Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0000-0000 | EISSN 0000-0000 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Online Supplement: Modeling Financial Products and Their Supply Chains

Margrét V. Bjarnadóttir
University of Maryland mbjarnad@umd.edu,

Louiqa Raschid
University of Maryland, lraschid@umd.edu,

A. Brief Introduction to Topic Modeling

LDA is a statistical model that aims to explain a set of documents using unobserved topics. LDA is based on a generative statistical model for collections of discrete data (documents); it allows us to extract topics based on assumptions about the probability distributions that were used to generate the documents. At a high level, it is assumed that to create a document, one first randomly chooses a distribution over a collection of (unobserved) topics. Then, for each word in a document, one independently and randomly chooses a topic from the previously sampled mix of topics assigned to the document, then again independently and randomly draws a word from the (unobserved) word distribution for that topic. Based on these assumptions about the generative mechanism and the associated underlying probability distributions, topic models aim to extract the underlying topic structure via maximum likelihood estimation.

More specifically, let $P(z)$ represent the distribution for a topic z from K topics in a particular document. Let $P(w|z)$ represent the probability distribution over words w for the given topic z. Let $P(z_i = j)$ be the probability that the j-th topic was sampled for the i-th word token in the document. Let $P(w_i|z_i = j)$ be the probability of word w_i for topic j. We then have the following probability distribution for words within a document:

$$P(w_i) = \sum_{j=1}^{K} P(w_i|z_i = j) \times P(z_i = j) \tag{1}$$

LDA represents each document as a random mix over latent topics. Each topic is characterized by a distribution over words. Associated with these two distributions are the hyperparameters α and β, corpus level parameters assumed to be sampled once in the process of generating a corpus. The variable θ is a document level variable (a topic mixture over K topics, sampled from a Dirichlet distribution parameterized by α) sampled once per document. The variables

z_n and w_n are word level variables sampled once for each word in each document. Specifically, for each word, a topic z_n is sampled from a multinomial distribution conditioned on θ, and each word w_n is sampled from a multinomial distribution conditioned on z_n and parameterized by β.

One can rewrite the probability of a document with N words as follows:

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta))d\theta \tag{2}$$

where $\mathbf{w} = \{w_1, w_2, \ldots, w_N\}$ is a set of N words. Figure 1 is a graphical representation of the LDA process for a collection of M documents.

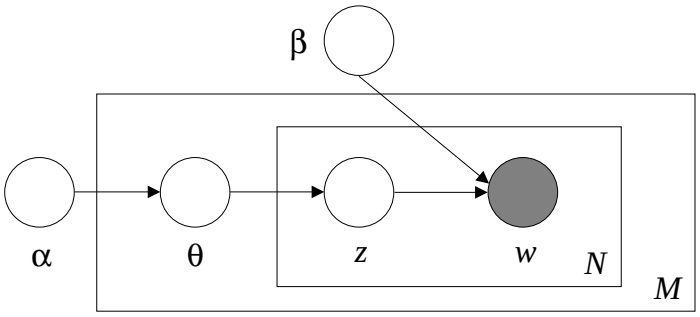


Figure 1 **A graphical representation of LDA for a text corpus (Blei et al. 2003). The boxes are "plates" representing replicates, where the outer plate represents M documents and the inner plate represents the repeated choice of topics and words within a document.**

The LDA process just described assumes that the documents are drawn from a static set of topics. Extensions that relax this assumption include a continuous non-Markovian model of topics over time (ToT) (Wang and McCallum 2006) and dynamic topic models (DTM) (Blei and Lafferty 2006). ToT assumes that each topic is associated with a continuous distribution over the time slices. This allows a modeler to model how some topics may be more popular over certain time periods. In that case, for each document, the mix of topics (θ in LDA) is influenced by both the word co-occurrence and the timestamps. In the DTM extension, which is the approach we adopt, the document collection is also divided by time, e.g., in our case a time slice for each year. As the Dirichlet distribution does not lend itself well to sequential modeling, the time dependent parameters evolve over time according to a Brownian motion. For example, $\beta_{t,k}$ can be expressed as follows:

$$\beta_{t,k}|\beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, \sigma^2 I). \tag{3}$$

The evolution of α and θ are expressed in a similar manner. We point the interested reader to Blei and Lafferty (2006) for additional technical details. Figure 2 is a graphical representation of the DTM process. Through the process described above, the topics can evolve over time, with the speed of evolution controlled by a tuning parameter. Setting the parameter at one extreme will convert DTM to a static topic model, while setting it at the other extreme will result in independent topics in each time slice.

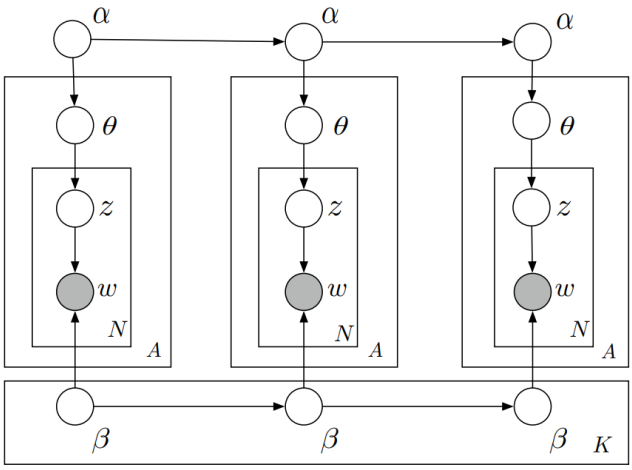


Figure 2 **A graphical representation of a dynamic topic model (DTM) for three time slices (Blei and Lafferty 2006). The parameters for each topic, $\beta_{t,k}$, as well as the parameter for the topic mix α_t, evolve over time.**

B. Data Details and Summary Statistics

Role	Frequency maximum: 4787	%	Frequency maximum: 3537	%
Issuer	3676	76	3183	89
Originator	1757	36	1713	48
Seller	3131	65	2588	73
Trustee	3722	77	3284	92
Servicer (many variants)	4335	90	3502	98
Depositor	4081	85	3314	93
Sponsor	1731	36	1651	46
Securities administrator	873	18	842	23
Custodian	733	15	706	19
Swap counterparty	516	10	507	14
Cap counterparty	424	8	392	11
Insurer	355	7	330	9
Underwriter	228	4	143	4

Table 1 Distribution of roles across prospectuses; 4,787 prospectuses (total) and 3,537 prospectuses (with Bloomberg data).

	Word Count	Distinct Word Count	Entity Count	Distinct Entity Count	Role Role	Distinct Role Count
mean	513	106	64	9.5	51	7.6
min	97	28	4	1	5	2
max	1767	234	318	25	190	13
var	72371.8	1382.4	1791.3	17.6	744	3.9

Table 2 Summary statistics for tokens extracted from the prospectuses in `resMBS`.

Attribute	Description
CUSIP:	Unique identifier.
Name:	Unique name.
Tranche Description:	Essential characteristics of multi-class mortgage and asset-backed securities (CMO, ABS, CMBS).
Original Principal:	The principal balance at issuance of the security.
Maturity:	Date the principal of a security is due and payable.
Count of Loans:	The current number of loans, created as collateral for the deal, which are still outstanding.
60+ DQ:	Percentage of loans which are 60 or more days delinquent, including loans in foreclosure, bankruptcy.
Cumulative shortfall:	Cumulative supported shortfall that has yet to be repaid.
2 shortfall:	Cumulative shortfall that will not be repaid.
Cumulative Loss:	Cumulative writedown on the principal.
Historic Cashflow:	A complete set of known historical cashflows starting from the bond issuance. The data includes the period number, date, coupon, interest, principal paid, and principal balance.
Ratings:	Multiple fields with the original and current ratings from Moody, Fitch, and SP; corresponding dates and a composite rating from Bloomberg L.P.

Table 3 Attributes for securities (Bloomberg).

	2002	2003	2004	2005	2006	2007	2008	Total
Prospectus count (initial)	452	640	881	1047	1019	667	81	4787
With issuer or originator and match in Bloomberg	231	398	577	744	923	593	71	3537
With >= 3 DISTINCT FE mentions								3146
With >= 5 (Role, FE) mentions								4472

Table 4 Count of prospectuses: with issuer and/or originator identified (4,787); with match in Bloomberg (3,537); further filtered for specific models.

	`resMBS`
Count of documents	4472
Count of distinct financial institutions	85
Count of distinct roles	27
Distinct (Role_FI) pairs	267
Count of (Role_FI) occurrences	41075

Table 5 Summary statistics for `resMBS` topic model experiments.

Features	Level	Number of Features	Description
Moody's initial rating	Security	11	Over 60 initial rating values are grouped together into 9 higher level rating levels, in addition to a value for "not rated" and "no value."
Payoff characteristics	Security	73	The payoff structure of a security is captured with 73 binary indicators.
Initial mortgage amount	Security	1	The initial mortgage amount of the pool of mortgages serving the security.
Security class	Security	3	The class of the security, _A_, _M_, or _B_.
Presence of SSUP	Prospectus	1	A binary indicator for whether any security within a prospectus is a `SSUP` security.
Class distribution in a prospectus	Prospectus	6	Six variables that capture the fraction of the number securities within each class (_A_, _M_, or _B_) and the volume (in USD) within the same.
Financial communities	Prospectus	30	Binary indicator variables for the topic that each prospectus is assigned to using the largest topic weight.
Annual controls	Security	6	Binary indicators for the year of issuance.

Table 6 Overview of the features used in the study.

C. Detailed LASSO Model Results

All Securities

Model	Accuracy		F1-Score	
	FE	FNE	FE	FNE
Security Level Model	91.3%	84.7%	0.891	0.857
Prospectus Level Model	91.2%	85.0%	0.889	0.860
Comprehensive Community Level Model	91.6%	85.4%	0.894	0.864

Class _A_

Model	Accuracy		F1-Score	
	FE	FNE	FE	FNE
Security Level Model	95.8%	82.6%	0.734	0.623
Prospectus Level Model	95.7%	82.7%	0.732	0.639
Comprehensive Community Level Model	95.9%	82.6%	0.737	0.644

Class _M_

Model	Accuracy		F1-Score	
	FE	FNE	FE	FNE
Security Level Model	86.0%	85.0%	0.903	0.909
Prospectus Level Model	86.2%	86.0%	0.904	0.914
Comprehensive Community Level Model	86.5%	86.2%	0.906	0.916

Class _B_

Model	Accuracy		F1-Score	
	FE	FNE	FE	FNE
Security Level Model	87.2%	88.0%	0.912	0.930
Prospectus Level Model	86.5%	88.1%	0.907	0.930
Comprehensive Community Level Model	87.0%	89.3%	0.910	0.937

Table 7 Performance metrics for the three LASSO models, for the entire dataset and for each security class, _A_, _M_ and _B_.

Variable	FE	FNE
Intercept	-0.717	0.933
Class & Amount		
MTG.ORIG.AMT	-0.003	
IsA	-4.669	-3.307
IsB	0.079	0.046
Tranche Type		
AD	-0.449	-0.014
AFC	-0.248	0.305
AS	0.223	0.705
CMPLX	-0.198	
CPT	0.98	0.88
CSTR	0.31	0.65
DLY	0.602	0.066
EXCH	-0.194	0.666
EXE	-1.468	
FLT	-0.093	-0.135
FTV	-0.336	
INV	-0.145	
IRC	-0.06	-0.064
MEZ	0.971	
MR	-0.543	
NAS	1.019	0.698
NTL	-0.311	-0.519
OC	-3.215	
PAC1	-0.512	-0.24
PT	0.056	
R	0.139	
RAKE	-0.508	-0.303
RSTP	-0.071	
RTL		1.83
SC	-3.208	
SEQ	-1.243	-0.35
SSNR	-1.2	0.294
SSUP	4.413	2.188
STEP	-0.106	
SUB	1.433	0.496
SUP		-0.074
TAC.22.	-0.391	
W	-0.289	
Z		1.486

Variable	FE	FNE
Moody's Initial Rating		
A	0.198	
Aa	-0.599	-0.53
Aaa	-0.252	-0.063
B	0.144	0.005
Ba	0.81	0.735
Baa	0.517	0.572
C	1.827	0.065
Not rated	0.516	0.319
Annual Controls		
2002	-4.518	-3.297
2003	-3.15	-2.119
2004	-1.673	-1.292
2006	2.08	1.291
2007	2.113	1.364

Table 8 The regression coefficients for the security level model. Blank cells indicate that the variable was not retained by the FE or FNE model. Note that variables that are not retained by either the FE or FNE model are not included.

Variable	FE	FNE
Intercept	-0.635	1.260
Class & Amount		
MTG.ORIG.AMT	-0.00309	-0.00022
IsA	-4.658	-3.439
IsB	0.045	0.028
Tranche Type		
AD	-0.438	-0.136
AFC	-0.309	0.343
AS	0.062	0.665
CMPLX	-0.149	-0.023
CPT	0.946	1.216
CSTR	0.249	0.608
DGT		-0.279
DLY	0.622	0.157
EXCH	-0.107	0.759
EXE	-1.262	
FLT	-0.238	-0.270
FTV	-0.500	-0.152
INV	-0.171	
IRC	-0.063	-0.091
MEZ	0.987	
MR	-0.622	
NAS	0.824	0.587
NTL	-0.579	-0.838
OC	-3.307	
PAC1	-0.486	-0.436
PIP		-0.091
PT	0.050	-0.141
RAKE	-0.636	-0.975
RSTP	-0.309	-0.229
RTL		2.225
SC	-3.198	
SEQ	-1.272	-0.527
SSNR	-1.315	0.115
SSUP	4.196	2.037
STEP	-0.016	
SUB	1.386	0.477
SUP		-0.307
TAC.1.22.		0.338
TAC.11.		-0.041
TAC.2.22.		-0.096
TAC.22.	-0.434	
W	-0.231	
Z		1.636

Variable	FE	FNE
Moody's Initial Rating		
A	0.207	0.025
Aa	-0.611	-0.595
Aaa	-0.320	-0.050
B	0.168	0.218
Ba	0.863	0.932
Baa	0.554	0.682
C	1.932	0.513
Not rated	0.457	0.294
Prospectus Characteristics		
Frac A	-0.729	-0.748
HasSSUP	0.573	0.538
Vol A	0.282	0.070
Vol M	-0.487	-0.449
Annual Controls		
2002	-4.367	-3.365
2003	-3.081	-2.201
2004	-1.647	-1.379
2005	2.047	1.340
2006	2.017	1.381

Table 9 The regression coefficients for the prospectus level model. Blank cells indicate that the variable was not retained by the FE or FNE model. Note that variables that are not retained by either the FE or FNE model are not included.

Variable	FE	FNE
Intercept	-0.454	1.368
Class & Amount		
MTG.ORIG.AMT	-0.003	-6.02E-05
IsA	-4.9	-3.513
IsB	0.233	0.179
Tranche Type		
AD	-0.32	-0.031
AFC	-0.17	0.388
AS		0.588
CMPLX	-0.013	
CPT	0.858	1.221
CSTR	0.407	0.698
DGT		-0.004
DLY	0.547	0.239
EXCH	-0.085	0.768
EXE	-0.803	
FLT	-0.258	-0.399
FTV	-0.325	-0.198
INV	-0.166	-0.042
IRC	-0.009	-0.055
MEZ	0.833	
NAS	0.761	0.539
NTL	-0.563	-0.961
OC	-3.326	
PAC1	-0.211	-0.315
PIP		-0.468
PT	0.146	
RAKE	-0.293	-0.906
RSTP		-0.125
RTL		2.243
SC	-3.192	
SEQ	-0.929	-0.408
SSNR	-1.289	0.093
SSUP	4.317	2.093
SUB	1.096	0.373
SUP		-0.258
TAC.1.22.		0.032
TAC.11.		-0.206
TAC.22.	-0.186	
TAC.33.		-0.018
Z		1.694
Annual Controls		
2002	-4.206	-3.192
2003	-3.064	-2.164
2004	-1.611	-1.327
2006	1.995	1.351
2007	2.094	1.484

Variable	FE	FNE
Moody's Initial Rating		
A	0.188	0.01
Aa	-0.627	-0.613
Aaa	-0.43	-0.072
B	0.157	0.175
Ba	0.845	0.917
Baa	0.538	0.664
C	1.861	0.664
Not rated	0.462	0.266
Prospectus Characteristics		
FracA	-0.471	-0.56
HasSSUP	0.467	0.449
VolA	0.12	
VolM	-0.751	-0.661
Topics		
Topic 1		-0.667
Topic 10	0.371	
Topic 11	-0.108	-0.247
Topic 13		0.261
Topic 15	-1.34	-1.585
Topic 16	-0.316	-0.349
Topic 17	-0.285	-0.099
Topic 18	0.489	0.194
Topic 19	-1.076	
Topic 2	-0.662	-0.061
Topic 20	-0.335	0.027
Topic.21	-0.441	
Topic 22		-0.12
Topic 24	0.353	0.082
Topic 25	0.187	0.055
Topic 26	0.412	0.449
Topic 27	0.328	0.371
Topic 28	-0.122	-0.246
Topic 29	-1.308	-0.457
Topic 3	0.636	0.292
Topic 30	-0.249	-0.358
Topic 4	-1.09	-0.949
Topic 5	-0.242	-0.071
Topic 6		0.032
Topic 7	0.323	0.567
Topic 8	0.974	0.364
Topic.9	0.027	

Table 10 The regression coefficients for the comprehensive community level model. Blank cells indicate that the variable was not retained by the FE or FNE model. Note that variables that are not retained by either the FE or FNE model are not included.

D. Evaluation of BERT Embeddings on the `resMBS` Dataset

The state of the art for language understanding and document processing is represented by large language models (LLMs), e.g., Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019). We report on our evaluation using a BERT-like model for the task of clustering `resMBS` prospectuses based on their supply chain similarity.

Sentence embedding is an approach to map sentences from documents to vectors of real numbers. Sentence BERT is an efficient model that can identify pairs of the most similar sentences across a corpus. Such sentence similarity can be extended in a straightforward manner to address document similarity. Identifying all of the best matching sentence pairs can be computationally expensive. We use `all-mpnet-base-v2`, a sentence embedding approach that has been trained using Siamese BERT networks; it has been extensively evaluated and found to provide efficient performance in assessing sentence similarity (Reimers and Gurevych 2019). We refer to this model as mpnet-BERT.

While BERT-like models are very successful in exploiting general document similarity, several limitations of the `resMBS` dataset hinder the task of identifying the supply chain communities comprising financial entities and their roles. The following evaluation discusses these limitations, presents our results, and identifies opportunities for further research.

D.1. Strengths and Limitations of the `resMBS` Dataset for Use with BERT Sentence Embeddings

The `resMBS` dataset comprises prospectuses – legal contracts – that are hundreds to thousands of pages and contain details of the securities, the covenants of the legal contract, details of the mortgages, etc. Most prospectuses contain one or two pages of free text that specify the supply chain (financial institutions and their roles), and such content typically makes up less than 1% of the prospectus.

The prospectuses in the `resMBS` dataset are often a part of a *series* created and issued by the same financial institution. The text for each of the prospectuses in the series is often generated from an identical document template, leading to high *document similarity*. Our evaluation will reflect the strength of mpnet-BERT in exploiting this document similarity, derived from the use of a common document template, in order to identify and cluster similar prospectuses, within a series.

However, in contrast to document similarity, the DTM topic model identifies shared communities along the financial supply chain. Communities comprise (Role, Financial-entity) pairs extracted from the prospectuses. We label the similarity of prospectuses that share a community as *supply chain similarity*. These prospectuses are clustered within a topic by the DTM topic model. We differentiate pairs of prospectuses that share supply chain similarity from pairs of prospectuses that share high document similarity (because they were generated using the same document template). This is very important since our research indicates that it is the supply chain that can have an impact on the financial performance of securities within some prospectus.

We observe that the following characteristics of the `resMBS` dataset can hinder mpnet-BERT from utilizing supply chain similarity to cluster prospectuses:

- Most BERT models are pre-trained using general corpora, e.g., English Wikipedia or the Brown corpus. However, a specialized vocabulary with role terms (e.g., mortgage originator, prospectus issuer, servicer, trustee) plays an outsize role in the (Role, Financial-entity) pairs that are used to identify supply chain similarity. Such role terms may not be very common in the corpora used to train mpnet-BERT.

- The task of Named Entity Resolution (NER) – matching across the mentions of the names of financial institutions that occur in the community – is another challenge. For example, the same entity may be referred to as `Wells`, `Wells Fargo`, `Wells N.A.`, etc., across the corpus. However, the pre-processing pipeline for mpnet-BERT typically does not perform NER against the names of financial institutions.
- One serious challenge is text alignment. A significant percentage of the prospectuses employ a tabular layout to describe the (Role, Financial-entity) pairs. Both two- and four-column tabular layouts are used. This spatial alignment feature is very likely to introduce noise when matching the (Role, Financial-entity) pairs in the supply chain.

To address these challenges, we applied NLP pre-processing, used NER for the names of the financial institutions, and took advantage of text extraction tools that were tuned for tabular layouts. We then created a `proxy resMBS` dataset of (Role, Financial-entity) pairs for each prospectus. We exploited this proxy dataset for DTM topic modeling with the goal of identifying communities of (Role, Financial-entity) pairs. Unfortunately, this proxy dataset cannot be utilized with mpnet-BERT since it does not contain any sentences.

We note that some of these challenges could be overcome through training the mpnet-BERT model to identify communities. Unfortunately, we face the further limitation that the `resMBS` dataset only includes about four thousand documents, whereas there are several hundred (Role, Financial-entity) pairs that must be identified, as well as many potential communities. The limited data available for training is likely to result in overfitting.

D.2. Qualitative Evaluation Results with the mpnet-BERT Embedding

Our qualitative evaluation will make extensive use of the topics (supply chain communities) produced by the DTM topic model and the assignment of prospectuses to each topic. We note that we have manually verified the correctness of the topics and the assignment of prospectuses to topics. The verification included a manual check that each prospectus assigned to a topic indeed included a majority of the (Role, Financial-entity) pairs in the topic.

Our goal is to evaluate how well mpnet-BERT can cluster prospectuses that share a supply chain community, within the same topic, where the topic is identified by the DTM topic model. *The evaluation of mpnet-BERT rests on the hypothesis that a focal prospectus and each of its most similar neighbor prospectuses – as determined by mpnet-BERT – should have a high(er) probability of being assigned to the same DTM topic, if mpnet-BERT is correctly identifying supply chain similarity.*

The DTM Topic Model produced thirty topics. Eleven of the thirty topics were each dominated by a single issuer, e.g., `Bear Stearns` or `Wells Fargo`. We label these as homogeneous single issuer topics. Recall that prospectuses in a series are generated from an identical document template. The eleven single issuer homogeneous topics are closely associated with such document templates that can be used to determine document similarity. In other words, for a single issuer topic, document similarity can be a proxy to identify supply chain similarity. However, the other nineteen topics were not dominated by a single issuer. Typically, two or three financial institutions issued a majority of the prospectuses assigned to each topic. Multiple mortgage originators were also associated with some topics. We label these as heterogeneous topics. We use two single issuer homogeneous topics and five heterogeneous topics for the evaluation. We choose topics that each have over 100 prospectuses. The details of the topics are in Table 11.

For each topic, we consider all focal prospectuses assigned to it with a high level of confidence by the DTM Topic Model. The number of focal prospectuses for the seven topics ranges from 104 to 229 prospectuses as noted in Table

Topic ID	Prospectus Count	Type	Issuer(s)	Originators
4	176	Homo	Wells Fargo	–
9	130	Hetero	Merrill, American Home, Specialty Underwriting	No significant originator(s).
12	229	Homo	Bear Stearns	No significant originator(s).
25	191	Hetero	Maia, Goldman, Barclays	No significant originator(s).
26	104	Hetero	Ace, Deutsch, Fieldstone	Fremont, American Home
27	171	Hetero	Lehman, Struct Asset Sec	Lehman, Aurora
28	107	Hetero	Principal Residential, Citigroup, Renaissance	No significant originator(s).

Table 11 Topics Used for Evaluation.

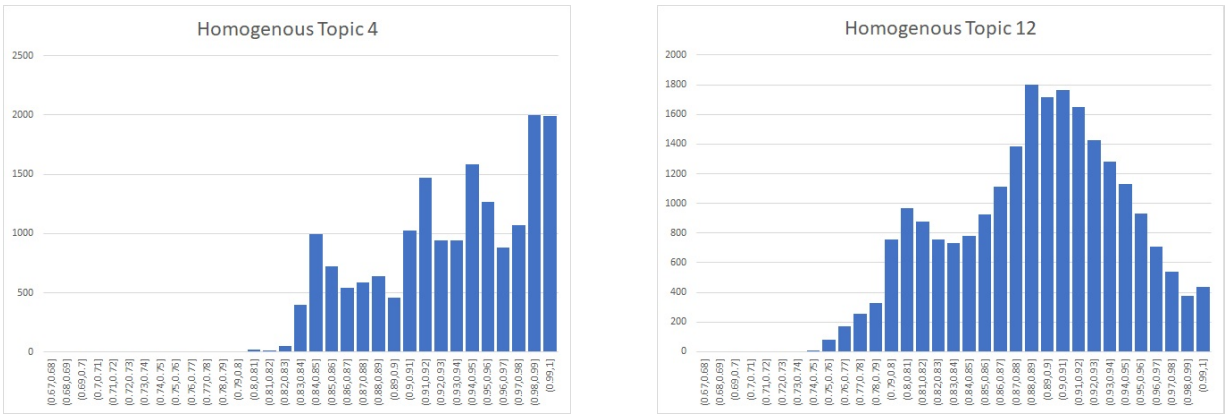
11. For each focal prospectus, we use the mpnet-BERT sentence embedding for that prospectus and compute pair-wise cosine similarity with respect to all of the other prospectuses in the corpus. We then identify the *top K* most similar (neighbor) prospectuses for each focal prospectus. For the evaluation, we choose the top 75 neighbors for the smaller topics and the top 100 neighbors for the larger topics. We note that the top 75 (top 100) represents approximately 50% of the potential neighbors for the prospectus within the small (large) topic. We therefore do not include neighbors with lower similarity scores in the evaluation.

We report on the following metrics for each focal prospectus and to its top K neighbors:

- The count of the Top K most similar neighbors with similarity scores that are > 0.90, > 0.80, etc. with respect to the focal prospectus. If the mpnet-BERT embedding was able to capture document similarity, then we would expect to observe very high pair-wise similarity scores.
- The count of the top K most similar neighbors that are assigned to the same topic as the focal prospectus. If the mpnet-BERT embedding accurately captured supply chain similarity, then the majority of the top K most similar neighbors would be assigned to the same topic as the focal prospectus.

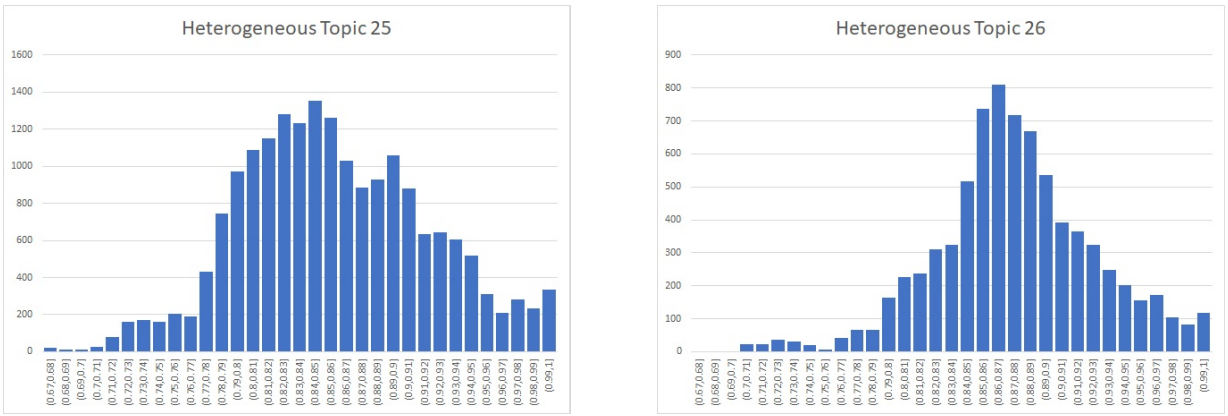
Figure 3 provides a histogram of the similarity scores of the top 100 neighbors for the prospectuses assigned to the homogeneous Topics 4 and 12; both are larger topics. The highest top 100 similarity scores occur for Topic 4; approximately two-thirds of the values are 0.9 or higher. These values reflect a high level of document similarity for this homogeneous topic that is dominated by the series of prospectuses issued by Wells Fargo. The similarity scores for the homogeneous Topic 12 are somewhat lower; approximately two-thirds of the values are 0.85 or higher. Figure 4 provides a histogram of the similarity scores of the top 100 neighbors for the prospectuses assigned to the larger heterogeneous Topic 25 and the top 75 neighbors for the smaller heterogeneous Topic 26. For Topic 25, approximately half of the scores are 0.85 or higher, but the lower half of the scores are as low as 0.7. We observe a similar distribution for the top 75 neighbors for Topic 26. This evaluation reflects that mpnet-BERT performs well on document similarity tasks and yields high pair-wise similarity scores. Its performance is better for the single issuer homogeneous topics.

Finally, Figure 5 reports on the count of the closest neighbors that occur in the same topic as the focal prospectus. Recall that our hypothesis is that a high quality mpnet-BERT embedding that successfully captures supply chain similarity will result in a high(er) percentage of the neighbors being assigned to the same topic as the focal prospectus. We note that while the previously reported similarity scores reflect excellent performance on document similarity, success in identifying supply chain similarity is only indicated by the co-occurrence of the closest neighbors in the same topic.



(a) Topic 4 (b) Topic 12

Figure 3 Histogram of pair-wise similarity scores for the top 100 neighbors for the prospectuses in Topic 4 and Topic 12.



(a) Topic 25 (b) Topic 26

Figure 4 Histogram of pair-wise similarity scores for the top 100 neighbors for the prospectuses in Topic 25 and the top 75 neighbors for Topic 26.

Recall that there is a wide variation in topic size, i.e., the count of prospectuses assigned to the topic by DTM. For simplicity, we pick a single top K and we report on the results for the top 75 neighbors. For the large topics such as Topics 12 and 25, 75 neighbors corresponds to approximately one-third of all of the potential DTM neighbors, i.e., the count of prospectuses assigned to the topic by DTM. For the smaller topics, 75 neighbors corresponds to approximately half of the potential DTM neighbors. We note that as we consider a larger top K, performance will further degrade.

In the ideal case, if mpnet-BERT was very successful at identifying supply chain similarity, then a *majority* of the top 75 neighbors of each focal prospectus would be assigned to the same topic as the focal prospectus. However, this is not the observed result; overall the counts are much lower than 75. The majority of the Top 75 neighbors are not assigned to the same topic, and this reflects that mpnet-BERT does not appear to be very successful in identifying supply chain similarity.

As we observe from the histograms of Figure 5 , there is wide variation across the topics. The best performance is for the homogeneous Topic 4. There is a large subgroup of prospectuses where 70 or more of the Top 75 neighbors co-occur in the same topic and a smaller second subgroup with moderately high counts of 50 to 60. Surprisingly, the performance for homogeneous Topic 12 is mixed. There appears to be one subgroup of prospectuses with moderately high counts of 50 to 60 of the Top 75 neighbors co-occurring in the same topic; however, this is offset by a second subgroup with low counts of 20 to 40.

As expected, mpnet-BERT performance is poor for the heterogeneous topics. The multiple issuers in these topics result in multiple document templates and lower document similarity scores. Other contributing factors that complicate the identification of supply chain similarity include the multiple issuers and originators, the lack of NER for financial entities, and the noise due to the tabular layout of (Role, Financial-entity) pairs. Yet we observe that despite all of these challenges, mpnet-BERT does provide reasonably good performance for Topic 27. The majority of prospectuses have 50 to 75 of the Top 75 neighbors co-occurring in the same topic. One potential explanation is that `Lehman`, `Aurora`, `Structured Asset`, etc. are financial entities that work(ed) closely together. This may have resulted in greater document similarity. Furthermore, there may have been a higher overlap of the (Role, Financial-entity) pairs within the supply chain community occurring across the prospectuses, simplifying the challenge of identifying supply chain similarity.

On the other hand, we find that mpnet-BERT provides very poor performance on Topic 26: the majority of prospectuses have 0 to 30 of the Top 75 neighbors co-occurring in the same topic. We note that this topic comprises multiple issuers and multiple originators; this increases the difficulty of identifying the community of (Role, Financial-entity) pairs. Topics 9 and 25 also provide a mixed/poor performance.

D.3. Summary of the mpnet-BERT Sentence Embedding Evaluation

This evaluation reflects our finding that the application of the mpnet-BERT embedding to the `resMBS` dataset is unable to identify communities of (Role, Financial-entity) pairs and to capture the underlying financial supply chain. We note that the poor performance is not an intrinsic limitation of BERT-like models. It essentially reflects a mismatch: The mpnet-BERT model was trained on domain independent document collections for a document similarity task. We then applied it to a specialized document collection like the `resMBS` dataset with the expectation that it would identify both document similarity and supply chain similarity.

We further note that this evaluation provides some insights into opportunities for future research in neuro-symbolic machine learning, where the neural learning model must additionally satisfy symbolic constraints (Hitzler and Sarker 2022). In this case, the symbolic constraints would be that prospectuses that are clustered together must exhibit both document similarity and supply chain similarity, i.e., there must be an overlap of the communities comprising (Role, Financial-entity) pairs across two similar prospectuses. We would thus train a model to predict when a pair of prospectuses exhibit high document similarity, while satisfying the constraint that both contain a specific (Role, Financial-entity) pair.

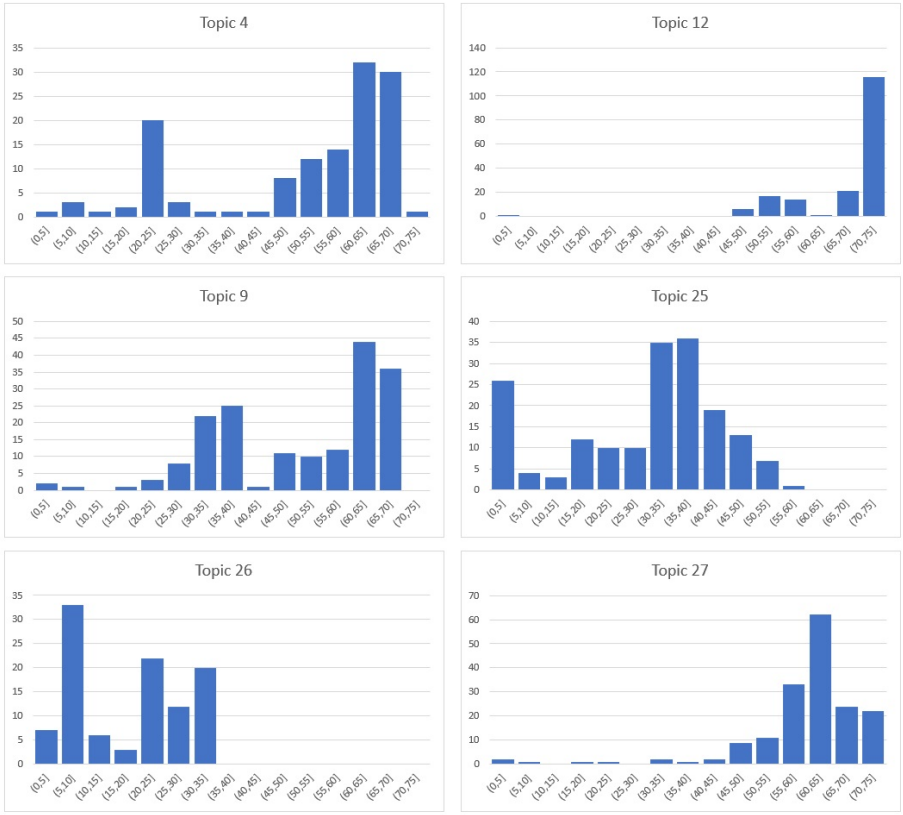


Figure 5 **Histogram of the count of top 75 neighbors in the same topic as the focal prospectus for homogeneous Topic 4 and Topic 12 and for heterogeneous Topic 9, Topic 25, Topic 26 and Topic 27.**

References

Blei D, Lafferty J (2006) Dynamic topic models. *Proceedings of the International Conference on Machine Learning (ICML)* 113–120.

Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–2003.

Bloomberg (2018) Bloomberg l.p. https://en.wikipedia.org/wiki/Bloomberg_Terminal.

Devlin J, Chang MW, Lee K, Toutanova K (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186, URL http://dx.doi.org/10.18653/v1/N19-1423.

Hitzler P, Sarker MK (2022) *Neuro-symbolic Artificial Intelligence: The State of the Art*. Frontiers in Artificial Intelligence and Applications, ISBN 9781643682440, URL https://books.google.com/books?id=jnL0zgEACAAJ.

Reimers N, Gurevych I (2019) Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, URL https://arxiv.org/abs/1908.10084.

Wang X, McCallum A (2006) Topics over time: A non-markov continuous-time model of topical trends. *Proceedings of the ACM Knowledge Discovery and Data Mining Conference* .