Future of FinTech

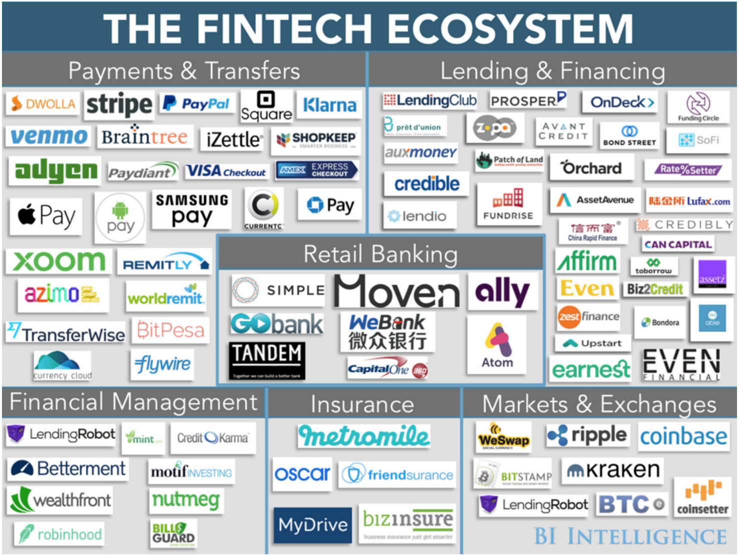
Sanjiv R. Das
Santa Clara University

San Francisco
November 2016

Role of Analytics

- Four Vs of Big Data: Volume, Velocity, Variability, Veracity are critical for competitiveness of a large institution. (A fifth commonly asserted V is Volatility.)
- Decisions based on big data are less judgmental. Creates a new culture of good judgment based on data.
- The decision process is repeatable and automatizeable.
- Analytics leads to new business opportunities through uncovering unexpected correlations in the data.
- Scope now extends to system wide risk and return, and generates useful data aggregates.
- Data models lead to real time decision-making.
- Better communication and information sharing across the organization.

Lumascape (example of a complex ecosystem in analytics)

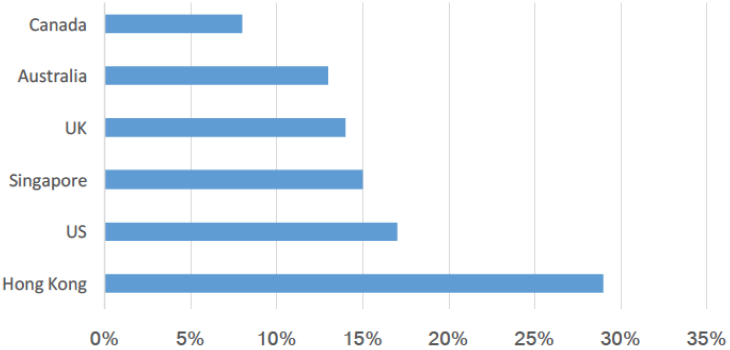


Benefits of Analytics for Large Banks

- Monitoring corporate buzz.
- Analyzing data to detect, analyze, and understand the more profitable customers or products.
- Targeting new clients.
- Customer retention.
- Lending activity (automated)
- Market prediction and trading.
- Risk management.
- Automated financial analysts.
- Financial forensics to prevent rogue employees.
- Credit cards: optimizing use, marketing offers.
- Fraud detection.
- Detecting market manipulation.
- Social network analysis of clients.
- Measuring institutional risk from systemic risk.

FinTech Landscape

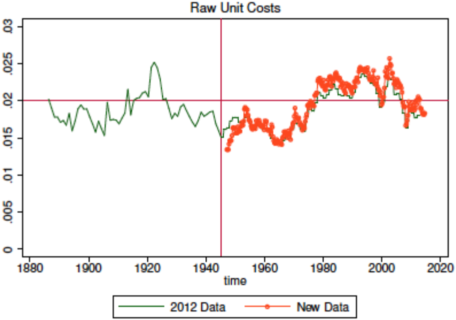
- 1,400 FinTech companies with $33 BN in funding.
- Losses from credit card fraud are $31 BN per year.
- FinTech adoption rates:

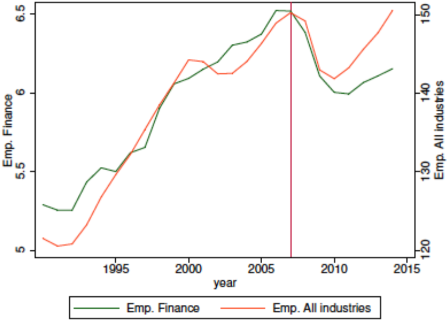


"Using Big Data to Detect Financial Fraud Aided by FinTech Methods"
- S. Srinivasan, Texas Southern U.

Unit Cost of Financial Intermediation

The high cost of financial intermediation is being dis-intermediated by data-driven technologies.





Philippon (2016)

Areas in FinTech 1

- Systemic risk: Espinosa-Vega (2010); Espinosa-Vega and Sola (2010); Billio, Getmansky, Lo, and Pelizzon (2012); Merton, Billio, Getmansky, Gray, Lo, and Pelizzon (2013); and Das (2016).
- Consumer finance: Wei, Yildirim, den Bulte, and Dellarocas (2015), application using social media interactions. Lin, Prabhala, and Viswanathan (2013) exploit friendship networks in peer-lending. Big data helps eliminate bias from small data, see Choudhry, Das, and Hartman-Glaser (2016), ills are outlined in detail in O?Neill (2016).
- Nowcasting: Evans (2005); Giannone, Reichlin, and Small (2008); and Babura, Giannone, Modugno, and Reichlin (2013).
- Text analytics: Das (2014); Jegadeesh and Wu (2013); Loughran and McDonald (2014). Topic analysis, Blei, Ng, and Jordan (2003). Opens up new areas of risk.

Areas in FinTech 2

- Cybersecurity is a massive application area. https://www.sans.org/media/critical-security-controls/critical-controls-poster-2016.pdf.
- Detecting financial fraud.
  - Financial fraud allows perpetrator to be removed from the scene of the crime.
  - Therefore, logging all financial activity to enable traceback is critical.
  - Limited defense at the authentication stage if there is a data breach.
  - Widespread use of machine learning.
  - Social media based; highly consumer-centric. Device usage, Email use, customer location at time of transaction.
  - Adaptive behavioral analytics, e.g., Bionym, EyeVerify, BioCatch.
- Payment systems: Bypassing the banks, e.g., Apple pay, Samsung pay, Google pay, Venmo, Dwolla.

Areas in FinTech 3

- High frequency trading: TradeWorx (http://www.tradeworx.com/) and Automated Trading Desk (ATD, bought by Citibank for $680M in 2007) were pioneers in the field. Algorithmic trading, 50% of executed trades in the equity markets, down from around 2/3 of stock trades in late 2000s, profits from algorithmic trading are under competitive pressure, and regulatory oversight.
  1. Since 2013, 2/3 of the top 30 cited papers on HFTs show positive market effects.
  2. Automated firms reduced trading costs, and contrary to popular opinion, improved market depth and stability.
  3. Much of the research is possible because this sort of FinTech data has become available.
  4. Work by Hendershott and Riordan (HFTs stabilize markets); Hasbrouk and Saar (HFTs improve market quality, reduce bid-ask spreads); Menkveld (HFTs reduce trading costs).

Areas in FinTech 4

- Blockchains. A decentralized record, with copies of the blockchain being maintained by several entities, with (hopefully) comprehensive security and consensus updates.
    1. Acronym DIST (a file that is Distributed, Immutable, Secure, and Trusted).
    2. Banks are experimenting with blockchains for automated settlement, and have formed consortiums such as R3 (https://r3cev.com/).
    3. USC (Utility Settlement Coin) from UBS and three other major banks, as well as SETL coin from Goldman Sachs.

Expected Benefits from Analytics Staff

Managerial focus: Students need to be aligned with top-level goals of the organizations they will go to work for:

- Develop a data-driven decision-making culture.
- Higher customer satisfaction.
- Streamline revenue acquisition; pare down costs.
- Translate small data driven gains into large bottom line because of the volume in banking.
- Profile customers and achieve better customer and price segmentation.
- Role of IT: not a cost center any more.
- Overall, a more competitive organization.

Pitfalls of Big Data Analytics

- Garbage in, garbage out.
- Collecting too much data and not using it correctly.
- Big data leads to bigger errors if misused.
- Confusing correlation with causality.
- May involve expensive infrastructure.
- Privacy issues.
- Excessive misdirected automation leading to poor client service.

Research Examples

1. Financial networks.
2. Zero-revelation prediction of bank malaise.
3. Market illiquidity.

Midas Project: Overview

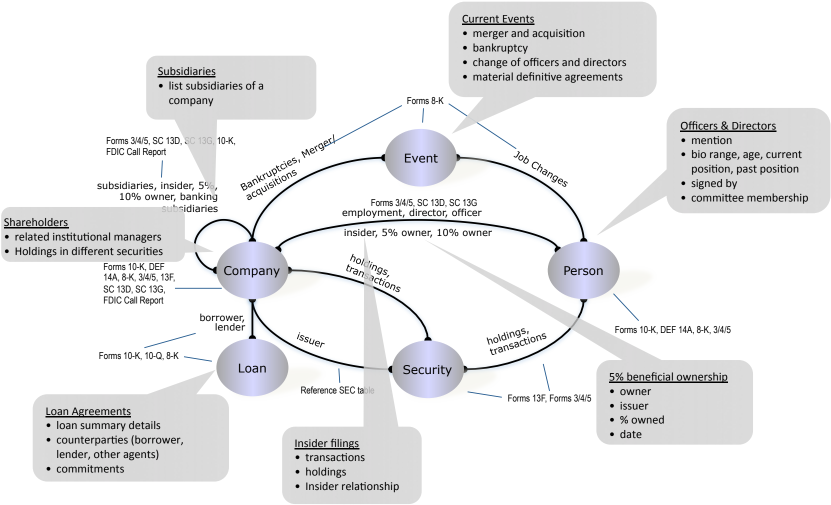
Joint work with IBM Almaden[1]

- Focus on financial companies that are the domain for systemic risk (SIFIs).
- Extract information from unstructured text (filings).
- Information can be analyzed at the institutional level or aggregated system-wide.
- Applications: Systemic risk metrics; governance.
- Technology: information extraction (IE), entity resolution, mapping and fusion, scalable Hadoop architecture.

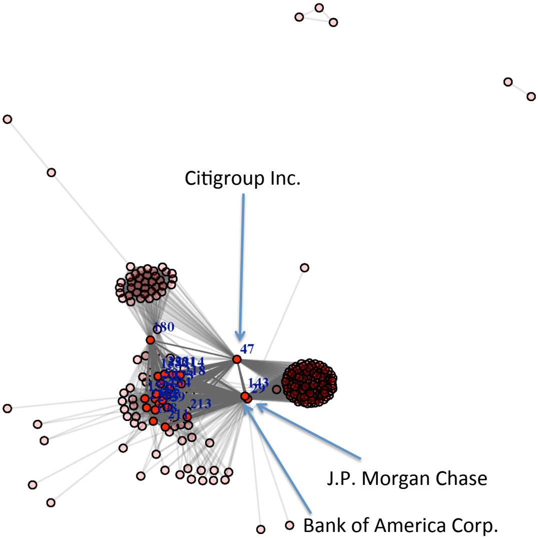
[1] "Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study," (2011), (with Douglas Burdick, Mauricio A. Hernandez, Howard Ho, Georgia Koutrika, Rajasekar Krishnamurthy, Lucian Popa, Ioana Stanoi, Shivakumar Vaithyanathan), *IEEE Data Engineering Bulletin*, 34(3), 60-67. [Proceedings WWW2010, April 26-30, 2010, Raleigh, North Carolina.]

Data

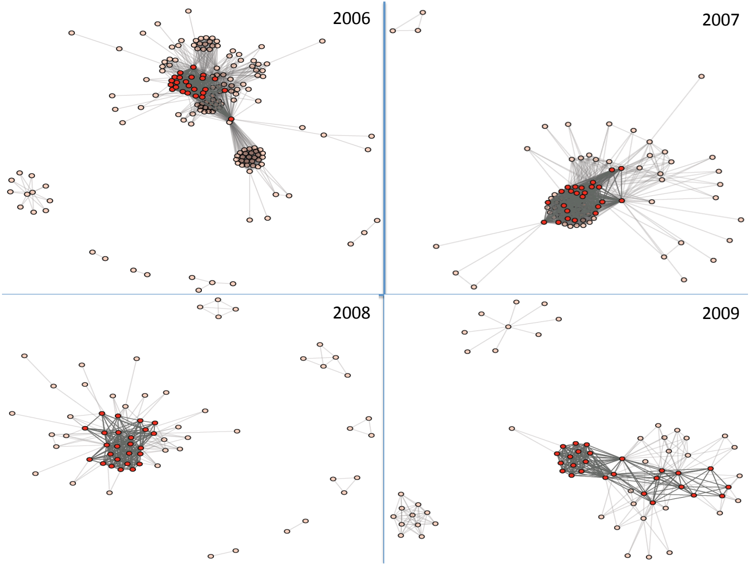
Midas provides Analytical Insights into company relationships by exposing information concepts and relationships within extracted concepts



Loan Network 2005



Loan Network 2006–2009



Systemically Important Financial Institutions (SIFIs)

Year	#Colending banks	#Coloans	Colending pairs	$R = E(d^2)/E(d)$	Diam.
2005	241	75	10997	137.91	5
2006	171	95	4420	172.45	5
2007	85	49	1793	73.62	4
2008	69	84	681	68.14	4
2009	69	42	598	35.35	4

(Year = 2005)

Node #	Financial Institution	Normalized Centrality
143	J P Morgan Chase & Co.	1.000
29	Bank of America Corp.	0.926
47	Citigroup Inc.	0.639
85	Deutsche Bank Ag New York Branch	0.636
225	Wachovia Bank NA	0.617
235	The Bank of New York	0.573
134	Hsbc Bank USA	0.530
39	Barclays Bank Plc	0.530
152	Keycorp	0.524
241	The Royal Bank of Scotland Plc	0.523
6	Abn Amro Bank N.V.	0.448
173	Merrill Lynch Bank USA	0.374
198	PNC Financial Services Group Inc	0.372
180	Morgan Stanley	0.362
42	Bnp Paribas	0.337
205	Royal Bank of Canada	0.289
236	The Bank of Nova Scotia	0.289
218	U.S. Bank NA	0.284
50	Calyon New York Branch	0.273
158	Lehman Brothers Bank Fsb	0.270
213	Sumitomo Mitsui Banking	0.236
214	Suntrust Banks Inc	0.232
221	UBS Loan Finance Llc	0.221
211	State Street Corp	0.210
228	Wells Fargo Bank NA	0.198

Risk Networks: Definitions and Risk Score

- Assume n nodes, i.e., firms, or "assets."
- Let $E \in R^{n \times n}$ be a well-defined adjacency matrix. This quantifies the influence of each node on another.
- E may be portrayed as a directed graph, i.e., $E_{ij} \neq E_{ji}$. $E_{jj} = 1$; $E_{ij} \in \{0, 1\}$.
- C is a $(n \times 1)$ risk vector that defines the risk score for each asset.
- We define the "risk score" as

$$S = \sqrt{C^\top \; E \; C}$$

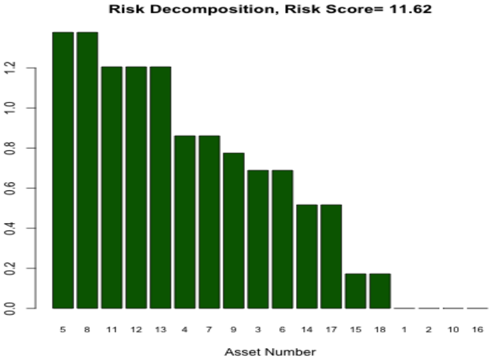
- $S(C, E)$ is linear homogenous in C.

Risk Decomposition

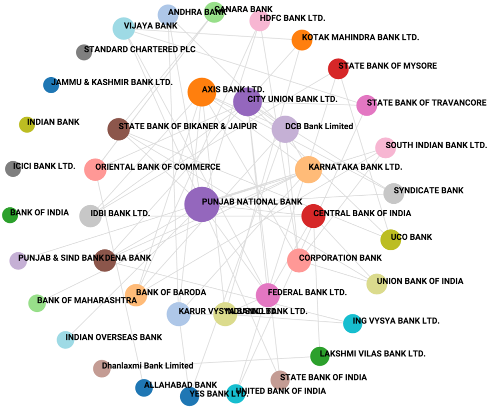
1. Exploits the homogeneity of degree one property of S.
2. Risk decomposition (using Euler's formula):

$$S = \frac{\partial S}{\partial C_1} C_1 + \frac{\partial S}{\partial C_2} C_2 + \ldots + \frac{\partial S}{\partial C_n} C_n$$

3. Plot:



Systemic Risk in Indian Banks



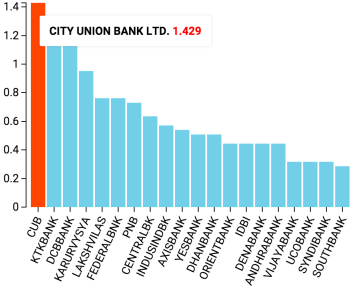
Fragility

2.91

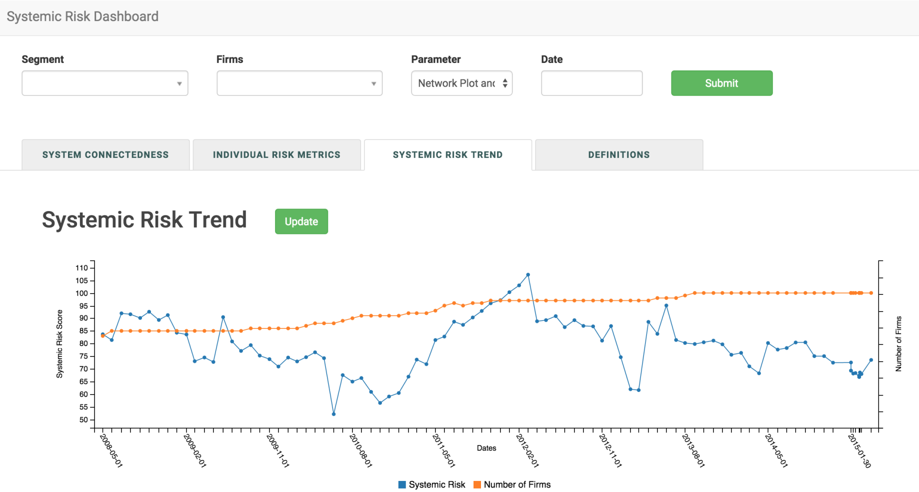
Systemic Risk Score

15.75

Risk Decomposition



Systemic Risk in India over time



Stochastic Risk Networks in a Structural Framework

- We use the Merton (1974) model to extend the static Das (2016) model to a stochastic network setting (Das, Kim, Ostrov, 2016, wip).
- We extend each node's properties to including size, in addition to the credit score.
- To do this we normalize the S measure.
- This model can be calibrated using the same methods used for the Merton model, or variants such as the Moody's KMV model.

Definitions

Model Data (standard Merton model inputs) for each firm:

- Equity price $= \mathbf{s} = \{s_1, s_2, ..., s_n\}$
- Equity volatility $= \boldsymbol{\sigma} = \{\sigma_1, \sigma_2, ..., \sigma_n\}$
- Number of shares $= \mathbf{m} = \{m_1, m_2, ..., m_n\}$
- Risk free rate $= r$

Model Variables (all derived from the Merton model):

- n = number of banks in the system
- $\mathbf{a}$ = n-vector with components a_i that represent the assets in bank i (derived from s, σ, m, r).
- $\boldsymbol{\lambda}$ = n-vector with components λ_i that represent the average yearly chance of bank i defaulting (from s, σ, r).
- $\mathbf{E} = n \times n$ matrix with components E_{ij} that represent the probability that if bank j defaults, it will cause bank i to default (from s, σ, r).

Model

- Define $\mathbf{c}$ to be an n-vector with components c_i that represent bank i's credit risk. More specifically, we define

$$\mathbf{c} = \mathbf{a} \odot \boldsymbol{\lambda},$$

  where $\odot$ represents component multiplication; that is, $c_i = a_i \lambda_i$.
- The aggregate systemic risk created by the n banks in our system is

$$R = \frac{\sqrt{\mathbf{c}^\top \mathbf{E} \mathbf{c}}}{\mathbf{1}^\top \mathbf{a}}, \tag{1}$$

  where $\mathbf{1}$ is an n-vector of ones, so the denominator $\mathbf{1}^\top \mathbf{a} = \sum_{i=1}^{n} a_i$ represents the total assets in the n banks.
- r is linear homogenous in $\boldsymbol{\lambda}$.

Network of Top 50 Financial Institutions

Dynamic Risk Networks (2016)

Upload a .csv file with data for banks (in columns) and attributes (in rows) Click here to see format

File input

Choose File no file selected

Compute Scores

Systemic Risk Score

5.5667

Please refer to following Paper published for some details Matrix Metrics: Network-Based Systemic Risk Scoring

Avg Correlation:

0 0.5 1

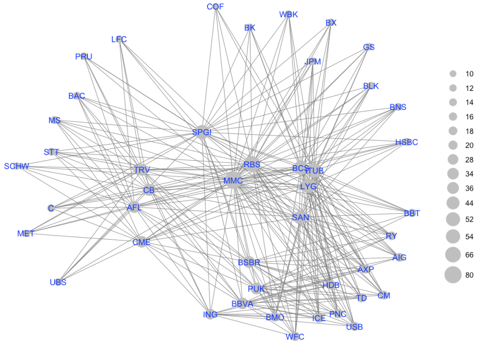
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Correlation threshold:

0 0.5 1

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Network Graph | Risk Decomposition



Computational Properties

$$R = \frac{\sqrt{\mathbf{c}^\top \mathbf{E} \mathbf{c}}}{\mathbf{1}^\top \mathbf{a}}, \quad \mathbf{c} = \mathbf{a} \odot \boldsymbol{\lambda}$$

- R is linear homogeneous in $\boldsymbol{\lambda}$: Let α be any scalar constant. If we replace $\boldsymbol{\lambda}$ with $\alpha\boldsymbol{\lambda}$, it immediately follows that **c** is replaced by $\alpha\mathbf{c}$, and, by our equation for R, we see that R is replaced by αR.
- Sensitivity of R to changes in $\boldsymbol{\lambda}$: Differentiating our equation for R with respect to $\boldsymbol{\lambda}$

$$\frac{\partial R}{\partial \boldsymbol{\lambda}} = \frac{1}{2} \frac{\mathbf{a} \odot [(\mathbf{E} + \mathbf{E}^T)\mathbf{c}]}{\mathbf{1}^T \mathbf{a} \sqrt{\mathbf{c}^T \mathbf{E} \mathbf{c}}}$$

  whose components represent the sensitivity of R to changes in each bank's value of λ. This is the basis of **Risk Decomposition**, equal to $\left(\frac{\partial R}{\partial \boldsymbol{\lambda}} \cdot \boldsymbol{\lambda}\right)$, a vector containing each bank's contribution to R.

Risk Decomposition: 50 Financial Institutions

Dynamic Risk Networks (2016)

Upload a .csv file with data for banks (in columns) and attributes (in rows) Click here to see format

File input

Choose File no file selected

Compute Scores

Systemic Risk Score

5.5667

Please refer to following Paper published for some details Matrix Metrics: Network-Based Systemic Risk Scoring

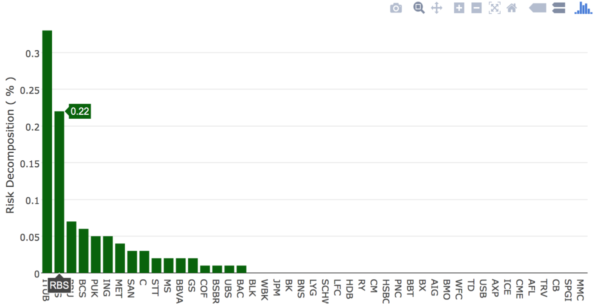
Avg Correlation:

0 0.5 1

Correlation threshold:

0 0.5 1

Network Graph | Risk Decomposition

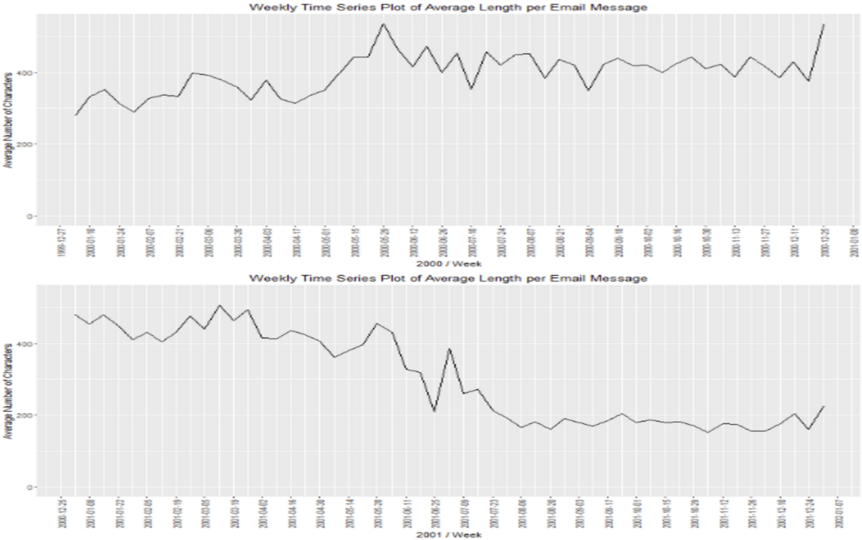


Analyzing Emails for Early Warnings of Failure

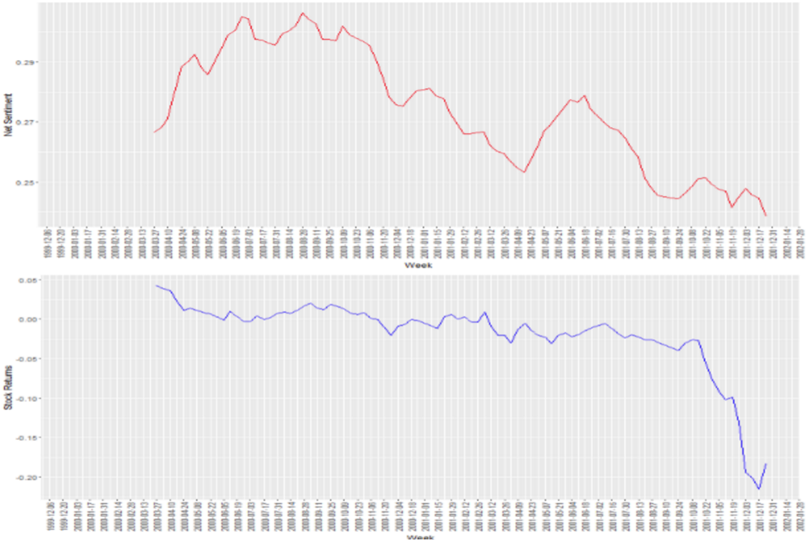
Title: "Zero-Revelation Linguistic Regulation: Detecting Risk Through Corporate Emails and News"

- Financials are often delayed indicators of corporate quality.
- Internal discussion may be used as an early warning system for upcoming corporate malaise.
- Emails have the potential to predict such events.
- Software can analyze vast quantities of textual data not amenable to human processing.
- Corporate senior management may also use these analyses to better predict and manage impending crisis for their firms.
- The approach requires zero revelation of emails.

Enron: Email Length



Enron: Sentiment and Returns



Enron: Returns and Characteristics

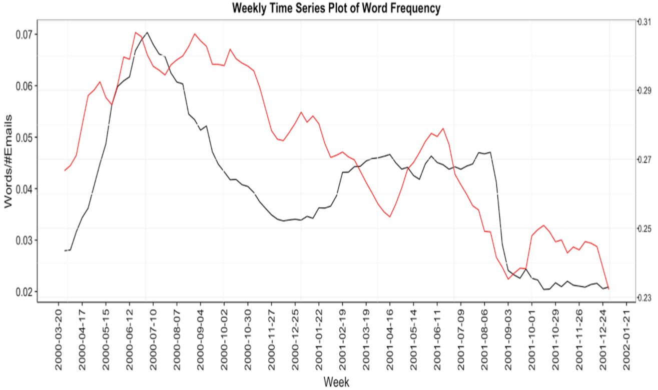
Variable	Coefficient Estimate (t-statistic)			
	(1)	(2)	(3)	(4)
MA Net Sentiment $_t$	XXX***	0.575	2.330***	-1.397
	(XXX)	(0.63)	(3.14)	(-1.25)
MA Email Length $_t$		0.584***		1.046***
		(2.97)		(4.19)
MA Total Emails $_t$			-0.004	-0.131***
			(-0.10)	(-2.83)
Intercept		-0.406*	-0.671***	0.117
		(-1.93)	(-3.08)	(0.43)
Adjusted R-squared	XXX		0.09	0.24
Number of observations	88	88	88	88

Enron: WordPlay

Select Word

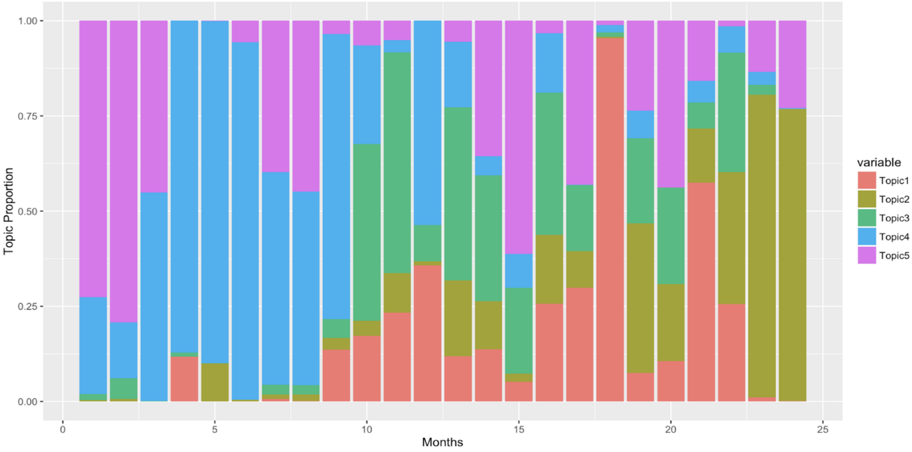
credit

Word over time | Word_vs_Sentiment | Word_vs_Return | Word_Sentiment_Correlation | Word_Return_Correlation



Black: words/#emails , Red: sentiment

Enron: Topic Analysis



BILLIQ: An Index-Based Measure of Illiquidity

Uses option pricing theory to derive a measure for the cost of immediacy:

$$BILLIQ = -10,000 \times \ln \left[\frac{NAV}{NAV + |ETF - NAV|} \right]$$