BANK OF CANADA
BANQUE DU CANADA

Technical Report/Rapport technique—121

Last updated: May 2, 2022

Historical Data on Repurchase Agreements from the Canadian Depository for Securities

by Maxim Ralchenko[1] and Adrian Walton[2]

[1] Corporate Services
[2] Financial Markets Department
Bank of Canada
mralchenko@bankofcanada.ca; awalton@bankofcanada.ca

DOI: https://doi.org/10.34989/tr-121 | **ISSN** 1919-689X

Acknowledgements

We acknowledge Narayan Bulusu for sharing his knowledge of the data and Venkat Balasubramanian for sharing his insights at the beginning of the project. We thank Johan Brännlund for reviewing an early draft of this document.

Abstract

We develop an algorithm that extracts information about sale and repurchase agreements (repos) from disaggregated settlement data in order to generate a new historical dataset for research. Data from Canada's fixed-income settlement authority, the Canadian Depository for Securities (CDS), is a valuable source of historical information on Canada's fixed-income markets, especially from 2003 to 2016 when few other data sources were available. However, the CDS does not contain details on the terms of trade for repos, such as the repo rate, term or haircut. In the data, each repo is recorded as two distinct settlements but, critically, the sale and repurchase legs of a repo are not explicitly associated. We use a variant of the Gale-Shapley algorithm to solve a "stable roommates" problem to link repos' sale and repurchase transactions and compute their terms of trade. We verify our algorithm by running it on a separate dataset that explicitly associates the sale and repurchase legs of a repo. In addition, we verify the computed repo terms of trade by comparing a subsample of the CDS data with a third dataset that reports terms of trade directly. The derived data are useful for researchers to study the evolution of fixed-income market structure and market conditions.

Topics: Econometric and statistical methods; Financial markets
JEL codes: C55, C81, G10

Résumé

Nous élaborons un algorithme capable d'extraire de l'information sur les opérations de pension à partir de données de règlement désagrégées pour générer un nouvel ensemble de données historiques servant à la recherche. Les données de la Caisse canadienne de dépôt de valeurs (CDS), l'autorité de règlement sur les marchés canadiens des titres à revenu fixe, sont une source précieuse d'information historique sur ces marchés – surtout pour la période de 2003 à 2016, où il existait peu d'autres sources de données. Toutefois, la CDS ne fournit pas le détail des modalités des opérations de pension, comme le taux, la durée ou la décote. Dans les données, chaque opération de pension correspond à deux règlements distincts, mais – fait important – les deux volets de l'opération, soit la vente et le rachat, ne sont pas explicitement appariés. Nous utilisons une variante de l'algorithme de Gale-Shapley pour résoudre un « problème des colocataires stables » afin de coupler les transactions de vente et de rachat des opérations de pension, et ainsi calculer leurs modalités d'exécution. Nous validons notre algorithme en l'appliquant à un ensemble de données distinct où les transactions de vente et de rachat sont explicitement appariées. De plus, nous validons les modalités calculées des opérations de pension en comparant un sous-échantillon des données de la CDS avec un troisième ensemble de données faisant directement état des modalités d'exécution. Les données dérivées sont utiles aux chercheurs qui étudient l'évolution de la structure et les conditions du marché des titres à revenu fixe.

Sujets : Méthodes économétriques et statistiques; Marchés financiers
Codes JEL : C55, C81, G10

1 Introduction

A repo, or sale and repurchase agreement, is a form of collateralized loan and an important source of short-term funding for large Canadian financial institutions. In a repo, one counterparty sells a security to a second counterparty and then repurchases the same security at a later date for the original price plus interest at an agreed-upon "repo rate." The repo market is one of Canada's core funding markets and a key market for the implementation of monetary policy (Garriott and Gray 2016; Fontaine, Selody and Wilkins 2009). Despite their importance, limitations in historical data before 2016 pose a challenge for researchers in their efforts to deepen their understanding of this market in Canada.[1]

Although historical data exist on Canada's repo market, the data do not contain critical information needed to compute quantities relevant for studying repos. The main source of historical securities transactions data before 2016 is the Canadian Depository for Securities (CDS), which is the central settlement authority for fixed-income securities (namely, bonds and bills). The CDS logs all transactions including repos and outright sales and purchases ("cash" trades). Repos are logged in two distinct settlements: the sale and the repurchase legs. From the perspective of the CDS, such logging is sufficient because it is concerned with settling transactions as opposed to any economic terms involved. Each settlement has information on trade and settlement time, the specific security issue, and prices and quantities. However, repo-specific details such as the repo rate, term, haircut and settlement are not logged. In particular, the data lack a unique identifier to connect the sale and repurchase legs to each other and allow one to compute the terms of trade of the repo. These variables are useful in analyzing, for example, how repos rates and quantities respond to market conditions and how other terms of trade respond to changes in market structure.

The goal of this work is to use the disaggregated data from the CDS to generate historical repo data that include key variables relevant to research of the repo market. This repo dataset will have labelled sale and repurchase legs that in turn allow the computation of repo rates, tenor, settlement delay, trade size, security issue, and haircut. This dataset is constructed by applying a novel algorithm that performs a global optimization to match the two transaction legs of a repo from the unlabelled securities transactions data.

[1] Since 2016, the Investment Industry Regulatory Organization of Canada has collected detailed data on repos from Canadian fixed-income dealers through the Market Trade Reporting System.

2 Similar problems

At a basic level, a repo-matching problem can be expressed in terms of matching members of one set to another. For repos, these sets are termed first and second legs of transactions. For clarity, in the first leg of a repo, a counterparty receives cash in return for some units of a bond; in the second leg, the same counterparty buys back those units of a bond for cash plus interest at an agreed-upon "repo rate." A comparable problem is the stable marriage problem, which asks how N men and N women can be paired given each party's preferences. Note that for repos, it is not known *a priori* which settlement is a first or second leg. This is equivalent, in the context of the stable marriage problem, to not knowing *a priori* whether someone is a man or a woman.

The Gale-Shapley algorithm guarantees a stable match for every couple in the stable marriage problem (Gale and Shapley 1962). In this algorithm, a man will propose to his highest-ranked woman, who will tentatively "accept" the proposal and be "engaged" to the man, such that he is her highest-ranked man so far. An unengaged man may propose to an engaged woman, and if she ranks him higher, the woman will trade up. The original fiancé will be no longer be engaged and will rejoin the pool of prospective suitors. This process continues until everyone is married, which is when no woman can trade up.

Our repo matching problem is similar to the stable marriage problem in that there exist two distinct types that will form pairs. In the CDS data, although settlements can be labelled as repos, it is not specified whether a settlement is a first or a second leg, nor who the counterparties are. The settlement labels are themselves occasionally inaccurate in that some repos appear to be labelled as cash trades, and vice-versa. From a practical perspective, a better analogy for the repo matching problem becomes a "stable roommates" problem (Irving 1985).

Stable matchings cannot be guaranteed for the roommates problem, as there is only the one set. With the CDS dataset, it is possible that a trade labelled a repo may not have a match for a variety of reasons, such as an error in labelling or because the repo was aborted and the second leg does not exist. For a repo matching problem, it must be acceptable to not match a given settlement in the event that no plausible partner settlement exists.

We build on prior work by Bulusu and Gungor (2021), who develop a similar algorithm to match repos using the CDS data. Other comparable work in economics has been done in particular for identifying uncollateralized loans using data from payment systems. The literature frequently follows Furfine (1999), who introduced an identification algorithm that consists of a series of filters, summarized as the following:

- finding a loan and repayment that involve the same set of counterparties and fall on different business days
- eliminating candidate pairs whose interest rate is unreasonable for the given market conditions
- ensuring the principal payment is in "round" increments as per market convention

Other researchers have suggested refinements to this approach, such as considering only interest rates that are integer values in basis points (bps) or are increments of 1/32 percentage points (Demiralp, Preslopsky and Whitesell 2006), or systematic tests for possible false positives involving *k*-nearest neighbours (Rempel 2016). Most similar previous work has been done with payments data, where counterparty information is available—unlike for the CDS data. However, a recent example of work on securities transaction data is a study by Garvin (2018), who reviewed Australian data to assess repo market structure using an algorithm similar to Furfine (1999).

3 Data description

The CDS dataset can be grouped by date into three vintages (2003–07, 2008–09, and 2009–present; vintage 1, 2, and 3 respectively), which represent different versions of the data supplied to the Bank of Canada by the CDS. The major distinction between these three vintages is slight variability in which key data channels are available (to be covered later in this section). As vintage 3 is the modern dataset and is being continuously synchronized between the CDS and the Bank servers, our description will focus on it, highlighting the subtle distinctions in the two historical vintages as needed.

The CDS dataset is a list of settlement instructions containing information that corresponds to transfers of securities and funds between two counterparties. Each row, corresponding to a single settlement, includes the traded bond characteristics, dates, and times relevant to the transactions as well as a variety of other miscellaneous attributes. For the purposes of matching the two legs of a repo transaction, we take a subset of these fields, along with some related or derived attributes, and store it in a relational database. The data used are listed in **Table 1**. The original data contain further fields that are not relevant to matching two legs of a repo; some examples of these omitted fields include the security's short name, currency or instrument type.

Vintage 1 data do not contain a trade type field. For the other vintages, a key step is to match for similar trade type (that is, cash or repo). While running the matching algorithm raises questions about the reliability of the trade type field as a truthful descriptor of the actual transaction,

experience suggests that cash and repo settlements will generally not be a valid match: that is, both legs of a repo will have the "cash" label if it is so mislabelled. Vintage 1 furthermore does not have an entry date time stamp—which gives the time that a trade was entered into the CDS database to the millisecond—but only a simple trade date.

Vintage 2 data differ from vintage 3 data in that they do not have a net_amount field: that is, the actual amount of cash transferred for a bond settlement. They have an original price (also known as the clean price), which is the cash value of the bond. This cash value is in turn converted to a dirty price, which accounts for the accrued interest. We use this dirty price in place of the net amount field for calculating the repo rate. There is some risk to using the original price, as the reporting convention is inconsistent and it may already be a dirty price. The original price may also be rounded, which affects the roundness of the repo rate. Unlike vintage 1, vintage 2 has both the trade type field and a precise time time stamp.

A more detailed remark is merited on the meanings of various trade types found in the CDS data. By way of example, consider the distribution of trade type for the entire vintage 3 dataset (**Table 2**). Most of the settlements are either cash (C)or repos, PRA or RPA. One peculiarity to note is that there are two labels—purchase resale agreement (PRA) and repurchase agreement (RPA) (Canadian Depository for Securities 2018)) for repo settlements, but they appear to be wholly interchangeable as far as matching pairs. Using these labels in a strict manner strongly worsened performance, whereas using these labels interchangeably produced pairs whose only "deficiency" was a mismatch on those labels. The rest of the labels do not tend to bear repo pairs.

Table 1: Data fields used for matching settlements as repos

Field	Data source	Description	Notes
Entry date time since epoch	Derived	Conversion of entry_date_time to Unix time	Using an integer time stamp, instead of a more complex date-time object, improves code readability and computational performance.
Entry date time	CDS	Date and time (to milliseconds) of entry of settlement into the CDS database; used as a trade time	It is possible this field represents when a settlement was entered into an older system, separate from the CDS database.
Value date	CDS	Settlement date	
Par quantity	CDS	Dollar amount of a bond paid at maturity	Maximum quantity is $50 million. Settlements can involve amounts greater than $50 million, but these will typically be split into chunks of

			$50 million plus a remainder, and only rarely into equal chunks.
Issue ISIN	CDS	International Securities Identification Number (ISIN) for bond	
Net amount	CDS	Actual amount of money in dollars and cents sent from one counterparty to another	
Coupon	Other	Bond coupon	A complete set of coupon dates can be used for correction of the interest rate if the repo spans a payment date. It is also useful for searching for false positives whose interest rate is related to the coupon payment.
CORRA	Other	Canadian Overnight Repo Rate Average	CORRA is used to calibrate the range of plausible repo rates.
Dirty price	Derived	Bond price including accrued interest	This price derives from original price. It corrects for ghost rates, and substitutes for net amount, which is missing in vintage 2.
Trade type	CDS	Type of trade, whether cash, repo or something else	The types are not strictly reliable; some settlements labelled cash are likely repos (that is, have all characteristics of repos except the trade type label) and some settlements labelled as repos are likely cash trades.
Master rowid	Derived	Primary key for a settlement	
Trade status	CDS	Whether a trade has been confirmed by both CDS members conducting the trade	The status implements restrictions on matching.

Note: CDS is Canadian Depository for Securities.

Table 2: Data fields used for matching settlements as repos

Trade type tag	Count	Percentage
A	91	0.0
C	25,330,650	73.3
DP	427,303	1.2
DPL	56,105	0.2
FR	2	0.0
NI	633,930	1.8
P	3,652	0.0
PRA	2,863,897	8.3
RPA	5,263,831	15.2
SPR	51	0.0
SPA	10	0.0

4 Matching algorithm

The basis of the matching algorithm, after matching on certain "hard" criteria, is that rankings are determined based on "soft" criteria, quantified by an affinity score, to establish which second settlement is most likely to be a matching repo pair for an initial settlement. A global search is done in which a given settlement is compared with all N as-of-yet unpaired settlements. As paired settlements are eliminated from further consideration (given that each leg would have been compared to all settlements, including paired settlements, at some point in the process), the complexity of the algorithm is at worst $O(N^2)$ with N decreasing, but not necessarily to zero, as the process executes. The algorithm is similar to the algorithm from Furfine (1999) but differs through:

- doing a global search and optimization for rate roundness and closeness in trade time subject to true-or-false filters for whether a pair can be a plausible repo
- focusing on the roundness of the repo rate as opposed to the amount loaned
- using securities settlements instead of payments data

Figure 1 presents a flowchart of the steps taken by the algorithm in assessing whether a candidate pair is matched as a repo. The flowchart shows a global search for a best matching partner for

Figure 1: Flowchart for matching algorithms

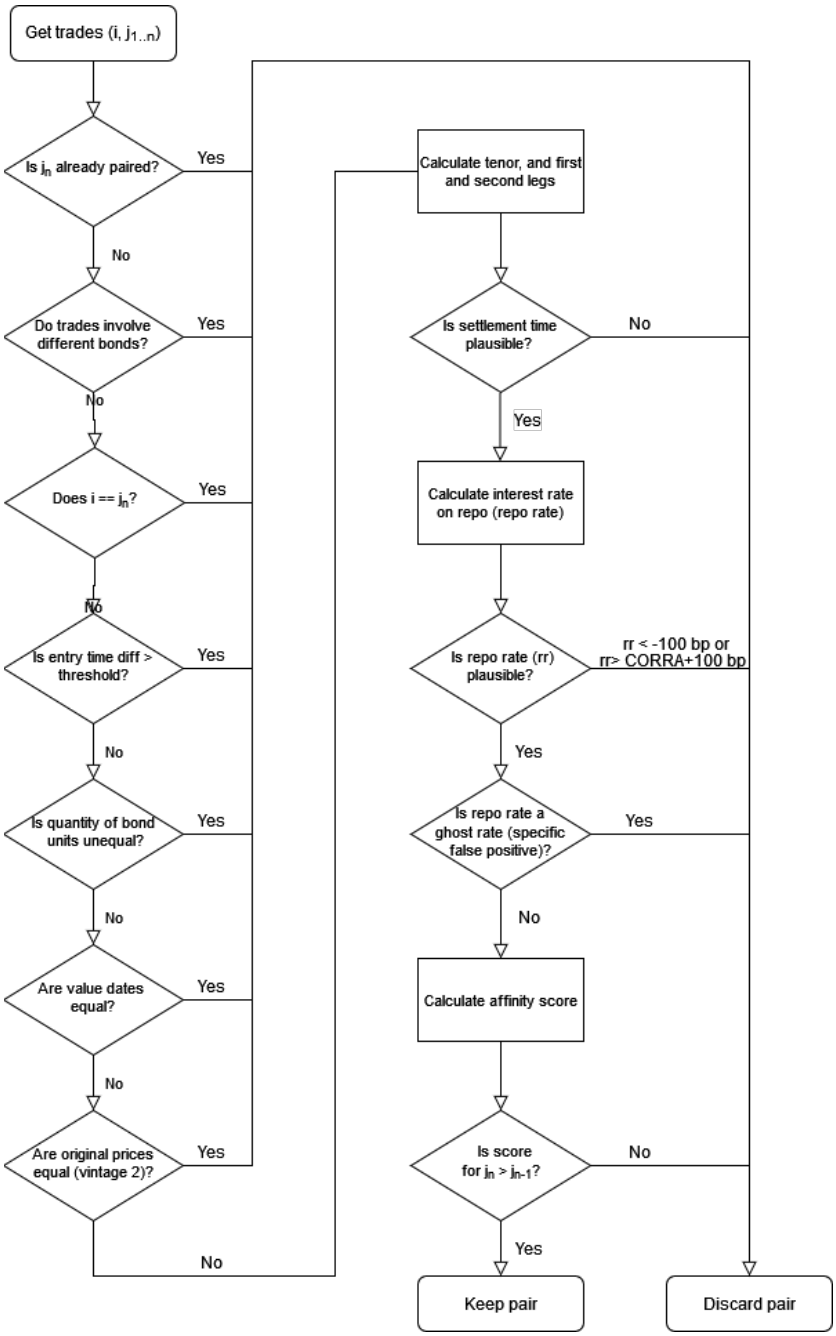
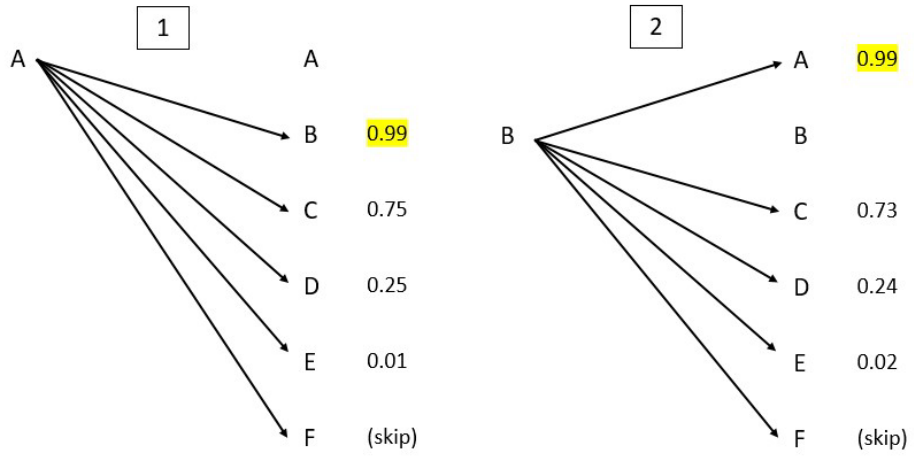


Figure 2: Example matching process for settlement A against five candidate partners



settlement *i* out of a pool of candidates $j_1...j_n$. This process is double-ended; once a partner *j* is found for settlement *i*, the same process repeats using settlement *j* as the base settlement and a pool of candidates ($i_1....i_n$). The (*i,j*) pairing is considered stable when both halves of the matching process return the same pair. If one of these settlements prefers a different settlement, the pairing is discarded. **Figure 2** shows an example of the two-sided matching process. In the first half, settlement A is compared with five candidate partners. Settlements B through E have progressively decreasing affinity scores; thus, B is more likely to be another repo leg relative to settlement A than any of settlements C through E. Note that settlement F is labelled as "skip": this notation signifies that settlement F was rejected at the hard criteria stage. Next, settlement B is compared with all the other settlements. Given that the preferred partner for settlement B is settlement A, these two settlements are considered a stable repo pair.

This process is run for all settlements. Once all settlements are exhausted, further rounds of this algorithm are run until a round returns no matched pairs. Further rounds will either identify pairs with very low affinity scores (unlikely to be a valid repo—type I error) or the consequence of matching for repos where the legs were split into chunks of $50 million in par quantity. This algorithm assumes that because each repo had to be somehow contracted, one other settlement should be a true fit. It is thought that type II errors (not matching for a repo where one exists) are uncommon; tests run on orphan settlements suggest that there is either no plausible partner settlement left (usually, after satisfying the hard criteria, the only possible settlements yield an unrealistic interest rate) or that a repo may be found by accounting for a coupon payment. Repos

that occur over coupon dates need different accounting, as there exists a convention that coupon payments are paid back to the *original* bond holder. This process is not always directly observable in the data but can be inferred from changes to the dirty price.

Several additional observations enrich Figure 1. At the fourth decision point ("is entry time diff > threshold"), the plausibility of the difference in trade time is assessed; in general, differences over two days are considered implausible. At the eighth decision point ("is settlement time plausible?"), the algorithm verifies that the candidate pair follows the usual convention of ≤ T1 settlement, although in practice, this part of algorithm was often relaxed to ≤ T5 settlement (T*n* settlement signifies that a trade settles *n* days after the transaction date). At the following decision point, the plausibility of the repo rate is assessed; the typical range for an acceptable repo rate is between -50 bps to the Canadian Overnight Repo Rate Average (CORRA) plus 100 bps. CORRA is a benchmark interest rate that serves as an estimate of average short-term general collateral repo rates.

A common source of false positive matches is cash-market trades with different settlement dates and close trade prices that produce plausible repo rates. We can use information from market conventions to catch some of these false positives. Clean prices of cash-market trades are commonly found at round increments of 1 cent. In contrast, clean prices used in repos tend not to have round increments since they are a function of the negotiated repo rates in increments of round bps. It is rare to find a round-increment repo rate where the clean prices of both matched legs have round increments of 1 cent. On a repo rate–time graph, these data points appear as a series of horizontal lines, resulting in the sobriquet "ghosts." We therefore exclude matches where the difference in clean prices between two legs is near a round number of cents. For most bonds, the following expression is checked for roundness such that when it evaluates between -0.006 to +0.006, the match is abandoned:

$$abs(dcp \text{ x } 200 \text{ x } round(dcp * 200))/2, \quad (1)$$

where

$$dcp = \text{dirty price leg2} - \text{dirty price leg1} - \text{term x coupon}/365. \quad (2)$$

For bonds with a coupon rate of zero (e.g., Government of Canada Treasury bills), we slightly modify the expression that is checked for roundness to account for market convention as follows:

$$abs(dcp \text{ x } 1000 - round(dcp \text{ x } 1000))/10. \quad (3)$$

A further explanation is merited for the affinity score, or the soft criteria. Most of the initial steps in the algorithm are in some way self-explanatory. For example, in our sample, the Canadian repo market functions by using specific bonds as collateral for each repo. Consequently, a repo pair may

not consist of settlements involving different bonds. Nevertheless, multiple candidate pairs will match based the initial hard criteria, which then have to be distinguished at the last decision point, that is, by the affinity score that combines the soft criteria.

By inspection, matching repo settlement legs are usually entered and scheduled in the CDS system within seconds, often with an interval that is less than one second. The second key observation is that repos are negotiated for interest rates with round (i.e., integer) basis point increments. The algorithm uses an affinity score that itself is a function of a score that reflects closeness in time and a score that reflects the closeness of a repo rate to an integer value in bps (i.e., roundness). In this way, the algorithm will run a global optimization for closeness in entry time and roundness to discriminate for more likely repo pairs that otherwise meet the hard criteria for a repo.

The affinity score is a function of two factors: closeness in entry time and the roundness of the repo rate weighted equally:

$$\textit{affinity\_score} = 0.5 * \textit{subscore\_(closeness in entry time)} + 0.5 * \textit{subscore (roundness of repo rate)} \quad (4)$$

These subscores are explained as follows. Recall that candidate pairs are rejected if the separation in trade time (dt) is greater than a threshold, usually two days. When calculating this subscore, we use the following process:

- If $dt < 10.0\ s$, dt is scaled linearly from 0.5 to 1, where a score of 1 is for $dt = 0$, and a score of 0.5 is for $dt = 10.0\ s$.
- If $dt > 10.0\ s$, an exponential decay function is fitted where the score is 0.5 for $dt = 10.0$ s and the score is 0.001 for $dt = 2$ days.

For the subscore for roundness, a similar approach is used. A residual is calculated as

$$\text{residual} = |\text{repo rate} - \text{round(repo rate)}|, \quad (5)$$

which is used as the subscore for roundness of the repo rate.

For example, if the repo rate is 99.9999 bps, the residual is 0.0001. The smaller the residual, the higher the score. We calculate as follows:

- If the residual < 0.01, it is scaled linearly from 0.05 to 10, where a score of 1 is for residual = 0 and a score of 0.05 is for residual = 0.01.

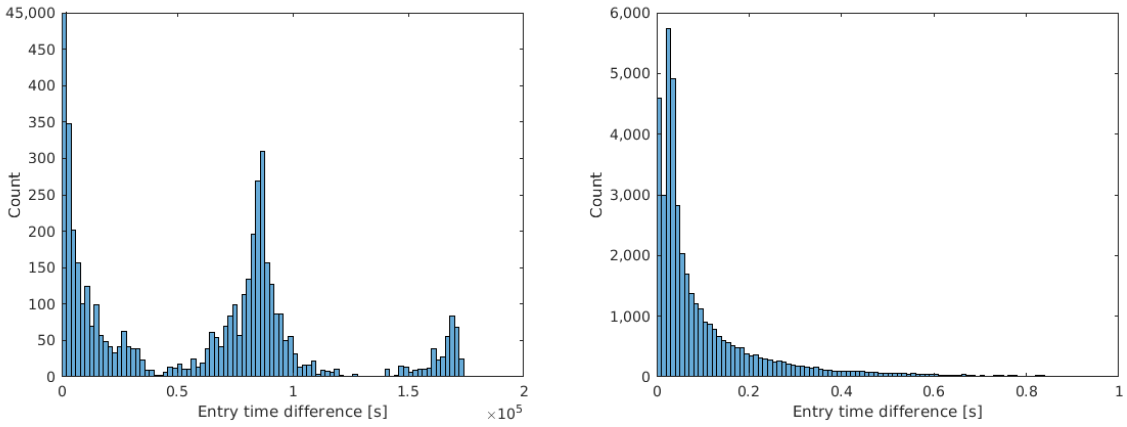
- If the residual > 0.01, an exponential decay function is fitted where the score is 0.5 for residual = 0.01 and the score is 0.01 for residual = 0.99.

The scaling functions and coefficient cutoffs are determined empirically through trial-and-error experiments. The goal of these experiments is to improve optimization for roundness of repo rate and closeness in trade time without creating too many false positive or false negatives. The piecewise scaling functions are a consequence of many candidate pairs having similarly small residuals and small time separations. The result of this optimization is seen in the distribution of entry time differences and roundness of a high-circulation Government of Canada bond (**Chart 1** and **Chart 2**).

Chart 1: Histograms of entry time difference for all repos using bond CA135087A610

a. Entry time differences less than 1 s.

b. Entry time differences for all repos; note the discontinuity in the y-axis



Note: Most repos have entry time differences less than 1 *s*, although some have greater difference, such as the cluster slightly prior to 1×10^5 *s*, which corresponds to a difference of 1 day (86,400 *s*).

Chart 2: Histogram of repo rate roundness for all repos using bond CA135087A610

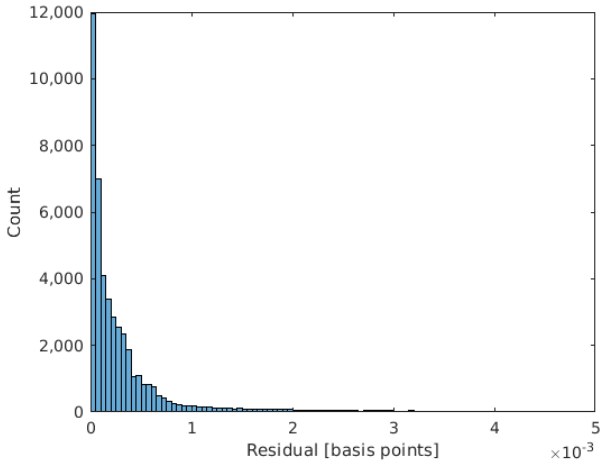
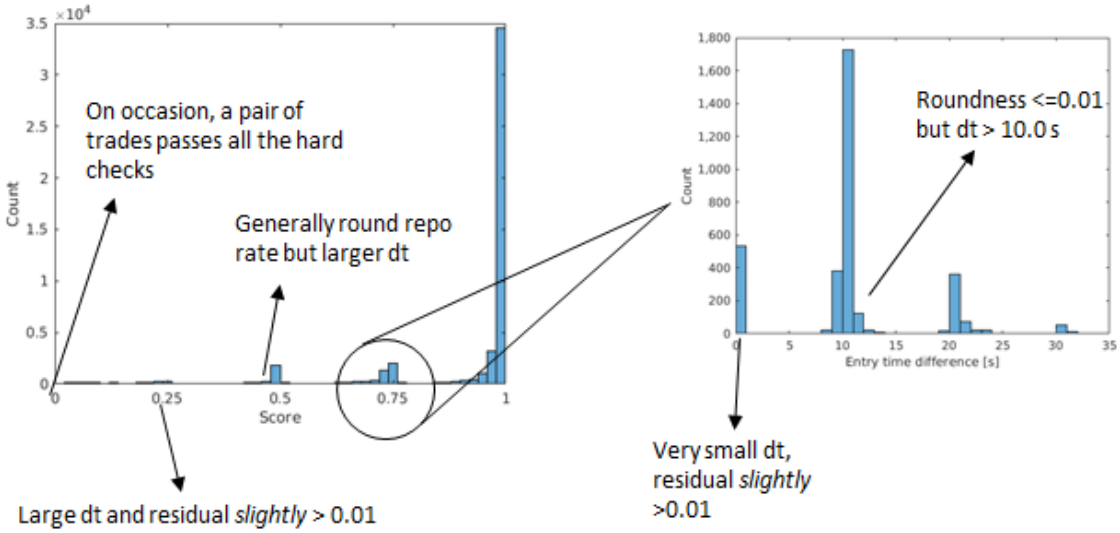


Chart 3: Distribution of affinity scores for all repos using bond CA135087A610

a. Histogram of all residuals (scores)

b. Histogram of entry time differences for repos with residuals of ~0.75.

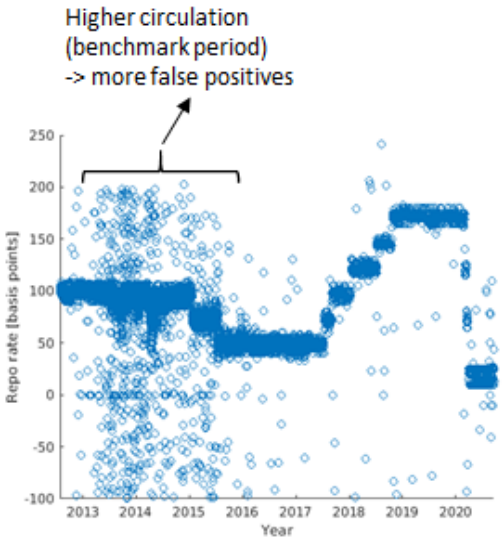


A consequence of the piecewise scaling functions is that it creates clustering of affinity scores around 1.00, 0.75, 0.50 and 0.25, which can have specific meanings (**Chart 3**). With a score of about 1.00, this signifies a very small residual and very small separation in entry time. At a score of 0.75, most of the repos will have a *dt* > 10.0 *s*, but not by much (there appears to be cluster in increments

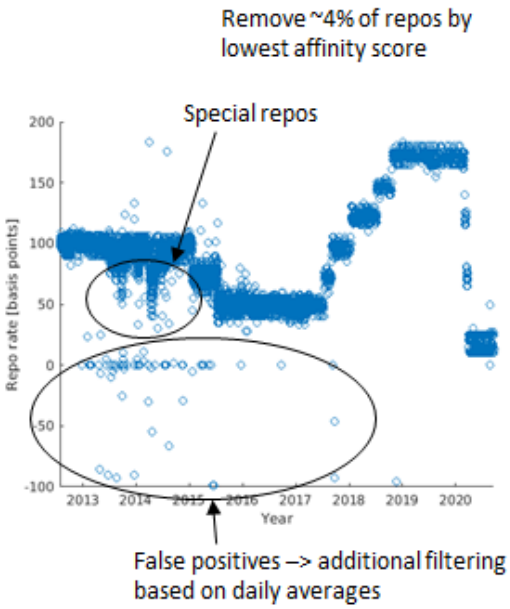
of 10 *s* for this bond), with a very round repo rate. In a minority of cases, they will have a residual that is slightly greater than 0.01 while still having a separation in time less than 1 *s*. At a score of 0.50, repos still generally have a round rate, but the separation in time increases more.

Chart 4: Repo rate over time for bond CA135087A610 with and without low-affinity-score pairs

a. Entire dataset



b. Dataset with low-affinity-score pairs removed

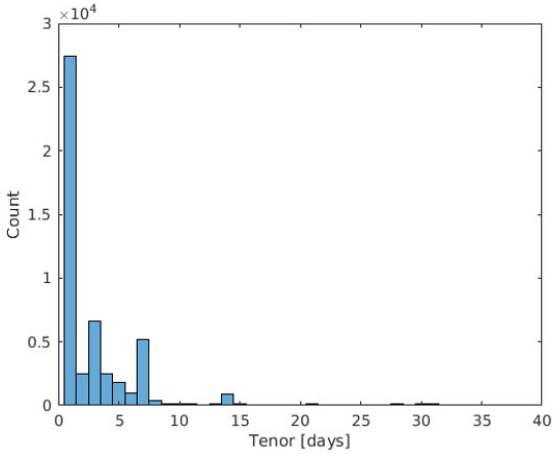


At a score of 0.25, there is always a large separation in entry time and a residual slightly above 0.01. The cases with residuals that are close to or slightly greater than 0.01 are possibly a result of net amounts being transacted to the dollar and not to the cent. Finally, when the score is much below 0.25, a smattering of repo pairs match all the hard criteria but otherwise do not have a strong affinity in terms of closeness in entry time or repo rate roundness. Such repos are likely false positives; removing 4% of the repos with lowest affinity scores will generally show a very consistent trend of the prevailing interest rate (**Chart 4**). Post-processing can also be applied by comparing the interest rate on a repo with the prevailing interest rate at the time. Outlier pairs with low affinity scores can be considered false positives.

Most repos are known to be overnight loans (one calendar day), over-the-weekend loans (three calendar days but also one business day) or over-the-week loans (seven calendar days). **Chart 5** shows the distribution of repo tenors (the difference in settlement dates) for bond CA135087A610. An early approach to the affinity score was to account for the tenor in the weighting. However, trial-

and-error work, as well as experimentation with a labelled dataset (i.e., repo pairs are already given and labelled), suggested that accounting for tenor in the affinity score at best did not improve classification and at worst resulted in false negatives and false positives.

Chart 5: Distribution of tenor (repo duration) for bond CA135087A610



To produce a final dataset for analysis, we perform the following additional steps. Only matches with scores greater than 0.2 are kept. This tends to eliminate false positive matches by inspection of, for example, **Chart 3**. We keep only matches with first legs with positive confirmation status (trade was not cancelled). The CDS has confirmed the necessity of trade confirmation for the first leg of a repo. Outliers are removed by dropping matches with repo rates more than 1.5 standard deviations from the weekly average for a given bond. Haircuts are computed by comparing the dirty price with a reference market price from a commercial dataset from FTSE-Russell at the International Securities Identification Number (ISIN) date level. To protect a lender against counterparty credit risk, we set the amount lent as less than the market value of the bond collateral. The difference between the market value of collateral and the loan amount, in percentage terms, is known as the haircut and computed as

$$haircut = 100 * dirty\ price\ leg\ 1/dirty\ price\ reference \quad (6)$$

and rounded to integer percentage points per market convention. The "dirty price reference" is the market value for the bond.

We take a step to improve estimates of repos' trade volume since the CDS splits settlements of size volume greater than $50 million par quantity into smaller settlements. Settlement volume is aggregated across matches with the same:

- ISIN, tenor, trade date and first leg settlement date

- dirty price, rounded to the nearest .0001 cents
- repo rate, rounded to the nearest basis point

Where entry time is available, aggregation excludes repos with gaps in first leg entry time greater than 20 seconds.

4.1 Verification of the algorithm

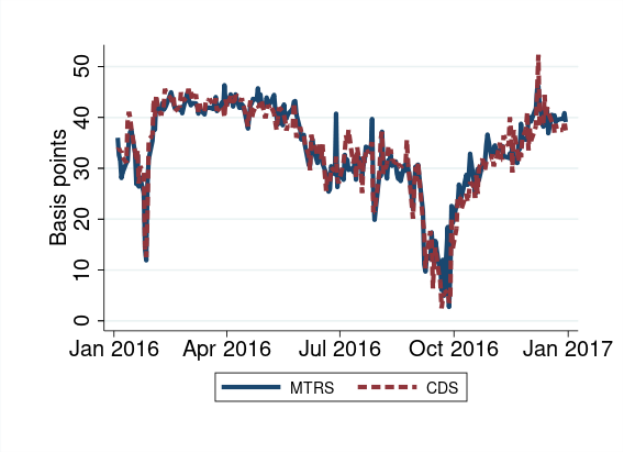
A labelled dataset was acquired from the Canadian Derivatives Clearing Corporation, which is a clearing house that clears a subset of repos conducted by Canadian dealers before they are settled in the CDS. The data are similar to the CDS sample but include an additional unique identifier for the two matching legs of a repo. Running the algorithm on this labelled dataset resulted in 100% replication of the labels when the bond coupons were accounted for. In some cases, a repo leg was split in increments of $50 million par quantity where there was some permutation of the labels. We ran a script to demonstrate that the permutations involved only identical leg segments. For clarity, consider two legs each split in two parts, such as $A = \{A_1 , A_2\}$ and $B = \{B_1, B_2\}$; the set $\{\{A_1, B_1\}, \{A_2, B_2\}\}$ is equally valid as $\{\{A_1, B_2\}, \{A_2, B_1\}\}$.

5 Example of results

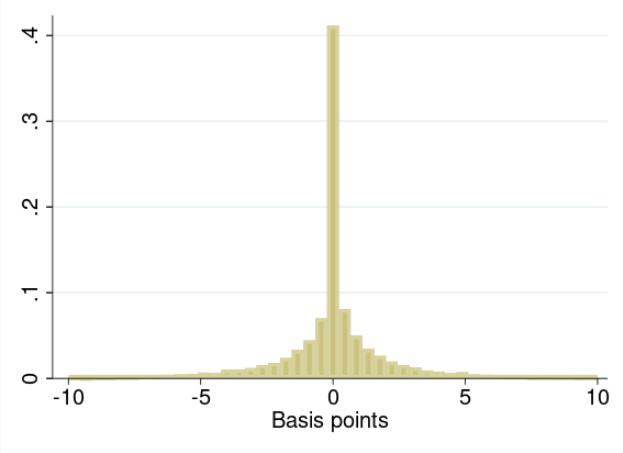
While the main focus of this document is to describe the algorithm used to find repo pairs in the CDS dataset, we can make a few brief exemplary observations about the possible uses of the data. **Chart 6** shows a comparison between the average repo rate time series obtained from the Market Trade Reporting System (MTRS) and the analyzed CDS data. MTRS data consist of all repos conducted by Canadian Government Securities Distributors and is collected by the Investment Industry Regulatory Organization of Canada. The overlapping sample period is from 2016 to 2018. Upon visual inspection, it is readily apparent that both time series are highly similar. A distribution of the daily difference in repo rate, shown in Chart 6, panel b, shows that our matching is unbiased (a mean difference of 0.1 bps) with low error (a standard deviation of 2.6 bps). **Chart 7** demonstrates consistency with the market convention for overcollateralized repos of 2%, as derived from the CDS data.

Chart 6: The average repo rate time series as determined from the MTRS data and CDS , with the distribution of differences between the two datasets

a. MTRS and CDS repo rates for a particular Government of Canada bond

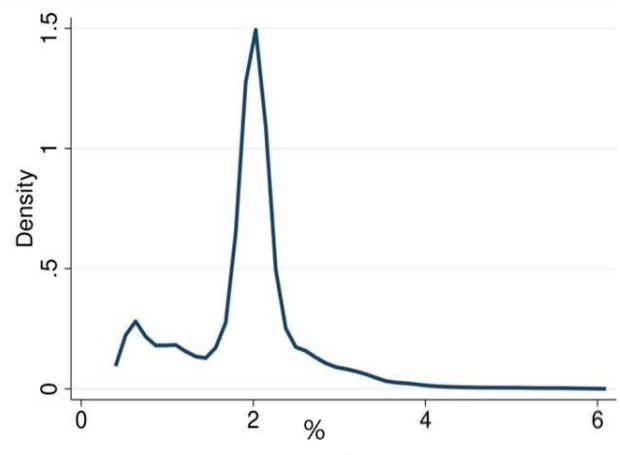


b. Distribution of daily differences in repo rate between MTRS and CDS data



Note: MTRS is the Market Trade Reporting System, and CDS is the Canadian Depository for Securities.

Chart 7: Distribution of repo haircuts for positive haircuts



Note: The distribution is estimated using a kernel density estimator.

6 Conclusion

We develop an algorithm to obtain matched repo pairs from CDS trading data. This algorithm initially runs candidate pairs through a series of true-false filters, also known as "hard" criteria, which assess the plausibility of a pair of securities settlements being matching repo legs. A global optimization is then done based on a set of "soft" criteria to eliminate otherwise plausible pairs from consideration. When two possible legs most prefer each other, a stable pairing is found. Comparison to a dataset that already has securities settlements labelled with the actual repo tags found a 100% correspondence, subject to allowing the permutation of identical repo legs for cases where more than $50 million in par quantity were transacted, which is a known limitation of the CDS settlement. The present work has a generated a rich dataset of labelled repos, which can be used in the future to study the evolution and possible future of the Canadian repo.

A caveat is needed for the accuracy of aggregate repo volume. Since vintages have differing levels of accuracy, aggregate volume may be biased to varying degrees. We recommend using the resulting data to study historical terms of trade and relative volume over time, but not aggregate volume.

References

Bulusu, N. and S. Gungor. 2021. "The Life Cycle of Trading Activity and Liquidity of Government of Canada Bonds: Evidence from Cash, Repo and Securities Lending Markets." *Canadian Journal of Economics/Revue canadienne d'économique* 54 (2): 557–581.

Demiralp, S., B. Preslopsky and W. Whitesell. 2006. "Overnight Interbank Loan Markets." *Journal of Economics and Business* 58 (1): 67–83.

Fontaine, J.-S., J. Selody and C. Wilkins. 2009. "Improving the Resilience of Core Funding Markets." Bank of Canada *Financial System Review* (December): 41–46.

Furfine, C. H. 1999. "The Microstructure of the Federal Funds Market." *Financial Markets, Institutions and Instruments* 8 (5): 24–44.

Gale, D. and L. S. Shapley. 1962. "College Admissions and the Stability of Marriage." *American Mathematical Monthly* 69 (1): 9–14.

Garriot, C. and K. Gray. 2016. "Canadian Repo Market Ecology." Bank of Canada Staff Discussion Paper No. 2016-8.

Garvin, N. 2018. "Identifying Repo Market Microstructure from Securities Transactions Data." Reserve Bank of Australia Research Discussion Paper No. 2018-09.

Irving, R. W. 1985. An Efficient Algorithm for the Stable Roommates Problem. *Journal of Algorithms* 6 (4): 577–595.

Rempel, M. 2016. "Improving Overnight Loan Identification in Payments Systems." *Journal of Money, Credit and Banking* 48 (2-3): 549–564.