Contractual Completeness in the CMBS Market: Insights from Machine Learning

Brent W. Ambrose* Yiqiang Han† Sanket Korgaonkar‡
Lily Shen§

January 3, 2020

Preliminary Draft

Abstract

A complete contract attempts to specify the rights and duties of the parties to the contract for every possible future contingency. The degree of contractual completeness is particularly salient in the commercial mortgage-backed securities (CMBS) market, where the Pooling and Servicing Agreement (PSA) governs the actions of the agents who are a party to the deal, and where investors have limited interaction with mortgage originators and servicers following deal inception. We first use a machine learning (ML) methodology to access the completeness of PSAs for conduit CMBS deals created between 2000 and 2019. Next, we analyze how the completeness of PSAs vary with features of the underlying CMBS pools. Our results indicate that CMBS PSA documents do reflect observable differences in the underlying collateral pools. In addition, we also show that PSA documents from the same issuer are more likely to be similar than PSAs from CMBS deals originated by different underwriters. Somewhat surprisingly, we do not see that PSA documents differ when deals have geographic dispersion in the underlying collater pools. Finally, consistent with the theory of complete contract, we find that loans in deals with more complete PSA documents have lower default rates.

*Penn State University, (814) 867-0066, bwa10@psu.edu
†Clemson University, yiqianh@clemson.edu
‡University of Virginia, sanketk@virginia.edu
§Clemson University, yannans@clemson.edu

Contractual Completeness in the CMBS Market: Insights from Machine Learning

1 Introduction

Contracting between agents is a pervasive feature of any economic setting. Seminal contributions to the theory of contracting have distinguished between complete and incomplete contracts (see Oliver Hart's Nobel Lecture (Hart (2017)) for a summary). A contract is considered to be complete if the rights and duties of the agents are specified for every possible future state of the world.

Contractual completeness is particularly salient for real estate assets such as mortgage-backed securities (MBS), where the Pooling and Servicing Agreement (PSA) governs the actions of the agents involved in the securitization process. Unlike the typical corporate structure where investors employ managers to manage the firm's assets on their behalf, assets (the mortgages) are placed into a trust for the benefit of the investors in a typical mortgage securitization structure. The investors then share in the cash flows from the assets with each investor's cash flow share determined by the PSA. Investors in MBS have limited interaction with originators and servicers following the inception of the deal, hence increasing the importance of spelling out future contingencies in the PSA.

However, the typical PSA is a lengthy legal document filed with the Securities and Exchange Commission (SEC) and, thus, is often viewed as mostly

containing boilerplate legal text. As a result, PSA contracts may appear similar across deals despite heterogeneity in the underlying collateral and deal structure. Given the highly technical and legalistic writing of the PSA, estimating the degree of contractual completeness is not trivial. Subsequently, the completeness of contracting can influence deal performance.

Previous studies of financial contracting, which examine either the use of various contractual features (Nini et al. (2009), Nini et al. (2012)) or examine their welfare implications (Matvos (2013)), have tended to focus on a few key words or paragraphs of a contract; thus excluding from the analysis information that is admittedly difficult to quantify. We bring to bear on this problem a new machine learning method that is capable of processing large quantities of textual data. More specifically, our methodology allows us to compute the similarity/dissimilarity of complex financial contracts (e.g., ones that are on average, 300 pages long) that belong to defined comparison groups. We marry this methodology with the following insight: In a world with a set of unique assets that are contracted over, if every contract is complete, then every contract must be unique. We then characterize how the completeness of contracts varies with the nature of the assets. Consequently, we are able to offer insights into the welfare implications of contractual incompleteness.

Our study focuses on conduit commercial mortgage-backed securities (CMBS), where the Pooling and Servicing Agreement (PSA) is a tri-party agreement written at deal inception that outlines the rights, duties, and responsibilities of various parties to the securitization. The PSA governs the selection of mortgages into the loan pools, the subsequent monitoring of the loans, and

the actions to be taken if a mortgage becomes seriously delinquent. Conduit CMBS deals typically comprise large pools of commercial real estate loans that were originated explicitly for inclusion in mortgage-backed securities. We hand collect the relevant PSAs from the Securities and Exchange Commission EDGAR database and Trepp.

We employ the machine learning (ML) algorithm introduced in Shen (2018) to convert a PSA's text into numerical vectors which capture not only the contents but also the semantic meaning of each document.[1] The algorithm delivers pairwise scores that measure the deviation between pairs of CMBS deal PSAs with lower scores indicating greater similarity between the PSA pairs. We also aggregate these pairwise scores to create contractual completeness measures across various comparison groups controlling for issuer and origination year cohorts. Finally, we bring into the numerical data on the individual mortgage contracts and collateral properties that comprise the security pools. Thus, we are able to measure the heterogeneity across the asset pools that provide the cash flows to the security investors.

The results reveal significant heterogeneity in the pairwise uniqueness across CMBS PSAs. We find greater similarity among deals from the same underwriter than when compared to other underwriters. We also find greater similarity across deals issued in the same year. Univariate regressions of pairwise uniqueness scores on pairwise differences in deal characteristics reveal that PSA content is correlated with average loan interest rates and average

[1]This represents an innovation from the standard "bag-of-words" approach used in the seminal applications of textual analysis in the finance literature (Tetlock, 2007, Loughran and McDonald, 2011).

loan-to-value (LTV) ratios, two key measures of mortgage risk. Our multivariate regressions reveal that deals originated by the same underwriter are significantly more similar, even after controlling for differences in observable characteristics. This provides evidence for contractual incompleteness in the CMBS market since observably different securities originated by the same firm share similar legal documents.

We also study the potential consequences of incomplete contracts by examining the underlying collateral performance. In a complete contracting environment, every contingency will be accounted for in the security design, and thus, loan originators will have limited ability to shirk or deliver loans that satisfy the minimum contractable elements. To test this hypothesis, we examine whether PSA uniqueness is correlated with the underlying loan-level risk (proxied by the *ex post* default rate). Regression estimates indicate that CMBS deals with more complete PSA documents (compared with deals issued in the same year by different underwriters) contain loans that have lower probabilities of default.

Our paper adds to four streams of the literature. First, we complement the emerging literature using new techniques in data science to explore economic and financial issues. Second, we provide a novel view of the value of optimal contracts in the market for securitized assets (DeMarzo, 2005; DeMarzo and Duffie, 1999; Longstaff and Rajan, 2008; Benmelech and Dlugosz, 2009; Gennaioli et al., 2013; Riddiough, 1997; Glaeser and Kallal, 1997; Hartman-Glaser et al., 2012; Lacker, 2001; Maskara, 2010; Hanson and Sunderam, 2013; Begley and Purnanandam, 2017; An et al., 2009; Malamud et al., 2013; Ambrose

et al., 2016; An et al., 2008, 2015; Beltran et al., 2017; Mooradian and Pichler, 2018). Third, our analysis has implications for the role of lenders misreporting asset risk in the securitization market (Piskorski et al., 2015; Downing et al., 2009; An et al., 2011; Demiroglu and James, 2012). Finally, the analysis also has implications for the role of rating agencies in asset securitization design (Pagano and Volpin, 2010).

Our paper proceeds as follows. Section 2 describes the data. Section 3 discusses the machine learning algorthim used to measure PSA pairwise similiarities, while Section 4 describes the emirical analysis we use to document variation across the CMBS PSAs. Section 5 presents analysis of the effect of contracctual completeness on deal performance and Section 6 concludes.

2 Data

Our primary data set consists of loan and deal level information on commercial mortgage-backed securities (CMBS) collected by Trepp. Trepp is cited as one of the real estate industry's largest providers of information on securitized commercial mortgages.[2] The data set contains extensive information about CMBS deals, bonds, as well as the commercial real estate mortgages that comprise the CMBS loan pools (including detailed data on loan terms and property characteristics). To create our sample, we first downloaded all PSAs available in the TREPP database for deals originated between 1998 and 2019. In addition, where possible we matched deal names from Trepp to the Secu-

[2]Trepp tracks over 1,500 CMBS deals comprising over 200,000 mortgages. More information about Trepp is available at `http://www.trepp.com/about-us`.

rities and Exchange Commission (SEC) EDGAR database.[3] To ensure fair comparison, only the main text body is used for language modeling, whereas the nomenclature chapter and Exhibits in the Appendix are removed during the pre-processing stage. The document screening process includes automated screening using regular expression matching technique and also proofreading by trained research assistants. After cleaning the files for errors and duplicates, our sample consists of 975 CMBS deals. Given data quality issues for deals and loans originated before 2000, we restricted the sample to the period 2000 to 2019. The final sample consists of 908 deals comprising 687 conduit deals, 45 agency deals, 70 single-asset deals, 36 large loan deals, and 68 miscellaneous deals.

Although our sample consists of multiple deal types, we choose to focus our primary analysis on the 687 conduit CMBS deals as these are securities created as Real Estate Mortgage Investment Conduits (REMIC) under the US Federal income tax law for the purpose of pooling and securitizing mortgage loans. These 687 CMBS deals contain more than 200,000 commercial mortgages.

Figure 1 shows the frequency of CMBS deals by the origination year. The graph clearly shows the structural break in the market during the Great Financial Crisis of 2008 through 2010 when the non-agency mortgage securitization market disappeared. Table 1 shows the summary statistics for the underlying loan characteristics making up the CMBS deals.

[3] `https://edgar.sec.gov/`

3 Machine Learning and Semantic Analysis

Extant studies in finance and economics largely rely on numerical data. However, in the era of "Big Data," numerical data account for less than 20 percent of available information while the other 80 percent are in textual format.[4] Natural Language Processing (NLP) is an Artificial Intelligence (AI) method designed to allow computers to process human language and convert textual data to numerical data for analysis. This is also called machine learning.

We adopt the unsupervised machine learning data vectorization algorithm introduced in Shen (2018) to convert each PSA document into a high–dimension numerical vector. These vectors capture the semantic meaning of the words and sentences found in the PSAs. One of the novel features of the algorithm is that it does not require any prior assumptions or specialized knowledge about the document being analyzed.

Our algorithm follows a fundamental linguistic principle: the meaning of words and sentences are defined by their contexts because the contextual elements often share syntactic and semantic relations with each other. This is commonly known as "You shall know a word by the company it keeps" (Firth, 1957). The algorithm recognizes patterns and connections in textual data and then translates these findings into a high–dimensional numerical vector that represents its semantic meaning. Thus, PSAs that are similar are closer to each other in vector space.

The algorithm uses a neural network model with three layers (input, hid-

[4] "Structured Data in a Big Data Environment", https://www.dummies.com/programming/big-data/engineering/structured-data-in-a-big-data-environment/ .

den, and output) to create the numerical vectors.[5] We train the neural network model iteratively to get a vector representation of each description. To generalize the idea, we define w_{ij}^{out} as ith output word (target) randomly selected from document j, and $\underline{w}_{ij}^{in}$ as a vector of input words from its context. We source the context words within a distance of L from the target word w_{ij}^{out}. The distance of L can also be regarded as the size of a sliding window, which defines the extent of the context that we would like to include in the word vector analysis. For an arbitrary element in the $\underline{w}_{ij}^{in}$, we use w_{ijl}^{in} to denote the lth specific element in the vector between 1 and L.

Figure 2 provides a simplified representation of our algorithm. First, we tokenize the entire pool of PSA documents into a vocabulary list. Each of the word, also called a token, in the list will have a unique numerical word vector also called one-hot vector) to represent its position in the list (i.e., one-hot vector is a list of many zero's and only one non-zero value "1" at the position of the word).

Each word from input list $\underline{w}_{ij}^{in}$ is then projected onto the n–dimensional space by a weighting matrix comprised of weighting vectors (v_k^w for an arbitrary word k). We are specifically interested in constructing this weighting matrix for the corresponding input. Since a document can be analyzed as a combination of sentences composed of words, we include a document vector (v_k^p for the PSA that contains arbitrary word k, also denoted as "D" on Figure 2) to represent the overall document weights between the input layer and the

[5]Our neural network is a version of the paragraph vector method. Dai et al. (2015) compare the paragaph vector method against other textual analysis algorithms, including the popular Bag-of-Words method, and conclude that the PV method strictrly outperforms other methods.

hidden layer. The entire collection of the weight vectors can be regarded as a transformation matrix between the input layer and the hidden layer. Inside the matrix, the weighting vectors v_k^w are each a "1 by A" vector containing the numeric value for each feature that describes a word k, where A is the number of features (also called the dimension of the space). The subscript $k \in [1, N]$ means it is the kth element in the total number of N words considered by the procedure. The document weighting vectors v_j^p are "1 by A" vectors containing the numeric value for each feature that describes a document j. Similarly, the subscript $j \in [1, M]$ means it is the jth element in the total number of M documents considered by the procedure. The superscript w denotes it is a word vector, while the superscript p represents a vector for a document. There is a counterpart of the abovementioned weighting matrix between the hidden layer and the output layer. We use μ_k^w and μ_k^p to represent the weight vectors for words and documents in the hidden layer, respectively.

The training is unsupervised since this process does not need human intervention, and samples do not need to be labeled. This ensures the learning can be performed objectively without additional human interpretation. The training includes a prediction process since we are trying to predict an output word that matches the masked ground-truth word. At the beginning of the training, the weighting matrix is randomly initialized. The model will adjust the weighting matrix through Bayesian iterations until converge.

Now, we define the index that describes the likelihood that an output word is the kth word in the population given the set of input words for the

ith randomly selected word from the jth PSA document.

$$x_{ijk} = \mu_k^{p\prime} v_j^p + \sum_{l=1}^{L} \mu_k^{w\prime} v_{w_{ijl}^{in}}^w \tag{1}$$

Correspondingly, the correct prediction should be:

$$x_{ijw_{ij}^{out}} = \mu_{w_{ij}^{out}}^{p\prime} v_j^p + \sum_{l=1}^{L} \mu_{w_{ij}^{out}}^{w\prime} v_{w_{ijl}^{in}}^w \tag{2}$$

The conditional probability of matching a correct output content w_{ij}^{out} given input words $\underline{w}_{ij}^{in}$ is denoted as $Pr\left[w_{ij}^{out}|\underline{w}_{ij}^{in}\right]$. The probabilities $Pr\left[w_{ij}^{out}|\underline{w}_{ij}^{in}\right]$ are evaluated iteratively using a log–linear Softmax function during every iteration until convergence, as depicted in the following equation:[6]. Now, the probability of the output word w_{ij}^{out} arising is written as below: the goal is to maximize the probability.

$$Pr\left[w_{ij}^{out}|\underline{w}_{ij}^{in}\right] = \frac{e^{x_{ijw_{ij}^{out}}}}{\sum_{k=1}^{K} e^{x_{ijk}}} \tag{3}$$

where the probabilities are initialized to sum to one for any word i from document j over all possible words k in the population.

Finally, assuming that N words are drawn randomly from M PSAs, the log likelihood problem can be written as:

$$Min_{\mu_k^p, \mu_k^w, v_j^p, v_k^w} \sum_{j=1}^{M} \sum_{i=1}^{N} -log\left(Pr\left[w_{ij}^{out}|\underline{w}_{ij}^{in}\right]\right) \tag{4}$$

[6]The Softmax function is a generalization of the logistic function to calculate categorical probability used in Artificial Neural Networks

Iteratively, we maximize the probability of getting the correct outputs through the fine–tuning of θ={ μ_k^p, μ_k^w, v_j^p, v_k^w } and minimizing the error term ϵ:

$$\begin{aligned} \epsilon &= \log \sum\nolimits_k \exp\left(x_{ijk}\right) - x_{ijw_{ij}^{out}} \\ &= log \sum\nolimits_o \exp\left(\mu_k^{p\prime} v_j^p + \sum_{l=1}^{L} \mu_k^{w\prime} v_{w_{ijl}^{in}}^w\right) - \mu_{w_{ij}^{out}}^{p\prime} v_j^p + \sum_{l=1}^{L} \mu_{w_{ij}^{out}}^{w\prime} v_{w_{ijl}^{in}}^w \end{aligned}$$

where $x_{ijw_{ij}^{out}}$ is the output vector corresponding to the ground truth output words. We also define a generic output score, x_{ijk}, to be the output probability based on a given input word. Thus, our algorithm iteratively modifies the vector representation, θ, until the model converges.

Figure 3 offers a 3–D demonstration of the vector space for the CMBS PSA documents. Each dot represents a PSA document. Although the axes in vector space do not hold physical meaning in human language, the relative distance between two dots indicates the relative semantic distance between the corresponding PSAs. Even in the relatively simple demonstration in Figure 3 that compares PSAs in three dimensions, we can identify clusters of similar documents. The precision of the similarity measure increases as the number of dimensions increases. Thus, the vector space employed in calculating our uniqueness scores uses more than 100-dimensions. This vector representation approach provides the basis for the creation of a numerical similarity/difference measure for the empirical analysis.

The content and semantic deviations across PSA documents are easily calculated once we obtain the vector representation of each PSA. Since the model

is trained to capture the semantic meanings of the documents, the relative cosine distance between the vectors represents the corresponding deviation:

$$\begin{aligned} U(\boldsymbol{v}_1, \boldsymbol{v}_2) &= 1 - cos(\boldsymbol{v}_1, \boldsymbol{v}_2) \\ &= 1 - \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{||\boldsymbol{v}_1|| \cdot ||\boldsymbol{v}_2||}. \end{aligned} \tag{5}$$

$U(v_1, v_2)$ is bounded between 0 and 1. A cosine distance score of 1 means the two documents are completely different from each other, whereas a distance score of 0 indicates an exact match. Therefore, the distance from a document to itself ($U(v_1, v_1)$) will always be 0.

The pairwise distance between two PSA documents increases as their semantic meanings deviates from each other. Since the relationship between the documents cannot be properly analyzed using simple methods based on keywords or word frequencies, the algorithm provides a substantial improvement in the ability to compare legal contracts.

Under the assumption of complete contracting, if underwriters view the mortgages in deal "A" to be significantly different from those in deal "B", and consequently reflect these differences in the PSA, we should observe a higher $U(v_A, v_B)$. However, if the PSAs are mostly legal boilerplate, then we would find a lower $U(v_A, v_B)$.

To provide context to how the algorithm works, we first demonstrate the methodology on a smaller scale by calculating pairwise uniqueness scores for Section 2.1, subsection (a) from five representative PSAs. This section identifies the "Conveyance" terms for transferring the mortgage pool from the under-

writer to the CMBS trust. Exhibits 1 through 5 in the Appendix correspond to these sections for the following CMBS deals: (1) Morgan Stanley Bank of America Merrill Lynch Trust 2012–C6; (2) LB-UBS Commercial Mortgage Trust 2007–C6; (3) Citigroup Commercial Mortgage Trust 2006–C5; (4) Banc of America Commercial Mortgage Inc. Commercial Mortgage Pass–Through Certificates, Series 2004-1; and (5) Banc of America Commercial Mortgage Inc., Commercial Mortgage Pass-Through Certificates, Series 2008–1.

The pairwise uniqueness score comparing Exhibits 1 and 2 is 0.55, indicating that these documents are relatively dissimilar. Likewise, the pairwise score comparing Exhibits 2 and 3 is 0.2, suggesting that these documents share a higher degree similarity compared to the pair in Exhibits 1 and 2. For instance, Exhibits 2 and 3 contain bullet points (i) to (iv) with the same subjects and order. Both list the conveyance between different parties and indicate the same end of the fiscal year for the trust. Finally, Exhibits 4 and 5 report Sections 2.1 (a) for two Banc of America deals. Not surprisingly, given that these deals are from the same underwriter, the pairwise uniqueness score is 0.015 revealing a high degree of overlap.

Based on this example, one may be concerned that the algorithm assigns a non-zero uniqueness score even if documents differ on trivial elements or elements that are already being captured by hard data (for example, origination year, number of loans, the geographic distribution of properties, etc.). To alleviate this concern, we perform the following exercise. We artificially constructed a comparison pseudo-PSA by altering the deal name, series numbers, and origination year for the PSA document associated with Banc of America

Commercial Mortgage Inc. Commercial Mortgage Pass-Through Certificates, Series 2004-1. In other words, we simply replaced the deal name, series identification numbers, and origination year in the pseudo-PSA to something completely different, leaving the rest of the document identical to the original. We then use the algorithm to calculate a uniqueness score for these documents. The uniqueness score between the original PSA and the corresponding pseudo-PSA is approximately 0, thus verifying that our algorithm correctly identifies these documents as being identical.[7]

We extend this methodology to the full sample and compute the pairwise scores between every single pair of the 975 CMBS deals in the full sample. The result is the 975 by 975 cosine distance matrix reported in Figure 4. We obtain Figure 4 by coloring each cell with its distance score magnitude. The deals are aligned alphabetically on the horizontal and vertical axes. Thus, the downward sloping diagonal represents the comparison of each deal with itself. The color white/cream represents a value of 0. We can easily identify that the low–distance scores are often found between deals that are located close to each other, which usually correspond to the PSAs from the same underwriter or similar time.

4 Documenting variation in PSAs

Our analysis starts with documenting the variation in the pairwise uniqueness score for each of the 975 PSAs in our sample. In particular, we document

[7]We report the PSA and pseudo-PSA comparison in our online Appendix (`http://yannans.people.clemson.edu/online_appendix.html`).

that variation across PSAs does exist, and we explore the variation within and across issuer and deal origination years.

Determing the control group

We begin by defining various comparison groups. For example, we can analyze the pairwise uniqueness of all deals originated by Bank of America or Wells Fargo to determine the exent that individual underwriters tailor the deal PSA to the differences in the underlying mortgage pools. Alternatively, we can aggregate deals across origination year cohorts to see how PSA documents evolve to reflect changing macroeconomic risk factors.

Let D represent the set of all deals in the sample ($N = 975$) with $D_i \subset D$ denoting the set of deals that have the same underwriter i. Similarily, we define $D_t \subset D$ as the set of deals originated in year t and $D_{it} \subset D$ as the set of deals originated by underwriter i in year t. Using the CMBS PSA deal vector notation from above, we let d_{jit} index each element of a respective set corresponding to a deal j by issuer i in year t. For example, we can compare the uniqueness of deal j to all other deals as:

$$U(d_i, d) \quad = \quad \frac{1}{|D| - 1} \sum_{d \in D} U(d_{jit}, d) \tag{6}$$

Our analysis considers the uniqueness of deal j with the following seven comparison groups: (1) all other deals regardless of issuer or origination year, (2) other deals from the same issuer (across all origination years), (3) deals from differnt issuers (across all years), (4) deals from the same origination year

(across all issuers), (5) deals from different years (across all issuers), (6) deals from the same issuer in the same year, and (7) deals from other issuers in the same year.

Figures 5 to 7 document the distribution of these deal level measures. Overall, the distributions confirm the intuition of the uniqueness score. For example, uniqueness scores are on-average lower (i.e. the PSAs are more similar) when the comparision groups are deals from the same underwriter (see Figures 5 and 7). However, substantial heterogeneity exists in the distributions with some deals being very different (high uniqueness score) when compared to deals from the same issuer. In addition, PSAs tend to be more similar for deals issued in the same year (Figure 6), although once again substantial heterogeneity exists.

Next, we further aggregate to the year level to observe time-series variation in these uniqueness scores. Figures 8 and 9 document these further aggregations. Figure 8 shows that, on average, deals did not become more or less similar to deals from other issuers over time. However, figure 8 shows a slight increase in the trend in PSA uniqueness over time for deals originated by the same issuer.[8] Figure 9 shows that there are a few underwriters (e.g. Banc of America, Citibank, Credit Suisse, UBS, and Goldman Sachs) who show substantial uniqueness in PSAs even among deals that they underwrote themselves.

[8]Deals after 2007 are removed due to small sample size.

What drives variation in PSAs?

Having established that there does exist meaningful variation in the PSAs across deals, we now explore what drives variation across CMBS deals. For this analysis, we focus on the conduit CMBS deals since the other deal types are either have too few observations for meaningful statistical analysis, or are backed by the government sponsored agencies and thus have vastly different credit risk. The analysis that follows will lend initial insight into the degree of contractual incompleteness inherent in these documents. The analysis thus far has ignored the underlying collateral, and thus, pairwise differences across PSAs may simply reflect differences in the underlying mortgage collateral. Therefore, in this section, we extend the analysis to map differences in the collateral to the uniqueness scores, assessing the extent to which mortgage characteristics predict differences in PSA documents.

Performing the analysis at the deal level, we regress our uniqueness score on the means and standard deviations of key loan-pool characteristics. We specifically focus on deal dispersion across observable characteristics such as loan-level contract terms, property geographic location, and property type. For example, a single CMBS deal may contain a mix of retail, multifamily, office, and other property types located in a large number of states. In fact, the rating agencies look to diversification across property type and geography as important factors in assessing the risk of mortgage-backed securities when assigning ratings. Thus, we ask whether differences in property-type concentrations between the mortgage pools of CMBS deals i and j are reflected in $U(d_i, d_j)$. For every pair (i, j), we compute the "distance" between the deal-

level observable variables ($|\Delta X_{ij}|$, where X is a matrix of the deal observables). We then test for the presence of contractual incompleteness by estimating the following regression:

$$U(d_i, d_j) = \alpha + \beta|\Delta X_{ij}| + \epsilon \tag{7}$$

Coefficient estimates on $|\Delta X_{ij}|$ that are significantly less than or equal to zero are consistent with contractual incompleteness—two PSAs will be more similar, despite having dissimilar mortgage pools. In contrast, significantly positive coefficients suggest contractual completeness along the dimensions of X. That is, as the observable difference between characteristics X increases, the PSAs for deals i and j become more unique.

We define ΔX_{ij} as follows:

$$\Delta X_{ij} = \frac{|X_i - X_j|}{\frac{1}{2}(X_i + X_j)} \tag{8}$$

Therefore, ΔX_{ij} is a vector of measures that compare the collateral underlying deals i and j. We use this normalization approach to facilitate comparisons across the estimated coefficients, β. The measure ΔX_{ij} has the support $[0, 2]$ for all values of X_i and X_j. Therefore $2 \times \beta_k$ predicts how $S(d_i, d_j)$ changes as two deals move from being perfectly identical ($\Delta X_{ij} = 0$) to drastically different ($\Delta X_{ij} = 2$), in terms of their underlying collateral.

We estimate equation 7 using a univariate and multivariate specification of X. We consider deal-level averages for the following key characteristics: loan balance, duration proxy (the difference between amortization term and

term to maturity, in months), interest rate, and the loan-to-value ratio. To examine differences in collateral type, we consider deal-level measures of the fraction of loans collateralized by officed, multifamily units, and retail properties. We also construct a HHI-based measure of property mix across all the property-type categories in the data. Additionally, we construct HHI measures of property location (MSA-level) to capture geographic dispersion of the underlying loan pool. To capture dispersion in the underlying collateral we include the standard deviation of *ln(Original loan balance)*, *ln(Duration proxy)*, loan-to-value ratio, and interest-rate. To capture differences in the size of the deal, we consider the *ln(Loan count for deal i)*.

We also include a set of fixed effects and dummy variables to capture deal pairs that have the same underwriter or were issued in the same year. To control for pairwise uniqueness scores driven by an underwriter's idiosyncratic tastes for particular contract types, we include a set of underwriter fixed effects for each deal. We include a set of issuance year fixed effects to control for the time-series trend in uniqueness scores.

Figure 10 plots the estimated β coefficients along with their confidence intervals for the univariate specification of X and reveals that several variables have a significant impact on deal similarity.

First, in terms of the overall magnitude of effect, the results indicate that PSA content responds to two key measures of mortgage risk: the average interest rate, and the average LTV ratio of the underlying collateral. Interpreting the coefficients, if two deals move from being exactly identical ($\Delta X_{ij} = 0$) to substantially different ($\Delta X_{ij} = 2$) with respect to average loan interest rate,

the uniqueness score is predicted to increase by approximately 0.5. This is a large increase considering that the score is distributed on the interval $[0, 1]$.

Second, Figure 10 shows that other variables describing the underlying pool collateral also have a positive, albeit smaller, impact on deal uniqueness. For example, we note that as the absolute difference increases in the number of loans in the pool, the average loan balance, and the property type concentration, the more different (unique) are the PSAs.

Third, confirming the graphical evidence for the distributions of uniqueness scores by underwriter and origination year we note that the coefficient on the 'same-underwriter' indicator variable is large and significantly negative indicating that two deals with the same underwriter have lower uniqueness scores, again suggesting that the PSA documents from the underwriter are more alike. Similarly, the negative and significant coefficient on origination year reveals that the contents of PSAs appear to be more similar for CMBS that were issued in the same year.

Interestingly, Figure 10 reveals that PSA pair uniqueness is not very responsive to deal-pair differences in geographic dispersion of the underlying properties or in differences in the percentages of property types represented in the pools. This is somewhat surprising given that rating agencies often focus on pool level diversification across geography and property type as measures of deal risk, which would suggest that PSA contractual completeness should increase along these dimensions.

Finally, we examine the effect of the dispersion in underlying collateral on PSA uniqueness by comparing the similarity of two deals based on the diversity

of their underlying collateral. We capture heterogeneity in the underlying collateral pools by the standard deviation of the LTV ratios, log loan balances, and interest rates of the underlying mortgages. ΔX_{ij} then compares a deal whose mortgages have very similar LTV ratios, for example, to another, which has a wide distribution of LTV ratios. As Figure 10 shows, PSA uniqueness increases as the dispersion in underlying collateral increases between deal pairs.

Table 2 repeats this analysis in a multivariate setting, sequentially adding regressors into the analysis. Each regression includes underwriter fixed effects and origination year fixed effects. Focusing on column (8), the complete regression specification, we note that the conclusions primarily remain the same as above. However, there are four interesting exceptions. In the multivariate setting, deal uniqueness decreases as average loan duration differences increase, and as the percentage of retail properties increase. We also see that deal uniqueness decreases as the dispersion increases in the underlying collateral with respect to average loan duration and LTV. We note that underwriter and year fixed effects by themselves explain over 21% of the variation in the uniqueness scores. Including all the covariates in the regression, as in Column 8, explains about a third of the variation.

5 Contractual Completeness and Deal Performance

Having established the existence of meaningful variation in CMBS PSA and the attributes associated with these differences, we now examine the potential

consequences of contractual incompleteness in CMBS. The PSA is a tri-party agreement that governs the relationship between the trustee (who represents the interests of the investors) and the relevant agents (the mortgage originators and servicers). The mortgage originators deliver collateral to the CMBS conduit and are held to the terms of the PSA. If the PSA is complete, then every requirement of the trustee and servicer will be explicitly spelled out, and every contingency accounted for in the security design. As a result, the loan originator would have limited ability to shirk on his actions or deliver mortgages that satisfy only the minimum contractible elements of the trustee's requirements. We examine this hypothesis by focusing on whether PSA completeness is related to the performance of mortgages contained in the CMBS pool and on whether servicer actions differ based on contractual completeness. Thus, when comparing individual CMBS deals with various comparison groups ($U(d_i, d)$) we hypothesize that conditional on the observable underlying collateral, loans in deals with more complete PSAs (i.e., higher $U(d_i, d_j)$) should exhibit lower ex-post probabilities of delinquency and lower yield spreads at origination.

We test our hypothesis by regressing loan origination yield $Spread_i$ and loan $Performance_i$ on $U(d_i, d)$ controlling for the characteristics of the deal's mortgages and pool. Second, we consider measures of servicing efficiency, for example, time spent in special servicing. We refine this approach by using a matching estimator. That is, for every deal in our sample, we find a "nearest-neighbor" deal, which contains observably similar mortgage collateral.

Is mortgage risk associated with PSA uniqueness?

To test whether our measure of PSA uniqueness captures whether the CMBS PSAs reflect observable differences in their collateral, we turn to a loan-level analysis. Table 1 shows the summary statistics for the underlying loan characteristics making up the CMBS deals. The average loan is $12.2 million with a loan-to-value (LTV) ratio of 67.3%. However, significant heterogeneity exists across LTV (ranging from 42% to 80% in the 5^{th} and 95^{th} percentiles, respectively). In addition, the loan-level risk premium (spread over 10-year Treasury) ranges between 71 basis points at the 5^{th} percentile to 295 basis points at the 95^{th} percentile. We also note significant heterogeneity in property types in the deals. For example, retail properties account for approximately 35% of the CMBS deal pool, while multi-family make up about 22%, and 17% are office properties. Finally, in terms of the *ex post* performance, we see that 12.9% of loans were transferred to the special servicer, and 11.9% entered delinquency within 9-years of origination.

Table 3 reports the estimation results for the loan-level regression of the probability of default (defined as transfer to special servicing within 7-years of origination) with the variable of interest being the deal-level uniqueness score. The master servicer typically transfers loans to special servicing in anticipation of default, and thus this measure is a more expansive measure of default risk. In this regression, each deal is compared to the deals originated in the same year cohort, and the standard errors are clustered at the deal level. The larger the uniqueness score, the less similar (more unique) the PSA is relative to the average deal PSA in the same year of origination. Column (1) reports

the baseline regression with no control variables. In this specification, the PSA uniqueness score is positive and statistically significant, indicating that loans in deals that are more unique have a higher likelihood of default. However, the adjusted R^2 is only 0.4%, indicating that the model without any controls is very weak. In columns (2) through (8), we systematically add various deal and loan level controls as well as fixed effects for origination year, underwriter, and location. Once we add control variables and fixed effects, the estimated coefficient for the similarity measure becomes negative and statistically significant, and the adjusted R^2 increases to 12.6% (column (8)). This indicates that loans in deals with PSA documents that are more unique from other deals have lower probabilities of default. The estimated coefficient is also economically significant. Since the mean default rate is 8.4%, the estimated coefficient reported in the fully specified model (column (8)) implies that a deal that has a fully unique PSA has a 50% lower ex post default risk than a deal that is identical to its comparison. This is consistent with the complete contracting hypothesis that, even after controlling for observable differences in the collateral, deals with more unique PSA contracts will have a lower risk. However, we note that the estimated coefficient on deal uniqueness, while still negative, is marginally significant (at the 10% level) when underwriter fixed effects are included. This suggests that the underwriter fixed effect soaks up the majority of the variation in uniqueness scores within the underwriter.

In Table 4, we repeat the analysis using the 60+ day delinquency status as the dependent variable as the proxy for default. Again, the negative coefficient for uniqueness confirms the finding that greater uniqueness is associated with

lower risk. As in model using the transfer to special servicing as the proxy for default, we note that inclusion of underwriter fixed effects (column (5)) soaks up the majority of the variation in uniqueness within underwriter. However, the coefficient remains negative.

Finally, in Table 5, we the results for the regression of the loan level interest rate spread at origination. In the first specification with only origination year fixed effects (column (1)), the coefficient for uniqueness is positive and statistically significant, indicating that underwriters tend to select loans with higher origination yield spreads for inclusion in deals that are more unique. As we add more controls and fixed effects (columns 2-5), the coefficient remains positive but is no longer statistically significant.

Is deal uniqueness associated with servicer efficiency?

In this section, we ask whether contracting (PSA uniqueness) affects the time between transfer to special servicing and default? In essence, we wish to know whether "better" contracting— or complete contracts—are associated with more efficient actions on the part of one of the contracting parties (the master server). The theory of complete contracts predicts that the master servicer should be quicker to transfer loans to special servicing in anticipation of default – resulting in a longer time between transfer and 60+ day delinquency. Furthermore, once transferred, the special servicer should be able to stave off default for longer. This suggests that better contracting should predict a larger number of months between the transfer to special servicing and default; whether through its effect on the former or the latter.

Servicing efficiency can be measured by looking at the outcome of transfers to special servicing. We begin this analysis by conditioning the sample on those loans that enter special servicing. For these set of loans, "better" (or more unique) contracting predicts a higher probability of a return from special servicing within 24 months.[9]

Table 6 shows that the coefficient on the deal level uniqueness measure is positive and statistically significant. The estimated coefficients suggest that, conditional on a loan being transferred to special servicing, loans in deals that are unique spend between 10 and 13 months longer in special servicing before delinquency than a similar loan in deals that are less unique. Thus, we find evidence consistent with the hypothesis that greater contracting uniqueness leads to more efficient monitoring by the master servicer.

Next, we examine the effectiveness of the special servicer by analyzing the probability that a loan is returned to performing status within 24-months of being referred to the special servicer. The estimated coefficients on deal uniqueness are negative and statistically significant (except when deal underwriter fixed effects are included). This suggests that loans in deals that are characterized as being more unique are less likely to return to performing status. This is consistent with the theory that the special servicer has less latitude to modify troubled loans in a complete contract environment. Thus, the more complete or unique the contract, the lower the probability that the special servicer will return the loan to performing status.

Finally, in Table 8, we repeat the analysis but restrict the sample to loans

[9]To account for censoring, we drop loans that enter special servicing after March 2017.

that were transferred to the special servicer and defaulted. In this more restricted set of loans, we continue to see that the estimated coefficient for deal uniqueness is negative but is generally no longer statistically significant. The interpretation remains the same: loans in deals that have more unique PSAs are less likely to return to performing status.

6 Conclusion

A complete contract specifies the rights and duties of the parties to the contract for all possible future states of the world. In the context of securitized mortgages, the Pooling and Servicing Agreement is the contract that determines the actions of the various parties to the security. The typical PSA is often viewed as containing legal boilerplate language and thus may appear similar despite differences in the underlying mortgages or deal structure.

In this study, we use the advantages of artificial intelligence to process large quantities of textual data. The advantage of this tool is that it allows the researcher to calculate the uniqueness of contracts relative to other deals. Our analysis reveals the following findings.

First, the univariate regression analysis demonstrates that Pooling and Servicing Agreements are incomplete contracts on average. The uniqueness scores between two deals do not depend on differences in those deal characteristics, which, a priori, we expect to be important; for example, geographic dispersion of property, the mix of property-type, or the size of the deal. This, despite the regressions capturing a substantial amount of variation (R-square of 0.28)

in uniqueness scores. At the same time, there does appear to be substantial variation across deals in the degree of contractual incompleteness.

Second, in examining the consequences of incomplete contracts, our analysis reveals that deals with more unique PSA documents have underlying loans that have lower ex post default rates. This suggests that underwriters and investors do respond to the underlying risk in the collateral pools by requiring that PSA documents reflect the uniqueness of the collateral.

Finally, our analysis offers insight into the efficiency of the mortgage servicers as a result of contractual completeness. The results indicate that when contracts are more complete, servicers are less likely to engage in modification activities that would return loans in default to performing status.

References

Ambrose, B. W., Sanders, A. B., and Yavas, A. (2016). Servicers and Mortgage-Backed Securities Default: Theory and Evidence. *REAL ESTATE ECONOMICS*, 44(2):462–489.

An, X., Deng, Y., and Gabriel, S. A. (2009). Value Creation through Securitization: Evidence from the CMBS Market. *JOURNAL OF REAL ESTATE FINANCE AND ECONOMICS*, 38(3):302–326. APRU Symposium on Real Estate Research, Singapore, SINGAPORE, 2007.

An, X., Deng, Y., and Gabriel, S. A. (2011). Asymmetric information, adverse selection, and the pricing of CMBS. *JOURNAL OF FINANCIAL ECONOMICS*, 100(2):304–325.

An, X., Deng, Y., Nichols, J. B., and Sanders, A. B. (2015). What is Subordination About? Credit Risk and Subordination Levels in Commercial Mortgage-backed Securities (CMBS). *JOURNAL OF REAL ESTATE FINANCE AND ECONOMICS*, 51(2):231–253. Maastricht-NUS-MIT (MNM) International Real Estate Finance and Economics Symposium, MIT, Cambridge, MA, OCT 25-26, 2013.

An, X., Deng, Y., and Sanders, A. B. (2008). Subordination Levels in Structured Financing. In Thakor, AV and Boot, AWA, editor, *HANDBOOK OF FINANCIAL INTERMEDIATION AND BANKING*, Handbooks in Finance, pages 41–60.

Begley, T. A. and Purnanandam, A. (2017). Design of Financial Securities: Empirical Evidence from Private-Label RMBS Deals. *REVIEW OF FINANCIAL STUDIES*, 30(1):120–161.

Beltran, D. O., Cordell, L., and Thomas, C. P. (2017). Asymmetric information and the death of ABS CDOs. *JOURNAL OF BANKING & FINANCE*, 76:1–14.

Benmelech, E. and Dlugosz, J. (2009). The alchemy of CDO credit ratings. *JOURNAL OF MONETARY ECONOMICS*, 56(5):617–634.

Dai, A. M., Olah, C., and Le, Q. V. (2015). Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*.

DeMarzo, P. (2005). The pooling and tranching of securities: A model of informed intermediation. *REVIEW OF FINANCIAL STUDIES*, 18(1):1–35.

DeMarzo, P. and Duffie, D. (1999). A liquidity-based model of security design. *ECONOMETRICA*, 67(1):65–99.

Demiroglu, C. and James, C. (2012). How Important is Having Skin in the Game? Originator-Sponsor Affiliation and Losses on Mortgage-backed Securities. *REVIEW OF FINANCIAL STUDIES*, 25(11):3217–3258.

Downing, C., Jaffee, D., and Wallace, N. (2009). Is the Market for Mortgage-Backed Securities a Market for Lemons? *REVIEW OF FINANCIAL STUDIES*, 22(7):2457–2494.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Gennaioli, N., Shleifer, A., and Vishny, R. W. (2013). A Model of Shadow Banking. *JOURNAL OF FINANCE*, 68(4):1331–1363.

Glaeser, E. and Kallal, H. (1997). Thin markets, asymmetric information, and mortgage-backed securities. *JOURNAL OF FINANCIAL INTERMEDIATION*, 6(1):64–86.

Hanson, S. G. and Sunderam, A. (2013). Are there too many safe securities? Securitization and the incentives for information production. *JOURNAL OF FINANCIAL ECONOMICS*, 108(3):565–584.

Hart, O. (2017). Incomplete contracts and control. *American Economic Review*, 107(7):1731–52.

Hartman-Glaser, B., Piskorski, T., and Tchistyi, A. (2012). Optimal securitization with moral hazard. *JOURNAL OF FINANCIAL ECONOMICS*, 104(1):186–202.

Lacker, J. (2001). Collateralized debt as the optimal contract. *REVIEW OF ECONOMIC DYNAMICS*, 4(4):842–859.

Longstaff, F. A. and Rajan, A. (2008). An empirical analysis of the pricing of collateralized debt obligations. *JOURNAL OF FINANCE*, 63(2):529–563.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Malamud, S., Rui, H., and Whinston, A. (2013). Optimal incentives and securitization of defaultable assets. *JOURNAL OF FINANCIAL ECONOMICS*, 107(1):111–135.

Maskara, P. K. (2010). Economic value in tranching of syndicated loans. *JOURNAL OF BANKING & FINANCE*, 34(5):946–955.

Matvos, G. (2013). Estimating the benefits of contractual completeness. *The Review of Financial Studies*, 26(11):2798–2844.

Mooradian, R. M. and Pichler, P. (2018). Servicer Contracts and the Design of Mortgage-Backed Security Pools. *REAL ESTATE ECONOMICS*, 46(3):698–738.

Nini, G., Smith, D. C., and Sufi, A. (2009). Creditor control rights and firm investment policy. *Journal of Financial Economics*, 92(3):400–420.

Nini, G., Smith, D. C., and Sufi, A. (2012). Creditor control rights, corporate governance, and firm value. *The Review of Financial Studies*, 25(6):1713–1761.

Pagano, M. and Volpin, P. (2010). Credit ratings failures and policy options. *ECONOMIC POLICY*, (62):401–431.

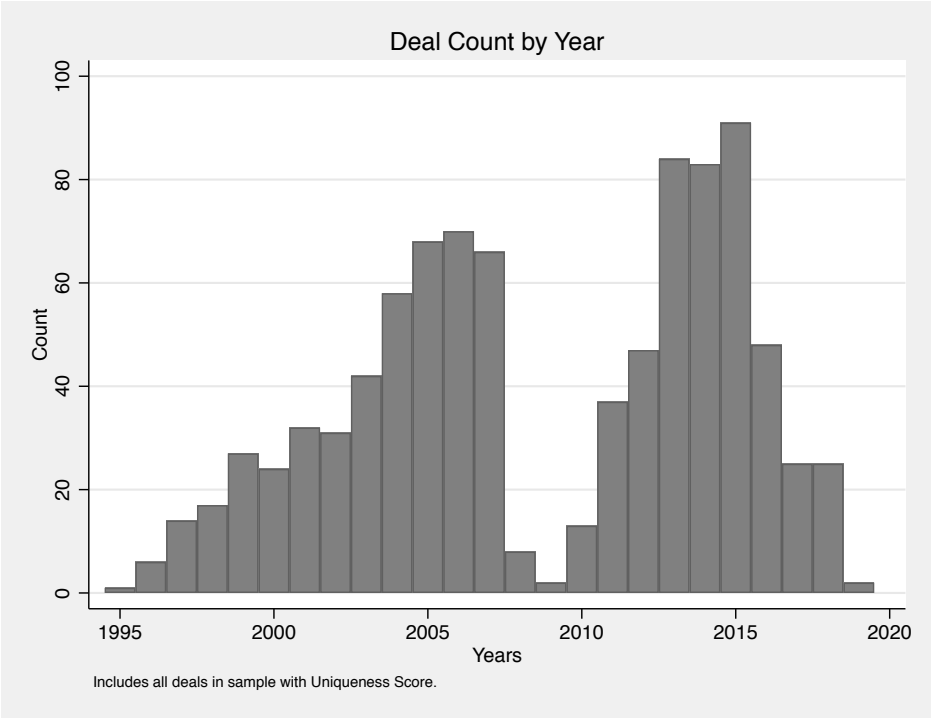
Piskorski, T., Seru, A., and Witkin, J. (2015). Asset Quality Misrepresentation by Financial Intermediaries: Evidence from the RMBS Market. *JOURNAL OF FINANCE*, 70(6):2635–2678.

Riddiough, T. (1997). Optimal design and governance of asset-backed securities. *JOURNAL OF FINANCIAL INTERMEDIATION*, 6(2):121–152.

Shen, Y. (2018). Information value of property description: A machine learning approach. *Available at:* `https://ssrn.com/abstract=3281223` *or* `https://dx.doi.org/10.2139/ssrn.3281223`.

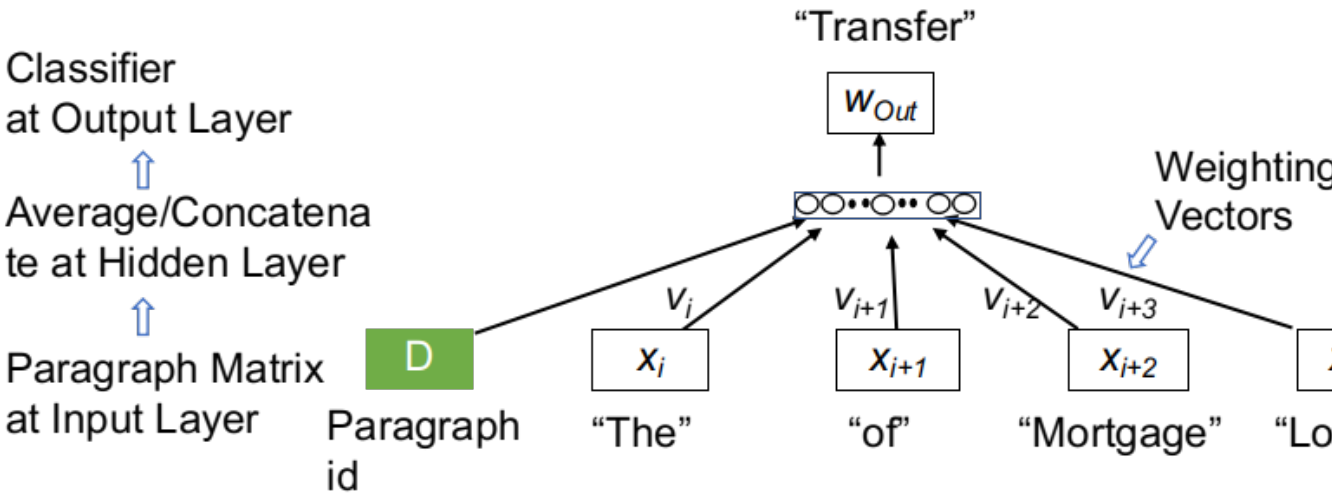
Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.

Figure 1: **Frequency Count of CMBS Deals by Origination Year**



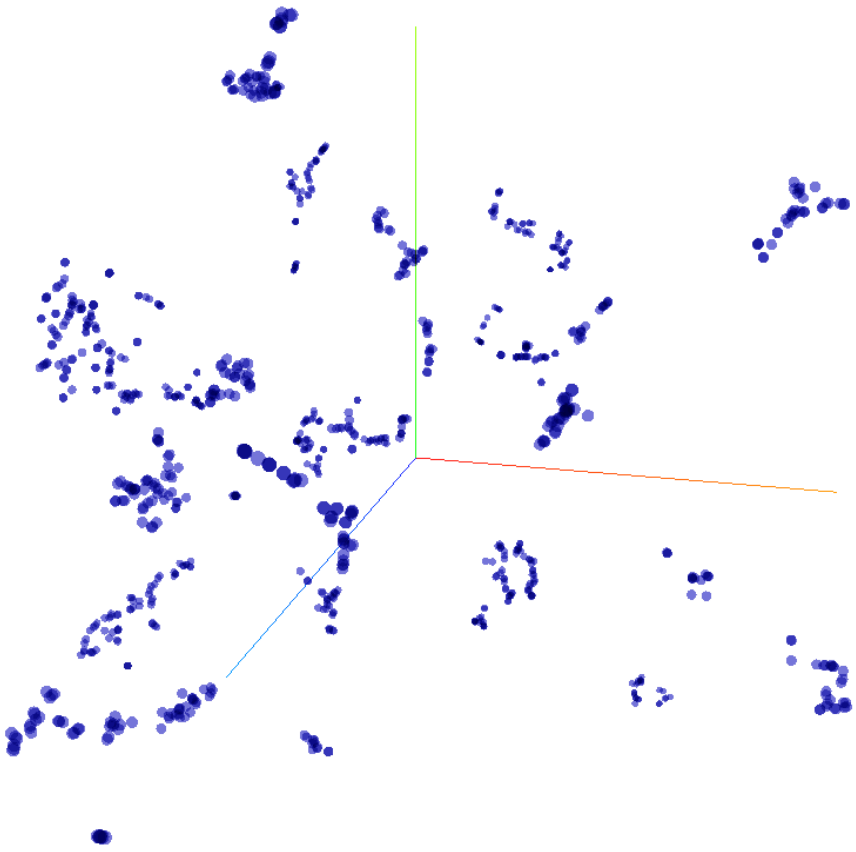
Note: In the figure, we plot the frequency count of the 975 CMBS deals by year of origination.

Figure 2: Schematic Algorithm



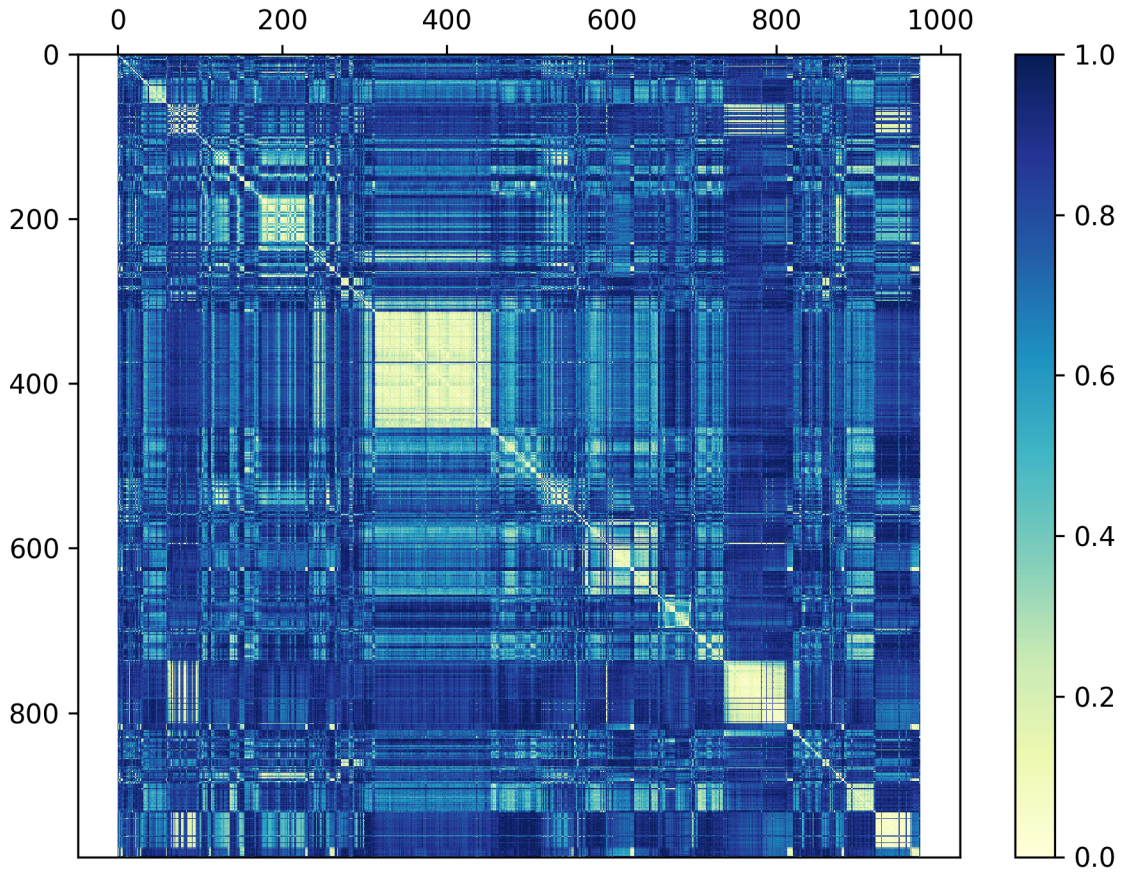
Notes: This figure illustrates the process of projecting a PSA into a 7–dimensional vector space. It is important to emphasize that this simplified example is only created for demonstration purposes. The actual learning algorithm is more sophisticated and projects PSA documents into a space with over 100 dimensions.

Figure 3: 3–D Representation of Vector Space for CMBS PSAs



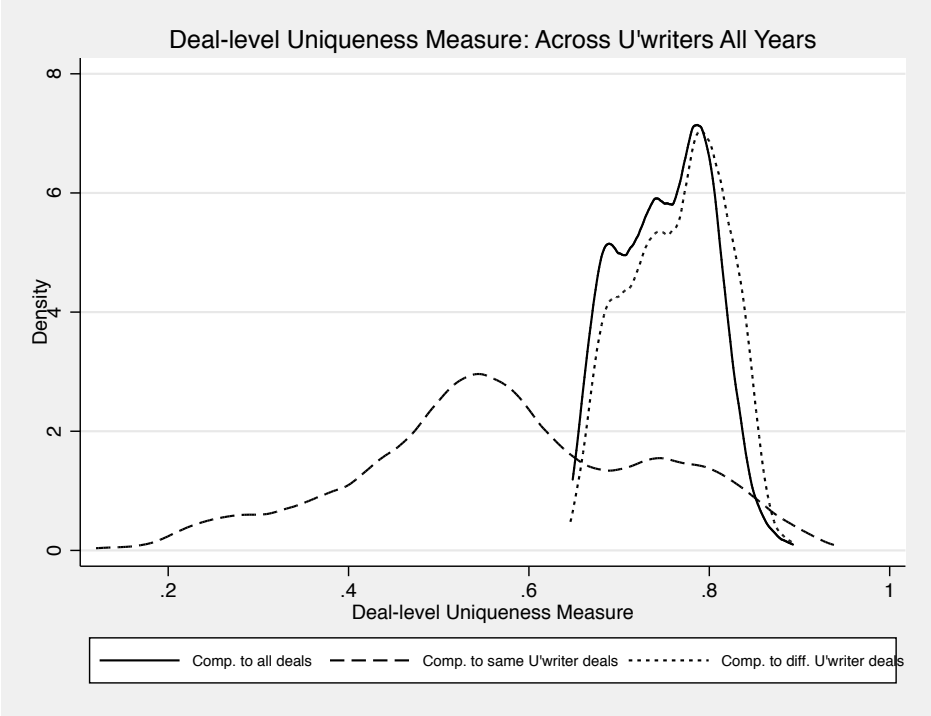
Notes: This Figure offers a 3–D demonstration of the vector space for the PSA files. In reality, our vector space has more than 100–dimensions. Every blue dot represents a unique PSA document. Each axis in the vector space does not hold a physical meaning in human language; the relative distance between two dots indicates the relative semantic distance between the corresponding PSAs.

Figure 4: **Distance Score Visualization of CMBS Deals**



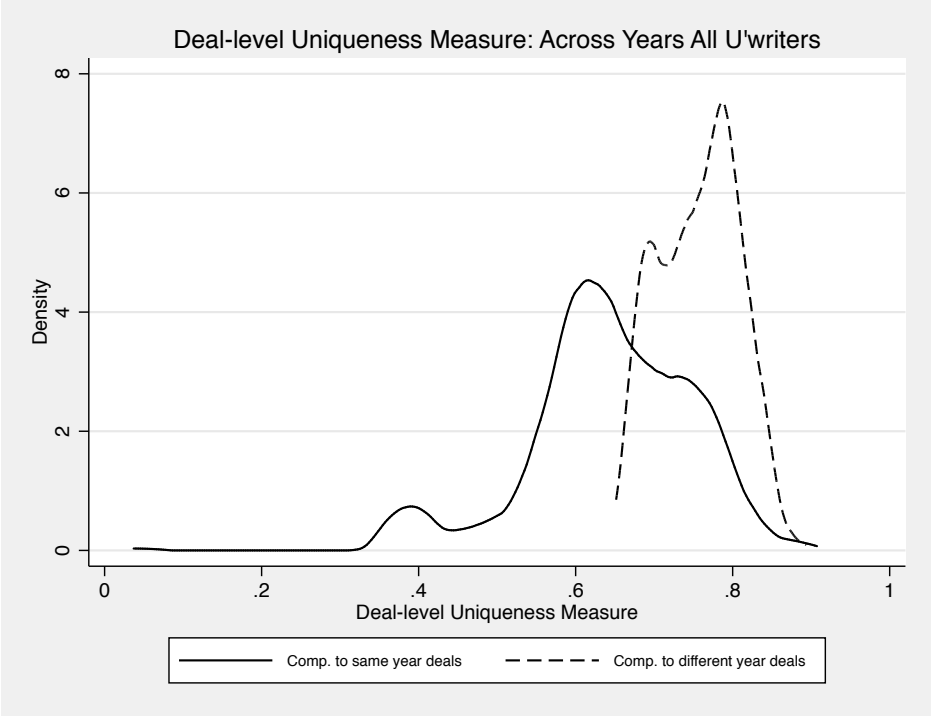
Note: Color scheme: White/Yellow: similar documents, Dark Blue: dissimilar documents. Distance score to itself will always result in a 0 value and similar documents are usually from the same underwriter. Therefore the white colors are often found close to the diagonal axis of the matrix.

Figure 5:



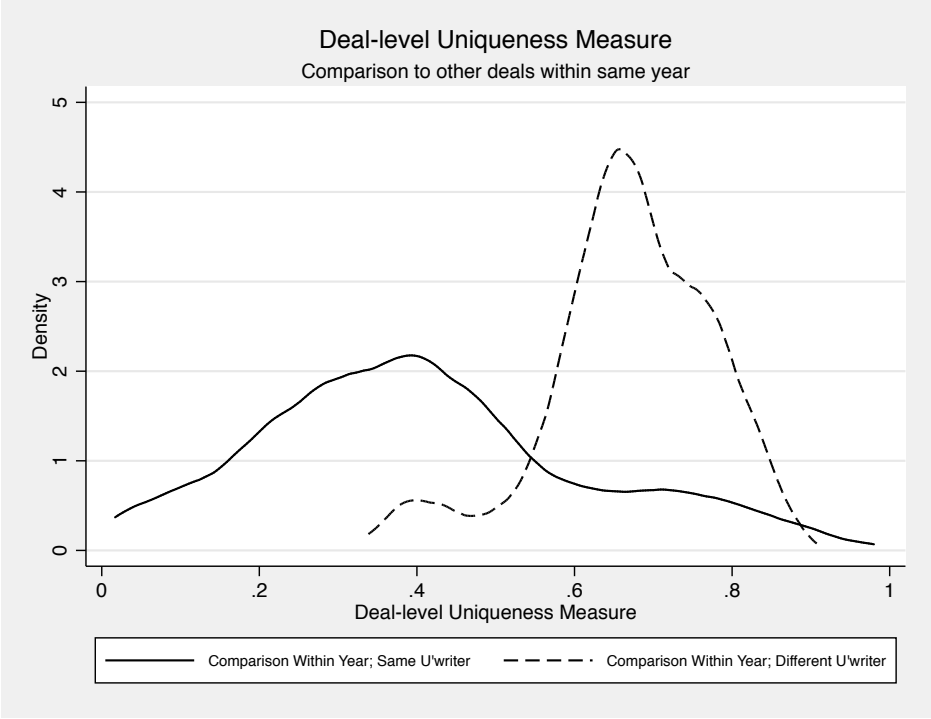
Notes:

Figure 6:



Notes:

Figure 7:



Notes:

Figure 8: Average Deal Level Uniqueness Measure Over Time

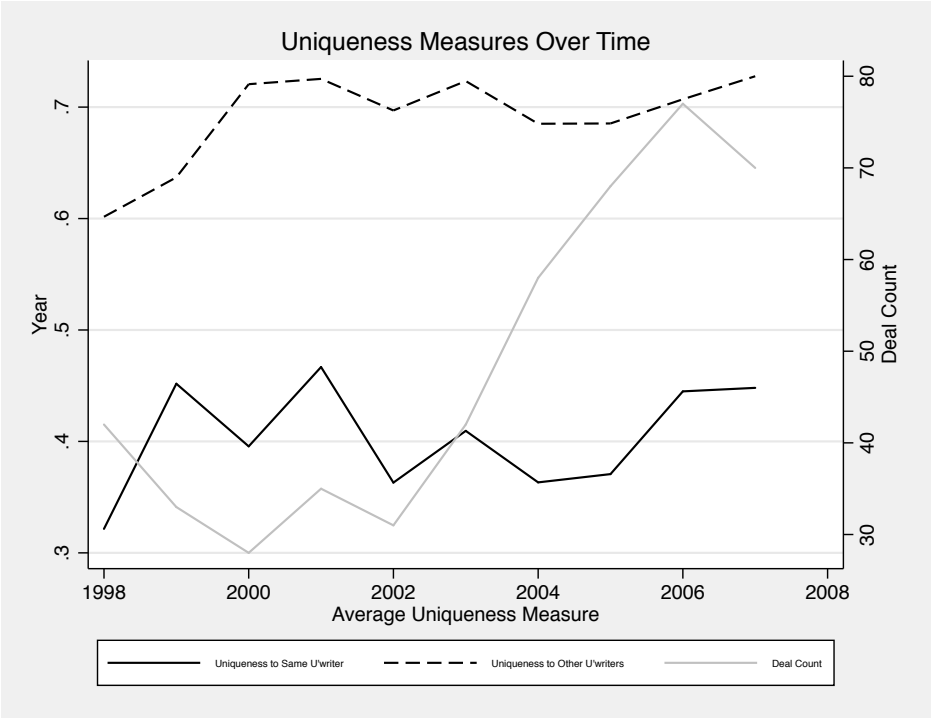
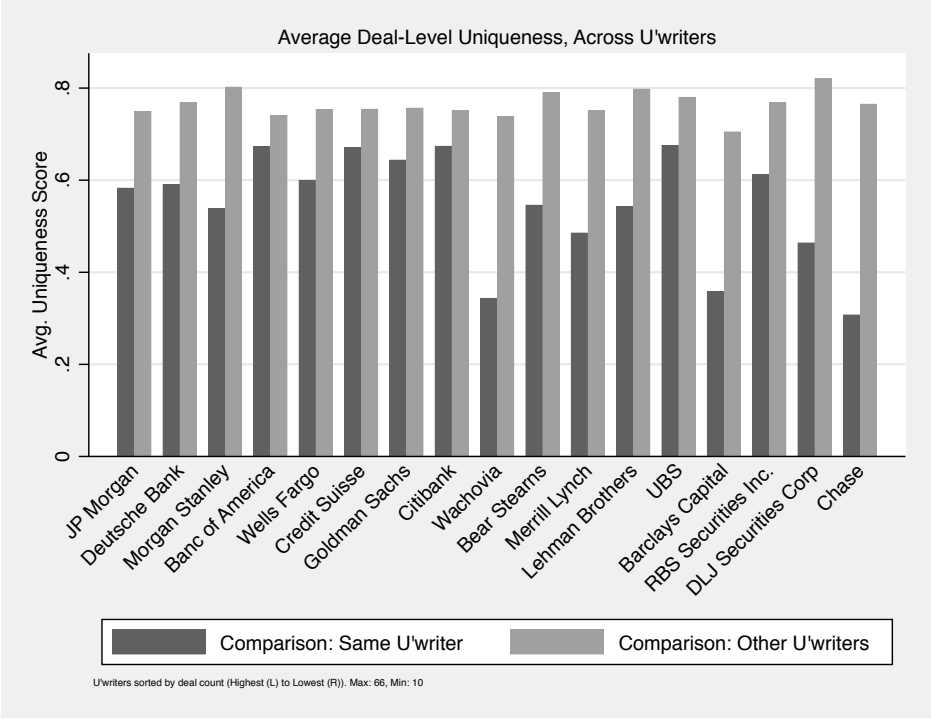
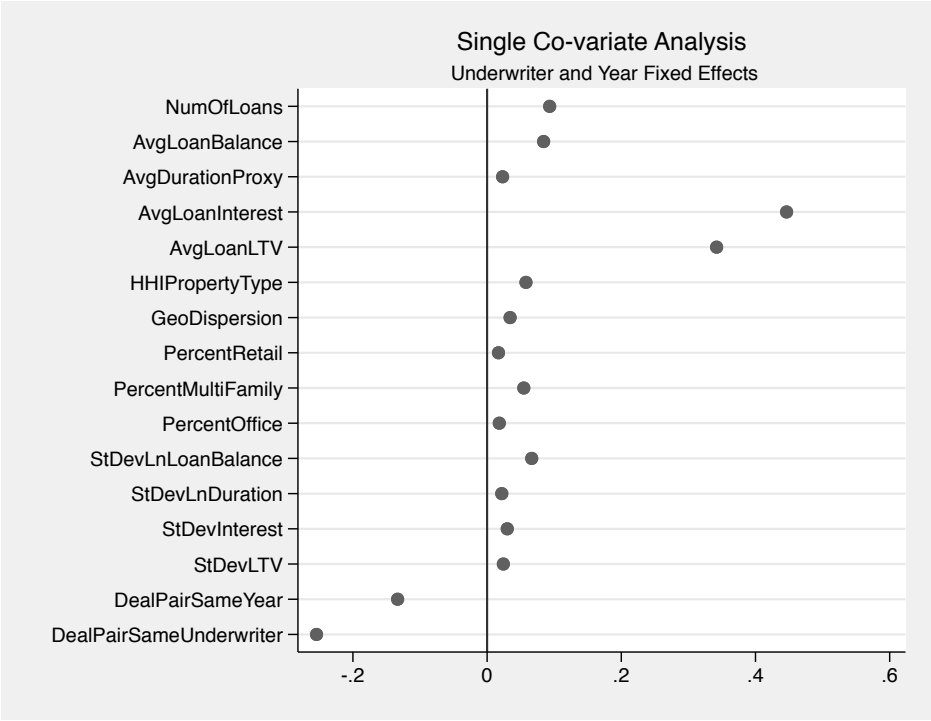


Figure 9: Average Deal Level Uniqueness Measure Across Issuers



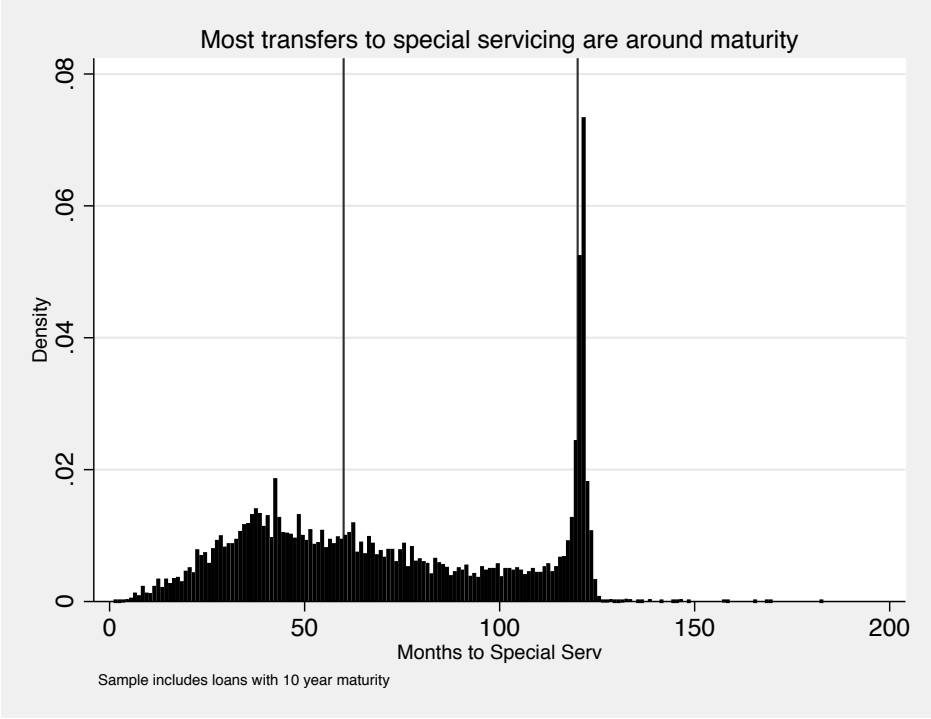
Notes:

Figure 10: Single Covariate Analysis; Deal-Pair Regression Coefficients



Notes:

Figure 11: Time to Transfer to Special Servicer



Notes:

Table 1: Loan Level Summary Statistics

VARIABLES	(1) N	(2) mean	(3) sd	(4) p5	(5) p50	(6) p95
LTV	62,726	0.673	0.134	0.420	0.708	0.799
Loan Balance Millions	62,726	12.23	24.68	1.225	5.850	41.10
Interest Rate (bps)	62,726	582.0	97.34	436.1	573.6	782
Spread to 10 Yr Treasury (bps)	62,707	173.9	71.44	79.70	160	295
Occupancy Rate	62,726	0.937	0.0878	0.743	0.970	1
Seasoned Loan Indicator	62,726	0.0777	0.268	0	0	1
Retail	62,726	0.346	0.476	0	0	1
Office	62,726	0.173	0.378	0	0	1
Multifamily	62,726	0.218	0.413	0	0	1
Transfer to Spec. Serv. w/in 9 yrs	62,726	0.129	0.335	0	0	1
60+ Days Delinquency w/in 9 yrs	62,726	0.119	0.324	0	0	1
Return from SS w/in 2 yrs	11,742	0.199	0.399	0	0	1
Time b.w. Delinquency and Special Serv.	4,785	-0.166	11.76	-16	1	9

Table 2: Multivariate Regression: Dependent variable is similarity score ($S(d_i, d_j)$)

Dependent Variable: Similarity Score								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AvgLoanBalance	0.026***		0.024***	0.023***		0.025***		0.008***
	(0.001)		(0.001)	(0.001)		(0.001)		(0.001)
AvgDurationProxy	-0.014***		-0.023***	-0.011***		-0.014***		-0.017***
	(0.001)		(0.002)	(0.002)		(0.002)		(0.002)
AvgLoanInterest	0.400***		0.397***	0.402***		0.404***		0.331***
	(0.003)		(0.003)	(0.003)		(0.003)		(0.003)
AvgLoanLTV	0.253***		0.247***	0.253***		0.359***		0.329***
	(0.004)		(0.004)	(0.004)		(0.005)		(0.005)
HHIPropertyType		0.054***	0.032***	0.033***		0.040***		0.037***
		(0.002)	(0.002)	(0.002)		(0.002)		(0.002)
GeoDispersion		0.031***	0.002**	0.003***		0.006***		0.002**
		(0.001)	(0.001)	(0.001)		(0.001)		(0.001)
PercentRetail				-0.016***		-0.017***		-0.016***
				(0.001)		(0.001)		(0.001)
StDevLnLoanBalance					0.064***	0.018***		0.010***
					(0.003)	(0.003)		(0.003)
StDevLnDuration					0.013***	-0.002		-0.002*
					(0.001)	(0.001)		(0.001)
StDevInterest					0.028***	0.006***		0.003**
					(0.002)	(0.002)		(0.001)
StDevLTV					0.016***	-0.062***		-0.062***
					(0.002)	(0.002)		(0.002)
DealPairSameUnderwriter							-0.246***	-0.243***
							(0.002)	(0.002)
DealPairSameYear							-0.109***	-0.064***
							(0.002)	(0.002)
Num Of Loans							0.076***	0.046***
							(0.0009)	(0.001)
Observations	243,951	249,571	241,860	241,454	241,164	238,735	251,695	238,735
R-squared	0.172	0.084	0.173	0.174	0.082	0.180	0.219	0.282
U'writer FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Observation: Deal Pair Dependent Variable: Pairwise Uniqueness Score (*** p<0.01, ** p<0.05, * p<0.1).

Table 3: Transfer to Special Servicing

Dependent Variable: Transfer to Special Servicer in 8–Year (d)						
	(1)	(2)	(3)	(4)	(5)	(6)
Uniqueness	0.173***	-0.167***	-0.115***	-0.114***	-0.106***	-0.048*
	(0.010)	(0.015)	(0.023)	(0.023)	(0.023)	(0.025)
Observations	62,693	62,674	62,674	62,674	56,617	56,617
R-squared	0.004	0.061	0.092	0.092	0.123	0.126
Deal Controls	No	No	No	Yes	Yes	Yes
Prop Type	No	No	Yes	Yes	Yes	Yes
Loan Controls	No	No	Yes	Yes	Yes	Yes
Year FE	No	Yes	Yes	Yes	Yes	Yes
Underwriter FE	No	No	No	No	No	Yes
MSA FE	No	No	No	No	Yes	Yes
Cluster	Deal	Deal	Deal	Deal	Deal	Deal
Mean Dep Var	0.129	0.129	0.129	0.129	0.130	0.130

Notes: Variable: Indicator for transfer to special servicing within 7 years of origination.
(*** p<0.01, ** p<0.05, * p<0.1).

Table 4: Serious Delinquency

Dependent Variable: Serious Delinquency					
	(1)	(2)	(3)	(4)	(5)
Uniqueness	-0.148***	-0.0983***	-0.0970***	-0.0927***	-0.0335
	(0.0144)	(0.0218)	(0.0216)	(0.0214)	(0.0232)
Observations	62,674	62,674	62,674	56,617	56,617
R-squared	0.058	0.087	0.087	0.119	0.122
Deal Controls	No	No	Yes	Yes	Yes
Prop Type	No	Yes	Yes	Yes	Yes
Loan Controls	No	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Underwriter FE	No	No	No	No	Yes
MSA FE	No	No	No	Yes	Yes
Cluster	Deal	Deal	Deal	Deal	Deal
Mean Dep Var	0.119	0.119	0.119	0.120	0.120

Notes: Dependent Variable: Indicator for entry into serious delinquency within 5 years of securitization. (*** p<0.01, ** p<0.05, * p<0.1).

Table 5: Interest Rate Spread to 10–year Treasury

Dependent Variable: Interest Rate Spread to 10–Year Treasur (bps)					
	(1)	(2)	(3)	(4)	(5)
Uniqueness	16.78***	7.069	8.756	10.44	7.800
	(1.833)	(9.472)	(9.246)	(9.046)	(10.06)
Observations	62,674	62,674	62,674	56,617	56,617
R-squared	0.685	0.730	0.732	0.738	0.742
Deal Controls	No	No	Yes	Yes	Yes
Prop Type	No	Yes	Yes	Yes	Yes
Loan Controls	No	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Underwriter FE	No	No	No	No	Yes
MSA FE	No	No	No	Yes	Yes
Cluster	Deal	Deal	Deal	Deal	Deal
Mean Dep Var	173.9	173.9	173.9	172.7	172.7

Notes: Variable: Dependent Variable: Interest rate Spread to 10 year treasury.
(*** p<0.01, ** p<0.05, * p<0.1).

Table 6: Time Between Transfer to Special Servicing and Default

Dependent Variable: Months to Default						
	(1)	(2)	(3)	(4)	(5)	(6)
Uniqueness	10.37***	14.71***	14.21*	14.17**	14.10**	13.06**
	(2.015)	(2.210)	(7.446)	(6.431)	(5.800)	(5.925)
Observations	4,785	4,785	4,785	4,785	4,349	4,349
R-squared	0.006	0.016	0.046	0.051	0.205	0.226
Deal Controls	No	No	No	Yes	Yes	Yes
Prop Type	No	No	Yes	Yes	Yes	Yes
Loan Controls	No	No	Yes	Yes	Yes	Yes
Year FE	No	Yes	Yes	Yes	Yes	Yes
Underwriter FE	No	No	No	No	No	Yes
MSA FE	No	No	No	No	Yes	Yes
Cluster	Deal	Deal	Deal	Deal	Deal	Deal
Mean Dep Var	-0.166	-0.166	-0.166	-0.166	-0.164	-0.164

Notes: Dependent Variable: Months to Default (Default Date – Special Servicing Transfer Date) (note, this could be negative). Observation: Loan Level Sample: Loans that went delinquency AND were transferred to special servicing (*** p<0.01, ** p<0.05, * p<0.1).

Table 7: Return from Special Servicing: All Loans

Dependent Variable: Returned from Special Servicing in 24 months (d)						
	(1)	(2)	(3)	(4)	(5)	(6)
Uniqueness	-0.148***	-0.161***	-0.133*	-0.133*	-0.154**	-0.0116
	(0.0479)	(0.0503)	(0.0748)	(0.0762)	(0.0729)	(0.109)
Observations	11,227	11,227	11,227	11,227	10,192	10,192
R-squared	0.001	0.005	0.022	0.023	0.072	0.079
Deal Controls	No	No	No	Yes	Yes	Yes
Prop Type	No	No	Yes	Yes	Yes	Yes
Loan Controls	No	No	Yes	Yes	Yes	Yes
Year FE	No	Yes	Yes	Yes	Yes	Yes
Underwriter FE	No	No	No	No	No	Yes
MSA FE	No	No	No	No	Yes	Yes
Cluster	Deal	Deal	Deal	Deal	Deal	Deal
Mean Dep Var	0.205	0.205	0.205	0.205	0.202	0.202

Notes: Sample: Loans that were transferred to special servicing. Indicator of whether loan returned from special servicing within 24 months
(*** p<0.01, ** p<0.05, * p<0.1).

Table 8: Return from Special Servicing: Delinquent Loans

Dependent Variable: Returned from Special Servicing in 24 months (d)					
	(1)	(2)	(3)	(4)	(5)
Uniqueness	-0.110	-0.113	-0.103	-0.167*	-0.0396
	(0.0765)	(0.0942)	(0.0981)	(0.0959)	(0.171)
Observations	4,678	4,678	4,678	4,253	4,253
R-squared	0.013	0.028	0.031	0.124	0.130
Deal Controls	No	No	Yes	Yes	Yes
Prop Type	No	Yes	Yes	Yes	Yes
Loan Controls	No	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Underwriter FE	No	No	No	No	Yes
MSA FE	No	No	No	Yes	Yes
Cluster	Deal	Deal	Deal	Deal	Deal
Mean Dep Var	0.196	0.196	0.196	0.196	0.196

Notes: Sample: Loans that were transferred to special servicing. Indicator of whether loan returned from special servicing within 24 months
(*** p<0.01, ** p<0.05, * p<0.1).

Appendix: Scoring Examples

This appendix provides an illustration of the ML based document comparison method applied to a brief section from the CMBS PSA. In the interest of brevity, we selected subsection (a) from Section 2.1 for five representative CMBS deals. The PSA Section 2.1 subsection (a) identifies the "Conveyance" terms for transferring the mortgage pool from the underwriter to the CMBS trust.

Exhibits 1 through 5 report the text used in this scoring example and correspond to the following CMBS deals: (1) Morgan Stanley Bank of America Merrill Lynch Trust 2012–C6; (2) LB-UBS Commercial Mortgage Trust 2007–C6; and (3) Citigroup Commercial Mortgage Trust 2006–C5.

The pairwise uniqueness score for Exhibits 1 and 2 is 0.55, indicating that these documents are relatively dissimilar. Likewise, the pairwise score for Exhibits 2 and 3 is 0.2, suggesting that these documents share a relatively high degree of common elements. For instance, both the two documents discuss bullet points from (i) to (iv) with the same subjects and order. They both list the conveyance between different parties and indicate the end of the fiscal year of the Trust is at the same time.

Note that the pairwise uniqueness scores reported above measure the similarities of Section 2.1 of these PSAs. When comparing full PSAs, our algorithm focuses on the overall contract completeness and minimizes any immaterial differences such as origination years and deal names. To show this, we artificially constructed pseudo-PSAs by altering the deal names/series numbers/origination years in the PSA to something completely different, leaving

the rest of the PSAs identical. The uniqueness scores between an original PSA and its corresponding pseudo-PSA is approximately 0. We show a comparison between two PSAs (main text body of a PSA for a CMBS deal and its modification) in our online Appendix (`http://yannans.people.clemson.edu/online_appendix.html`).

To extend the use of cosine distance scores, we calculated the pairwise scores between every single pair of the 975 CMBS deals. The result is a 975 by 975 cosine distance matrix. We obtain Figure 4 by coloring each cell with its distance score magnitude. From the color scheme, we can easily identify that the low–distance scores are often found between deals that are located close to each other, which usually correspond to the PSAs from the same underwriter or similar time. Such visualization can provide insights into the overall dataset and also used for characterizing the sample distance score distribution.

Exhibit 1: Morgan Stanley Bank of America Merrill Lynch Trust 2012-C6

ARTICLE II
DECLARATION OF TRUST;
ISSUANCES OF CERTIFICATES
Section 2.1 Conveyance of Mortgage Loans. (pages 107-108)

> (a) Effective as of the Closing Date, the Depositor does hereby establish a trust designated as "Morgan Stanley Bank of America Merrill Lynch Trust 2012-C6" and assign in trust to the Trustee, without recourse, for the benefit of the Certificateholders all the right, title and interest of the Depositor, in, to and under (i) the Mortgage Loans identified on the Mortgage Loan Schedule including the related Mortgage Notes, Mortgages, security agreements and title, hazard and other insurance policies, including all Qualifying Substitute Mortgage Loans, all distributions with respect thereto payable after the Cut-Off Date, the Mortgage File and all rights, if any, of the Depositor in the Distribution Account, all REO Accounts, the Collection Account and the Reserve Accounts, (ii) the Depositor's rights under each Mortgage Loan Purchase Agreement that are permitted to be assigned to the Trustee pursuant to Section 14 thereof, (iii) the Initial Deposit, (iv) the Depositor's rights under any Intercreditor Agreement, Non-Serviced Mortgage Loan Intercreditor Agreement and the related Non-Serviced Mortgage Loan Pooling and Servicing Agreement with respect to any Non-Serviced Mortgage Loan, (v) with respect to the EC Trust Certificates, each of the EC Trust REMIC III Regular Interests and (vi) all other assets included or to be included in REMIC I or the Class J Grantor Trust. Such assignment includes all interest and principal received or receivable on or with respect to the Mortgage Loans and due after their respective Due Dates in October 2012. The transfer of the Mortgage Loans and the related rights and property accomplished hereby is absolute and is intended by the parties to constitute a sale. In connection with the initial sale of the Certificates by the Depositor, the purchase price to be paid includes a portion attributable to interest accruing on the Certificates from and after October 1, 2012. The transfer and assignment of any Non-Serviced Mortgage Loans to the Trustee and the right to service such Mortgage Loans are subject to the terms and con-

ditions of the related Non-Serviced Mortgage Loan Pooling and Servicing Agreement and the related Non-Serviced Mortgage Loan Intercreditor Agreement, and the Trustee, by the execution and delivery of this Agreement, hereby agrees that such Mortgage Loans remain subject to the terms of the related Non-Serviced Mortgage Loan Intercreditor Agreement and, with respect to each Serviced Pari Passu Mortgage Loan and Serviced Companion Loan, the related Intercreditor Agreement. The transfer and assignment of any A Notes and Serviced Pari Passu Mortgage Loans to the Trustee and the right to service such Mortgage Loans are subject to the terms of the related Intercreditor Agreements, and the Trustee, by the execution and delivery of this Agreement, hereby agrees, that such Mortgage Loans remain subject to the terms of the related Intercreditor Agreements (or with respect to a Joint Mortgage Loan treated as a Loan Pair in accordance with Section 8.30 hereof, the applicable Mortgage Loan documents and Section 8.30 hereof).

Exhibit 2: LB-UBS Commercial Mortgage Trust 2007-C6

ARTICLE II

CONVEYANCE OF TRUST MORTGAGE LOANS; REPRESENTATIONS AND WARRANTIES;
ORIGINAL ISSUANCE OF CERTIFICATES

SECTION 2.01. Creation of Trust; Conveyance of Trust Mortgage Loans. (page 122)

(a) It is the intention of the parties hereto that multiple common law trusts be established pursuant to this Agreement and the laws of the State of New York and that such trusts be designated as: "LB-UBS Commercial Mortgage Trust 2007-C6", in the case of the Mortgage Trust individually or all the subject trusts collectively, as the context may require; "Class A-2FL Grantor Trust", in the case of Grantor Trust A-2FL; and "Class A-MFL Grantor Trust", in the case of Grantor Trust A-MFL. LaSalle is hereby appointed, and does hereby agree, to act as Trustee hereunder and, in such capacity, to hold the Trust Fund in trust for the exclusive use and benefit of all present and future Certificateholders.

The Depositor, concurrently with the execution and delivery hereof, does hereby assign, sell, transfer, set over and otherwise convey to the Trustee in trust, without recourse, for the benefit of the Certificateholders, all the right, title and interest of the Depositor in, to and under (i) the Trust Mortgage Loans, (ii) the UMLS/Depositor Mortgage Loan Purchase Agreement(s), (iii) any Co-Lender Agreement(s), and (iv) all other assets included or to be included in the Trust Fund. Such assignment includes all interest and principal received or receivable on or with respect to the Trust Mortgage Loans and due after the Cut-off Date and, in the case of each Trust Mortgage Loan that is part of a Loan Combination, is subject to the provisions of the related Co-Lender Agreement. With respect to each Trust Mortgage Loan that is part of a Loan Combination, the Trustee, on behalf of the Trust, assumes the obligations of the holder of such Trust Mortgage Loan and the related Mortgage Note under, and agrees to be bound by, the related Co-Lender Agreement.

The parties hereto acknowledge and agree that, notwithstanding Section 11.07, the transfer of the Trust Mortgage Loans and the related rights and property accomplished hereby is absolute and is intended by them to constitute a sale.

The Trust Fund shall constitute the sole assets of the Trust. Except as expressly provided herein, the Trust may not issue or invest in additional securities, borrow money or make loans to other Persons. The fiscal year end of the Trust shall be December 31.

Exhibit 3: Citigroup Commercial Mortgage Trust 2006-C5

ARTICLE II

CONVEYANCE OF MORTGAGE LOANS; REPRESENTATIONS AND WARRANTIES; ORIGINAL ISSUANCE OF CERTIFICATES

SECTION 2.01 Conveyance of Trust Mortgage Loans. (page 92)

(a) The Depositor, concurrently with the execution and delivery hereof, does hereby establish a common law trust under the laws of the State of New York, designated as "Citigroup Commercial Mortgage Trust 2006-C5", and does hereby assign, sell, transfer, set over and otherwise convey to the Trustee, in trust, without recourse, for the benefit of the Certificateholders (and for the benefit of the other parties to this Agreement as their respective interests may appear) all the right, title and interest of the Depositor, in, to and under (i) the Trust Mortgage Loans and all documents included in the related Mortgage Files and Servicing Files, (ii) the rights of the Depositor under Sections 1, 2, 3 and 5 (and to the extent related to the foregoing, Sections 8 through 17 and 19) of each of the Mortgage Loan Purchase Agreements, (iii) the rights of the Depositor under each Co-Lender Agreement and (iv) all other assets included or to be included in the Trust Fund. Such assignment includes all interest and principal received or receivable on or with respect to the Trust Mortgage Loans and due after the Cut-off Date and, in the case of each Trust Mortgage Loan that is part of a Loan Combination, is subject to the provisions of the corresponding Co-Lender Agreement. The Trustee, on behalf of the Trust, assumes the rights and obligations of the holder of the Mortgage Note for each Combination Mortgage Loan under the related Co-Lender Agreement; provided that Master Servicer No. 2 and the Special Servicer shall, as further set forth in Article III, perform the servicing obligations of the holder of the Mortgage Note for each A-Note Trust Mortgage Loan under the related Co-Lender Agreement. The transfer of the Trust Mortgage Loans and the related rights and property accomplished hereby is absolute and, notwithstanding Section 11.07, is intended by the parties to constitute a sale.

The Trust Fund shall constitute the sole assets of the Trust. Except as expressly provided herein, the Trust may not issue or invest in additional securities, borrow money or make loans to other Persons. The fiscal year end of the Trust shall be December 31.

Finally, Exhibits 4 and 5 report Sections 2.1 (a) for two Banc of America deals. Not surprising, given that these deals are from the same underwriter, the pairwise uniqueness score is 0.015 revealing a high degree of overlap, which is one order of magnitude lower than the other sample comparisons. This comparison can also serve as a sanity check for the proposed algorithm, which demonstrates that the model can pick out very fine details between documents and quantify them at a basis that can be shared across the entire document pool.

Exhibit 4: Banc of America Commercial Mortgage Inc. Commercial Mortgage Pass-Through Certificates, Series 2004-1

> It is the intention of the parties hereto that a common law trust be established pursuant to this Agreement and further such trust be designated as "Banc of America Commercial Mortgage Inc. Commercial Mortgage Pass-Through Certificates, Series 2004-1". Wells Fargo Bank, N.A. is hereby appointed, and does hereby agree to act, as Trustee hereunder and, in such capacity, to hold the Trust Fund in trust for the exclusive use and benefit of all present and future Certificateholders. It is not intended that this Agreement create a partnership or a joint stock association.

Exhibit 5: Banc of America Commercial Mortgage Inc., Commercial Mortgage Pass-Through Certificates, Series 2008-1

> It is the intention of the parties hereto that a common law trust be established pursuant to this Agreement and further such trust be designated as "Banc of America Commercial Mortgage Inc., Commercial Mortgage Pass-Through Certificates, Series 2008-1". Wells Fargo Bank, N.A. is hereby appointed, and does hereby agree to act, as Trustee hereunder and, in such capacity, to hold the Trust Fund in trust for the exclusive use and benefit of all present and future Certificateholders. It is not intended that this Agreement create a partnership or a joint-stock association.