resMBS: Constructing a Financial Supply Chain from Prospectus

Doug Burdick
IBM
drburdic@us.ibm.com

Soham De
University of Maryland
sohamde@cs.umd.edu

Louisa Raschid
University of Maryland
louisa@umiacs.umd.edu

Mingchao Shao
University of Maryland
shao@cs.umd.edu

Zheng Xu
University of Maryland
xuzh@cs.umd.edu

Elena Zotkina
University of Maryland
ezotkina@umiacs.umd.edu

ABSTRACT

Understanding the behavior of complex financial supply chains is usually difficult due to a lack of data capturing the interactions between financial institutions (FIs) and the roles that they play in financial contracts (FCs). resMBS is an example supply chain corresponding to the US residential mortgage backed securities that were critical in the 2008 US financial crisis. In this paper, we describe the process of creating the resMBS graph dataset from financial prospectus. We use the SystemT rule-based text extraction platform to develop two tools, ORG NER and Dict NER, for named entity recognition of financial institution (FI) names. The resMBS graph comprises a set of FC nodes (each prospectus) and the corresponding FI nodes that are extracted from the prospectus. A Role-FI extractor matches a role keyword such as originator, sponsor or servicer, with FI names. We study the performance of the Role-FI extractor, and ORG NER and Dict NER, in constructing the resMBS dataset. We also present preliminary results of a clustering based analysis to identify financial communities and their evolution in the resMBS financial supply chain.

Keywords

Rule based text extraction, SystemT, named entity recognition, mortgage backed securities (MBS), clustering, community detection.

1. INTRODUCTION

There has been significant interest in understanding the behavior of complex financial eco-systems and supply chains. An example is the supply chain comprising US residential mortgage backed securities, resMBS [1, 5, 8]. This system combined with the subprime mortgage crisis to lead to the 2008 U.S. financial crisis. Unfortunately, information about the complex financial networks that comprise such a supply chain have not historically been made available to regula-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DSMM'16, June 26-July 01 2016, San Francisco, CA, USA

© 2016 ACM. ISBN 978-1-4503-4407-4/16/06...\$15.00

DOI: <http://dx.doi.org/10.1145/2951894.2951895>

Issuing Entity	Bear Stearns ALT-A Trust 2006-3.
Title of Series	Bear Stearns ALT-A Trust, Mortgage Pass-Through Certificates, Series 2006-3.
Cut-off Date	April 1, 2006.
Closing Date	On or about April 28, 2006.
Depositor	Structured Asset Mortgage Investments II Inc.
Sponsor	EMC Mortgage Corporation, an affiliate of the depositor.
Master Servicer	Wells Fargo Bank, National Association.
Servicers	EMC Mortgage Corporation and various other servicers, none of which will originate more than 10% of the mortgage loans in either loan group I or loan group II. Countrywide Home Loans Servicing LP will service all of the group III mortgage loans.
Originators	EMC Mortgage Corporation and various other originators, none of which will originate more than 10% of the mortgage loans in either loan group I or loan group II. Countrywide Home Loans, Inc. will originate all of the group III mortgage loans.
Cap Contract Provider	ABN AMRO Bank N.V.
Trustee	U.S. Bank National Association.
Securities Administrator	Wells Fargo Bank, National Association.
Distribution Dates	Distributions on the offered certificates will be made on the 25th day of each month, or, if such day is not a business day, on the next succeeding business day, beginning in May 2006.

Figure 1: Summary of a Prospectus

tors, analysts or academic researchers. Our proposition is that through text analytics, financial supply chains can be successfully extracted from prospectus that are filed with the Securities and Exchange Commission (SEC). Creating such public *financial big data* collections would then enable new streams of *data science for finance* research.

Figure 1 illustrates the summary section of a resMBS prospectus; typically such a prospectus can be over 500 pages. The summary describes key particulars of the prospectus including the following: the title of the series and the nominal value of the contract; the names of financial institutions (FI names) that play a key role in the contract; the roles, e.g., issuing entity, depositor, sponsor, servicer, originator, trustee, etc.; important dates, e.g., cut-off date, distribution dates, etc.

In this paper, we describe the process and tools to create the resMBS dataset. We make extensive use of the SystemT information extraction platform [2, 4]. SystemT provides a high-performance scalable runtime environment for executing rule-based extractors, implemented in the declar-

ative AQL language, over unstructured text. Our extraction pipeline uses two tools, ORG NER and Dict NER, developed in AQL for named entity recognition (NER), to extract FI names. ORG NER is a general purpose extractor while Dict NER is customized to use dictionaries to extract FI names.

The resMBS graph comprises a set of financial contract (FC) nodes, one for each prospectus, and the corresponding FI nodes that are extracted from the prospectus. A Role-FI extractor matches a role keyword such as originator, sponsor or servicer, with FI names. This creates a set of (Role, FI) pairs, where the FI plays the specified role in the FC. From example, from Figure 1, we can say that the (Role, FI) pair of (Servicer, EMC Mortgage Corporation) should be extracted from the prospectus and associated with the corresponding FC node.

We evaluate the effectiveness of the Role-FI extractor and Dict NER versus ORG NER, over a collection of 5000+ resMBS prospectus that were filed with the SEC between 2000 and 2008¹. We map each of the FI names to a corpus of FI names that was obtained from ABSNet². Our evaluation reveals that while ORG NER and Dict NER provide similar performance for approximately 2100+ documents, they extract different (Role, FI) pairs for the other documents. This suggests that both approaches are needed to successfully create the resMBS dataset.

As an exemplar of novel data science for finance research, we present preliminary results of clustering the resMBS graph dataset. Our analysis illustrates the emergence and evolution of communities along the resMBS supply chain. First, we apply clustering to the financial institutions based on pairwise FI-FI similarity. This allows us to observe community evolution from 2002 to 2008. We note that there are multiple strong communities in 2004, but by 2007, two financial institutions, Wells Fargo and Countrywide dominate the resMBS supply chain. We also cluster the financial institutions based on FC-FC similarity; this allows us to drill down further to observe communities at a more granular level. This drill down reveals that while Wells Fargo and Countrywide do indeed dominate in 2007, each also participates in several smaller communities. This is a first step to more sophisticated data science for finance research to understand the role played by communities on the performance of financial contracts.

2. EXTRACTION PIPELINE

Figure 2 provides an overview of the SystemT based extraction pipeline. AQL rules are used to identify the header (cover sheet) and the summary section of the prospectus. Both sections were identified by domain experts to contain relevant information. Next, FI names of participant financial institutions are extracted from the header and the summary, using Dict NER and ORG NER. The Role-FI extractor then matches roles with (adjacent) FI names to produce (Role, FI) pairs, each of which corresponds to a Role (FI, FC) labeled edge. Finally, we map the FI names to the ABSNet corpus of FI names.

2.1 Named Entity Recognition (NER)

ORG NER is a sophisticated general purpose rule-based

¹We downloaded the documents from the SEC website <http://www.sec.gov/>

²<http://www.absnet.net/ABSNet/>

Figure 2: Named Entity Recognition (NER) and Role-FI Extraction Pipeline

NER tool that is built on the SystemT platform. It has been shown to achieve state-of-the-art performance on several standard NER tasks [3]. ORG NER has multiple customization points which are exposed as user-defined dictionaries. These dictionaries allow ORG NER to be tuned to capture a variety of specialized root and suffix values across many specialized domains. Dict NER [8] was developed for the specialized task of extracting financial institution (FI) names and named entity resolution of FIs. It was designed to meet the following assumptions:

- The FI names are composed of a root fragment and a suffix. The root fragment tends to be distinct and does not show much variation. There are a limited number of suffix options.
- An (almost) complete list(s) of formal complete names for financial institutions (FI names) will be provided.
- A version of the formal FI name will appear at least once in the document so that Dict NER can extract at least one mention that will match the formal name.

2.2 Role-FI Extractor

The first task is the identification of the relevant context, i.e., sentences in the prospectus header and/or summary. We summarize the most common cases, layout, etc., as follows:

- A two column table or a list-like layout (see Figure 1) where one column includes a role (keyword) and the second has one or more mentions of FI names.
- Lists with unstructured text; this corresponds to the text describing servicers and originators in Figure 1.
- Detection of double or multi-column layouts, e.g., the two columns in Figure 3. We use heuristics, e.g., identifying two roles which occur one after the other, in the same sentence, without a comma or 'and' separating them. As can be seen in Figure 1, there are also double columns that fail this heuristic, e.g., **Depositor** followed by **Seller and Sponsor**.

While we use these cases above to refine and modify the logic behind our Role-FI extraction and matching rules, we found from experience that we had better performance with a single set of rules that cover all cases, compared to implementing separate rules for each case.

The next task is Role-FI extraction and matching. We consider the following cases:

Header section of WACHOVIA_1382260_0001068238-06-001265.htm	
Wachovia Mortgage Loan Trust, Series 2006-ALT1	
Issuing Entity	
Wachovia Mortgage Loan Trust, LLC	Wachovia Bank, National Association
Depositor	Seller and Sponsor
Wells Fargo Bank, N.A. and National City Mortgage Co.	American Mortgage Network, Inc.
Servicers	Seller
\$456,709,000 (Approximate)	
Asset-Backed Certificates, Series 2006-ALT1	
Common pairs from Dict NER and ORG NER	
(Issuing Entity, Wachovia Mortgage Loan Trust, Series 2006-ALT)	(Depositor, Wachovia Mortgage Loan Trust, LLC)
** (Seller, Wachovia Mortgage Loan Trust, LLC)	** (Sponsor, Wachovia Mortgage Loan Trust, LLC)
Not extracted	Well Fargo Bank, N. A.
Dict NER only	ORG NER only
(Sponsor, Wachovia Bank, National Association)	** (Seller, American Mortgage Network, Inc. National City Mortgage Co.)
(Seller, Wachovia Bank, National Association)	** (Servicers, American Mortgage Network, Inc. National City Mortgage Co.)
(Seller, American Mortgage Network, Inc.)	
** (Servicers, American Mortgage Network, Inc.)	
** (Seller, National City Mortgage Co.)	
(Servicers, National City Mortgage Co.)	

Figure 3: Summary of Role-FI pairs extracted by Dict NER and ORG NER.

- The base case is where the role keyword and the mention of the FI name are in proximity, in the same sentence, with a 'FI as role' structure. An example is the following sentence:
Lehman Brothers Inc., as underwriter, will purchase from Lehman ABS Corporation, as depositor, ...
A variation of this case is where phrases such as will sell, will service replace the role keyword.
- The next case is where an organization may play multiple roles; this may result in the following sentence structure: '*fi-1 role-1 role-2*'.
An example is the following:
Accredited Home Lenders, Inc. (Sponsor and Servicer) ...
Variations include the case where the roles precede the FI name mention.
- An outlier case is where the FI name mention is not in the same sentence as the role keyword. To address this case, we extend the matching so that adjacent FI name mentions and roles will be matched, despite not being located in the same sentence. As expected, heuristics for such complex cases can be difficult to develop and may not be always successful.

2.3 Sample Role-FI Extraction

Figure 3 provides an example of the (Role, FI) pairs extracted by Dict NER and ORG NER. The top section of the figure displays the sample text from the prospectus header. The middle section shows (Role, FI) pairs that were successfully extracted by both NERs, as well as the (Role, FI) pairs that both missed. The bottom section of the image shows the (Role, FI) pairs that are extracted only by Dict NER (on the left) and those pairs extracted only by ORG NER (on the right). Incorrect (Role, FI) pairs are in red and preceded with **.

While the general purpose ORG NER performs well on most documents, it has difficulty with a multi-column layout. Dict NER can handle such structures and is better at detecting the boundaries of entity mentions. For example, ORG NER incorrectly extracted **American Mortgage Network, Inc. National City Mortgage Co.**, while Dict NER correctly extracted two separate FI names. Wells Fargo

	Dict NER	ORG NER
Failed to extract	17	43
Count (Role, FI) pairs	53235	55040
Duplicates	15029	14369
Distinct count	38206	40671
Distinct count (matching ABSNet)	32942	32826
Common (Role, FI) pairs	29472	29472
Excess (Role, FI) pairs	3470	3354

Table 1: Summary statistics for extracted (Role, FI) pairs for ORG NER and Dict NER.

as a servicer was not extracted by both approaches. Dict NER extracted National City Mortgage Co., but it was wrongly labeled as both seller and servicer, instead of only servicer.

3. EFFECTIVENESS OF EXTRACTION

We evaluate the effectiveness of the Role-FI extractor, and Dict NER versus ORG NER, over the collection of 5000+ resMBS prospectus. For this evaluation, we map each of the FI names to the corpus of FI names from ABSNet. We make the following assumptions:

- An FI name that does not match an entry from the corpus represents an incorrect (Role, FI) pair. This is reasonable since ABSNet maintains a fairly comprehensive collection of asset backed products.
- We eliminate duplicates; they are usually extracted from both the header and the summary.
- When an FI name from the (Role, FI) pair matches an entry in the ABSNet corpus correctly, then this is a valid (Role, FI) pair.

This last assumption is the most difficult to validate since our experience, as illustrated in Figure 3 is that there can be errors in the extraction process. In ongoing research we are performing an extensive manual evaluation of (Role, FI) pairs over randomly sampled documents, to estimate the actual proportion of valid (Role, FI) pairs that are extracted.

Next, for each document, we examine the (Role, FI) pairs extracted by ORG NER and Dict NER. Table 1 provides summary statistics of the (Role, FI) pairs for ORG NER and Dict NER, respectively. We see that the extraction failed for a small number of documents. After eliminating duplicates, we retained only those (Role, FI) pairs that match FI names from ABSNet. We then determine the (Role, FI) pairs that were common to both ORG NER and Dict NER and those *excess* pairs that were in extracted by each tool.

Table 2 compares the extraction statistics of the excess (Role, FI) pairs, for ORG NER and Dict NER, respectively. As can be seen, ORG NER and Dict NER performance is identical, and there is no excess of (Role, FI) pairs, for 2100+ documents. ORG NER has an excess ranging from 1, up to 5 or more (Role, FI) pairs, for approximately 1000+ documents. Similarly, Dict NER has a similar excess of (Role, FI) pairs, for approximately 1400+ documents.

What is really noteworthy is that there are also documents in which *both* ORG NER and Dict NER produce an excess of (Role, FI) pairs. For example, for 180 documents, ORG NER excess is 1 and Dict NER excess is 1. For a further 100+ documents, ORG NER excess ranges from 2 to 4+,

ORG NER Excess	0	1	2	3	4	5+
Dict NER Excess	0	0	0	0	0	0
Count	2164	429	179	89	81	126

ORG NER Excess	0	0	0	0	0	0
Dict NER Excess	0	1	2	3	4	5+
Count	2164	885	271	144	73	35

ORG NER Excess	1	2	3	4+
Dict NER Excess	1	1	1	1
Count	180	48	20	30

ORG NER Excess	1	1	1	1
Dict NER Excess	1	2	3	4+
Count	180	117	39	20

Table 2: Distribution of excess Role-FI pairs, for ORG NER and Dict NER, across the 5000+ financial contracts (FC).

Year	2002	2003	2004	2005
FC Count	475	666	969	1184
FI-FC edges	4037	5900	8367	10447
\$ Value (10^9)	273	414	723	1019

Year	2006	2007	2008
FC Count	1051	680	71
FI-FC edges	13922	9300	1026
\$ Value (10^9)	895	510	16

Table 3: Summary statistics for the financial contracts (FC).

while Dict NER excess remains at 1. Similarly, for a further 180+ documents, Dict NER excess ranges from 2 to 4+, while ORG NER excess remains at 1.

To summarize, the performance of ORG NER and Dict NER is identical in 2100+ documents. In approximately 2400+ documents, either ORG NER or Dict NER produces an excess of 1 to 5+ (Role, FI) pairs. Finally, in about 500+ documents, both ORG NER and Dict NER produce an excess of (Role, FI) pairs, for the same document. This suggests that both approaches are needed to successfully create the resMBS dataset.

4. ANALYTICS PIPELINE AND PRELIMINARY RESULTS

Table 3 provides summary statistics for the resMBS graph dataset for each year, from 2002 through 2008. As can be seen, 2002 through 2003 is an emerging period. The number of contracts reach a peak in 2005 and then decline while the supply chain dries up with the 2008 crisis.

Figure 4 provides an overview of the analysis steps. The first step is data cleaning. This will remove some FC and FIs with degree less than 3 and 20, respectively. We further filtered the dataset to only include the following four roles: Issuer, Sponsor, Originator and Servicer. While this significantly reduced the size of the graph, it also allowed us to focus on more important roles and interactions.

We then compute the pairwise FI-FI and FC-FC similarity. Both FI-FI and FC-FC similarity are based on co-occurrence. FI-FI similarity is proportional to the count of documents in which the pair of FIs co-occur. FC-FC simi-

Figure 4: Analysis of the Role(FI, FC) resMBS Graph

larity is proportional to the count of FIs that co-occur in the pair of FCs. The similarity is normalized to between 0 and 1. We then performed three types of clustering as follows:

- **FI-FI similarity clustering:** This is a direct clustering of the financial institutions based on FI-FI similarity. We use the normalized cut (NCut) method [7]. This clustering will reveal dominant associations among financial institutions.
- **FC-FC similarity:** This is an indirect clustering based on FC-FC similarity using the NCut method. We then use the (FI, FC) labeled edges associated with each FC to produce clusters of FIs.
- **(FI-FC) edge partition using semEP** [6]: Unlike node based partitioning, this is a partitioning of the edges. The resulting clusters will provide granular insights into the key relationships among financial institutions across groups of contracts.

4.1 Observations from Clustering based on FI-FI Similarity

Figures 5 and 6 present the output of clustering FIs based on FI-FI similarity, for 2004 and 2007, respectively. We selected 2004 since we consider 2002 and 2003 as an emerging period, where the number of financial contracts increased rapidly and partnerships were being forged among the FIs. The number of contracts peaked in 2005 and declined in 2006 and 2007. 2007 is considered the terminal period since the number of financial contracts rapidly decreased with the onset of the financial crisis in 2008. We chose the number of clusters to be eight for ease of visualization.

In 2004, we observe that there are three major communities (purple, light green and light blue), three moderate communities (orange, dark green and light orange), and two small communities (red and pink) in Figure 5. Among the major communities, Wells Fargo dominates one (light green) and GreenPoint is a major partner. Another community (light blue) is dominated by Countrywide with Bear Stearns as a partner. The third (purple) community is dominated by Chase Manhattan with Candant as a major partner. The three moderate communities are dominated by Aurora (light orange), GMAC (orange) and National City and Goldman

Figure 5: Direct clustering of the FIs from 2004 into eight clusters using NCut.

Sachs and IndyMac (dark green). Finally, among the two small communities, Lehman is prominent in one (red) and Morgan Stanley in the other (pink).

We observe a very different pattern in 2007 in Figure 6. For example, there is one large community (light green), 2 moderate communities (purple and light blue), and multiple smaller communities. The large (light green) community is dominated by Countrywide and Wells Fargo. Among the moderate communities, one (purple) includes GMAC, Deutsche Bank and American Home Mortgage, while the second (light blue) includes GreenPoint, Aurora, Lehman and IndyMac.

4.2 Clustering based on FC-FC Similarity

As was just discussed, clustering FIs based on FI-FI similarity does not provide insight at a deeper level about smaller and more connected communities. As an alternative, we can use FC-FC similarity to cluster financial contracts, and then induce communities of FIs that are connected to those contracts. Figure 8 presents communities that were obtained following this approach.

As can be seen, Figure 8 reveals many smaller communities, compared to Figure 6, for the same 2007 dataset. We note that the direct clustering of financial institutions of Figure 6 restricted each financial institution to one community. In contrast, the FC-FC based clustering of Figure 8 does not have that restriction.

Consider that Countrywide and Wells Fargo dominated the largest community in Figure 6. We can now observe in Figure 8 that Countrywide participates in three communities, with Wells Fargo (light green), with GMAC and Deutsche Bank (dark red) and with Morgan Stanley and Saxon (dark orange). Similarly, Wells Fargo also participates in a (pink) community with Bank of America and Washington Mutual and Sun Trust. The (purple) community of GMAC, Deutsche Bank and American Home Mortgage of Figure 6 is retained in Figure 8, with the addition of Wells Fargo into that community. Finally, the (light blue) community of Figure 6 is further divided; Aurora and

Figure 6: Direct clustering of the FIs from 2007 into eight clusters using NCut.

Lehman now participate in a (light orange) community in Figure 8.

4.3 Edge Partitioning

Finally, we can apply edge partitioning to the Role (FI, FC) bipartite graph to further drill down to community details, i.e., the role played by the FIs. Figure 7 illustrates a community of eight FCs where SACO is the issuer (green edge) and where multiple other FIs are the originators (pink edges). We observe that all eight contracts in this community are also associated with Bear Stearns & Co. and its affiliate, EMC Mortgage.

5. SUMMARY

We describe the use of text analytics and the SystemT platform to successfully create the resMBS financial big data dataset, corresponding to the MBS supply chain. Both ORG NER and Dict NER are needed for good performance during the extraction for FI names. Preliminary analysis using a variety of clustering methods illuminate the emergence and evolution of financial communities along this supply chain. This is a first step to more sophisticated research to understand the role played by communities.

We now discuss ongoing and future research. Inspired by the Latent Dirichlet Allocation (LDA) and topic models, we have developed two probabilistic financial community models [9]. Our models are based on an intuitive assumption that FIs will form communities within an FC, and FIs within a community are more likely to collaborate with other FIs in that community, and play the same role, in another FC. Results from [9] indicate that these communities are indeed created, and that they can be used to describe and characterize the resMBS financial supply chain. In future research, we will explore if specific (Role, FI) pairs are correlated with the performance of the securities in the financial contracts. Alternatively, network characteristics of the resMBS graph may reveal issues related to contagion or stability under var-

Figure 8: Indirect clustering of the FIs from 2007 into eight clusters using NCut.

Figure 7: FI FC Bipartite graph; this illustrates a community of FCs with a single issuer SACO (green edges) and multiple originators (pink edges).

ious stress conditions.

6. ACKNOWLEDGMENTS

The authors would like to thank Nancy Wallace and Paulo Issler from the Haas School of Business, UC Berkeley, and Rajasekar Krishnamurthy and Howard Ho from IBM Research, for many discussions and insightful feedback. This research was partially supported by awards NSF CNS1305368, NIST 70NANB15H194 and the Smith School of Business.

7. REFERENCES

- [1] D. Burdick, M. Franklin, P. Issler, R. Krishnamurthy, L. Popa, L. Raschid, R. Stanton, and N. Wallace. Data science challenges in real estate asset and capital markets. In *Proceedings of the International Workshop on Data Science for Macro-Modeling*, pages 1–5. ACM, 2014.
- [2] D. Burdick, M. A. Hernández, H. Ho, G. Koutrika, R. Krishnamurthy, L. Popa, I. Stanoi, S. Vaithyanathan, and S. R. Das. Extracting, linking and integrating data from public sources: A financial case study. *IEEE Data Eng. Bull.*, 34(3):60–67, 2011.
- [3] L. Chiticariu, R. Krishnamurthy, Y. Li, F. Reiss, and S. Vaithyanathan. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1002–1012. Association for Computational Linguistics, 2010.
- [4] L. Chiticariu, Y. Li, and F. R. Reiss. Rule-based information extraction is dead! long live rule-based information extraction systems! In *EMNLP*, pages 827–832, 2013.
- [5] J. Hunt, R. Stanton, and N. Wallace. U.s. residential mortgage transfer systems: A data management crisis. In M. Brose, M. Flood, D. Krishna, and B. Nichols, editors, *Handbook of Financial Data and Risk Information II: Software and Data*, 2014.
- [6] G. Palma, M. Vidal, and L. Raschid. Drug-target interaction prediction using semantic similarity and edge partitioning. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference Proceedings, Part I*, pages 131–146, 2014.
- [7] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.
- [8] Z. Xu, D. Burdick, and L. Raschid. Exploiting lists of names for named entity identification of financial institutions from unstructured documents. *arXiv preprint arXiv:1602.04427*, 2016.
- [9] Z. Xu and L. Raschid. Probabilistic financial community models with latent dirichlet allocation for financial supply chains. In *SIGMOD DSMM*. ACM, 2016.