プロパティグラフによる社会データにおける常識の考慮
「ビッグデータからインテリジェンス作成」

久野遼平[1,2]
1. 東京大学大学院情報理工学系研究科ソーシャルICT研究センター　2. キヤノングローバル戦略研究所

[概要]
本研究の狙いは企業・人・商品などに関する情報を複数ソースから収集し、異質情報ネットワークとして分析することで、単一ソースの情報だけでは予想することや分析することが困難な諸課題に対して、予想精度の向上や未知の洞察の発見などの実証分析がどれくらい改善するかを検証することにある。無論これではあまりにも抽象的なため、本研究では特にテーマをESG（環境・社会・ガバナンス）投資に用いられる投資除外リストと制裁リストの予想に問題設定を絞った。ここでは特にESG投資に関連する成果を概説する。

問題定義

図1(a)：投資除外リスト・制裁リスト

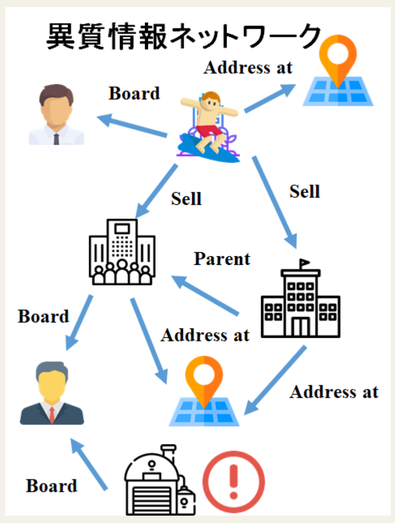
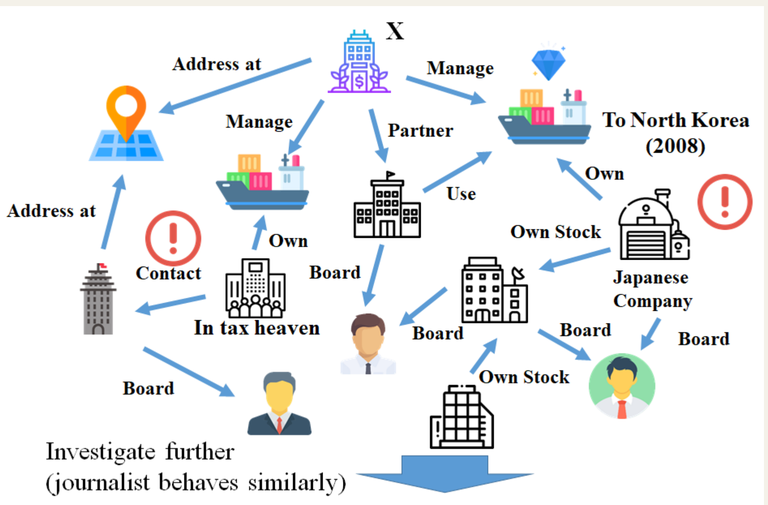


図1(b)：異質情報ネットワークの例

複数ソースからの情報の必要性

●制裁リスト（スマート制裁）に関する国連捜査
‣センシティブ情報の利用（米国財務省・SWIFTデータ）の他に<u>複数情報（登記簿、船舶、企業関係、所有）を手動で取得・検証し捜査を進める</u>(Furukawa 2017)
‣センシティブ情報はデータマイニング不能なことが多い（Zarate 2015）。前者をビッグデータで真似れるか。



ESG投資と投資除外リスト

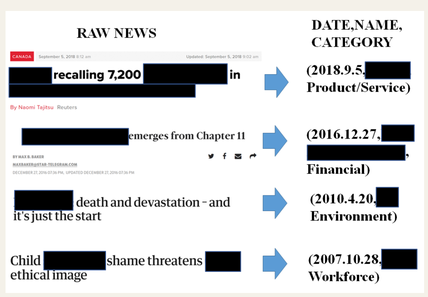
●特に機関投資家は国際社会の持続的な発展に寄与するように配慮して投資することが求められており（United Nations 2006; OECD 2017）、近年ESG（環境・社会・ガバナンス）投資に関心が高まっている。

Region	2014		2016
Europe	59%	! →	53%
United States	18%		22%
Canada	31%		38%
Australia/New Zealand Asia	17%		51%
Asia			1%
Japan	1%		3%
Global	30%		26%

ESG投資/全投資
[Sherwood and Pollard (2018)]

●投資除外リストの作成法
(i) 企業が自主公開した情報
‣偽装可能性（エンロンの粉飾決算）
(ii) 格付け情報
‣利益相反（サブプライム問題）
(iii) 過去のニュース情報を用い作成
‣単に報道されていない企業を無視
➔ ESG投資に対する関心が高まっている反面、それを支える情報環境はまだ追いついていないことが問題視されている。
●ここでは(iii)に注目し今あるリストの情報（図1(a)）とファクト（図1(b)）を用いることで<u>将来時点にリストに追加されるものを予想できるかを検証（～ニュース予想とも見れる）</u>
➔ 情報環境の改善への一歩

(B) 投資除外リスト作成用のデータ

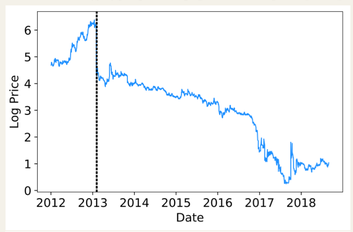


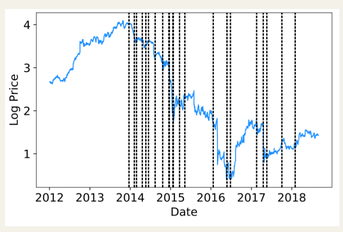
Label	Raw count	Unique firms
Product-Service	20,637	8,779
Regulatory	21,652	7,552
Financial	22,754	3,310
Fraud	14,489	3,997
Workforce	7,523	3,963
Management	11,220	4,063
Anti-Competitive	7,748	3,620
Information	6,401	2,873
Workplace	6,827	2,492
Discrimination-Workforce	6,477	2,426
Environmental	4,083	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

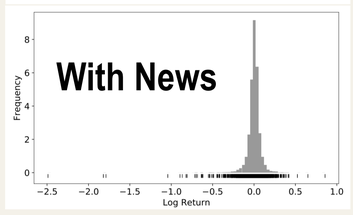
● ダウジョーンズアドバースメディアデータ（2012年1月-2018年5月）
● カテゴリは「製品」「規制」「財政状況」「詐欺」「環境」「ジェンダー」など17種類

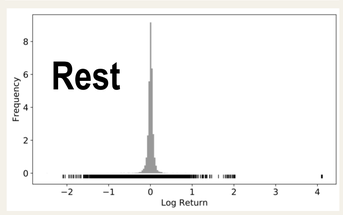
株価への影響

●今回対象にした<u>負のニュース</u>は株価への影響も顕著
‣ESG投資が利益に組するかは諸説ある









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

図1(b)に対応するデータのソース

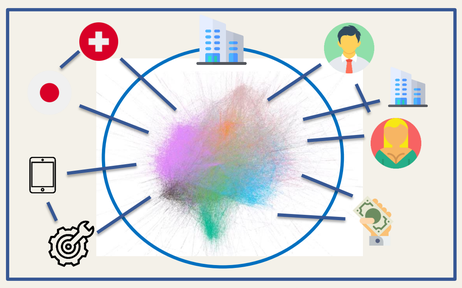
● 投資家が用いるデータ7種と公開情報2種を組み合わせ異質情報ネットワークを作成

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

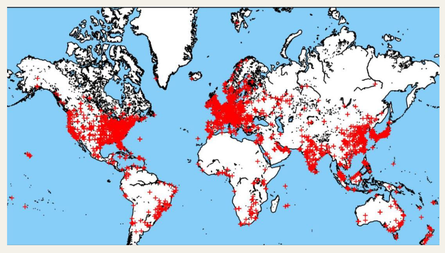
●関係型（頻出上位25）

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

● 企業関係以外も豊富



ノード数:5千万、エッジ数: 4億
予想対象ノード数: 35,657
予想対象間エッジ数：322,138



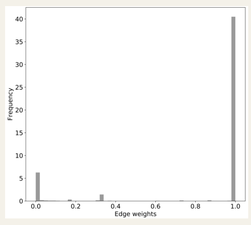
● 予想対象（35,657は世界中に散らばっている

提案手法

● ラベル伝播法の亜種
● ポイント：予想対象間の無向性エッジ（一つでも関係があればエッジがあるとする）の重みを調整することで各ラベルに適合するようにする

Algorithm 1 Slight Variation of Label Propagation

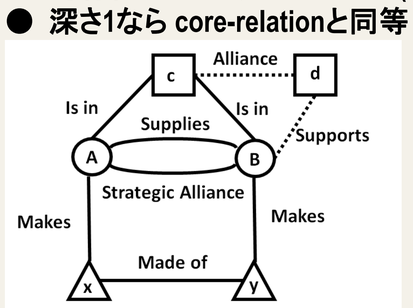
(1) For each edge in the core network set, $w_{ij} = f_\theta(x_{ij})$, where x_{ij} denotes features from the network.
(2) Compute diagonal degree matrix D by $D_{ii} = \Sigma_j 1_{ij \in E}$.
(3) Compute $A_{ii} = I_l(i) + D_{ii}$, where $I_l(i)$ indicates i's known label.
(4) Initialize $Y^0 = (y_1, ..., y_l, 0, ..., 0)$, 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 θ in f_θ using gradient descent.
(8) Repeat until convergence.



訓練後：エッジの重み

エッジ特徴量

● A-B間のエッジの重みに関連する特徴量を定義する
(1) core-relation
● A-B間の関係型だけに注目しバイナリの特徴量を作成（下記例なら(A, Supplies, B) と (A, Strategic Alliance, B)
➔ [1,0,0,1,0,...]).
(2) path
● 深さ4までの各パスに対してバイナリで特徴量を作成
● 計算量の都合で頻出上位3000のみ使用
(3) path-segments
● パスの各セグメントに出現している関係型に注目
● 対称性 ➔ 6 セグメントのみ (i.e. 1,2,3:1,3:2,4:1,4:2)
● 深さ1なら core-relationと同等



● 高次の関係を考慮する時は低次で出現したノードはもう辿らないようにする。
●これにより国や産業などスーパーノードによる影響を軽減できる。

比較手法と予想期間の設定

● ゴール：複数ソースの情報を活用すると予想精度が向上するか検証すること

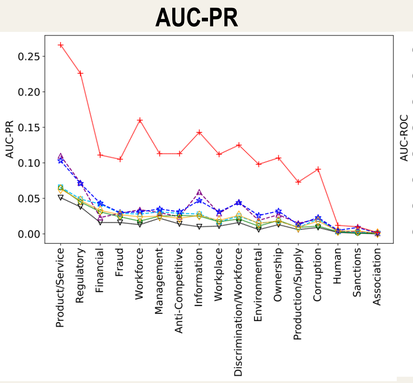
Low ↓ Info ↓ High

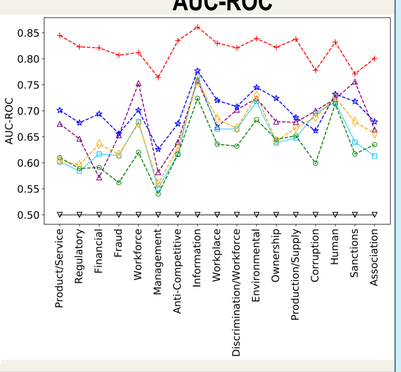
Methods	Approach	Features	Edge weights	Learning Label Patterns	Label Correlation
Random Forest	Non-Network	Country and Industry Classification	-	Yes	No
LP-fixed	Network	-	Fixed	No	No
LP-mult	Network	-	Fixed	No	Yes
LP-core-relation	Network	Relation types among watch list firms	**Learned**	Yes	No
LP-path	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

● 関係型の半数以上はタイムスタンプがない
● そこで予想期間は私が関係型のデータを取得して以降の日付とする（2017年2月以降）（リストは2018年4月まで）

予想精度

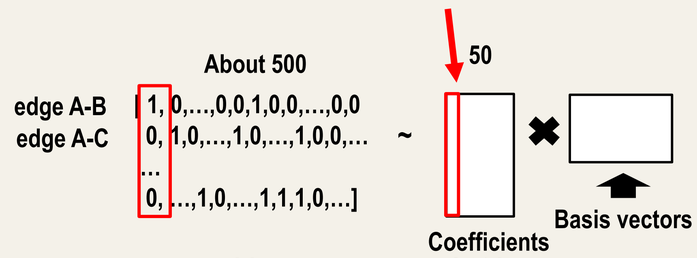
● 異質情報ネットワークのデータを用いると飛躍的向上
● 黒：random guessing, 紫: random forest,
水色:LP-fixed, 緑:LP-mult, 青:LP-core-relation,
オレンジ: LP-path, 赤:LP-path-segment





解釈性

● 特徴量の相関が強いため直接的に解釈を与えるのは難しい
● そのため特徴量をバイナリの非負行列分解を用い特徴量の行列を50次元に削減。そして非負行列分解で得られた基底の方向に沿って部分従属分析を行う。



● Product/Service（製品）ラベルの分析結果

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

● 基底 4: 負の効果: License
● 基底 13: 正の効果: Buyer-seller

なぜうまくいくか

(1) 似た企業は同様の問題を抱えていることが多い
● 似ている＝異質情報ネットワークで定義
● さらにラベルごとに近さを調整している
(2) 捜査　～　調査報道
● 問題が発覚するとその周囲に目が行きやすい

参考文献

● Hisano, R., Sornette, D., and Mizuno, T. (2018). Social Blacklist Prediction using a Heterogeneous Information Network. https://arxiv.org/abs/1811.12166. Submitted.
● Furukawa, K. (2017). Kitacyosen Kaku no Shikengen Kokuren Sousa no Hiroku [Funding Source of North Korea: A Note on United Nation's Investigation]. Tokyo Shincyosya. Tokyo. Japan
● OECD. (2017). Responsible business conduct for institutional investors: Key considerations for due diligence under the OECD Guidelines for Multinational Enterprises. OECD guidlines (2017). https://mneguidelines.oecd.org/RBC-for-Institutional-Investors.pdf
● Sherwood, M. W. & Pollard, J. (2018). Responsible Investing An Introduction to Environmental, Social, and Governance Investments. Routledge. ISBN 9781351361927.
● United Nations. (2006). Principles for Responsible Investment. 2006. www.unpri.org. Accessed March 10, 2019.
● Zarate, J.C. (2015). Treasury's war : the unleashing of a new era of financial warfare. PublicAffairs. New York. ISBN　1610391160, 9781610391160.