Predicting Adverse Media Risk using a Heterogeneous Information Network

https://arxiv.org/abs/1811.12166

Ryohei Hisano[1,2],

Didier Sornette[3], Takayuki Mizuno[4,2]

[1] Social ICT Research Center, Graduate School of Information Science and Technology, The University of Tokyo,

[2] The Canon Institute for Global Studies,

[3] ETH Zürich [4] National Institute of Informatics

Dec 4, 2018

The two keywords

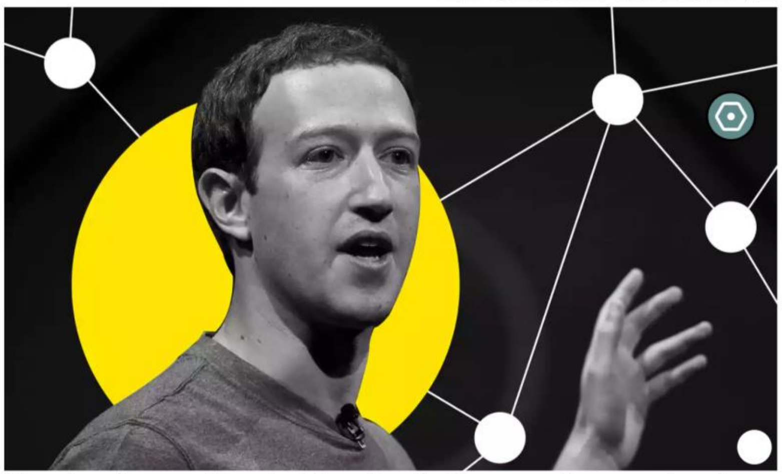
Adverse	Heterogeneous
Media	Information
Risk	Network

Adverse Media Risk

- Negative media coverage may lead to huge risk

Facebook–Cambridge Analytica data scandal

The Guardian



Facebook says Cambridge Analytica may have gained 37m more users' data

Company reveals up to 87m people may have been affected as Mark Zuckerberg takes responsibility for 'a huge mistake'

Olivia Solon *in San Francisco*

Wed 4 Apr 2018 23.01 BST

US imposes sanctions against Russian oligarchs and government officials

By Donna Borak, CNN

Updated 2234 GMT (0634 HKT) April 6, 2018

Sanctioned firms

Among the companies targeted
By the US include GAZ Group …

Database on adverse media risk

- Factiva (Dow and Jones), RepRisk etc
- Gathered for financial investment
- Jan 2012 – May 2018

Num firms under our watch list: 35657, 17 Label

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

Date	Name	Adverse Media Label
2012/1/3	FCA	Management
2012/1/3	Daimler Trucks North America	Product/Service
2012/1/10	Atlas Fibre	Regulatory
2012/1/11	Tokyo Electric Power Company	Workplace
2012/1/16	Air India Regional	Management
…	…	…

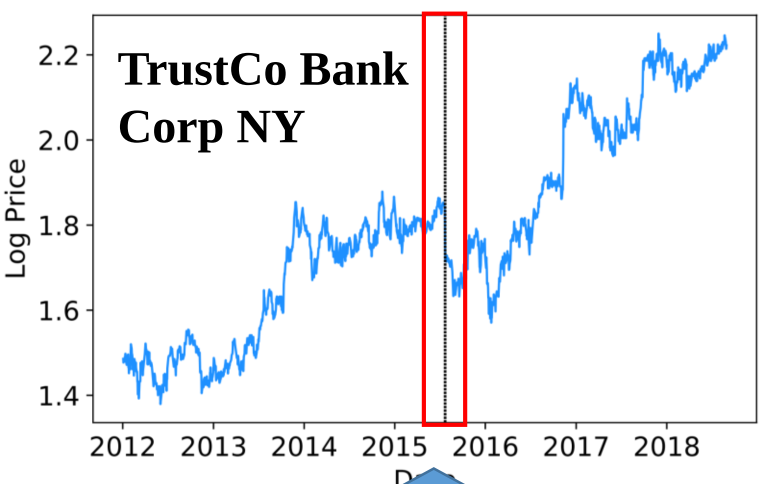
BUT WHY CARE TO PREDICT?

Measuring the effect of media on returns

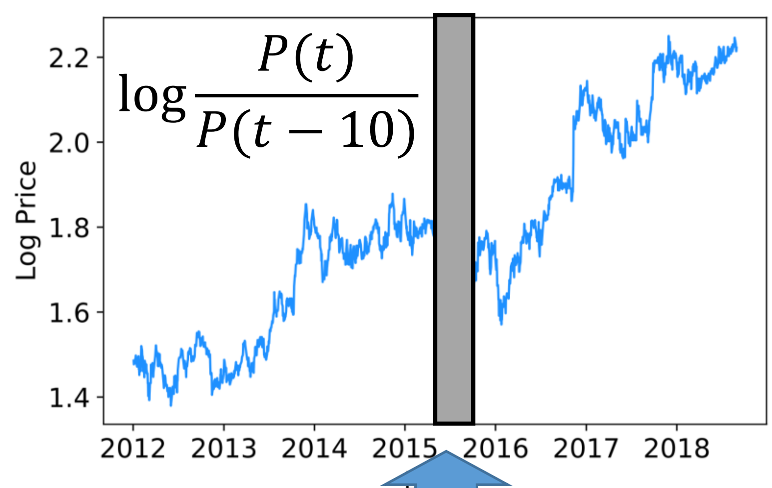
- For all US stocks in the list, we gather price for the period 2012.1–2018.5.
  - <u>**1,139 stocks in total**</u>

Date	Name	Adverse Media Label
2012/1/3	FCA	Management
2012/1/3	Daimler Trucks North America	Product/Service
2012/1/10	Atlas Fibre	Regulatory
2012/1/11	Tokyo Electric Power Company	Workplace
2012/1/16	Air India Regional	Management
...	...	...

- For each date in the adverse media label list, employ a 10-day window centered on the specified date. We then take a log return of the start and end dates (10 trading days difference).
- We compare the above log return to that of 10 trading days log return outside the windows.



News date, Calc $\log \frac{P(t+5)}{P(t-5)}$

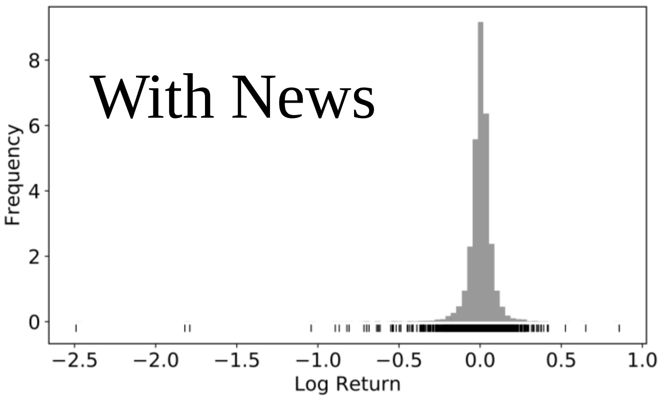


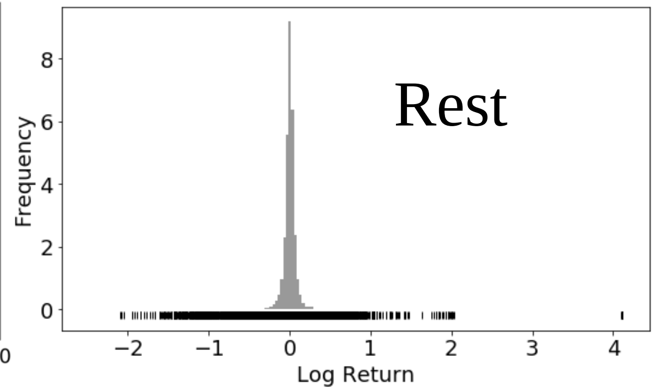
Ignore

Result

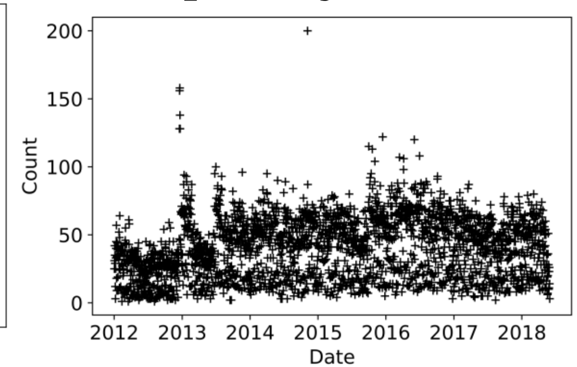
Indeed there is an effect

- **Left:** a histogram of log returns inside the time windows. **Middle:** same thing outside the time windows.
- We could see that the negative tail distribution is more stretched while the positive tail is shrunk compared to the middle.





Frequency of news



	N	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
2 sample KS-test p-value=7*10^-7							

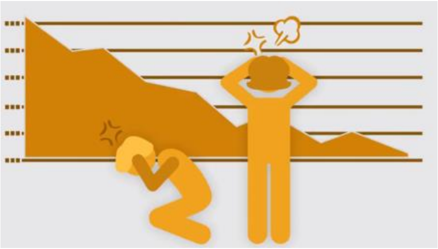
Other reasons

<u>(2) Watchdog adversarial role of the press</u>

- Media plays a central role in monitoring powerful institutions and identifying any activities harmful to the public.
  - Identifying problems = adverse media
  - **<u>Social responsible investment</u>**

(3) Human nature

- People tend to prioritize negative information more.
  - Psychology: Impression formulation voting behavior
  - Economics: Loss aversion macroeconomic behavior



Can we predict future adverse media?

Obviously past label info is not enough to predict future patterns

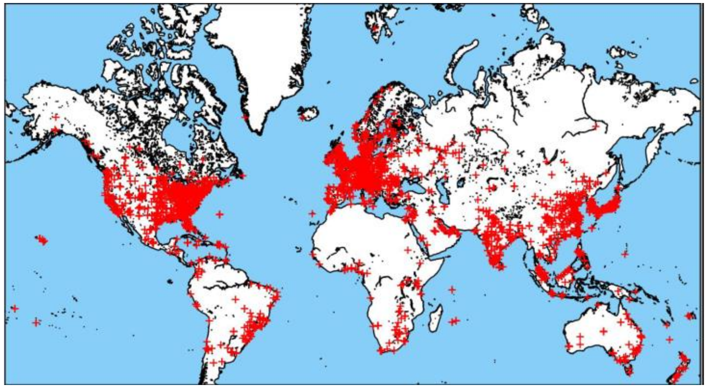
Date	Name	Adverse Media Label
2012/1/3	FCA	Management
2012/1/3	Daimler Trucks North America	Product/Service
2012/1/10	Atlas Fibre	Regulatory
2012/1/11	Tokyo Electric Power Company	Workplace
2012/1/16	Air India Regional	Management
…	…	…

Our approach: Construct a heterogeneous information network combining data from different sources and perform label propagation utilizing this HIN

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

Overview of the HIN

- we have more than just a network of major firms

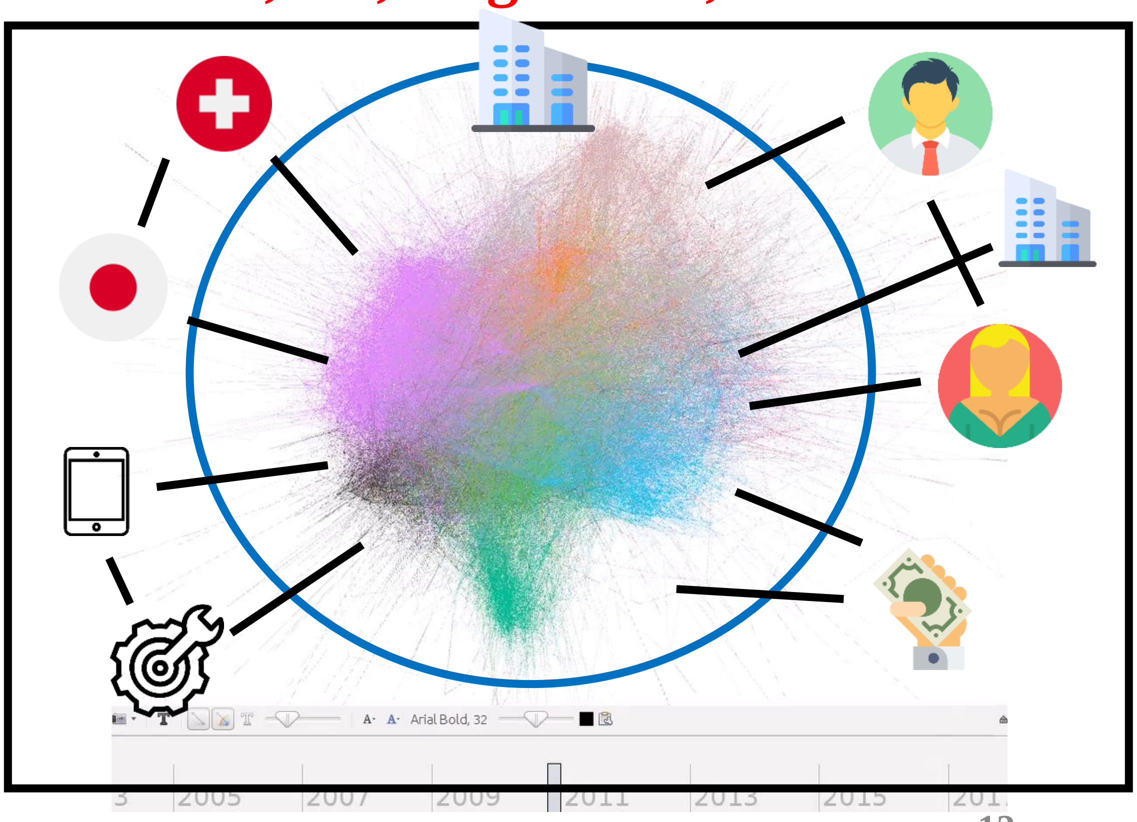


Top 25 / **216** relation types

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

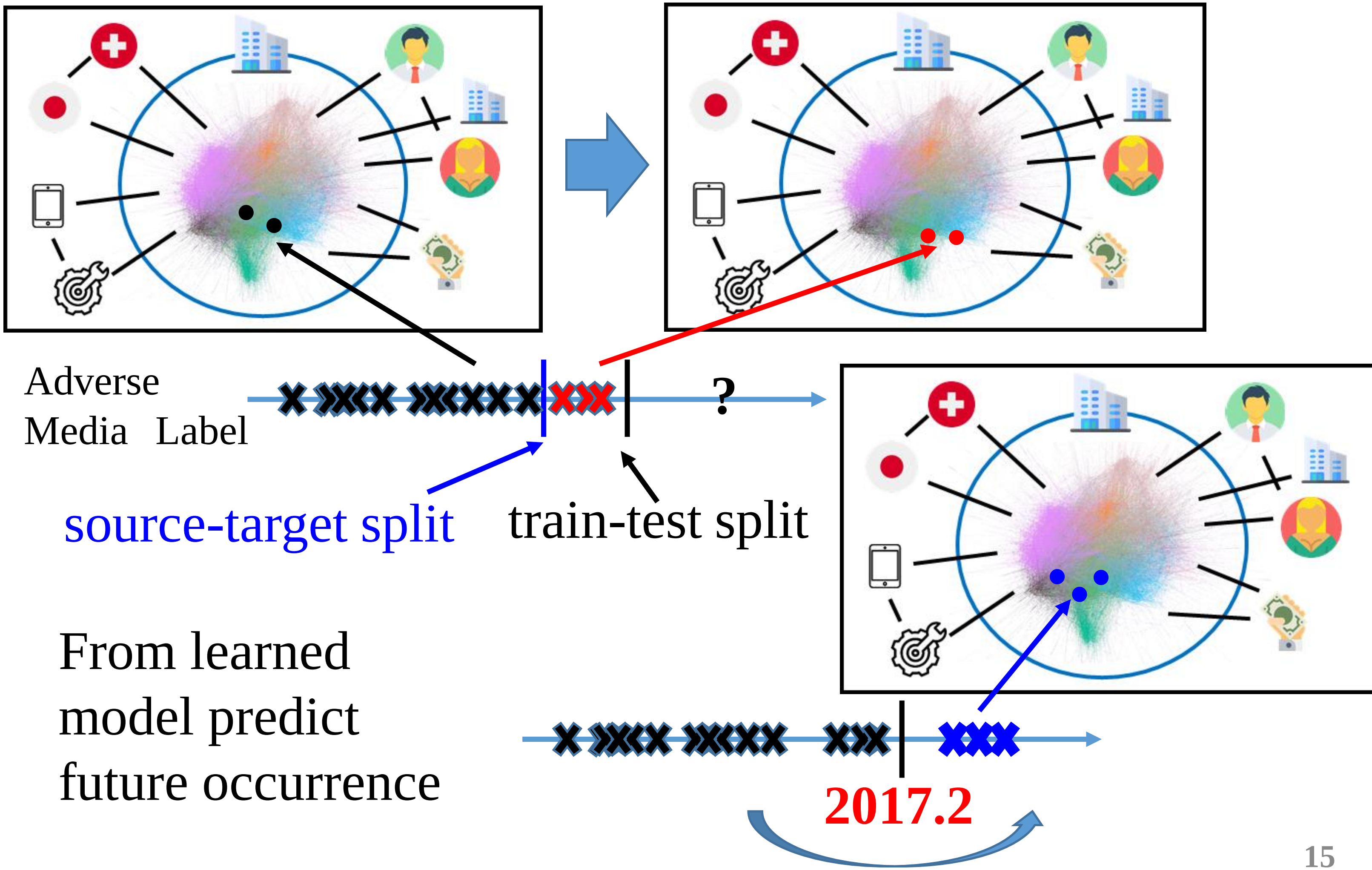
Nodes:50 mil, Edges: 400 mil

Core: 35,000, Edges: 320,000



Schematic Figure

- Using adverse media label occurrence patterns and HIN we want to learn how to propagate labels to predict future occurrences



Two ingredients of the model

- (1) Propagation model
  - that could adaptively adjust to each label
  - Slight variation of LP with edge weight learning

- (2) Edge features

Propagation Model

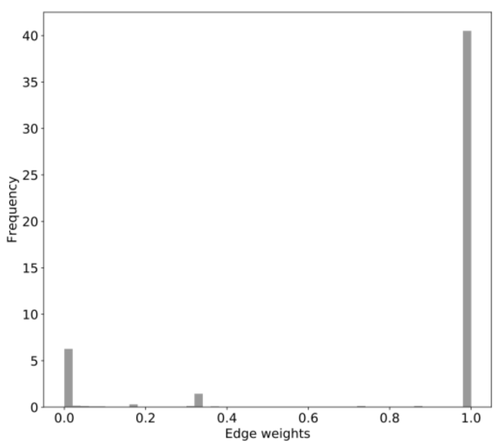
- We model edge weights $w_{ij} = f_\theta(x_{ij})$ using edge features. We **enforce $\mathbf{0 \leq w_{ij} \leq 1}$** by using a sigmoid function.

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 $\boxed{D_{ii} = \Sigma_j 1_{ij \in E}}$.
(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 θ in f_θ using gradient descent.
(8) Repeat until convergence.

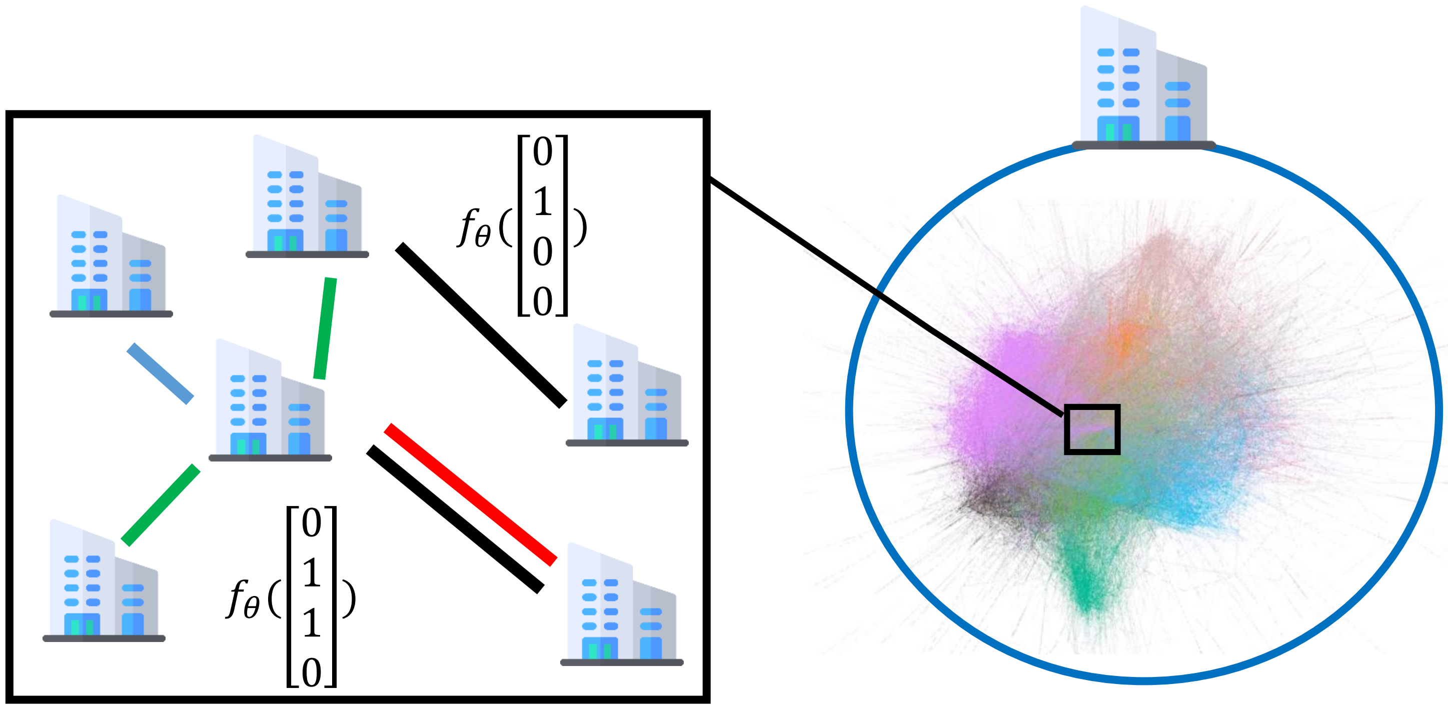
$\cancel{D_{ii} \leftarrow \Sigma_j w_{ij}}$

Learned
Edge weights



Edge features 1: core-relation

- Relation types in the network among firms in the watch list



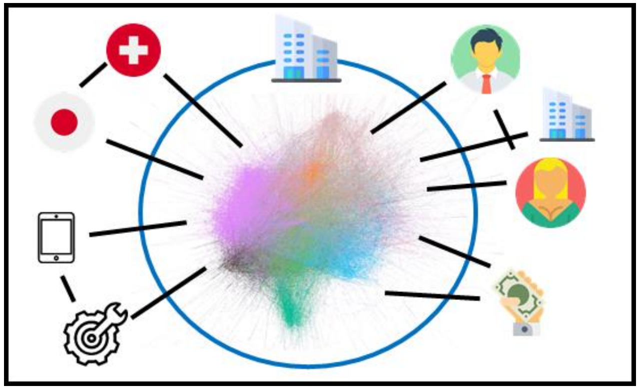
Edge features 2: path

We also focus on paths that could be reached ignoring nodes that we reached in the previous path lengths

- Path Ranking Algorithm [Lao,Cohen2010]

uses path to perform knowledge graph competition

- We use the top 3,000 frequent paths and use them as an one hot features
  - We use path length up to 4



- For example if firm A and firm B has

The following relationships

  - Path length 1: (A,supplies,B)
  - Path length 2: (A,is_in,c,is_in,B)
  - Path length 3: (A,makes,x,is_made_of,y,imports,B)

(A,supplies,B) **(A,is_in,c,is_in,B)** **(A,makes,x,is_made_of,y,imports,B)**

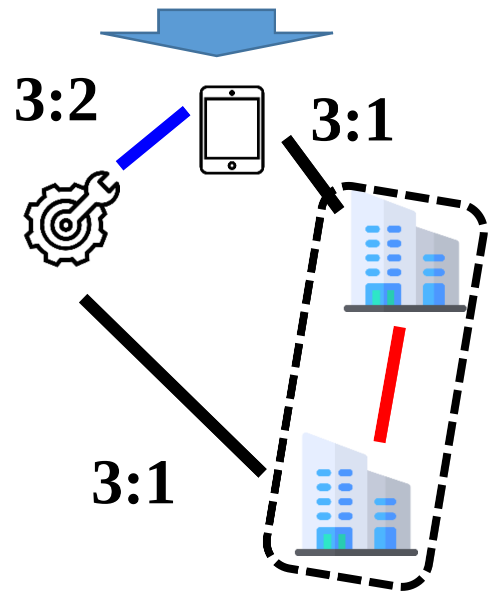
$$\text{edge A-B}\quad [\underbrace{1,0,\ldots,0}_{\text{Path length 1}},\underbrace{0,1,0,0,\ldots,0}_{\text{Path length 2}},\underbrace{0,\ldots,1,0,\ldots,0}_{\text{Path length 3}},\ldots]$$

Path length 1 **Path length 2** **Path length 3**

Edge features 3:path-segment

- We record the occurrence of relation types along the path’s segments
  - <u>We use path length up to 4</u>
- For example if firm A and firm B has the following relationships
  - Path length 1: (A,supplies,B)
  - Path length 2: (A,is_in,c,is_in,B)
  - Path length 3: (A,makes,x,is_made_of,y,makes,B)
  - Path length 3: (A,supplies,C,supplies,D,supplies,B)
- We record it in a binary format as follows

We distinguish relation types occurring along path segments. However, since our network is undirected there is a symmetry.



(supplies) (is_in) (supplies) (makes) (supplies) (is_made_of)

edge A-B [1,0,…,0,0,1,0,0,…,0,0,…,1,0,…,1,0,…,1,0,…,1,0,…]

Path length 1 **Path length 2** **Path length 3:1** **Path length 3:2**

Train test split time

No Info from the future

- We split our data using **2017.1.31** as our last day of training
- Because we want to avoid any information coming from the future to contaminate our HIN
- The problem here is **half of the edges in our database has no timestamp**. So in order to really ensure that all the edges our from the past, we set the test date after the latest date when we acquired the data (which is Jan 2017)

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

Summary of Compared Methods

Information

Low

Methods	Approach	Features	Edge weights	Learning Label Patterns	Label Correlation
Random Forest	Non-Network	Country and Industry Classification	-	Yes	No
LP-fixed	Network	-	Fixed	No	No
LP-mult	Network	-	Fixed	No	Yes
LP-core-relation	Network	Relation types among watch list firms	Learned	Yes	No
LP-path	HIN	Paths relating two nodes	Learned	Yes	No
LP-path-segment	HIN	Occurrence of relation types among path segments relating two nodes	Learned	Yes	No

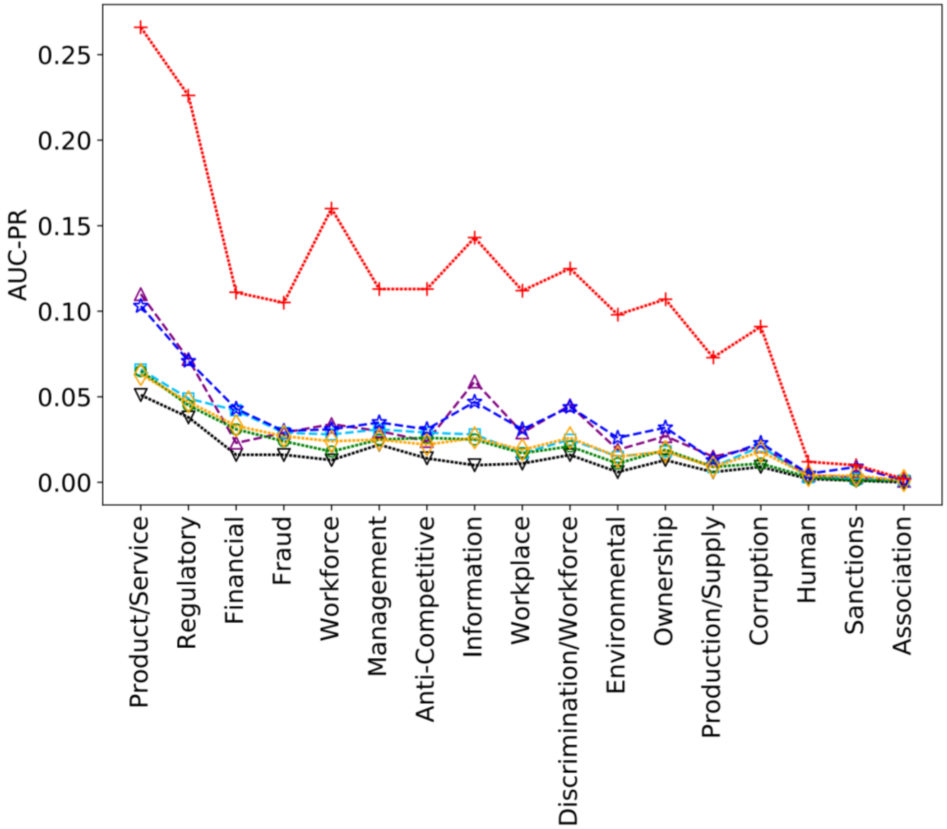
High

It's enough to prove that the HIN approach beat other methods

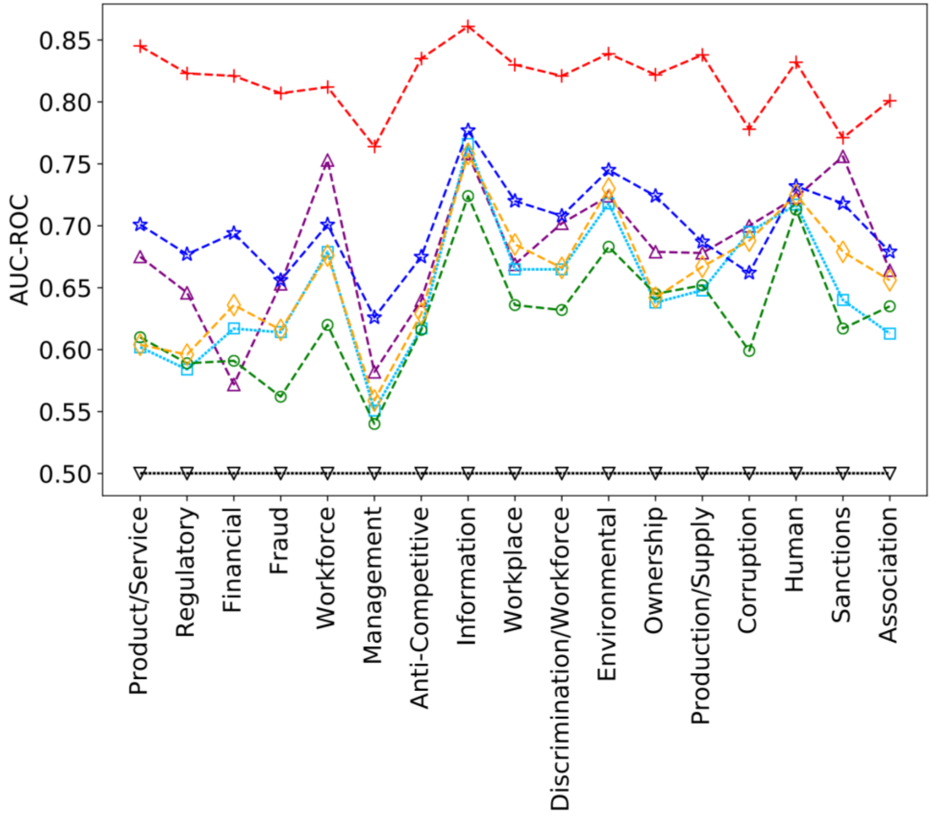
Results as figures

Black：random guessing, Purple: random forest
Light blue:LP-fixed, Green:LP-mult, Blue:LP-core-relation,
Orange: LP-path, Red:LP-path-segment

AUC-PR

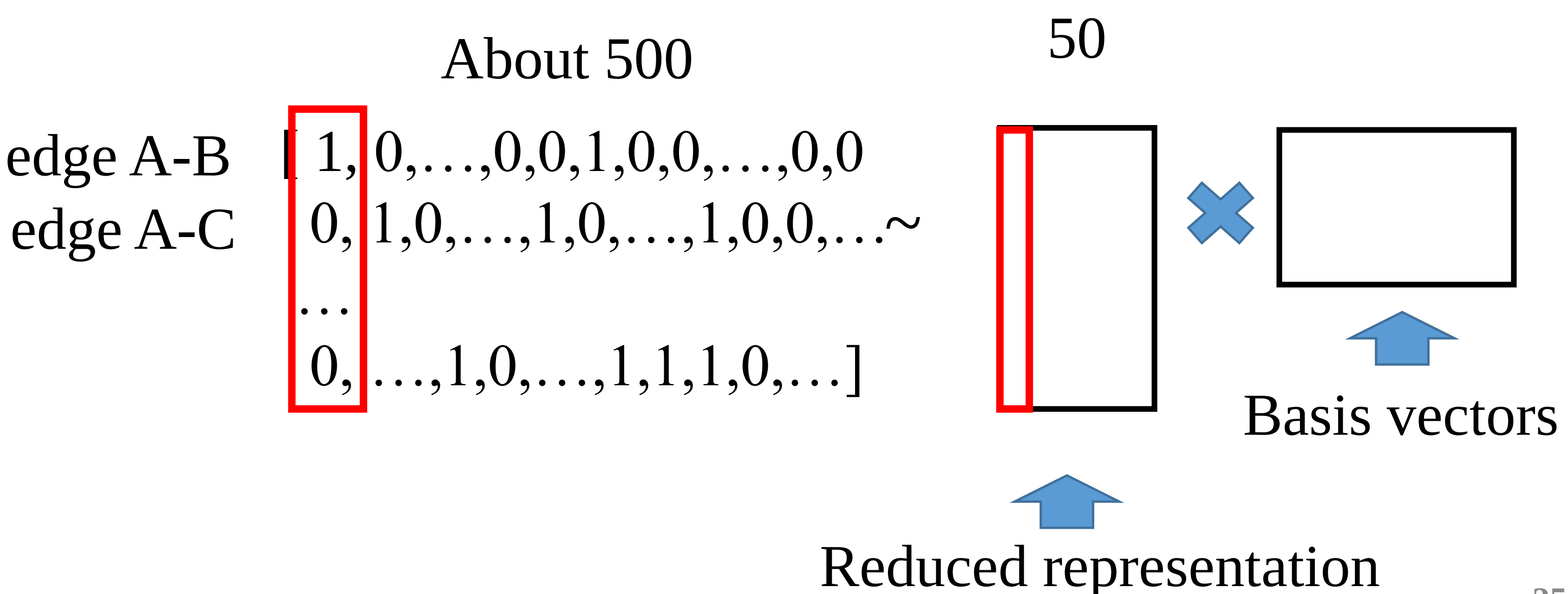


AUC-ROC



Interpreting the learned model

- Too many correlated features making it difficult to analyze what our models have learned directly.
- Thus, we reduce the number of features using nonnegative matrix factorization to 50 and perform the usual partial dependency analysis along the basis of the matrix obtained by binary NMF

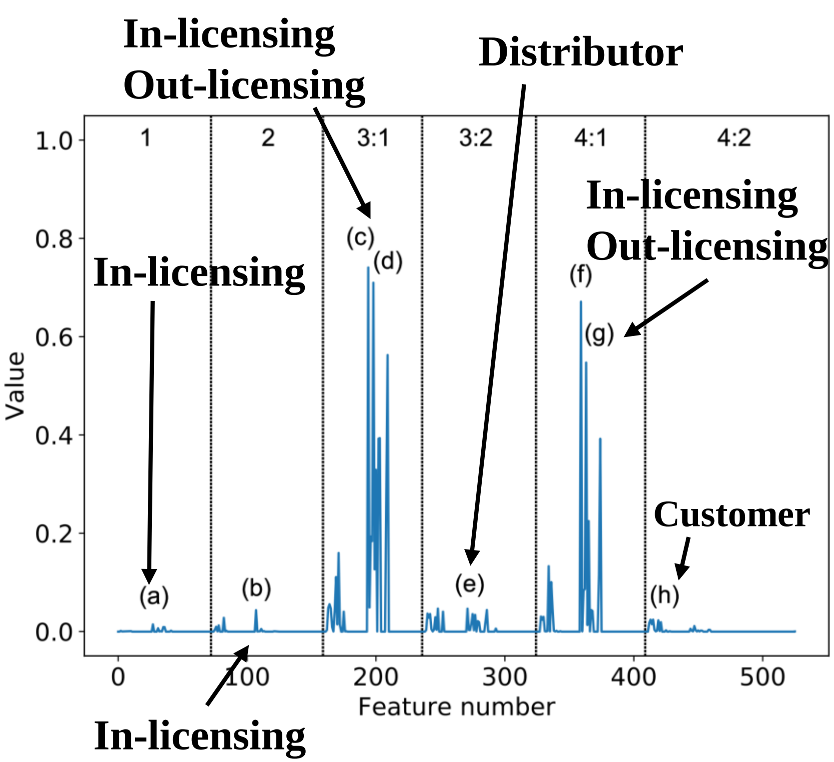


Product/Service Label

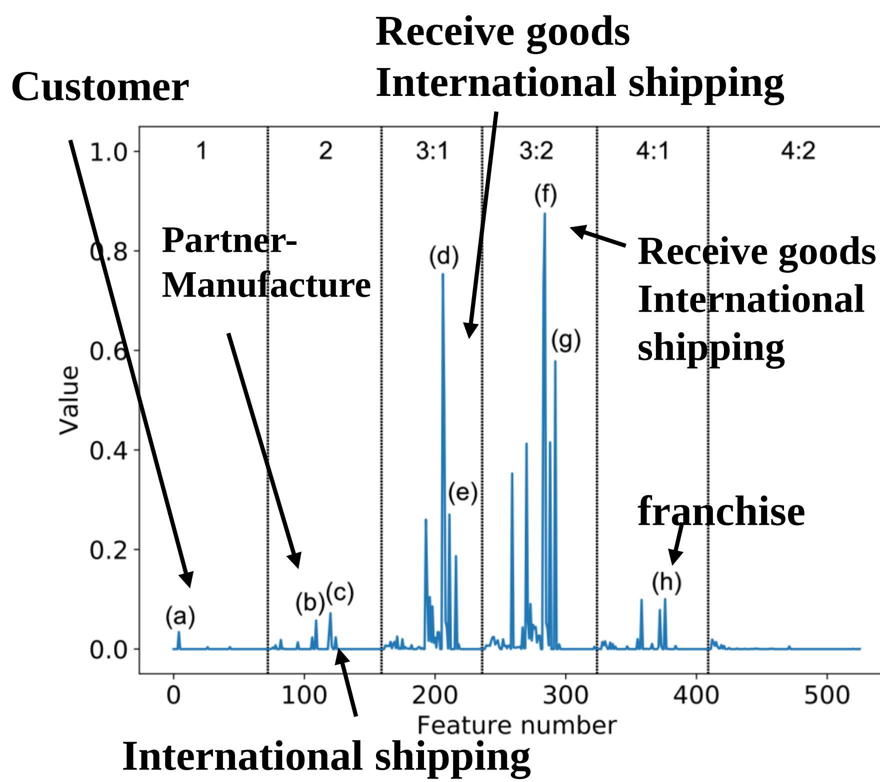
- Basis 4: Top Negative effect
- Basis 13: Top Positive effect

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

Basis 4 : license

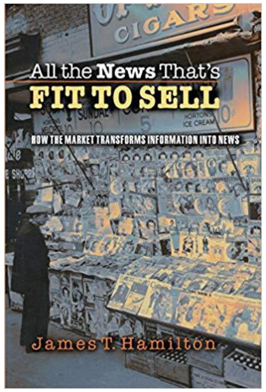


Basis 13: buyer-seller



Why does our method work?

- (1) When a problem occurs, it is likely that similar firms are also in trouble.
  - **Similarity**: closeness in information network
  - Moreover, **<u>we adjust for the closeness measure using past adverse media label patterns</u>**
- (2) Media does not look for news at random. They search for nearby firms for follow-up stories
  - Watchdog role of the press
  - "All the news that's fit to sell"



Adverse Media Prediction

Heterogeneous Info Net

Significance

- **Finance:** Many "news -> financial impact", but very few focusing on predicting news itself
  - **35,657 -> 8,795(firms with ticker)/46,583 (world total)**
- **CS/Network:** New frontier of HIN (knowledge graph)
- **Management:** Adverse media risk score
  - (a) Firms could plan counter measure (CSR/PR)
  - (b) Journalism to find next possible target
  - (c) (Social Responsible) Investment
- **Media studies:** Adverse media prediction
- **Society in general:** Created ways to monitor dominant multinational institutions in the era of information technology (**Cyber watchdog**)