Estimating Latent Structure of Acquisitions from Text

Jiayi Li [*]
Tepper School of Business
Carnegie Mellon University
Pittsburgh, PA 15213
jiayil@andrew.cmu.edu

Abstract

This paper uses machine learning tools to analyze press releases announcing company mergers and acquisitions. We find that acquisitions can be clustered into different groups based on their text similarities. The result of this paper shows that existing tools on document clustering can work on structurally highly similar documents, and benefit economic researchers in understanding mergers and acquisitions.

1 Introduction

1.1 Background

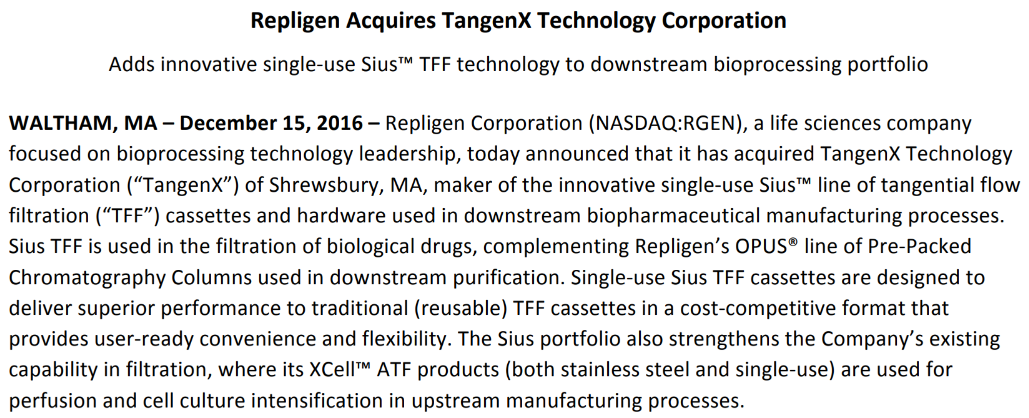
Researchers in economics, finance, and managerial science have long been interested in various aspects of mergers and acquisitions (M&A), including their reasons, patterns, and aftermath [7, 8, 17]. To have a better understanding of these aspects empirically, it is necessary to look into how merger pairs are formed, and in what way the characteristics of the buyers and targets are correlated. Since firms are not transparent by nature, it is not obvious what data we should use in our statistical analysis to answer each related question. Various empirical studies has been conducted using hard data about the merger pairs in known acquisitions [11, 12, 14]. Researchers have also utilized information from text to analyze different aspects of these transactions [16]. However, less attention has been paid to information disclosed by the firms about their own transactions [2]. This paper study the press releases announcing mergers and acquisitions.

Among all M&A transactions, the ones with particular interest are those with "technology synergies". Most typically, one party in such a merger owns a technology that can benefit the other's business, the combination of the two can then create extra value compared with when they are separate firms. The relationship between these takeover activities and the innovation activities within firms thus become especially intriguing [5, 9, 15, 20]. Here we ask if firms specifically reveal their consideration of "technology synergies" in the announcement press releases, and we seek to provide such information to benefit the studies about innovation activities.

Figure 1 shows part of a typical press release of a "tech-oriented" acquisition[2]. Besides the information about the buyer firm and the target firm (their industry, size, location, etc.), a typical press release also describes when and how the transaction takes place (cash or share exchange), and how this transaction could affect both firms in the future. Interestingly, different documents spend different efforts in describing each of the elements above. For example, one document could focus on how firm B provides technology solution to firm A, or how the combined firm has a greater share of the

[*] I thank Byran Routledge and Alan Montgomery for their advice on this project.

[2] See https://www.repligen.com/files/2014/8176/4231/RGEN_TangenX_PR__15Dec16.pdf for the full text of this document.



Repligen Acquires TangenX Technology Corporation

Adds innovative single-use Sius™ TFF technology to downstream bioprocessing portfolio

WALTHAM, MA – December 15, 2016 – Repligen Corporation (NASDAQ:RGEN), a life sciences company focused on bioprocessing technology leadership, today announced that it has acquired TangenX Technology Corporation ("TangenX") of Shrewsbury, MA, maker of the innovative single-use Sius™ line of tangential flow filtration ("TFF") cassettes and hardware used in downstream biopharmaceutical manufacturing processes. Sius TFF is used in the filtration of biological drugs, complementing Repligen's OPUS® line of Pre-Packed Chromatography Columns used in downstream purification. Single-use Sius TFF cassettes are designed to deliver superior performance to traditional (reusable) TFF cassettes in a cost-competitive format that provides user-ready convenience and flexibility. The Sius portfolio also strengthens the Company's existing capability in filtration, where its XCell™ ATF products (both stainless steel and single-use) are used for perfusion and cell culture intensification in upstream manufacturing processes.

Figure 1: An example of acquisition press release[2]

market. Meanwhile another document could focus on describing how many shares of firm A stock that firm B shareholders are going to receive per share of firm B stock they own, which is financially important. Despite possible strategic disclosure by firms to fool the market, here we assume the most emphasized aspect in the corresponding press release reflects the most important feature of a merger. As a consequence, intrinsically similar transactions will generate similar documents, which motivates us to study the similarities between press releases in the hope that it could lead us to some latent similarities between acquisitions. Specifically, the goal of this project is to apply document clustering in such way that it aligns with the cluster of acquisitions to the largest extent.

1.2 Approach

There are various existing ways of document clustering. [1] provides a comprehensive survey within this domain. While there are direct applications of general clustering algorithms [18, 22, 25], there are also algorithms that borrow from language modeling [4, 23]. Essentially all clustering algorithms seek to largely reduce the dimension of the feature vector representing a document, during which process interpretability will be inevitably lost, to different degrees. On the other hand, topic modeling [6], as well as other generative processes with latent variables, is known to be able to produce interesting qualitative and interpretable results. [24] integrated standard topic modeling with document clustering, where the interpretability is largely preserved, and the flexibility of the model is increased.

This paper applies two document clustering methods, both featuring topic modeling, on the target dataset. Firstly, we set LDA+Kmeans as the benchmark approach. Under this approach, naive LDA is first applied on the documents as in [6], and each document is represented as a distribution over the generated topics. Then we apply k-means++ algorithm [3] to cluster the documents. The second approach is the application of Multi-Grain Clustering Topic Model (MGCTM) in [24]. We compare the economic interpretability of the clusters generated by the two approach, since there is no ex-ante good criterion of evaluating a cluster assignment. More details of the second approach as well as deeper economic motivation will be discussed in Section 3.

2 Data Description

Transaction records are obtained from S&P Capital IQ[3]. In this paper, we restrict to M&A transactions announced between 2000-01-01 and 2016-12-31, where the buyer firm is a publicly traded firm in the United States. This dataset is hand-collected and contains identifying information of buyers and targets, and other features of the transactions, with possible errors, missing values, and even missing observations. Among the subset of 31,741 transactions, we try to locate the announcement press release of each of them through scraping the SEC Edgar database[4]. A match occurs if a within

[3] `https://www.spglobal.com/marketintelligence/en/solutions/transactions`

[4] `https://www.sec.gov/edgar.shtml`

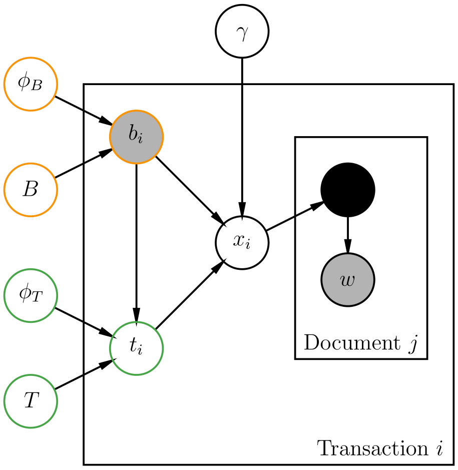


Figure 2: Generative model of transactions

the five-day window of the first announcement date[5] of a transaction, the buyer/target firm files an 8-K report[6] that mentions the name of the target/buyer firm, and attaches a press release as one of the exhibits. We then separate the exhibits containing the press releases and parse them into pure text documents. This part is done by imperfect hard coding and brings noise into the dataset. A public firm is only obligated to disclose its "significant" acquisitions through 8-K, which means the majority of the transactions are disclosed through other channels, or not disclosed at all. Following research with access to more complete or clean dataset from other sources are welcome to replicate the experiments in this paper.

The above data collection process gives us 6,237 press releases, each announces a transaction. The average length of a document is 1149 words. The raw dataset contains 84,475 unique words in the vocabulary. To reduce the effect of firm names on clustering, we replace firm names (full names and possible nicknames and acronyms) with the strings "buyerfirm" and "targetfirm" respectively. This ensures that we are focusing on the difference in content between documents. After removing stopwords, very common and uncommon words, we are left with a dictionary with 2,844 unique words. The average length of a document is now reduced to 495 words. Each document is then represented as a bag-of-words(bow) vector.

Buyer and target hard information are obtained from Compustat Fundamentals North America[7]. It contains columnized information from the 10-Q and 10-K reports of firms that file these reports. This dataset, again, suffers from missing values and missing observations. Since all buyers are public firms, and presumably big enough to make acquisitions, almost all of them can be found in the Compustat dataset. On the other hand, only around 10% of the targets can be found in the dataset, which are the public firms or relatively big private ones. We collect the available quarterly information about the buyers and targets in the text dataset. Specifically, for each transaction, we collect data from two years before the announcement of the transaction to two years after the announcement.

3 Model

3.1 Model of transactions

We formalize the problem in 1.1 into the following generative process of acquisitions in Figure 2.

[5] Some transanctions are first announced when the definitive agreements are signed, while others are first announced upon closure of the transactions.

[6] `https://en.wikipedia.org/wiki/Form_8-K`

[7] `https://www.spglobal.com/marketintelligence/en/?product=compustat-research-insight`

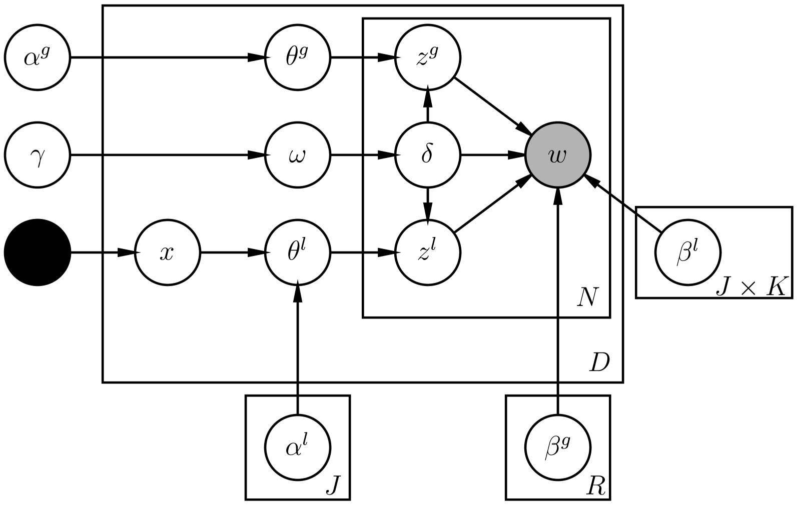


Figure 3: Generative model of documents

B is the pool of all potential buyers. Each time a transaction happens, a buyer b_i is drawn from B according to the distribution ϕ_B. Similarly, a target t_i is drawn from the pool of targets T by a distribution conditioned on the buyer's characteristics. In other words, buyers are choosing targets to some degree. The buyer and the target then jointly determine the type of this transaction $x_i \in 1, 2, \cdots, K$. γ is the functional parameter that maps b_i and t_i into x_i. Each type of transactions has a black box of document generator. Note that although b_i is marked as "observed" in the figure, the data is actually incomplete for some of the data, as is illustrated in Section 2. On the other hand, although t_i is marked as "unobserved", we do know very limited things such as names of the targets. And for some of the targets, we have as much information as the buyers.

The question that interest economists is the arrow pointing from b_i to t_i in Figure 2. We are curious how M&A pairs are formed, how buyer features and target features within a pair is correlated. This problem is noisy, since we know very little about the targets, which is reflected by the graph. By inducing a discrete type x_i and assuming that there are finitely many types of M&A transactions, we have reduced the complexity of the model relative to the limited data we have, which makes solving the problem more hopeful. Nevertheless, the type variable x_i is unobserved as well. This paper then attempts to extract the information about the x_i's from the text data. Although one way to pursue this is to combine the two processes together (i.e. Figure 2 and 3 which is to be explained) and estimate the gigantic latent structure of acquisitions and its press releases utilizing available data of the buyers and the documents, we take a step back and separate the problem into two halves. This project only focuses on the second half, document clustering, and is leaving the first half to future economic research. However, one particular type of mergers, the "tech oriented" ones, especially interests the author. Therefore an important criterion in evaluating a clustering method is whether it provides insights on the measurement of how tech-oriented each transaction is, or whether it is tech-oriented at all.

3.2 Model of documents

The benchmark approach is Kmeans, and it is not a generative model.This approach can be itself viewed as a two-step estimation, where in the first step LDA topics are estimated, and in the second step Kmeans is applied on the document-topic vectors. It is worth comparing this approach with a "full" or one-step method, where topic modeling and clustering are trained together, and can interact with each other. For this purpose, we borrow the MGCTM model from [24], and the document generating process is represented in Figure 3.

A type $x \in 1, 2, \cdots J$ is drawn from some unknown distribution, which can be thought of as the economic model in Figure 2, and can be simplified into a multinomial with parameter π. Each type has K local topics, and a dirichlet prior α^l_j over the topics. Given the type x and the prior of this type, a multinomial θ^l over the K local topics is drawn. At the same time, all types share R global topics, which has a dirichlet prior α^g. A multinomial θ^g is drawn accordingly. Further, a Bernoulli parameter ω is drawn from a Beta distribution with parameter γ. x, θ^l,ω,θ^g are document-specific parameters. To generate a word in such a document, a binary indicator δ is drawn from $Bernoulli(\omega)$. If $\delta = 1$, this word will be generated from a local topic. A topic z^l is drawn from $Multinomial(\theta^l)$. Each topic is described by a multinomial $\beta^l_{j,k}$, a word is then drawn from this distribution. Similarly, if $\delta = 0$, this word will be generated from a global topic z^g based on the distribution β^g.

The above model differs in the benchmark approach Kmeans in the following ways. Firstly, the proportion of "global" words in a document is independent with its type. This kind of flexibility might be a solution to the noises present in the documents. On the other hand, the model is limiting each type to its own local topics, which might lead to overall loss in fit. Taken together, the comparison of the performances of this model and the naive LDA+Kmeans can help us better understand the structure of our documents.

3.3 Challenges

It is challenging to cluster documents specific to this dataset and this task, where interpretability is the greater concern, instead of standard metrics for evaluating clustering. The clusters that best fit our purpose might not be the ones with the highest score. The following specifics demonstrate why this could particularly be a concern for press releases.

In general, the composition of all documents disclosed by public firms, not restricted to press releases, is decided by the joint force of mandatory disclosure regulations and voluntary disclosure incentives. The mandatory part results from SEC regulations ensuring firms do not lie, mislead investors, or hide important facts. The voluntary part refers to firms' flexibility of adding content to the documents that investors could potentially be interested in. Note that different firms conditional on their status are subject to different levels of regulatory restrictions, and have different voluntary disclosure incentives. As a result, documents can differ in their proportion of mandatory and voluntary information. The following two extreme situations demonstrate why naive document clustering might not be successful.

- Transactions A and B have similar economic features, which both of them disclose in their documents. Due to regulatory requirements, half of A's document is safe harbor statements[8], while B's document has no safe harbor statements. Acquisition C has completely different economic features than A and B, while half of its document is also safe harbor statements.

- Transactions A and B have similar economic features. Despite both documents have disclosed these features and facts, A's document is strategically extended (legally) and includes a CEO interview about how beneficial the transaction is, while B's document misses this section. Acquisition C has completely different economic features than A and B, yet it has a "bragging" section just like A.

To handle these cases, an ideal algorithm in this task should have the flexibility in discarding "unimportant" features when measuring similarities between documents. The intuition is similar to removing stopwords in standard text processing, however we do not have the set of stopwords here, and do not have enough data to figure it out. In fact, if we had a large enough dataset, it would be interesting to experiment existing models with attention gates [10, 13] on the documents. On the other hand, it is as well likely that all the differences are caused by omitted fundamental economic features (e.g. bigger firms have longer safe harbor statements[9]), in which case the most distinct feature in a document will reflect, maybe implicitly, the most distinctive economic feature of a transaction. This is especially a concern since there are fundamental features about the firms that are unobserved by nature, among which are "innovation ability", "organization efficiency", and "long term profitability", etc. Therefore, we should also remain cautious in relevance judgment.

[8] See `https://en.wikipedia.org/wiki/Forward-looking_statement` for more information about safe harbor statements.

[9] Just an example, not necessarily true.

4 Experiments

4.1 LDA+Kmeans

We first run naive LDA[10] with 35[11] topics on the documents. We show the 10 words with the highest probabilities in 10 of the converged topics in Table 1. With the converged topics, we represent each document as a vector of distribution over topics. We then apply Kmeans[12] to assign the documents in to 5 clusters. The final "centers" of the clusters can be seen as 5 different vectors over topics. To visualize the cluster centers, we then collapse each of the center into a distribution over the vocabulary, which resembles a "topic". Figure 4 shows the weighted word cloud of the 5 cluster centers.

Based on the keywords and a little economics background, we recognize that cluster 0 is about "technology solutions", cluster 1 is about "income", cluster 2 is about "banks", cluster 3 is about "shareholders" and "proxy statements" (labeled as "stock"). However, cluster 4 is relatively ambiguous to label based on the top words only. The clustering assignment of each document-topic distribution vector is shown by the 2 dimensional t-sne graph in Figure 5. We observe that the most popular "unnamed" group, which refers to cluster 4, are mainly in the middle of all documents, while the other four clusters are different distinct corners of the distribution. Next we try to validate this economically using the outside hard information.

Firstly, we notice that cluster 1 mentions earnings related terms more frequently. The structural reason behind this can be found by the classification of their 8-K reports. 79% of the transactions within this cluster are disclosed under the item "Results of Operations and Financial Condition", i.e. firms mention their recent acquisitions when announcing their earnings of the current quarter. In other words, the press releases that we collect about these transaction are mainly talking about earnings and income, instead of the acquisitions. Firms choose to disclose acquisitions in this way when they don't feel necessary, and are not obliged, to make disclosures about the acquisitions individually. These acquisitions are usually not significant, and within the usual business operations of the firms. In other words, theses acquisitions are small, or less important relative to the size of the buyer. As a comparison, less than 5% of the documents are under the item "Results of Operations and Financial Condition" in other four clusters. In fact, the majority of our documents are under the item "Entry into a Material Definitive Agreement".

Next, by looking at the top words of cluster 2, we infer that these transactions are bank mergers. In fact, around 87% of the buyers in cluster 3 have SIC codes within "6020", "6021", and "6022", which refer to "Commercial Banks", "National Commercial Banks", and "State Commercial Banks". The remaining buyers are mostly within SIC codes "6035" and "6036", which refer to "Federal savings institutions" and "Savings institutions, except federal".

We now try to identify the economic features of cluster 3. One clue that we have is that documents in this cluster mention "proxy statement" and "stockholders" more often than other documents. When a potential buyer offers to acquire a target, the shareholders have to vote on whether to accept this offer. A proxy statement is required for soliciting the votes. Other top words in this cluster also indicate the discussion of the transaction details and paperwork, instead of the fundamentals of the companies. Naturally, the transaction process gets more serious and worth talking about when the merger is larger. A quick way to check this is to see how many of these targets can be found in the Compustat dataset (recall that most targets are not in this dataset). Figure 6a shows the proportion of targets in Compustat for each cluster. We can see that less than 10% of the targets in the other 3 clusters are in Compustat, while bank targets have a higher chance. Targets in cluster 3 have a much larger probability of being in Compustat. It is also worth checking whether buyers in this cluster are larger than other buyers. We compare their average sales two years before the merger in Figure 6b. Apparently buyers in this cluster is also larger than other buyers in terms of sales. Firm size might not be the most direct variable linking to this cluster, but the above validation at least shows there is some correlation between this cluster and the economic features of firms.

Documents in cluster 0 mention technology related vocabulary more frequently. To validate this, we check whether the buyer firms of these transactions are more active in R&D activities than other buyers. Figure 7a shows the proportion of firms active in R&D (i.e. firms that report R&D expense

[10] We use the ldamodel package in gensim [21].

[11] The number of topics chosen here is to match the model complexity in MGCTM.

[12] We use the kmeans cluster package in sklearn [19].

within the two year window prior to the transaction) in each cluster. In the bank cluster, almost none of the buyers conduct any R&D, which is consistent with our knowledge. In cluster 0, over 80% of the buyers are active in R&D activities. In the remaining 3 clusters, around 50% of the firms are active. Next we compare in Figure 7b the average R&D intensity[13] among R&D active firms while filtering out the firms with R&D intensity above 1. We can see that the average R&D intensity of cluster 0 is still higher than other clusters.

Finally, we admit there is probably nothing special about cluster 4. Unlike transactions in other clusters, these mergers do not have a distinct economic, nor textual feature, which is why they tend to lie in the middle of all transactions.

4.2 MGCTM

As a contrast, We follow [24] and estimate the model using variational inference and EM algorithm. The model parameters are π,α^l,α^g,γ,β^l,β^g in Figure 3. In the E-step, we fix the model parameters and update the latent variables x,ω,θ^l,θ^g,δ,z^l,z^g. In the M-step, we update the model parameters by maximizing the lower bound of the likelihood taking the converged latent variables as given in E-step. The two steps are iterated until convergence. we set the number of groups to be 5, number of global topics to be 10, and number of local topics of each group to be 5.

The final weighted global topics and local topics of each group is represented as word clouds in Figure 8 and 9. This time the clusters are not as interpretable as in LDA+Kmeans, although we do see some overlaps. For example, group 0 resembles both center 2 and 3, which are about banks and proxy statements. Group 1 resembles center 1, both of which are about income and earnings. To better compare the two clustering assignment, we plot the group assignment in the same t-sne graph in Figure 10. This approach is not able to separate the "technology" cluster from the average group. Several factors might have caused the underperformance of the MGCTM approach. Firstly, we have a relatively small dataset. Secondly, the model might not be a very good characterization of the data. Technically speaking, this could mean the objective function is relatively flat around the optimal solution or have local maxima. Thirdly, there is the possibility that all hyperparameters are not properly tuned.

The comparison of the results from the two approaches has given us some insights. Firstly, we have understood that transactions disclosed in earnings releases should not be treated the same way as other M&A press releases. This might be a trivial finding ex-post, but it is still a relief to see that the data confirms our expectation. Similarly, we have learned that bank mergers have very different press releases than others, which is highly reasonable. On the other hand, we also do not see other industry-specific clusters at this level (may be they will reveal as we increase the number of clusters), which might partly explain why financial economists study bank mergers separately from regular M&As. Secondly, an interesting result is the formation of the cluster featuring "proxy statements" and "shareholders". A preliminary validation shows these transactions happen between bigger firms. One potential explanation would be that as the two parties in a merger become larger, the negotiation process gets more complicated, and more important to investors and shareholders. In other words, the details of the agreement start to dominate other economic fundamentals of the merger, and become the most distinguishable feature of a transaction, as well as its press release. Finally, as an economist particularly interested in tech-mergers, one lesson to draw is that we should probably restrict our attention to non-bank and smaller transactions disclosed in a press release discussing the mergers solely. The result from MGCTM shows that, instead of considering tech-mergers as generated by a different process than other M&As, we should consider the how tech-oriented a transaction is as a quantitative matter, although it is not entirely clear what the "tech" cluster is capturing in our LDA+Kmeans result.

5 Conclusion

This paper applies document clustering methods on press releases announcing M&A deals. We are able to obtain sensible clusters out of the LDA+Kmeans approach, which could benefit future research studying the structures of M&As. Specifically, we find that some mergers are disclosed in earnings releases, which are very different from other acquisition announcement press releases. We

[13] R&D expense over Sales

also find that bank mergers have identifiable press releases among mergers in all industries. Some mergers emphasize the negotiation process and legal process of the merger in their press releases, one characteristic of which is that the firms involved are typically larger. Finally, we find that some information about the technology considerations of mergers is revealed in press releases. What these clusters are essentially capturing remains to be explored in further studies.

References

[1] Charu C Aggarwal and ChengXiang Zhai. A survey of text clustering algorithms. In *Mining text data*, pages 77–128. Springer, 2012.

[2] Kenneth R Ahern and Denis Sosyura. Who writes the news? corporate press releases during merger negotiations. *The Journal of Finance*, 69(1):241–291, 2014.

[3] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[4] Florian Beil, Martin Ester, and Xiaowei Xu. Frequent term-based text clustering. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 436–442. ACM, 2002.

[5] Jan Bena and Kai Li. Corporate innovations and mergers and acquisitions. *The Journal of Finance*, 69(5):1923–1960, 2014.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[7] Robert Bruner. Where m&a pays and where it strays: A survey of the research. *Journal of Applied Corporate Finance*, 16(4):63–76, 2004.

[8] Robert F Bruner. Does m&a pay? a survey of evidence for the decision-maker. *Journal of applied finance*, 12(1):48–68, 2002.

[9] Bruno Cassiman and Reinhilde Veugelers. In search of complementarity in innovation strategy: Internal r&d and external knowledge acquisition. *Management science*, 52(1):68–82, 2006.

[10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[11] C Edward Fee and Shawn Thomas. Sources of gains in horizontal mergers: evidence from customer, supplier, and rival firms. *Journal of Financial Economics*, 74(3):423–460, 2004.

[12] Kathleen Fuller, Jeffry Netter, and Mike Stegemoller. What do returns to acquiring firms tell us? evidence from firms that make many acquisitions. *The Journal of Finance*, 57(4):1763–1793, 2002.

[13] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural Computation*, 12(10):2451–2471, 2000.

[14] Alexander S Gorbenko and Andrey Malenko. Strategic and financial bidders in takeover auctions. *The Journal of Finance*, 69(6):2513–2555, 2014.

[15] Matthew J Higgins and Daniel Rodriguez. The outsourcing of r&d through acquisitions in the pharmaceutical industry. *Journal of financial economics*, 80(2):351–383, 2006.

[16] Gerard Hoberg and Gordon Phillips. Product market synergies and competition in mergers and acquisitions: A text-based analysis. *The Review of Financial Studies*, 23(10):3773–3811, 2010.

[17] Marina Martynova and Luc Renneboog. A century of corporate takeovers: What have we learned and where do we stand? *Journal of Banking & Finance*, 32(10):2148–2177, 2008.

[18] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[20] Gordon M Phillips and Alexei Zhdanov. R&d and the incentives from merger and acquisition activity. *The Review of Financial Studies*, 26(1):34–78, 2013.

[21] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

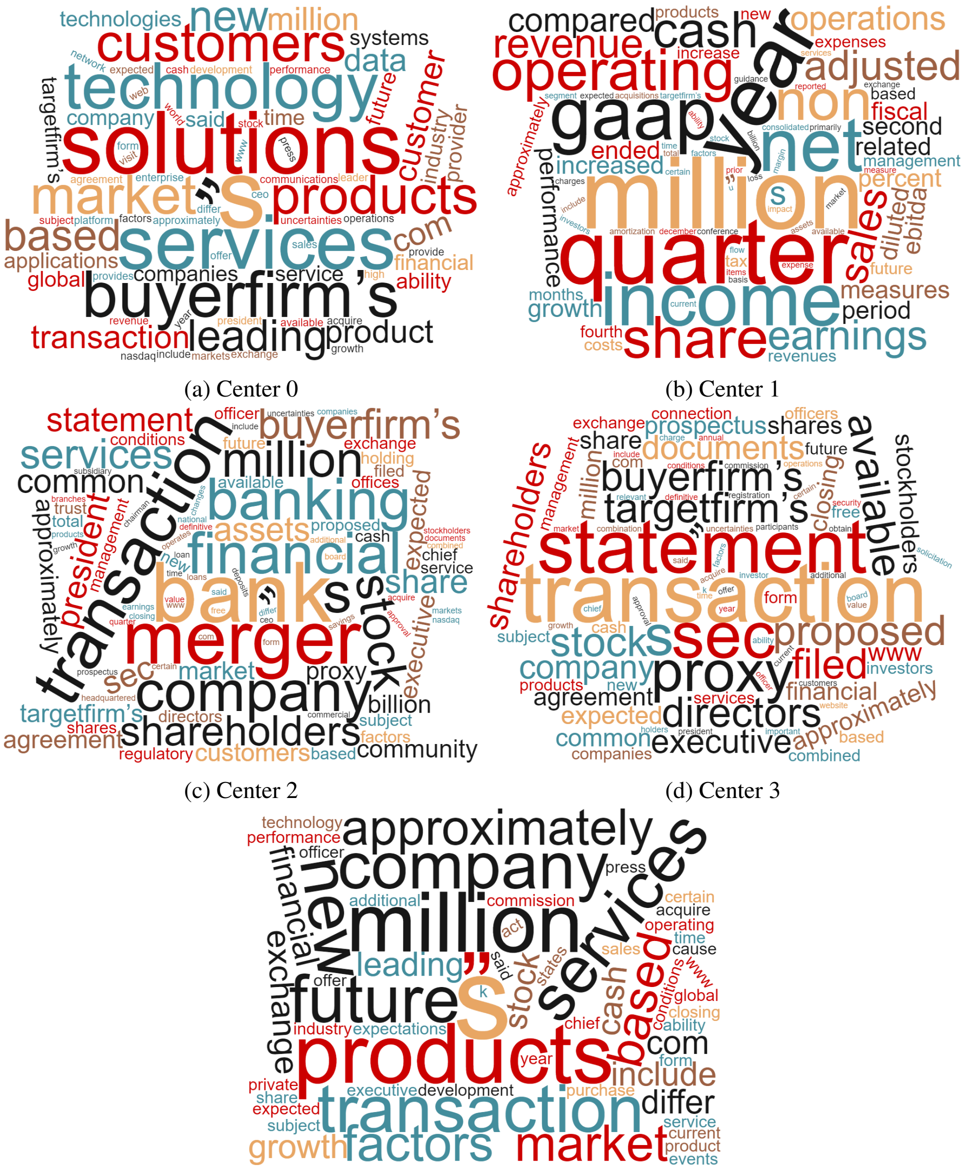
[22] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[23] Noam Slonim and Naftali Tishby. Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 208–215. ACM, 2000.

[24] Pengtao Xie and Eric P Xing. Integrating document clustering and topic modeling. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 694–703. AUAI Press, 2013.

[25] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM, 2003.

A Table of contents


technologies new million
customers systems data
technology
company said time
solutions
market ’s products
based services com
applications financial
global companies service ability
buyerfirm’s
transaction leading product
targetfirm’s customer industry provider future
compared cash operations
revenue increase expenses
operating adjusted
gaap year non fiscal
net second related management
increased ended performance approximately
million percent
quarter sales diluted ebitda
income measures future
months growth period
fourth costs share earnings revenues
statement officer buyerfirm’s
services million exchange
common banking holding filed offices
approximately trust total president management
assets proposed cash expected
financial chief service share
bank ’s stock
transaction merger executive
sec market proxy
company billion
targetfirm’s directors subject
agreement shareholders factors
regulatory customers based community
connection officers
exchange prospectus shares
share documents future available
buyerfirm’s
targetfirm’s stockholders
shareholders management million closing
statement
transaction form
stock sec proposed
subject company filed www investors
products proxy
agreement services financial
expected directors approximately
common executive based
companies combined
technology performance approximately
officer company press
financial new additional commission certain acquire
million operating
leading ’s services sales cause
exchange offer
future stock cash based
industry expectations chief
products com
year include
private share development purchase
expected transaction differ
subject service current
growth factors market product events

(a) Center 0

(b) Center 1

(c) Center 2

(d) Center 3

(e) Center 4

Figure 4: centers

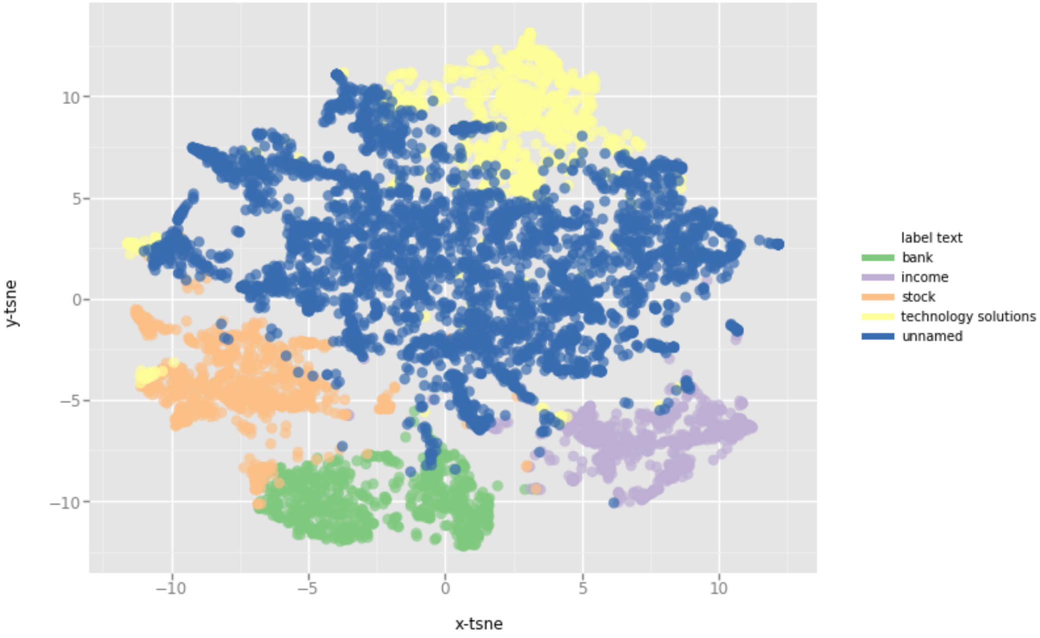
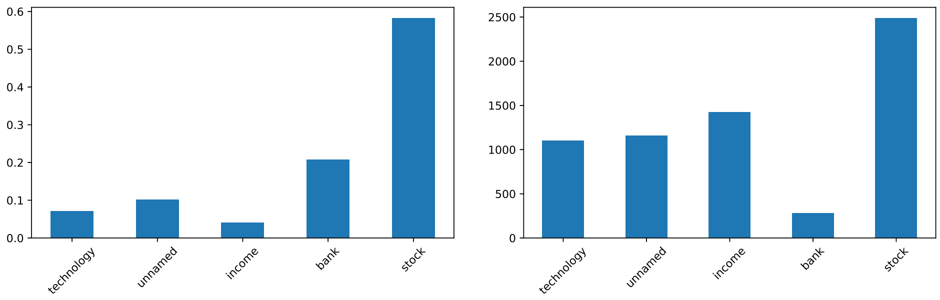


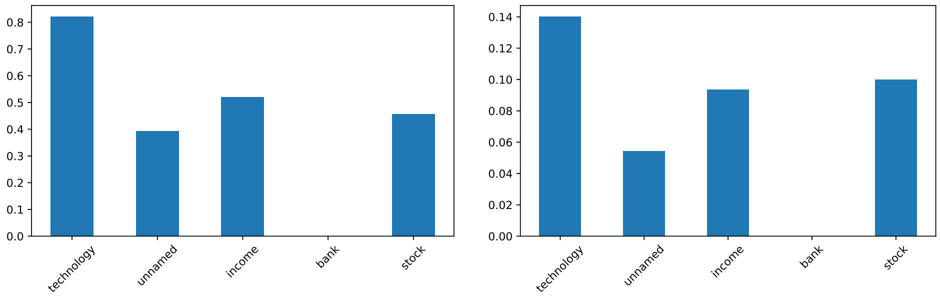
Figure 5: 2 dimensional t-sne graph



(a) Proportion of targets in Compustat

(b) Average sales of buyers

Figure 6: Firm size



(a) Proportion of R&D active firms

(b) Average R&D intensity among research firms

Figure 7: R&D activities

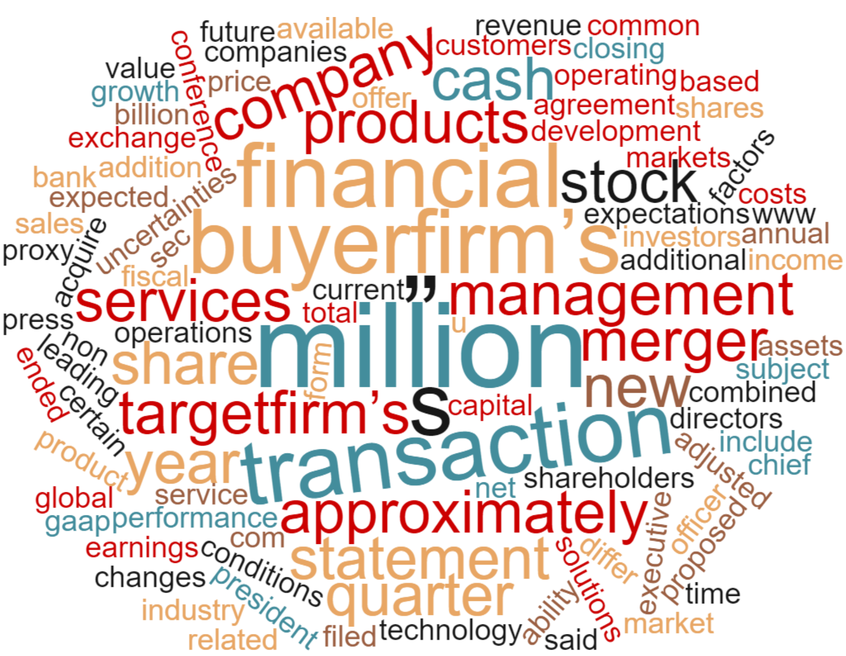
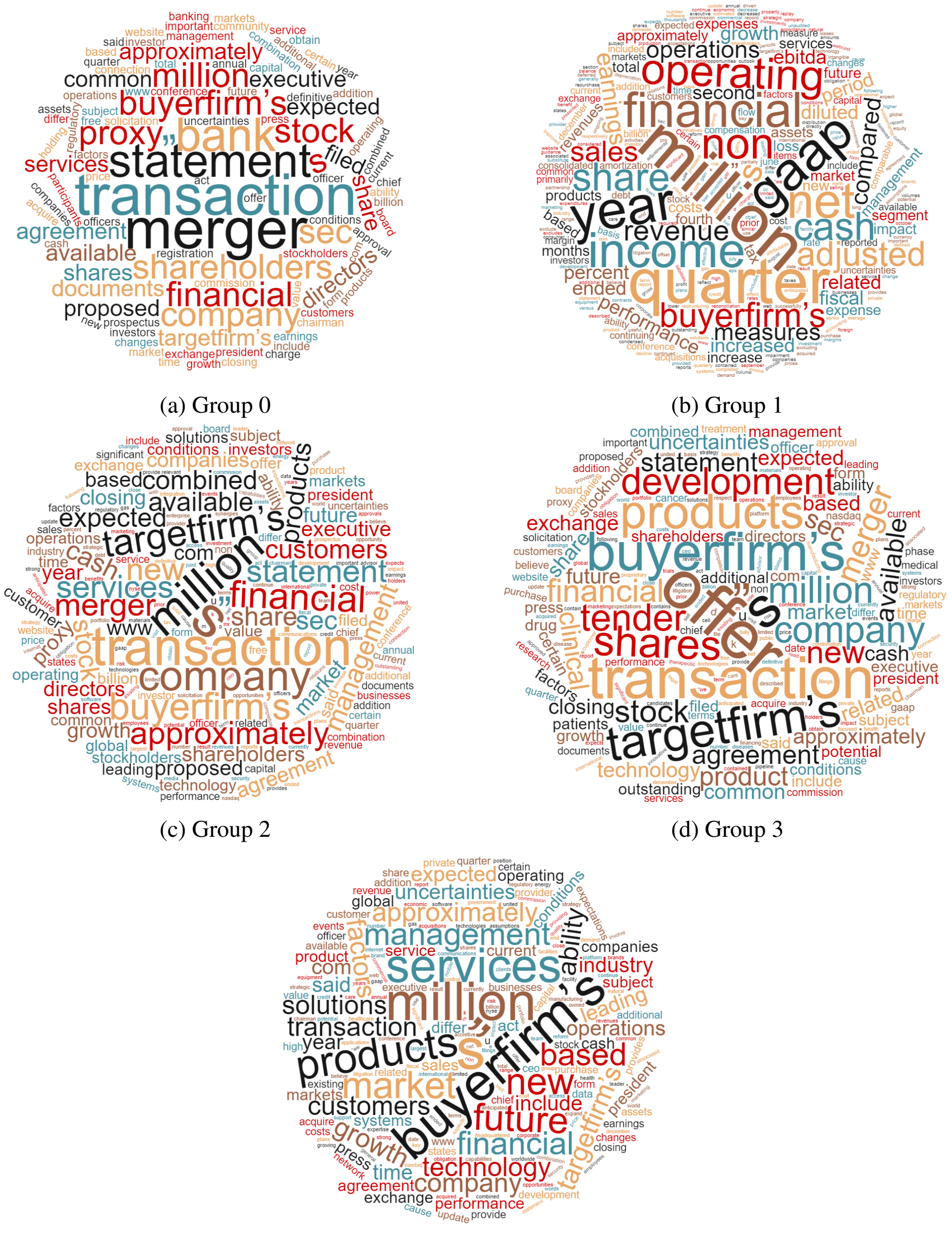


Figure 8: Global Topics

Topic	Top 10 words
27	vehicle automotive parts homes vehicles repair new land market ford
3	offer tender shares statement sec transaction solicitation purchase targetfirm’s documents
29	gas oil production natural reserves drilling exploration properties mining texas
0	offer shall registration sale laws sell jurisdiction solicitation act offering
26	price ’s prices q estimated market value changes total losses
18	stock shares common merger agreement shareholders share board outstanding directors
23	’s com www u o r contact new e president
13	" buyerfirm’s targetfirm’s we future factors form "the words events
17	transaction company combined expected growth companies closing approximately billion value
16	future factors uncertainties press act exchange differ expectations events cause

Table 1: Some converged topics



(a) Group 0

(b) Group 1

(c) Group 2

(d) Group 3

(e) Group 4

Figure 9: Local Topics

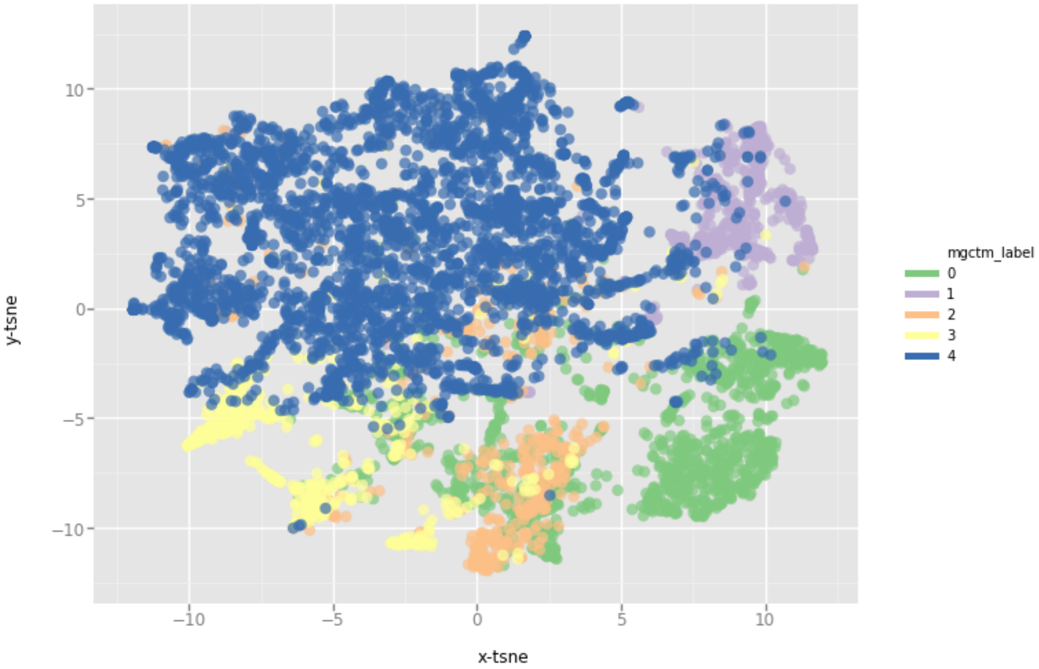


Figure 10: Group assignment of MGCTM in identical t-sne