UTI between jurisdictions
  - ESMA estimates pairing rate at 87% = but newly reported trades only

Method

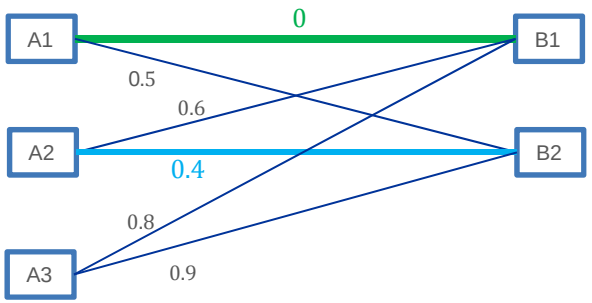
- The trades are split into groups with same values of the **grouping** variables
- The procedure calculates the **matching distance** between each member trade of the group



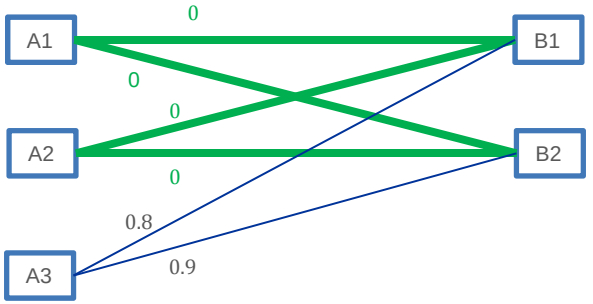
Classification of trades

- Depending on outcome, exclusive categories

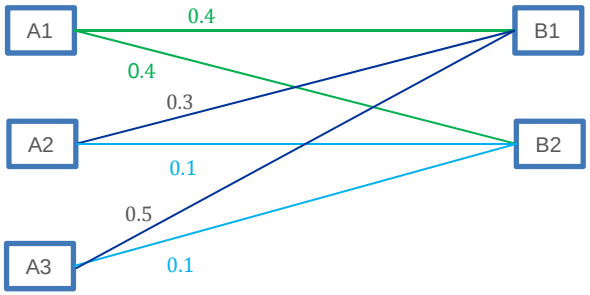
Perfect match / imperfect match



Perfect matching group



Ambiguous



No match

Implementation

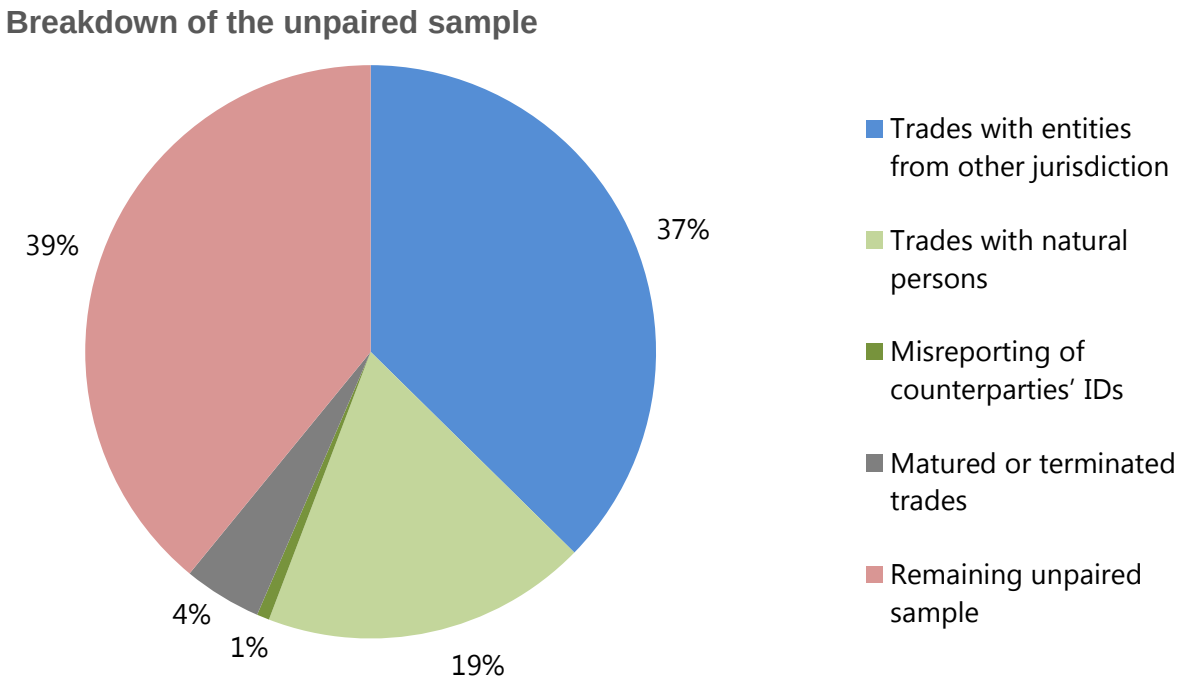
- Sample paired with UTI was used to calibrate the parameters
- **Grouping variables**:
  - Counterparties' IDs
  - Asset class
  - Contract type
  - Clearing status
  - Execution date
- **Matching distance weights**: function of fidelity (how good) and specificity (how revealing) of the variable
- **Thresholds of the distance function**: to accept 98% of the observations in the paired sample

Implementation

- Our procedure has limited impact in the non-paired sample
- Most of these trades don't have any counterpart in their group
  ➔ other reporting issues are at stake

	Paired sub-sample	Non-paired sub-sample
Perfect match	48.74%	0.48%
Imperfect match	28.39%	2.88%
Perfect matching group	12.99%	0.08%
Ambiguous	5.53%	4.78%
No match	4.35%	91.79%

Some things we will never be able to pair



- Remaining 40% of the unpaired sample:
  - may be a result of underreporting
  - may constitute invalid reports

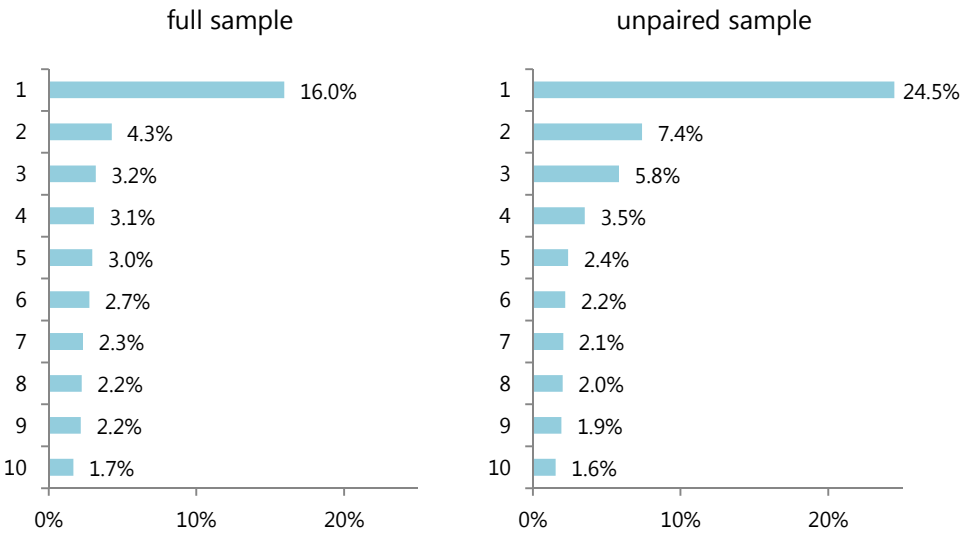
What can't we pair?

Logit regression - Odds ratios, probability of being paired

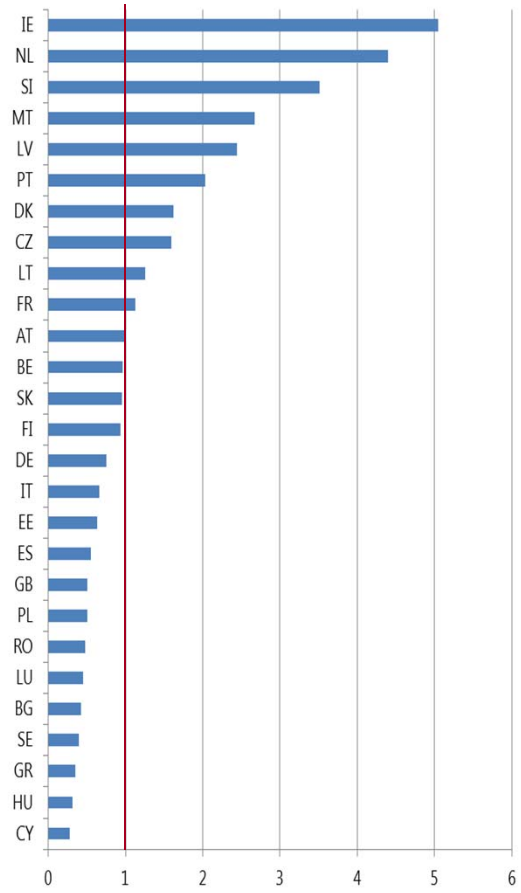
Variable	Odds ratio	Variable	Odds ratio
Asset class		**Location of other counterparty**	
Commodity	1	Euro area	1
Credit	0.547	Other EU	0.405
Currency	0.508	RoW	0.00676
Equity	0.619	**Nature of reporting party**	
Interest rate	0.745	CCP	1
Other	0.324	Financial	1.378
Missing	1.371	Non-financial	4.740
Contract type		Other	2.600
Contracts for difference	1	Missing	5.220
Forwards	1.906	**Execution date**	
Forward rate agreements	1.919	<= 2013	0.247
Futures	0.879	2014-2017	0.403
Option	1.197	2017	0.645
Other	0.420	2018 Q1	1.178
Swap	1.770	2018 Apr-May	1.197
Swaption	1.899	2018 Jun	1
Missing	0.173	2018 Jul	0.599
Clearing Status		> Aug 2018	0.608
No	1	**Contract value missing**	
Yes	1.720	No	1
Missing	0.439	Yes	0.590
Intra group		Notional amount (log)	1.036
No	1		
Yes	4.228		
Missing	0.669		

What can't we pair?

Share of largest reporters in the unpaired and the full sample



Pairing success by country of the other counterparty *(odds ratio)*



Conclusions

- Caution is recommended when **making assumptions** about the unpaired sample **to compute aggregates**
- A **significant share of the non-paired sample is difficult to interpret,** and cannot be easily reconciled
- There exist some **clear patterns** between some characteristics of the contracts and **probability of being paired**
- The unpaired sample exhibits **higher concentration** with regards to reporting entities
- A **focused data quality management** process may bring **significant benefits with limited effort**