ISSN 1936-5330

kc FED Research Working Papers

The U.S. Syndicated Loan Market: Matching Data

Gregory J. Cohen, Melanie Friedrichs, Kamran Gupta, William Hayes, Seung Jung Lee, W. Blake Marsh, Nathan Mislang, Maya Shaton, and Martin Sicilian

December 2018
RWP 18-09
https://dx.doi.org/10.18651/RWP2018-09

Federal Reserve Bank of Kansas City

The U.S. Syndicated Loan Market: Matching Data

Gregory J. Cohen°, Melanie Friedrichs†, Kamran Gupta★, William Hayes§,

Seung Jung Lee*, W. Blake Marsh‡, Nathan Mislang⊤, Maya Shaton*, and Martin Sicilian*

December 2018

Abstract

We introduce a new software package for determining linkages between datasets without common identifiers. We apply these methods to three datasets commonly used in academic research on syndicated lending: Refinitiv LPC DealScan, the Shared National Credit Database, and S&P Global Market Intelligence Compustat. We benchmark the results of our match using results from the literature and previously matched files that are publicly available. We find that the company level matching is enhanced by careful cleaning of the data and considering hierarchical relationships. For loan level matching, a tailored approach based on a good understanding of the data can be better in certain dimensions than a more pure machine learning approach. The R package for the company level match can be found on Github at https://github.com/seunglee98/fedmatch.

JEL Classification: C55, C88, E44, G21
Keywords: bank credit, syndicated loans, probabilistic matching, company level matching, loan level matching.

We thank Mary Chen, Danno Lemu, Nicholas Stewart, and Cristhian Vera for excellent research assistance. We thank Mark Carey for many helpful discussions and seminar participants at the Federal Reserve Board and the Federal Reserve Bank of Kansas City for comments. All remaining errors and omissions are our own. This work was completed while Friedrichs, Gupta, Hayes, and Mislang were employed at the Federal Reserve Board. The views expressed are our own and not the views of the Federal Reserve Bank of Chicago, the Federal Reserve Bank of Kansas City, the Board of Governors of the Federal Reserve System, nor anyone else associated with the Federal Reserve System. Some of the data used here are confidential and were processed solely within the Federal Reserve System.
Affiliations:
° Federal Reserve Bank of Chicago, 230 S LaSalle St, Chicago, IL 60604
† Stern School of Business, New York University, 44 West 4th Street, New York, NY 10012
★ Booz Allen Hamilton, 901 15th St. NW, Washington, DC 20005
§ Columbia Law School, Columbia University, 435 West 116th Street, New York, NY 10027
* Board of Governors of the Federal Reserve System, 20th and C Sts., NW, Washington, DC 20551
‡ Federal Reserve Bank of Kansas City, 1 Memorial Drive, Kansas City, MO 64198
⊤ Economics Department, Arts and Sciences, Cornell University, 109 Tower Road, 404 Uris Hall, Ithaca, NY 14853
Correspondence: blake.marsh@kc.frb.org

1 Introduction

Over the last decade, the syndicated loan market has proven to be a robust laboratory for exploring corporate finance and banking topics due to its unique features and the availability of loan level data. Highly influential papers have explored such fundamental corporate finance topics as asymmetric information and loan pricing (Lee and Mullineaux, 2004; Sufi, 2007; Ivashina, 2009), borrower reputation (Beatty, Liao, and Zhang, 2015; Ivashina and Kovner, 2011; Chaudhry and Kleimeier, 2015), and the balance sheet effects of loan covenant violations [Chava and Roberts (2008), Nini, Smith, and Sufi (2012), Roberts and Sufi (2009)]. Additionally, researchers have investigated issues related to financial stability [Ivashina and Scharfstein (2010), Aramonte, Lee, and Stebunovs (2015)], monetary policy transmission [Ippolito, Ozdagli, and Perez (2013), Cohen, Lee, and Stebunovs (2016), Lee, Liu, and Stebunovs (2017)] and the effect of credit market shocks on firm employment [Chodorow-Reich (2014)] and investment [Correa, Sapriza, and Zlate (2012)].

Syndicated loan research tends to be driven by three key datasets: Refinitiv LPC DealScan, the Shared National Credit database, and the S&P Global Market Intelligence Compustat database. While each of these are valuable in their own right, in combination they can provide a comprehensive look at loan pricing at origination, changes in loan terms and ownership over time, and borrower balance sheet health, respectively. Indeed, many researchers have combined these datasets in their work, though most are not publicly available and the matching methods are unknown. Those that are available, most prominently Chava and Roberts (2008)'s Refinitiv LPC DealScan–S&P Global Market Intelligence Compustat match, which forms the basis of most syndicated loan research, are infrequently updated and, therefore, may be insufficient for answering research questions on recent developments in the syndicated loan market.

This paper attempts to address these limitations by providing methods to match observations across datasets that lack common identifers. The paper has four key goals. First, we hope to reduce the manual matching burden that researchers face when assembling matched datasets. Second, we hope to update existing matched datasets using these methodologies to further the research agenda. Third, we hope to provide a common methodology for assembling syndicated loan market data that will provide a benchmark for research going forward. Finally, our methods can be applied to more generic string matching problems outside the scope of syndicated lending that involve linking different information from varying sources for a common business entity. Thus, we seek to encourage dialogue among researchers by proposing open and replicable methods for constructing research datasets.

The remainder of the paper is organized as follows: Section 2 reviews major data sources available for conducting research using syndicated loan data. Section 3 describes the general matching methods we have developed. Section 4 discusses the application of these methods to the Refinitiv LPC DealScan–S&P Global Market Intelligence Compustat link first provided by Chava and Roberts (2008). Section 5 discusses the incorporation of company hierarchical information into

the company level match algorithm. Section 6 discusses the application of these methods to the SNC–DealScan link, which provides a new loan level data match. Section 7 concludes.

2 Syndicated Loan Data Sources

Academic research on the syndicated loan market commonly relies on three data sources. Refinitiv LPC DealScan is a commercially available dataset that provides information on syndicated loan originations. These data are widely used by both academics and market practitioners. The Refinitiv LPC DealScan data are frequently paired with firm level balance sheet data from S&P Global Market Intelligence Compustat using the links provided by Chava and Roberts (2008). Researchers at federal bank supervisory agencies have access to data from the Shared National Credit (SNC) program, which records outstanding syndicated loans at various intervals. Each of these datasets have specific strengths, so, in combination, their various weaknesses can be overcome. Each dataset is discussed in turn below.

2.1 Refinitiv LPC DealScan

Refinitiv LPC DealScan ("DealScan") is the most commonly used syndicated loan market database due to its commercial availability. DealScan identifies loan originations from borrower SEC filings, arrangers and other lenders, and various public sources. Lenders are willing to submit data for use in constructing "league tables" which are bank level rankings of new originations that are helpful for attracting new investment banking clients. DealScan's data collection began in the early 1990s and has since been backfilled with prior year originations [Ivashina (2009), Murfin and Pratt (forthcoming)]. The current coverage includes loans originated as early as 1981 and the database is updated quarterly.

DealScan collects loan pricing and covenant information observed at origination. Participant information and borrower balance sheets are available for a subset of loans. For private borrowers, company information is not available for all records

2.2 S&P Global Market Intelligence Compustat

S&P Global Market Intelligence Compustat ("Compustat") is a commercially available database of publicly listed company filings. The data include annual company filings back to the 1950s and quarterly statements beginning in the 1960s. The data primarily include company income, balance sheet, and supplementary information. Compustat data standardize the reporting across items and filing types. Compustat includes only publicly listed company information, which narrows the scope of the data sample when matched to loan level sources. Compustat includes broad debt categories such as short or longterm debt and other liabilities, but does not have an indicator specifically for "bank debt." Compustat's balance sheet data is detailed and provides information on borrower

health and ability-to-repay when combined with loan level datasets. An immediate challenge one encounters when attempting to match borrowers to Compustat data are company name changes over time. Compustat provides current company data whereby new company names and locations overwrite historical company names and locations. Researchers' needs for historical company data has been addressed through the Center for Research in Security Prices, CRSP/Compustat Merged Database available through the Wharton Research Data Services ("CCM") which includes historical information taken directly from Compustats ftp files.[1] Details on the construction of the historical Compustat data used in the analysis can be found in Appendix A.

2.3 Shared National Credit Program

Researchers at federal bank regulatory agencies – the Federal Reserve, the Office of the Comptroller of the Currency, and the Federal Deposit Insurance Corporation – have access to a confidential, loan-level supervisory dataset. The Shared National Credit (SNC) database has been collected annually since 1993 and contains all outstanding syndicated commitments of at least $20 million that include three or more federally regulated syndicate members. The largest market participants have submitted quarterly data on all syndicated loans on which they are an agent since 2009:Q4.

Usage of the SNC database overcomes some of DealScan's major limitations. Specifically, the SNC database includes the full syndicate participant membership for each recorded loan as well as each participant's outstanding and committed shares of the loan. Moreover, because regulators collect the same loan across time, syndicate membership changes can be observed in these data. In addition, the database includes both lender generated and regulatory loan ratings. However, unlike in DealScan, little pricing or covenant information is available on most loans. Borrower financial information is also not available.

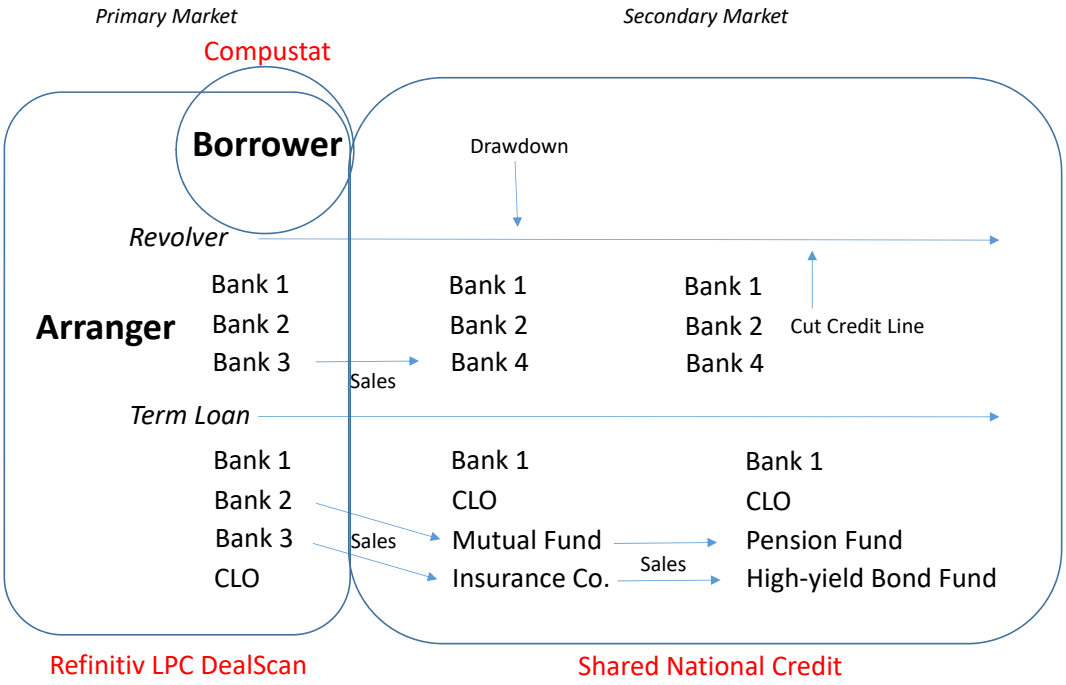
Figure 1 shows each dataset's respective syndicated loan market specialty. Compustat has firm level information. DealScan has package and facility level information associated with those firms at origination. Shared National Credit provides an understanding of revolving credit behavior through the evolution of credit line commitments and drawdowns through time. SNC also provides secondary market information, which particularly affects term loans, by tracking the syndicate participant distribution of individual credits over time. Participant tracking includes both banks and nonbanks that are active in the secondary market.

3 Methodology

Our methods are informed by earlier work on probabilistic record matching by Fellegi and Sunter (1969) and, in particular, by Winkler (2006). Traditional probabilistic record linkage, as discussed by Sayers, Ben-Shlomo, Blom, and Steele (2016), defines a matchscore for all possible record pairs

[1] Documentation on the CCM Database can be found here.

Figure 1: The Syndicated Lending Market and Relevant Data



across datasets based on the match success rate and comparisons across common fields. First, given common fields i across different datasets, define an indicator variable $\gamma_{i,j}$ for each record pair j. This indicator takes a value of one when the fields i match across the two records and zero otherwise as shown in equation 1.

$$\gamma_{i,j} = \begin{cases} 1 & if \ \ match \\ 0 & otherwise \end{cases} \tag{1}$$

Together, these field-specific indicators define the field agreement set γ_j for each record pair j across all common fields used in matching from $i = 1, 2, ..., N$ as shown in equation 2.

$$\gamma_j = [\gamma_{1,j}, \gamma_{2,j}, ..., \gamma_{N,j}] \tag{2}$$

Assuming conditional independence, the m–probability (defined as m_j) is the probability that a

record pair j has agreement set γ_j given that record pair j is a true a match ($M_j = 1$). Similarly, the u–probability (defined as $u-j$) is the probability that record pair j has agreement set γ_j given that record pair is not a true match ($U_j = 1$). The m–probability (u–probability) can be calculated as the product of the conditional probability that field i matches ($\gamma_{i,j} = 1$) given that the record pair j is a true match (not a true match). Intuitively, the m–probability represents how well the fields i identify true matches, while the u–probability represents the extent that fields i coincidentally match across records. The matchscore for record pair j, R_j, is defined as the likelihood ratio of the m– and u–probabilities when the fields match and one minus these probabilities when the fields do not match. These definitions are given in Equations (4) – (5).

$$m_j = \prod_i P(\gamma_{i,j}|M_j = 1) \tag{3}$$

$$u_j = \prod_i P(\gamma_{i,j}|U_j = 1) \tag{4}$$

$$R_{i,j} = \begin{cases} \frac{m_j}{u_j} & if \ \ \gamma_{i,j} = 1 \\ \frac{1-m_j}{1-u_j} & if \ \ \gamma_{i,j} = 0 \end{cases} \tag{5}$$

The matchscore for each record pair j is typically expressed as a log likelihood ratio as shown by equation 6. Matches can be determined by setting acceptable matchscore thresholds.

$$matchscore_j^{PRL} = \sum_i log_2 R_{i,j}. \tag{6}$$

Probabilistic record linkage methods have been very successful when employed to match records across personal names, and applications are readily found in medical records and census data [Gill (1999), Winkler (2006), Méray, Reitsma, Ravelli, and Bonsel (2007)]. For our purposes, however, these methods are not ideal. First, many common fields used for matching across personal names– such as family names, city of birth, or birth year– are conditionally independent which is an important assumption of the Fellegi and Sunter (1969) model. Company name matching, on the other hand, often relies exclusively on address and industry information. Address fields– such as state or zip code– are almost perfectly correlated while many industries are concentrated within certain geographic areas, closely linking industry classification to address information.

A more immediate problem is the assignment of the conditional probabilities. Generally, these probabilities are calculated on a pre-matched dataset that is known to be very high quality. While Chava and Roberts (2008) have provided a verified DealScan–Compustat matched dataset, no comparable SNC–matched datasets were available. Therefore, determining match success rate probabilities requires first hand–matching a minimum number of records.[2] Perhaps a more problematic

[2]For a small subset of loans, we use CUSIPs available in both DealScan and SNC to identify a training sample.

issue is that data errors and disagreements are common across similar fields in these datasets. For example, in company–level datasets, names may have minor differences, timing of reporting name changes may vary, or borrowers may be reported at different hierarchical levels. For loan level datasets, reporting lags may result in loan amount changes between origination and SNC reporting, and key loan dates– such as origination or maturity dates– may also differ by days, weeks, or even months. Given this high variability, we should expect that m-probabilities will be low, resulting in low matchscores.

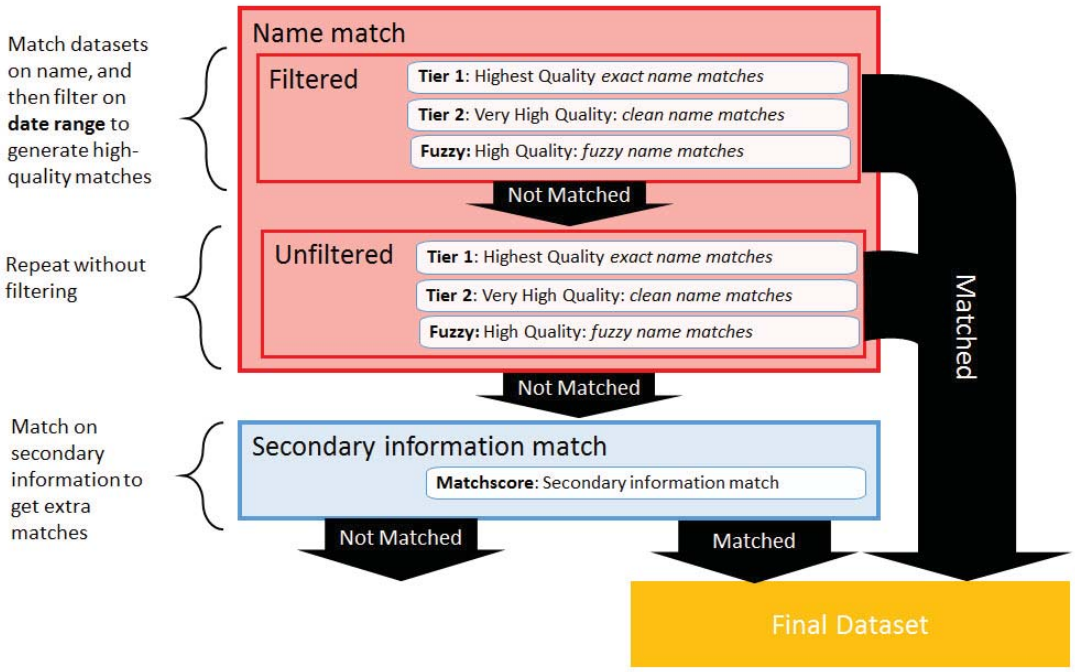
To overcome these issues when generating a matched dataset, we use rigorous data cleaning, exact matching, and probabilistic record linkage methods. Using these methods, we develop a "waterfall" approach by generating matched subsets of data where the subsets are defined by gradually looser match identification criteria. While each successive set of criteria produces a larger potential match set, the overall confidence level in those matches is lower. The algorithm, as constructed, removes observations matched in the current step before moving to the next, looser criterion. This process is iterated until all observations are matched or further loosening criteria would result in a match of unacceptable quality. This tiered approach to generating match sets assists hand verification, if desired, by providing the researcher a general idea of each potential match's quality and allows the researcher to discard matches below a given threshold. Conceptually, the approach outputs a set of potential matches along with summary information on the match quality allowing researchers to make informed decisions on the match validity. While this general framework applies to all matches we construct, the details, which are discussed below, vary for company and loan level matches.

3.1 Company Level Matching Methodology

Company level matching refers to constructing company pair matches across datasets. In our case, we hope to match sets of loans originated to a given company to the balance sheet data of that company collected by Compustat. To do so, we develop a two stage matching method that pairs traditional string matching techniques with probabilistic record linkage methods. In the first stage, company names are increasingly scrubbed of extraneous information and then matched across datasets using exact and fuzzy matching methods. In the second stage, we identify potential matches using non–name information– such as location, industry, or trading index identifiers– to generate a matchscore and filter accordingly. The overall approach is summarized in Figure 2.

The first stage of the company level match has several substeps based on the string cleaning methods outlined in Sayers, Ben-Shlomo, Blom, and Steele (2016). First, all alphabetic characters are lower–cased. Second, special characters with common word meanings are replaced with their respective word. For example, symbols such as "$", "%", and "&" become "dollar", "percent", and "and", respectively. Third, punctuation is removed. Fourth, common abbreviations are identified using a word frequency list and then replaced with their long-form words. For example, "co" is

Figure 2: Match Methodology



replaced with "company" while "corp" is replaced with "corporation". Finally, multiple spaces and tabs are reduced to single spaces and leading and trailing blanks around the string are removed. These actions standardize elements common to company names but do not significantly alter the string's composition. The two databases are merged on the standardized strings to complete the first substep.

Additional string cleaning is performed on a more judgmental level using the remaining set of unmatched records in the second substep. Common words found across databases, such as those standardized in the first substep, are eliminated on the basis that these words are uninformative when judging potential matches. While common word removal is a typical string cleaning technique, it presents a unique problem in a company match application. As an example, one may decide to remove the common words "american", "international", "group", and "incorporated" for all corporate names because, individually, each of these words are uninformative. However, when combined, they identify a large multinational insurance company. If removed collectively, then the AIG, Inc. string is empty.[3] To remedy this situation, we sequentially remove common words and perform a match after each drop. In our experience, this sequential dropping results in more correct

[3]Additional examples from the DealScan and Compustat data abound. The clothing store, "The Limited" is one example. In addition, there may be only one unique identifying word so that the two companies cannot be identified separately. For example, Zimmer Corp and Zimmer Inc both reduce to Zimmer when corporation and incorporated are removed.

matches because there are fewer empty strings generated and fewer duplicate matches because more identifying information is kept. In each pass through of the common words, we perform an exact match using the cleaned string.

The third substep consists of using fuzzy matching techniques to generate near matches that are not found using exact merges. We have found that the Jaro–Winkler method performs relatively well with a high degree of confirmed matches when the penalty parameters are set appropriately [Jaro (1989), Winkler (1989)]. By default, we have selected the parameters conservatively so that the algorithm errs on the side of falsely rejecting true matches rather than returning doubtful or highly suspect matches. A result of this decision is that fuzzy matching returns relatively few matches, but with more experimentation and confirmation, this element could be more powerful.

Throughout the first stage for the Compustat–DealScan match, we use date information to separate high and low confidence matches. During the first pass over the data, we apply a filter that removes any matches where the facility start date does not lie between the first and last instance of that company record in Compustat. In practice, this filter is rather conservative because company names change frequently and those changes are not well aligned across datasets. Therefore, we run two iterations of the first stage: once with the filter and once without.

The second stage attempts to match any remaining unmatched company names using secondary information such as company tickers, industry codes, and address information. Match quality is lower in this stage compared to exact name matching because secondary information fields are less unique. For example, companies that share geographies and industries– such as financial firms on Wall Street or oil companies in Houston– may all match to the same industry and zip codes. To determine matches in this stage, matchscores are calculated using distances between company names and other common fields across the two datasets. Raw distance measures depend on the field's data type: Jaro–Winkler string distance for names, number of intervening days for dates, and indicators for most other variables. Raw distance between industry and zip codes is proportional to the number of subsequent equal digits. Distances are normalized to a $[0, 1]$ interval and weights are judgementally assigned due to the lack of a pre–matched, verified dataset in most cases. The matchscore formula is shown in Equation (7) as the sum of the standardized distance ($norm.\ distance$) times the applicable weight (w_i) across all comparison variables i.

$$matchscore^{Fedmatch} = \sum_{i=1}^{N} norm.\ distance \times w_i \tag{7}$$

3.2 Loan Level Matching Methods

Loan level matching is a fundamentally different, and more difficult, problem than company level matching. In a company level match, there is one clear primary piece of information, the company name. While secondary information may agree or disagree, company names must be similar or the

match success is unlikely. This single primary variable suggests an obvious ordering for the company level match; exact and fuzzy matches on increasingly cleaned strings followed by matches based on secondary information. Loans, on the other hand, are not characterized by a single datapoint and there is no obvious informational hierarchy or order structure. Instead, multiple data points– including borrower name, loan amount, origination and maturity dates, loan type, and lender– must roughly agree.

Further, syndicated loan contracts are complicated and parsing them into relational databases is difficult. Differences in data sources and collection methodologies across datasets increase the possibility of differences across common fields for the same loan. For example, loans recorded in DealScan are more likely to have approximate loan amounts and loan dates. Syndicated loan contracts are also frequently structured as multiple loan facilities within a single package. Primary information for facilities originated in the same package is often similar and difficult to differentiate. Finally, syndicated loan contracts are often amended or renegotiated, causing primary information to differ between a loan's origination date and its later records.

Confronted with these challenges, we combine established matching methods with dataset specific techniques in two broad steps. First, a "loan match candidate set" is constructed using borrower names matched across each loan level dataset using the methods outlined in Section 3.1. In this step, however, the fuzzy match parameters are loosened significantly, more string content is removed, and importantly, observations are rematched in each successive iteration to generate the largest potential match set. A larger match possibility set generates a greater number of potential loan matches which is a concern when loans are assigned to different entities in the corporate hierarchy. Importantly, however, the more advanced techniques used later act as filters for any incorrect company matches generated in this step. Second, all candidate pairs are classified as either a match or not a match. We classify candidate pairs using a tailored approach that combines subjective, dataset–specific methods and more generalized probabilistic methods.

We begin the tailored approach by removing candidate pairs that have origination dates more than five years apart or loan amounts that differ by more than 500 percent. Next, for each loan in one dataset, we rank each of the possible matches to loans in the other dataset based on five comparable fields: (1) loan amount (ratio difference), (2) origination date and (3) maturity date (absolute difference in days), (4) the type of loan (revolver or not, dummy), and (5) lender name (dummy). We repeat this process using the second dataset as the reference dataset and combine difference rankings and hard thresholds to categorize potential loan matches into ordered "cases". For example, a case may include all matches that ranked highest from each perspective based on all five reference variables. While the selection of reference variables is subjective, it leverages dataset–specific expertise. Using a training set generated from loans with non–missing CUSIPs, we then assign m– and u– probabilities and weights for each comparison variable and calculate a matchscore for each candidate pair.

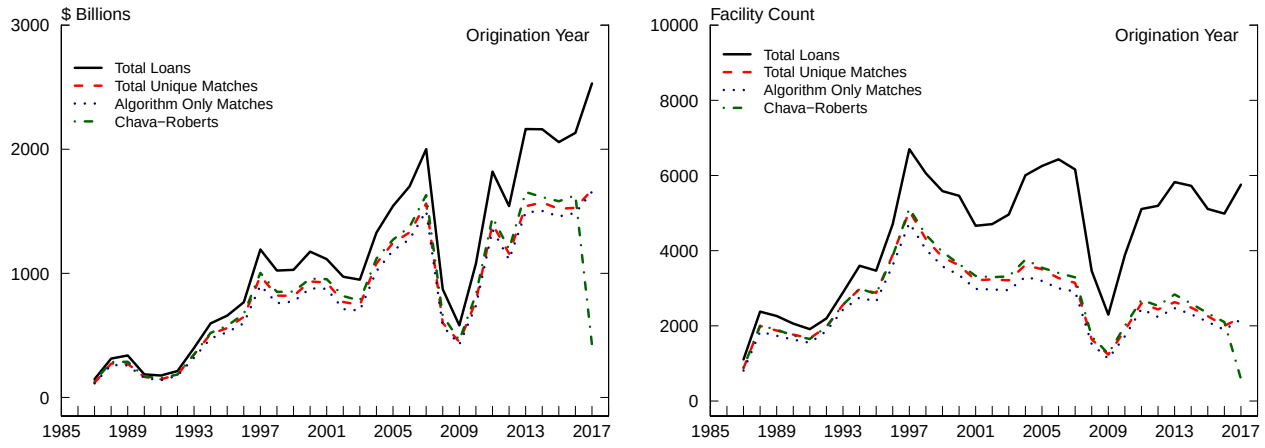
Next, the matchscore, the case information, distances between specific comparison variables, and alternative match counts for a given loan are fed into a neural net algorithm trained on CUSIP verified data based on 20,213 loan–level observations. This neural net generates a score for each candidate pair, with higher scores corresponding to higher confidence that the pair is a match. In order to classify pairs as matches or non matches, the researcher must decide on a threshold score, such that pairs with a score below the threshold will be classified as non matches and pairs with scores meeting or exceeding the threshold will be classified as matches. We use the CUSIP–verified match sample to generate a receiver operating characteristic (ROC) curve for the neural net model. The ROC curve, which plots the true positive rate against the false positive rate for each potential threshold score, is a useful tool both for evaluating a binary classification model and for choosing a threshold score that is appropriate for the needs and preferences of the researcher. To the extent that the CUSIP–verified sample is representative of the matched data as a whole, the ROC curve provides a good picture of the tradeoff between true positives and false positives.

To show that subjective dataset–specific matching methods contribute significant value, we compare results from the tailored approach to results from an assortment of comparatively naïve machine learning models. Inputs to these models exclude the judgmental assessments included in the tailored approach, relying solely on the raw comparison variables (not the measures of distance between those variables). A variety of random forest, neural net, and generalized linear models were compared, along with a diverse set of data pre–processing methods. The models were trained and evaluated on the same set of CUSIP–verified data that was used for the neural net in the tailored approach.

After deciding on a threshold classification value, the match is many-to-many and must be narrowed to a one-to-one match. Hard cutoffs based on loan amount, origination date, loan type, and lender are used to narrow down the pairs classified as matches. In cases where a loan from one dataset is matched to multiple loans from the other dataset, the set of matches is narrowed to a one-to-one match set in a way that optimizes first the total number of matches and second the total neural net score of the resulting match set. This is done by treating the problem as a maximum weighted bipartite graph matching problem [Christen (2012)] and employing the push relabel maximum flow algorithm.[4] Compared to a simple greedy algorithm, which sequentially goes through the highest neural net scores one at a time, the optimal approach results in a match set with 0.4 percent higher total neural net score and approximately 0.48 percent more matches, but is more computationally expensive. In the context of our loan level matches, the different methods of narrowing down to a set of one-to-one uniquely matched loans made little difference to the final outcome. However, for other types of matches, the methods could potentially make a far greater difference. The match reduction process is described in greater detail in Appendix B.

[4]For more information about solving maximum weighted bipartite graph problems efficiently, see Appendix B.

Figure 3: DealScan–Compustat Match Share



(a) Commitments

(b) Facilities

Source: Refinitiv LPC DealScan, S&P Global Market Intelligence Compustat, Center for Research in Security Prices, CRSP/Compustat Merged Database, Wharton Research Data Services, and authors' calculations.

4 DealScan–Compustat Match

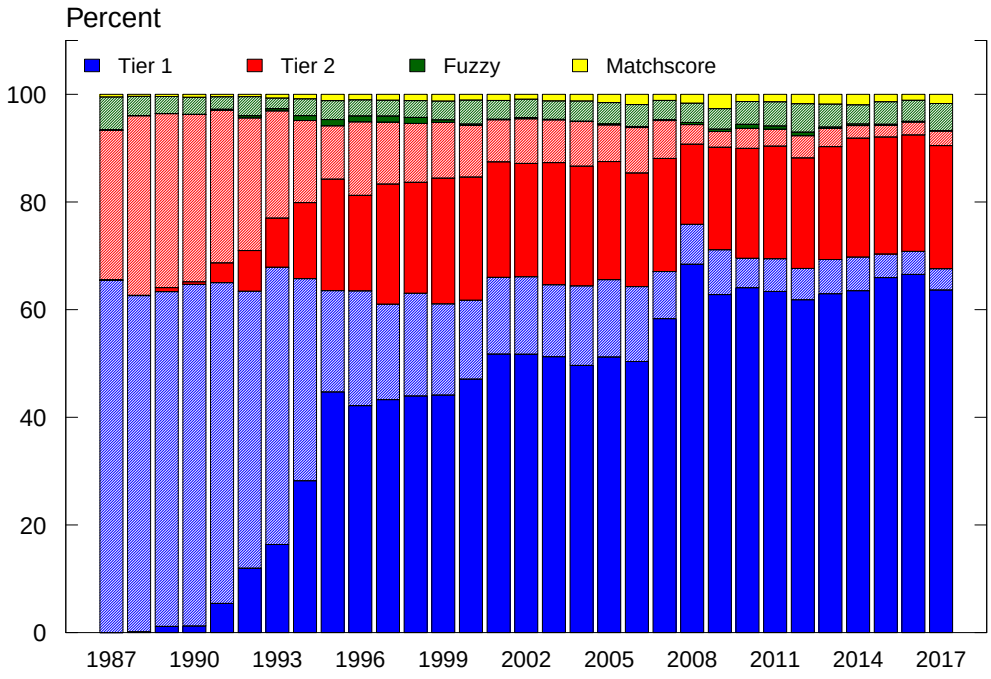
The DealScan sample is constructed with only minimal cuts to the raw data. We keep loans to all U.S. borrowers based on the country identifer and drop any loans with origination dates prior to 1980, which are likely data errors. Our final raw DealScan dataset includes 138,369 facilities originated between April 1982 and December 2017. The Compustat sample is constructed using the CRSP–Compustat Merge (CCM) as described in Appendix A. We match the company names across DealScan and Compustat using the methods described in Section 3.1.

Figure 3a shows results by commitment size for all facilities with a borrower match to Compustat. The black line represents the total commitments of all facilities in our raw DealScan dataset. Of those facilities, the algorithm finds company matches for a large fraction of total commitment amounts as shown by the orange line. On average, the algorithm finds a slightly smaller amount of Compustat matches by commitment size as found in the Chava and Roberts (2008) dataset (shown by the dashed blue line). The red line shows the total unique facilities by commitment size matched using the algorithm including incorporating the corporate family tree hierarchical information (to be discussed in Section 5) and a small hand–verified set of matches that we collected over time. However, there are unique matches across these datasets. Particularly in the late 1990s to mid–2000s, a total match set (not shown) contains slightly more matches than any of the three methods alone.

Figure 3b shows matches by facility count. On average, the algorithm finds company matches

for a slightly smaller number of facilities per year as found in Chava and Roberts (2008)'s dataset. The facility count metric, however, shows that a large share of facilities are unmatched starting in the late–1990s. This is due to the fact that syndicated loan market activity increased significantly starting around this time and became a key financing source for smaller, non–publicly–listed companies. Given that Compustat mainly collects publicly–listed company filings, we should not expect to find matches for many of these smaller companies. Figure 3a shows however, that the majority of commitment dollars are matched, suggesting that most of the unmatched facilities are small. Repeating the analysis for companies with "public" indicator equals one improves the match rate significantly as expected.

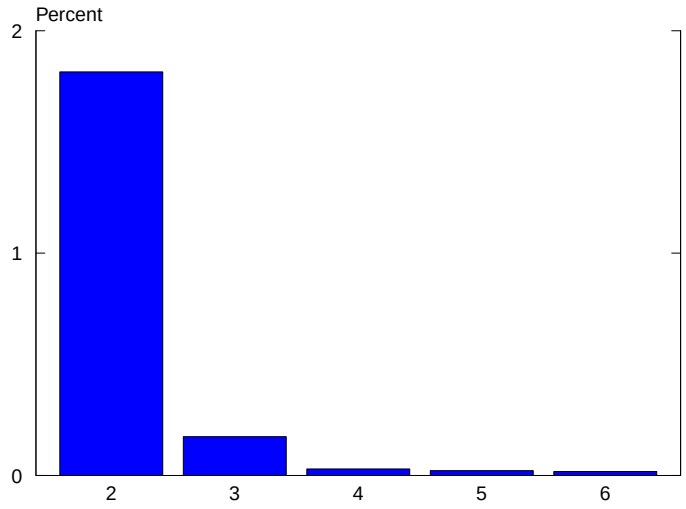
Figure 4: Matches by Match Type



Source: Refinitiv LPC DealScan, S&P Global Market Intelligence Compustat, CRSP/Compustat Merged Database, Wharton Research Data Services, and authors' calculations.

Figure 4 shows the share of matched facilities by matching method. Solid colors represent matches that also passed the date filtering screen while shaded bars denote unfiltered matches. The figure shows that the majority of matches are exact matches made with minimal string cleaning. In later years, nearly all of these matches also pass date filtering checks suggesting that a majority

Figure 5: Duplicate Matched Facilities



Source: Refinitiv LPC DealScan, S&P Global Market Intelligence Compustat, CRSP/Compustat Merged Database, Wharton Research Data Services, and authors' calculations.

of DealScan to Compustat matches are of the highest quality. In the earliest years, many matches are filtered out, but this likely reflects limitations in the historical data file rather than poor match quality. Nearly all of the remaining matches are generated in the second stage when generic words are dropped from the company names. Again, the filter works best on later years of the sample. Finally, around 5 percent of the total matches come from fuzzy name matching, nearly all of which are unfiltered, and a very small portion from match scoring.

One caveat to the algorithimic method is that some borrower names may be matched to duplicate Compustat company names as strings are subsequently dropped. Figure 5, however, shows that less than 3 percent of all facilities are matched to duplicate Compustat ids. Of those that are, most have only two possible matches making it fairly easy to hand sort through these duplicates.

Finally, table 1 shows how our matches compare to those of Chava and Roberts (2008). Between 1980 and 2016, the latest full year of facilities included in Chava and Roberts (2008), we find 77,569 total facility matches. Of these, about 2 percent are duplicates, similar to what we found for the whole sample, resulting in 75,722 unique facility matches. The majority of these matches are found in the Chava and Roberts (2008) dataset though the share varies by the match method. Nearly all of the matches collected by the exact matching function are found in Chava and Roberts (2008) while one half of those found using fuzzy matching or matchscoring are found. Nonetheless, among the facility matches found in both datasets, the agreement rate is very high. For the filtered matches and the exact name matches, the agreement rate is more than 95 percent. The unfiltered fuzzy

Table 1: Match Agreement to Chava and Roberts (2008): 1987 - 2016

	Algorithm			Chava and Roberts (2008)	
Filtered	Total	Unique	Duplicate (%)	Count	Agree (%)
Tier 1	33,991	33,741	0.73	33,595	96.7
Tier 2	14,389	13,788	4.2	12,587	82.5
Fuzzy	347	347	0.0	241	73.0
Unfiltered					
Tier 1	16,392	15,860	3.2	15,666	96.2
Tier 2	8,723	8,359	4.2	7,591	87.6
Fuzzy	2,729	2,729	0.0	1,471	61.5
Matchscore	998	898	10.0	760	88.9

Source: Refinitiv LPC DealScan, S&P Global Market Intelligence Compustat, CRSP/Compustat Merged Database, Wharton Research Data Services, and authors' calculations.

matches and matchscoring don't fare as well, with agreement rates around 70 percent, but these are a small share of the total matches. Moreover, the overall agreement rate is in line with our expectations about the match quality, and further analysis shows that these matches appear to correspond to different companies in the same corporate hierarchy.

5 Incorporating Hierarchical Information

In order to quantify how family tree information can contribute to further agreement with Chava and Roberts (2008) and even provide additional matches, we bring in other proprietary datasets that have comprehensive data on corporate and business structures and focus on disagreements between our match and Chava and Roberts (2008), as well as the names that were not matched by our algorithm.

The Walls & Associates National Establishment Time Series ("NETS") is a panel dataset compiled from annual establishment data collected by Dun and Bradstreet. To capture the hierarchy information, we roll up company names from the NETS corporate hierarchy to their parent and ultimate parent. This is done on the entire time series of corporate hierarchies provided by NETS from 1990 to 2014. After removing duplicates, we link the NETS identifier to Compustat GVKEYs using a mapping file provided by S&P Global Market Intelligence Business Entity Cross Reference ("BECRS") files as of 2017. Despite the small data gap, the resulting dataset contains a plethora of names associated with each GVKEY through almost the entire length of our sample.

After filtering to the unmatched sample of Compustat GVKEYs and DealScan FacilityIDs, we

plug these names into our standard name matching algorithm.[5] To create a reasonable sample for manual verification, we filter matches that agree on state, leaving us with 8,604 unique GVKEY–FacilityID pairs. Finally, we examine the relationship between the matched FacilityID and the company whose GVKEY was picked up through a name from the NETS hierarchy trees. We hand-sorted the matches into 7 buckets as shown in Table 2.

Table 2: Extra Matches Using NETS

	Frequency	**Percent Share**
False Match	783	9.10
Hierarchical Relationship	5,947	69.12
Unclear Relationship	34	0.40
Name Change	719	8.36
Mergers and Acquisitions	250	2.91
Name Difference	59	0.69
Other	812	9.44
Total	8,604	100

Note. *False Match* indicates a relationship was not identified. *Hierarchical Relationship* indicates a hierarchical relationship. *Unclear Relationship* indicates that firms or names appear to be related, but unclear if they are part of different hierarchy or names were written differently. *Name Change* indicates that a name change was a reason the original algorithm did not pick up the match. *Mergers and Acquisitions* indicate that the original algorithm appears to have failed because of a merger or acquisition that was not accounted for. *Name Difference* indicates that the same firm is referred to with different names, such as nicknames, across different data sources. *Other* includes matches that were picked up with NETS identifier link to Compustat GVKEYs using a mapping file provided by BECRS files, but could not be located in our Compustat database of U.S. firms, perhaps due to foreign affiliated companies. Source: Refinitiv LPC DealScan, S&P Global Market Intelligence Compustat, CRSP/Compustat Merged Database, Wharton Research Data Services, Walls & Associates, National Establishment Time-Series (NETS) Database 2014, S&P Global Market Intelligence, Business Entity Cross Reference Service (BECRS), and authors' calculations.

Of these 8,604 unique GVKEY FacilityID pairs, 3,211 agree with Roberts, 77.6 percent of which have been categorized as based on hierarchical relationships. This provides evidence that corporate family tree information is useful in reconciling the output of our main algorithm with Chava and Roberts (2008). Another 11.6 percent were picked up due to additional historical names, while the remaining observations fall into other categories.

Our baseline match proposed 83,774 unique GVKEY–FacilityID pairs from 1982 to data available as of 2018. The Chava and Roberts (2008) verified subsample of the NETS match extends the match by 3.56 percent (2,983 observations not marked as "False Match" or "Other"), while the general verified subsample extends the match by 8.37 percent (7,009 observations not marked as "False Match" or "Other").

We conclude that hierarchy information plays an important role in identifying matches. While this exercise was strict and limited in scope, it still yielded a significant contribution to the overall

[5] We include in this set matches generated through fuzzy matching and matchscoring where the match success rate is found to be lower.

match. This exercise improves the historical portion of our match, and motivates further work in using other, more up–to–date hierarchy trees to enhance our matching algorithms. Figures 3a and 3b indicate that after incorporating family tree information and names from other data sources, "Total Unique Matches" becomes almost indistinguishable to the Chava–Roberts match.

6 SNC–DealScan Match

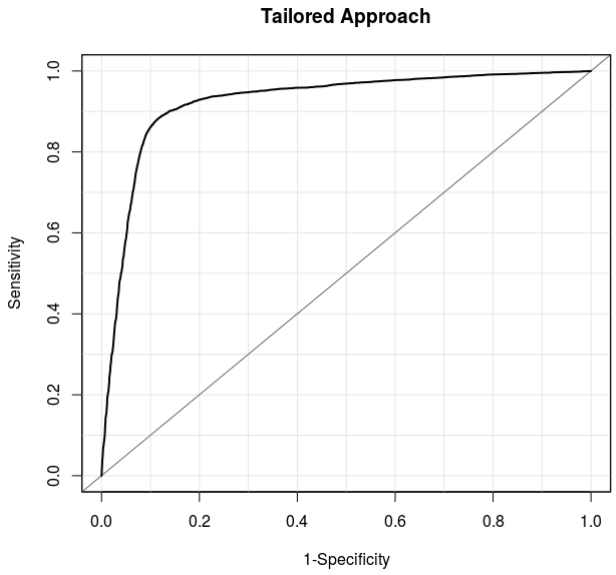
The SNC sample is made up of all exam and agent reported SNC credits that have a credit ID and that have a non-missing report date and obligor ID. Our final sample includes 103,944 credits reported between 1992 and 2017. We match this sample to DealScan facilities using the methods described in Section 3.2.

All attempted neural net and generalized linear models significantly underperformed the tailored approach, but random forests tended to fare better overall. Figures 6 and 7 show ROC curves for the tailored approach after the neural net stage and the best–performing random forest model, respectively. These curves were generated from CUSIP verified loan data. The tailored approach has an area under the curve (AUC) of 0.924 and the nonjudgmental machine learning method has an AUC of 0.944. However, since false positives are relatively worse than false negatives for loan matching, we should look at the region on the left where the false positive rates are relatively small. In the very beginning, the machine learning approach picks up relatively more true positives given very small false positive rates. However, this criteria will result in too few matches to do meaningful analysis. Allowing for slightly larger false positive rates, the tailored approach outperforms the machine learning approach by a relatively large margin: a 12 percent false positive rate (specificity 0.88) corresponds to approximately 0.89 true positive rate (sensitivity) in the tailored approach and only about 0.86 sensitivity in the machine learning approach. This shows that, if a researcher places a high emphasis on minimizing false positives, while also assuring enough loan–level matches are retrieved, the tailored approach is superior.

Using the methods described in Section 3.2, based on our tailored approach, we match over 75 percent of committed loan dollars reported in recent years in SNC to a DealScan origination. Figure 8 shows the share of unique SNC loans, utilized amounts, and committed amounts matched to DealScan by report year. All three numbers generally grow over time. That the share of unique loans matched is always lower than the share of utilized and committed loan amounts shows that larger loans are significantly more likely to be matched than smaller loans.

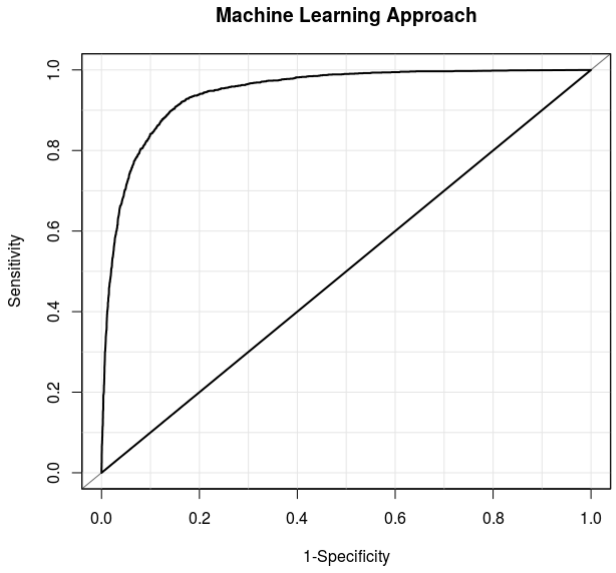
Figure 9 shows the same numbers by origination year. We are able to match only a small fraction of loans originated in the 1980s and early 1990s, but by the mid 2010s we successfully find a match for 75 percent of committed loan amounts. The resulting dataset has 50,196 unique matched loans.

Figure 6: ROC Curve – Tailored Approach



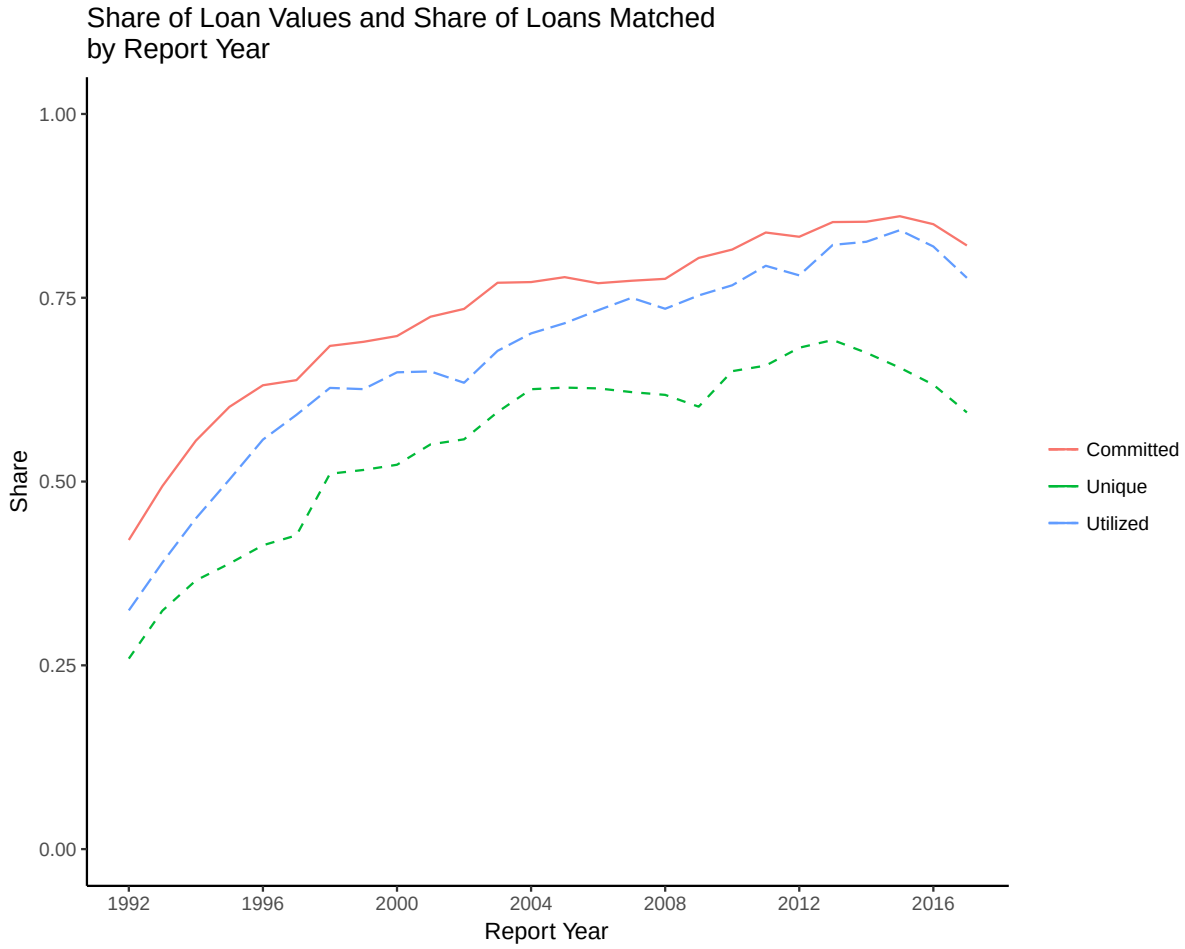
Source: Shared National Credit Database, Refinitiv LPC DealScan, Center for Research in Security Prices, CRSP/Compustat Merged Database, Wharton Research Data Services, and authors' calculations.

Figure 7: ROC Curve – Machine Learning Approach



Source: Shared National Credit Database, Refinitiv LPC DealScan,Center for Research in Security Prices, CRSP/Compustat Merged Database, Wharton Research Data Services, and authors' calculations.

Figure 8: DealScan Loans Matched to SNC Loans – by Report Date

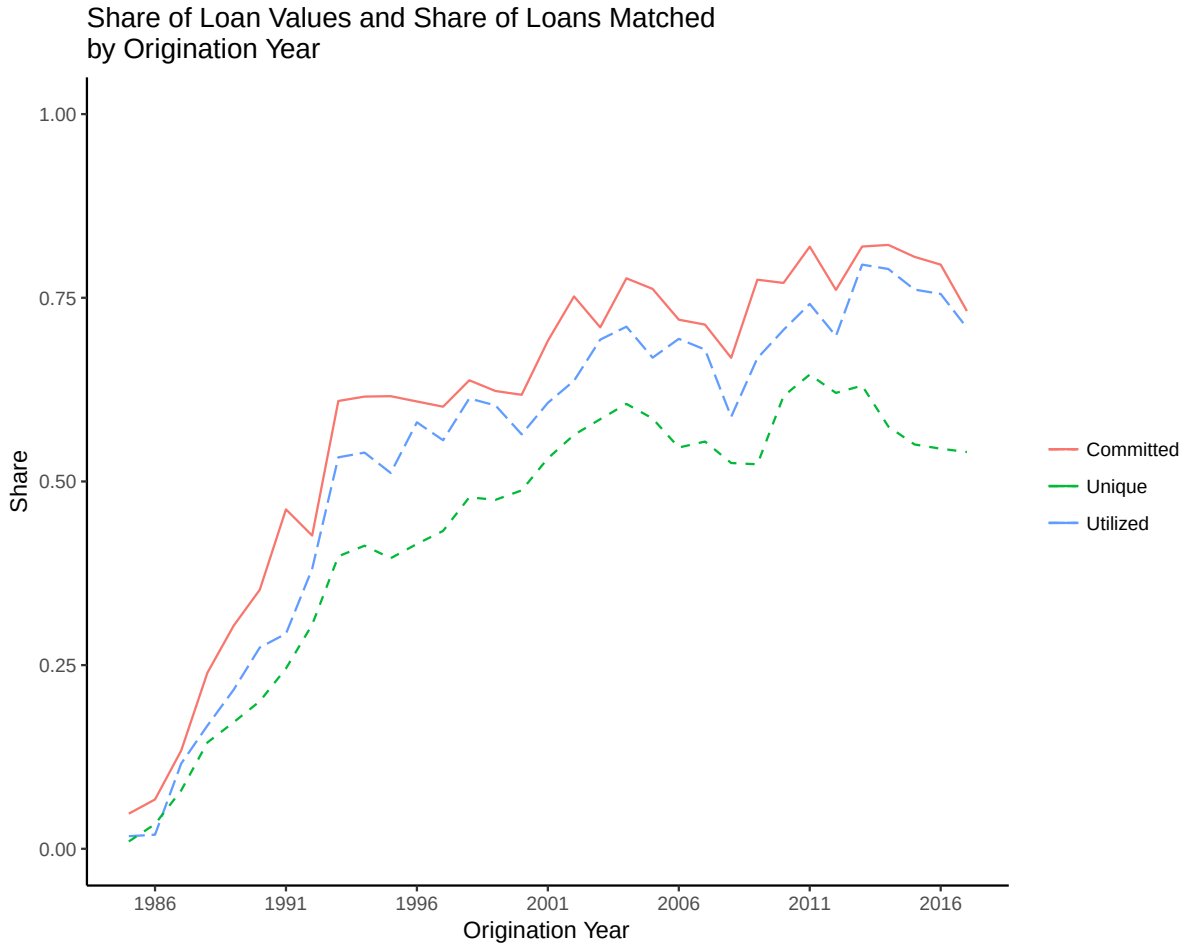


Source: Shared National Credit Database, Refinitiv LPC DealScan, Center for Research in Security Prices, CRSP/Compustat Merged Database, Wharton Research Data Services, and authors' calculations.

7 Conclusion

We have developed a simple and effective method for matching corporate loan databases that lack unique observational identifiers. Using our algorithm for borrower level matching, we can match at least 90 percent of the facilities from a hand matched dataset and more than 95 percent of the most–widely used linkages for company level data. In these cases, our matches agree with those in the baseline datasets about 95 percent of the time. Much of the disagreement can be resolved by incorporating hierarchical information associated with the borrowers in each dataset. In such cases, we have provided some thoughts for further filtering. Using our algorithm for loan level matching, we can match about 72 percent of committed loan amounts in the SNC universe of loans, which

Figure 9: DealScan Loans Matched to SNC Loans – by Origination Date



Source: Shared National Credit Database, Refinitiv LPC DealScan, Center for Research in Security Prices, CRSP/Compustat Merged Database, Wharton Research Data Services, and authors' calculations.

provide a useful sample for research and policy purposes.

Going forward, it is our hope that others that perform research on syndicated loans will modify and improve our methodologies in an effort to automate the matching process. Moreover, we hope that these improvements will be publicly disclosed following the tradition established by Chava and Roberts (2008). Having replicable and robust methods of generating key datasets is an important part of the scientific process. In addition, wider data access will encourage new research directions in this field.

Appendix

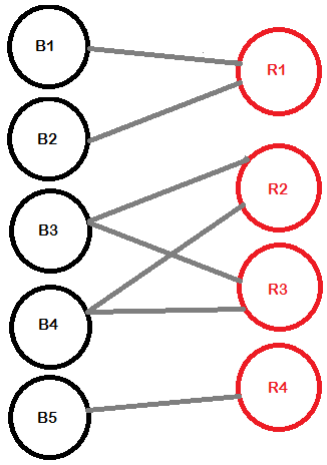
A CRSP–Compustat Data Preparation

We use the Center for Research in Security Prices, CRSP/Compustat Merged Database ("CCM") from Wharton Research Data Services to construct unique historical information on company name and location changes. The database includes historical header information taken directly from Compustat's ftp files for each quarterly release. The data are stored in separate files: one for pre April 2007 data and one for all subsequent data following the inclusion of securities level data in Compustat. Post April 2007 data include permno changes to incorporate securities level information so are not directly comparable to pre April 2007 data.

First, we uniformly reformat the common fields on each table. End dates on the the pre April 2007 table (cst_hist) are set to April 13, 2007, the last date before the permno change. By default, the CCM database sets the end dates for current entities to the most recent as of date. We move this date to a generic December 31, 9999, a future date that indicates the data are currently valid. Next, the two tables are stacked which provides a uniform set of historical, company level information.

In the final step, we examine all the historical information for a given company to denote changes. This is done is several short steps. First, we sort the data by company and historical date. Next, each rows company information is compared to its lagged value from the previous row within the company. Rows are collapsed depending on when a field's value changes. For example, a company that had identical field values between March 1980 and December 2000 would be collapsed to a single row who's company characteristics were the field values, a start date of March 1980 and an end date of December 2000. A new row would be created starting for January 2001 when the first field changed with an end date corresponding to the month just before the next characteristic change. Thus our final dataset is a company characteristic database that is unique by company and start and end date intervals. No changes were recorded in the CCM database between the start and end date intervals for the fields considered. This dating system is analogous to the bank structure data provided by the National Information Center. Moreover, it makes writing the filter functions used in the matching algorithm almost trivial.

Figure A: Example Bipartite Graph



B Loan Match many-to-many to one-to-one Reduction

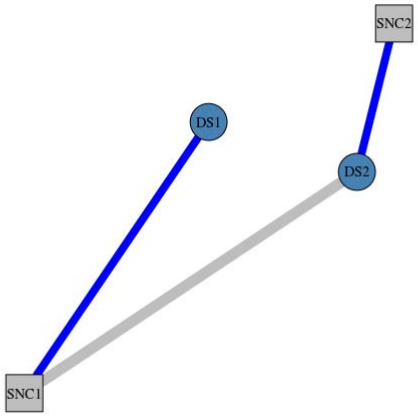
Narrowing an many-to-many match to a one-to-one match in a way that optimizes the total neural net score of the resulting match set is equivalent to solving the maximum weighted bipartite graph matching problem (Christen, 2012). As such, it will be helpful to think of the many-to-many match set that results from the matching process described in Section 3.2 as a bipartite graph. Each SNC and DealScan (DS) loan is a node, and matches between loans are edges connecting the corresponding nodes. Our graph is bipartite because SNC loans can only be connected to DS loans, and vice versa. The weight of an edge connecting two nodes is equal to the neural net score of the match between those two loans. Figure A provides a fabricated example of a bipartite graph.

A bipartite graph can be decomposed into a number of unconnected subgraphs. A decomposition of the graph in Figure A results in three subgraphs, where the first includes nodes B1, B2, and R1, the second includes B3, B4, R2, and R3, and the third includes B5 and R4. It is useful to decompose our match set this way because the optimal one-to-one solutions for each subgraph are independent of each other, and solving them separately can greatly reduce the computational complexity of the problem.

One way to narrow the match from many-to-many to one-to-one is to use a greedy algorithm. Generally, a greedy algorithm makes locally–optimal decisions without looking ahead to see if those decisions will preclude a globally–optimal solution. In the present case, that means looping through SNC or DS loans, taking the highest–scored pair first and then removing the pairs that share a loan with it before moving onto the next loan. This process is repeated until all loans have been matched or until there are no remaining pairs. A greedy approach is computationally easier but will perform worse– it may end up choosing weaker matches instead of stronger ones, or fewer matches instead of more.

One could find the optimal one-to-one match within a bipartite graph by calculating the total neural net score of each possible configuration that satisfies the one-to-one constraint and choosing the configuration with the largest score. This approach is computationally prohibitive for large match sets. Fortunately, there are a number of algorithms available that efficiently approximate an optimal solution. The problem can be solved through auction algorithms, the Hungarian algorithm,

Figure B: SNC-DS Bipartite Graph



the Ford–Fulkerson algorithm, or a variety of others. We use the push relabel algorithm, which has already been implemented in the igraph package in R.

Due to computational constraints, it is impossible to operate the push relabel algorithm on the entire many-to-many match set. There are two steps that can be taken before employing the push relabel algorithm in order to reduce the computational difficulty. First, the researcher can simply remove all one-to-one subgraphs from the match set, and put them into the final one-to-one match set. In Figure A, this would be equivalent to removing B5 and R4. Second, the researcher can remove all 1:m subgraphs from the match set and use a greedy approach to pair them down to one-to-one matches, which would then be put into the final match set. In Figure A, one would compare the match strength of the two pairs, B1–R1 and B2–R1. The stronger match would be put into the final match set and the other would be discarded. Both steps preserve optimality and lower the difficulty of reducing the many-to-many match set into an optimal one-to-one set in R.

The majority of subgraphs in the original SNC–DS match set are one-to-one or 1:m matches. After one-to-one and 1:m matches are removed from the many-to-many match set, then, the remaining set of matches is small enough to run through the max_bipartite_match() function in the igraph R package. If a researcher is left with an many-to-many set that is still too large, she can decompose the graph as described above, solve each subgraph separately, and then combine the results into a final match set.

Figure B is a representative example of a many-to-many subgraph from the SNC–DS match, and will also provide an illustration of how a greedy algorithm for reducing the match set to one-to-one might be suboptimal. It shows two SNC loans, SNC1 and SNC2, two DS loans, DS1 and DS2, and three matches among them. The width of the edges corresponds to the neural net score associated with the match; thicker lines represent stronger matches. The blue line shows the matches that are kept by the push relabel algorithm, and the grey line shows the match that would be kept under a greedy algorithm.

References

Aramonte, S., S. J. Lee, and V. Stebunovs (2015): "Risk Taking and Low Longer-term Interest Rates: Evidence from the U.S. Syndicated Loan Market," Finance and Economics Discussion Series No. 2015-068, Board of Governors of the Federal Reserve System.

Beatty, A., S. Liao, and H. Zhang (2015): "The Effect of Banks' Financial Reporting on Syndicated Loan Structure," Working Paper.

Chaudhry, S. M., and S. Kleimeier (2015): "Lead Arranger Reputation and the Structure of Loan Syndicates," *Journal of International Financial Markets, Institutions and Money*, 38, 116–126.

Chava, S., and M. R. Roberts (2008): "How Does Financing Impact Investment? The Role of Debt Covenants," *The Journal of Finance*, 63(5), 2085–2121.

Chodorow-Reich, G. (2014): "The Employment Effects of Credit Market Disruptions: Firm-Level Evidence from the 2008-9 Financial Crisis.," *Quarterly Journal of Economics*, 129(1), 1–59.

Christen, P. (2012): *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer-Verlag Berlin Heidelberg.

Cohen, G. J., S. J. Lee, and V. Stebunovs (2016): "Limits to Monetary Policy Transmission at the Zero Lower Bound and Beyond: The Role of Nonbanks," Working paper.

Correa, R., H. Sapriza, and A. Zlate (2012): "Liquidity Shocks, Dollar Funding Costs, and the Bank Lending Channel During the European Sovereign Crisis," International Finance Discussion Papers, No. 2012-1059, Board of Governors of the Federal Reserve System.

Fellegi, I. P., and A. B. Sunter (1969): "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64(328), 1183–1210.

Gill, L. E. (1999): "OX-LINK: The Oxford Medical Record Linkage System," in *Record Linkage Techniques - 1997: Proceedings of an International Workshop and Exposition*, Washington, DC. Committee on Applied Theoretical Statistics, National Research Council, National Academy Press, Federal Research Committee on Statistical Methodology, Office of Management and Budget.

Ippolito, F., A. Ozdagli, and A. Perez (2013): "Is Bank Debt Special for the Transmission of Monetary Policy? Evidence from the Stock Market," Center for Economic Policy Research, Discussion Paper No. 9696.

Ivashina, V. (2009): "Asymmetric Information Effects on Loan Spreads," *Journal of Financial Economics*, 92(2), 300–319.

Ivashina, V., and A. Kovner (2011): "The Private Equity Advantage: Leveraged Buyout Firms and Relationship Banking," *Review of Financial Studies*, 24(7), 2462–2498.

Ivashina, V., and D. S. Scharfstein (2010): "Loan Syndication and Credit Cycles," *American Economic Review*, 100(2), 57–61.

Jaro, M. A. (1989): "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 84(406), 414–420.

Lee, S. J., L. Q. Liu, and V. Stebunovs (2017): "Risk Taking and Interest Rates: Evidence from Decades in the Global Syndicated Loan Market," International Finance Discussion Papers, No. 1188, Board of Governors of the Federal Reserve System.

Lee, S. W., and D. J. Mullineaux (2004): "Monitoring, Financial Distress, and the Structure of Commercial Lending Syndicates," *Financial Management*, 33(3), 107–130.

Méray, N., J. B. Reitsma, A. C. Ravelli, and G. J. Bonsel (2007): "Probabilistic Record Linkage is a Valid and Transparent Tool to Combine Databases Without a Patient Identification Number," *Journal of Clinical Epidemiology*, 60(9), 883.e1–883.e11.

Murfin, J., and R. Pratt (forthcoming): "Comparables pricing," *The Review of Financial Studies*.

Nini, G., D. C. Smith, and A. Sufi (2012): "Creditor Control Rights, Corporate Governance, and Firm Value," *Review of Financial Studies*, 25(6), 1713–1761.

Roberts, M. R., and A. Sufi (2009): "Control Rights and Capital Structure: An Empirical Investigation," *The Journal of Finance*, 64(4), 1657–1695.

Sayers, A., Y. Ben-Shlomo, A. W. Blom, and F. Steele (2016): "Probabilistic Record Linkage," *International Journal of Epidemiology*, 45, 954–964.

Sufi, A. (2007): "Information Asymmetry and Financing Arrangements: Evidence from Syndicated Loans," *The Journal of Finance*, 62(2), 629–668.

Winkler, W. E. (1989): "Advances in Record Linkage Methodology as Applied to the 1985 census of Tampa Florida," *Journal of the American Statistical Association*, 84(406), 414–420.

——— (2006): "Overview of Record Linkage and Current Research Directions," Census Bureau.