Using semantic fingerprinting in finance

Feriha Ibriyamova, Samuel Kogan, Galla Salganik-Shoshan and David Stolin [*]

March 2016

Abstract

Researchers in finance and adjacent fields have increasingly been working with textual data, a common challenge being analyzing the content of a text. Traditionally, this task has been approached through labor- and computation-intensive work with lists of words. In this paper we compare word list analysis with an easy-to-implement and computationally efficient alternative called semantic fingerprinting. Using the prediction of stock return correlations as an illustration, we show semantic fingerprinting to produce superior results. We argue that semantic fingerprinting significantly reduces the barrier to entry for research involving text content analysis, and we provide guidance on implementing this technique.

Keywords: textual analysis, industries, stock returns, semantic fingerprint

JEL codes: G10

[*] Kogan and Ibriyamova are at Leiden University, Salganik-Shoshan is at Ben-Gurion University, and Stolin is at University of Toulouse, Toulouse Business School. Contact e-mail: d.stolin@tbs-education.fr. We thank Erik Graf, Søren Tjagvad Madsen and Francisco Webber for their help and participants at the 2016 FIBA conference as well Olivier Galaud, Bernd Hanke, Jerry Hoberg, Bill McDonald, Laura Obreja Braşoveanu, Crina Pungulescu and Maxim Zagonov for comments. We are very grateful to Jerry Hoberg and to Paul Ma for sharing their data. All errors are ours.

1. Introduction

Finance and economics researchers are deluged with data, much of it of non-quantitative nature. Accordingly, there is much interest in how such data – usually in the form of text – can be used in explaining and predicting financial and economic phenomena. Our paper conducts a first assessment of the predictive power of an emerging development in textual analysis: semantic fingerprinting.[1]

A series of papers co-authored by Hoberg and Phillips (2010, 2015a, 2015b) pioneered the use of textual data to measure similarity between firms' products and therefore their proximity in the business space. These papers document that text-derived measures of competitive proximity outperform traditional industry classifications in a wide variety of predictive and explanatory specifications. Specifically, these papers use text sources such as 10-K business descriptions to create descriptive word lists for individual firms. Each firm is then represented by a vector of 1s and 0s indicating the presence and absence, respectively, of a given word in the text. The cosine similarity between the vectors captures the overlap between the word lists and is therefore a measure of proximity between firms.

Recent advances in the semantic analysis of texts make it potentially possible to improve on such 'word list' analyses. For example, if one text uses the term 'online' while another relies on the term 'internet', a word list process would not conclude that the two texts are talking about the same thing. Semantic fingerprinting, on the other hand, is 'trained' on a large body of text so that it can record concepts that a given word is associated with, similarly to how the

[1] Semantic fingerprinting can be experienced interactively on http://www.cortical.io/static/demos/html/fingerprint-editor.html. Appendix A shows the semantic fingerprint of the word 'fund'.

human neocortex processes and stores related concepts. These related concepts form the so-called ‘semantic fingerprint’ of a given word. For example, the terms ‘online’ and ‘internet’ will have similar semantic fingerprints. Accordingly, the semantic fingerprints of two texts each of which relies on one but not both of two related words, will overlap. Therefore, the two related words’ presence will contribute to the two texts’ proximity measure if the measure is derived from their semantic fingerprints but not if it is derived from word counts. We thus expect that semantic fingerprinting has the potential to improve on word list methods, at least when it comes to measuring document similarity – and therefore the similarity of two firms when the documents in question describe the firms’ business activities.

The premise underlying our tests is simple. If semantic fingerprinting improves on word list-based analyses when measuring firm relatedness, it should be useful for predicting future stock return correlations, as similar firms can be expected to have similar stock price responses to relevant news. We therefore conduct regressions of pairwise correlations for a group of firms from a variety of industries on textual proximity measures derived from semantic fingerprints and from word lists, with prior correlations and other similarity measures as controls. We find that similarities based on semantic fingerprints are better at predicting correlations than are those based on word lists. This finding, combined with the ease of use, computational advantage and visual interpretation inherent to semantic fingerprinting suggest that this method merits attention from finance and economics researchers. To this end, Appendix B of our paper provides detailed information about how to implement semantic fingerprinting.

The remainder of our paper has the following structure. In Section 2 we overview semantic fingerprinting and introduce our data. In Section 3 we use semantic fingerprinting to predict stock return correlations. Section 4 discusses possible applications of semantic fingerprinting in finance, and Section 5 concludes.

2. Method and data

Our interest is in extracting valuable information from textual data about companies. Specifically, we wish to measure similarity between firms based on their descriptions. The traditional method finance researchers use to compare documents (e.g. Hoberg and Phillips 2010, 2015a, 2015b; Box 2015) is based on word lists. Under this approach, documents are first reduced to a list of words that are deemed to have potential relevance to the documents' meaning (which in practice means that the most commonly used words in a library of documents are dropped). Each document is then represented as an N-dimensional vector, where N is the number of distinct words in the library, and the nth element of the vector is either a 0/1 indicator variable capturing whether the nth word is present in the document (e.g. Hoberg and Phillips 2010, 2015a, 2015b) or the frequency of the word in the document (Box 2015). The similarity of two documents is then obtained as the cosine similarity measure, which can be interpreted as the cosine of the angle between the two vectors in the N-dimensional space:

$$\cos(\boldsymbol{v_i}, \boldsymbol{v_j}) = \frac{\boldsymbol{v_i}^{\boldsymbol{T}} \boldsymbol{v_j}}{|\boldsymbol{v_i}||\boldsymbol{v_j}|}$$

Since the elements of the vectors representing each document are all non-negative, the cosine similarity is restricted to the [0,1] interval. As Loughran and McDonald (2015) point out, if the word vectors are first mean-adjusted, then the cosine similarity equals the Pearson correlation between them.

Semantic fingerprinting is a new method whose manner of processing text is modelled on that of the human neocortex (see Numenta 2011). Due to its focus on learning links between related concepts, semantic fingerprinting has the potential to be more powerful in document comparisons than are word list-based analyses. The leading proponent of semantic fingerprinting is Cortical (www.cortical.io). While Cortical's white paper (Webber 2015) and online documentation supply in-depth information on semantic fingerprinting, we provide a brief description below.

First, Cortical's so-called 'retina' engine is trained on a large number of single-topic texts (retina version 2.2.0, which we used, was trained on 4.4 million pre-selected Wikipedia articles). The retina then identifies 2^{14} (or 16,384) clusters of co-occurring words; Cortical calls these clusters 'semantic contexts'. Each word is linked to a small fraction of the contexts which therefore define the word. This means that each word is characterized by a sparse 16,384-row vector of 0s and 1s, with "1" indicating that the word is linked to the particular semantic context, and "0" indicating the absence of a link. Semantic contexts are further organized into a 128-by-128 matrix, where related contexts (those that tend to be linked to the same words) are adjacent to one another. Thus, the semantic fingerprint of a word can be visualized as a 128-by-128 matrix of 0s and 1s (or equivalently, a 16,384-dimensional vector of 0s and 1s). The semantic fingerprint for a text is derived by Cortical's engine from the semantic fingerprints of the words comprising the text. This allows us to obtain semantic fingerprints for individual companies based on their business descriptions. Once semantic fingerprints are obtained, the similarity between the two companies is calculated using the cosine similarity between the two matrices just as it is for word list-based measures of Hoberg and Phillips.

We note that semantic similarity in the finance literature has also previously been studied with latent semantic analysis (Boukus and Rosenberg 2006; see also Loughran and McDonald 2015) which, like semantic fingerprinting, takes into consideration connectedness between different words. Semantic fingerprinting has important theoretical and practical advantages over latent semantic analysis a discussion of which goes beyond the scope of this paper (see Webber, 2015). One of these is that semantic fingerprinting relies exclusively on binary vectors, making all processing several orders of magnitude faster. In addition, semantic fingerprinting is easier to use for a non-specialist.

Although Cortical allows users to curate their own specialized libraries on which to train the retina, we have not done so in this project. Clearly, the power of semantic fingerprinting in application to business-related texts would increase if such texts are used to create the retina. This is analogous to Loughran and McDonald (2015) reporting that word lists are more useful in working with 10-K filings when the lists themselves are derived from such filings.

It is worth addressing a potential disadvantage that semantic fingerprinting shares with other artificial intelligence / machine learning methods: from the perspective of a finance researcher, it is largely a ‘black box’. This makes it difficult, for example, to examine the robustness of the results to changes in the textual analysis algorithm. We believe, however, that this (arguable) disadvantage is outweighed by the advantages. Machine learning and artificial intelligence have made enormous strides in the recent past, and will make even bigger ones in the near future. Accordingly, they are beginning to transform business practice in general and finance in particular, as the rise of such firms such as Numenta or Kensho illustrates. Side-stepping the use of such techniques simply because of their complexity would mean academic researchers lagging behind cutting-edge practice, instead of leading it. Indeed, our approach – comparing a proprietary machine-learning methodology with a more transparent word-list one – echoes Heston and Ranjan Sinha's (2015) comparison of the performance of Thomson-Reuters' NewsScope with word-list methods in the related context of sentiment analysis. Further, in Appendix B we describe the semantic fingerprinting procedure we use in full detail, making it easy for others to replicate our results and to use semantic fingerprinting in their own projects.

Our data sample consists of the 30 Dow Jones Industrial Average (DJIA) constituents. The reason for this choice is that in order to illustrate semantic fingerprinting, we wanted to use a small, well-defined group of well-known stocks that is balanced across sectors. Specifically, we use the DJIA constituent set as of end-2012, with corresponding 10-K forms that are filed during the 2013 calendar year. The motivation for this is due to the combination of the following considerations. First, the retina we are using is trained on Wikipedia articles that are accessed through end-2013, and we wanted to avoid look-ahead bias. Second, Hoberg-Phillips cosine similarities, which we benchmark our semantic fingerprint-based similarities against, at the time of writing were only available through 2013. Third, we wanted to use recent data, while still being able to calculate stock return correlations one year ahead (i.e. for the full 2014 calendar year).

For each of the firms in our sample, we retrieve business descriptions from their 10-K filings from the SEC's EDGAR site. The descriptions correspond to Item 1 in the 10-K filings, which are variable in structure and length and can range from just over two thousand words (Johnson & Johnson) to over 40 thousand words (American Express). We do not further pre-process these descriptions, and simply obtain the corresponding semantic fingerprints using the procedure described in Appendix B. The actual semantic fingerprints for several of the firms are shown in Appendix C. It is easy to see that pairs of companies involved in similar business activities (JP Morgan and Bank of America; Merck and Pfizer; Microsoft and Intel; United Technologies and Boeing) have visually similar semantic fingerprints.

In the final step, we calculate the pairwise cosine similarity measure based on semantic fingerprints (henceforth s^{SF}) for each of the 30*(30-1)/2=435 pairs of firms. For comparison, we also use Hoberg and Phillips' pairwise cosine similarity measure based on word lists (henceforth s^{WL}) kindly provided by them. Due to an apparent RA error, Hoberg-Phillips data exclude Du Pont, hence our subsequent regression analyses are based on 29 stocks only, i.e. on 406 pairs of firms.

Note that Hoberg and Phillips exclude the 25 percent most common words before calculating their cosine similarities. We follow their approach by excluding the 25 percent most common semantic contexts before calculating our cosine similarities.

Lastly, we calculate stock return correlations for every pair of stocks using daily stock returns for 2013 and 2014.

[Table 1 here]

Descriptive characteristics of our sample are contained in Table 1. The first row of Table 1 pertains to 3M which, in its 10-K filing, defines itself as '*a diversified technology company with a global presence in the following businesses: Industrial and Transportation; Health Care; Consumer and Office; Safety, Security and Protection Services; Display and Graphics; and Electro and Communications*'. Columns 2 through 4 of Table 1 state that 3M's ticker symbol is MMM, its GICS sector is Industrials, and its fiscal year ended in December (hence the business description we use is as of 31 December, 2012). Column 5 reports that, according to the sample companies' semantic fingerprints, the company most similar to 3M is Du Pont (stock ticker DD, self-definition '*DuPont brings world-class science and engineering to the global marketplace in the form of innovative products, materials and services.*'). The company most similar to 3M based on Hoberg-Phillips word-list similarities (Column 6) is General Electric, which describes itself as '*one of the largest and most diversified infrastructure and financial services corporations in the world*'. (Recall that Du Pont happens to be excluded from the Hoberg-Phillips sample). On the basis of 2013 daily stock return correlations (Column 7), however, 3M is closest to United Technologies which, according to its 10-K, '*provides high technology products and services to the building systems and aerospace industries worldwide*.'

The first row of Column 8 shows that the average value of s^{SF} for firms in the same Industrials sector as 3M (namely, BA, CAT, GE, and UTX) is 0.063, while the corresponding average for firms in other sectors, in Column 11, is actually higher, at 0.115; this result is counterintuitive. The average values of s^{WL} for same-sector firms (in Column 9) and for other-sector firms (in Column 12) are 0.053 and 0.013, respectively, which is consistent with same-sector companies having more similar descriptions. Lastly, Columns 10 and 13 indicate that the average correlation of 3M's stock return with its Industrials sector peers is 0.511, while its correlation with other DJIA stocks is, as we would expect, lower, at 0.414.

Although the lower semantic fingerprint-based similarity between 3M and its GICS sector peers than between 3M and other-sector firms is counter to expectations, this is rather exceptional – the only other company in our sample for which this is the case is Du Pont. For the full sample the expected ordering is restored, as shown in the last row of the table. Specifically, SF similarities for same-sector vs. different-sector firms average 0.327 vs. 0.114; WL similarities average 0.090 vs. 0.015; and stock return correlations average 0.434 vs. 0.316, respectively. All of these differences are significant at the 1 percent level.

Casual perusal of columns 5 and 6 of the table reveals several surprises. While the company closest to Exxon Mobil according to SF similarity (and stock return correlation) is Chevron, Exxon's nearest match according to the WL criterion is Boeing. The latter is probably due to the important role played by fuel costs in Boeing's operations. On the other hand, in the case of General Electric, while WL picks out United Technologies as the most similar company, SF points to JP Morgan, a bank. Although surprising at first, the latter result makes more sense if we recall that General Electric stresses its provision of financial services in its self-definition.[2]

The closest matches on the basis of stock return correlations (Column 7) tend to pick out plausible pairs, but on other occasions may appear puzzling – as is the case for the predominance of United Technologies and 3M in that column, with 6 and 4 appearances,

[2] Indeed, GE was even classified as a systematically important financial institution (SIFI) by the Financial Stability Oversight Council.

respectively. However, given that these are highly diversified companies, it makes sense that they would have substantial correlations with multiple firms via their correlations with the market index. In fact, of the stocks in our sample, United Technologies has the highest correlation with the DJIA index (0.776) and 3M the second highest (0.758); the third highest is Travelers (0.717), which makes two appearances in Column 7. This does raise a question about how well stock return correlation identifies similar companies. While we are not arguing that stock return correlation is the best single measure of company similarity, it is widely perceived to be related to similarity, in addition to being an important parameter in investment decision-making. Therefore, our subsequent comparisons of WL-based and SF-based similarity measures will focus on how well they predict future stock return correlations.

In addition to text-based measurement of company similarity, two other measures have recently attracted attention for their ability to group together companies with similar financial characteristics, which we therefore include in some of our subsequent analyses. One is the common analyst (CA) measure of Kaustia and Rantala (2013). To calculate it, for each of our sample firms we use I/B/E/S data to create a vector representing the sell-side analysts that followed it in 2013, and obtain the CA-based similarity measure for a pair of firms, s^{CA}, as the cosine similarity for the corresponding pair of vectors.

The second measure is the Lee, Ma and Wang (2015) search-based (SB) measure. Lee et al. report that a dataset of IP addresses from which SEC filings of US companies are consulted can be used to produce particularly meaningful peer groupings. Specifically, according to the SB measure, peer companies are those which tend to be searched for from the same IP addresses. For comparability with our other similarity measures, we would ideally use a cosine similarity measure based on vectors of unique IP addresses used for each firm.

However, Lee et al. argue that for their purposes the best measure is the search fraction f_{ij}, representing the ratio of unique visitors who searched for company j after searching for company i to those who searched for any company after searching for company i. Lee et al. kindly provided us with 2012 (the latest year available) f_{ij} values for each company, covering its top ten co-searches. As all of our other measures of similarity are symmetric, we force symmetry by defining search-based similarity as $s^{SB} = \frac{f_{ij}+f_{ji}}{2}$.

In order to convey a sense of how the different similarity measures – s^{SF}, s^{WL}, s^{CA} (all of which are based on 2013 data) and s^{SB} (based on 2012 data), as well the pairwise daily stock return correlations in 2012, 2013, 2014 – interrelate, we report correlations between them in Table 2.

[Table 2 here]

As we would expect, all of the correlations are positive and (not reported in the table) highly statistically significant (p-values < 0.01). Since s^{SF} and s^{WL} are derived from the same source, it is unsurprising that the correlation between them is quite high, namely, 0.665. Interestingly, though, the correlation between s^{CA} and s^{SB} is higher still, at 0.817 – even though the CA and SB measures are obtained from different sources, over different years, and using different metrics. This suggests that while usage-based similarity measures are likely to lead to the same conclusion (after all, many of the sell-side analysts common to two firms are likely to be searching for both firms on the SEC site), there is more scope for divergence – and, by implication, improvement – when measuring company similarity with different methods on the basis of textual descriptions. We note also that pairwise stock return correlations are quite stable over time (their autocorrelations are in excess of 0.5) and that they are more highly correlated with s^{SF} than they are with s^{WL}.

[Table 3 here]

To supply further insight into our SF-based firm similarity measure, we present this measure for all pairs of stocks in Table 3. The highlighted off-diagonal terms identify pairs of companies in the same GICS sector. It is easy to see that SF similarity for most same-sector firms tends to be higher than the previously reported different-sector average of 0.114, and is often substantially higher – in excess of 0.3, 0.4 and even 0.5. Further, when same-sector firms have low SF similarity, it is far from clear that a human analyst would have identified these firms as similar. This is the case, for example, for 3M and Boeing, which, despite co-existing in the Industrials sector have SF similarity score of 0.032 – and whose stocks have no sell-side analysts in common.

In short, prima facie evidence suggests that similar companies are likely to have more similar semantic fingerprints. The question of whether SF or WL captures similarity better, and whether they do so after controlling for other similarity measures, will be addressed in the next section.

3. Predicting stock return correlations with firm similarity measures

In this section we assess the usefulness of semantic fingerprint-based cosine similarity between firms in the financial context and compare it with that of Hoberg-Phillips word list-based cosine similarity. A natural way to do this is to examine the ability of the two similarity measures to predict correlations between stock returns for pairs of firms. The reason for this is that the more similar two firms are, the more similarly they will respond to demand shocks, new regulations, and other developments. Thus, Box (2015) studies the similarity of newswire texts for pairs of firms and finds that it helps predict future stock return correlations even after taking into consideration contemporaneous correlations. While our research uses 10-K filings rather than newswires, the context and purpose of our investigation are similar to Box's, so we largely adopt his methodology. Firstly, to address the bounded nature of the correlation coefficients, we apply the Fisher transformation, $z = 0.5\ln\frac{1+\rho}{1-\rho}$. Secondly, because correlations are covariances scaled by the product of standard deviations, our predictive regression needs to include the product of standard deviations of the two firms' stock returns to account for this mechanical association. Accordingly, our basic regression takes the form

$$z_{i,j,t} = \beta_0 + \beta_1 z_{i,j,t-1} + \beta_2 \sigma_{i,t-1}\sigma_{j,t-1} + \beta_3 s^k_{i,j,t-1} + \varepsilon_{i,j,t} \qquad (1)$$

where $z_{i,j,t}$ is the correlation between the stock returns of company *i* and company *j* in period *t*; $\sigma_{i,t}$ is the standard deviation of the stock return of company *i* in period *t*; and $s^{SF}_{i,j,t}$ ($s^{WL}_{i,j,t}$) is the semantic fingerprint-based (respectively, word list-based) cosine similarity between company *i* and company *j* in period *t*.

However, to facilitate intuitive interpretation of the results, we also report on the simpler regression specification,

$$\rho_{i,j,t} = \beta_0 + \beta_1 \rho_{i,j,t-1} + \beta_2 s^k_{i,j,t-1} + \varepsilon_{i,j,t} \qquad (2)$$

[Table 4 here]

Our results are presented in Table 4. Panel A shows the results of specification (2). The regressions reported in Column 1 and 2 show that past correlations strongly predict future correlations, with R-squared around 0.3. Column 3 indicates that semantic similarity as measured by s^{SF} is also a highly significant predictor of future correlation, although its R-squared is only 0.092. The strong statistical significance of s^{SF} is confirmed even if past correlations are also included in predicting future correlations, as is the case for the regression results in Columns 4 and 5.

The results of regressions based on our preferred specification (1) are presented in Panel B of Table 4. They confirm the results of Panel A. In particular, Column 4 shows that, when predicting (the Fisher transformation of) the 2014 pairwise correlation between companies' stock returns, s^{SF} is statistically significant (p-value = 0.02) even after controlling for the 2013 pairwise correlation.

[Table 5 here]

Table 5 shows the results of a 'horse race' among different measures of proximity for pairs of companies used to predict their stock return correlations. In addition to our own semantic fingerprint-based similarity measure, s^{SF}, and Hoberg-Phillips' word-list based measure s^{WL}, we also use the Kaustia and Rantala (2013) common-analyst similarity measure, s^{CA}, and a search-based measure, s^{SB}, based on the work of Lee, Ma and Wang (2015).

As in the previous table, we report on specification (2) in Panel A, and on our preferred specification (1) in Panel B. The tenor of the results in the two panels is similar, so we focus the discussion below on Panel B. Our main interest is in comparing the performance of the WL-based similarity measure with the SF-based one. Column 1, for ease of comparison, restates the results of Column 5 of Table 4, Panel B, showing s^{SF} to be a highly significant predictor of future stock return correlation even after including past correlation. Column 2 shows that WL-based similarity is likewise significant, but with a lower R-squared (0.388 vs. 0.382) and Akaike Information Criterion (-676.9 vs. -681.1). The 4.2 difference in the Akaike Information Criterion exceeds the commonly used threshold of 2 and is therefore suggestive of significantly higher predictive content of SF similarity as compared to WL similarity.

For completeness, in Columns 3 and 4 we report regressions with CA- and SB-based similarities, both of which are highly significant predictors of future correlations, as would be expected. Interestingly, though, even if we include all four similarity measures simultaneously (Column 5), the coefficient estimate of s^{SF} is significant (p-value = 0.02) while that of s^{WL} is not (p-value = 0.46). In all, our regressions suggest that semantic fingerprinting 1) outperforms word list-based similarity and 2) has incremental explanatory power even over CA- and SB-based similarity measures. The next section puts this finding into perspective.

4. Discussion

The ground-breaking work of Hoberg and Phillips showed that valuable quantitative content can be extracted from listed firms' business descriptions. Our results indicate that, compared to the straightforward word list-based procedure or Hoberg and Phillips, semantic fingerprinting performs better, albeit by a narrow margin. Is the use of semantic fingerprinting worth it?

Before considering the pros, we will enumerate the cons. First, SF analysis is more of a black box than its WL counterpart: it is harder (although possible) for a researcher to deconstruct an SF similarity measure between a pair of texts than to deconstruct a WL one. Second, given that the semantic fingerprinting retina initially needs to be trained on a set of texts, a different retina needs to be used for different cross-sections of texts if one is to avoid look-ahead bias while keeping the retina up-to-date.

On the other hand, SF has important benefits. First, unlike the WL method, it requires no pre-processing e.g. to remove common words. Second, SF is easy to implement. Third, SF comes with a powerful visual interpretation. Fourth, SF is fast, which can be a particularly strong advantage in automated trading applications. Fifth, SF (at least in the setting we used) produces stronger results. Sixth, SF, unlike WL, has significant scope for further improvement.

In fact, while the retina we used was trained on general texts, one would expect performance to improve significantly if business and financial texts were used to train it. Another major improvement would result from exploiting the topology of SF. A cursory glance at the fingerprints in Appendix C is sufficient to make it clear that the shapes of the constellations of contexts corresponding to Merck and Pfizer are very similar, while those corresponding to Merck and Microsoft are not. The cosine similarity measure is the same whether or not the 16,384 possible semantic contexts are arranged such that related contexts are nearby, and so does not leverage this two-dimensional topology of SF (which WL in any case does not possess); a measure capable of comparing two-dimensional patterns would likely produce much better results.

While we used SF to predict stock return correlations, we did so merely for expository reasons. Of course, better prediction of correlations is worthwhile, not least in the context of portfolio optimization, but the potential uses of semantic fingerprinting stretch well beyond this application. Here are but a few possibilities in the finance context (it is easy to extend this list to other areas of business and economics research):

- Using SF similarities between sell-side analysts' professional profiles and the descriptions of companies covered by them as weights may produce more accurate aggregate analyst forecasts.
- Using SF to compare company director profiles may produce better measures of board diversity.
- Using SF to identify unusual news patterns may produce better predictions of volatility than sequential textual analysis as in Mamaysky and Glasserman (2015).
- Using SF to fingerprint news stories and to correlate them to the fingerprint of an investor's portfolio may be a valuable addition to a risk management system.

In short, semantic fingerprinting has a number of promising potential applications in finance. Implementing them and evaluating their performance is left to future research.

5. Conclusion

Among researchers in finance and in adjacent disciplines, the race is on to extract the most value from textual data. We argue that semantic fingerprinting is a powerful new tool in this race. As an illustration, we compare the efficacy of semantic fingerprinting with that of the word list-based approach in predicting stock return correlations for pairs of DJIA constituent firms. Even in the absence of human intervention, fine-tuning, or methodological improvements, semantic fingerprinting produces superior results. In addition, the method is easy to set up and use, fast, improvable, and intuitive. We therefore argue that the promise of semantic fingerprinting is worth exploring in a variety of other settings requiring analysis of textual content.

References

Boukus, Ellyn, and Joshua V. Rosenberg, 2006, The information content of FOMC minutes, Working Paper, Federal Reserve Bank of New York.

Box, Travis, 2015, Stock price comovement and the market for information, Working Paper, University of Mississippi.

Heston, Steven L., and Nitish Ranjan Sinha, 2015, News versus sentiment: Predicting stock returns from news stories, Working Paper, University of Maryland.

Hoberg, Gerard and Gordon Phillips, 2010, Product market synergies and competition in mergers and acquisitions: A text-based analysis, *Review of Financial Studies* 23, 3773-3811.

Hoberg, Gerard and Gordon Phillips, 2015a, Text-based network industries and endogenous product differentiation, *Journal of Political Economy*, forthcoming.

Hoberg, Gerard and Gordon Phillips, 2015b, Text-based industry momentum, Working Paper, University of Southern California.

Kaustia, Markku, and Ville Rantala, 2013, Common analyst-based method for defining peer firms, Working Paper, Aalto University.

Lee, Charles M. C., Paul Ma, and Charles C. Y. Wang, 2015, Search-based peer firms: Aggregating investor perceptions through internet co-searches, *Journal of Financial Economics* 116, 410-431.

Loughran, Tim and Bill McDonald, 2015, Textual analysis in accounting and finance: A survey, Working Paper, University of Notre Dame.

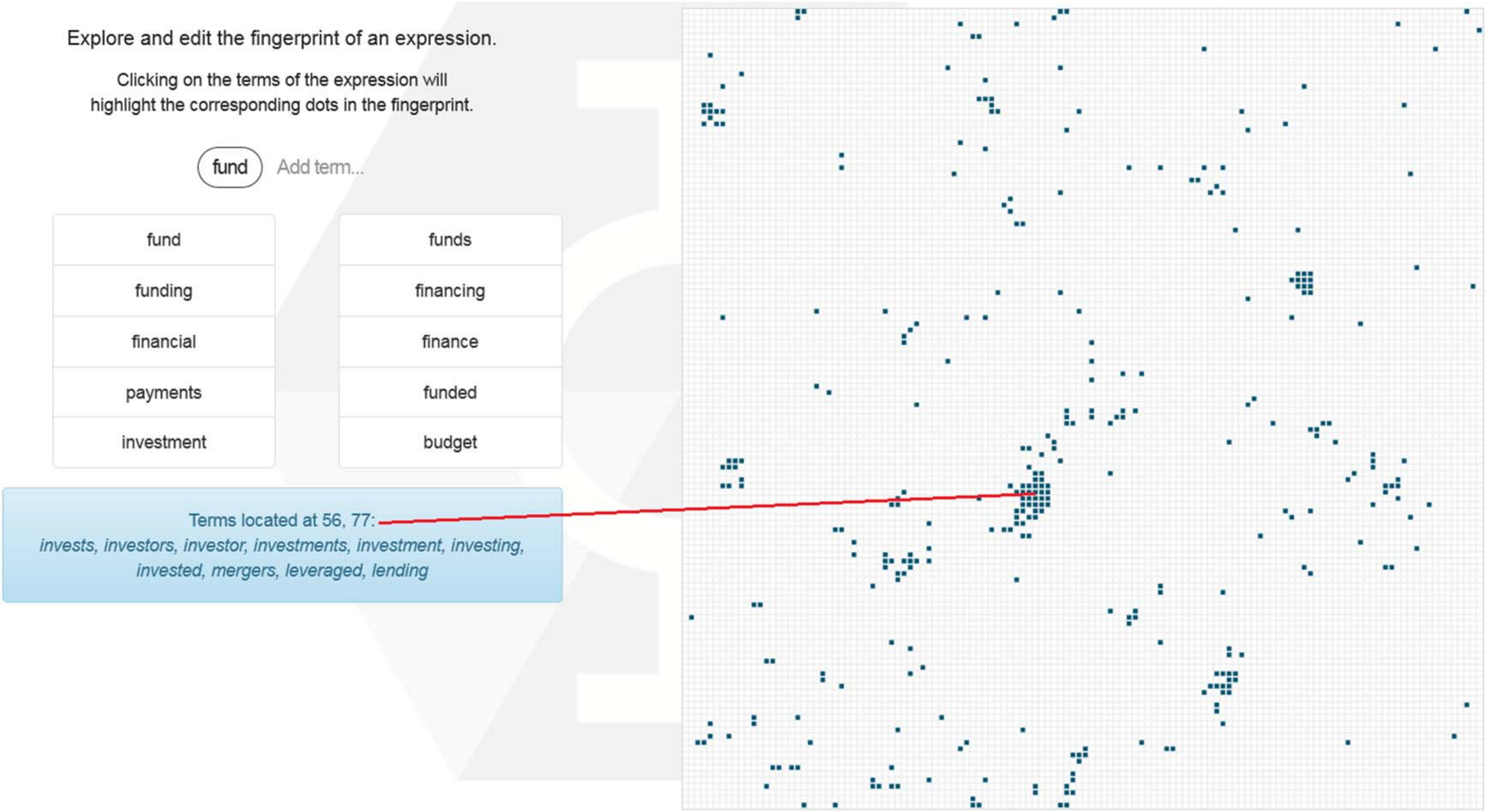
Mamaysky, Harry and Paul Glasserman, 2015, Does unusual news forecast market stress?, Working Paper, Columbia Business School.

Numenta, 2011, Hierarchical temporal memory including HTM cortical learning algorithms, Technical report, Version 0.2.1.

Webber, Francisco Eduardo de Sousa, 2015, Semantic folding and its application in semantic fingerprinting, Cortical.io White Paper, Version 1.0, available from http://www.cortical.io/static/downloads/semantic-folding-theory-white-paper.pdf

Appendix A. The semantic fingerprint of the word ‘fund’

(from http://www.cortical.io/static/demos/html/fingerprint-editor.html)



Explore and edit the fingerprint of an expression.

Clicking on the terms of the expression will highlight the corresponding dots in the fingerprint.

fund	funds
funding	financing
financial	finance
payments	funded
investment	budget

Terms located at 56, 77:
invests, investors, investor, investments, investment, investing, invested, mergers, leveraged, lending

Appendix B: Obtaining semantic fingerprints

In order to obtain the distance between two companies' semantic fingerprints, we proceed as follows. As mentioned in the methodology section of this paper we are using the Cortical API to conduct our semantic analysis of texts. The API uses REST and you can interact with it via HTTP. You can find information about and use the Cortical API at http://api.cortical.io. In writing this paper we used Cortical API version 2.2.0.

One can interact with the API in various ways. You can use the web client or a local client. For the purposes of this paper we used a local client. Cortical have local clients for the Java, PHP, and Python programming languages. Due to our familiarity with the Python programming language we have chosen to use the Python client. You can download a folder with the files and scripts used in this research paper[3].

The tools used for this project were:

- Cortical API (v2.2.0)
- A Cortical API key[4]
- Python 2.7 with Requests[5] extension (required for local Cortical client)[6]
- Cortical Python client (v2.2.0)[7]
- SAS software for statistical analysis

Method and code

We now describe the procedure for acquiring a semantic fingerprint, a list of keywords and a fingerprint image generated for a given text.

The following Python script will only run if you are using the folder which includes the Cortical Python client and all the text files. You must also add your cortical API key by replacing with it the text *Your_API_key_here* in the code below.

The script will generate a semantic fingerprint vector, a keywords list, and an image of the fingerprint (in this order). It will generate these for the text file whose name should be used in place of *15_UTX.txt*.

```
#####################################################
################### Setup code ######################
#####################################################
import os
# Adding cortical client to Python path
import sys
client_path = '%s/CorticalPython_client/' %
os.path.dirname(os.path.realpath(__file__))
```

[3] The information and the code below, as well as sample companies' business descriptions needed for the semantic fingerprinting are available from https://github.com/SamKogan/Semantic_Fingerprinting.

[4] The API key can be requested from http://www.cortical.io/resources_apikey.html

[5] http://docs.python-requests.org/en/latest/

[6] The Anaconda Python distribution (http://continuum.io/downloads) contains both along with the Spyder code editor (IDE).

[7] https://github.com/cortical-io/python-client-sdk

```
sys.path.append(client_path)

# Importing cortical API to script
from cortical.client import ApiClient

#import TextApi for fingerprint vector and keywords
from cortical.textApi import TextApi

#import ImageApi for generating fingerprint image
from cortical.imageApi import ImageApi
client = ApiClient(apiKey="Your_API_key_here",
apiServer="http://api.cortical.io/rest")

# Body contains string of text to be analysed

# Code to get fingerprints from a string
# body = "Semantic fingerprints are cool."

# Code to get fingerprints from a .txt file put filename
file_name = "15_UTX.txt"

with open (file_name, "r") as myfile:
       body = myfile.read().replace('\n', '')

####################################################
######### Code for fingerprint (vector) ############
####################################################

# Chose either en_synonymous or en_associative retina
text = TextApi(client).getRepresentationForText("en_synonymous", body)
print text[0].positions

####################################################
############# Code for keywords list ###############
####################################################

# Chose either en_synonymous or en_associative retina
terms = TextApi(client).getKeywordsForText("en_synonymous", body)
print terms

#####################################################
########## Code for fingerprint (image) #############
#####################################################

body = '{"text":"%s"}' % body

# Chose either en_synonymous or en_associative (default) retina, image scalar
(default: 2), square or circle (default) image, encoding type, and sparsity
terms = ImageApi(client).getImageForExpression("en_synonymous", body, 2,
"square","base64/png", '1.0')

# Chose image name
image_name = file_name.replace(".txt","")
fh = open(image_name + "_fpImage.png", "wb")
fh.write(terms.decode('base64'))
fh.close()

print(image_name + ' fingerprint image saved to %s') %
os.path.dirname(os.path.realpath(__file__))

#####################################################
################## End of Script ####################
#####################################################
```

When the above script is run on the UTX description, it outputs the list of semantic contexts associated with the text followed by the list of keywords extracted from it.

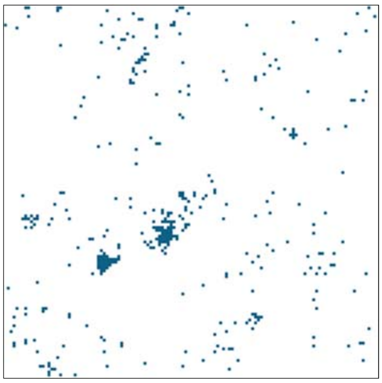
```
[7, 18, 19, 61, 125, 128, 146, 163, 255, 293, 319, 371, 377, 380, 384, 392, 424,
504, 511, 551, 558, 739, 755, 768, 786, 888, 900, 940, 1020, 1141, 1148, 1194,
1212, 1269, 1317, 1324, 1332, 1402, 1438, 1448, 1515, 1569, 1683, 1768, 1822, 1823,
1831, 1918, 1924, 2035, 2053, 2097, 2111, 2137, 2176, 2397, 2493, 2497, 2516, 2517,
2522, 2525, 2616, 2651, 2719, 2782, 2864, 2891, 3104, 3158, 3225, 3286, 3371, 3372,
3451, 3456, 3565, 3772, 3816, 3836, 3945, 3995, 4216, 4219, 4377, 4378, 4421, 4765,
4796, 4888, 4925, 4956, 5016, 5023, 5030, 5114, 5130, 5134, 5183, 5265, 5268, 5269,
5274, 5365, 5398, 5401, 5472, 5523, 5524, 5525, 5526, 5527, 5540, 5656, 5659, 5660,
5661, 5662, 5663, 5664, 5671, 5731, 5781, 5782, 5785, 5792, 5810, 5820, 6010, 6051,
6052, 6068, 6069, 6171, 6176, 6317, 6578, 6610, 6957, 6981, 7219, 7233, 7338, 7427,
7453, 7494, 7606, 7812, 7864, 7917, 8028, 8136, 8212, 8252, 8284, 8326, 8381, 8385,
8389, 8411, 8465, 8486, 8491, 8509, 8510, 8516, 8625, 8639, 8666, 8677, 8721, 8769,
8770, 8780, 8792, 8805, 8875, 8891, 8894, 8895, 8898, 8899, 8919, 8965, 9006, 9011,
9014, 9022, 9023, 9026, 9027, 9046, 9105, 9136, 9144, 9145, 9152, 9153, 9154, 9179,
9198, 9225, 9272, 9276, 9277, 9281, 9302, 9322, 9350, 9351, 9352, 9354, 9360, 9366,
9399, 9405, 9406, 9407, 9484, 9522, 9523, 9525, 9526, 9529, 9530, 9535, 9540, 9609,
9610, 9615, 9651, 9662, 9663, 9706, 9734, 9735, 9737, 9770, 9779, 9790, 9791, 9837,
9842, 9868, 9909, 9910, 9911, 9912, 9917, 10031, 10039, 10040, 10045, 10090, 10145,
10146, 10166, 10167, 10168, 10170, 10171, 10258, 10266, 10269, 10270, 10283, 10284,
10285, 10292, 10293, 10295, 10296, 10298, 10299, 10317, 10413, 10415, 10416, 10424,
10465, 10484, 10535, 10544, 10545, 10553, 10586, 10649, 10663, 10673, 10674, 10675,
10699, 10755, 10782, 10785, 10808, 10809, 10844, 10935, 10936, 10988, 10989, 10993,
11068, 11073, 11081, 11095, 11112, 11168, 11172, 11175, 11193, 11246, 11296, 11300,
11301, 11414, 11424, 11426, 11427, 11504, 11505, 11518, 11555, 11628, 11681, 11682,
11701, 11702, 11711, 11806, 11836, 11842, 11852, 11879, 11883, 11886, 11900, 12003,
12004, 12165, 12263, 12417, 12523, 12651, 12664, 12774, 12897, 12931, 12961, 13024,
13025, 13026, 13034, 13155, 13156, 13161, 13283, 13289, 13310, 13316, 13319, 13325,
13412, 13418, 13419, 13447, 13540, 13541, 13575, 13590, 13653, 13654, 13661, 13669,
13670, 13699, 13727, 13783, 13784, 14026, 14031, 14047, 14083, 14088, 14108, 14183,
14293, 14333, 14412, 14413, 14505, 14536, 14546, 14625, 14667, 14698, 14727, 14855,
14888, 14970, 14996, 15049, 15082, 15094, 15179, 15181, 15182, 15209, 15210, 15243,
15462, 15491, 15563, 15684, 15696, 15721, 15739, 15755, 15783, 15884, 15885, 15887,
15902, 15906, 15907, 15933, 15943, 15998, 16009, 16016, 16017, 16020, 16077, 16143,
16201, 16250, 16259, 16271, 16280, 16327]

[u'pratt', u'whitney', u'sales', u'aerospace', u'businesses', u'sikorsky',
u'products', u'contracts', u'controls', u’subject']

UTX fingerprint image saved to /Current/Directory/
```

Appendix C: Semantic fingerprints of several DJIA constituents

The figures below show the keywords and semantic fingerprints for eight of the 30 constituents of the Dow Jones Industrial Average based on the business descriptions in their 2013 10-K filings.

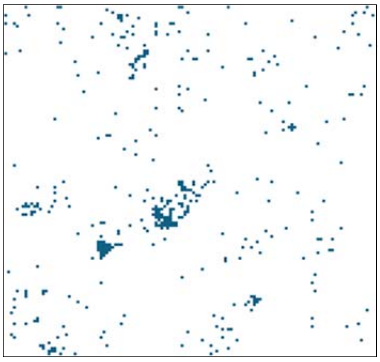
JP Morgan

```
[
  "federal reserve",
  "jpmorgan chase",
  "securities",
  "subsidiaries",
  "banking",
  "bank",
  "fdic",
  "assets",
  "equity",
  "firm"
]
```



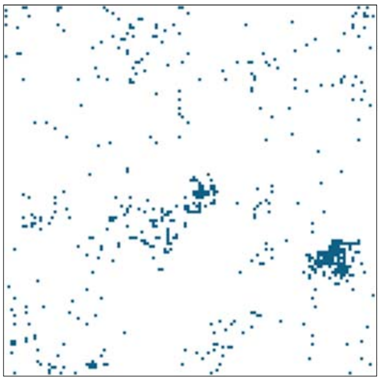
Bank of America

```
[
  "fdic",
  "banking",
  "banks",
  "bank",
  "subsidiaries",
  "investment",
  "federal reserve",
  "insurance",
  "subject",
  "prudential"
]
```



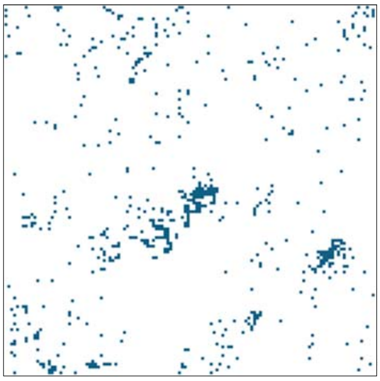
Merck

```
[
  "fda",
  "merck",
  "treatment",
  "vaccine",
  "patients",
  "health",
  "care",
  "medicines",
  "markets",
  "products"
]
```



Pfizer

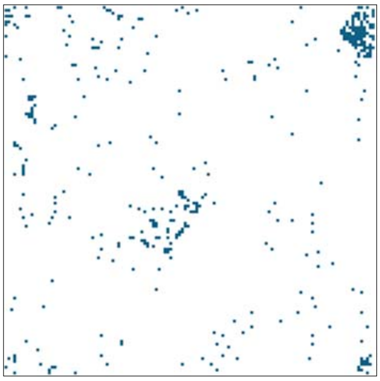
```
[
  "healthcare",
  "products",
  "revenues",
  "pfizer",
  "medicines",
  "medicaid",
  "medicare",
  "drugs",
  "health",
  "fda"
]
```



Appendix C continued.

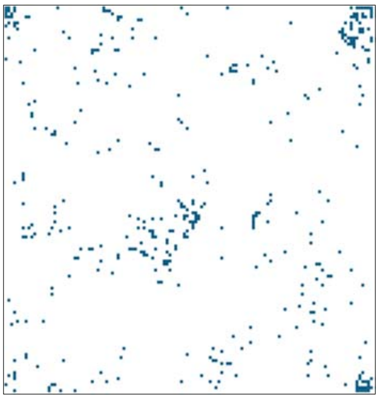
Microsoft

```
[
  "microsoft",
  "software",
  "server",
  "hardware",
  "products",
  "applications",
  "oems",
  "windows",
  "tools",
  "online"
]
```



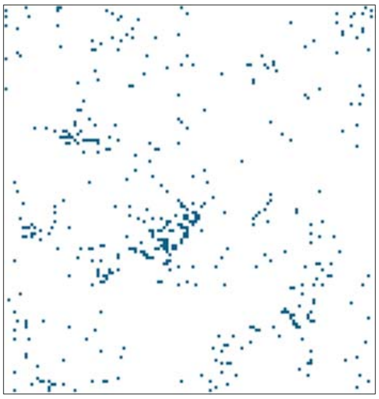
Intel

```
[
  "intel",
  "products",
  "computing",
  "processor",
  "microprocessors",
  "software",
  "devices",
  "customers",
  "smartphones",
  "processors"
]
```



United Technologies

```
[
 'pratt',
 'whitney',
 'sales',
 'aerospace',
 'businesses',
 'sikorsky',
 'products',
 'contracts',
 'controls',
 'subject'
]
```



Boeing

```
[
 'airborne',
 'customers',
 'costs',
 'customer',
 'contracts',
 'products',
 'markets',
 'boeing',
 'statements',
 'subject'
]
```

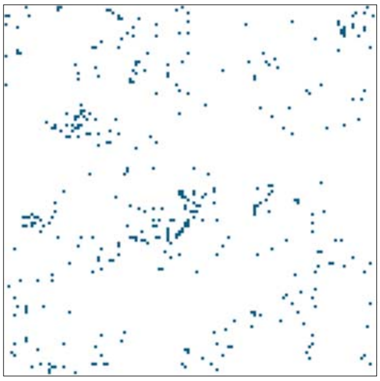


Table 1. Descriptive statistics

This table shows our sample of DJIA constituents. For each company, in columns 1-4, it gives the name, ticker, GICS sector and the fiscal year end month of the company's 2012 10-K report on which textual similarities are based. For every pair of firms, we obtain the textual similarity measure based on semantic fingerprinting, s^{SF}; the textual similarity measure based on word lists from Hoberg and Phillips (2010, 2015a, 2015b), s^{HP}; and the correlation between the companies' 2013 daily stock returns, ρ. For each sample company, columns 5-7 show the ticker of the most similar company according to s^{SF}, s^{HP}, and ρ measures, respectively. Columns 8-10 show, for each company *i*, the average values of $s^{SF}(i,j)$ $s^{HP}(i,j)$ and ρ_{ij}, respectively, across all companies *j* that are in the same GICS sector as company *i*. Columns 11-12 show the average values of $s^{SF}(i,j)$ $s^{HP}(i,j)$ and ρ_{ij}, respectively, across all companies *j* whose GICS sector is different from that of company *i*. Lastly, column 14 (respectively, 15) shows, for each company *i*, the correlation coefficient of ρ with s^{SF} (respectively, s^{HP}) averaged across all other companies in the sample.

				Closest firm based on			Mean for same-sector firms			Mean for other-sector firms		
Company	Ticker	GICS sector	FYE month	s^{SF}	s^{WL}	ρ	s^{SF}	s^{WL}	ρ	s^{SF}	s^{WL}	ρ
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
3M	MMM	Industrials	December	DD	GE	UTX	0.063	0.053	0.511	0.115	0.013	0.414
ALCOA	AA	Materials	December	CVX	CVX	CAT	0.198		0.344	0.135		0.254
AMERICAN EXPRESS	AXP	Financials	December	TRV	JPM	UTX	0.473	0.097	0.480	0.205	0.016	0.364
AT&T	T	Telecommunications Services	December	VZ	VZ	VZ	0.422	0.173	0.673	0.134	0.018	0.333
BANK OF AMERICA	BAC	Financials	December	JPM	JPM	JPM	0.466	0.029	0.559	0.108	0.009	0.312
BOEING	BA	Industrials	December	UTX	UTX	UTX	0.137	0.062	0.368	0.088	0.015	0.262
CATERPILLAR	CAT	Industrials	December	GE	GE	MMM	0.158	0.024	0.391	0.103	0.010	0.305
CHEVRON	CVX	Energy	December	AA	AA	XOM	0.365	0.062	0.724	0.133	0.015	0.399
CISCO SYSTEMS	CSCO	Information Technology	July	HPQ	VZ	GE	0.444	0.163	0.256	0.095	0.021	0.220
COCA-COLA	KO	Consumer Staples	December	PG	PFE	PG	0.326	0.033	0.472	0.174	0.019	0.351
DISNEY	DIS	Consumer Discretionary	September	T	MSFT	UTX	0.115	0.027	0.438	0.093	0.010	0.396
DU PONT	DD	Materials	December	KO		UTX	0.198		0.344	0.095		0.392
EXXON MOBIL	XOM	Energy	December	CVX	BA	CVX	0.365	0.062	0.724	0.129	0.018	0.388
GENERAL ELECTRIC	GE	Industrials	December	JPM	UTX	MMM	0.175	0.065	0.442	0.124	0.012	0.343
HEWLETT-PACKARD	HPQ	Information Technology	December	CSCO	CSCO	CSCO	0.475	0.155	0.217	0.076	0.013	0.118
HOME DEPOT	HD	Consumer Discretionary	January	IBM	MSFT	UTX	0.088	0.027	0.388	0.095	0.015	0.310
IBM	IBM	Information Technology	December	CSCO	MSFT	MCD	0.319	0.138	0.240	0.107	0.023	0.261
INTEL	INTC	Information Technology	December	MSFT	CSCO	AXP	0.416	0.140	0.252	0.089	0.012	0.284
JOHNSON & JOHNSON	JNJ	Health Care	December	MRK	MRK	PFE	0.494	0.110	0.448	0.075	0.015	0.391
JPMORGAN CHASE	JPM	Financials	December	BAC	AXP	BAC	0.478	0.119	0.565	0.099	0.008	0.363
MCDONALD'S	MCD	Consumer Discretionary	December	KO	KO	KO	0.119	0.028	0.356	0.115	0.015	0.317
MERCK	MRK	Health Care	December	JNJ	JNJ	PFE	0.462	0.132	0.289	0.071	0.013	0.256
MICROSOFT	MSFT	Information Technology	June	INTC	CSCO	DD	0.412	0.164	0.212	0.066	0.026	0.206
PFIZER	PFE	Health Care	December	UNH	MRK	JNJ	0.504	0.119	0.419	0.154	0.018	0.339
PROCTER & GAMBLE	PG	Consumer Staples	June	KO	WMT	T	0.361	0.073	0.509	0.142	0.009	0.310
TRAVELERS	TRV	Financials	December	AXP	AXP	MMM	0.462	0.066	0.502	0.180	0.009	0.395
UNITED TECHNOLOGIES	UTX	Industrials	December	BA	BA	MMM	0.215	0.093	0.538	0.087	0.013	0.411
UNITEDHEALTH	UNH	Health Care	December	PFE	PFE	JPM	0.424	0.060	0.276	0.115	0.013	0.239
VERIZON	VZ	Telecommunications Services	December	CSCO	CSCO	T	0.422	0.173	0.673	0.124	0.029	0.266
WAL-MART	WMT	Consumer Staples	January	PG	PG	PG	0.251	0.086	0.402	0.100	0.014	0.280
					Average		0.327	0.090	0.434	0.114	0.015	0.316

Table 2. Correlations between firm similarity measures.

This table shows correlations between the semantic fingerprint-based similarity measure, s^{SF}, Hoberg and Phillips' word list-based measure s^{WL}, Kaustia and Rantala's (2013) common-analyst similarity measure, s^{CA}, and the Lee, Ma and Wang (2015) search-based measure, s^{SB}, as well as pairwise correlations for 2014, 2013, and 2012.

	s^{SF}	s^{WL}	s^{CA}	s^{SB}	ρ_{2014}	ρ_{2013}	ρ_{2012}
s^{SF}	1.000						
s^{WL}	0.665	1.000					
s^{CA}	0.386	0.443	1.000				
s^{SB}	0.474	0.556	0.817	1.000			
ρ_{2014}	0.303	0.284	0.384	0.374	1.000		
ρ_{2013}	0.275	0.070	0.266	0.226	0.535	1.000	
ρ_{2012}	0.169	0.154	0.286	0.329	0.552	0.549	1.000

Table 3. Semantic fingerprint-based cosine similarity measures for all pairs of DJIA stocks.

Highlighted cells represent pairs of stocks in the same GICS sector.

	MMM	AA	AXP	T	BAC	BA	CAT	CVX	CSCO	KO	DIS	DD	XOM	GE	HPQ	HD	IBM	INTC	JNJ	JPM	MCD	MRK	MSFT	PFE	PG	TRV	UTX	UNH	VZ	WMT
MMM	1.000	0.128	0.149	0.172	0.056	0.032	0.059	0.081	0.160	0.156	0.082	0.184	0.087	0.059	0.149	0.095	0.147	0.153	0.142	0.042	0.079	0.105	0.073	0.105	0.137	0.115	0.102	0.076	0.139	0.069
AA	0.128	1.000	0.208	0.111	0.062	0.151	0.221	0.512	0.052	0.227	0.097	0.198	0.322	0.115	0.030	0.138	0.036	0.097	0.074	0.069	0.124	0.093	0.011	0.089	0.089	0.265	0.206	0.058	0.085	0.108
AXP	0.149	0.208	1.000	0.235	0.414	0.169	0.186	0.216	0.201	0.340	0.159	0.097	0.239	0.232	0.156	0.139	0.238	0.164	0.204	0.420	0.204	0.159	0.156	0.342	0.274	0.585	0.140	0.319	0.189	0.218
T	0.172	0.111	0.235	1.000	0.156	0.121	0.019	0.114	0.239	0.124	0.391	0.009	0.123	0.056	0.195	0.056	0.160	0.239	0.072	0.105	0.136	0.067	0.170	0.116	0.117	0.155	0.104	0.111	0.422	0.089
BAC	0.056	0.062	0.414	0.156	1.000	0.131	0.135	0.074	0.069	0.212	0.012	0.000	0.145	0.323	0.045	0.000	0.089	0.017	0.129	0.597	0.152	0.071	0.033	0.271	0.187	0.387	0.070	0.257	0.034	0.082
BA	0.032	0.151	0.169	0.121	0.131	1.000	0.062	0.149	0.092	0.069	0.043	0.029	0.100	0.078	0.078	0.056	0.077	0.060	0.047	0.099	0.083	0.060	0.048	0.097	0.059	0.145	0.377	0.119	0.060	0.068
CAT	0.059	0.221	0.186	0.019	0.135	0.062	1.000	0.210	0.028	0.161	0.018	0.179	0.179	0.346	0.064	0.069	0.126	0.062	0.023	0.122	0.089	0.046	0.024	0.111	0.158	0.164	0.166	0.042	0.062	0.084
CVX	0.081	0.512	0.216	0.114	0.074	0.149	0.210	1.000	0.032	0.280	0.108	0.179	0.365	0.140	0.022	0.095	0.043	0.067	0.078	0.092	0.122	0.080	0.005	0.139	0.160	0.279	0.166	0.048	0.090	0.130
CSCO	0.160	0.052	0.201	0.239	0.069	0.092	0.028	0.032	1.000	0.049	0.123	0.079	0.042	0.042	0.534	0.136	0.407	0.418	0.034	0.040	0.056	0.027	0.416	0.061	0.071	0.099	0.024	0.100	0.462	0.050
KO	0.156	0.227	0.340	0.124	0.212	0.069	0.161	0.280	0.049	1.000	0.087	0.327	0.247	0.199	0.056	0.125	0.055	0.061	0.153	0.193	0.233	0.186	0.052	0.350	0.436	0.361	0.114	0.191	0.079	0.216
DIS	0.082	0.097	0.159	0.391	0.012	0.043	0.018	0.108	0.123	0.087	1.000	0.033	0.095	0.009	0.096	0.085	0.065	0.130	0.021	0.012	0.146	0.017	0.162	0.042	0.117	0.075	0.053	0.048	0.203	0.200
DD	0.184	0.198	0.097	0.009	0.000	0.029	0.179	0.179	0.079	0.327	0.033	1.000	0.184	0.036	0.030	0.065	0.088	0.035	0.085	0.000	0.190	0.120	0.000	0.146	0.271	0.118	0.062	0.020	0.011	0.078
XOM	0.087	0.322	0.239	0.123	0.145	0.100	0.179	0.365	0.042	0.247	0.095	0.184	1.000	0.142	0.040	0.079	0.093	0.061	0.062	0.139	0.175	0.059	0.046	0.167	0.178	0.231	0.106	0.093	0.078	0.099
GE	0.059	0.115	0.232	0.056	0.323	0.078	0.346	0.140	0.042	0.199	0.009	0.036	0.142	1.000	0.064	0.092	0.111	0.049	0.000	0.420	0.051	0.028	0.024	0.221	0.158	0.276	0.215	0.073	0.098	0.135
HPQ	0.149	0.030	0.156	0.195	0.045	0.078	0.064	0.022	0.534	0.056	0.096	0.030	0.040	0.064	1.000	0.116	0.384	0.497	0.019	0.023	0.043	0.039	0.483	0.039	0.041	0.081	0.042	0.088	0.309	0.028
HD	0.095	0.138	0.139	0.056	0.000	0.056	0.069	0.095	0.136	0.125	0.085	0.065	0.079	0.092	0.116	1.000	0.209	0.163	0.055	0.016	0.092	0.022	0.085	0.067	0.131	0.054	0.080	0.126	0.104	0.182
IBM	0.147	0.036	0.238	0.160	0.089	0.077	0.126	0.043	0.407	0.055	0.065	0.088	0.093	0.111	0.384	0.209	1.000	0.244	0.094	0.078	0.073	0.030	0.242	0.106	0.100	0.147	0.041	0.190	0.243	0.028
INTC	0.153	0.097	0.164	0.239	0.017	0.060	0.062	0.067	0.418	0.061	0.130	0.035	0.061	0.049	0.497	0.163	0.244	1.000	0.022	0.009	0.025	0.030	0.506	0.047	0.086	0.067	0.075	0.061	0.403	0.032
JNJ	0.142	0.074	0.204	0.072	0.129	0.047	0.023	0.078	0.034	0.153	0.021	0.085	0.062	0.000	0.019	0.055	0.094	0.022	1.000	0.121	0.053	0.551	0.021	0.501	0.137	0.241	0.030	0.432	0.007	0.030
JPM	0.042	0.069	0.420	0.105	0.597	0.099	0.122	0.092	0.040	0.193	0.012	0.000	0.139	0.420	0.023	0.016	0.078	0.009	0.121	1.000	0.099	0.058	0.000	0.292	0.119	0.415	0.070	0.229	0.035	0.083
MCD	0.079	0.124	0.204	0.136	0.152	0.083	0.089	0.122	0.056	0.233	0.146	0.190	0.175	0.051	0.043	0.092	0.073	0.025	0.053	0.099	1.000	0.067	0.070	0.147	0.209	0.188	0.121	0.084	0.057	0.167
MRK	0.105	0.093	0.159	0.067	0.071	0.060	0.046	0.080	0.027	0.186	0.017	0.120	0.059	0.028	0.039	0.022	0.030	0.030	0.551	0.058	0.067	1.000	0.017	0.504	0.105	0.222	0.056	0.332	0.018	0.056
MSFT	0.073	0.011	0.156	0.170	0.033	0.048	0.024	0.005	0.416	0.052	0.162	0.000	0.046	0.024	0.483	0.085	0.242	0.506	0.021	0.000	0.070	0.017	1.000	0.039	0.096	0.037	0.041	0.064	0.309	0.062
PFE	0.105	0.089	0.342	0.116	0.271	0.097	0.111	0.139	0.061	0.350	0.042	0.146	0.167	0.221	0.039	0.067	0.106	0.047	0.501	0.292	0.147	0.504	0.039	1.000	0.302	0.398	0.087	0.508	0.047	0.177
PG	0.137	0.089	0.274	0.117	0.187	0.059	0.158	0.160	0.071	0.436	0.117	0.271	0.178	0.158	0.041	0.131	0.100	0.086	0.137	0.119	0.209	0.105	0.096	0.302	1.000	0.273	0.073	0.106	0.070	0.286
TRV	0.115	0.265	0.585	0.155	0.387	0.145	0.164	0.279	0.099	0.361	0.075	0.118	0.231	0.276	0.081	0.054	0.147	0.067	0.241	0.415	0.188	0.222	0.037	0.398	0.273	1.000	0.154	0.297	0.076	0.157
UTX	0.102	0.206	0.140	0.104	0.070	0.377	0.166	0.166	0.024	0.114	0.053	0.062	0.106	0.215	0.042	0.080	0.041	0.075	0.030	0.070	0.121	0.056	0.041	0.087	0.073	0.154	1.000	0.063	0.085	0.116
UNH	0.076	0.058	0.319	0.111	0.257	0.119	0.042	0.048	0.100	0.191	0.048	0.020	0.093	0.073	0.088	0.126	0.190	0.061	0.432	0.229	0.084	0.332	0.064	0.508	0.106	0.297	0.063	1.000	0.054	0.083
VZ	0.139	0.085	0.189	0.422	0.034	0.060	0.062	0.090	0.462	0.079	0.203	0.011	0.078	0.098	0.309	0.104	0.243	0.403	0.007	0.035	0.057	0.018	0.309	0.047	0.070	0.076	0.085	0.054	1.000	0.075
WMT	0.069	0.108	0.218	0.089	0.082	0.068	0.084	0.130	0.050	0.216	0.200	0.078	0.099	0.135	0.028	0.182	0.028	0.032	0.030	0.083	0.167	0.056	0.062	0.177	0.286	0.157	0.116	0.083	0.075	1.000

Table 4. Predicting stock return correlations with semantic fingerprint-based similarity measures.

This table shows the results of predicting pairwise correlations between sample firms with semantic fingerprint-based firm similarity measures. Panel A uses past and current pairwise correlations as dependent and explanatory variables, while Panel B uses the Fisher transformation of these variables, together with the product the two firms' standard deviations, $\sigma(i,t\text{-}2)\sigma(j,t\text{-}2)$, as in Box (2015). The key explanatory variable in both panels is $s^{SF}(i,j,t\text{-}1)$, the semantic fingerprint-based cosine similarity measure between pairs of companies. The regressions are based on 406 observations, representing all different pairs of the 29 DJIA constituent stocks as of end-2012 (excluding Du Pont, as explained in the text). Coefficient estimates are followed by t-statistics and p-values in italics.

Panel A: The dependent variable is $\rho_{i,j}$.

	(1)			(2)			(3)			(4)			(5)		
Constant	0.208	*16.67*	*0.000*	0.168	*11.42*	*0.000*	0.316	*39.80*	*0.000*	0.199	*15.97*	*0.000*	0.153	*10.51*	*0.000*
$s^{SF}(i,j,t\text{-}1)$							0.283	*6.39*	*0.000*	0.158	*3.94*	*0.000*	0.201	*5.29*	*0.000*
$\rho(i,j,t\text{-}1)$	0.462	*12.71*	*0.000*							0.422	*11.36*	*0.000*			
$\rho(i,j,t\text{-}2)$				0.469	*13.29*	*0.000*							0.438	*12.63*	*0.000*
R-squared	0.286			0.304			0.092			0.312			0.349		

Panel B: The dependent variable is $z_{i,j}$.

	(1)			(2)			(3)			(4)			(5)		
Constant	0.024	*0.92*	*0.356*	0.154	*8.32*	*0.000*	0.329	*33.83*	*0.000*	0.031	*1.18*	*0.239*	0.136	*7.41*	*0.000*
$s^{SF}(i,j,t\text{-}1)$							0.359	*6.64*	*0.000*	0.110	*2.34*	*0.020*	0.242	*5.33*	*0.000*
$z(i,j,t\text{-}1)$	0.700	*16.68*	*0.000*							0.662	*14.75*	*0.000*			
$\sigma(i,t\text{-}1)\sigma(j,t\text{-}1)$	0.089	*7.78*	*0.000*							0.082	*7.02*	*0.000*			
$z(i,j,t\text{-}2)$				0.485	*14.45*	*0.000*							0.452	*13.68*	*0.000*
$\sigma(i,t\text{-}2)\sigma(j,t\text{-}2)$				0.009	*1.29*	*0.197*							0.009	*1.31*	*0.191*
R-squared	0.411			0.345			0.098			0.419			0.388		

Table 5. Predicting stock return correlations with different firm similarity measures

This table shows the results of a ‘horse race’ among different measures of proximity for pairs of companies used to predict their stock return correlations. In addition to our own semantic fingerprint-based similarity measure, s^{SF}, and Hoberg and Phillips’ word-list based measure s^{WL}, we also use the Kaustia and Rantala (2013) common-analyst similarity measure, s^{CA}, and the Lee, Ma and Wang (2015) search-based similarity measure, s^{SB}. Panel A uses past and current pairwise correlations as dependent and explanatory variables, while Panel B uses the Fisher transformation of these variables, together with the product the two firms’ standard deviations, $\sigma(i,t\text{-}2)\sigma(j,t\text{-}2)$, as in Box (2015). The key explanatory variable in both panels is $s^{SF}(i,j,t\text{-}1)$, the semantic fingerprint-based cosine similarity measure between pairs of companies. The regressions are based on 406 observations, representing all different pairs of the 29 DJIA constituent stocks as of end-2012 (excluding Du Pont, as explained in the text). Coefficient estimates are followed by t-statistics and p-values in italics.

Panel A: The dependent variable is $\rho_{i,j}$.

	(1)			(2)			(3)			(4)			(5)		
Constant	0.153	*10.51*	*0.000*	0.166	*11.55*	*0.000*	0.187	*12.87*	*0.000*	0.185	*12.61*	*0.000*	0.171	*11.28*	*0.000*
$s^{SF}(i,j,t\text{-}1)$	0.201	*5.29*	*0.000*										0.114	*2.29*	*0.023*
$s^{WL}(i,j,t\text{-}1)$				0.587	*4.99*	*0.000*							0.177	*1.08*	*0.279*
$s^{CA}(i,j,t\text{-}1)$							0.326	*5.93*	*0.000*				0.282	*3.12*	*0.002*
$s^{SB}(i,j,t\text{-}1)$										2.437	*5.08*	*0.000*	-0.542	*-0.64*	*0.521*
$\rho(i,j,t\text{-}2)$	0.438	*12.63*	*0.000*	0.442	*12.74*	*0.000*	0.409	*11.57*	*0.000*	0.408	*11.26*	*0.000*	0.405	*11.41*	*0.000*
R-squared	0.349			0.345			0.360			0.346			0.381		
AIC(Akaike'sIC)	-823.9			-820.9											

Panel B: The dependent variable is $z_{i,j}$.

	(1)			(2)			(3)			(4)			(5)		
Constant	0.136	*7.41*	*0.000*	0.151	*8.37*	*0.000*	0.181	*10.03*	*0.000*	0.177	*9.66*	*0.000*	0.163	*8.65*	*0.000*
$s^{SF}(i,j,t\text{-}1)$	0.242	*5.33*	*0.000*										0.138	*2.34*	*0.020*
$s^{WL}(i,j,t\text{-}1)$				0.691	*4.90*	*0.000*							0.143	*0.74*	*0.460*
$s^{CA}(i,j,t\text{-}1)$							0.443	*6.72*	*0.000*				0.392	*3.67*	*0.000*
$s^{SB}(i,j,t\text{-}1)$										3.249	*5.62*	*0.000*	-0.526	*-0.53*	*0.598*
$z(i,j,t\text{-}2)$	0.452	*13.68*	*0.000*	0.457	*13.77*	*0.000*	0.409	*12.09*	*0.000*	0.412	*11.82*	*0.000*	0.405	*11.89*	*0.000*
$\sigma(i,t\text{-}2)\sigma(j,t\text{-}2)$	0.009	*1.31*	*0.191*	0.009	*1.31*	*0.191*	0.008	*1.19*	*0.235*	0.009	*1.24*	*0.217*	0.008	*1.22*	*0.223*
R-squared	0.388			0.382			0.411			0.393			0.428		
AIC(Akaike'sIC)	-681.1			-676.9											