Zero-Revelation RegTech: Detecting Risk through Corporate Emails

Seoyoung Kim
Santa Clara University

@ R/Finance Conference
Chicago, IL
May 19, 2017

Joint work with:
Sanjiv Das (SCU) and Bhushan Kothari (Google Inc.)

Big Picture

- Financials are often delayed indicators of corporate quality
- Internal discussion (e.g., emails) may be used as an early warning system
- An automated platform that parses emails and produces summary statistics would be highly valuable, since…
  - It can analyze vast quantities of textual not amenable to human processing
  - It does not require revelation of individual email content explicitly to monitors/regulators

Our Purpose

- Our purpose is to explore the predictive power of information conveyed by employee emails

- Specifically, we are interested in:
  - The sentiment conveyed by email content
  - The information conveyed by structural characteristics, such as email volume or length
  - Other non-verbal indicators of potential trouble (e.g., shifting email network patterns)

Preview of Results

- We find that the net sentiment conveyed by Enron employee email content is a significant predictor of stock-return performance

- Interestingly, email length was a stronger predictor of subsequent price declines than the net sentiment conveyed by the message body itself.

- We also identify other potential indicators/predictors of escalating risk or malfeasance.

Data

The Enron Email Corpus

The Enron Email Corpus

- Initial Sample:
  - Approximately 500,000 emails

The Enron Email Corpus

- Initial Sample:
  - Approximately 500,000 emails
  - January 2000 through December 2001

The Enron Email Corpus

- Initial Sample:
  - Approximately 500,000 emails
  - January 2000 through December 2001
  - First made publicly available by the Federal Energy Regulatory Commission (FERC) during its investigation of Enron

The Enron Email Corpus

- Initial Sample:
  - Approximately 500,000 emails
  - January 2000 through December 2001
  - First made publicly available by the Federal Energy Regulatory Commission (FERC) during its investigation of Enron
  - Subsequently culled and distributed by the Carnegie Mellon CALO project

The Enron Email Corpus

- Initial Sample:
  - Approximately 500,000 emails
  - January 2000 through December 2001
  - First made publicly available by the Federal Energy Regulatory Commission (FERC) during its investigation of Enron
  - Subsequently culled and distributed by the Carnegie Mellon CALO project

- Caveats / Redactions
  - The Enron corpus has been scrubbed over time for legal reasons and to honor requests from affected employees.

The Enron Email Corpus

- Initial Sample:
  - Approximately 500,000 emails
  - January 2000 through December 2001
  - First made publicly available by the Federal Energy Regulatory Commission (FERC) during its investigation of Enron
  - Subsequently culled and distributed by the Carnegie Mellon CALO project

- Caveats / Redactions
  - The Enron corpus has been scrubbed over time for legal reasons and to honor requests from affected employees.
  - Ex(1): user “fastow-a” is notably missing

The Enron Email Corpus

- Initial Sample:
  - Approximately 500,000 emails
  - January 2000 through December 2001
  - First made publicly available by the Federal Energy Regulatory Commission (FERC) during its investigation of Enron
  - Subsequently culled and distributed by the Carnegie Mellon CALO project

- Caveats / Redactions
  - The Enron corpus has been scrubbed over time for legal reasons and to honor requests from affected employees.
  - Ex(1): user “fastow-a” is notably missing
  - Ex(2): Email chatter surrounding Mr. Skilling’s sudden resignation on 8/14/2001 has been expunged.

The Enron Email Corpus

- Initial Sample:
  - Approximately 500,000 emails
  - January 2000 through December 2001
  - First made publicly available by the Federal Energy Regulatory Commission (FERC) during its investigation of Enron
  - Subsequently culled and distributed by the Carnegie Mellon CALO project

- Caveats / Redactions
  - The Enron corpus has been scrubbed over time for legal reasons and to honor requests from affected employees.
  - Ex(1): user “fastow-a” is notably missing
  - Ex(2): Email chatter surrounding Mr. Skilling’s sudden resignation on 8/14/2001 has been expunged.
  - Overall, details regarding exclusion criteria have not been made public, and our analyses should be viewed as exploratory and prescriptive

Curing the Data

- We focus on “sent” emails (rather than all emails) in order to…
  - Analyze content specifically written by Enron employees
  - Avoid processing the same content more than once
  - i.e., if user “lay-k” sends an email to “skilling-j”

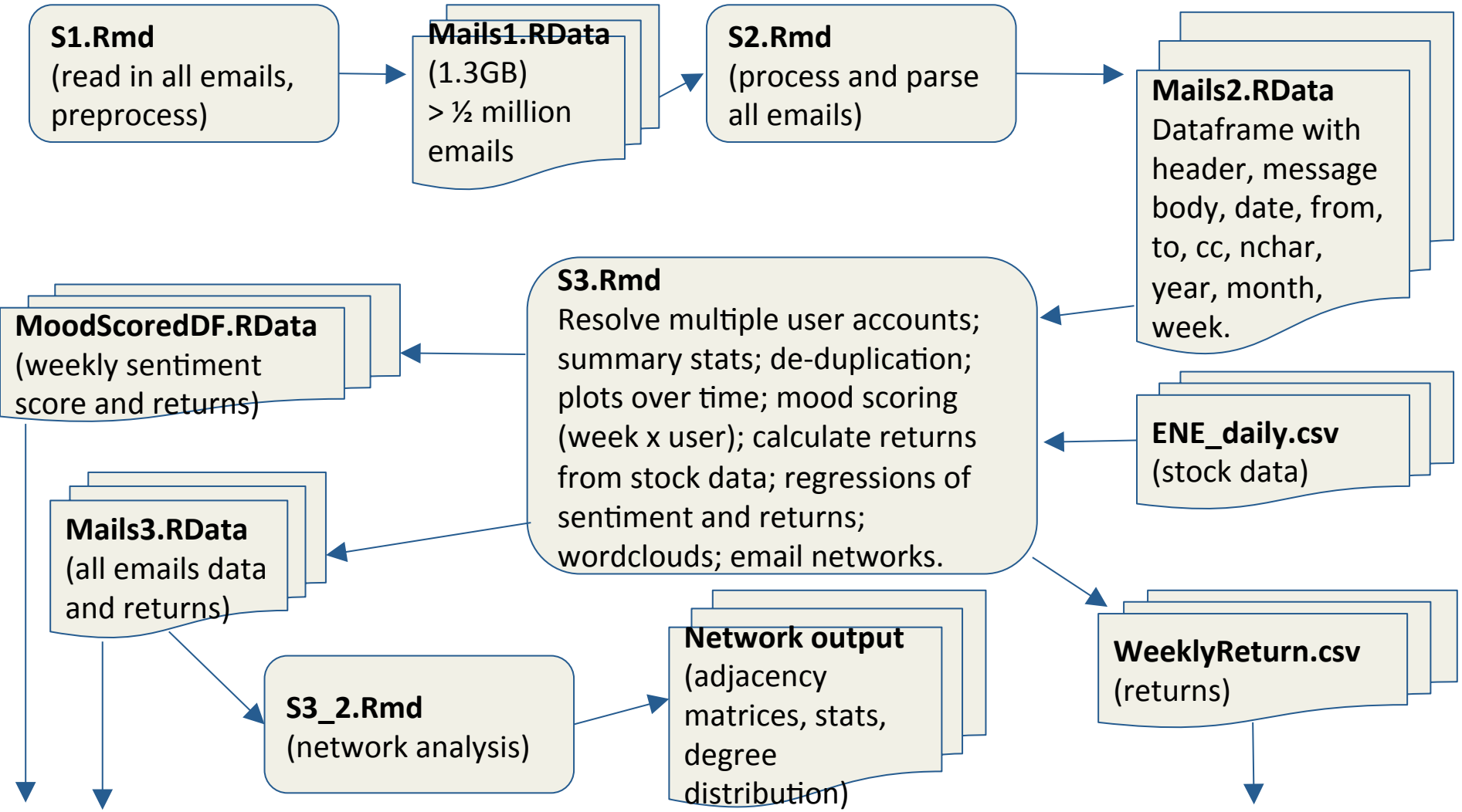
- Other filters applied to remove noisy (junk) mail:
  - Emails greater than 3,000 characters in length
  - Emails sent to more than 20 recipients

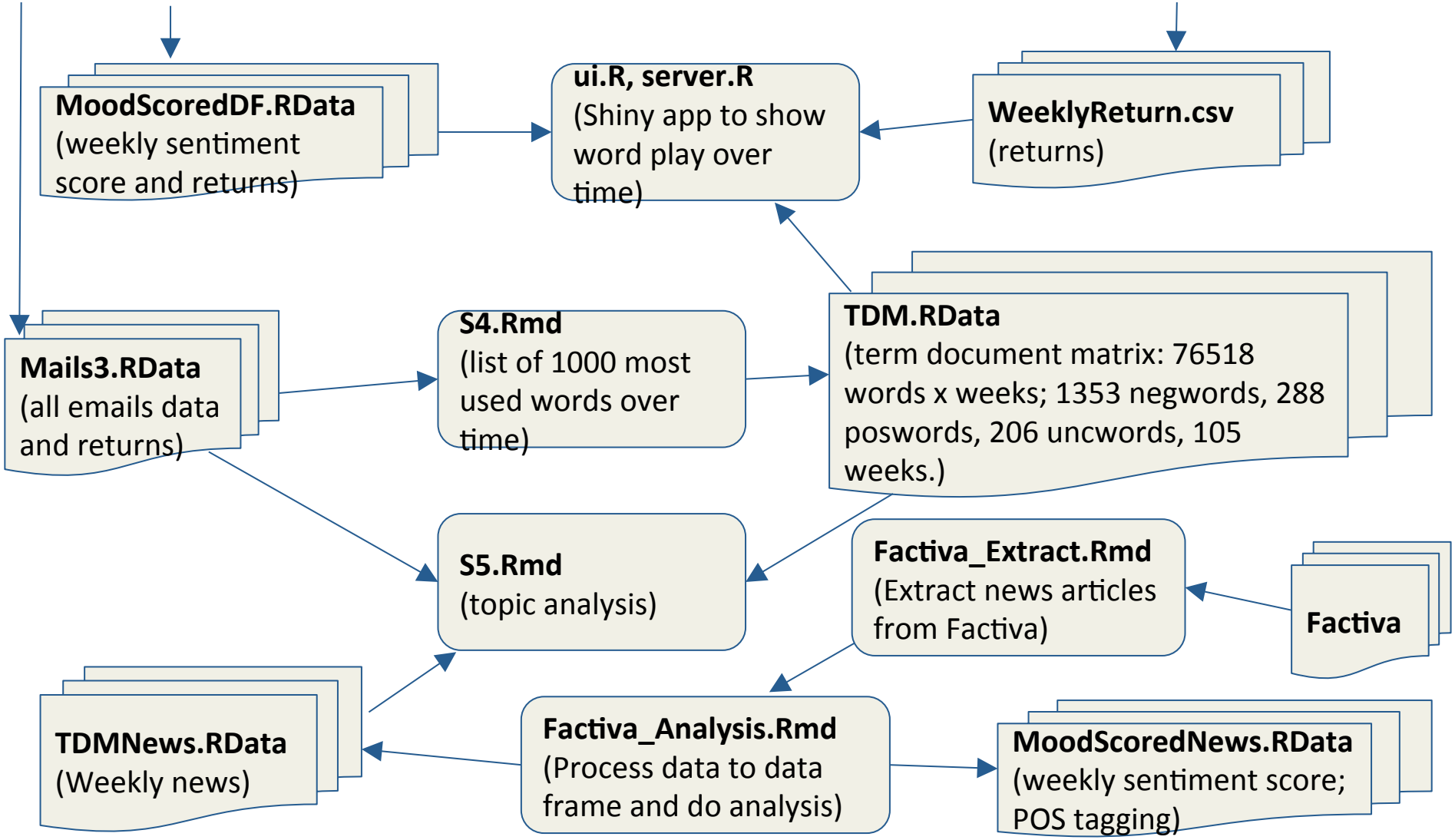
Our Final Sample

- Overall, we obtain…
  - The Enron email corpus from the Carnegie Mellon CS site
  - Stock price and stock return information from CRSP
  - News articles from Factiva PR Newswire
  - Sentiment word dictionaries from the Harvard Inquirer and the Loughran and McDonald sentiment word lists

- Final Sample:
  - 144 distinct employees
  - 113,266 sent emails
  - January 2000 through December 2001

Enron Code Pipeline


S1.Rmd
(read in all emails, preprocess)
Mails1.RData
(1.3GB)
> ½ million emails
S2.Rmd
(process and parse all emails)
Mails2.RData
Dataframe with header, message body, date, from, to, cc, nchar, year, month, week.
S3.Rmd
Resolve multiple user accounts; summary stats; de-duplication; plots over time; mood scoring (week x user); calculate returns from stock data; regressions of sentiment and returns; wordclouds; email networks.
MoodScoredDF.RData
(weekly sentiment score and returns)
ENE_daily.csv
(stock data)
Mails3.RData
(all emails data and returns)
S3_2.Rmd
(network analysis)
Network output
(adjacency matrices, stats, degree distribution)
WeeklyReturn.csv
(returns)


MoodScoredDF.RData
(weekly sentiment score and returns)
ui.R, server.R
(Shiny app to show word play over time)
WeeklyReturn.csv
(returns)
Mails3.RData
(all emails data and returns)
S4.Rmd
(list of 1000 most used words over time)
TDM.RData
(term document matrix: 76518 words x weeks; 1353 negwords, 288 poswords, 206 uncwords, 105 weeks.)
S5.Rmd
(topic analysis)
Factiva_Extract.Rmd
(Extract news articles from Factiva)
Factiva
TDMNews.RData
(Weekly news)
Factiva_Analysis.Rmd
(Process data to data frame and do analysis)
MoodScoredNews.RData
(weekly sentiment score; POS tagging)

Analyses

Table 1. Summary Statistics of Sent Mail

Panel A. Characteristics by Employee ($N = 144$)

Variable	Mean	Min	P25	Median	P75	Max
Emails per Person	787	2	105	349	891	8,793
Average "Connectedness"	1.62	1	1.21	1.44	1.76	4.47
Average Length per Person	279.92	19.15	160.45	227.90	338.07	944.23

Panel B. Email Characteristics ($N = 113,266$)

Variable	Mean	Min	P25	Median	P75	Max
Length of Email (# of characters)	362	0	46	163	466	2,998
Direct Recipients per Email ("to")	1.44	0	1	1	1	20
Indirect Recipients per Email ("cc")	0.					19
Total Recipients per Email	1.77	1	1	1	2	20

The average email is 362 characters in length, with a median of 163 characters…

Table 1. Summary Statistics of Sent Mail

Panel A. Characteristics by Employee ($N = 144$)						
Variable	Mean	Min	P25	Median	P75	Max
Emails per Person	787	2	105	349	891	8,793
Average "Connectedness"	1.62	1	1.21	1.44	1.76	4.47
Average Length per Person	279.92	19.15	160.45	227.90	338.07	944.23
Panel B. Email Characteristics ($N = 113,266$)						
Variable	Mean	Min	P25	Median	P75	Max
Length of Email (# of characters)	362	0	46	163	466	2,998
Direct Recipients per Email ("to")	1.44	0	1	1	1	20
Indirect Recipients per Email ("cc")	0.32	0	0	0	0	19
Total Recipients per Email	1.77	1	1	1	2	20

… with an average of 1.77 recipients per sent mail.

Table 1. Summary Statistics of Sent Mail

Panel A. Characteristics by Employee ($N = 144$)

Variable	Mean	Min	P25	Median	P75	Max
Emails per Person	787	2	105	349	891	8,793
Average "Connectedness"	1.62	1	1.21	1.44	1.76	4.47
Average Length per Person	279.92	19.15	160.45	227.90	338.07	944.23

Panel B. Email Characteristics ($N = 113,266$)

Variable	Mean	Min	P25	Median	P75	Max
Length of Email (# of characters)	362	0	46	163	466	2,998
Direct Recipients per Email ("to")	1.44	0	1	1	1	20
Indirect Recipients per Email ("cc")	0.32					
Total Recipients per Email	1.77	1	1	1	2	20

Many emails (close to 11%) are simply forwarded without added text.

Figure 1. Average Email Length over Time

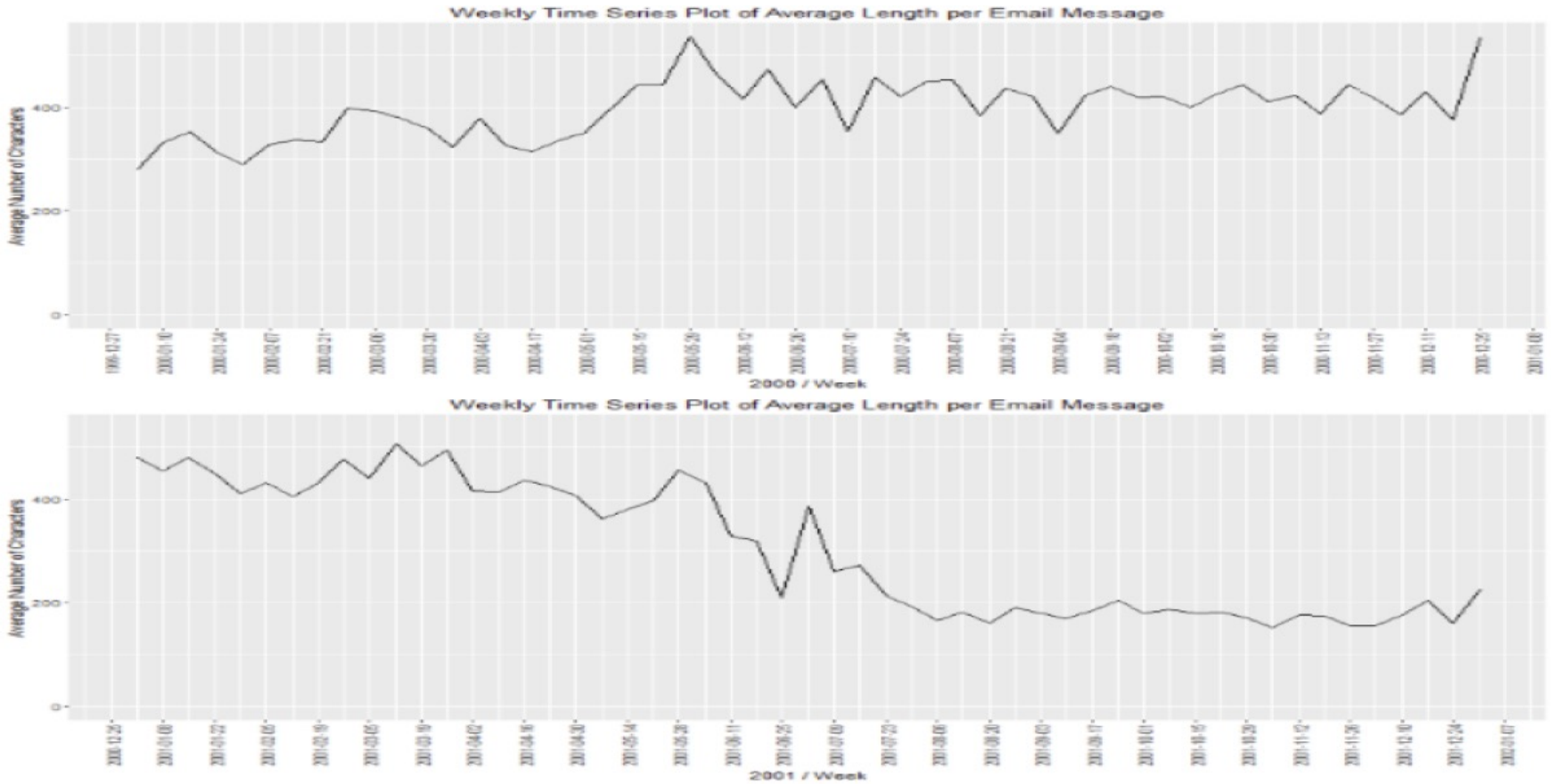
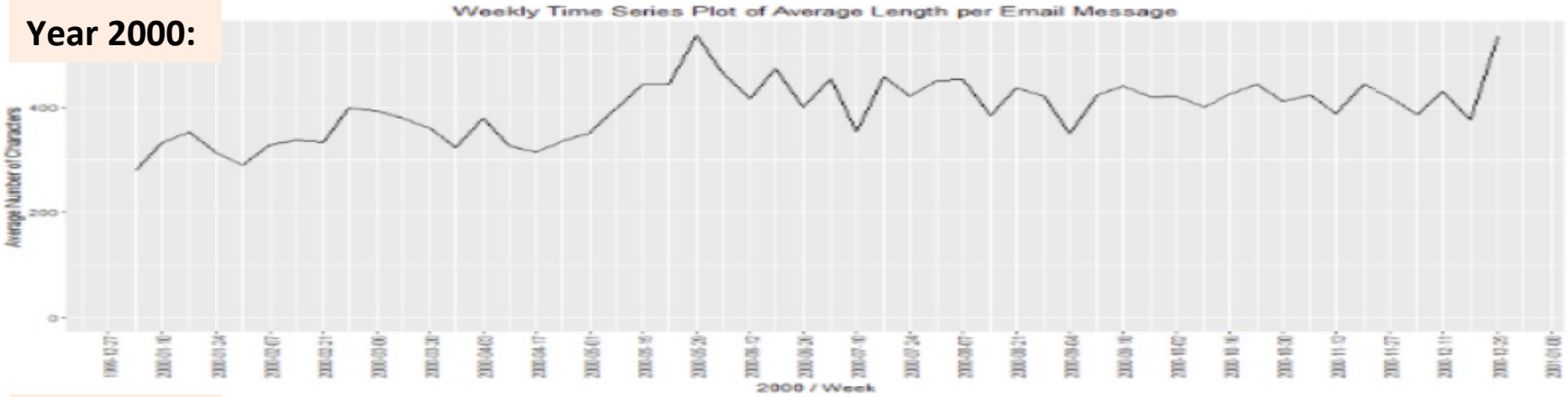


Figure 1. Average Email Length over Time

Year 2000:



Year 2001:

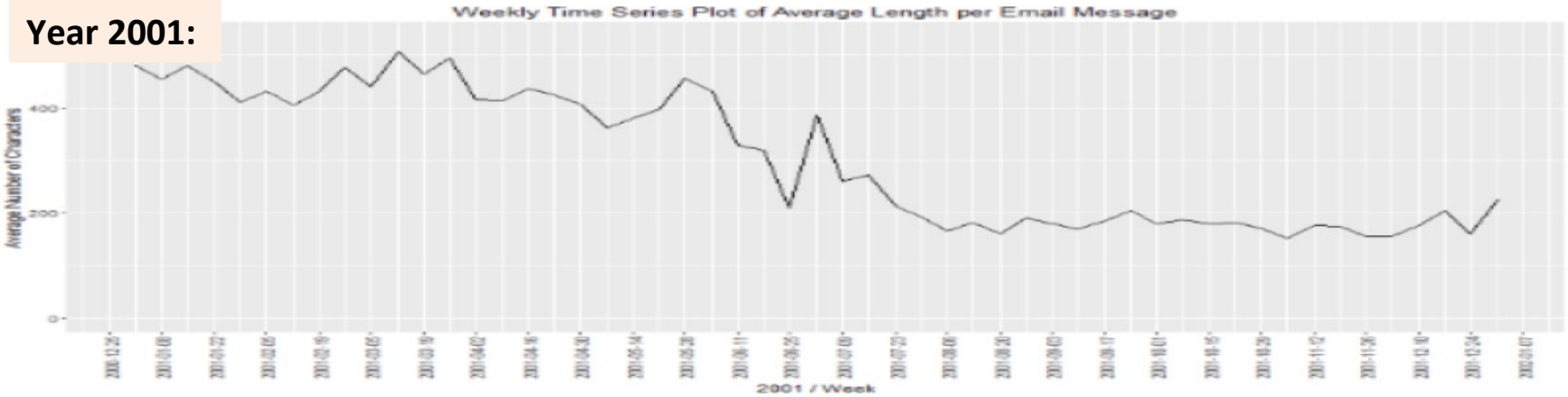
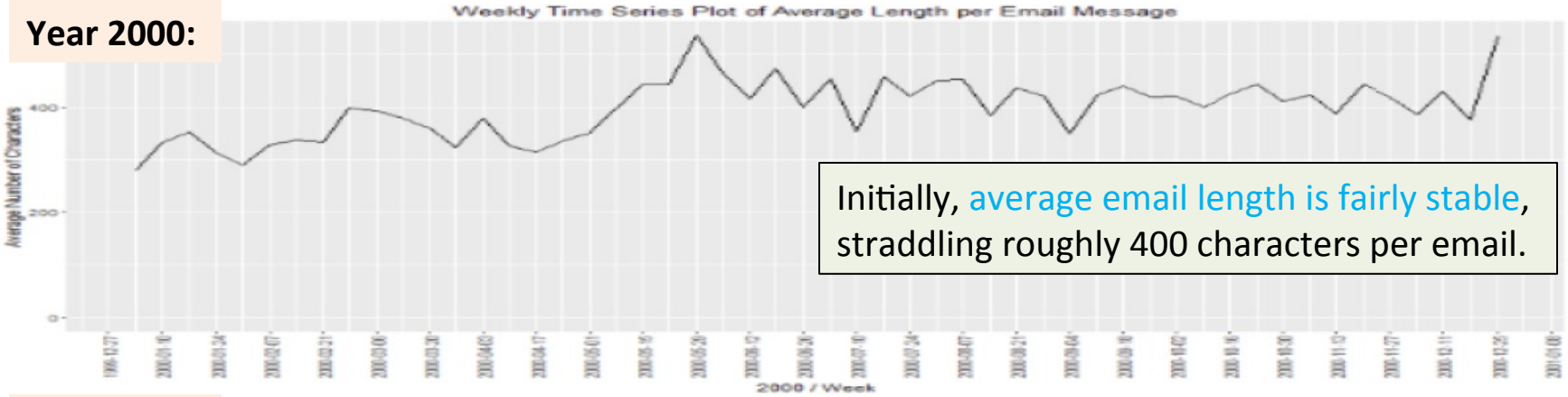


Figure 1. Average Email Length over Time

Year 2000:



Initially, average email length is fairly stable, straddling roughly 400 characters per email.

Year 2001:

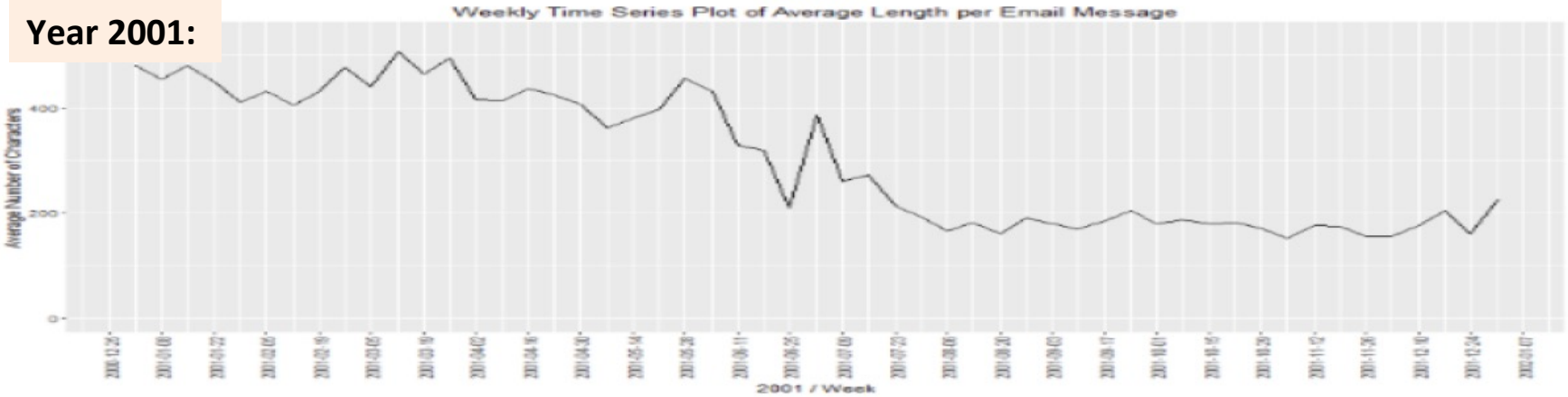
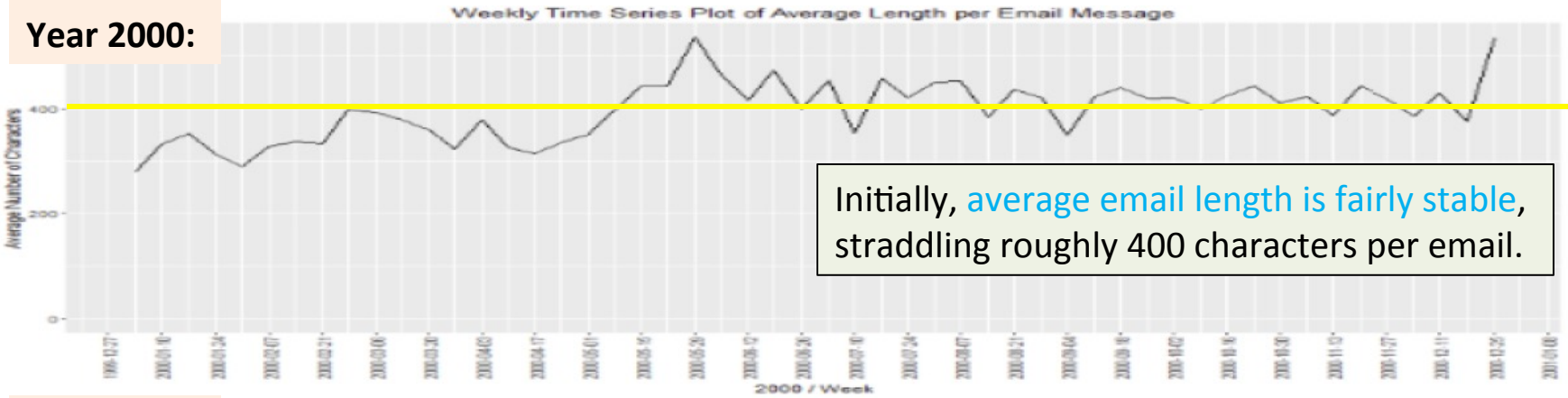


Figure 1. Average Email Length over Time

Year 2000:



Initially, average email length is fairly stable, straddling roughly 400 characters per email.

Year 2001:

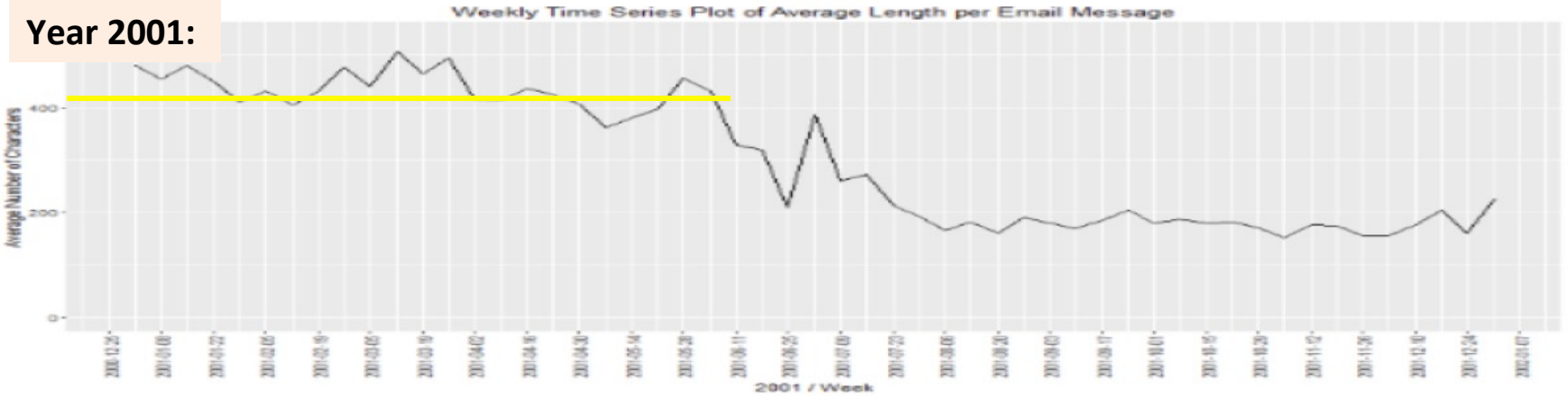
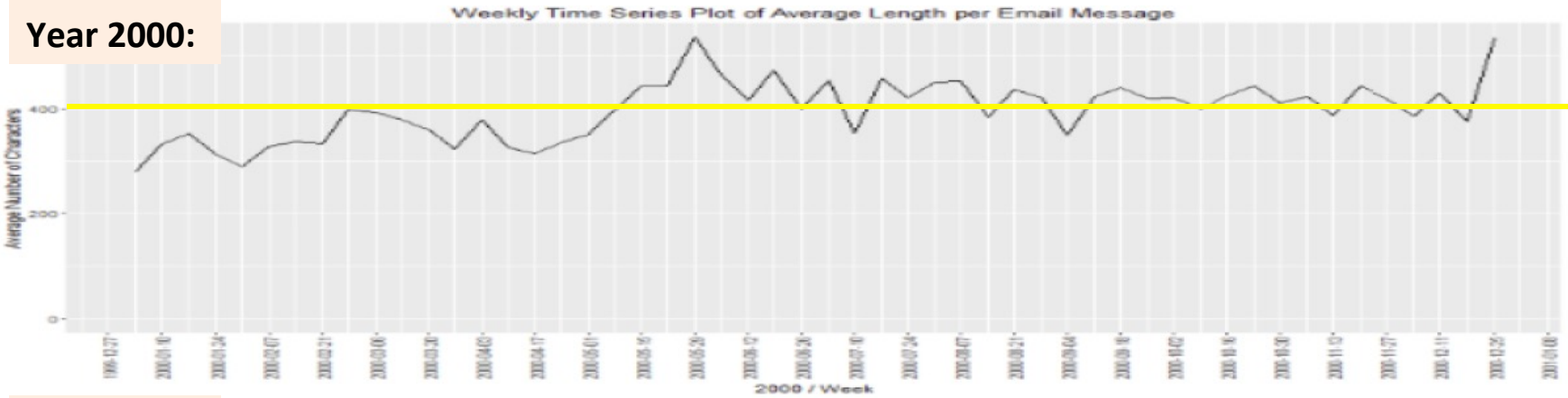
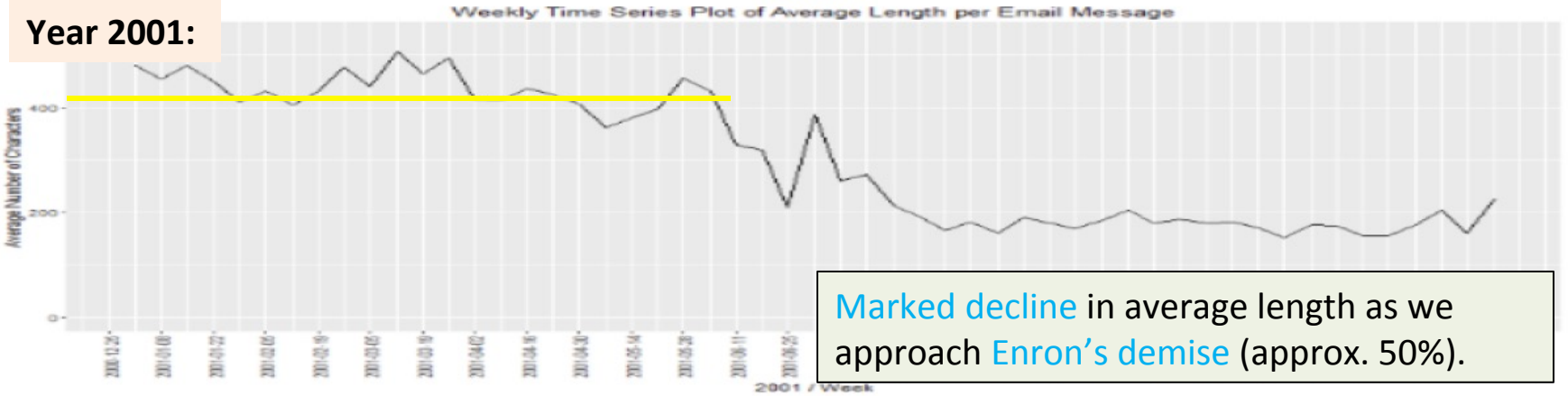


Figure 1. Average Email Length over Time

Year 2000:



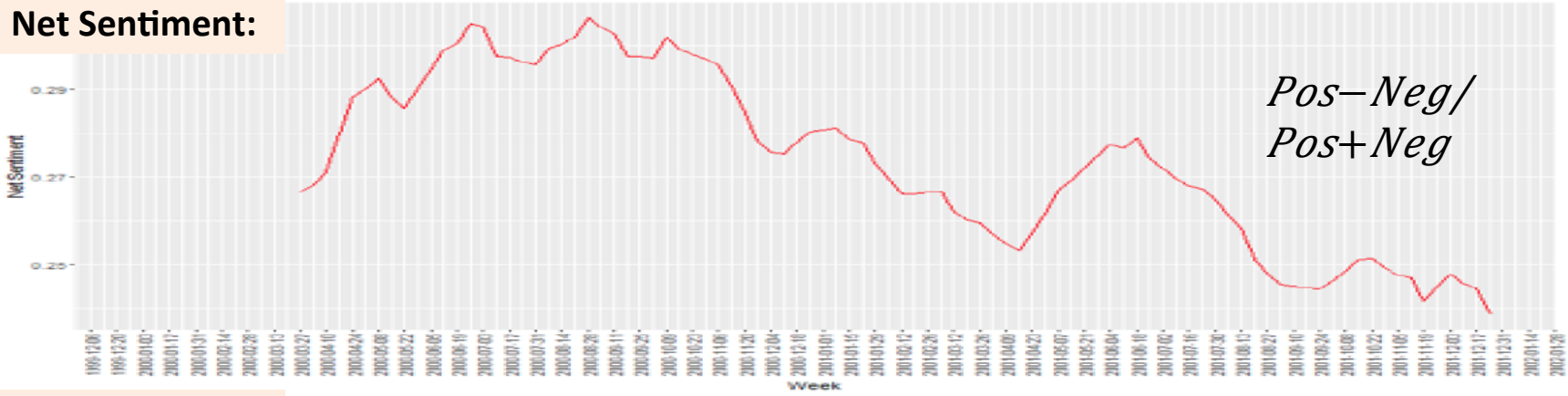
Year 2001:



Marked decline in average length as we approach Enron's demise (approx. 50%).

Figure 2. Email Sentiment and Disagreement over Time

Net Sentiment:



Disagreement:

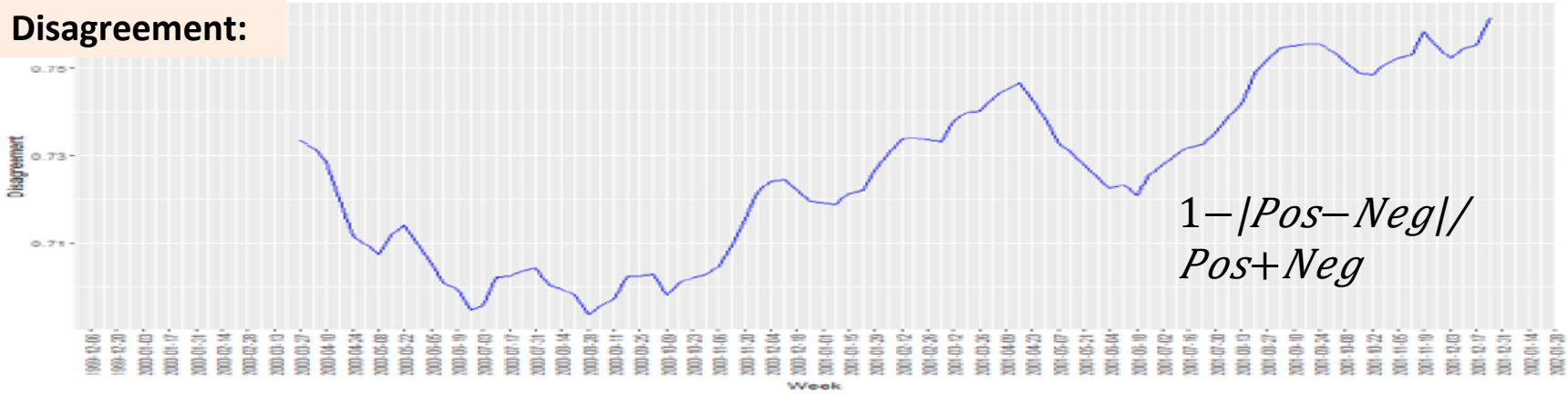
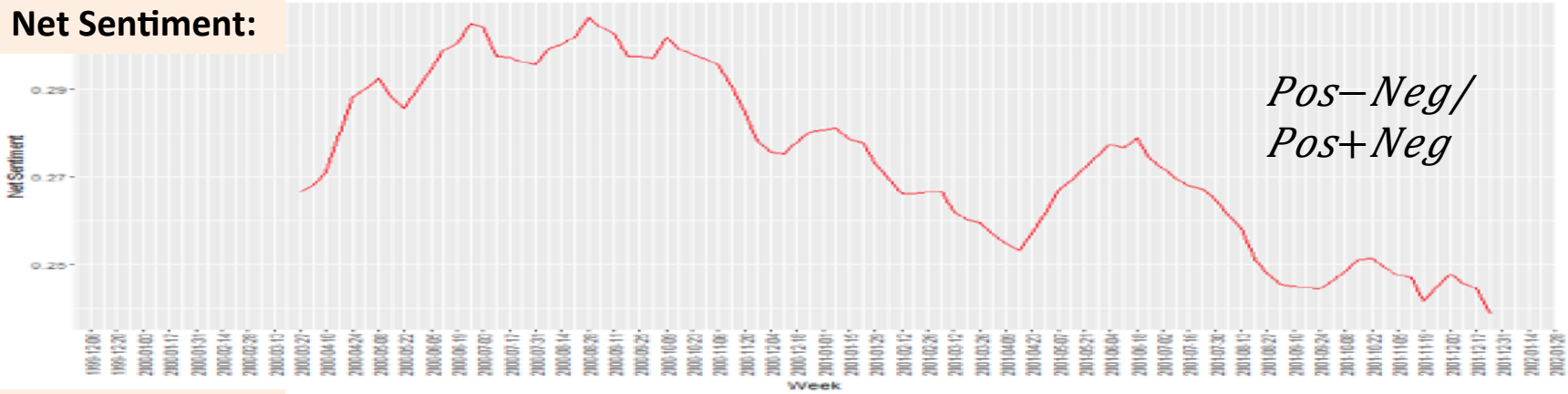
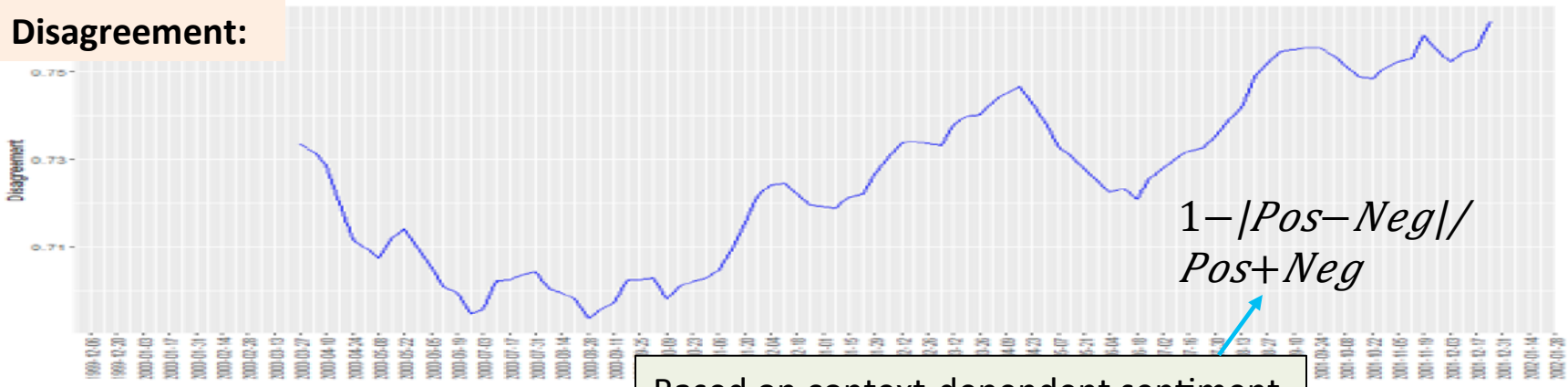


Figure 2. Email Sentiment and Disagreement over Time

Net Sentiment:



Disagreement:



Based on context-dependent sentiment dictionaries for word classification

Figure 3. Factiva News Coverage over Time

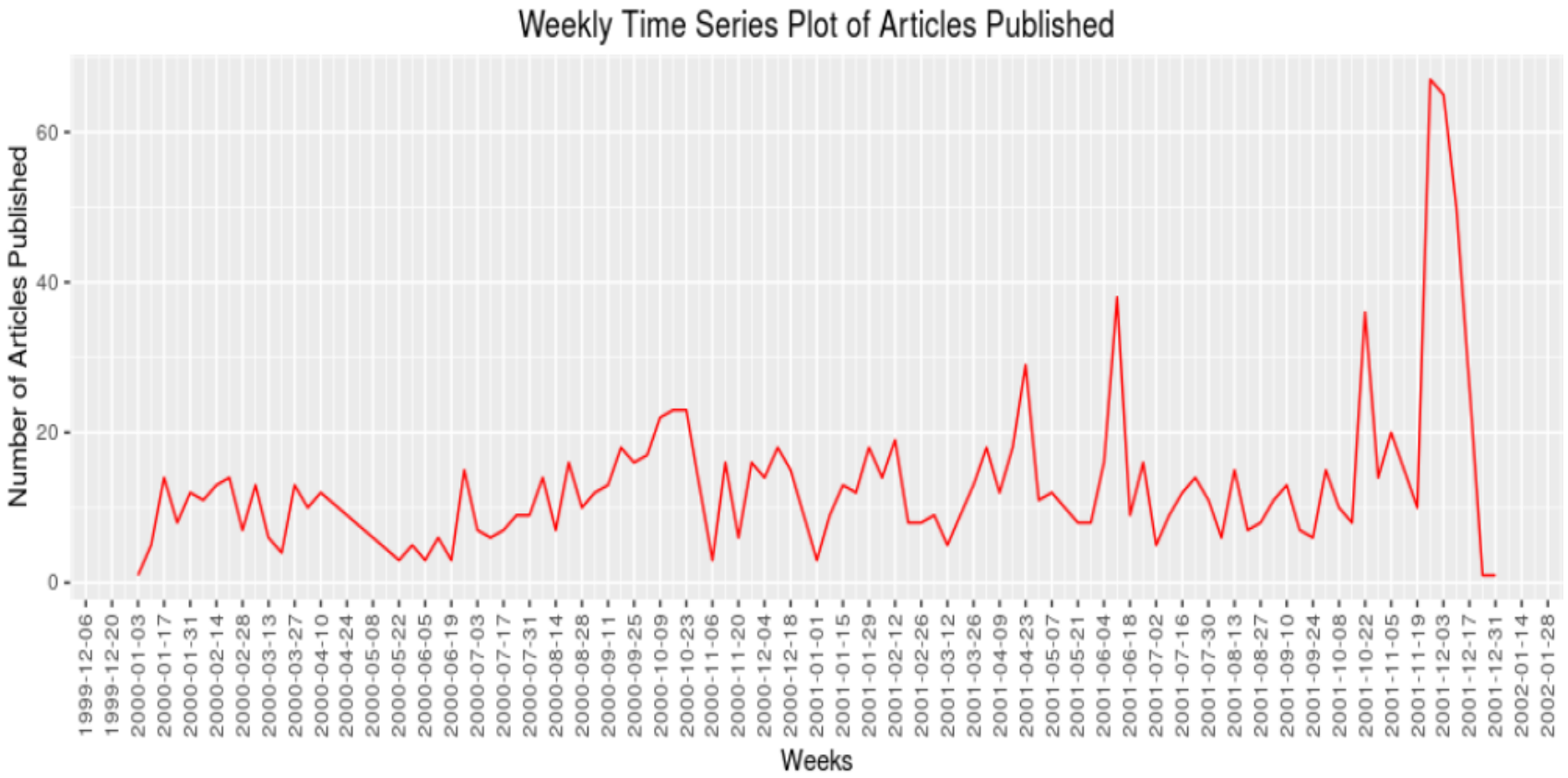


Figure 3. Factiva News Coverage over Time

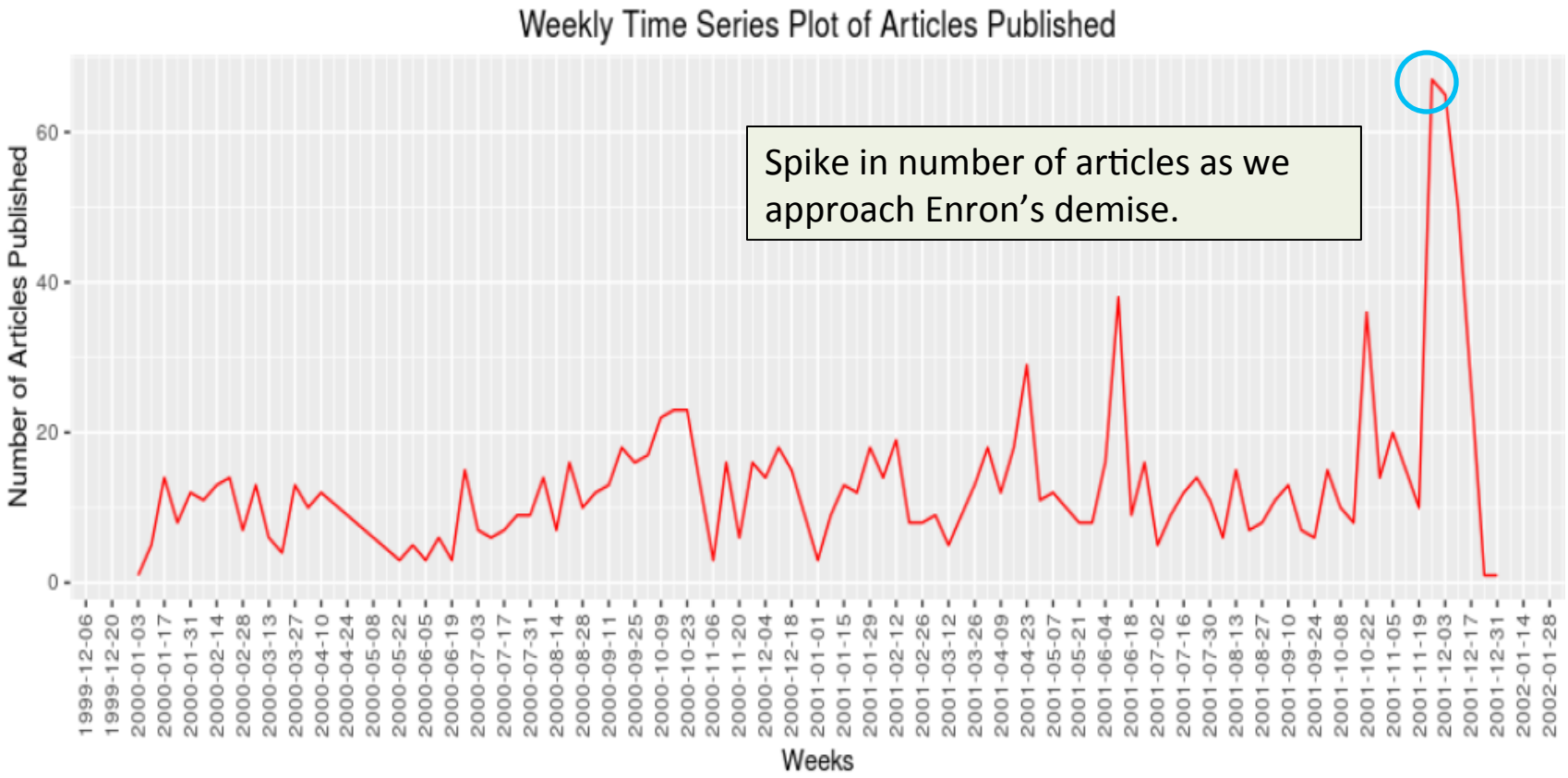


Figure 4. Factiva News Sentiment over Time

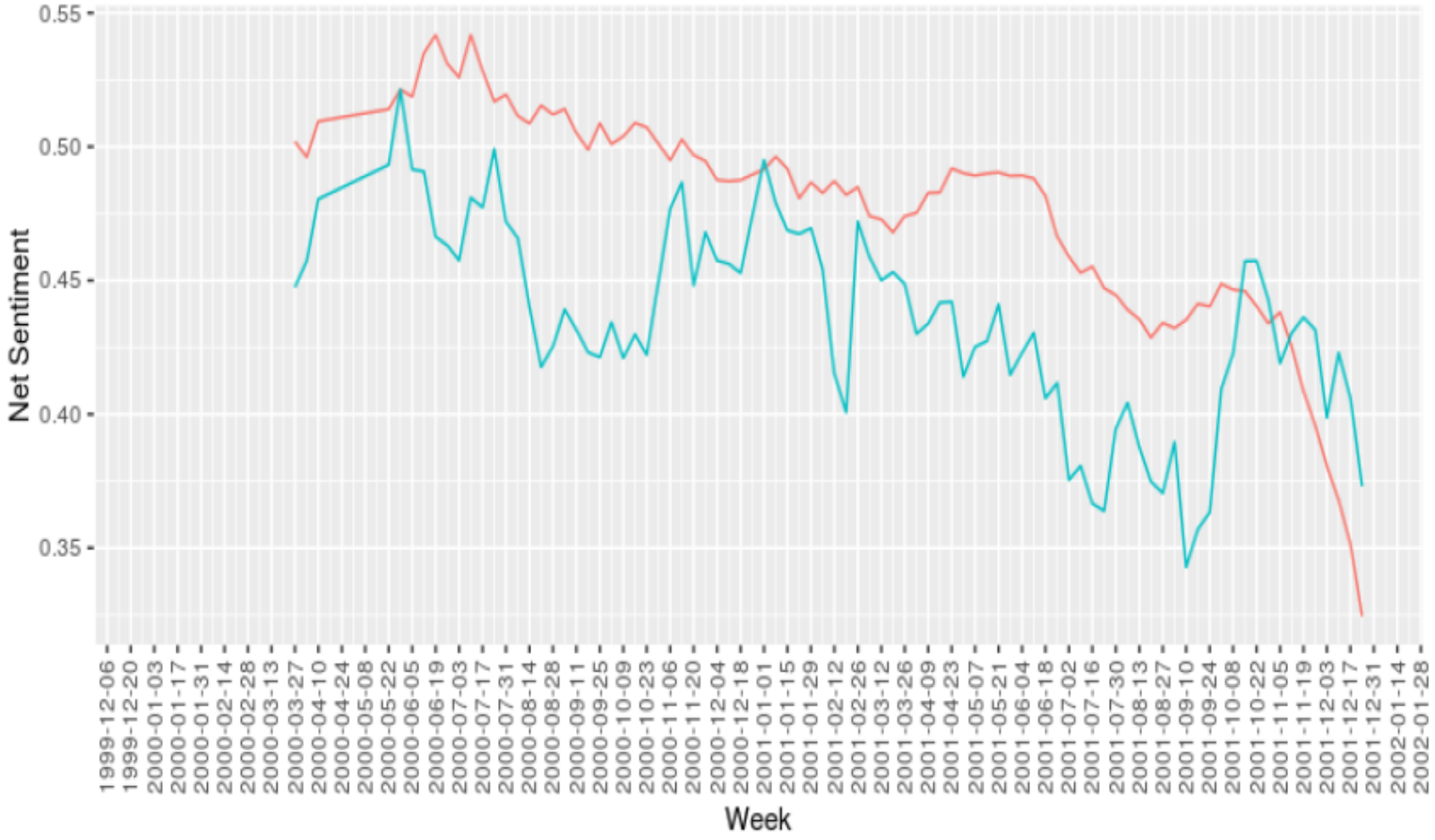


Figure 4. Factiva News Sentiment over Time

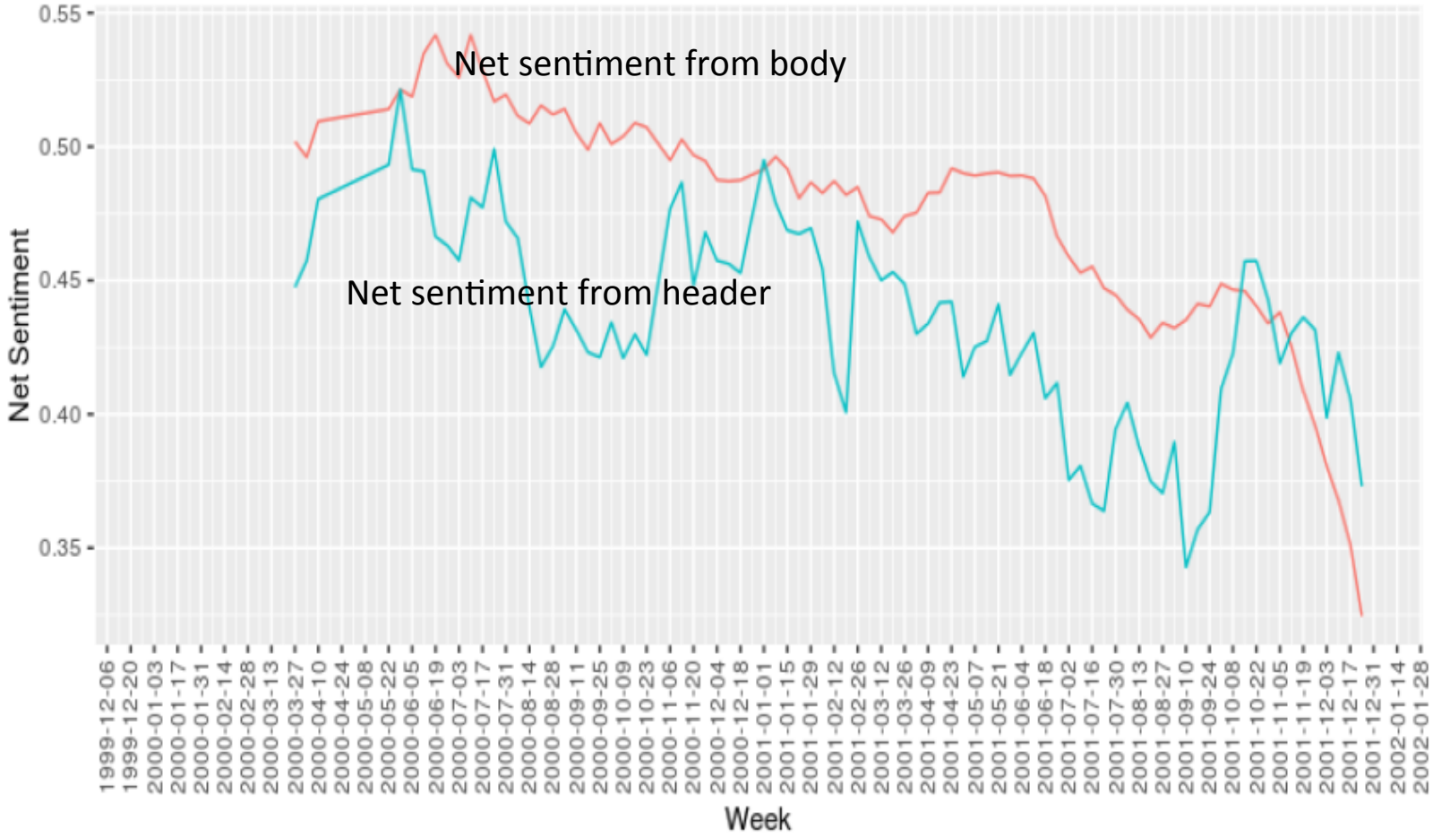
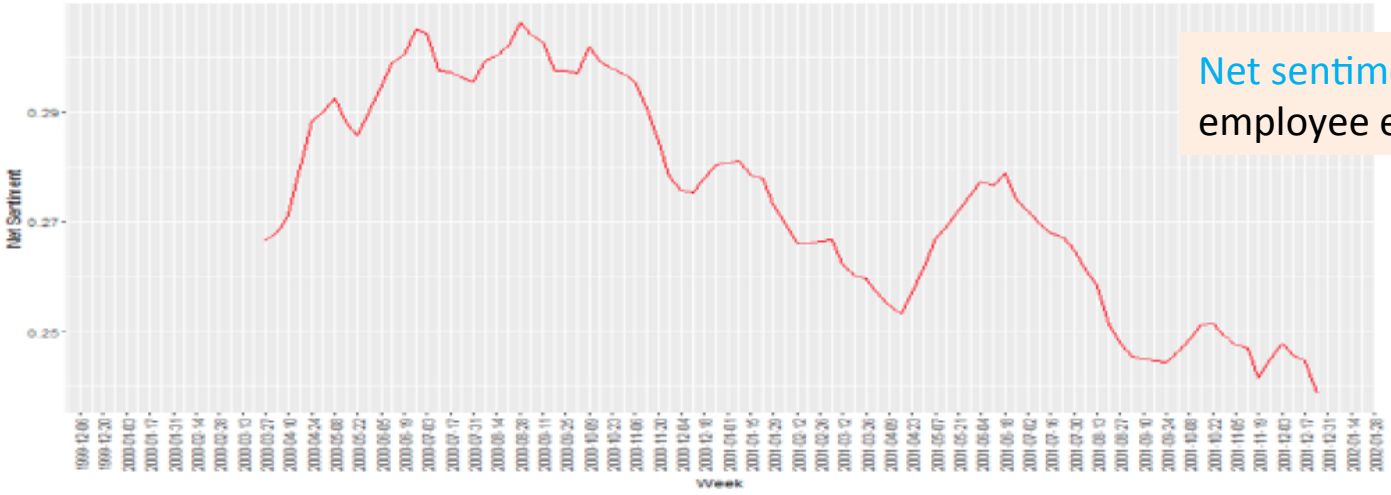
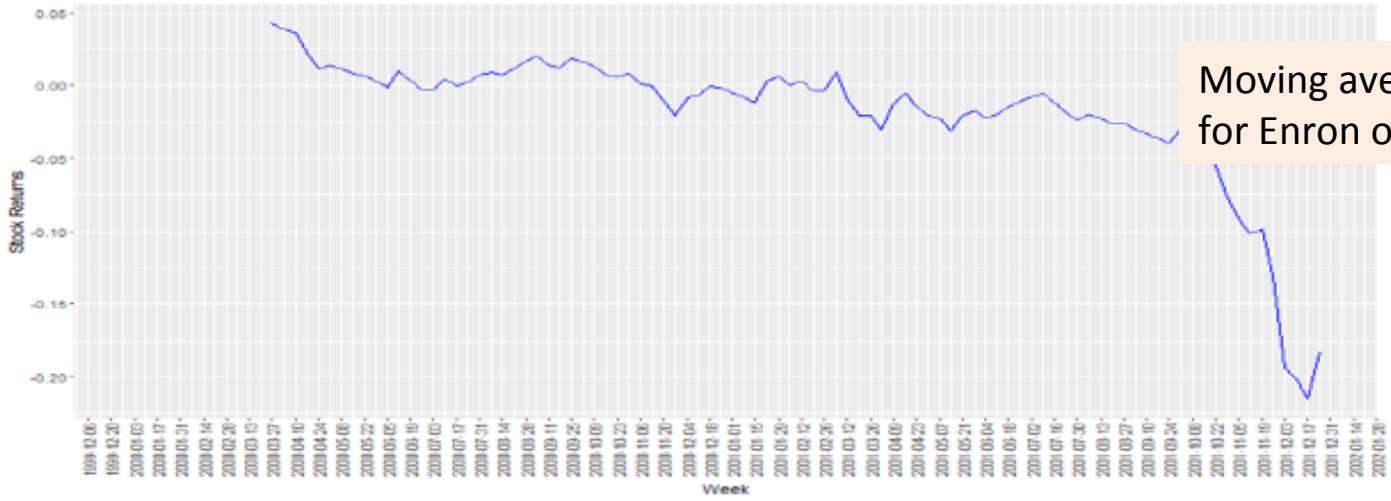


Figure 5. Stock Returns and Net Sentiment over Time

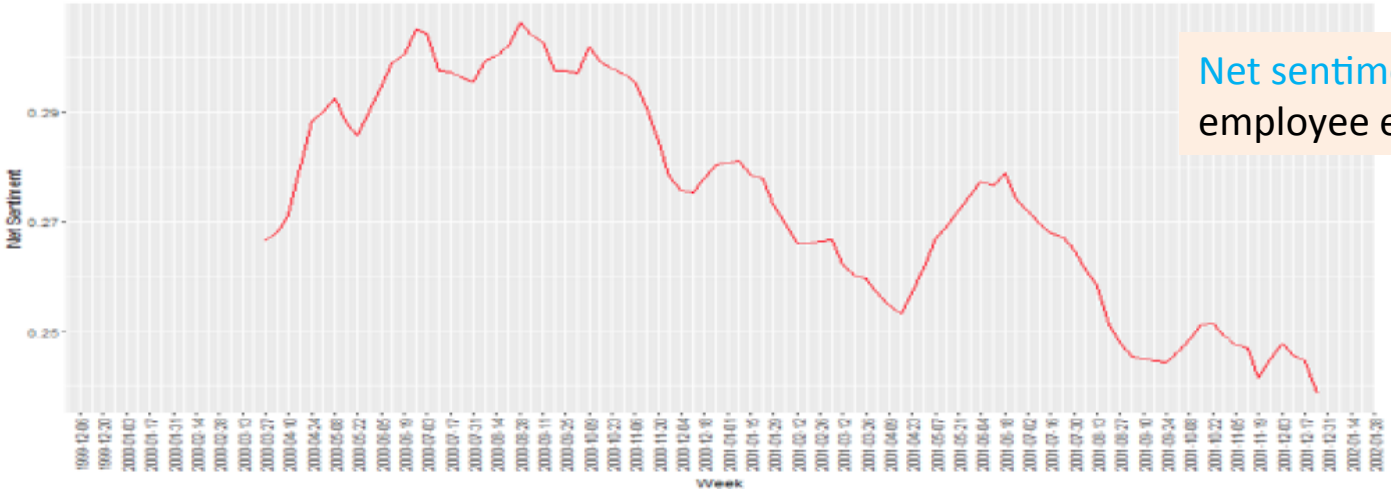


Net sentiment across Enron employee emails over time

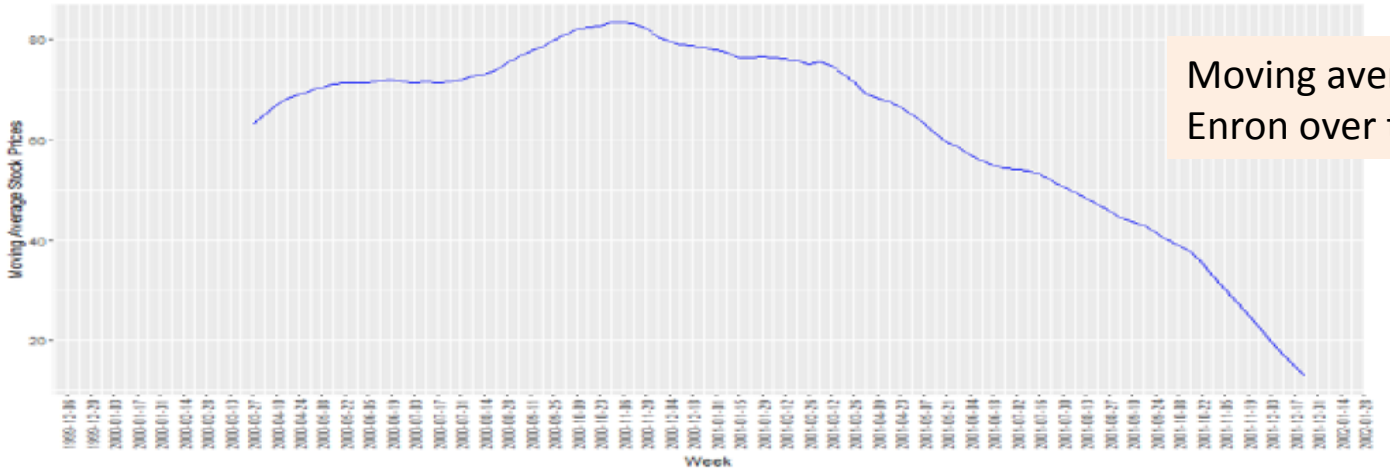


Moving average stock returns for Enron over time

Figure 6. Stock Prices and Net Sentiment over Time

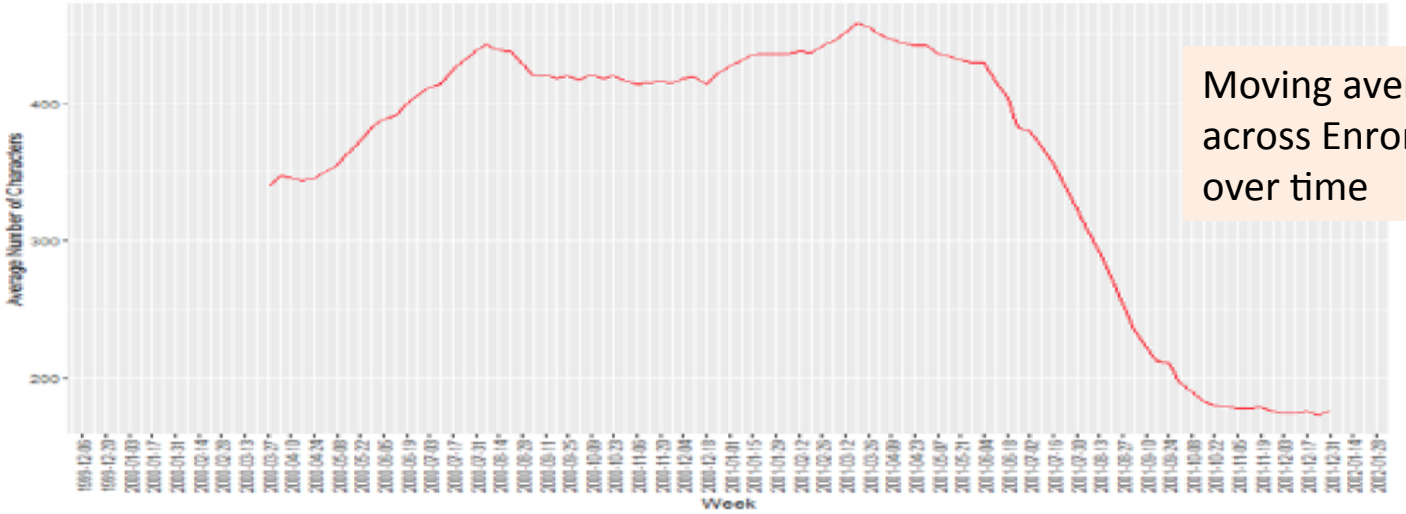


Net sentiment across Enron employee emails over time

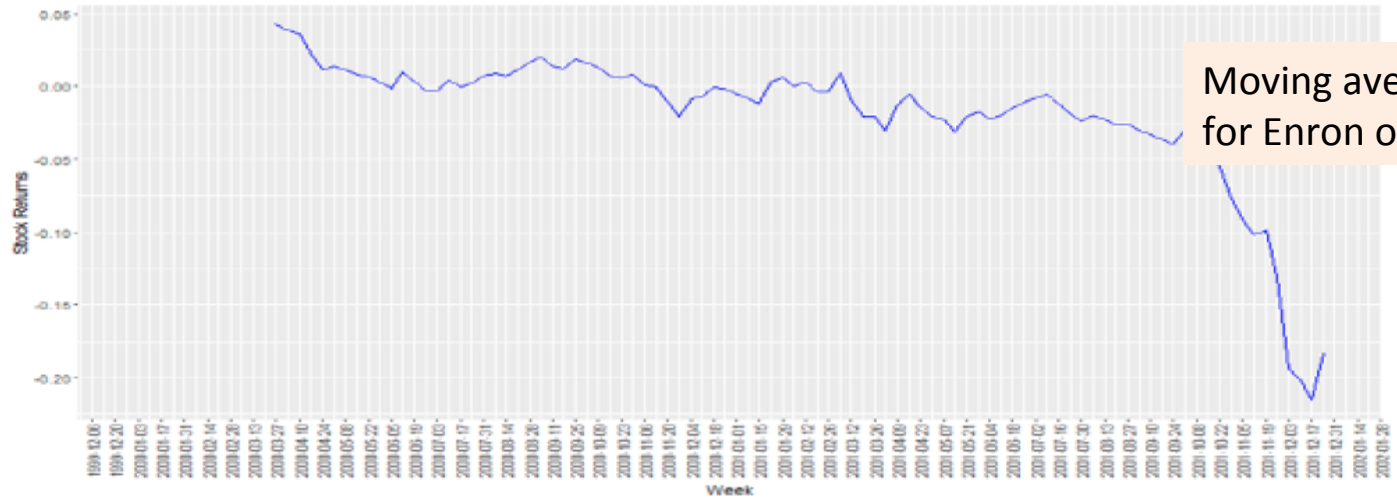


Moving average stock price for Enron over time

Figure 7. Stock Returns and Email Length over Time

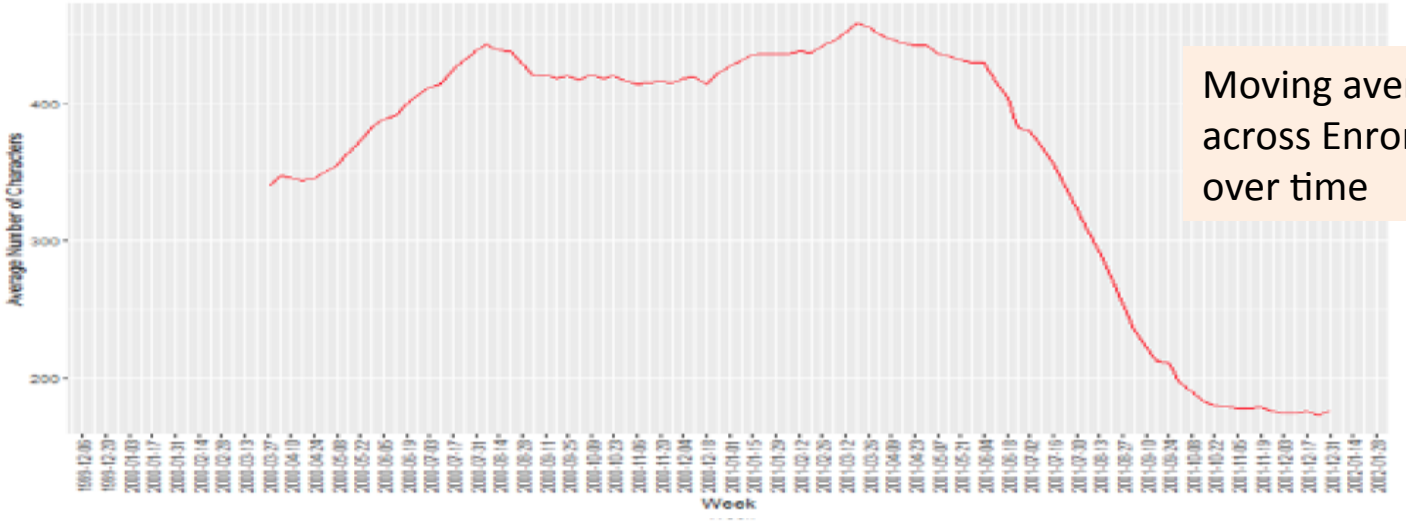


Moving average email length across Enron employee emails over time

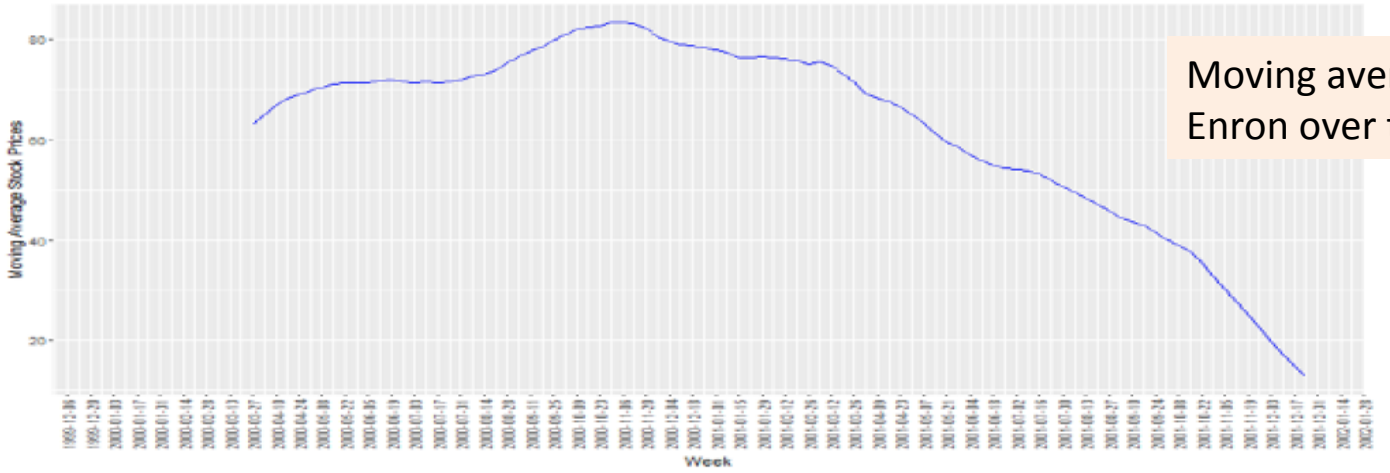


Moving average stock returns for Enron over time

Figure 8. Stock Prices and Email Length over Time



Moving average email length across Enron employee emails over time

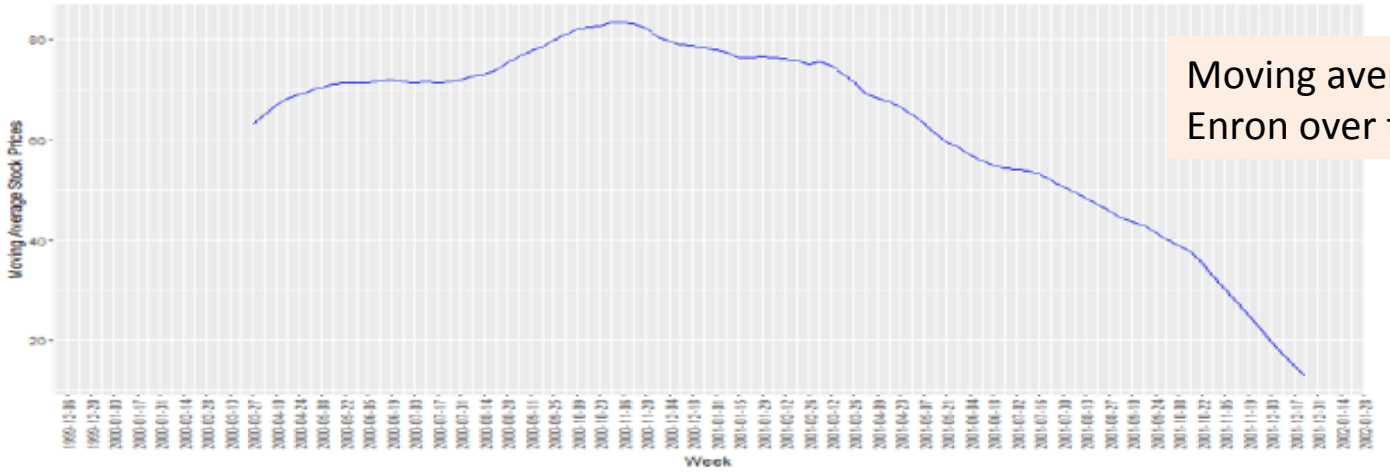


Moving average stock price for Enron over time

Figure 8. Stock Prices and Email Length over Time

Moving average email length across Enron employee emails over time

Perhaps, as risk/malfeasance escalates, emails become shorter, as employees are less likely to include details in message sent via the corporate server



Moving average stock price for Enron over time

Table 2. Email Content and Stock Returns

Dependent Variable = *Stock Returns*$_t$

Variable	Coefficient Estimate (t-statistic)			
	(1)	(2)	(3)	(4)
MA Email Sentiment$_t$	2.347***	0.575	2.330***	-1.397
	(3.27)	(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	-0.680***	-0.406*	-0.671***	0.117
	(-3.45)	(-1.93)	(-3.08)	(0.43)
Adjusted R^2	0.10	0.18	0.09	0.24
No. of observations	88	88	88	88

Table 2. Email Content and Stock Returns

Dependent Variable = *Stock Returns*$_t$

Variable	Coefficient Estimate (t-statistic)			
	(1)	(2)	(3)	(4)
MA Email Sentiment$_t$	2.347***	0.575	2.330***	-1.397
	(3.27)	(0.63)	(3.14)	(-1.25)
MA Email Length$_t$				
MA Total Emails$_t$			-0.004	-0.131***
			(-0.10)	(-2.83)
Intercept	-0.680***	-0.406*	-0.671***	0.117
	(-3.45)	(-1.93)	(-3.08)	(0.43)
Adjusted R^2	0.10	0.18	0.09	0.24
No. of observations	88	88	88	88

One stdev (i.e., 0.019) decrease in Net Sentiment is associated with a 4.5% decline in stock returns…

Table 2. Email Content and Stock Returns

Dependent Variable = *Stock Returns*$_t$

… but no longer significant when we control for email length.

Variable	Coefficient			
	(1)	(2)	(3)	(4)
MA Email Sentiment$_t$	2.347***	0.575	2.330***	-1.397
	(3.27)	(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	-0.680***	-0.406*	-0.671***	0.117
	(-3.45)	(-1.93)	(-3.08)	(0.43)
Adjusted R^2	0.10	0.18	0.09	0.24
No. of observations	88	88	88	88

Table 2. Email Content and Stock Returns over Time

Dependent Variable = *Stock Returns*$_t$

Variable	Coefficient Estimate (t-statistic)			
	(1)	(2)	(3)	(4)
MA Email Sentiment$_t$	2.347***	0.575	2.330***	-1.397
	(3.27)	(0.63)	(3.14)	(-1.25)
MA Email Length$_t$		0.584***		1.046***
		(2.97)		(4.19)
MA Total Emails$_t$				
Intercept	-0.680***	-0.406*	-0.671***	0.117
	(-3.45)	(-1.93)	(-3.08)	(0.43)
Adjusted R^2	0.10	0.18	0.09	0.24
No. of observations	88	88	88	88

Overall, 20-character decline in moving average email length is associated with a 1.17% decline in stock returns.

Table 3. Email Content versus Factiva News Content

Dependent Variable = *Stock Returns*$_t$

Panel B. News Header Sentiment and Returns					
MA Header Sentiment$_t$	-0.795	-1.136*	-0.772	-1.210**	-0.893
	(-1.31)	(-1.96)	(-1.34)	(-2.03)	(-1.61)
MA Email Sentiment$_t$		2.628***	0.705	2.566***	-1.254
		(3.30)	(0.66)	(3.18)	(-1.03)
MA Email Length$_t$			0.560**		1.026***
			(2.59)		(3.93)
MA Total Emails$_t$				-0.024	-0.138***
				(-0.59)	(-2.91)
Intercept	0.307	-0.256	-0.096	-0.178	0.485
	(1.15)	(-0.84)	(0.75)	(-0.54)	(1.39)
Adjusted R^2	0.01	0.12	0.18	0.11	0.25
No. of observations	81	81	81	81	81

Table 3. Email Content versus Factiva News Content

Dependent Variable = *Stock Returns*$_t$

Panel B. News Header Sentiment and Returns					
MA Header Sentiment$_t$	-0.795	-1.136*	-0.772	-1.210**	-0.893
	(-1.31)	(-1.96)	(-1.34)	(-2.03)	(-1.61)
MA Email Sentiment$_t$		2.628***	0.705	2.566***	-1.254
		(3.30)	(0.66)	(3.18)	(-1.03)
MA Email Length$_t$			0.560**		1.026***
MA Total Emails$_t$				-0.024	-0.138***
				(-0.59)	(-2.91)
Intercept	0.307	-0.256	-0.096	-0.178	0.485
	(1.15)	(-0.84)	(0.75)	(-0.54)	(1.39)
Adjusted R^2	0.01	0.12	0.18	0.11	0.25
No. of observations	81	81	81	81	81

Email content contains more information than news-header content….

Table 3. Email Content versus Factiva News Content

Dependent Variable = *Stock Returns*$_t$

Panel B. News Header Sentiment and Returns					
MA Header Sentiment$_t$	-0.795	-1.136*	-0.772	-1.210**	-0.893
	(-1.31)	(-1.96)	(-1.34)	(-2.03)	(-1.61)
MA Email Sentiment$_t$		2.628***	0.705	2.566***	-1.254
		(3.30)	(0.66)	(3.18)	(-1.03)
MA Email Length$_t$			0.560**		1.026***
			(2.59)		(3.93)
MA Total Emails$_t$				0.024	0.128***
Intercept	0.307	-0.256	-0.096	-0.178	0.485
	(1.15)	(-0.84)	(0.75)	(-0.54)	(1.39)
Adjusted R^2	0.01	0.12	0.18	0.11	0.25
No. of observations	81	81	81	81	81

…. But neither is significant when accounting for email length.

Table 3. Email Content versus Factiva News Content

Dependent Variable = *Stock Returns*$_t$

Panel A. News Body Sentiment and Returns					
MA Body Sentiment$_t$	1.410***	1.501**	0.657	1.505**	-0.827
	(3.95)	(2.49)	(0.87)	(2.48)	(-0.92)
MA Email Sentiment$_t$		-0.245	0.377	-0.284	-1.293
		(-0.19)	(-0.29)	(-0.22)	(-1.02)
MA Email Length$_t$			0.486*		1.380***
			(1.81)		(3.34)
MA Total Emails$_t$				-0.009	-0.164***
				(-0.24)	(-2.77)
Intercept	-0.711***	-0.688***	-0.426*	-0.668***	0.399
	(-4.18)	(-3.27)	(-1.69)	(-2.94)	(1.04)
Adjusted R^2	0.15	0.14	0.17	0.13	0.23
No. of observations	81	81	81	81	81

Table 3. Email Content versus Factiva News Content

Dependent Variable = *Stock Returns*$_t$

Panel A. News Body Sentiment and Returns					
MA Body Sentiment$_t$	1.410***	1.501**	0.657	1.505**	-0.827
	(3.95)	(2.49)	(0.87)	(2.48)	(-0.92)
MA Email Sentiment$_t$		-0.245	0.377	-0.284	-1.293
		(-0.19)	(-0.29)	(-0.22)	(-1.02)
MA Email Length$_t$					
MA Total Emails$_t$					
Intercept	-0.711***	-0.688***	-0.426*	-0.668***	0.399
	(-4.18)	(-3.27)	(-1.69)	(-2.94)	(1.04)
Adjusted R^2	0.15	0.14	0.17	0.13	0.23
No. of observations	81	81	81	81	81

On the other hand, email content contains less information than content from the news body…

(could this be due to redactions on the Enron email corpus?)

Table 3. Email Content versus Factiva News Content

Dependent Variable = *Stock Returns*$_t$

Panel A. News Body Sentiment and Returns					
MA Body Sentiment$_t$	1.410***	1.501**	0.657	1.505**	-0.827
	(3.95)	(2.49)	(0.87)	(2.48)	(-0.92)
MA Email Sentiment$_t$		-0.245	0.377	-0.284	-1.293
		(-0.19)	(-0.29)	(-0.22)	(-1.02)
MA Email Length$_t$			0.486*		1.380***
			(1.81)		(3.34)
MA Total Emails$_t$					
Intercept	-0.711***	-0.688***	-0.426*	-0.668***	0.399
	(-4.18)	(-3.27)	(-1.69)	(-2.94)	(1.04)
Adjusted R^2	0.15	0.14	0.17	0.13	0.23
No. of observations	81	81	81	81	81

…. But, again, neither is significant when accounting for email length.

Summary and Implications

- Thus far, we have shown that the net sentiment conveyed by employee sent mails is a significant predictor of stock-return performance
- Interestingly, email length was a stronger predictor of subsequent price declines than the net sentiment conveyed by the message body itself.
- Overall, email content may be controlled or manipulated
  - Thus, we are also (and perhaps even more!) interested in the non-verbal, interaction- or network-based indicators of potential trouble.

Additional Explorations

Other dimensions ripe for investigation….

Figure 11. Email Networks

Year 2000, Q4:

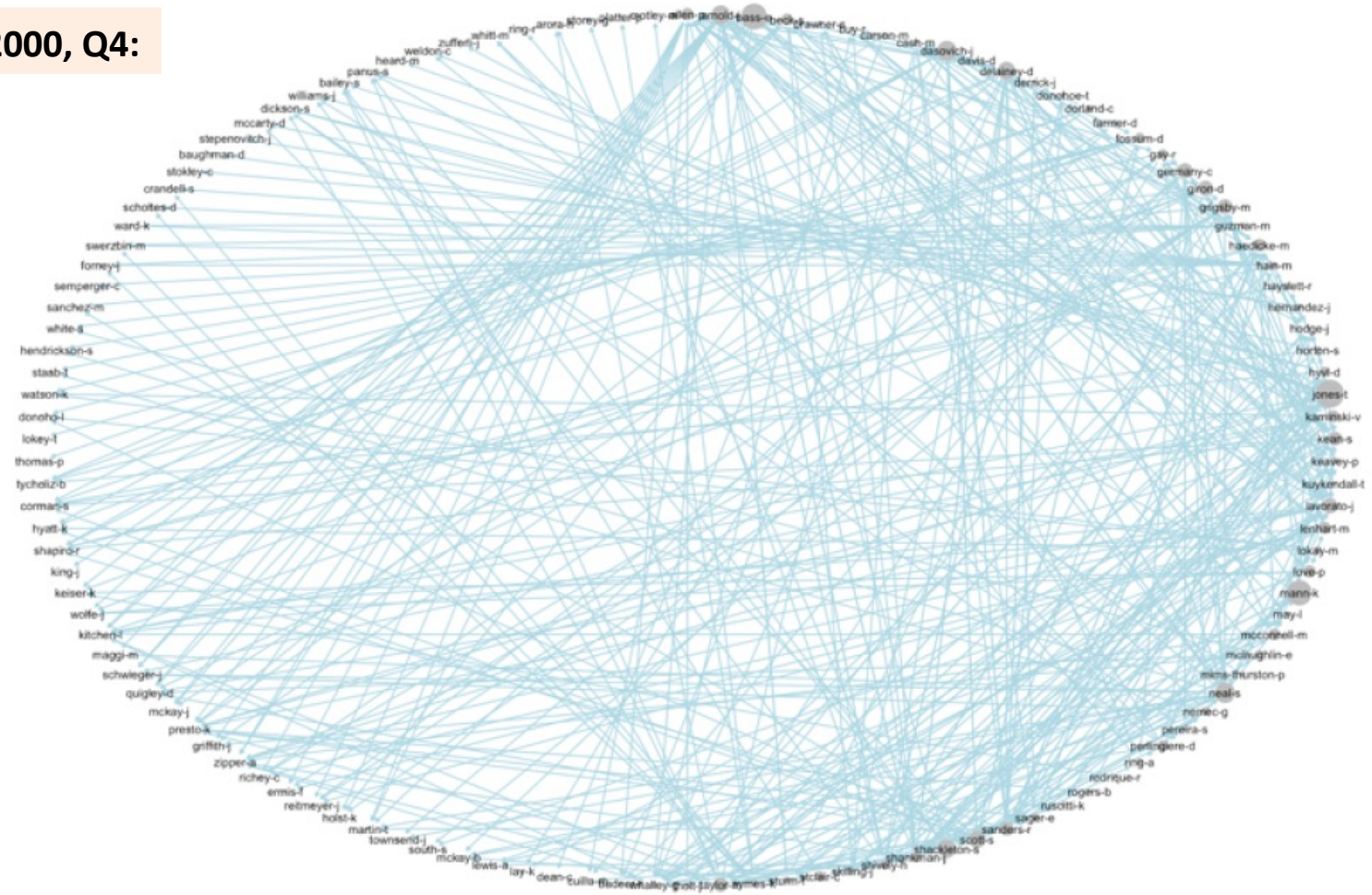


Figure 11. Email Networks

Year 2001, Q4:

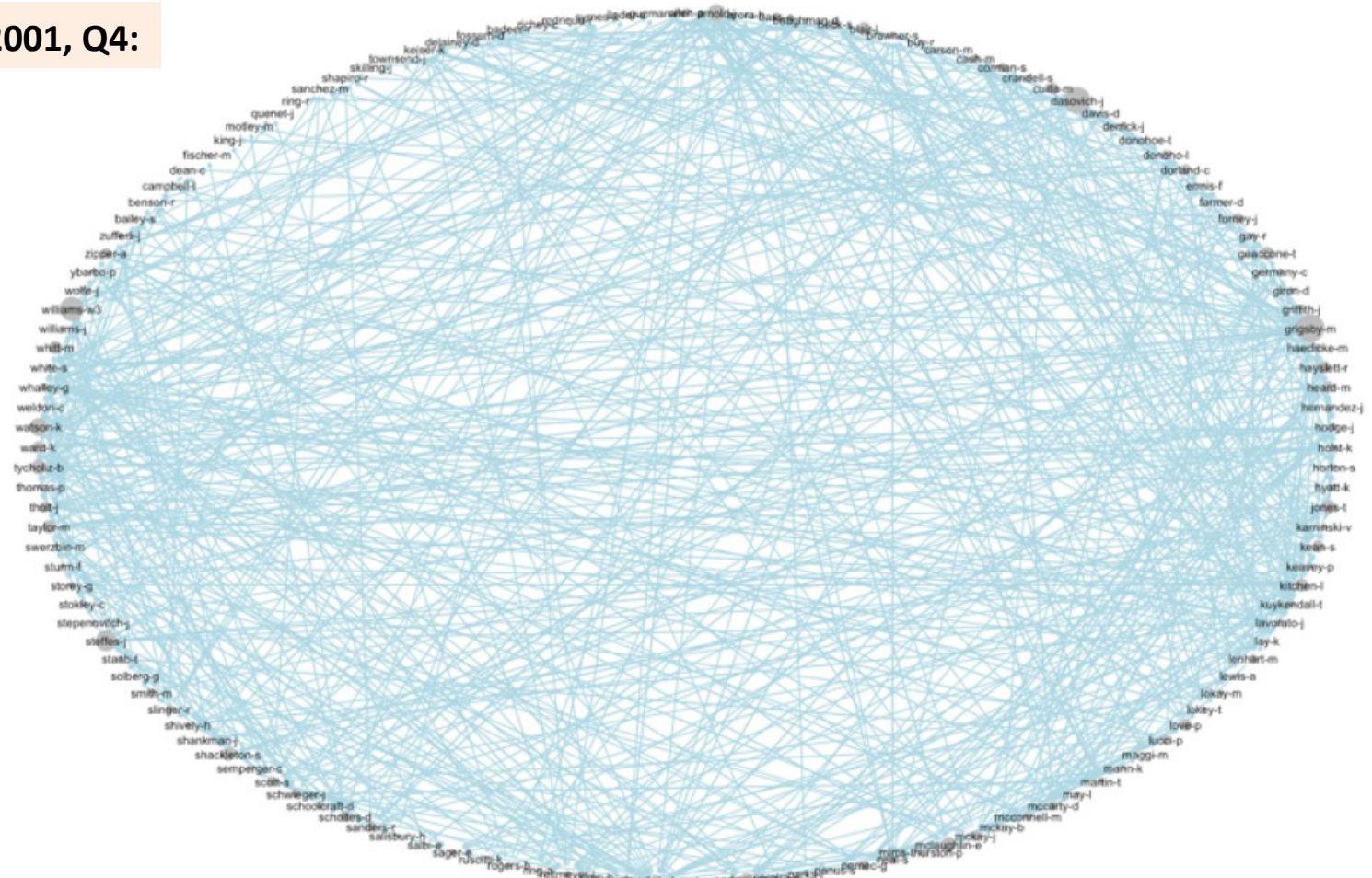


Figure 13. Vocabulary Trends

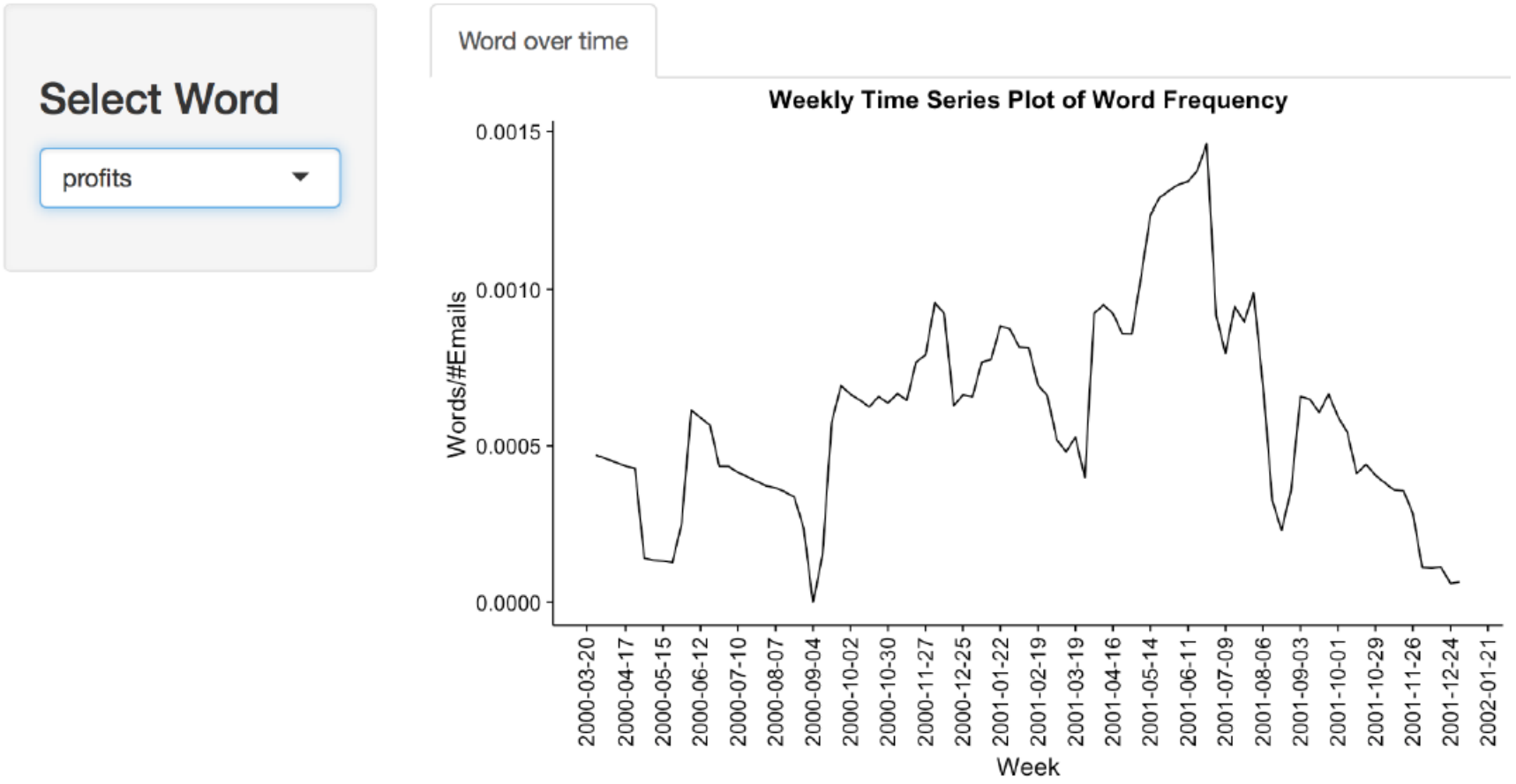


Figure 13. Vocabulary Trends

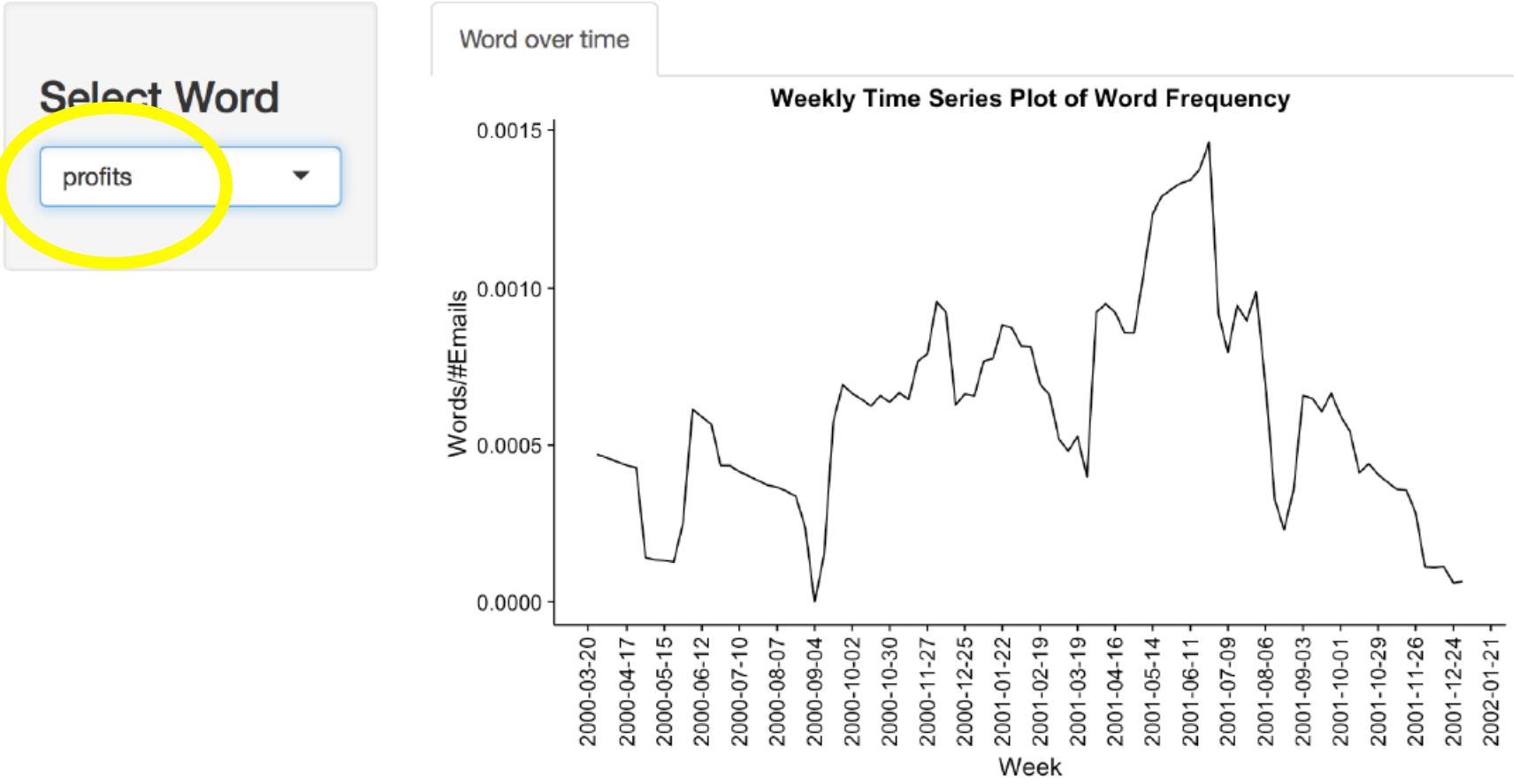


Figure 13. Vocabulary Trends

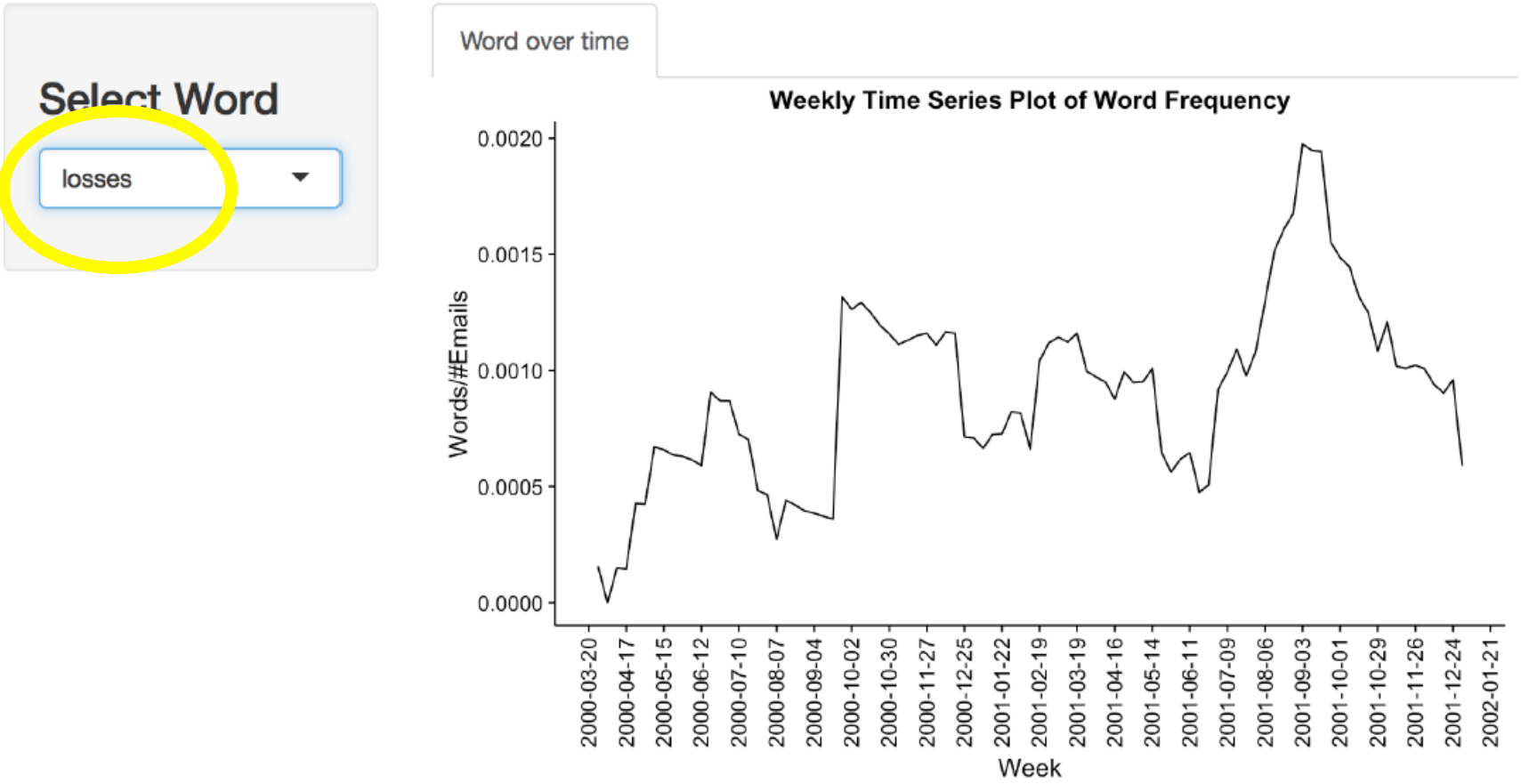


Figure 14. Topic Analysis over Time

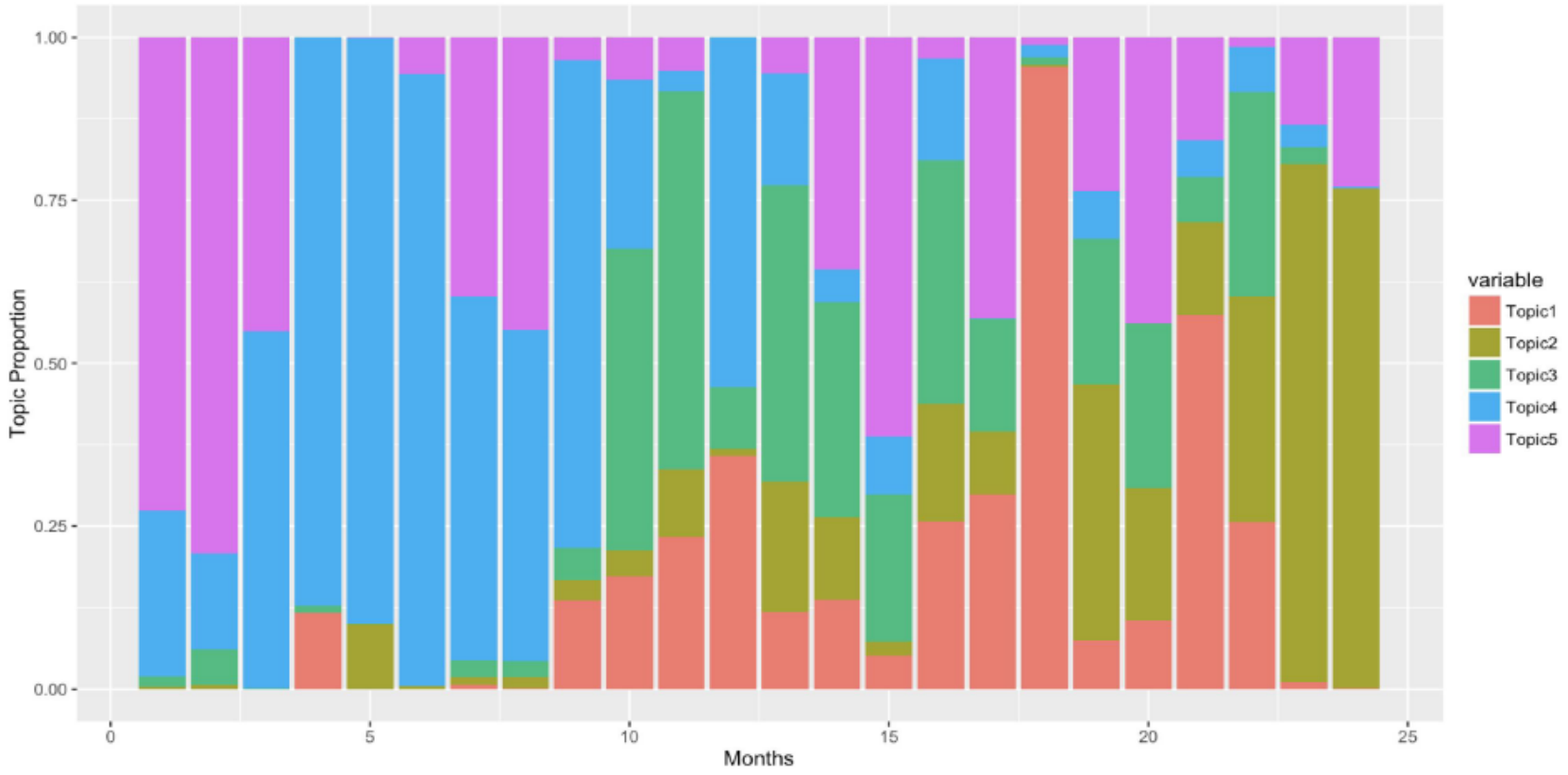
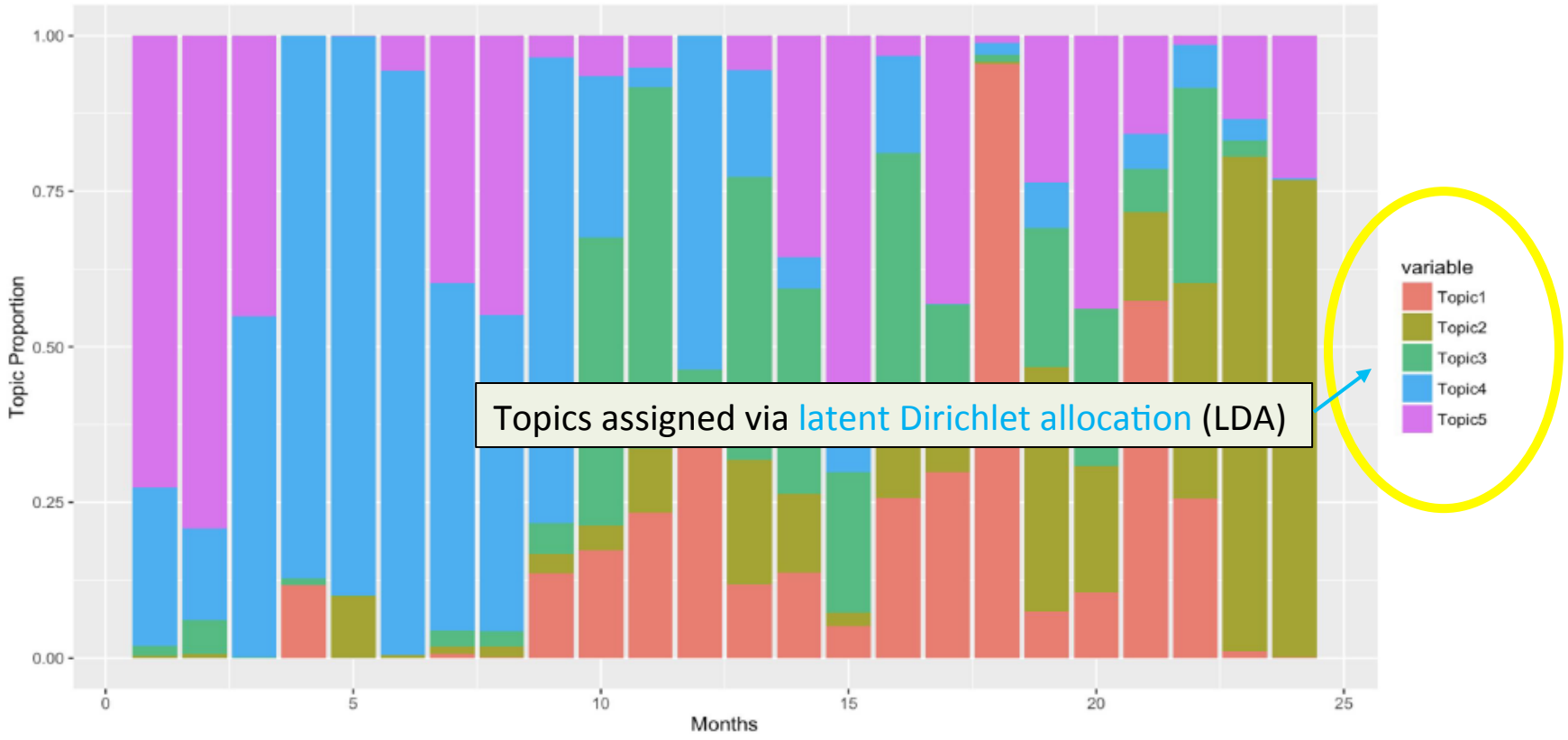


Figure 14. Topic Analysis over Time



Concluding Remarks

- We introduce an automated platform to parse corporate email content, and we find that the net sentiment conveyed by employee sent mails is a timely indicator of stock-return performance.

- Non-verbal indicators, such as email length and network structure, are particularly promising avenues to explore.

- Overall, we suggest the promise of a regulatory technology (RegTech) approach by which to systematically parse email content and network structure to detect indicators of risk or malfeasance on an ongoing and more timely basis.

Thank you.