CENTER FOR
ADVANCED
RESEARCH
IN FINANCE

**CARF Working Paper**

CARF-F-449

# Predicting Adverse Media Risk using a Heterogeneous Information Network

Ryohei Hisano
Social ICT Research Center
Graduate School of Information Science and Technology
The University of Tokyo

Didier Sornette
ETH Zürich
Department of Management Technology and Economics

Takayuki Mizuno
National Institute of Informatics

November, 2018

CARF is presently supported by The Dai-ichi Life Insurance Company, Limited, Nomura Holdings, Inc., Sumitomo Mitsui Banking Corporation, MUFG Bank, The Norinchukin Bank, The University of Tokyo Edge Capital Co., Ltd., Finatext Ltd., Sompo Holdings, Inc., Tokio Marine & Nichido Fire Insurance Co., Ltd. and Nippon Life Insurance Company. This financial support enables us to issue CARF Working Papers.

CARF Working Papers can be downloaded without charge from:
https://www.carf.e.u-tokyo.ac.jp/research/

Working Papers are a series of manuscripts in their draft form. They are not intended for circulation or distribution except as indicated by the author. For that reason Working Papers may not be reproduced or distributed without the written consent of the author.

# Predicting Adverse Media Risk using a Heterogeneous Information Network

Ryohei Hisano
Social ICT Research Center
Graduate School of Information Science and Technology
The University of Tokyo

Didier Sornette
ETH Zürich
Department of Management Technology and Economics

Takayuki Mizuno
National Institute of Informatics

## Abstract

The media plays a central role in monitoring powerful institutions and identifying any activities harmful to the public interest. In the investing sphere constituted of 46,583 officially listed domestic firms on the stock exchanges worldwide, there is a growing interest "to do the right thing", i.e., to put pressure on companies to improve their environmental, social and government (ESG) practices. However, how to overcome the sparsity of ESG data from non-reporting firms, and how to identify the relevant information in the annual reports of this large universe? Here, we construct a vast heterogeneous information network that covers the necessary information surrounding each firm, which is assembled using seven professionally curated datasets and two open datasets, resulting in about 50 million nodes and 400 million edges in total. Exploiting this heterogeneous information network, we propose a model that can learn from past adverse media coverage patterns and predict the occurrence of future adverse media coverage events on the whole universe of firms. Our approach is tested using the adverse media coverage data of more than 35,000 firms worldwide from January 2012 to May 2018. Comparing with state-of-the-art methods with and without the network, we show that the predictive accuracy is substantially improved when using the heterogeneous information network. This work suggests new ways to consolidate the diffuse information contained in big data in order to monitor dominant institutions on a global scale for more socially responsible investment, better risk management, and the surveillance of powerful institutions.

## 1 Introduction

Adverse media coverage sometimes leads to fatal results for a company. In the press release sent out by Cambridge Analytica on May 2, 2018, the company wrote that "Cambridge Analytica has been the subject of numerous unfounded accusations, ... media coverage has driven away virtually all of the company's customers and suppliers" [5]. This is just one recent example highlighting the impact of adverse media coverage on a firm's fate. In another example, the impact of adverse media coverage on Swiss bank profits was estimated to be 3.35 times the median annual net profit of small banks and 0.73 times that of large banks [3]. These numbers are significant, indicating how adverse media coverage can cause huge damage to a bank. Moreover, a new factor, priced as the "no media coverage premium" [10], has been identified to help explain financial returns: stocks with no media coverage earn higher returns than stocks with high media coverage. Within the rational-agent framework, this may result from impediments to trade and/or from lower investor recognition leading to lower diversification [10]. Another mechanism could be associated with the fact that most of the coverage of mass media is negative [15, 23].

On one side of the spectrum, the dominance of negative news is pushed by readers and the human psychology of impression formation [24] and loss aversion [17], which values negative information more than positive information. Thus, consumers buy news and magazines with negative information more than those with positive information [22]. On the other side of the spectrum, the principal role of the media in a liberal democratic society is to hold powerful institutions, such as the

| Date | Name | Adverse Media Label |
|---|---|---|
| Jan 3 2012 | Firm A | Management |
| Jan 3 2012 | Firm B | Product/Service |
| Jan 10 2012 | Firm C | Regulatory |

Table 1: Sample of the Dow Jones Adverse Media Entity dataset.

government and firms, accountable. This role is fulfilled by identifying any problems that they potentially have. In other words, the media act as the "Fourth Estate" [6], fulfilling their watchdog role of monitoring the dominant institutions in control of our society. Hence, predicting an adverse media coverage event is important, not only because of the direct adverse effects on a firm but also from the viewpoint of the watchdog role of the press. Furthermore, international institutions such as the OECD are advocating for more socially responsible investment, where an investor has the responsibility of divesting from firms that might have unwanted problems such as fraud and environmental issues [19]. Therefore, it is desirable, not only for investors' interests but for the sake of the public, to monitor and stop global money from flowing into unethical global firms.

Here, our focus is precisely to predict the future adverse media coverage of firms worldwide. Given the size of the sample (46,583 officially listed domestic companies worldwide in 2017 [20]) and the huge magnitude as well as highly heterogeneous nature of the data, organizing this information requires the development of machine-assisted methods. We thus construct a heterogeneous information network that gathers and combines information for firms worldwide. We merge curated data from several sources and store all the information in one heterogeneous information network (with 50 million nodes and 400 million edges in total). We introduce a variation of the label propagation method with edge weight learning using past media label patterns as supervisory signals, and the occurrence of relation types along path segments in the heterogeneous information network as features. We compare our method with a state-of-the-art method with and without using the network and show that our model obtains notably better performance in a prediction task, and is also interpretable.

There are many studies in computer science, dating back to at least the 1980s, on building a heterogeneous information network by gathering information from various sources [11]. More recent prominent work includes the network of Google [9] and Wikipedia's DBpedia [2], which are used for search engine optimization. Using web-based data, these databases are expanding rapidly. Some researchers even claim that the knowledge graph should be the default data model for learning heterogeneous knowledge [28]. However, the social impact of creating such a database beyond search engine optimization is not yet known. Our work provides positive support for this argument, showing that information concerning firms worldwide could be mapped into one heterogeneous information network and a machine-assisted method can learn patterns that can predict the occurrence of future adverse media coverage.

From the viewpoint of management science, our results can be interpreted as defining an adverse media risk score. Studies on management and finance show that firms could improve media coverage using firm-initiated press releases and investor relations firms [21, 1]. Moreover, recent studies show that firms engaging in corporate social responsibility or environmental social governance activities receive better media coverage [12] and this relationship between corporate social responsibility and positivity in media coverage is stronger for firms in controversial industries [4]. By understanding where the searchlight of a watchdog is aimed at, firms can determine the necessity to communicate with the public for fair coverage.

## Datasets

### Adverse Media Coverage

We use the Dow Jones Adverse Media Entity data from Jan 2012 to May 2018 as our basic data. The data consists of the name of the firm, date of the media coverage, and 17 categories classifying the adverse media coverage. Table 1 shows a sample of the dataset. Further details are provided in the Supplementary Information (SI).

We first test whether adverse media coverage has a financial impact by checking its relationship with a cross-section of the returns. We do this using the following steps. For all US stocks in the dataset, we gather their prices from Jan 2012 to May 2018. There are 1,139 such stocks in total. For each date in the adverse media label list, we employ a 10-day window centered on the specified date. We then calculate the log return between the start and end dates of the 10-day window, and we compare these returns with the ten trading day log returns outside such windows.

Table 2 compares the distributions of stock returns with and without adverse media coverage. Normalized histogram and rug plots are also provided in Fig. S2. The quantiles and skewness show that the negative tail of the log-return distribution is

| Group | Samples | 0.01 | 0.05 | 0.5 | 0.95 | 0.99 | Skewness |
|---|---|---|---|---|---|---|---|
| News | 8685 | -0.233 | -0.102 | 0.005 | 0.098 | 0.191 | -6.521 |
| Rest | 1667616 | -0.218 | -0.109 | 0.005 | 0.110 | 0.207 | 0.165 |

Table 2: Comparison of 10 trading day log returns with and without news events. Numbers indicate quantiles given in the first row.

| Source | Date of Acquisition | Node types | Relation types | Num Nodes | Num Edges |
|---|---|---|---|---|---|
| Dow Jones Adverse Media Entity | Dec 2016 | Firm | Location, Homepage | 132,127 | 390,320 |
| Dow Jones State Owned Companies | Dec 2016 | State Owned Firms | VIP, Employee, Owner | 280,995 | 702,172 |
| Dow Jones Watchlist | Dec 2016 | VIPs, Specially interested person | Social relations | 1,826,273 | 8,322,560 |
| Capital IQ Company Screening Report | Dec 2016 | Firms | Buyer-Seller, Borrower etc | 505,789 | 2,916,956 |
| FactSet | Dec 2015 | Firm, Goods, Industry | Parent-child firm, Issue Stock | 613,422 | 8,213,225 |
| FactShip | Jan 2017 | Firm, Goods, Invoice etc | Overseas trade etc | 16,137,550 | 36,345,381 |
| Reuters Ownership | Dec 2016 | Owners, Stocks | Issue, Own | 1,560,544 | 121,769,151 |
| Panama papers | Jan 2017 | Entities, Officers | Shareholder of, Director of | 888,630 | 1,371,984 |
| DBpedia | Apr 2016 | Various | Various | 35,006,127 | 249,429,771 |

Table 3: Summary of the dataset used in this study

more stretched than the positive tail, which agrees with previous studies that argue that negative information has a negative impact on financial returns. We also performed a two-sample Kolmogorov-Smirnov test for the null hypothesis that the two distributions are from the same distribution. This was rejected with a p-value below $10^{-6}$. Case studies of adverse media labels focusing on the top-4 negative returns are described in the SI.

## Heterogeneous information network

Besides the label information, the Dow Jones Adverse Media Entity data contains basic information on the location and homepage of each firm. However, this information is not sufficient to predict the labels. Hence, our strategy is to assemble data from other professional curated sources in the form of a heterogeneous information network. Table 3 summarized the dataset used in the paper.

We note several points about the data. First, in order to remove duplicates when combining node information from several sources, we did not just consider the name of the firm. In addition to name similarity, we determined two firms from different datasets to be the same if any of the following information was precisely the same: i) their homepage information, ii) the longitude and latitude information of their addresses (found using Google Place API), or iii) their stock symbol. We manually inspected our strategy and found that it leads to smaller "false positive" errors (i.e., incorrectly identifying different nodes as duplicates), but larger "false negative" errors (i.e., missing nodes that are duplicates). This stems from the fact that we could not remove duplicate firms that do not have a homepage, address, or stock symbol information.

Second, half of the relational information in our datasets does not include a timestamp. This is problematic in the sense that it is difficult to ensure that no future information is used when we perform our prediction. To avoid any information from the future contaminating our heterogeneous information network, we only predict the future occurrences of adverse media coverage after Feb 1, 2017, which occurs after the latest date at which we acquired data (Table 3). Finally, for the relational information in the Dow Jones Adverse Media Entity dataset, we used the Dec 2016 version and updated only the media label information to May 2018.

We also removed relation types that appear too many times in our dataset, in order to avoid computational overload. These relation types include "http://dbpedia.org/ontology/wikiPageWikiLink" and "http://purl.org/dc/terms/subjects", which create about 175 million and 22 million edges, respectively. We also ignored relation types that only appear in the dataset less than 100 times. Furthermore, some of the edges in our dataset have multiple timestamps, and we unified them into one relationship. These include relation types such as "own stocks" and "sends goods", of which the former are on a quarterly basis while the latter includes the timestamp information of when they passed through US customs. For "own stocks", we further restrict to relationships with at least 5% ownership. After the removal of duplicates and data cleaning, a total of about 3.7 million nodes and 9.1 million edges with 216 relation types remained. Table 4 shows ten examples of relations from the top-25 relation types found in our dataset (the full table is provided in Table S2). There are many relation types connecting the firms, but there are also relation types such as those for (i) associations and employees, which relate firms to people, (ii)

| Rank | Relation | Number |
|---|---|---|
| 2 | customer | 717,019 |
| 4 | own_stock | 493,316 |
| 5 | belongs_to_industry | 359,425 |
| 8 | receive_goods | 330,311 |
| 10 | issue_stock | 187,498 |
| 11 | make_products | 181,574 |
| 13 | part_of_industry | 172,621 |
| 15 | domain | 131,153 |
| 19 | associated-person | 100,699 |

Table 4: Selected examples of the top 25 relation types.

own stocks, which relates firms or individuals to a stock symbol, and (iii) domain, which relates firms and individuals to a homepage.

We restricted our prediction targets to firms that are found in the Dow Jones Adverse Media Entity dataset and for which we have at least one item of relational information among them. We call the network of our prediction target the core network. The core network is a weighted undirected network $G = (V, E, W)$ consisting of a set of nodes V, a set of edges E, and edge weights W. We assume that there is an edge between two nodes in the core network if there is at least one relation type connecting the two nodes. There are 35,657 firms with 322,138 edges in the core network. In Fig. S4, we also show a scatterplot of the longitude and latitude information of their addresses, indicating that our prediction targets are scattered worldwide. We restrict our attention to the core network because we only have limited information for firms outside of this network. Restricting our prediction target to the core network strikes a reasonable balance between improving the "reach" [26] of our prediction while assuring that we have enough information for prediction.

## Model

### Label propagation model

Using the core network defined in the last section, we define a non-negative weight function $f_\theta : X \to [0, 1]$, where $X$ defines the set of features for an edge $i, j$ extracted from the heterogeneous information network. We define $f_\theta$ to be a simple multilayer perceptron with 30 hidden units and a sigmoid layer for our output function. In addition, $\theta$ denotes the parameters of the model.

We combined the core network defined above with the adverse media labels using a slight variation of label propagation with Jacobi iteration [7]. Our strategy is to split the nodes into the source and target nodes depending on the date of the last adverse media coverage. We trained our model to minimize the loss of predicting the labels of our target nodes. The exact steps connecting $X$, the set of features for an edge $i, j$, to the loss is described in algorithm 1. Note that our model is not exactly a label propagation model because we set $D_{ii} = \Sigma_j 1_{ij\in E}$ instead of $D_{ii} = \Sigma_j w_{ij}$. The diagonal dominance condition [7] that ensures the Jacobi iteration will converge still holds because $\Sigma_j 1_{ij\in E} \geq \Sigma_j w_{ij}$, which stems from the fact that we defined $0 \leq w_{ij} \leq 1$. Note that our model is exactly equivalent to the usual label propagation when all $w_{ij}$ equals 1, but after learning the edge weights, the spectral radius of $A^{-1}W$ becomes smaller than the usual label propagation, leading the model to focus on propagating the labels to nearby nodes. A histogram of the learned edge weights is shown in Fig. S6.

After learning the parameters of the model, we treat both the source and target nodes as known labels and predict the future occurrence of adverse media coverage events after the last date of the training data (i.e., Feb 1, 2017) to May 31, 2018. A schematic figure of this procedure is shown in Fig. S5. The duration that separates target nodes from source nodes in the training data was set to 31 days before the last date of the training data for most of the adverse media category types for which we have sufficient label information and 182 days for labels with less information (e.g., sanction, human, and association labels). In Table S5, we also show results varying the last date of the training data to Aug 1, 2017.

As is evident in the procedure, we restrict our attention to predicting only the first occurrence of adverse media coverage. Instead of using all the adverse media label information from Jan 1, 2012 for training, we could also restrict this set by varying the start date as well. However, as shown from a case study of the top-4 negative log returns in the SI, there are cases

**Algorithm 1** Slight Variation of Label Propagation

(1) For each edge in the core network set, $w_{ij} = f_\theta(x_{ij})$, where $x_{ij}$ denotes features from the network.
(2) Compute diagonal degree matrix $D$ by $D_{ii} = \Sigma_j 1_{ij \in E}$.
(3) Compute $A_{ii} = I_l(i) + D_{ii}$, where $I_l(i)$ indicates $i$'s known label.
(4) Initialize $Y^0 = (y_1, ..., y_l, 0, ..., 0)$, where $l$ is the number of known labels.
(5) Iterate $Y^{t+1} = A^{-1}(WY^t + Y^0)$ until convergence
(6) Calculate loss by taking the mean squared error of $Y^{target} = (y_{l+1}, ..., y_{l+m}, 0, ..., 0)$ and $Y^T = (y^T_{l+1}, ...)$.
(7) Update $\theta$ in $f_\theta$ using gradient descent.
(8) Repeat until convergence.

for which the same type of adverse media label repeatedly appears due to follow up articles on the same issue. It is not yet known how much time should pass before we can assume that there are no ongoing allegations. Moreover, predicting that firms that have had previous adverse media coverage will suffer from future adverse media coverage might be an easier task than predicting adverse media coverage for firms that have never been adversely covered before, and we want to exclude these cases as much as possible. Hence, we only use the starting date of our data in the present study, but in Table S6 we also report results obtained by varying the start date.

### Edge features

For our model to work, we need to define the features for each edge. We used either the occurrence of relation types in the core network, a path in the overall heterogeneous information network connecting the two nodes [18], or the relation types along path segments connecting the two nodes as our features. We respectively denote each model as LP-core-relation, LP-path, and LP-path-segment, where LP stands for "label propagation". Instead of using the raw number counts of each relation type or path, we use a binary indicator to describe whether a specific feature exists.

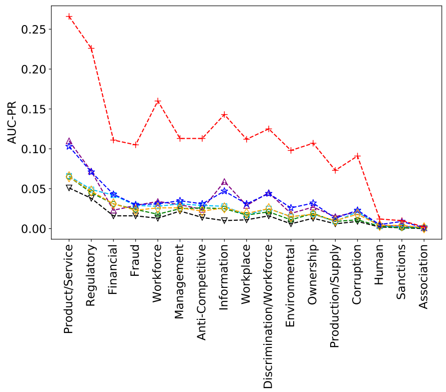
To be more specific, suppose that edge $A$, $B$ has the following two direct relations and two paths between them: (A,supplies,B), (A,strategic alliance,B), (A,is in,c,is in,B), and (A,makes,x,is made of,y,makes,B). In LP-core-relation, we only pay attention to (A, supplies, B) and (A,strategic alliance,B) and hence use $[0, ..., 1, 0, 1, 0...]$ as our feature, where the two 1's correspond to the supplies and strategic alliance relation types. The LP-path works similarly, but instead of creating a one-hot vector for each relation type, we create a one-hot vector for each path. We restrict our attention to the top-3,000 paths found with a length no larger than 4 for computational reasons. We also ignored the direction of each relation type.

Moreover, we discarded paths connecting two nodes that are already connected by shorter paths. Using our example above, paths with lengths 1 and 2 are not affected by this restriction but, starting from paths of length 3, there might be a path of length 3 such as (A,is in,c,alliance with,d,supports,B) that also connects A and B. We ignore these paths because node c already appeared in a path of length 2 (i.e., (A,is in,c,is in,B)). We use this additional restriction to keep super-nodes (such as industries) from contaminating our path features.

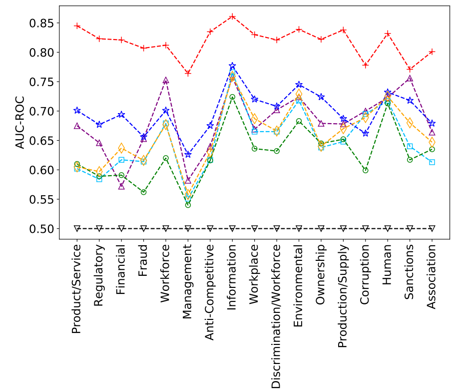
Features in LP-path-segment were created by distinguishing relation types occurring along the path segments. This can be thought of as a collapsed version of LP-path with relation-type one-hot vectors for each path segment. A naive implementation of this would result in 10 segments for path lengths of up to 4. However, because the core network is undirected, we can exploit the symmetry and reduce the number of segments. For example, there is no difference between starting a path from A or starting from B in (A, is in,c, is in, B). Hence, we do not need to distinguish path segments for paths of length 2, e.g., 2:1 and 2:2, but instead we could combine them, creating only one feature of path length 2. We use path lengths of up to 4, and there are six possible path segments in total, which we denote as 1, 2, 3:1, 3:2, 4:1, and 4:2.

### Other models compared

We compared our models against the following state-of-the-art methods, both using and not using the network. For the model that does not use the network, we added country and industry categories to Table 1, transformed it into one-hot vectors and used a random forest for classification. We call this model the "random forest". For a model that uses the network but not edge weight learning, we directly performed label propagation on the core network. We call this the "LP-fixed model". We further compared our method with methods that can incorporate multi-label correlation. Many previous studies combine multi-label correlation with label propagation [27]. However most of these methods are computationally very expensive, and hence we use the method of Ref. [27], which turned out to be computationally reasonable. However, Ref. [27] uses a KNN




(a) AUC-PR




(b) AUC-ROC

Figure 1: Comparison of predictive performance for random guessing (black inverted triangles), random forest (purple triangles), LP-fixed (light-blue squares), LP-mult (green circles), LP-core-relation (blue stars), LP-path (orange diamonds), and LP-path-segment (red crosses).

graph that is not available in our case. Instead, we use the core matrix, multiplying it by an additional parameter to ensure that the spectral radius of the entire matrix is below 1.

# Results

## Quantitative Comparison

Our prediction problem is a standard binary classification problem (whether there is adverse media coverage from Feb 1, 2017, to May 31, 2018, or not), so we used the area under the receiver operator characteristics (AUC-ROC) for evaluation. Because our labels are highly imbalanced, we also evaluated performance using the area under the precision-recall curves (AUC-PR) [8]. Because of space limitations, the results are shown in the form of graphics (see figure 1), while the full table is reported in Table S4.

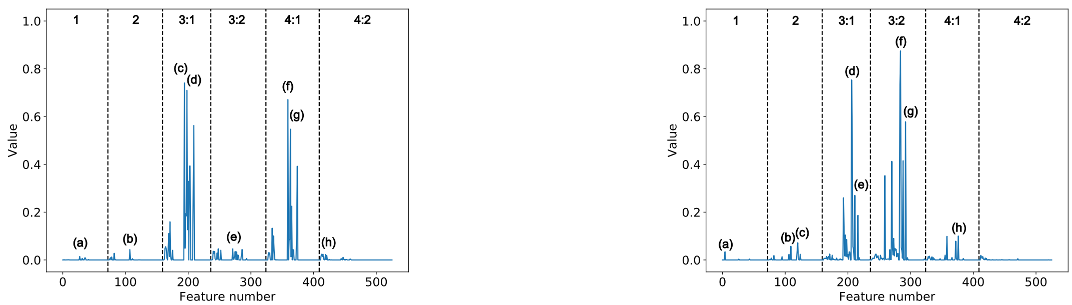
We first note that there seems to be predictability just by performing label propagation on the core network (i.e., LP-fixed). However, its performance is slightly worse than that of the random forest baseline using country and industry indicators. The performance of the network approach improves when the adaptive edge weighting scheme is used. This is apparent because LP-core-relation performs better than LP-fixed almost all the time. It is possible that LP-path performs less well than LP-core-relation because we only use the top-3,000 paths for computational reasons. LP-mult does not seem to improve performance when compared with LP-fixed. Whether this stems from the particular algorithm used or because not much information is added by incorporating multi-label correlation needs further investigation. Finally, comparing LP-path-segment to all the other methods, we find that it performs substantially better, outperforming all the methods for all the labels compared in this paper. We also provide a plot summarizing how the predictions of the methods differ in Fig. S7. In summary, our result shows that using the information stored in the heterogeneous information network leads to a substantially better predictive accuracy.

## Interpretability

To understand what our models have learned, we perform the partial dependency analysis on our learned model [16]. However, because the features used by LP-path-segment are highly correlated, calculating the importance measure for each feature might not be a reasonable approach. Hence, we first reduce the dimensionality of the feature space to 50 using a standard binary nonnegative matrix factorization (BNMF) technique [25] and then perform the usual partial dependence analysis along the basis of the matrix obtained by the standard BNMF method (see SI for a full description). The BNMF finds similar relation types among the different path segments that can be aligned together to make interpretation of the results possible. Usually, the sample standard deviation of the fitted values of the partial dependency plot is used as a measure of

| Rank | Basis | $E_{\hat{\theta}}[f(x_{0.99}) - f(x_{0.01})]$ | $\lvert E_{\hat{\theta}}[f(x_{0.99}) - f(x_{0.01})]\rvert$ |
|---|---:|---:|---:|
| 1 | 4 | -0.096 | 0.096 |
| 4 | 13 | 0.040 | 0.040 |

Table 5: Top positive and negative features among the top-five important features. Full list in Table S7.




(a) 1a

(b) 1b

Figure 2: Comparison of basis vector 4 and basis vector 13. The dotted vertical lines divide each path segment. Because there are relation types that does not appear in some path segments, the total number of features is 526 instead of 1,296 (216 $\times$ 6). Peaks in basis vector 4: (a) in-licensing, (b) in-licensing, (c) in-licensing, (d) out-licensing, (e) distributor, (f) in-licensing, (g) out-licensing, and (h) customer. Peaks in basis vector 13: (a) customer, (b) partner-manufactures, (c) international shipping (d) receive goods, (e) international shipping, (f) international shipping (g) receive goods, (h) franchise.

feature importance [14]. However, since our feature matrix is binary, we instead focus on the absolute difference of the response at the 0.99 and 0.01 quantile of the coefficients vector corresponding to each basis vector (see Fig. S8). We also take the average value of the importance measure repeating the training and partial dependency analysis step 30 times using different initial parameters to mitigate the effect of fluctuation stemming from the learning process.

Table 5 shows the top positive and negative basis among the top-five important features learned for the "Product/Service" label. We see that basis vector 4 seems to have the most negative effect while basis vector 13 seems to have the most positive effect on the weights. Note that features in higher path segments are likely to have a higher value in the basis vector because our feature matrix is a binary matrix taking one if there is at least one relation type in a particular path segment. Thus, we must pay attention to the relation type in each segment when interpreting the result and, in Fig. 2, we report the top relation types for each path segment for basis vector 4 and basis vector 13. While the path segments of basis vector 4 include more relation types that are related to the license relation, basis vector 13, which has a positive effect, focuses more on the buyer-seller and partnership–manufacture relations. Because "Product/Service" is more related to news about the specific products of a firm (see the case studies in SI), our model learned to value those relation type in the path segments more. Further details and analysis of the "Financial" label are provided in the SI.

## Discussion

Figure 1 and the SI have demonstrated a remarkable over-performance of our methods, which requires some explanation. First, when a problem occurs for a firm, it is likely that the firms it is related to or similar firms are also in trouble. The similarity of firms could be quantified by the closeness in the heterogeneous information network, which includes various information concerning a firm. Moreover, instead of using the raw closeness measure that our heterogeneous information network suggests, we adjust for the closeness measure using past adverse media label patterns, resulting in high predictive performance. Perhaps more importantly, when a problem catches the eye of the public, the media itself searches for nearby firms for follow-up stories. By doing so, they can claim that the first problem they reported is not just confined to one firm, but to a more general issue in need of more attention. Hence, it might not be surprising that our method works.

The misclassifications of our model can be organized into four categories, as shown in Table 6. The inaccuracy stemming from our model or data limitations could result in both false positive and false negative errors. There are exogenous events in false negatives that are impossible to predict from our approach of simply learning past adverse media coverage patterns. Exogenous events always constitute an intrinsic limit to prediction methods. However, on the positive side, there might be cases of false positive misclassifications that correspond to unrealized or uncovered events. From a journalist's point of view, the list of firms in this category might be the next possible target for further investigation. From a firm's point of view, this score might be a good diagnostic to follow to take timely actions for fair media coverage [4, 12].

Moreover, instead of using the media labels as the data vendor provides it, we could investigate further into the text to pick up news that had a significant impact (e.g., arrest, lawsuits) instead of just a shallow allegation. We could also take into account node information (e.g., firm size) to focus on firms that are too big to fail or the banking sector for which the effect of adverse media coverage is already well-known [3]. This might open another way to tackle the fight against fake news, and this is left for future work.

| | | Real | |
|---|---|---|---|
| | | False | True |
| Prediction | False | Correct | FN: Model error/Data limit<br>Exogeneous events |
| | True | FP: Model error/Data limit<br>Not realized/Not covered | Correct |

Table 6: Model prediction and the real world. FP stands for false positive and FN stands for false negative

## Acknowledgement

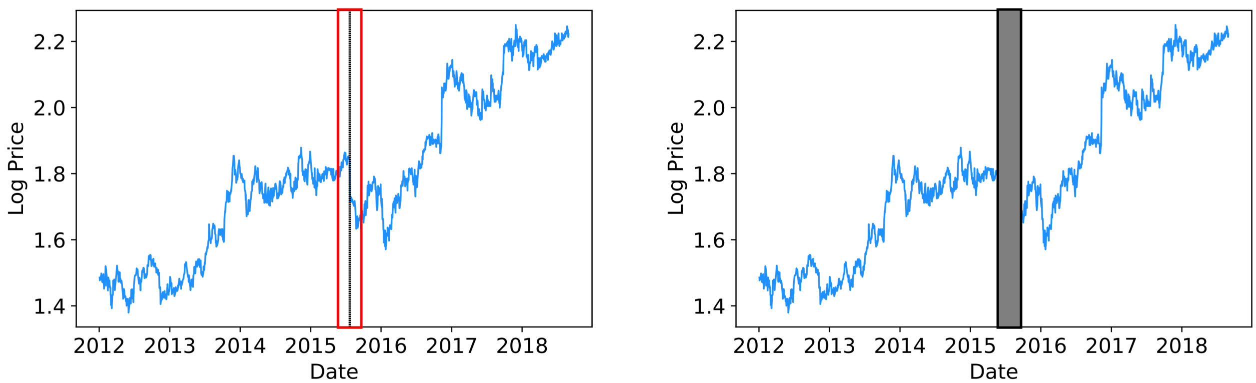
We would like to thank Tsutomu Watanabe, Takaaki Ohnishi, Hiroshi Iyetomi, Daisuke Inoue, Taichi Kiwaki and Kohei Miyaguchi for vigorous discussions. R.H. was supported by ACT-i, JST. T.M. was supported by Grant-in-Aid for Young Scientists (A) #16H05904, JSPS. Financial supports from CARF at the University of Tokyo is also acknowledged. We thank Kim Moravec, Ph.D., from Edanz Group for editing a draft of this manuscript.

## 2 Supplementary Material

## Materials

### Adverse media label and financial returns

In this section, we provide a more detailed account of how we investigated the relationship between adverse media labels and financial returns. We performed the analysis using the following steps. We first listed all firms with a US stock symbol in our dataset that had at least one adverse media coverage. Then we collected all the adjusted closing prices using Yahoo's API. We ignored stocks that had zero trading volume to focus on liquid stocks. This procedure resulted in a total of 1,139 stocks for investigation. For each date in the adverse media label list, we considered a 10 trading day window centered on the specified date, as shown in Fig. S1a. We take the log return for the start and end dates (the difference in the log price over 10 trading days) and classified this as a return with news. We compared the returns collected by this procedure with the 10-trading day log returns outside these windows, as shown in Fig. S1b.

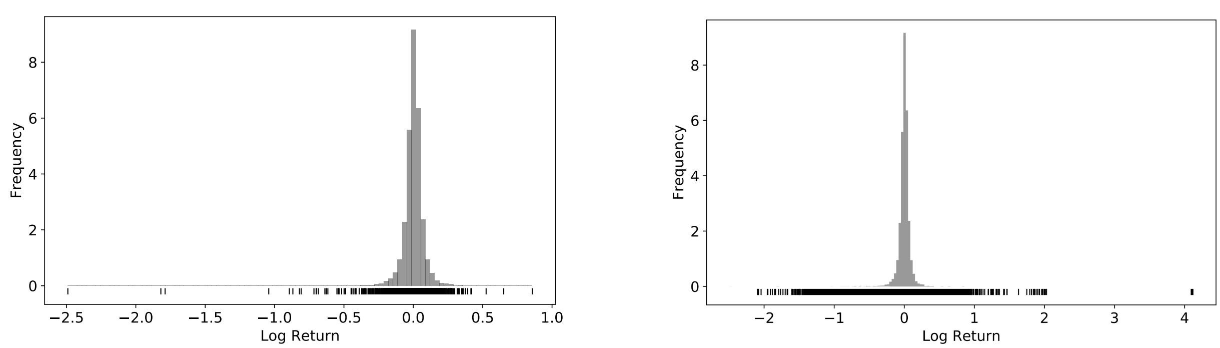



(a) Price movement with news

(b) Price movement without news

Figure S1: Separation of log returns with news events and without news events (the rest). The above figure is from TrustCo Bank Corp NY, and the dashed lines correspond to June 1, 2015.

In Fig. S2, we report normalized histograms with rug plots that show the difference of the distributions of log returns with news events and without news events (the rest) for the 1,139 stocks investigated in this study. We confirm that the negative tail of the log return distribution is more stretched than the positive tail, as shown in Table 2 in the main text.



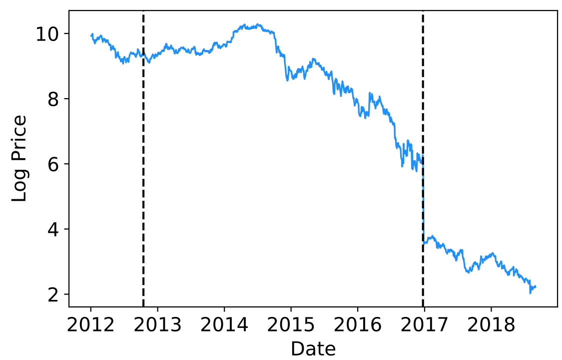

(a) With adverse media coverage

(b) Without adverse media coverage

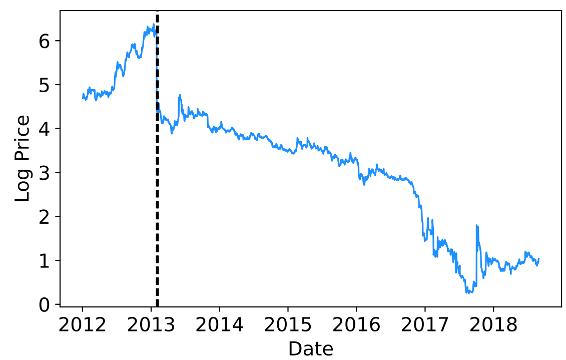
Figure S2: Normalized histogram of the 10 trading day log returns with rug plots.

## Case studies of the top-4 negative returns

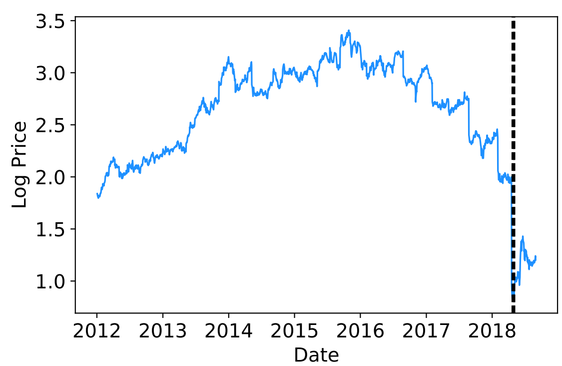
In Fig. S3, we show four case studies corresponding to the top four negative returns shown in the rug plot in Fig. S2 along with the adverse media label information corresponding to these negative returns (Table S1). Because these events are enormous falls in returns, performing a Google search readily provides insights into what happened to these firms. The Basic Energy Services fall in returns on Dec 23, 2016, corresponds to the date when they emerged from chapter 11 bankruptcy protection. In Celsion's case, there was a large amount of adverse reporting concerning an anti-cancer product called ThermoDox in early February 2013. For Aceto Agricultural Chemicals, May 3, 2018, was the date that the company announced that they would take proactive steps to address their business and financial challenges. In Ocwen Loan's case, the extreme fall occurred in April 20, 2017. The date corresponds to when the Consumer Financial Protection Bureau announced that they would sue Ocwen Loan for their subprime-related mortgage loan services. Note that the adverse media coverage connecting Ocwen Loan and subprime problems had been around (according to our dataset) from late 2013. Ocwen Loan's case is one example of how many follow up articles can appear on primarily the same issue.



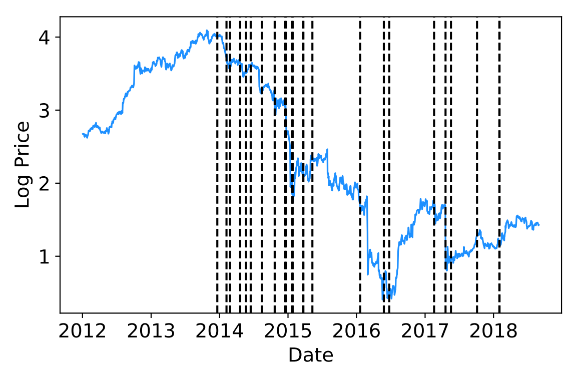

(a) Basic Energy Services




(b) Celsion




(c) Aceto Agricultural Chemicals




(d) Ocwen Loan Servicing

Figure S3: Time series of selected stock prices. The dashed lines correspond to dates when there was an adverse media coverage event.

## Top-25 relation types in our heterogeneous information network

In Table S2, we show the full list of the top-25 relation types in our heterogeneous information network.

| Firm | Label | Date |
|---|---|---|
| Basic Energy Services Inc | Regulatory | 2012/10/15 |
| Basic Energy Services Inc | Financial | 2016/12/23 |
| Celsion co | Ownership | 2013/2/6 |
| Celsion co | Product/Service | 2013/2/1 |
| Aceto Agricultural Chemicals co | Anti-Competitive | 2018/4/24 |
| Aceto Agricultural Chemicals co | Ownership | 2018/5/3 |
| Aceto Agricultural Chemicals co | Ownership | 2018/5/3 |
| Ocwen Loan Servicing llc | Regulatory | 2013/12/20 |
| ... | ... | ... |
| Ocwen Loan Servicing llc | Regulatory | 2017/4/20 |
| ... | ... | ... |
| Ocwen Loan Servicing llc | Discrimination/Workforce | 2018/2/2 |

Table S1: Part of the adverse media label data that corresponds to Fig. S3.

| Rank | Relation | Number | Description |
|---|---|---|---|
| 1 | located_in | 2,723,162 | relates firms to country |
| 2 | customer | 717,019 | buyer-seller relation |
| 3 | supplier | 713,434 | buyer-seller relation |
| 4 | own_stock | 493,316 | relates person or firm to stock ticker symbol |
| 5 | belongs_to_industry | 359,425 | relates firms to industry |
| 6 | strategic_alliance | 348,352 | relation among firms |
| 7 | creditor | 339,184 | relation among firms |
| 8 | recieve_goods | 330,311 | relates firms to goods received over US customs |
| 9 | send_goods | 319,292 | relates firms to goods sent over US customs |
| 10 | issue_stock | 187,498 | relates firms to stock ticker symbol |
| 11 | make_products | 181,574 | relates firms to goods |
| 12 | competitor | 174,487 | relation among firms |
| 13 | part_of_industry | 172,621 | relates industry to industry |
| 14 | borrower | 153,203 | relation among firms |
| 15 | domain | 131,153 | relates firms or person to homepage |
| 16 | distributor | 116,262 | relation among firms |
| 17 | subsidiary | 107,119 | relation among firms |
| 18 | parent-company | 107,117 | relation among firms |
| 19 | associated-person | 100,699 | relates firms to people |
| 20 | international_shipping | 95,050 | relation among firms |
| 21 | associate | 72,685 | relation among firms |
| 22 | landlord | 62,904 | relation among firms |
| 23 | http://dbpedia.org/ontology/party | 55,653 | relates people to party |
| 24 | employer | 47,901 | relates people to firms |
| 25 | employee | 47,184 | relates firms and people |

Table S2: Selected examples of the top 25 relation types.

### Location of firms investigated in this study

Figure S4 is a scatterplot showing the locations of the 35,657 firms investigated in this study. They are clearly scattered worldwide.

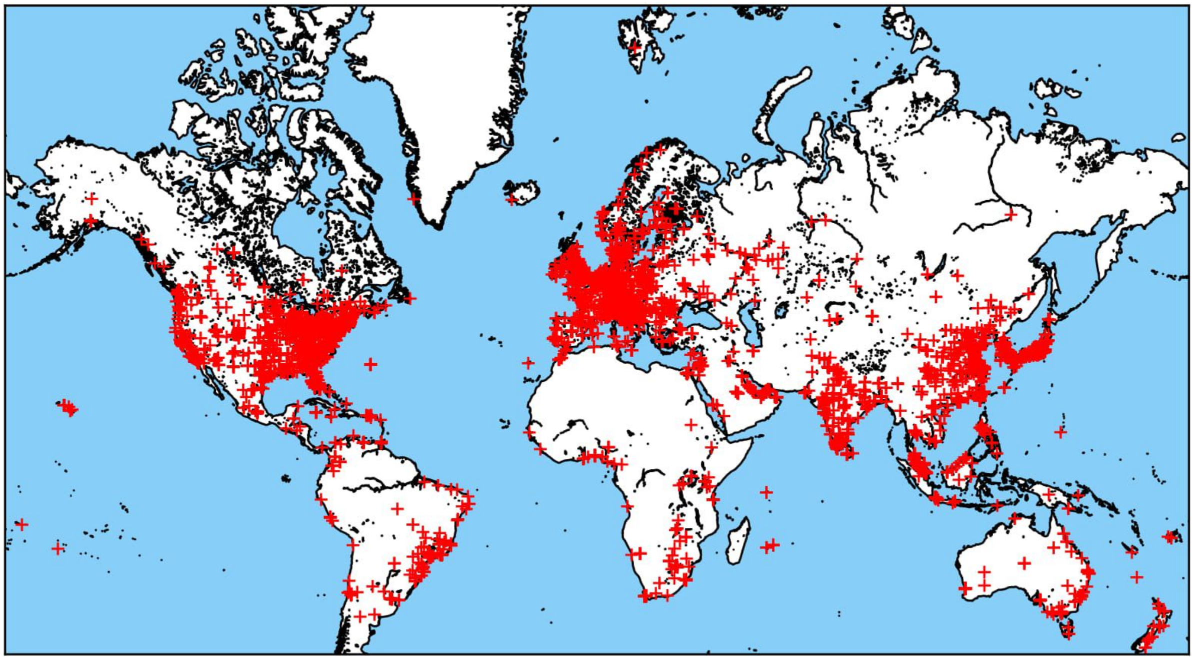


Figure S4: Scatterplot showing the longitudes and latitudes of the locations of firms investigated in this study.

### Adverse media label categories

In Table S3, we present the number of adverse media coverage for the 35,657 firms analyzed in this study from Jan 2012 to May 2018. "Raw count" denotes the total number of adverse media coverage for a particular adverse media category. "Unique firms" denotes the total number of unique firms tagged with a particular adverse media category at least once. In the table, "Raw count" is sometimes much higher than "Unique firms" which indicates that some firms are tagged with the same adverse media label multiple times.

## Methods

### Our approach

In Fig. S5, we provide a schematic figure describing our approach. The core network described in the main text corresponds to the network depicted in the center, and the rest of the heterogeneous information network is shown by a schematic icon for computational reasons. Our approach is to learn how to propagate labels by dividing past adverse media coverage patterns into source nodes and target nodes, where the duration that separates these two sets is either 31 days or 182 days (depending on the number of adverse media labels in the dataset) from the date that divides our training and test periods.

### Edge weights

In Fig. S6, we provide a normalized histogram showing the learned edge weights for LP-path-segment, where LP stands for Label Propagation, for predicting the Product/Service label. Our algorithm tends to separate edge weights into values of either 1 or 0.

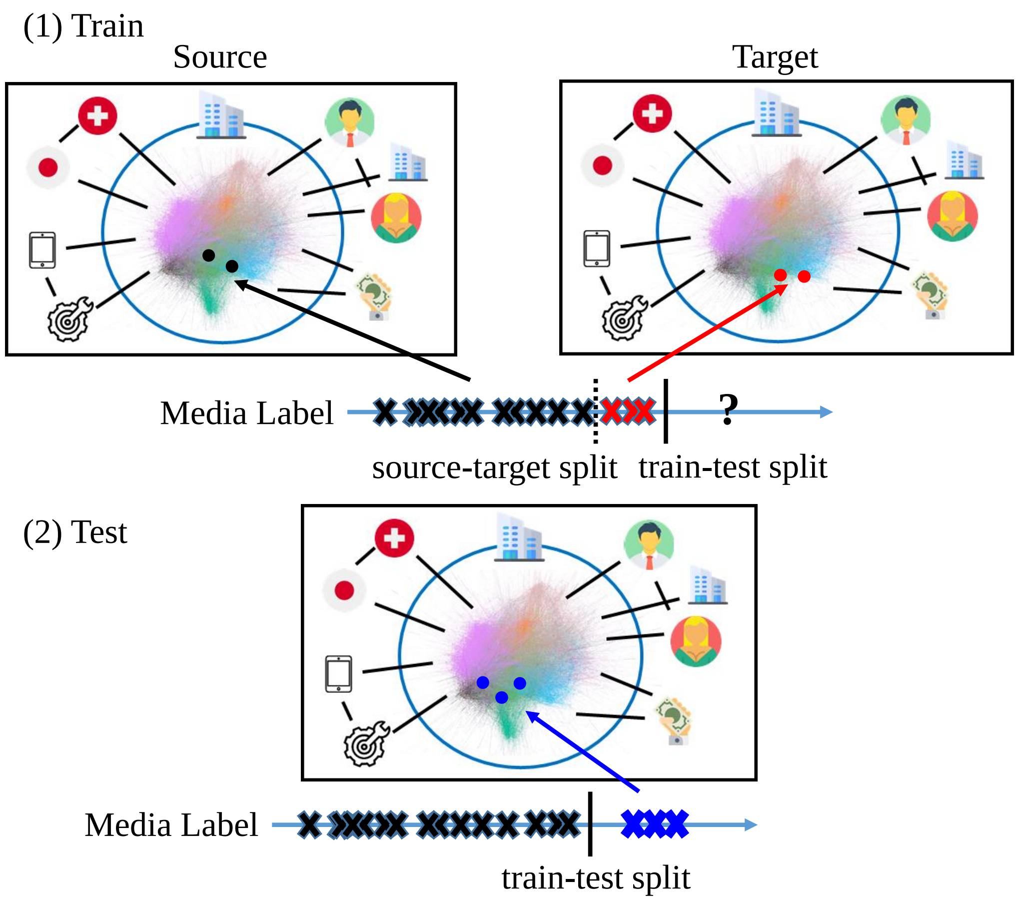



Figure S5: A schematic figure illustrating our approach.

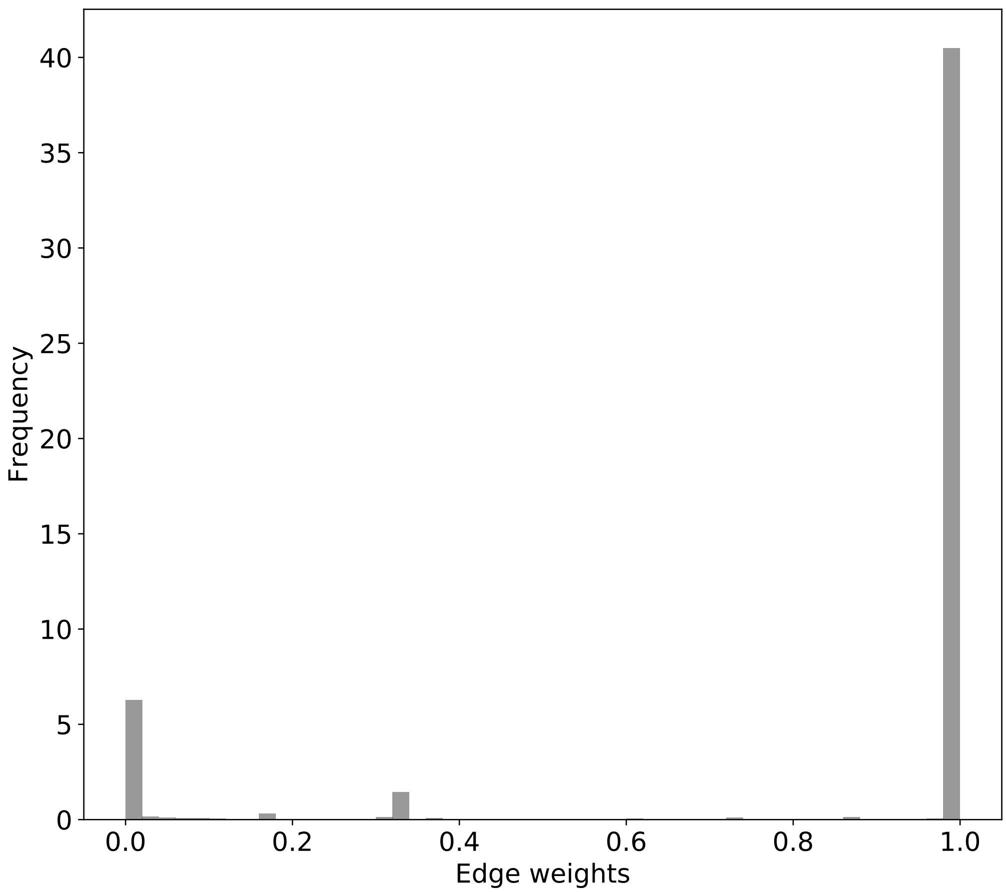



Figure S6: Normalized histogram showing the edge weights of the Product/Service label for LP-path-segment.

| Label | Raw count | Unique firms |
|---|---|---|
| Product-Service | 20,637 | 8,779 |
| Regulatory | 21,652 | 7,552 |
| Financial | 22,754 | 3,310 |
| Fraud | 14,489 | 3,997 |
| Workforce | 7,523 | 3,963 |
| Management | 11,220 | 4,063 |
| Anti-Competitive | 7,748 | 3,620 |
| Information | 6,401 | 2,873 |
| Workplace | 6,827 | 2,492 |
| Discrimination-Workforce | 6,477 | 2,426 |
| Environmental | 4,083 | 1,887 |
| Ownership | 4,124 | 2,615 |
| Production-Supply | 2,878 | 1,869 |
| Corruption | 3,621 | 1,578 |
| Human | 496 | 302 |
| Sanctions | 254 | 157 |
| Association | 247 | 90 |

Table S3: Number of adverse media coverage events from Jan 2012 to May 2018 among the 35,657 firms investigated in this study. "Raw count" denotes the total number of adverse media coverage for a particular adverse media category. "Unique firms" denotes the total number of unique firms tagged with a particular adverse media category.

## Predictive accuracy

Table S4 shows the actual numbers used to plot the result reported in Fig. 2 in the main text. The results obtained by varying the last date of the training data to Aug 1, 2017 are provided in Table S5 and the results varying the start date from Jan 1, 2012 to Jan 1, 2013 are provided in Table S6.

## Comparison of prediction

In Fig. S7, we compare the output of our prediction for the methods compared in this study. We only show results for random forest, LP-fixed, LP-core-relation, and LP-path-segment because the results of LP-mult are similar to those of LP-fixed and the results of LP-path are similar to those of LP-core-relation.

## Interpreting LP-path-segment using nonnegative matrix factorization and partial dependency plots

To understand what our models have learned, we perform the partial dependency analysis on our learned model [16]. A partial dependency plot is estimated by averaging out the effects of all the other variables using the learned model as follows:

$$\bar{f}_s(x_s) = \frac{1}{n}\Sigma_{i=1}^{n}\hat{f}(x_s, x_{i,c}), \tag{1}$$

where $x_{i,c}(i = 1, ..., n)$ are the values of $x_c$ that occur in the data and $x_s$ is the variable of interest. However, because the features used by LP-path-segment are highly correlated, calculating the importance measure for each feature might not be a reasonable approach. Hence, we first reduce the dimensionality of the feature space to 50 using a standard binary nonnegative matrix factorization technique [25] and then perform the usual partial dependence analysis along the basis of the matrix obtained by the standard binary nonnegative matrix factorization method. In mathematical terms, we first decompose the raw feature matrix using binary nonnegative matrix factorization as follows:

$$X = PH, \tag{2}$$

| Category | Random | | Random forest | | LP-fixed | | LP-mult | | LP-core-relation | | LP-path | | LP-path-segment | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PR | ROC | PR | ROC | PR | ROC | PR | ROC | PR | ROC | PR | ROC | PR | ROC |
| Product-Service | 0.051 | 0.5 | 0.110 | 0.675 | 0.066 | 0.602 | 0.065 | 0.610 | 0.103 | 0.701 | 0.065 | 0.605 | **0.266** | **0.845** |
| Regulatory | 0.038 | 0.5 | 0.072 | 0.646 | 0.049 | 0.584 | 0.045 | 0.589 | 0.071 | 0.677 | 0.047 | 0.597 | **0.226** | **0.823** |
| Financial | 0.016 | 0.5 | 0.023 | 0.572 | 0.042 | 0.617 | 0.031 | 0.591 | 0.043 | 0.694 | 0.031 | 0.637 | **0.111** | **0.821** |
| Fraud | 0.016 | 0.5 | 0.029 | 0.653 | 0.029 | 0.614 | 0.024 | 0.562 | 0.030 | 0.656 | 0.023 | 0.616 | **0.105** | **0.807** |
| Workforce | 0.013 | 0.5 | 0.034 | 0.753 | 0.028 | 0.679 | 0.018 | 0.620 | 0.031 | 0.701 | 0.026 | 0.677 | **0.160** | **0.812** |
| Management | 0.022 | 0.5 | 0.030 | 0.582 | 0.031 | 0.551 | 0.025 | 0.540 | 0.035 | 0.626 | 0.026 | 0.558 | **0.113** | **0.764** |
| Anti-Competitive | 0.014 | 0.5 | 0.024 | 0.640 | 0.029 | 0.617 | 0.026 | 0.616 | 0.031 | 0.675 | 0.022 | 0.630 | **0.113** | **0.835** |
| Information | 0.010 | 0.5 | 0.059 | 0.758 | 0.028 | 0.766 | 0.025 | 0.724 | 0.047 | 0.777 | 0.026 | 0.759 | **0.143** | **0.861** |
| Workplace | 0.011 | 0.5 | 0.029 | 0.669 | 0.017 | 0.665 | 0.017 | 0.636 | 0.031 | 0.720 | 0.019 | 0.687 | **0.112** | **0.830** |
| Discrimination-Workforce | 0.016 | 0.5 | 0.045 | 0.702 | 0.025 | 0.665 | 0.021 | 0.632 | 0.044 | 0.708 | 0.025 | 0.666 | **0.125** | **0.821** |
| Environmental | 0.006 | 0.5 | 0.019 | 0.724 | 0.015 | 0.718 | 0.011 | 0.683 | 0.026 | 0.745 | 0.015 | 0.729 | **0.098** | **0.839** |
| Ownership | 0.013 | 0.5 | 0.027 | 0.679 | 0.018 | 0.638 | 0.019 | 0.645 | 0.032 | 0.724 | 0.018 | 0.640 | **0.107** | **0.822** |
| Production-Supply | 0.006 | 0.5 | 0.015 | 0.678 | 0.009 | 0.648 | 0.009 | 0.652 | 0.013 | 0.687 | 0.010 | 0.670 | **0.073** | **0.838** |
| Corruption | 0.009 | 0.5 | 0.021 | 0.700 | 0.021 | 0.695 | 0.011 | 0.599 | 0.023 | 0.662 | 0.018 | 0.689 | **0.091** | **0.778** |
| Human | 0.002 | 0.5 | 0.004 | 0.723 | 0.003 | 0.717 | 0.003 | 0.713 | 0.005 | 0.732 | 0.004 | 0.725 | **0.012** | **0.832** |
| Sanctions | 0.001 | 0.5 | 0.003 | 0.756 | 0.002 | 0.640 | 0.001 | 0.617 | 0.009 | 0.718 | 0.004 | 0.680 | **0.010** | **0.771** |
| Association | 0.000 | 0.5 | 0.001 | 0.664 | 0.001 | 0.613 | 0.001 | 0.635 | 0.001 | 0.679 | 0.001 | 0.647 | **0.002** | **0.801** |

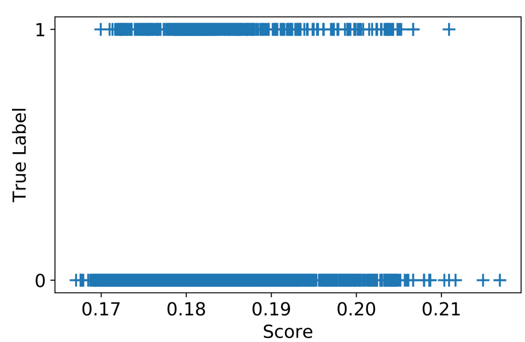
Table S4: Predictive accuracy comparison. PR stands for AUC-PR (area under the precision-recall curve) and ROC stands for AUC-ROC (area under the receiver operating characteristic curve). This is the actual numbers used to plot the result reported in Fig. 2 in the main text.

| Category | Random | | Random forest | | LP-fixed | | LP-mult | | LP-core-relation | | LP-path | | LP-path-segment | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PR | ROC | PR | ROC | PR | ROC | PR | ROC | PR | ROC | PR | ROC | PR | ROC |
| Product-Service | 0.032 | 0.5 | 0.075 | 0.683 | 0.040 | 0.597 | 0.042 | 0.615 | 0.071 | 0.710 | 0.039 | 0.602 | **0.236** | **0.839** |
| Regulatory | 0.023 | 0.5 | 0.044 | 0.658 | 0.031 | 0.588 | 0.028 | 0.597 | 0.048 | 0.689 | 0.028 | 0.596 | **0.209** | **0.855** |
| Financial | 0.010 | 0.5 | 0.013 | 0.566 | 0.029 | 0.614 | 0.020 | 0.574 | 0.030 | 0.680 | 0.023 | 0.636 | **0.098** | **0.819** |
| Fraud | 0.010 | 0.5 | 0.018 | 0.658 | 0.023 | 0.607 | 0.015 | 0.559 | 0.020 | 0.670 | 0.015 | 0.613 | **0.115** | **0.843** |
| Workforce | 0.007 | 0.5 | 0.024 | 0.772 | 0.015 | 0.682 | 0.011 | 0.639 | 0.018 | 0.704 | 0.013 | 0.681 | **0.142** | **0.837** |
| Management | 0.013 | 0.5 | 0.019 | 0.581 | 0.023 | 0.560 | 0.016 | 0.557 | 0.024 | 0.652 | 0.017 | 0.570 | **0.135** | **0.789** |
| Anti-Competitive | 0.008 | 0.5 | 0.012 | 0.611 | 0.021 | 0.610 | 0.012 | 0.597 | 0.020 | 0.668 | 0.013 | 0.625 | **0.109** | **0.762** |
| Information | 0.006 | 0.5 | 0.042 | 0.778 | 0.016 | 0.770 | 0.014 | 0.718 | 0.030 | 0.801 | 0.015 | 0.760 | **0.112** | **0.845** |
| Workplace | 0.007 | 0.5 | 0.024 | 0.671 | 0.010 | 0.659 | 0.011 | 0.627 | 0.026 | 0.726 | 0.012 | 0.686 | **0.112** | **0.858** |
| Discrimination-Workforce | 0.011 | 0.5 | 0.036 | 0.734 | 0.018 | 0.670 | 0.015 | 0.643 | 0.035 | 0.736 | 0.017 | 0.674 | **0.111** | **0.854** |
| Environmental | 0.003 | 0.5 | 0.013 | 0.728 | 0.007 | 0.718 | 0.007 | 0.701 | 0.016 | 0.757 | 0.008 | 0.733 | **0.072** | **0.849** |
| Ownership | 0.008 | 0.5 | 0.015 | 0.690 | 0.012 | 0.625 | 0.010 | 0.625 | 0.019 | 0.724 | 0.010 | 0.623 | **0.113** | **0.827** |
| Production-Supply | 0.004 | 0.5 | 0.010 | 0.692 | 0.005 | 0.655 | 0.006 | 0.675 | 0.009 | 0.724 | 0.006 | 0.684 | **0.057** | **0.804** |
| Corruption | 0.005 | 0.5 | 0.014 | 0.708 | 0.014 | 0.715 | 0.007 | 0.620 | 0.019 | 0.681 | 0.011 | 0.715 | **0.097** | **0.837** |
| Human | 0.001 | 0.5 | 0.003 | 0.704 | 0.001 | 0.722 | 0.002 | 0.780 | 0.002 | 0.748 | 0.001 | 0.732 | **0.007** | **0.843** |
| Sanctions | 0.001 | 0.5 | 0.003 | 0.842 | 0.002 | 0.738 | 0.001 | 0.630 | 0.007 | 0.798 | 0.003 | 0.753 | **0.028** | **0.853** |
| Association | 0.000 | 0.5 | 0.000 | 0.657 | 0.000 | 0.568 | 0.000 | 0.623 | 0.000 | 0.689 | 0.000 | 0.643 | **0.002** | **0.813** |

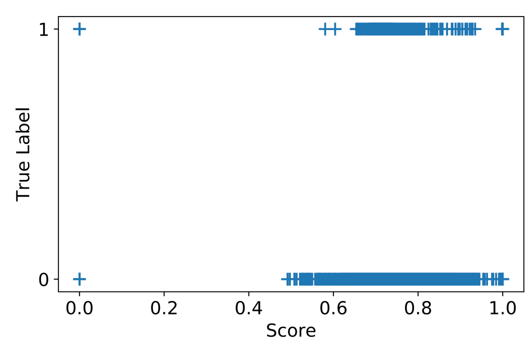
Table S5: Predictive accuracy comparison varying the last date of the training data to be Aug 1, 2017. PR stands for AUC-PR (area under the precision-recall curve) and ROC stands for AUC-ROC (area under the receiver operating characteristic curve).

| Category | Random | | Random forest | | LP-fixed | | LP-mult | | LP-core-relation | | LP-path | | LP-path-segment | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PR | ROC | PR | ROC | PR | ROC | PR | ROC | PR | ROC | PR | ROC | PR | ROC |
| Product-Service | 0.052 | 0.5 | 0.112 | 0.675 | 0.068 | 0.607 | 0.067 | 0.611 | 0.103 | 0.698 | 0.066 | 0.603 | **0.254** | **0.844** |
| Regulatory | 0.038 | 0.5 | 0.073 | 0.646 | 0.049 | 0.588 | 0.045 | 0.589 | 0.070 | 0.675 | 0.047 | 0.596 | **0.211** | **0.811** |
| Financial | 0.016 | 0.5 | 0.021 | 0.569 | 0.041 | 0.619 | 0.031 | 0.599 | 0.042 | 0.691 | 0.031 | 0.639 | **0.106** | **0.818** |
| Fraud | 0.017 | 0.5 | 0.029 | 0.653 | 0.030 | 0.616 | 0.024 | 0.568 | 0.031 | 0.654 | 0.024 | 0.617 | **0.102** | **0.802** |
| Workforce | 0.013 | 0.5 | 0.034 | 0.753 | 0.027 | 0.683 | 0.019 | 0.624 | 0.032 | 0.698 | 0.025 | 0.681 | **0.149** | **0.837** |
| Management | 0.022 | 0.5 | 0.032 | 0.583 | 0.031 | 0.550 | 0.025 | 0.539 | 0.035 | 0.627 | 0.026 | 0.561 | **0.103** | **0.778** |
| Anti-Competitive | 0.015 | 0.5 | 0.027 | 0.653 | 0.028 | 0.615 | 0.026 | 0.617 | 0.030 | 0.665 | 0.022 | 0.626 | **0.110** | **0.808** |
| Information | 0.011 | 0.5 | 0.059 | 0.769 | 0.031 | 0.771 | 0.027 | 0.729 | 0.050 | 0.791 | 0.028 | 0.760 | **0.130** | **0.853** |
| Workplace | 0.011 | 0.5 | 0.031 | 0.667 | 0.017 | 0.660 | 0.018 | 0.638 | 0.031 | 0.716 | 0.020 | 0.682 | **0.100** | **0.829** |
| Discrimination-Workforce | 0.016 | 0.5 | 0.044 | 0.706 | 0.025 | 0.668 | 0.022 | 0.634 | 0.043 | 0.706 | 0.026 | 0.670 | **0.111** | **0.820** |
| Environmental | 0.007 | 0.5 | 0.020 | 0.726 | 0.016 | 0.724 | 0.012 | 0.688 | 0.026 | 0.744 | 0.015 | 0.731 | **0.096** | **0.846** |
| Ownership | 0.013 | 0.5 | 0.027 | 0.675 | 0.019 | 0.643 | 0.019 | 0.646 | 0.032 | 0.721 | 0.018 | 0.646 | **0.104** | **0.816** |
| Production-Supply | 0.006 | 0.5 | 0.017 | 0.689 | 0.010 | 0.659 | 0.010 | 0.660 | 0.014 | 0.688 | 0.010 | 0.677 | **0.074** | **0.789** |
| Corruption | 0.009 | 0.5 | 0.022 | 0.705 | 0.021 | 0.697 | 0.012 | 0.605 | 0.023 | 0.662 | 0.019 | 0.690 | **0.085** | **0.776** |
| Human | 0.002 | 0.5 | 0.004 | 0.702 | 0.003 | 0.706 | 0.003 | 0.704 | 0.005 | 0.723 | 0.003 | 0.711 | **0.011** | **0.819** |
| Sanctions | 0.001 | 0.5 | 0.035 | 0.781 | 0.021 | 0.662 | 0.002 | 0.610 | 0.032 | 0.715 | 0.011 | 0.694 | **0.068** | **0.790** |
| Association | 0.000 | 0.5 | 0.001 | 0.666 | 0.001 | 0.623 | 0.001 | 0.656 | 0.001 | 0.699 | 0.001 | 0.665 | **0.002** | **0.808** |

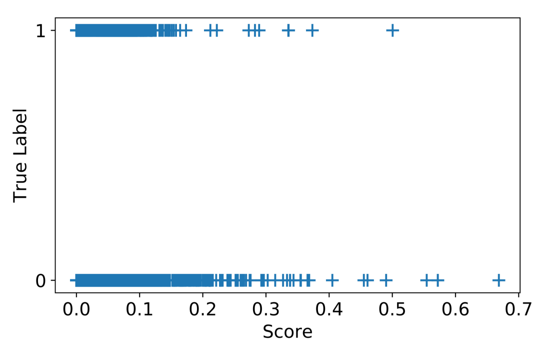
Table S6: Predictive accuracy comparison varying the start date to be Jan 1, 2013. PR stands for AUC-PR (area under the precision-recall curve) and ROC stands for AUC-ROC (area under the receiver operating characteristic curve).



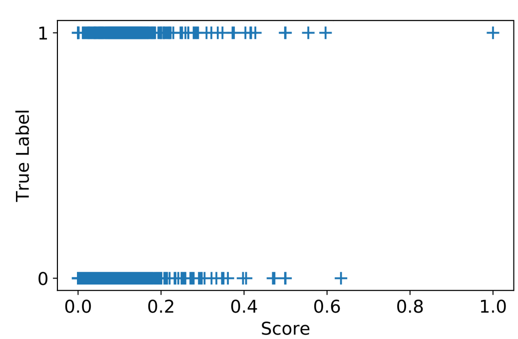

(a) random forest




(b) LP-fixed




(c) LP-core-relation




(d) LP-path-segment

Figure S7: Comparison of the predictions obtained by the methods compared in this study.

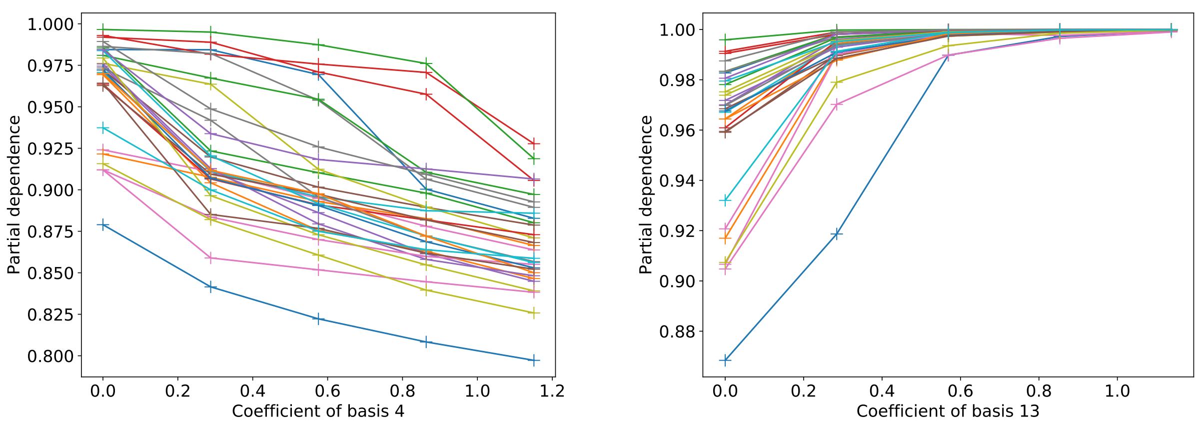
where $X$ denotes the raw feature matrix, $P$ the positive coefficient matrix, $H$ the positive basis matrix and the dimension of $P$ is much lower than that of $X$. We then rewrite Eq. 1 as follows:

$$\bar{f}_s(p_s) = \frac{1}{n}\Sigma_{i=1}^{n}\hat{f}(p_s, p_{i,c}, h_s), \tag{3}$$

where $s$ is the the row number of the basis matrix (i.e. $H$), obtained by the standard binary nonnegative matrix factorization method, under investigation and we vary $p_s$ (i.e. the $s$th column of the coefficient matrix $P$) instead of $x_s$ in Eq. 1. We could interpret our model by examining the partial dependency plot and the corresponding basis vector as was done in the main text.

In Fig. S8 we show the partial dependency plots of basis vector 4 and 13 of which the basis vectors were already examined in the main text (Fig. 2) and of which corresponds to the feature that had the most negative or the most positive effect respectively as is shown below (see Table S7). Each line corresponds to a partial dependency plot from a different run of model training varying the initial parameters. We repeated the training step and partial dependency analysis for 30 times. We confirm that although there are fluctuations among the partial dependency plots among the different learned parameters, they seem to exhibit similar behavior.

In a partial dependency plot analysis, we usually use the sample standard deviation of the fitted values of the partial dependency plot as a valid measure of feature importance [14, 13]. However, since our feature matrix is binary, we instead focus on the absolute difference of the response at the 0.99 quantile and 0.01 quantile of the coefficients vector corresponding to each basis vector. We also take the average value of the importance measure repeating the training and partial dependency analysis step 30 times to mitigate the effect of fluctuation stemming from the learning process.




(a) Partial dependency plot of basis vector 4.

(b) Partial dependency plot of basis vector 13.

Figure S8: Partial dependency plot. Each line corresponds to a partial dependency plot from a different run of model training varying the initial parameters. We repeated the training step and partial dependency analysis for 30 times. We confirm that although there are fluctuations among the partial dependency plots among the different learned parameters, they seem to exhibit similar behavior.

In Table S7, we show the top-five important features learned for the "Product/Service" label. We also report analysis for basis vector 26 (Fig. S9), basis vector 30 (Fig. S10) and basis vector 7 (Fig. S11). For basis vector 26, we see that it focuses more on same industry relations since "belongs to" relates a firm to industry classification and "send goods" relates a firm to a specific product code that they sent through the US customs. In order to understand what basis vector 30 seems to represent, we first note that we sorted the relation types in descending order from left to right using the number of occurrences reported in Table S2. The fact that it peaks mostly at the left of each path segments implies that this basis corresponds to common relation types in the dataset. Basis vector 7 is a little bit more tricky, but it seems to focus on indirect partnership relationships, adding more weights to an edge if its surrounding edges are tightly connected with partnership relationships.

In Table S8, we show the top-five important features and the top important feature that had a negative effect learned for the "Financial" label. We see that all the top-five features have a positive effect on the edge weights. We also report analysis for basis vector 34 (Fig. S12), basis vector 10 (Fig. S13), basis vector 21 (Fig. S14) and basis vector 20 (Fig. S15). Basis

| Rank | Basis | $E_{\hat{\theta}}[f(x_{0.99}) - f(x_{0.01})]$ | $\lvert E_{\hat{\theta}}[f(x_{0.99}) - f(x_{0.01})]\rvert$ |
|---|---|---|---|
| 1 | 4 | -0.096 | 0.096 |
| 2 | 26 | -0.070 | 0.070 |
| 3 | 30 | -0.057 | 0.057 |
| 4 | 13 | 0.040 | 0.040 |
| 5 | 7 | 0.039 | 0.039 |

Table S7: Top-five important features for the Product/Service label. We see that basis vector 4 is the most important feature with a negative effect on the edge weight and basis vector 13 is the most important feature among the basis which had a positive effect on the edge weight.

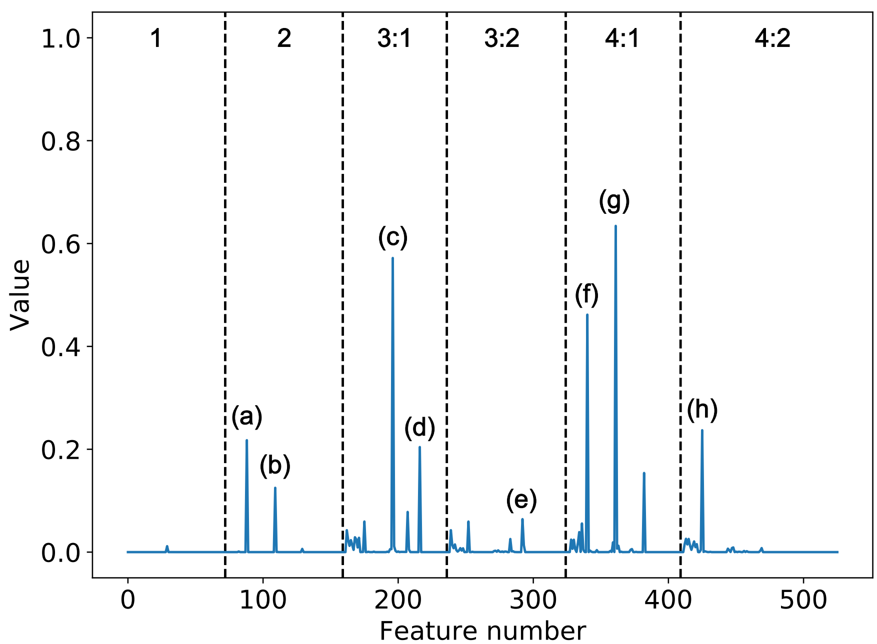



Figure S9: Basis vector 26. The dotted vertical lines divide each path segment. Because there are relation types that does not appear in some path segments, the total number of features is 526 instead of 1,296 (216 $\times$ 6). Peaks in basis vector 26: (a) belongs to, (b) partner-manufacture, (c) partner-manufacture, (d) send goods, (e) send goods, (f) belongs to, (g) partner-manufacture, (h) belongs to.

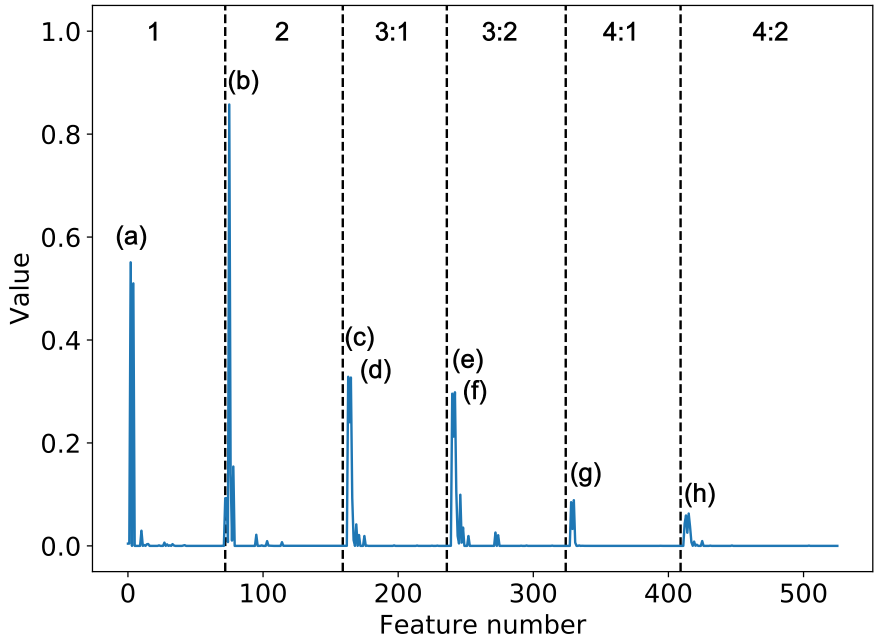



Figure S10: Basis vector 30. The dotted vertical lines divide each path segment. Because there are relation types that does not appear in some path segments, the total number of features is 526 instead of 1,296 (216 × 6). Peaks in basis vector 7: (a) supplier, (b) located in, (c) supplier, (d) customer, (e) customer, (f) supplier, (g) customer, (h) customer.

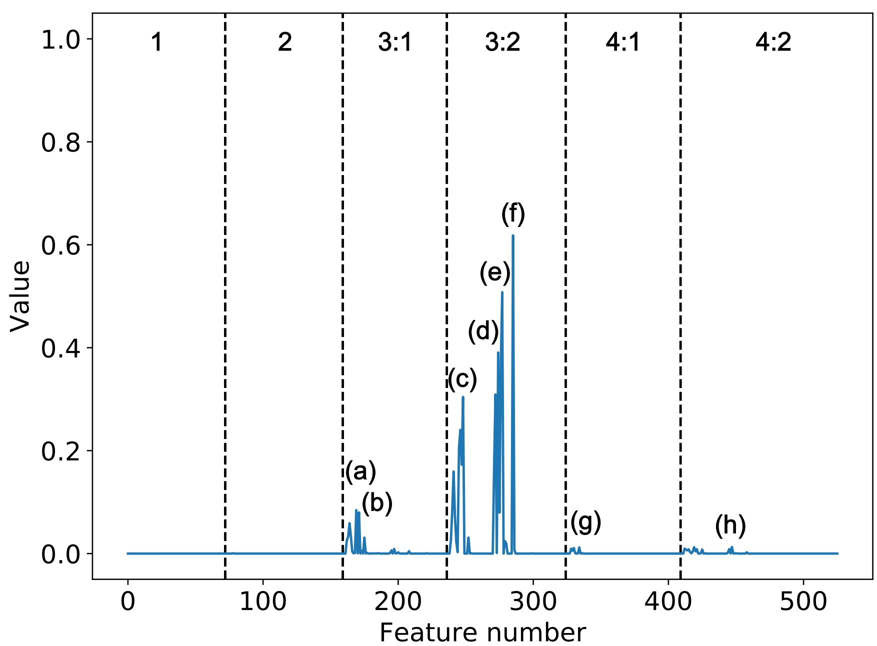



Figure S11: Basis vector 7. The dotted vertical lines divide each path segment. Because there are relation types that does not appear in some path segments, the total number of features is 526 instead of 1,296 (216 × 6). Peaks in basis vector 7: (a) competitor, (b) distributor, (c) partner-product-bundle, (d) partner-research-collaboration, (e) partner-marketing, (f) partner-technology, (g) competitor, (h) partner-research-collaboration.

vector 7 (indirect partnership relationships) and basis vector 30 (common relation types) are already analyzed above, so we omit it here. For basis vector 34 and 10, we see that they focus more on creditor-borrower relationships. Since "Financial" label news is about ownership and board related issues (see Agricultural Chemicals above), it makes sense that these relation types come on top. Furthermore, basis vector 20 focuses on competitor relationship while basis vector 21 focuses on general partnership relationships.

| Rank | Basis | $E_{\hat{\theta}}[f(x_{0.99}) - f(x_{0.01})]$ | $\lvert E_{\hat{\theta}}[f(x_{0.99}) - f(x_{0.01})]\rvert$ |
|---|---|---|---|
| 1 | 34 | 0.090 | 0.090 |
| 2 | 7 | 0.089 | 0.089 |
| 3 | 10 | 0.089 | 0.089 |
| 4 | 21 | 0.088 | 0.088 |
| 5 | 20 | 0.081 | 0.081 |
| ... | ... | ... | ... |
| 10 | 30 | -0.051 | 0.051 |

Table S8: Top-five important features and the top important feature that had a negative effect for the "Financial" label.

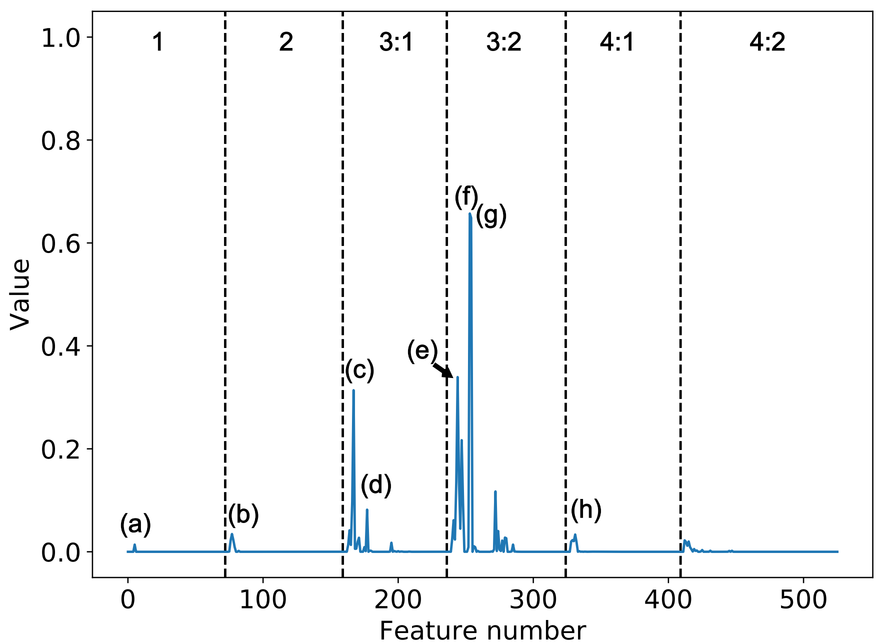



Figure S12: Basis vector 34. The dotted vertical lines divide each path segment. Because there are relation types that does not appear in some path segments, the total number of features is 526 instead of 1,296 ($216 \times 6$). Peaks in basis vector 34: (a) creditor, (b) strategic alliance, (c) borrower, (d) creditor, (e) borrower, (f) tenant, (g) landlord, (h) creditor.

## References


[1] Kenneth R. Ahern and Denis Sosyura. Who Writes the News? Corporate Press Releases During Merger Negotiations. *Journal of Finance*, 69(1), April 2014.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, pages 722–735, Berlin, Heidelberg, 2007. Springer-Verlag.

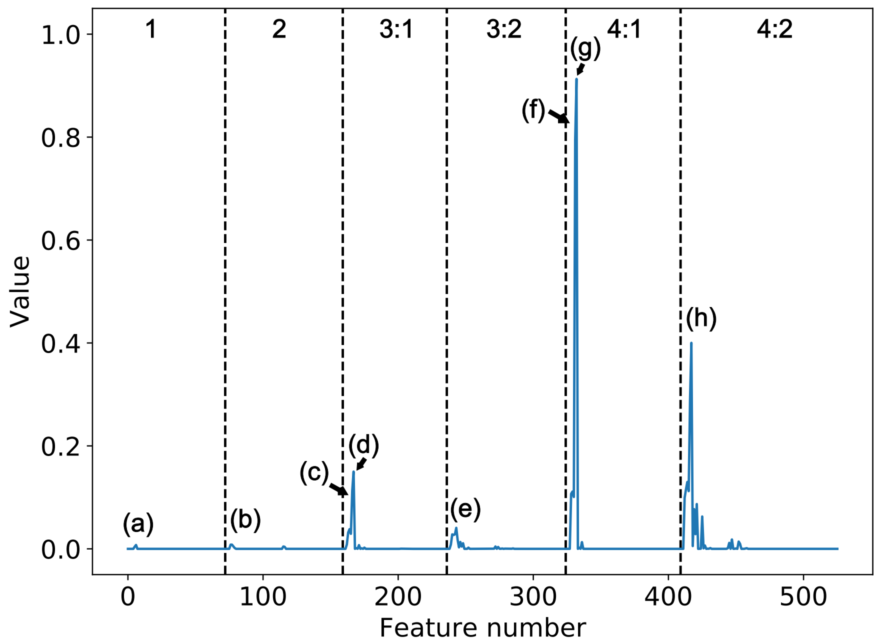



Figure S13: Basis vector 10. The dotted vertical lines divide each path segment. Because there are relation types that does not appear in some path segments, the total number of features is 526 instead of 1,296 (216 × 6). Peaks in basis vector 34: (a) borrower, (b) strategic alliance, (c) creditor, (d) borrower, (e) creditor, (f) creditor, (g) borrower, (h) borrower.

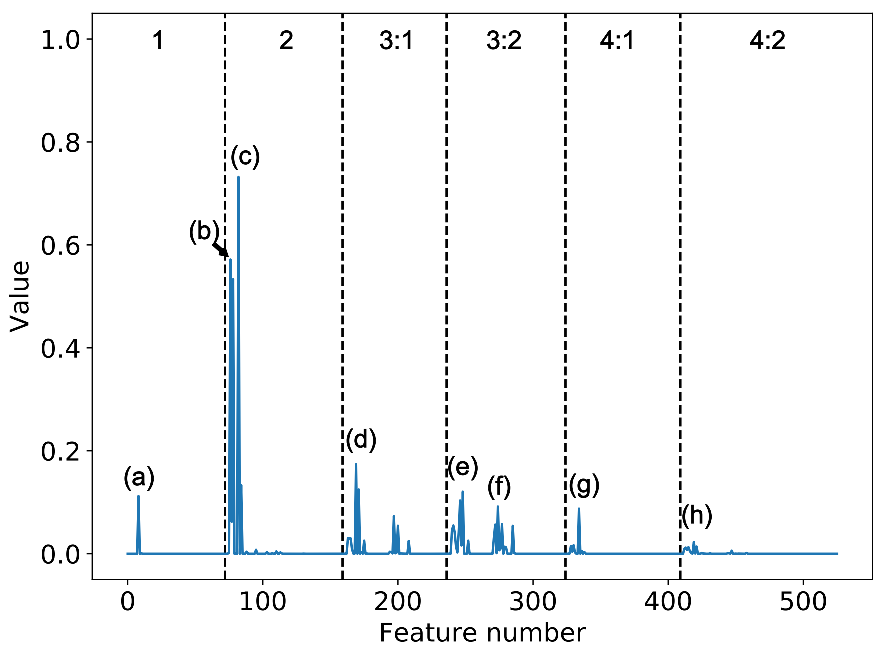



Figure S14: Basis vector 21. The dotted vertical lines divide each path segment. Because there are relation types that does not appear in some path segments, the total number of features is 526 instead of 1,296 (216 × 6). Peaks in basis vector 21: (a) competitor, (b) supplier, (c) competitor, (d) competitor, (e) distributor, (f) competitor, (g) competitor, (h) competitor.

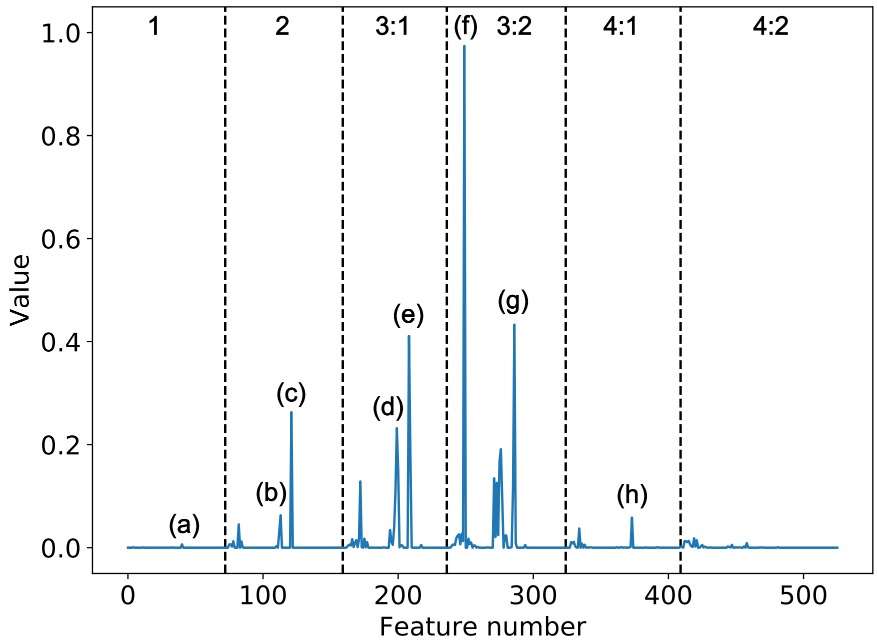



Figure S15: Basis vector 20. The dotted vertical lines divide each path segment. Because there are relation types that does not appear in some path segments, the total number of features is 526 instead of 1,296 ($216 \times 6$). Peaks in basis vector 20: (a) partner-technology, (b) partner-marketing, (c) partner-technology, (d) partner-product-bundle, (e) partner-technology, (f) partner-unknown, (g) partner-patent-license, (h) partner-technology.

[3] Urs Birchler, René Hegglin, Michael R. Reichenecker, and Alexander F. Wagner. Which swiss gnomes attract money? efficiency and reputation as performance drivers of wealth management banks. *Swiss Finance Institute Research Paper*, No. 16-28, 2016.

[4] Steven F. Cahan, Chen Chen, Li Chen, and Nhut H. Nguyen. Corporate social responsibility and media coverage. *Journal of Banking and Finance*, 59(C):409–422, 2015.

[5] Cambridge Analytica. Cambridge analytica and scl elections commence insolvency proceedings and release results of independent investigation into recent allegations. Press Release, May 2018. `https://ca-commercial.com/news/cambridge-analytica-and-scl-elections-commence-insolvency-proceedings-and-release-results-`

[6] T. Carlyle. *On Heroes, Hero-worship, and the Heroic in History*. Frederick A. Stokes, 1893.

[7] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

[8] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. ACM.

[9] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 601–610, New York, NY, USA, 2014. ACM.

[10] Lily Fang and Joel Peress. Media coverage and the cross-section of stock returns. *The Journal of Finance*, 64(5):2023–2052, 2009.

[11] Michael Färber, Basil Ell, Carsten Menne, and Achim Rettinger. A comparative survey of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web Journal, July*, pages 1–26, 2015.

[12] Ali Fatemi, Martin Glaum, and Stefanie Kaiser. Esg performance and firm value: The moderating role of disclosure. *Global Finance Journal*, 2017.

[13] B. M. Greenwell, B. C. Boehmke, and A. J. McCarthy. A Simple and Effective Model-Based Variable Importance Measure. *ArXiv e-prints*, May 2018.

[14] Brandon M. Greenwell. pdp: An r package for constructing partial dependence plots. *The R Journal*, 9(1):421–436, 2017.

[15] David E. Harrington. Economic news on television the determinants of coverage. *Public Opinion Quarterly*, 53(1):17–40, 1989.

[16] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[17] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–91, 1979.

[18] Ni Lao, Tom Mitchell, and William W. Cohen. Random walk inference and learning in a large scale knowledge base. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 529–539, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[19] OECD. Responsible business conduct for institutional investors: Key considerations for due diligence under the oecd guidelines for multinational enterprises. *OECD guidlines*, 2017.

[20] World Federation of Exchanges. Wfe annual statistics guide 2017. 2017.

[21] David H. Solomon. Selective Publicity and Stock Prices. *Journal of Finance*, 67(2):599–638, April 2012.

[22] S.N. Soroka. *Negativity in Democratic Politics: Causes and Consequences*. Cambridge Studies in Public Opinion and Political Psychology. Cambridge University Press, 2014.

[23] Stuart N. Soroka. Good news and bad news: Asymmetric responses to economic information. *Journal of Politics*, 68(2):372–385, 2006.

[24] James Uleman and Laura M. Kressel. *A brief history of theory and research on impression formation*, pages 53–73. Oxford University Press, 2013.

[25] Marinka Žitnik and Blaž Zupan. Nimfa: A python library for nonnegative matrix factorization. *J. Mach. Learn. Res.*, 13(1):849–853, March 2012.

[26] Chang Wan, Xiang Li, Ben Kao, Xiao Yu, Quanquan Gu, David Cheung, and Jiawei Han. Classification with active learning and meta-paths in heterogeneous information networks. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 443–452, New York, NY, USA, 2015. ACM.

[27] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1225–1234, New York, NY, USA, 2016. ACM.

[28] Xander Wilcke, Peter Bloem, and Victor De Boer. The knowledge graph as the default data model for learning on heterogeneous knowledge. *Data Science*, 1(1-2):39–57, 12 2017.