Text Mining and Textual Analysis of Corporate Filings for Developing Predictive Models and Risk Assessments

Rajendra P. Srivastava

PhD (Physics, 1972), PhD (Accounting, 1982)

Professor Emeritus, Ex EY Professor, and Ex Director of EY CARAT

University of Kansas; and

CEO, SeekEdgar, LLC

rsrivastava@ku.edu

48th World Continuous Auditing and Reporting Symposium

Banco de España • Online via Webex • Madrid, Spain

15:30 Session 3

Thursday, September 24, 2020

Outline

- Text Mining
- Textual Analysis
  - Counts: Word, Sentence, Phrases, & Proximity Counts
  - Readability Indices
  - Risk Sentiment (overall risk, financial risk, litigation risk, tax risk, etc.)
  - Competition Metric
  - Cosine similarity measure
  - Word variation over time
  - Sentiment analysis (Positive, Negative, Sentiments Spread)
- Financial Fraud Assessment Models
- Conclusion

Textual Analysis: *Seven Readability Indices*

Readability Indices

1. Gunning-Fog Index https://en.wikipedia.org/wiki/Gunning_fog_index
2. Smog Index https://en.wikipedia.org/wiki/SMOG
3. Flesch Reading Ease https://en.wikipedia.org/wiki/Flesch–Kincaid_readability_tests
4. Flesch-Kincaid Grade Level https://en.wikipedia.org/wiki/Flesch–Kincaid_readability_tests
5. Automated Readability Index https://en.wikipedia.org/wiki/Automated_readability_index
6. Coleman-Liau Index https://en.wikipedia.org/wiki/Coleman–Liau_index
7. Bog Index https://kelley.iu.edu/bpm/activities/bogindex.html

1. Gunning-Fog Index (Robert Gunning, 1952)

Gunning-Fog Index = 0.4[(Words/Sentences) + 100(Complex words/Words)]

- 17 College graduate
- 16 College senior
- - - - -
- 12 High school senior
- - - - -
- 10 High school sophomore
- - - - -
- 6 Sixth grade

7. Bog Index

A plain English measure of financial reporting readability

by

Bonsall IV, Leone, Rennekamp

in

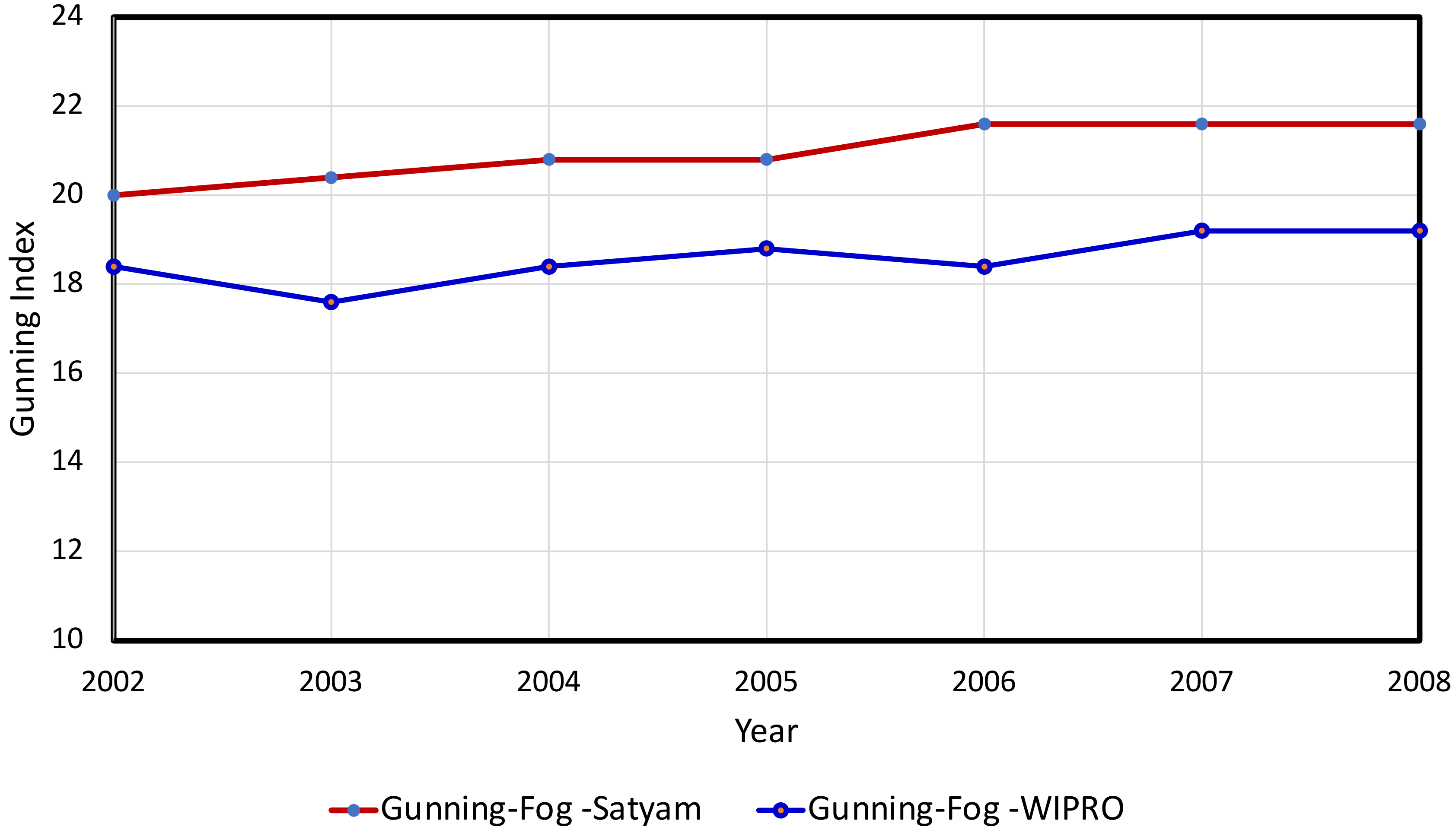
Journal of Accounting and Economics, 63 (2017) pp. 329–357

Example: Readability Indices for Satyam and WIPRO

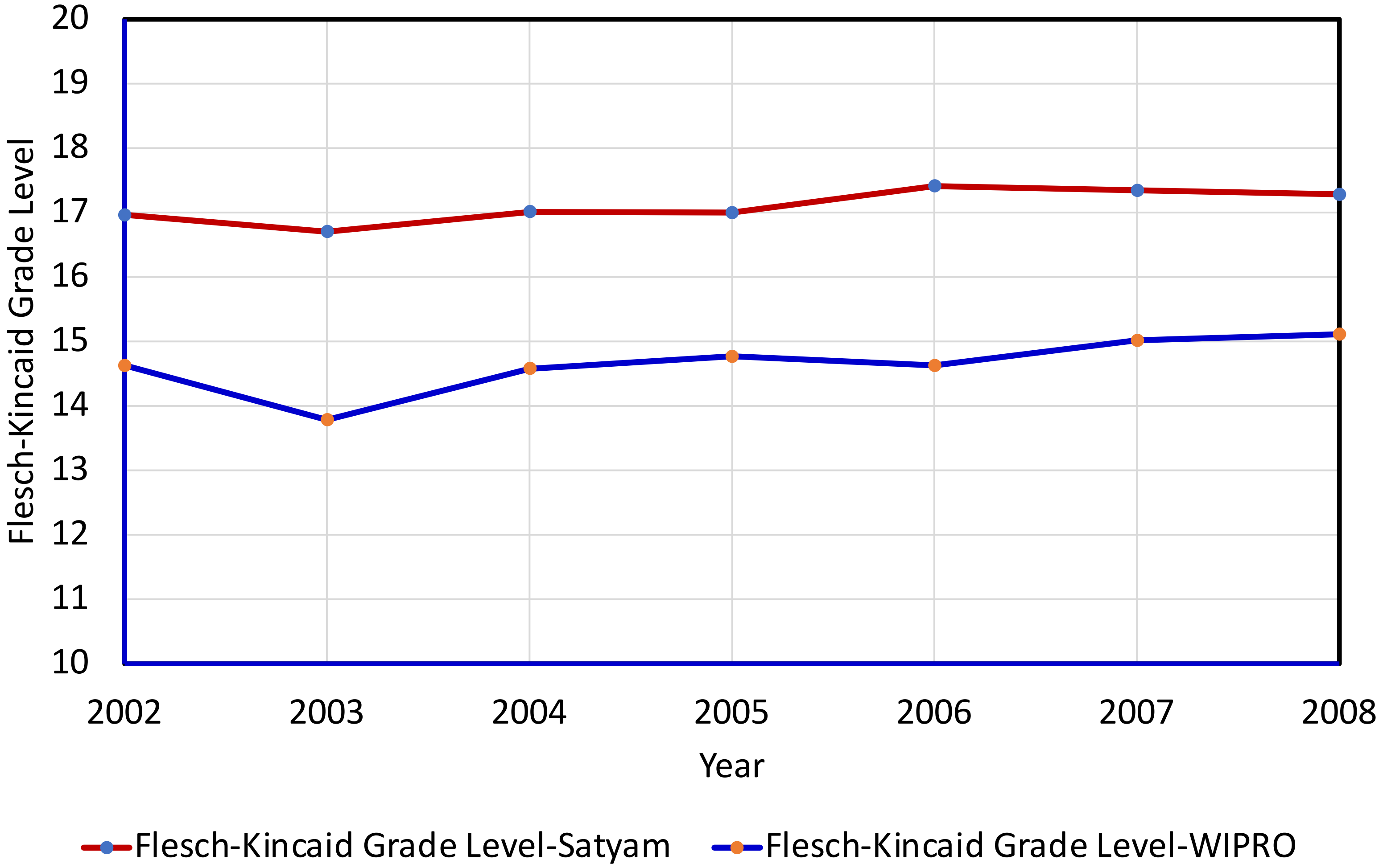
Satyam Textual Analysisç							
Year	2008	2007	2006	2005	2004	2003	2002
Total Word Count	81258	85673	80785	58473	67858	70837	259828
Total Word Count without numerics	74833	79145	74881	54641	60675	63526	227833
Sentence Count	2642	2770	2575	1966	2175	2368	5770
Gunning-Fog Index	21.6	21.6	21.6	20.8	20.8	20.4	20
Smog Index	18.666	18.762	18.73	18.459	18.394	18.18	13.618
Flesch Reading Ease	21.777	21.777	22.212	22.893	22.92	23.236	51.699
Flesch-Kincaid Grade Level	17.281	17.344	17.411	17.001	17.014	16.704	16.962
Automated Readability Index	17.759	17.819	17.908	17.316	17.383	16.964	13.404
Coleman-Liau Index	14.439	14.357	14.145	14.2	14.239	14.386	0.293

WIPRO LTD							
Year	2008	2007	2006	2005	2004	2003	2002
Total Word Count	93966	99464	96763	101922	87781	75005	120396
Total Word Count without numerics	85584	90570	88177	93798	78915	66793	104844
Sentence Count	3624	3894	3865	4080	3511	3290	4656
Gunning-Fog Index	19.2	19.2	18.4	18.8	18.4	17.6	18.4
Smog Index	17.059	16.935	16.644	16.797	16.625	16.004	16.688
Flesch Reading Ease	28.928	28.956	30.938	30.24	30.684	32.505	30.451
Flesch-Kincaid Grade Level	15.113	15.017	14.628	14.771	14.579	13.785	14.629
Automated Readability Index	15.082	15.012	14.573	14.709	14.397	13.423	14.471
Coleman-Liau Index	14.04	14.182	13.916	13.97	13.9	14.05	13.959

Example: Graph of Readability Indices for Satyam and WIPRO



Example: Graph of Readability Indices for Satyam and WIPRO



Risk Sentiment measure by Feng Li

Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports?

Definition of Risk Sentiment:

- $RS_t = \ln(1+NR_t)$

Change of risk sentiment as

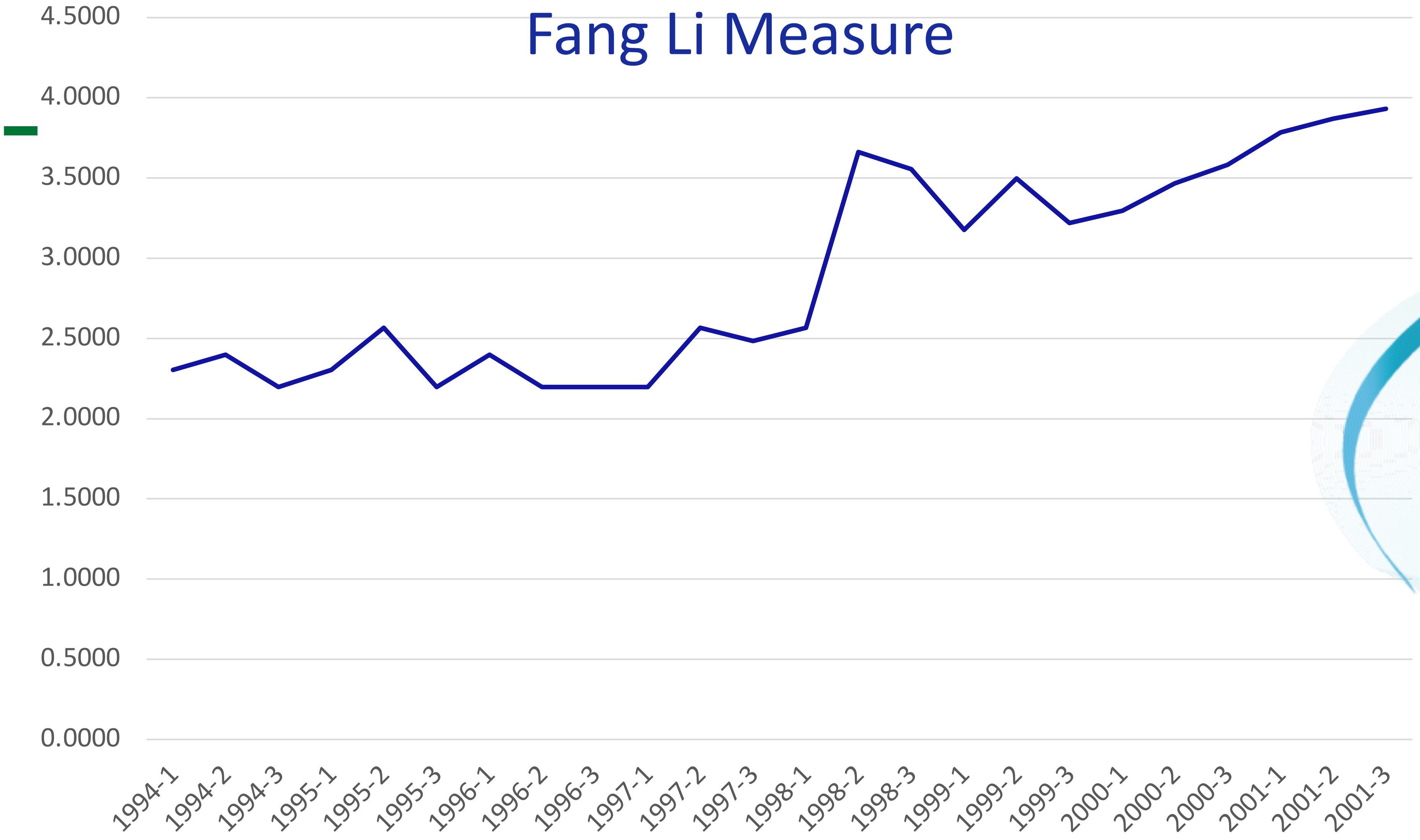
- $\Delta RS_t = \ln(1+NR_t) - \ln(1 + NR_{t-1})$

where NR_t and NR_{t-1} are the numbers of occurance of risk-related words in year t and year $t - 1$ respectively.

- risk", "risks", "risky", "uncertain", "uncertainty", and "uncertainties

Enron Risk Sentiments = RSt = Ln(1+NRt)

Fang Li Measure



The information content of mandatory risk factor disclosures in corporate filings (Item 1A)

by

John L. Campbell • Hsinchun Chen •
Dan S. Dhaliwal • Hsin-min Lu • Logan B. Steele

In

Rev Account Stud (2014) 19:396–455

Word List for Financial Risk

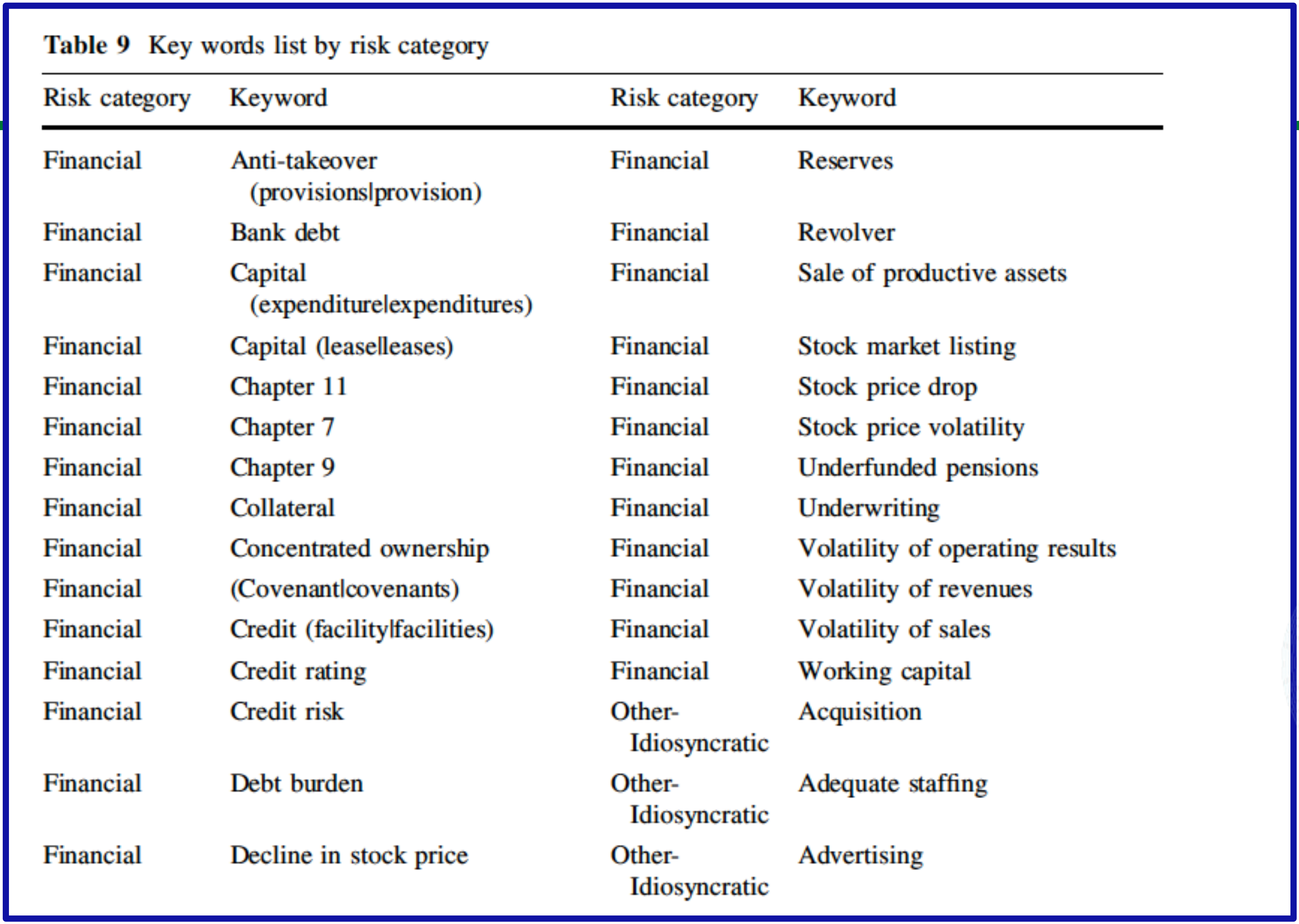


Table 9 Key words list by risk category

Risk category	Keyword	Risk category	Keyword
Financial	Anti-takeover (provisions|provision)	Financial	Reserves
Financial	Bank debt	Financial	Revolver
Financial	Capital (expenditure|expenditures)	Financial	Sale of productive assets
Financial	Capital (lease|leases)	Financial	Stock market listing
Financial	Chapter 11	Financial	Stock price drop
Financial	Chapter 7	Financial	Stock price volatility
Financial	Chapter 9	Financial	Underfunded pensions
Financial	Collateral	Financial	Underwriting
Financial	Concentrated ownership	Financial	Volatility of operating results
Financial	(Covenant|covenants)	Financial	Volatility of revenues
Financial	Credit (facility|facilities)	Financial	Volatility of sales
Financial	Credit rating	Financial	Working capital
Financial	Credit risk	Other-Idiosyncratic	Acquisition
Financial	Debt burden	Other-Idiosyncratic	Adequate staffing
Financial	Decline in stock price	Other-Idiosyncratic	Advertising

Word List for Litigation Risk

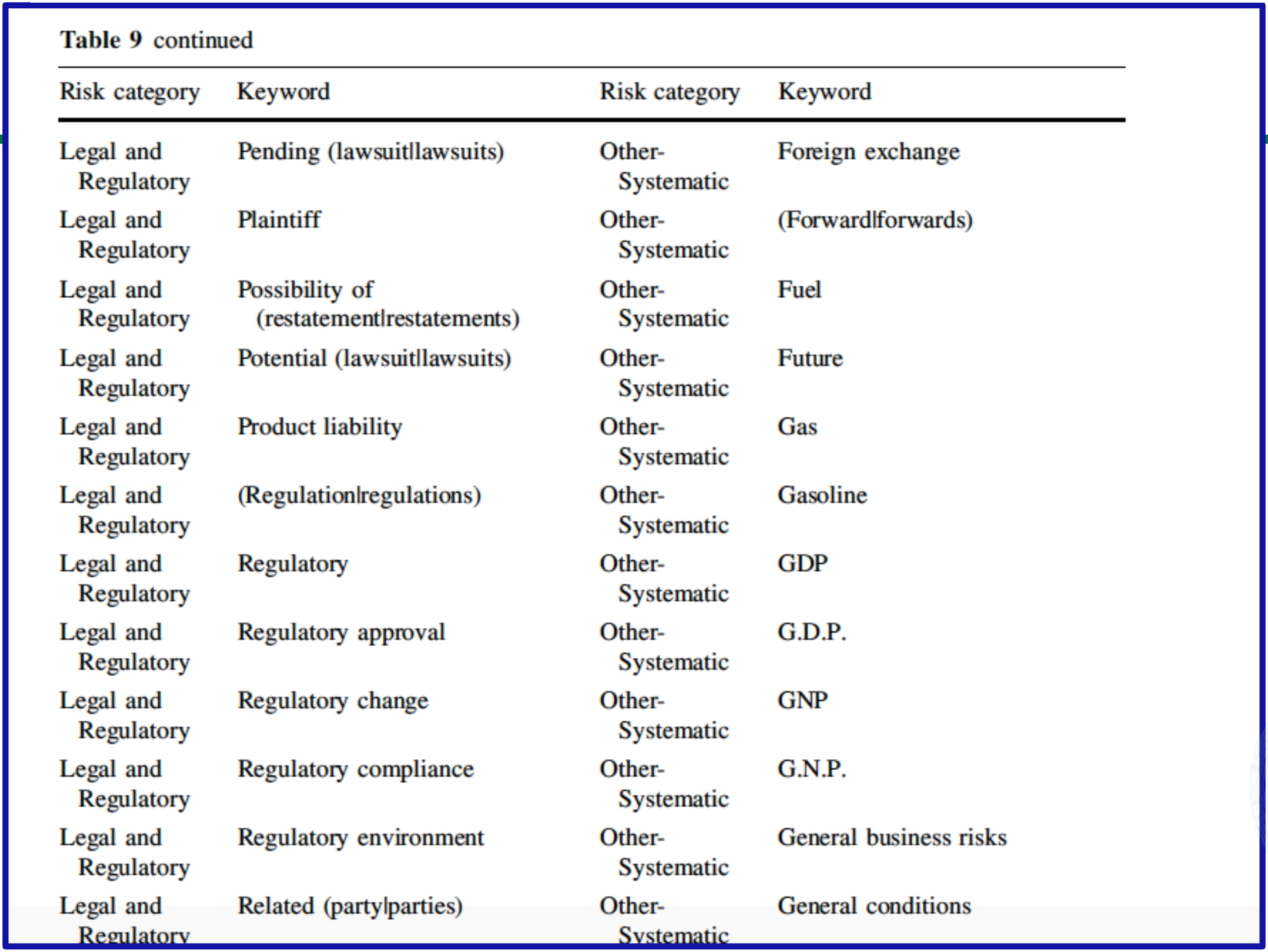
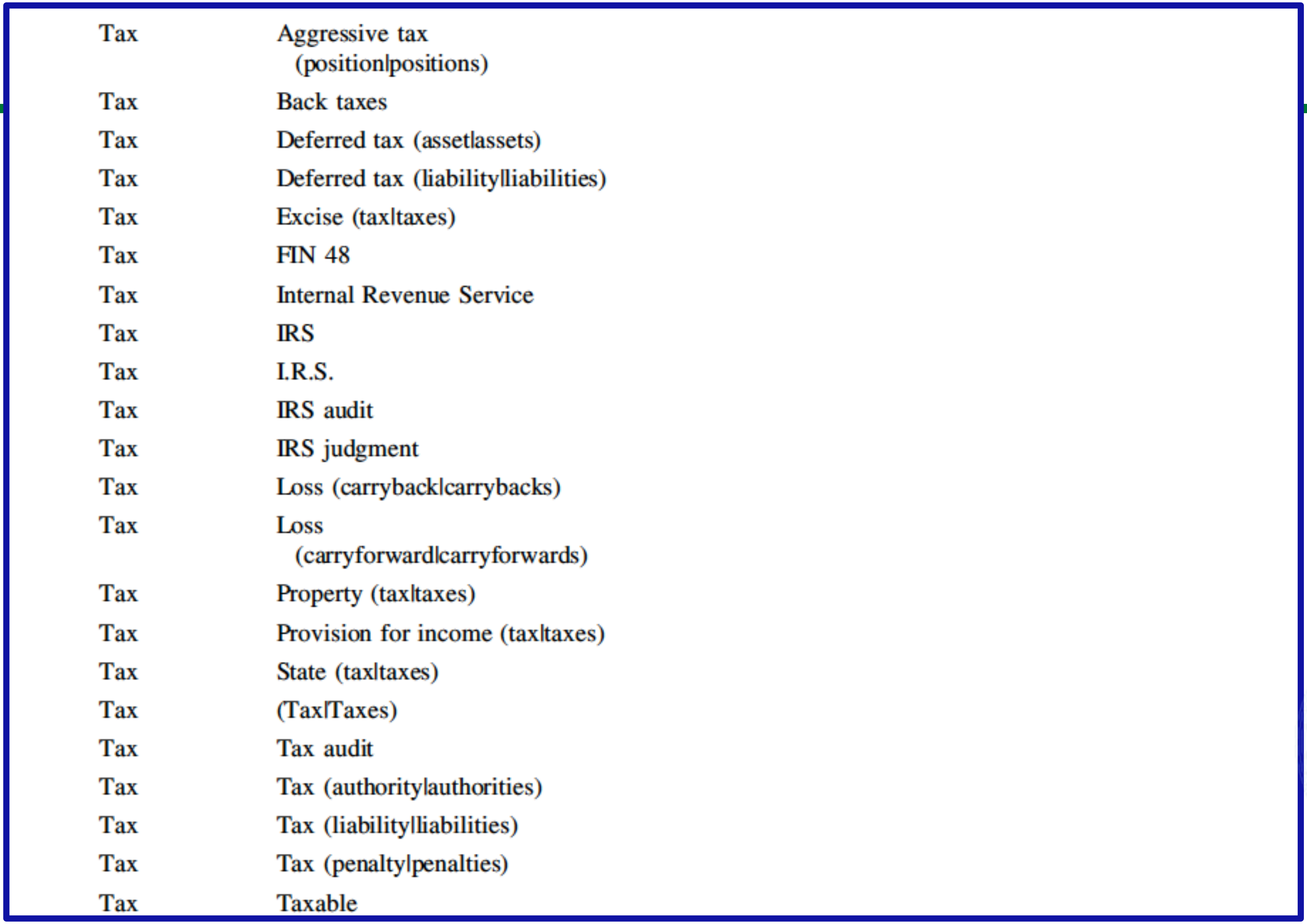


Table 9 continued

Risk category	Keyword	Risk category	Keyword
Legal and Regulatory	Pending (lawsuit|lawsuits)	Other-Systematic	Foreign exchange
Legal and Regulatory	Plaintiff	Other-Systematic	(Forward|forwards)
Legal and Regulatory	Possibility of (restatement|restatements)	Other-Systematic	Fuel
Legal and Regulatory	Potential (lawsuit|lawsuits)	Other-Systematic	Future
Legal and Regulatory	Product liability	Other-Systematic	Gas
Legal and Regulatory	(Regulation|regulations)	Other-Systematic	Gasoline
Legal and Regulatory	Regulatory	Other-Systematic	GDP
Legal and Regulatory	Regulatory approval	Other-Systematic	G.D.P.
Legal and Regulatory	Regulatory change	Other-Systematic	GNP
Legal and Regulatory	Regulatory compliance	Other-Systematic	G.N.P.
Legal and Regulatory	Regulatory environment	Other-Systematic	General business risks
Legal and Regulatory	Related (party|parties)	Other-Systematic	General conditions

Word List for Litigation Risk



Tax	Aggressive tax (position|positions)
Tax	Back taxes
Tax	Deferred tax (asset|assets)
Tax	Deferred tax (liability|liabilities)
Tax	Excise (tax|taxes)
Tax	FIN 48
Tax	Internal Revenue Service
Tax	IRS
Tax	I.R.S.
Tax	IRS audit
Tax	IRS judgment
Tax	Loss (carryback|carrybacks)
Tax	Loss (carryforward|carryforwards)
Tax	Property (tax|taxes)
Tax	Provision for income (tax|taxes)
Tax	State (tax|taxes)
Tax	(Tax|Taxes)
Tax	Tax audit
Tax	Tax (authority|authorities)
Tax	Tax (liability|liabilities)
Tax	Tax (penalty|penalties)
Tax	Taxable

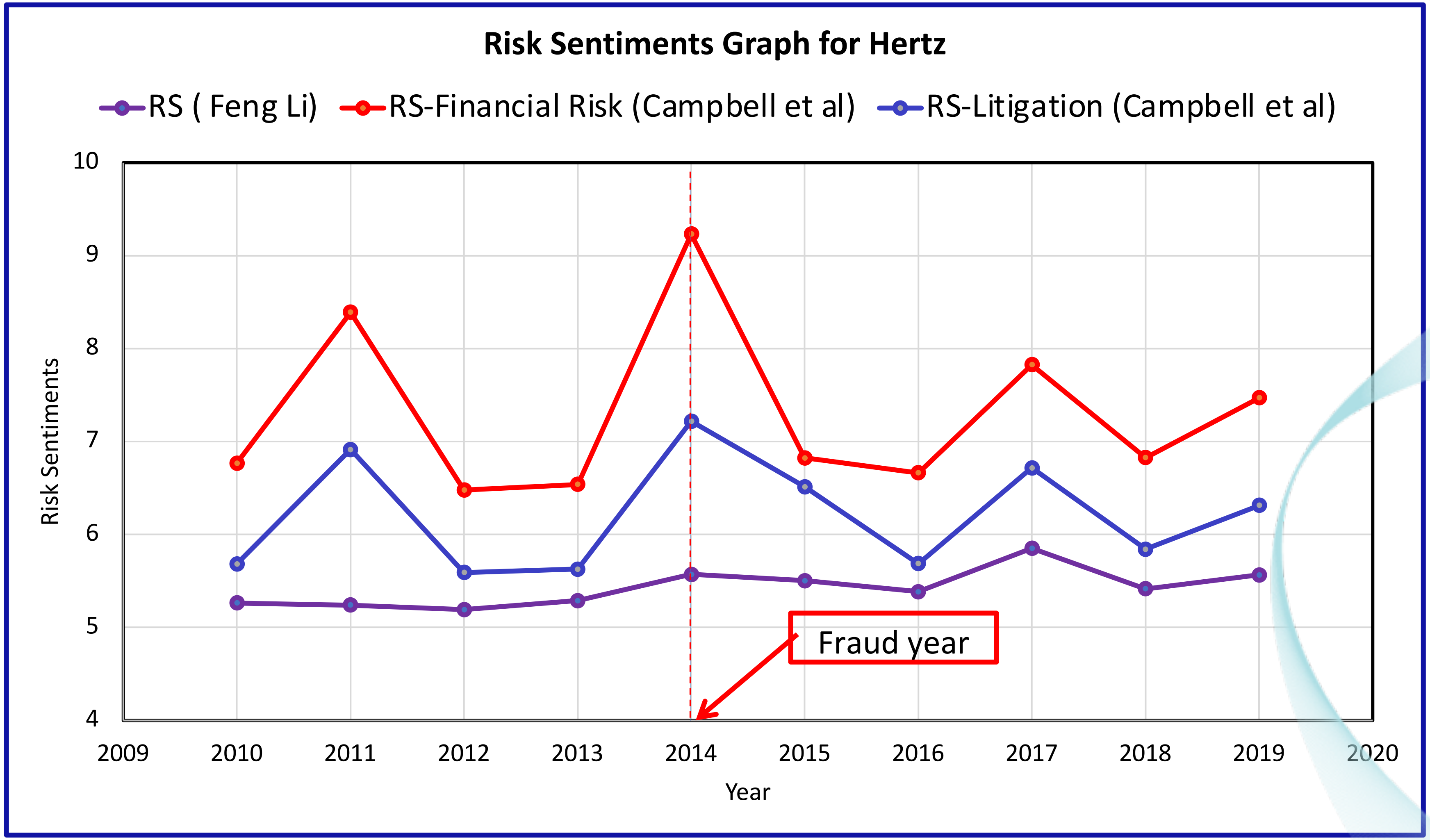
Textual Analysis with More Built-in Features

8. **Risk Sentiment Metrics**

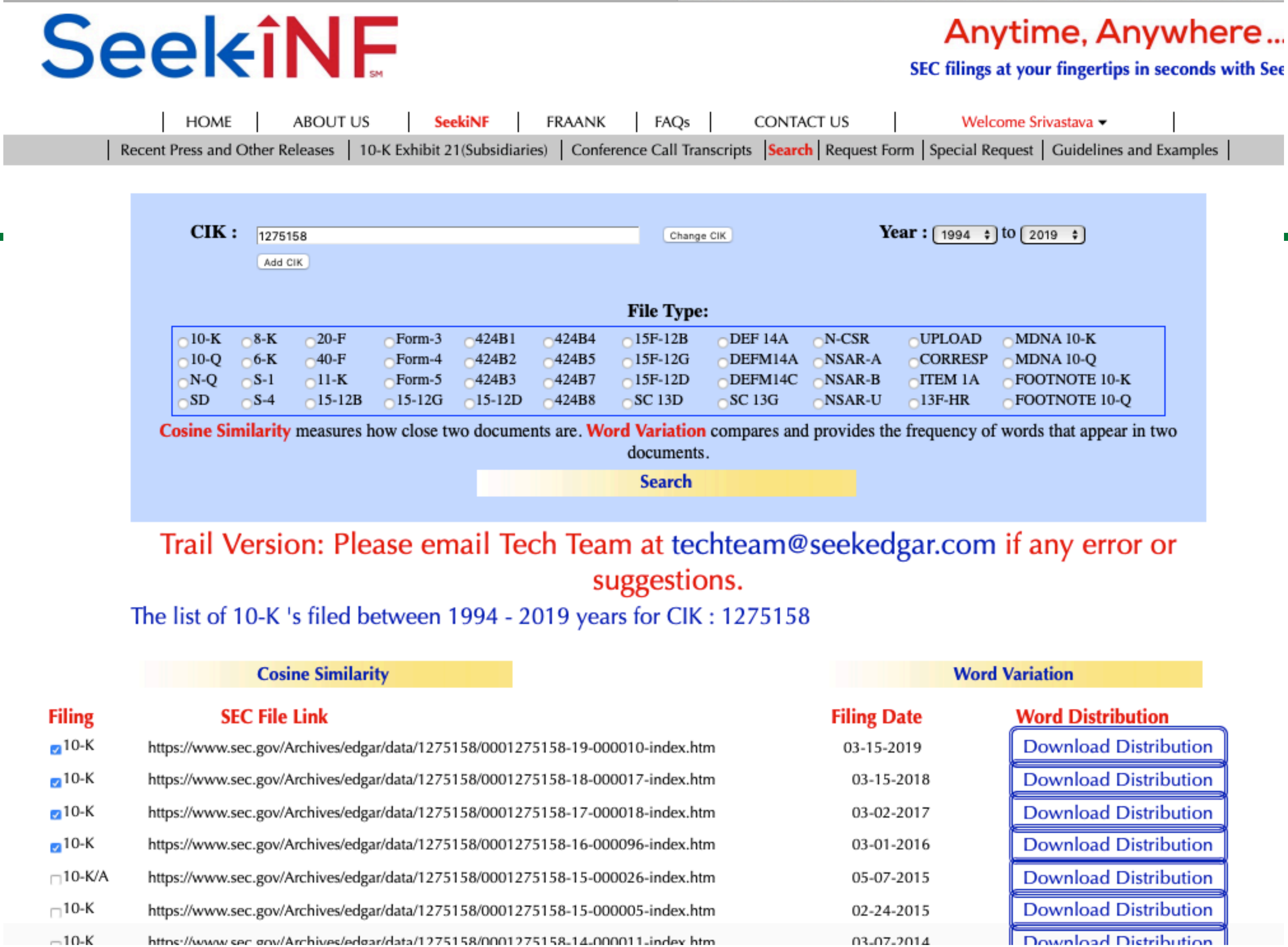
- Risk Sentiment (Feng Li Model) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=898181
- Risk Sentiments (Campbell et al. Model) https://link.springer.com/article/10.1007/s11142-013-9258-3

a. Risk Sentiment (Financial)
b. Risk Sentiment (Legal and Regulatory, i.e., Litigation)
c. Risk Sentiment (Tax)
d. Risk Sentiment (Systematic, economy)
e. Risk Sentiment (Idiosyncratic, specific to firm)
f. Risk Sentiment (Overall)

Risk Sentiments for Hertz Based on 10K



Cosine Measure of Similarity



Anytime, Anywhere...

SEC filings at your fingertips in seconds with See

| HOME | ABOUT US | SeekiNF | FRAANK | FAQs | CONTACT US | Welcome Srivastava ▾ |
| Recent Press and Other Releases | 10-K Exhibit 21(Subsidiaries) | Conference Call Transcripts | Search | Request Form | Special Request | Guidelines and Examples |

CIK : 1275158 Year : 1994 to 2019

File Type:

○ 10-K	○ 8-K	○ 20-F	○ Form-3	○ 424B1	○ 424B4	○ 15F-12B	○ DEF 14A	○ N-CSR	○ UPLOAD	○ MDNA 10-K
○ 10-Q	○ 6-K	○ 40-F	○ Form-4	○ 424B2	○ 424B5	○ 15F-12G	○ DEFM14A	○ NSAR-A	○ CORRESP	○ MDNA 10-Q
○ N-Q	○ S-1	○ 11-K	○ Form-5	○ 424B3	○ 424B7	○ 15F-12D	○ DEFM14C	○ NSAR-B	○ ITEM 1A	○ FOOTNOTE 10-K
○ SD	○ S-4	○ 15-12B	○ 15-12G	○ 15-12D	○ 424B8	○ SC 13D	○ SC 13G	○ NSAR-U	○ 13F-HR	○ FOOTNOTE 10-Q

Cosine Similarity measures how close two documents are. **Word Variation** compares and provides the frequency of words that appear in two documents.

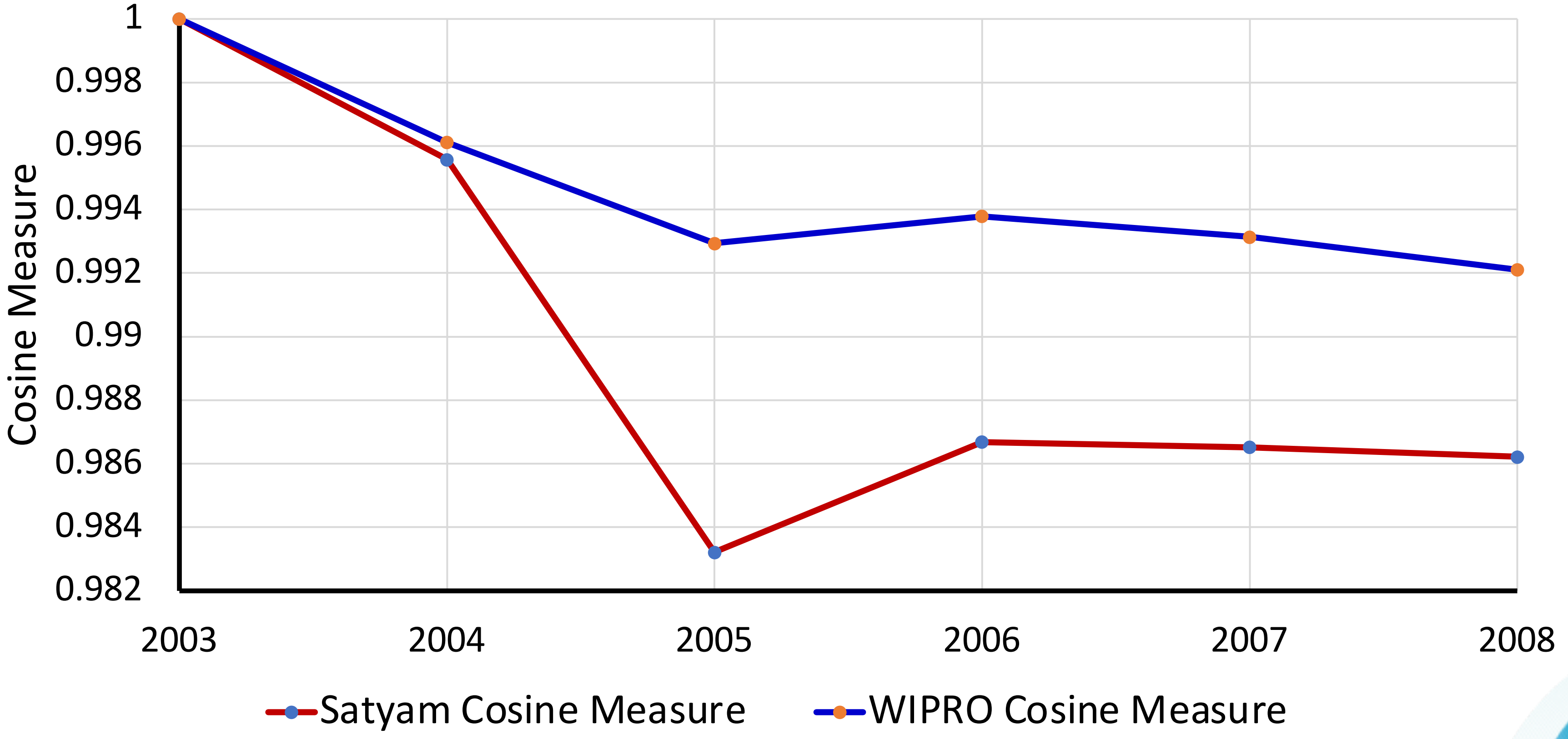
Trail Version: Please email Tech Team at techteam@seekedgar.com if any error or suggestions.

The list of 10-K 's filed between 1994 - 2019 years for CIK : 1275158

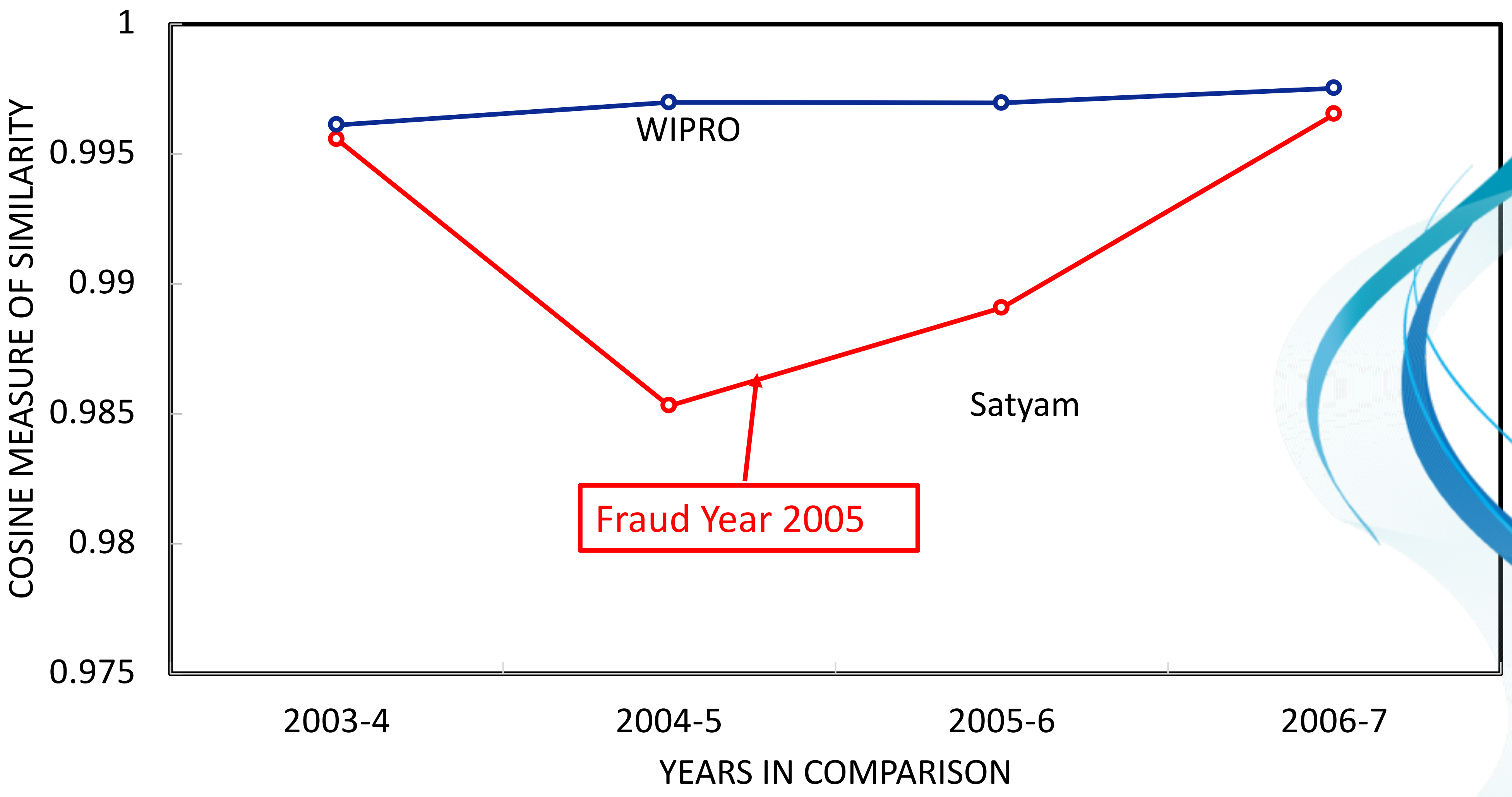
Filing	SEC File Link	Filing Date	Word Distribution
☑ 10-K	https://www.sec.gov/Archives/edgar/data/1275158/0001275158-19-000010-index.htm	03-15-2019	Download Distribution
☑ 10-K	https://www.sec.gov/Archives/edgar/data/1275158/0001275158-18-000017-index.htm	03-15-2018	Download Distribution
☑ 10-K	https://www.sec.gov/Archives/edgar/data/1275158/0001275158-17-000018-index.htm	03-02-2017	Download Distribution
☑ 10-K	https://www.sec.gov/Archives/edgar/data/1275158/0001275158-16-000096-index.htm	03-01-2016	Download Distribution
□ 10-K/A	https://www.sec.gov/Archives/edgar/data/1275158/0001275158-15-000026-index.htm	05-07-2015	Download Distribution
□ 10-K	https://www.sec.gov/Archives/edgar/data/1275158/0001275158-15-000005-index.htm	02-24-2015	Download Distribution
□ 10-K	https://www.sec.gov/Archives/edgar/data/1275158/0001275158-14-000011-index.htm	03-07-2014	Download Distribution

Graph of Cosine Similarity for Satyam and WIPRO

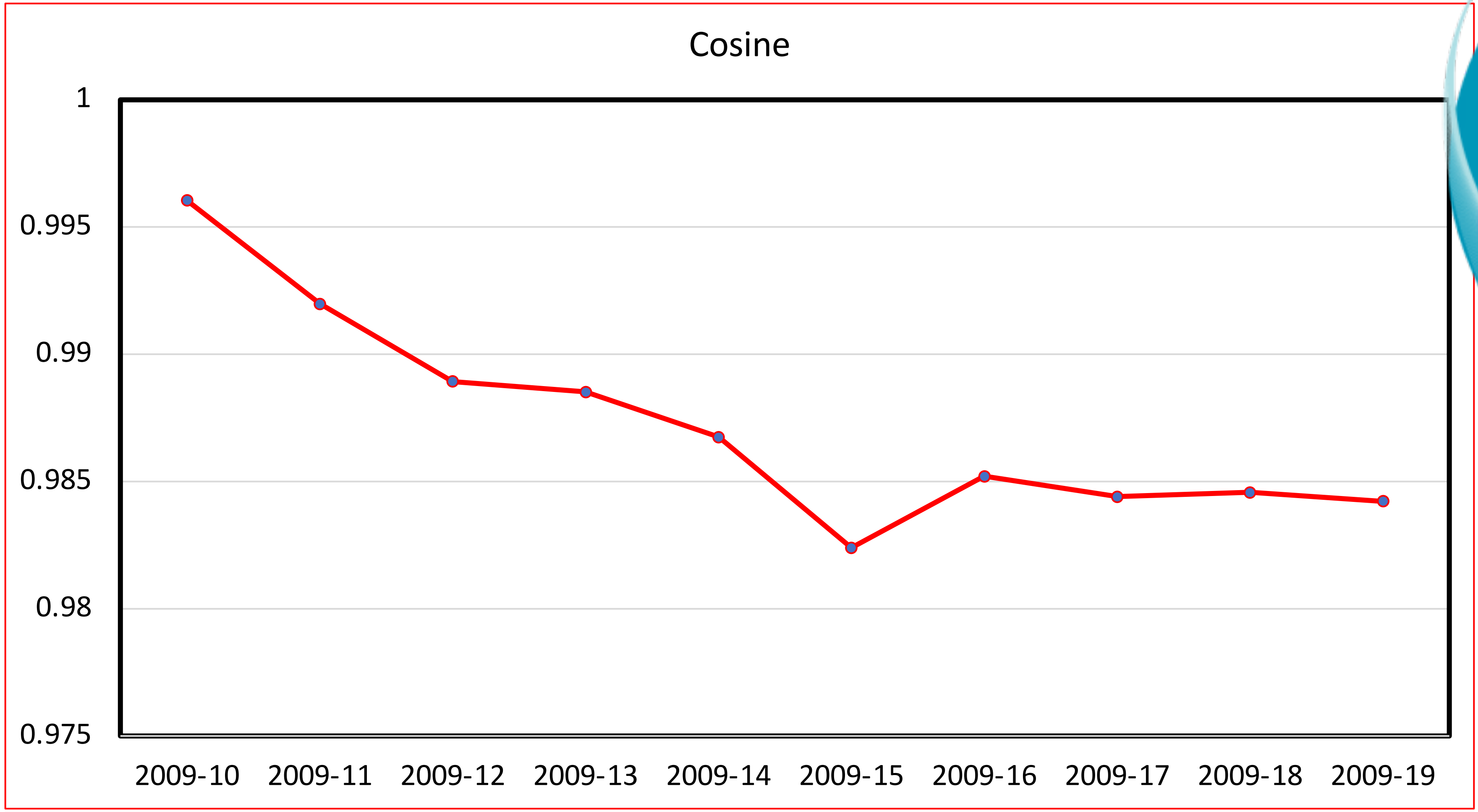
Cosine Measure of Similarity for Satyam and WIPRO in relation to 2003 20-F



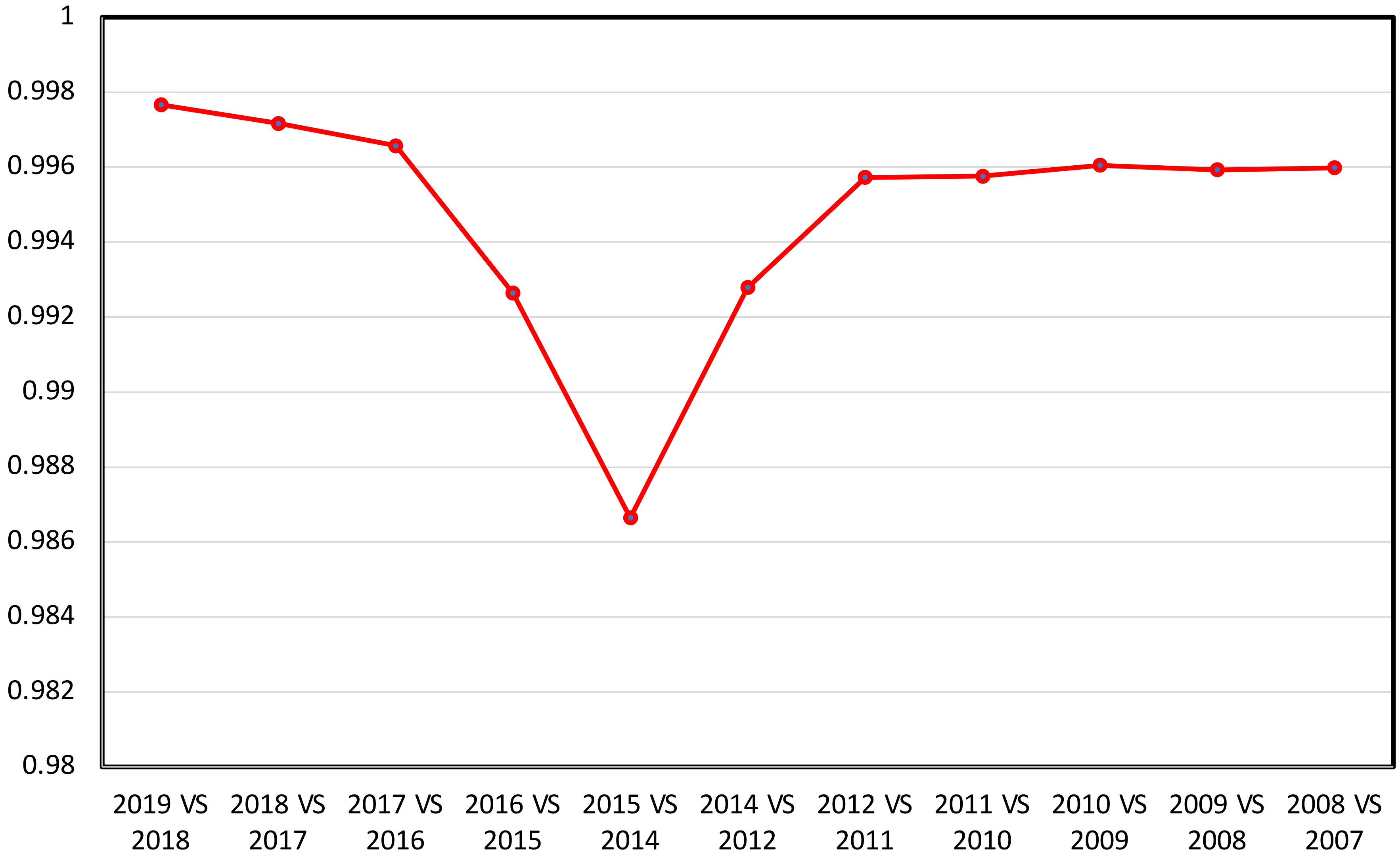
Change in Cosine Measure of Similarity for Satyam and WIPRO by Year (20-F)



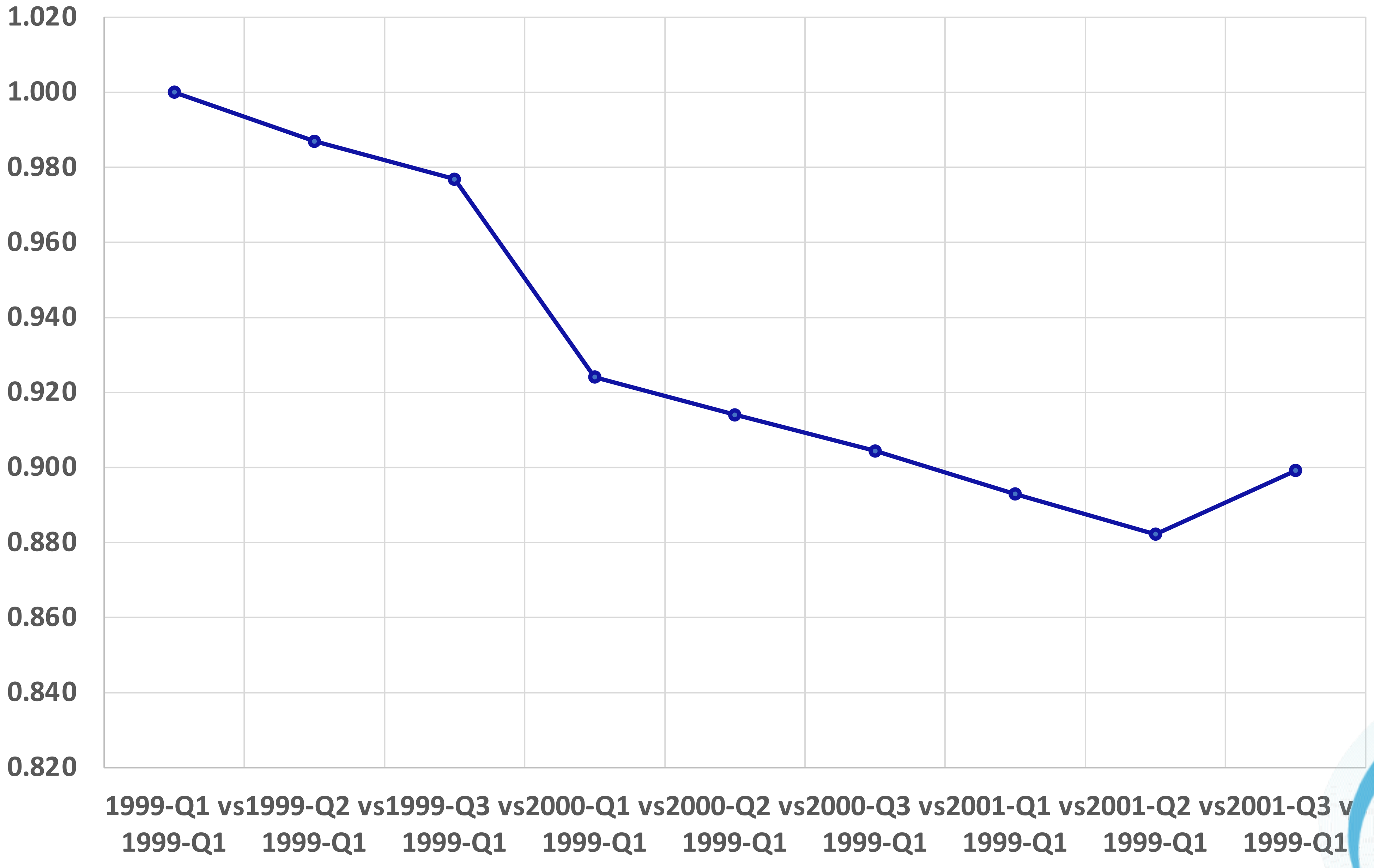
Graph of Cosine Similarity for Bancorp Inc. with respect to 2009 10K



Change in Cosine Similarity Measure for Bancorp



Enron -Cosine Measure w.r.t. 1999 10Q1



Measure of Competition

Measure of Competition

Li, Lundholm, and Minnis *JAR*, 2013, p. 399

Li, Lundholm, and Minnis (2013) develop a model to compute management's perception of the intensity of competition using textual analysis of firms’ 10-K filings.

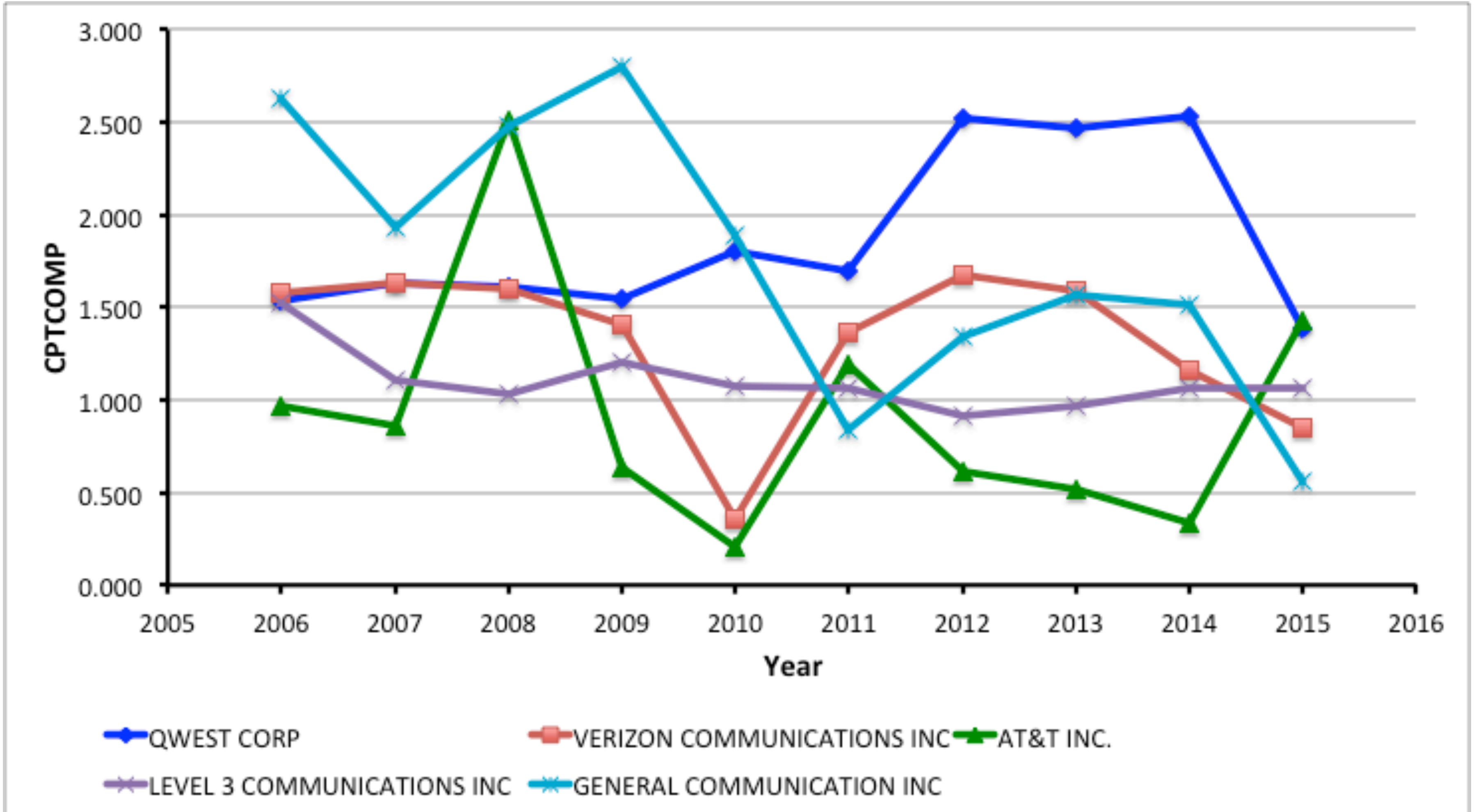
- Measure of competition varies across-industry and within-industry
- It is related to the firm’s future rates of diminishing marginal returns.
- This measure is based on the count of the number of words like “competition, competitor, competitive, compete, competing,” including those words with an "s" appended, less any case where "not," "less," "few," or "limited" precedes the word by three or fewer words.

$$PCTCOMP = 1000*NCOMP/NWORDS$$

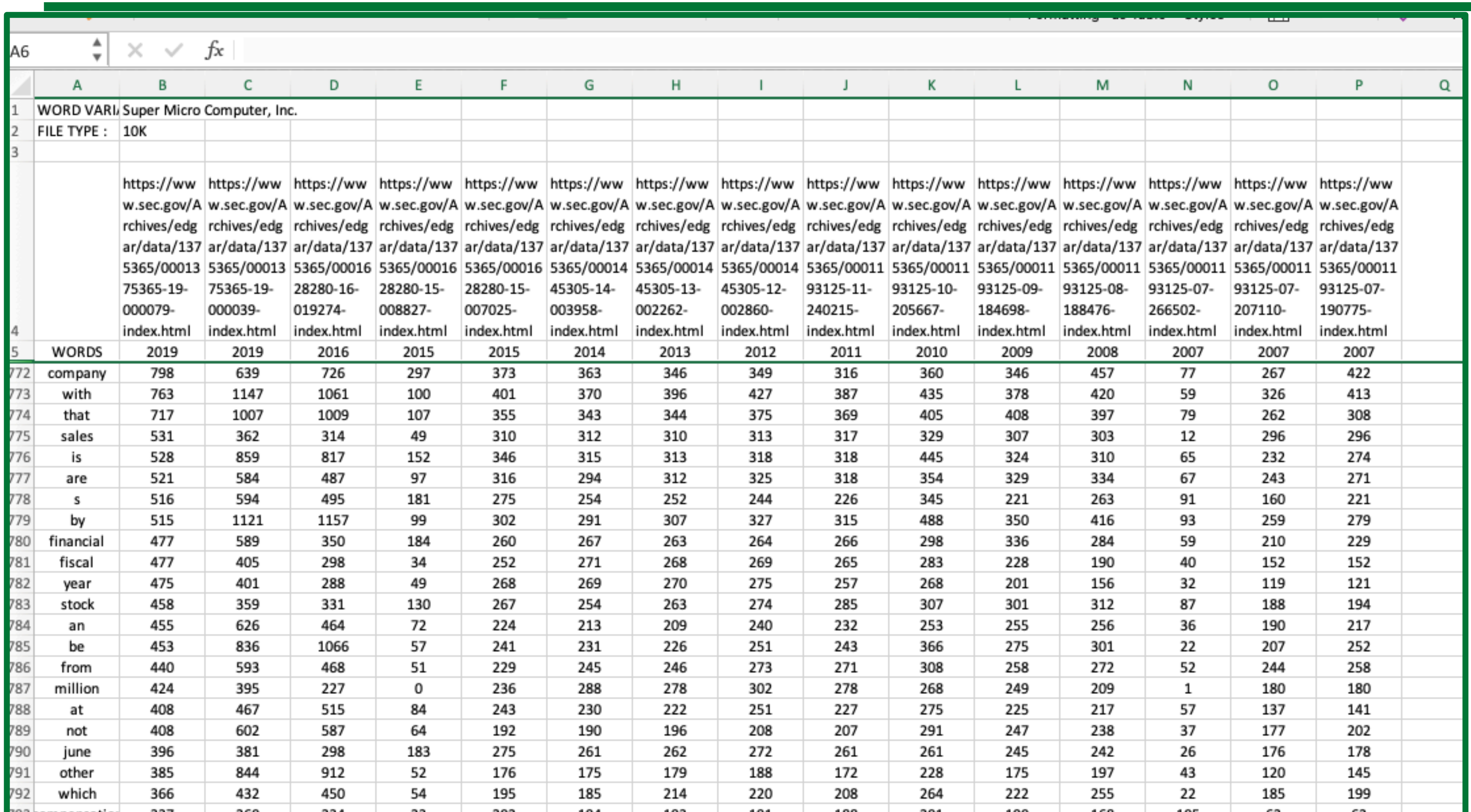
where NCOMP = number of words in 10-K as described above and NWORDS = Total number of words without numbers.

SeekiNF℠

Competition Metric for Five companies for 10 years



10 Years Word Variations in 10K



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	WORD VARI	Super Micro Computer, Inc.															
2	FILE TYPE :	10K															
3																	
4		https://www.sec.gov/Archives/edgar/data/1375365/0001375365-19-000079-index.html	https://www.sec.gov/Archives/edgar/data/1375365/0001375365-19-000039-index.html	https://www.sec.gov/Archives/edgar/data/1375365/0001628280-16-019274-index.html	https://www.sec.gov/Archives/edgar/data/1375365/0001628280-15-008827-index.html	https://www.sec.gov/Archives/edgar/data/1375365/0001628280-15-007025-index.html	https://www.sec.gov/Archives/edgar/data/1375365/0001445305-14-003958-index.html	https://www.sec.gov/Archives/edgar/data/1375365/0001445305-13-002262-index.html	https://www.sec.gov/Archives/edgar/data/1375365/0001445305-12-002860-index.html	https://www.sec.gov/Archives/edgar/data/1375365/0001193125-11-240215-index.html	https://www.sec.gov/Archives/edgar/data/1375365/0001193125-10-205667-index.html	https://www.sec.gov/Archives/edgar/data/1375365/0001193125-09-184698-index.html	https://www.sec.gov/Archives/edgar/data/1375365/0001193125-08-188476-index.html	https://www.sec.gov/Archives/edgar/data/1375365/0001193125-07-266502-index.html	https://www.sec.gov/Archives/edgar/data/1375365/0001193125-07-207110-index.html	https://www.sec.gov/Archives/edgar/data/1375365/0001193125-07-190775-index.html	
5	WORDS	2019	2019	2016	2015	2015	2014	2013	2012	2011	2010	2009	2008	2007	2007	2007	
772	company	798	639	726	297	373	363	346	349	316	360	346	457	77	267	422	
773	with	763	1147	1061	100	401	370	396	427	387	435	378	420	59	326	413	
774	that	717	1007	1009	107	355	343	344	375	369	405	408	397	79	262	308	
775	sales	531	362	314	49	310	312	310	313	317	329	307	303	12	296	296	
776	is	528	859	817	152	346	315	313	318	318	445	324	310	65	232	274	
777	are	521	584	487	97	316	294	312	325	318	354	329	334	67	243	271	
778	s	516	594	495	181	275	254	252	244	226	345	221	263	91	160	221	
779	by	515	1121	1157	99	302	291	307	327	315	488	350	416	93	259	279	
780	financial	477	589	350	184	260	267	263	264	266	298	336	284	59	210	229	
781	fiscal	477	405	298	34	252	271	268	269	265	283	228	190	40	152	152	
782	year	475	401	288	49	268	269	270	275	257	268	201	156	32	119	121	
783	stock	458	359	331	130	267	254	263	274	285	307	301	312	87	188	194	
784	an	455	626	464	72	224	213	209	240	232	253	255	256	36	190	217	
785	be	453	836	1066	57	241	231	226	251	243	366	275	301	22	207	252	
786	from	440	593	468	51	229	245	246	273	271	308	258	272	52	244	258	
787	million	424	395	227	0	236	288	278	302	278	268	249	209	1	180	180	
788	at	408	467	515	84	243	230	222	251	227	275	225	217	57	137	141	
789	not	408	602	587	64	192	190	196	208	207	291	247	238	37	177	202	
790	june	396	381	298	183	275	261	262	272	261	261	245	242	26	176	178	
791	other	385	844	912	52	176	175	179	188	172	228	175	197	43	120	145	
792	which	366	432	450	54	195	185	214	220	208	264	222	255	22	185	199	

Assessment of Financial Risk and Fraud Risk using Textual Analysis

- “Detect Fraud Before Catastrophe” by Lee, Churyk, and Clinton, Strategic Finance, March 2013, p. 33.
  - Proactive content analysis techniques can help management accountants prevent catastrophic financial fallout.
- “Using Nonfinancial Measures to Assess Fraud Risk” by Brazel, Jones, and Zimbelman, JAR 2009, p. 1135.
- SEC: Corporate Filers Beware: New “RoboCop” is On Patrol
  - Based on AQM and Text Analytics (not used yet, some companies are working on it)

Fraud Risk Assessment Model using Textual Analysis

Lee, Churyk and Clinton (*Strategic Finance* , 2013, p. 33)

Fraud detection model based on the textual, i.e., content, analysis of MD&A in 10-K:

$$\begin{aligned} \text{Fraud}_i = {} & 2.89757 - 0.83408\,(\text{Positive Emmotion}_i) \\ & - 0.48315\,(\text{Present Tense}_i) \\ & + .0001\,(\text{Total Words}_i) \\ & - 2.80753(\text{Colons}_i) \end{aligned}$$

"Conventional fraud detection measures using ratio analysis and other financial data were either unable to detect the fraud or unable to detect it soon enough to avoid catastrophic outcomes".

Text Mining: Fraud Risk Assessment Model using Nonfinancial Measures

Brazel, Jones, and Zimbelman (*JAR*, December 2009)

Del Global Technologies (1997, Fraud)

Income: Overstated	$3.7 million.
Revenue:	25% from PY. ↑
Employees:	6% (440 to 412) ↓
Distribution Dealers:	38% (400 to 250) ↓

Fischer Imaging Corp (1997, No Fraud): ↓

Revenue:	27% ↓
Employees:	20% ↓
Distribution Dealers:	7%

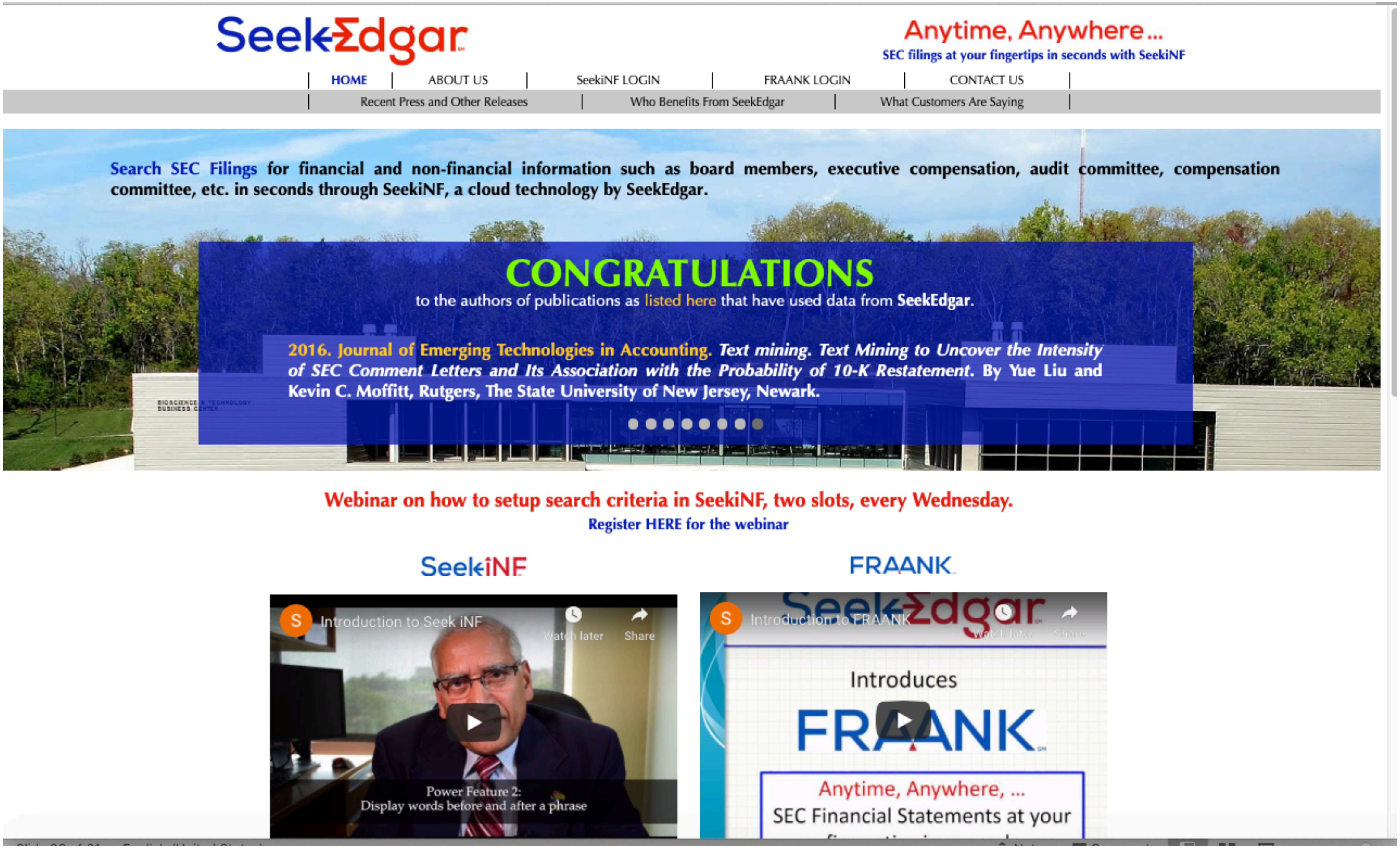
Liu and Moffitt

(*Journal of Emerging Technology in Accounting, 2016*)

- Textual analysis of SEC Comments Letters and developed a measure of intensity based on the modality of comment letters.
- Observed that the intensity of comment letters is positively associated with the probability of a restatement of the reviewed 10-K filings.
- Moreover, textual analysis and text mining techniques provide information about companies' performance that is not available otherwise.

Tone Analysis and Tone Dispersion

1. Loughran and Mcdonald. 2011. When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, Vol. 6, Issue 1, February: 35-65.
   - Develop an alternative negative word list, along with five other word lists, that better reflect tone in financial text. They link the word lists to:
     - 10-K filing returns, trading volume, return volatility, Fraud, material weakness, and unexpected earnings
2. Allee, K.D., and M. D. Deangelis. 2015. The Structure of Voluntary Disclosure Narratives: Evidence from Tone Dispersion. *Journal of Accounting Research,* Vol. 53 No. 2, p. 241. Tone dispersion is associated with
   - Analysts' and investors' responses to conference call narratives.
   - Reflects and affects the information that managers convey through their narratives.



SeekEdgar

Anytime, Anywhere…

SEC filings at your fingertips in seconds with SeekiNF

| HOME | ABOUT US | SeekiNF LOGIN | FRAANK LOGIN | CONTACT US |
| Recent Press and Other Releases | Who Benefits From SeekEdgar | What Customers Are Saying |

Search SEC Filings for financial and non-financial information such as board members, executive compensation, audit committee, compensation committee, etc. in seconds through SeekiNF, a cloud technology by SeekEdgar.

CONGRATULATIONS

to the authors of publications as listed here that have used data from SeekEdgar.

2016. Journal of Emerging Technologies in Accounting. *Text mining. Text Mining to Uncover the Intensity of SEC Comment Letters and Its Association with the Probability of 10-K Restatement.* By Yue Liu and Kevin C. Moffitt, Rutgers, The State University of New Jersey, Newark.

Webinar on how to setup search criteria in SeekiNF, two slots, every Wednesday.

Register HERE for the webinar

SeekiNF

FRAANK

SeekiNF℠

18 million Filings

and

33 million Documents

(1994-2020)

Daily Updated

2019-2020 Subscribers

1. Australian National University
2. Arizona State University, USA
3. Bentley University, USA
4. City University of Hong Kong
5. Fordham University, USA
6. Georgetown University, USA
7. Indian Inst. of Mgt. Ahamedabad
8. Macquarie University, Australia
9. Massey University, New Zealand
10. McMaster University, Toronto
11. Nanyang Technological University, Singapore
12. National Central University, Taiwan
13. National Taiwan University, Taiwan
14. New York University
15. Rutgers University-Newark, USA
16. University of Arkansas, USA

16. University of Bocconi, Italy
17. University of Chicago, USA
18. University of Illinois at Chicago, USA
19. University of Kansas, USA
20. University of Montreal, Canada
21. University of Nebraska – Lincoln, USA
22. University of New South Wales, Australia
23. University of Queensland, Australia
24. University of Southern California, USA
25. University of Sydney, Australia
26. University of Texas – San Antonio, USA
27. University of Waterloo, Canada
28. Washington University in St. Louis, USA
29. Xavier University, USA
30. Yale University, USA
31. BuzzFeed News, USA

SeekiNF℠

Questions?

Thanks!

rsrivastava@ku.edu