WRDS Corporate Bond Database:
Data Overview and Construction Manual

WRDS Research April 2017

1. Introduction

The WRDS Bond Database is a unique cleaned database for US Corporate Bond research. It incorporates two feeds: FINRA's TRACE (Trade Reporting and Compliance Engine) data for bond transactions, and Mergent FISD data for bond issue and issuer characteristics.

The objective of the WRDS Bond Database is to allow researchers to easily and effectively use a comprehensive source of corporate bond information. The database has the following three major components:

- Cleaned Corporate Bond Transaction Data:
    - "Cleaned" bond transaction data for both Trace Enhanced and Trace Standard data
- Individual Corporate Bond Return:
    - Report returns calculated using both "clean" and "dirty" bond prices
    - Various other transaction based characteristics (e.g., Volume, Yield, Spread)
- Bond CRSP Linking Table:
    - Linkage between bond data to equity data provided by CRSP

Our main source of corporate bond transaction-level data is TRACE, covering over-the-counter transactions for the US corporate bond market. Two separate data packages are provided by FINRA: TRACE Standard (Market Data) and TRACE Enhanced.

1.1 TRACE Standard

The Standard TRACE package was launched in July 2002, and was deployed to full scale in multiple stages:

- July 1, 2002: included only Investment Grade bonds > $1bn in original issue size
- March 1, 2003: added Investment Grade bonds rated A and higher that were > $1mil in original issue size
- April 14, 2003: added remaining Investment Grade bonds
- Sept 30, 2004: added all other Investment Grade bonds and high yield bonds

In addition to the phased launching scheme, the Standard data package has another feature worth mentioning. The trading volume of bonds reported is capped at certain level: 5 million cap for investment grade bonds and 1 million cap for high yield bonds. In other words, if a high yield bond is traded at 10 million on a given day, all the user will see from the Standard data package is simply a capped volume of 1 million.

1.2 TRACE Enhanced

As an improvement over the Standard data package, TRACE later introduced the Enhanced data, which is very similar to the Standard data, but has more coverage and less bias. More specifically, all TRACE-eligible securities (except Rule 144A) have been included since day one, July 2002. This uniform dissemination offers many more trade records that span the entire over-the-counter market from early on. Trade volumes are reported accurately; they are not capped at certain levels based on bond ratings. However, the downside of the Enhanced data package is that this information is only available at an 18 months' delay. Therefore, in spite of the fact that it contains richer information than the Standard data package, if researchers are in need of the most recent corporate bond transaction data, Enhanced alone is not sufficient.

2. Data Filtering Steps

TRACE data by FINRA is a platform that brings transparency to the US corporate bond market, enabling individual investors and market professionals to access the bond transaction information.

As the records in TRACE are self-reported by bond dealers, errors are anticipated in the data. When filers realize that a mistake was made in the filing, TRACE allows them to file a new trade report to correct the mistake. These correction reports can be filed the same day as the original report, or they can be filed days later.

The three most common types of errors in the trade reports are:

- Cancellation:
  - Voids a previous same-day trade report and does not replace it with updated information
- Correction:
  - Voids a previous same-day trade report and replaces it with updated information
- Reversal:
  - Voids a previous trade report (filed earlier than the reversal filing date) and replaces it with updated information

There are several other factors that may affect the quality of the trade report. For instance, in the TRACE Enhanced data package, agency transaction reporting conventions can lead to “double counting” in the trade records. There could also be other minor corrections necessary to address data issues such as missing prices/volume/date ranges, etc. Ultimately, it is clear the *raw* TRACE data, whether it’s Enhanced or Standard, needs to be cleaned before the information can be used to calculate bond returns and other trade-based research items. The WRDS Bond Database follows the data cleaning procedures outlined in Asquith, Covert and Pathak (2013) and Dick-Nielsen (2009, 2014) to clean TRACE Enhanced and Standard data.

The TRACE Enhanced and TRACE Standard data packages incorporate somewhat different reporting systems and field names. We will begin by using the TRACE Enhanced version of the data package to illustrate the data cleaning process.[1] The logic behind the cleaning process for both the Enhanced and Standard data packages, however, is essentially the same.

[1] For data manual for TRACE Enhanced pre 2012/02/06, please refer to the data manual at: http://www.finra.org/industry/trace/historic-file-layout. For data manual of TRACE Enhanced post 2012/02/06, please refer to the manual at: http://www.finra.org/industry/trace/historic-data-02062012.

Table 1: Outline of TRACE Enhanced cleaning steps

TRACE Enhanced Cleaning Steps		Observations
0	TRACE Enhanced Raw Data as of September 2016	145,720,692
1	Pre 2012/02/06 Format	
1.1	Initial Sample	99,281,350
1.2	Remove Cancellation (C) and original Trade (T) reports by matching 7 keys[2] and MSG_SEQ_NB	
1.2	Remove chained correction (W) and original Trade (T) reports[3]	
1.3	Remove Reversal (asof_cd=R) and original Trade (T) reports[4]	
	Percentage Cleaned = 6.3%	93,016,614
2	Post 2012/02/06 Format	
2.1	Initial Sample	46,439,342
2.2	Remove Trade Cancellation (X), Cancelled Correction (C) and their matched Trade (T) reports	
2.3	Remove Reversals by matching 7 keys and MSG_SEQ_NB	
	Percentage Cleaned = 5.7%	43,797,014
3	Remove double counting of agency trades (keep only Sell records by Dealers)	
	Percentage Cleaned = 22.8%	
4	Final Cleaned TRACE Enhanced Sample	105,569,765
	Total Percentage Cleaned = 27.6%	

The TRACE Standard data package incorporates a slightly different reporting format and different variable fields, but the logic of data cleaning is the same: the goal is to clear erroneous reports due to Cancellations, Corrections and Reversals.

[2] 7 keys include: Cusip_ID, Execution Date and Time, Quantity, Price, Buy/Sell Indicator, Contra Party

[3] A Correction (W) Report W_2 for a particular bond could be filed to correct a previous Correction Report W_1, which was in turn used to correct an original Trade Report (T). Therefore, in the cleaning process it is important to clean all the intermediate correction reports, on top of the original Trade Report and the final Correction Report.

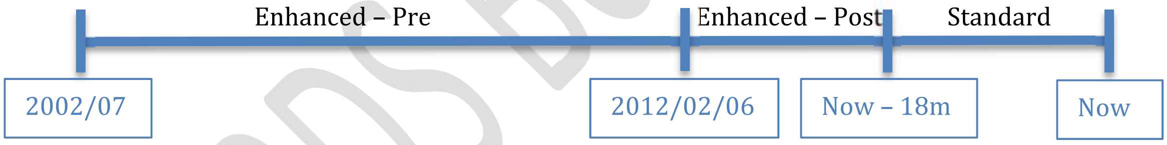
[4] While matching by all 7 keys produces the most strictly matched reversal and original trade report pairs, some reversal reports are left without a matched trade report as the execution time stamp is not always entered correctly. Hence, we drop the time dimension of the trade characteristic key, and only match by the remaining 6 keys (Cusip_ID, Execution Date, Quantity, Price, Buy/Sell Indicator, Contra Party). If more than one trade is mapped using the 6 keys linkage, we remove the earliest trade reports.

Table 2: Outline of TRACE Standard cleaning steps

TRACE Standard Cleaning Steps		Observations
0	TRACE Standard Sample as of September 2016	110,949,773
1	Remove Cancellation (C) and matched Trade (T) reports: matching by CUSIP_ID, TRD_EXCTN_DT, TRD_EXCTN_TM, RPTD_PR, ASCII_RPTD_VOL_TX, MSG_SEQ_NB	
2	Remove Correction (W) and matched Trade (T) reports: link chain of correction cases: by CUSIP_ID, TRD_EXCTN_DT, MSG_SEQ_NB	
3	Remove Reversals (asof_cd=R) and matched Trade (T) reports: matching by CUSIP_ID, TRD_EXCTN_DT, RPTD_PR, ASCII_RPTD_VOL_TX, RPT_SIDE_CD	
4	Final Cleaned TRACE Standard Sample	104,802,971
	Total Percentage Cleaned = 5.5%	
Note	No need to clean double counting in agency trades in TRACE Standard	

3. Return Construction

Although TRACE Enhanced has the benefit of accurate volume information, its eighteen-month lag in data is a downside relative to the TRACE Standard. In order to create the longest and most updated return and other aspects of time series data for WRDS Bond Database, the following supplementing is incorporated in the final output:



Whenever possible, TRACE Enhanced is used as the primary data source for computing bond returns. For the most recent time period where TRACE Enhanced data is not yet available, we use the transaction data obtained from TRACE Standard. The timeline above illustrates the arrangement.

While TRACE provides all transaction related bond data, information regarding other bond characteristics is extracted from the Mergent FISD database. More specifically, FISD provides information on bond type, issue and maturity date, coupon rate and payment frequency, bond rating, default and reinstatement date if applicable, etc.

Once we combine the transaction data from TRACE and bond characteristics data from FISD by matching 9-digit CUSIP, we are able to construct the bond return database.

Using the Bond Type information from FISD, we further restrict our bond sample to meeting the following criteria:

- Coupon Type equals Fixed or Zero (not variable coupons)

- Not under Rule 144a
- Bond Type is equal to US Corporate Convertible (CCOV), US Corporate Debentures (CDEB), US Corporate Medium Term Note (CMTN), US Corporate Medium Term Note Zero (CMTZ), or US Corporate Paper (CP).

Unlike the equity market, the corporate bond market does not display a high level of trading activity, and many bonds do not trade more than once per month. Therefore, we construct the WRDS Bond Return database only at monthly frequency. The key pricing related variables available from the database are as follows:

Table 3: Key variables from WRDS Bond Return Database

Variable Name	Description
PRICE_EOM	Last price at which bond was traded in a given month (e.g. 8/31, 8/28, 8/4)
PRICE_LDM	Price on last trading day of the month if available, missing if bond didn't trade on that day (e.g. 8/31)
PRICE_L5M	Last price at which the bond was traded in a given month, if that day falls within the last 5 trading days of the month (e.g. 8/31, 8/30, …, 8/25)
RET_EOM	Monthly return calculated based on PRICE_EOM and accrued coupon interest
RET_LDM	Monthly return calculated based on PRICE_LDM and accrued coupon interest
RET_L5M	Monthly return calculated based on PRICE_L5M and accrued coupon interest
T_VOLUME	Total par-value volume in a given month
T_DVOLUME	Total dollar volume traded in a given month
T_SPREAD	Average trade-weighted bid-ask spread
T_YLD_PT	Average trade-weighted yield point

Note the three price variables in the table. With these three forms of monthly "clean" bond prices, we then construct the three monthly bond returns (EOM, LDM and L5M) as month-over-month percentage price change of the corresponding "dirty" prices, where the "dirty" price is the "clean" price plus the accrued coupon interest between coupon payment dates. All bond return measures are winsorized at top and bottom 1% to filter out extreme returns.

3.1 Discussion on Different Price and Return Variables

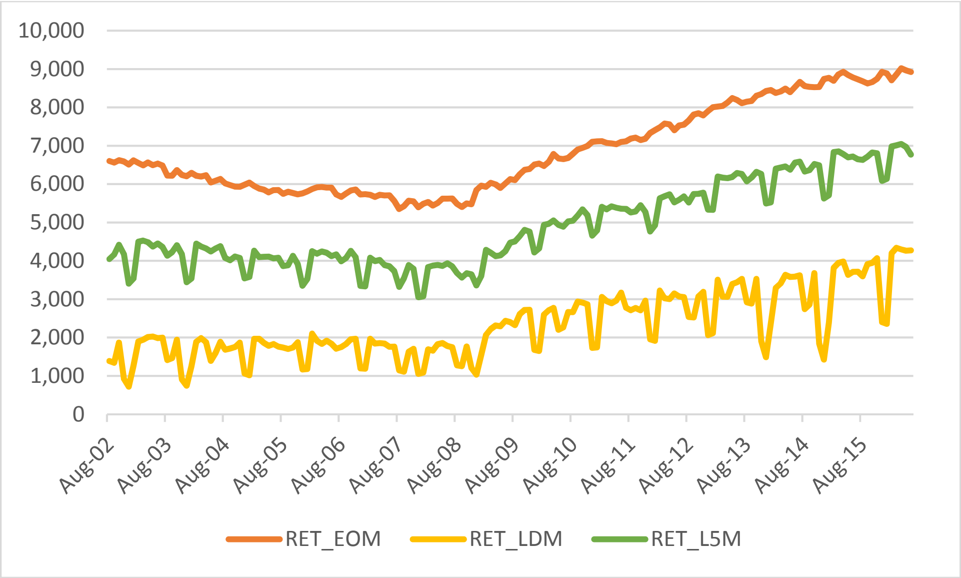
As explained above, three different versions of month-end bond prices are created, and hence three returns are calculated based on the prices. It is worth pointing out that due to the nature of the different month-end prices, the three returns measures are populated very differently.

Not surprisingly, the most well-populated return measure is RET_EOM, as it only requires capturing the last bond trade price in a month, with the greatest flexibility which day this trade may fall onto. For example, if a bond was last traded on the 10th of a month, this recorded price would be used for RET_EOM calculation.

In contrast, RET_L5M and RET_LDM impose stricter requirements, with RET_LDM being the most strict, as it requires a bond to trade exactly on the last trading day of a given month. As a result of this restriction, these two return variables are less populated relative to RET_EOM.

Figure 1 demonstrates the difference in the coverage of these three different bond returns. Notice that the orange line (RET_EOM) has a significantly larger number of non-missing returns relative to the other two measures.

Figure 1: Time Series of Monthly Bond Returns



Another observation worth of highlighting from Figure 1 is the "seasonality" of the RET_LDM and RET_L5M measure, where the number of non-missing returns for these two measures dips seasonally. This dip occurs during the month of December and January.

As the measure RET_LDM requires the bond to trade exactly on the last trading day of the month, this day falls in December either on New Year's Eve or the weekend before the New Year. Bonds trade significantly less on this last trading day of the year.

Figure 2: Number of Bond Prices on Last Trading Day by Calendar Month

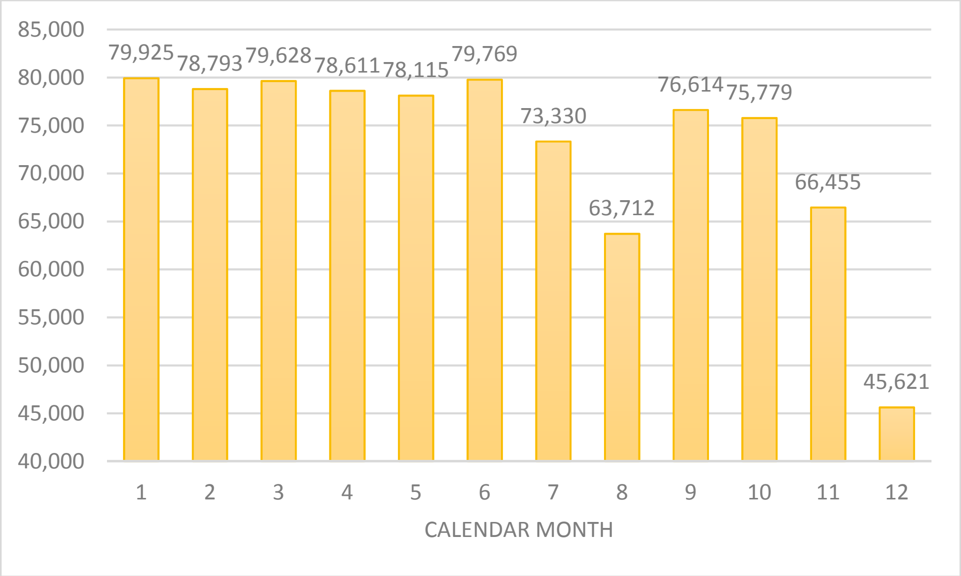
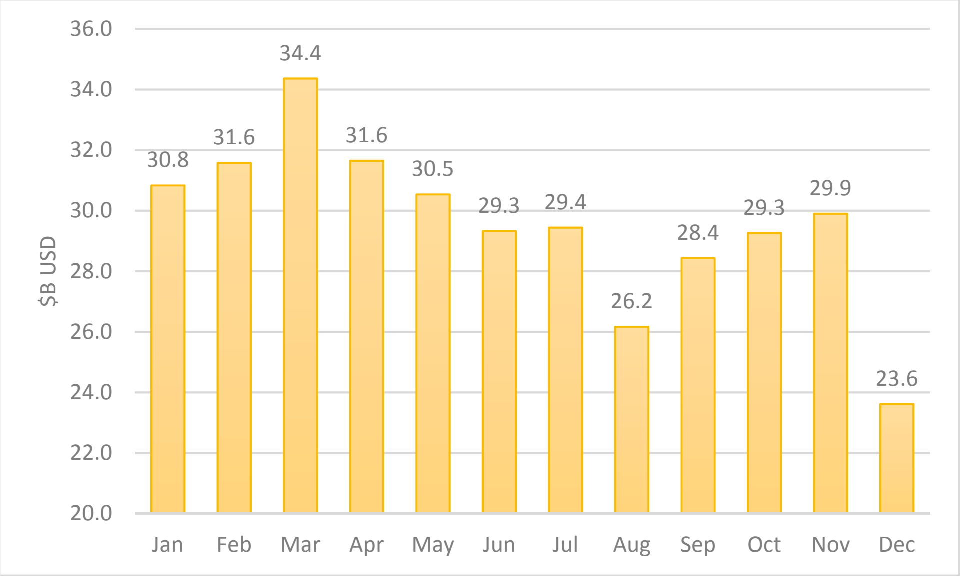


Figure 2 plots the number of non-missing bond prices on the last trading day of each calendar month. As we expect, the month of December has the lowest non-missing bond prices on last trading day. We obtained the average daily bond trading volume from SIFMA of 2016 by calendar month (Figure 3), and the evidence supports the observation we see in our sample.

Figure 3: Average Daily Corporate Bond Trading Volume by Calendar Month 2016



Source: SIFMA http://www.sifma.org/research/statistics.aspx

Using the same logic, as RET_L5M requires valid bond prices within the last five trading days of the month, and given the fact that the last 5 trading days of December fall around the Christmas and New Year holidays, the trading volume during that particular window is much thinner compared to other months, resulting in missing PRICE_LDM and PRICE_L5M for December. Furthermore, as the return measures of January rely on the month-end prices of December, when PRICE_LDM and PRICE_L5M are missing for December, the corresponding return measures are missing for January.

In addition to the pricing and return metrics, the WRDS Bond Return Database also pre-calculates various other monthly bond transaction-related variables, including Total Volume, Total Dollar Volume, Average Bid Ask Spread, and Trade Weighted Yield.

3.2 Bond Default Returns

Like equities, corporate bonds occasionally default prior to reaching maturity. If default returns are simply treated as missing observations, return estimates can be overstated, particularly for high-yield bonds. To address this potential return bias, we follow Cici Gibson and Moussawi (2017) methodology in computing a composite default returns for all defaulted bonds.

Unlike stocks, bond issues continue trading after merger and acquisitions events, which leaves defaulting to be the main delisting event for bonds. Before calculating monthly bond returns, we generate post-default prices for any bonds that defaulted. We search for any price information on defaulted issues after the default event. In our investigation, we were able to find pricing information on 492 defaulted issues out of the 1000+ issues that defaulted after July 2002.

We computed the median return on these defaulted issues in the (-1, +1) month window around the default date and found the median is equal to -40.17% for defaulting Investment Grade issues and -17.67% for defaulting High Yield issues, which reflect higher expected default probability for high yield ex-ante.[5] We then included these delisting return averages as proxies for delisting returns for all defaulting issues at the month of default, and we dropped all observations for defaulted issues after their first default event. Using the in-sample composite default-month returns for defaulting bonds sharing similar credit quality一but without valid post-default pricing information 一 enables us to avoid delisting bias that has been documented in previous research on equity returns (Shumway (1997)).

4. Linking Bond Data with CRSP Equity Data

The WRDS Bond Database also includes a data package that enables a linkage between the corporate bond data and the equity data issued by the same company. This linkage is important for empirical research as it allows users to directly link fixed income data at the individual bond level to the equity data from the CRSP database.

4.1 TRACE Identifiers

In TRACE Enhanced data packages, four identifiers are reported:

- CUSIP_ID (9-digit)
- BOND_SYM_ID (Bond Symbol)
- BSYM_ID (Bloomberg ID)
- COMPANY_SYMBOL

The first three are at individual corporate bond level, while the last one is at the company level. CUSIP_ID and BSYM_ID are mostly permanent identification that follow a bond throughout its lifespan. BOND_SYM_ID, on the other hand, can change over time for various reasons (e.g., M&A, issuing company Ticker change). Therefore, in order to trace a bond correctly over time, we use the most commonly used permanent identifier CUSIP_ID as the primary identification.

At each trade report in TRACE, both CUSIP_ID and COMPANY_SYMBOL are recorded, therefore providing us with a snapshot of CUSIP_ID/COMPANY_SYMBOL pairs historically. While most CUSIP_IDs are associated with only one COMPANY_SYMBOL, a small fraction are associated with multiple COMPANY_SYMBOLs. Table 4 below reports the distribution of CUSIP_ID – COMPANY_SYMBOL linking pairs, showing that over 120,000 observations have only one mapped COMPANY_SYMBOL throughout the lifespan, while 39 CUSIP_IDs are linked to four COMPANY_SYMBOLs throughout time.

Table 4: Distribution of CUSIP_ID – COMPANY_SYMBOL links

No of company_symbol	No of Obs
1	121,307
2	5,191
3	510
4	39

There are various reasons why a particular bond can experience change in corresponding COMPANY_SYMBOL. Two main reasons are M&A and Ticker Change. For example, the bond with cusip_id=000886AE1 experienced

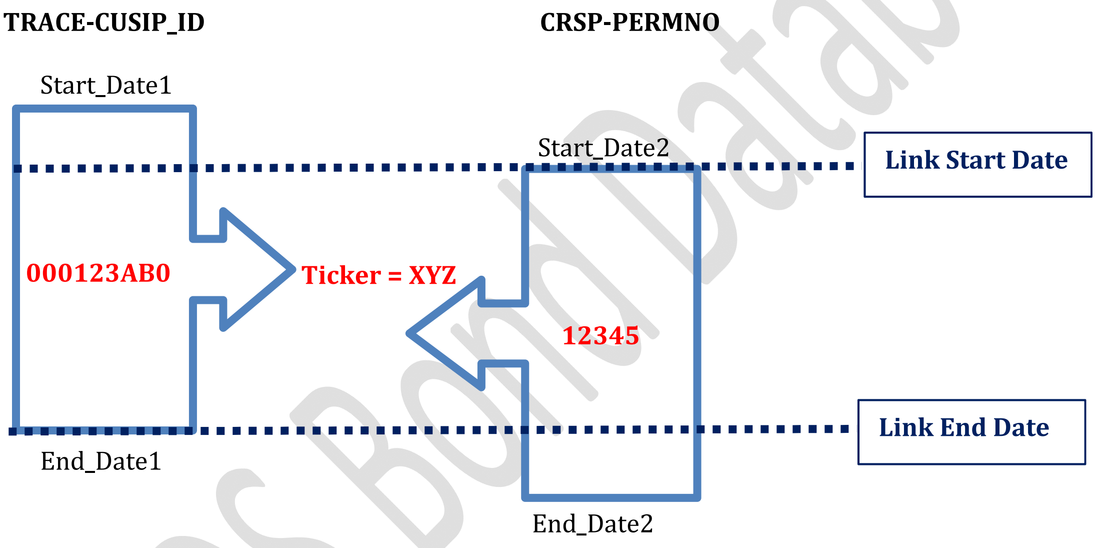
[5] The figure of median return by investment grade is calculated based on the sample period of 2002/07 - 2016/03.

a company symbol change from “ADCT” to “TEL” due to the acquisition of ADC Telecom (Ticker: ADCT) by Tyco Electronics (TEL). In another example, the bond with CUSIP_ID=126117AM2 had its company_symbol change from “CNA” to “L” in August 2012. This was purely due to a trading symbol change by the holding company.

4.2 Linking Logic

We use the trading symbol of the issuing company as a bridge to link the corporate bond identifiers (cusip_id) to the equity identifier (permno). From the TRACE side, we obtain the date range for a given cusip_id to be associated with a company_symbol. We repeat the same exercise for the CRSP side to build a time range for each permno and corresponding trading symbol. We then match the two data sources by the trading symbol[6], and impose a conservative date range, where the link start date is set to be the latter of the two start dates in TRACE and CRSP, and link end date is set to be the earlier of the two end dates in the two databases.

Figure 4: Link Start and End Dates



Occasionally, a false duplicate matched pair can be created following this logic. This is primarily due to the recycling of trading symbols and non-synchrony in symbol updating in TRACE and CRSP. Table 5 illustrates an example of such a false duplicate match.

Table 5: Example of False Matching

CUSIP	COMPANY_SYMBOL	PERMNO	PERMCO	Link_StartDt	Link_EndDt	LinkWk
023551AM6	AHC	28484	20064	2002/07/01	2006/05/08	100
023551AM6	AHC	92528	52924	2008/02/11	2009/10/14	44
023551AM6	HES	28484	20064	2009/10/22	2014/12/29	136

In May 2006, Amerada Hess Corporation changed its name to Hess Corporation, and hence updated the trading symbol from AHC to HES. The ticker AHC was later reused by A. H. Belo Corporation starting in January 2008.

[6] We use trading symbol match as our primary matching criteria. For the unmatched ones, we also use a fuzzy name matching algorithm to create secondary linkage by company names. The trading symbol matching method counts for over 98% of the matched pairs.

TRACE database, on the other hand, was slower to update the ticker change from AHC to HES for the Hess Corporation, and hence led to the false match of CUSIP=023551AM6 to PERMNO=92528.

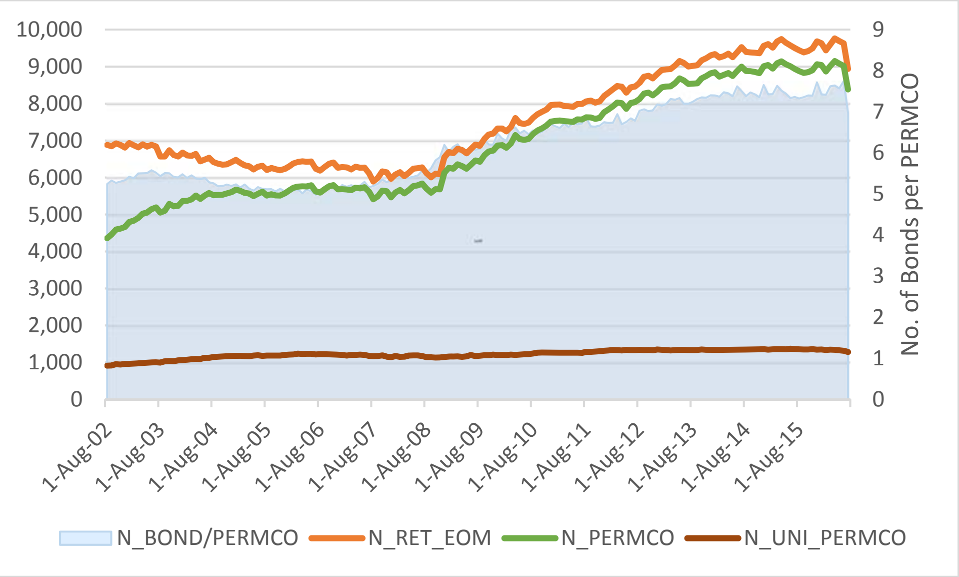
To address this issue, we calculate the link duration of each matched pair. Assuming the latency in updating trading symbols will eventually be addressed, a false match should, in theory, have a shorter link duration than the true link. Therefore, if more than one linked PERMNO is found for a given CUSIP/Company_Symbol combination, we keep only the one with the longer link duration. So for the example illustrated above, since the link duration for the second observation is only 44 weeks, thereby shorter than the 100 weeks for the first observation, only the first linked pair (and the third pair) will be kept in the final linking table. The variable descriptions for this final linking table are presented in Table 6.

Table 6: Variable Description of Bond-CRSP Linking Table

Variable Name	Description
CUSIP	CUSIP ID at individual bond level
STARTDT	TRACE Start Date
ENDDT	TRACE End Date
PERMNO	CRSP Permno
PERMCO	CRSP Permco
NAMEDT	CRSP Start Date
NAMEENDT	CRSP End Date
LINK_STARTDT	Link Start Date
LINK_ENDDT	Link End Date

Figure 5, below, illustrates the coverage of bond records with valid PERMCO information. The N_RET_EOM line represents the number of bonds non-missing RET_EOM each month. We then join the bond records with CRSP PERMCO info using the linking table. The N_PERMCO line reports the number of bond records with non-missing RET_EOM that have matched PERMCO. Lastly, the N_UNI_PERMCO presents the unique number of PERMCOs observed in this valid return sample per month, which is a little over 1000.

Figure 5: Coverage of Bond Records with PERMCO



Lastly, the difference between the number of bond records and number of unique PERMNOs is explained by the shaded area, N_BOND/PERMCO. This series reports the average number of bonds (bond CUSIPs) that each PERMNO linked to over the entire sample. On average, a PERMCO is associated with approximately six bond CUSIPs at a given time, and this number grows over time.

5. List of Data Tables

The following SAS datasets are contained in the ***/wrds/bond/sasdata/*** directory. Some datasets can also be accessed through the web query. Here is a description of the two tables available as part of the WRDS Bond Database:

Table 7: WRDS Bond Database Table Description

Table	**Description**	**Source**	**Access**
BondRet	Monthly pricing (price, return, volume, etc.) and bond characteristics data	Created by WRDS	TRACE + Mergent FISD
BondCrsp_Link	Linking table of bond cusip and CRSP Permno	Created by WRDS	TRACE + CRSP

6. Reference

Asquith, Paul and Covert, Thomas R. and Pathak, Parag A., The Effects of Mandatory Transparency in Financial Market Design: Evidence from the Corporate Bond Market (September 4, 2013). Available at SSRN: https://ssrn.com/abstract=2320623 or http://dx.doi.org/10.2139/ssrn.2320623

Cici, Gjergji, Scott Gibson, and Rabih Moussawi, 2017, "Explaining and Benchmarking Corporate Bond Returns," Working Paper.

Dick-Nielsen, Jens, Liquidity Biases in TRACE (June 3, 2009). *Journal of Fixed Income*, Vol. 19, No. 2, 2009. Available at SSRN: https://ssrn.com/abstract=1424870 or http://dx.doi.org/10.2139/ssrn.1424870

Dick-Nielsen, Jens, How to Clean Enhanced TRACE Data (December 3, 2014). Available at SSRN: https://ssrn.com/abstract=2337908 or http://dx.doi.org/10.2139/ssrn.2337908

Shumway, Tyler (1997), The Delisting Bias in CRSP Data*. The Journal of Finance*, 52: 327–340. doi:10.1111/j.1540-6261.1997.tb03818.x