



CAIA
ASSOCIATION®



FDP
INSTITUTE™

Financial Data Professional Institute

Analyzing Text to Detect Risk

Seoyoung Kim, Professor of Business Analytics at Santa Clara University
Mehrzaad Mahdavi, Executive Director, FDP Institute
Kathy Wilkens, Senior Advisor, FDPI Curriculum
Mirjam Dekker, Project Manager, FDP Institute

www.fdpinstitute.org

February 19, 2020

Agenda

- Welcome
- Introductions



Seoyoung Kim
Professor of Business
Analytics at Santa Clara
University



Katherine Wilkens
Sr. Curriculum
Advisor FDPI



Mehrzad Mahdavi
Executive Director
FDPI



Mirjam Dekker
Project Manager
FDPI

- Seoyoung Kim's research presentation
- FDP Curriculum
- Q & A

Zero-Revelation RegTech: Detecting Risk through Corporate Emails

Seoyoung Kim
Santa Clara University

@ FDP Institute Webinar Series
Santa Clara, CA
February 19, 2020

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

Zero-Revelation RegTech: Detecting Risk through Corporate Emails

Seoyoung Kim
Santa Clara University

@ FDP Institute Webinar Series
Santa Clara, CA
February 19, 2020

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

Analyses

Table 1. Summary Statistics of Sent Mail

<i>Panel A. Characteristics by Employee (N = 144)</i>						
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
<i>Panel B. Email Characteristics (N = 113,266)</i>						
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

<i>Panel A. Characteristics by Employee (N = 144)</i>						
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
<i>Panel B. Email Characteristics (N = 113,266)</i>						
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

<i>Panel A. Characteristics by Employee (N = 144)</i>						
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
<i>Panel B. Email Characteristics (N = 113,266)</i>						
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	Many emails (close to 11%) are simply forwarded without added text.				
Total Recipients per Email	1.77	1	1	1	2	20

Figure 1. Average Email Length over Time

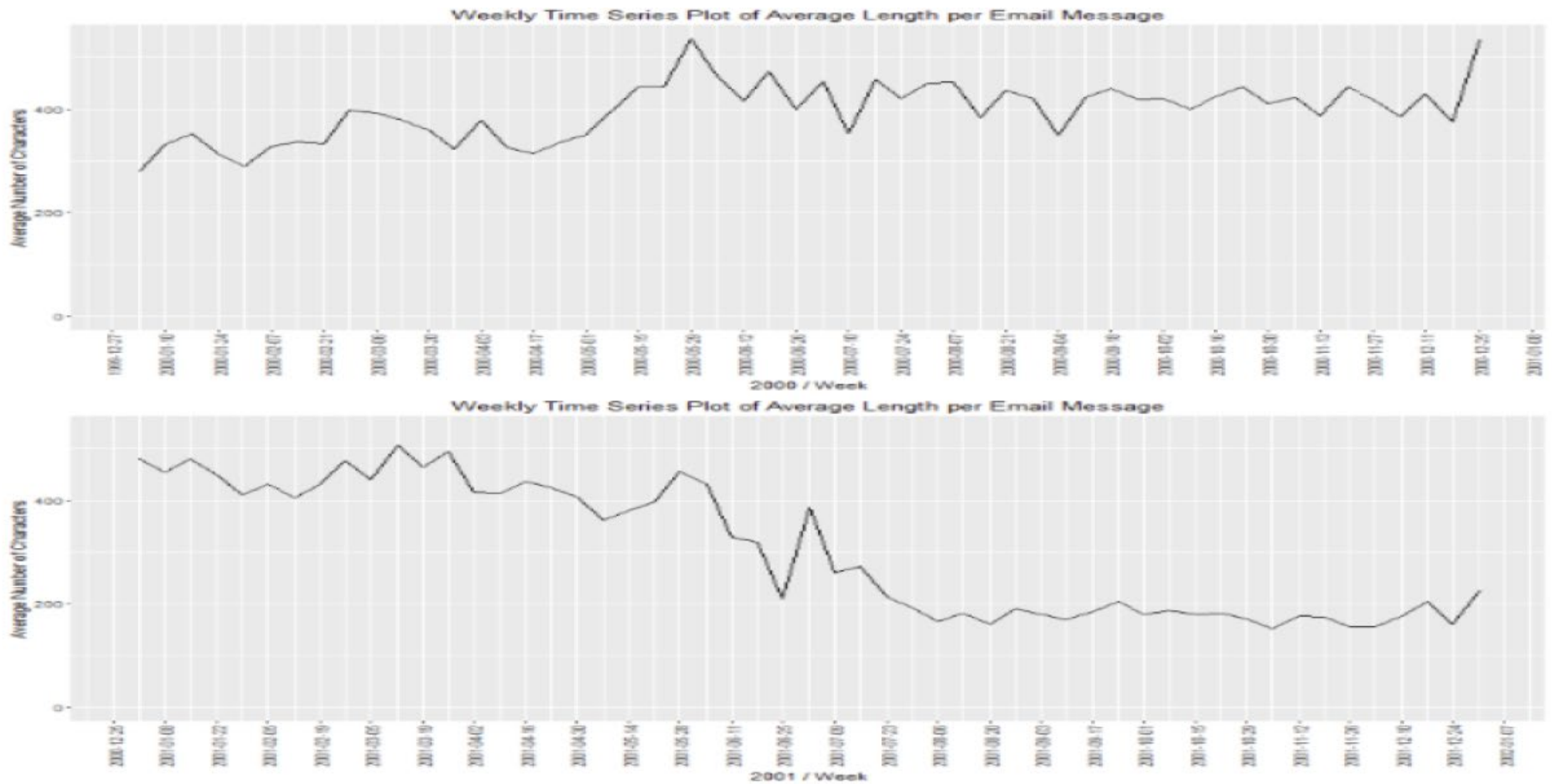
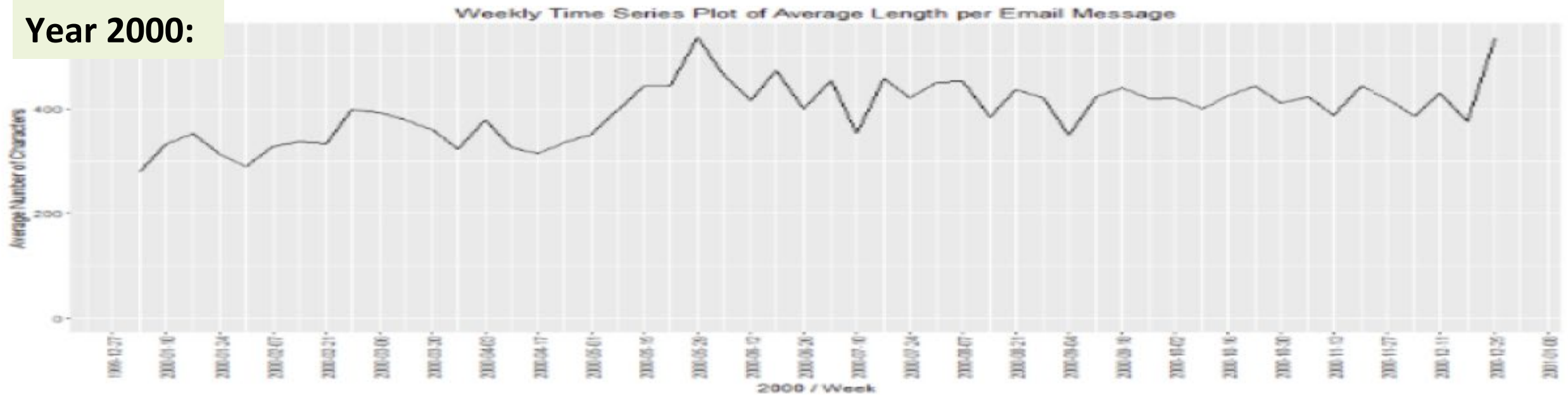


Figure 1. Average Email Length over Time

Year 2000:



Year 2001:

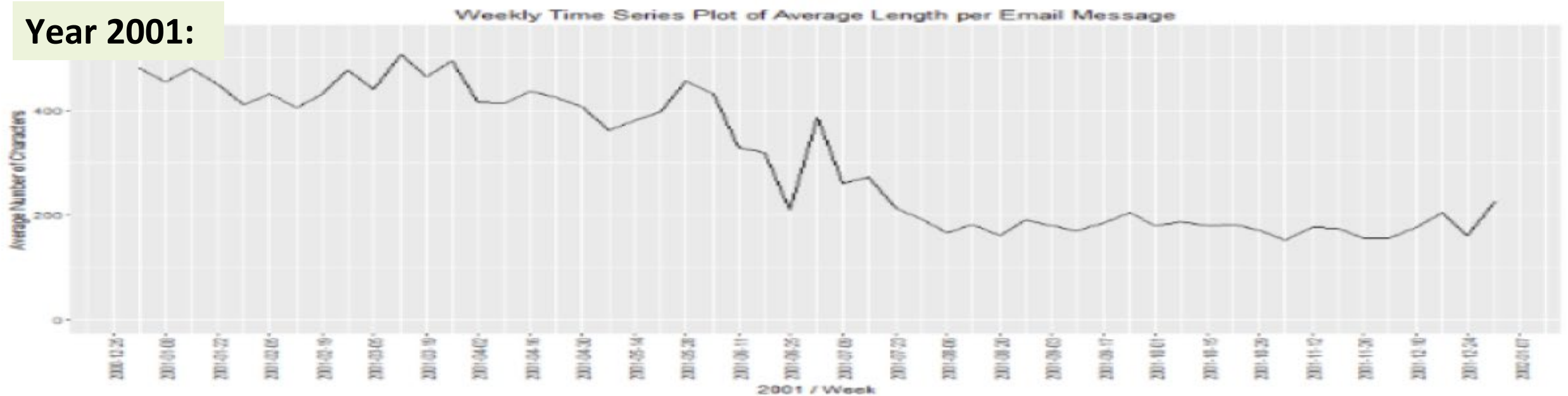
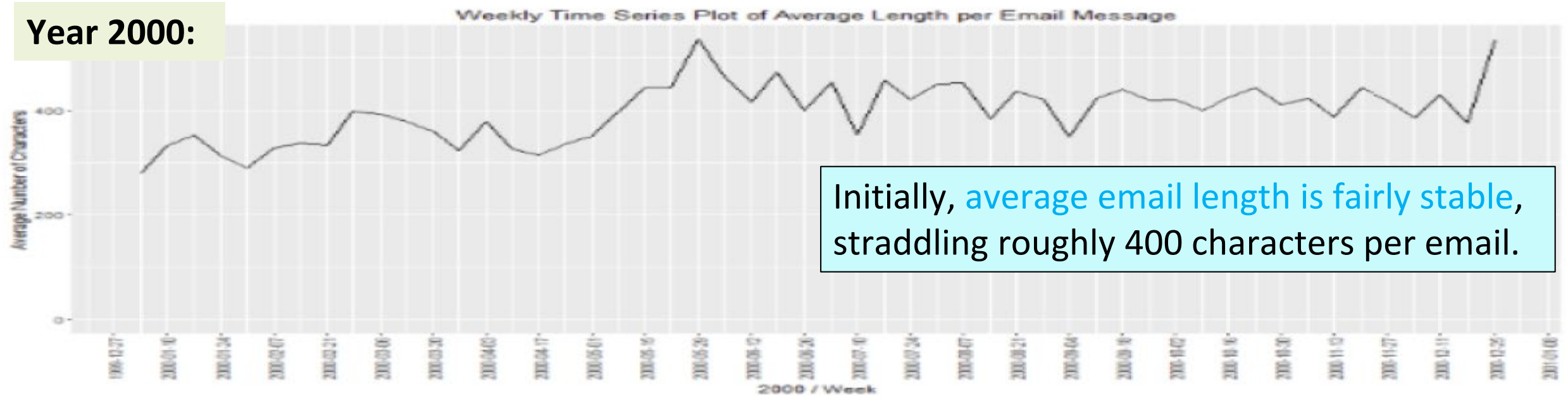


Figure 1. Average Email Length over Time

Year 2000:



Year 2001:

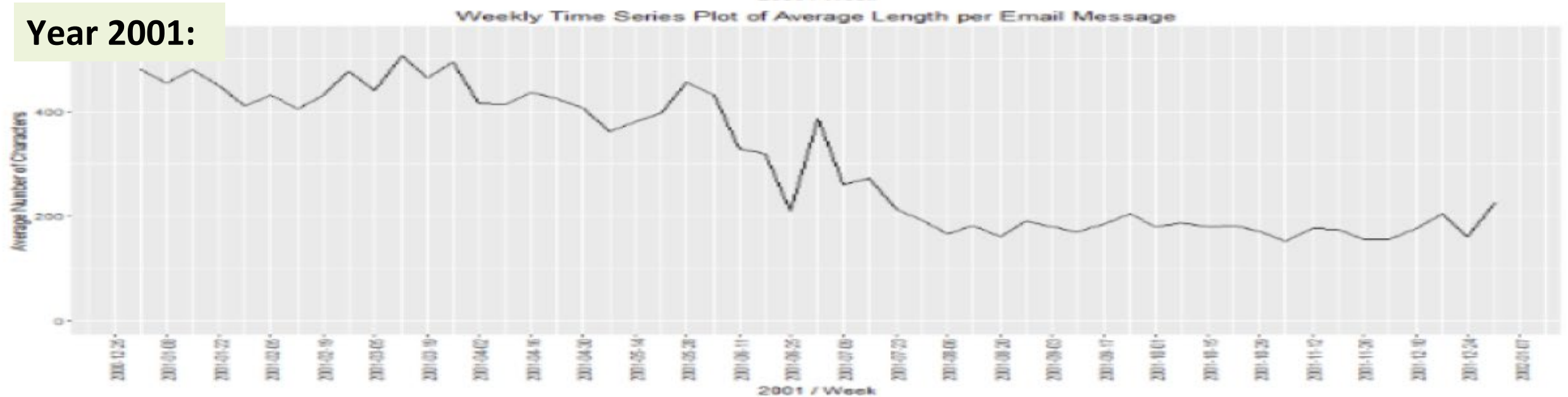
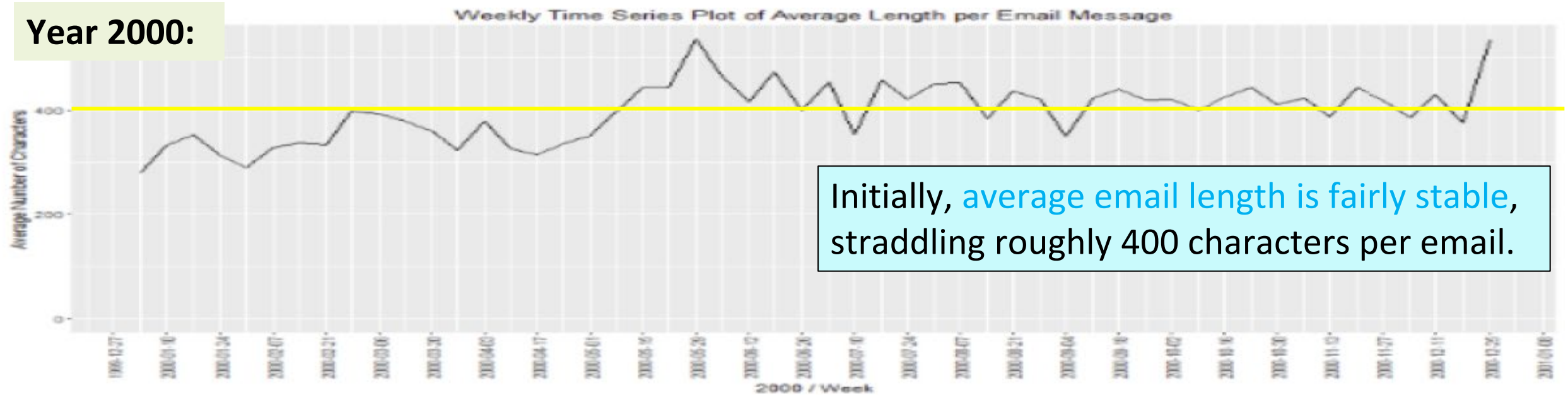


Figure 1. Average Email Length over Time

Year 2000:



Year 2001:

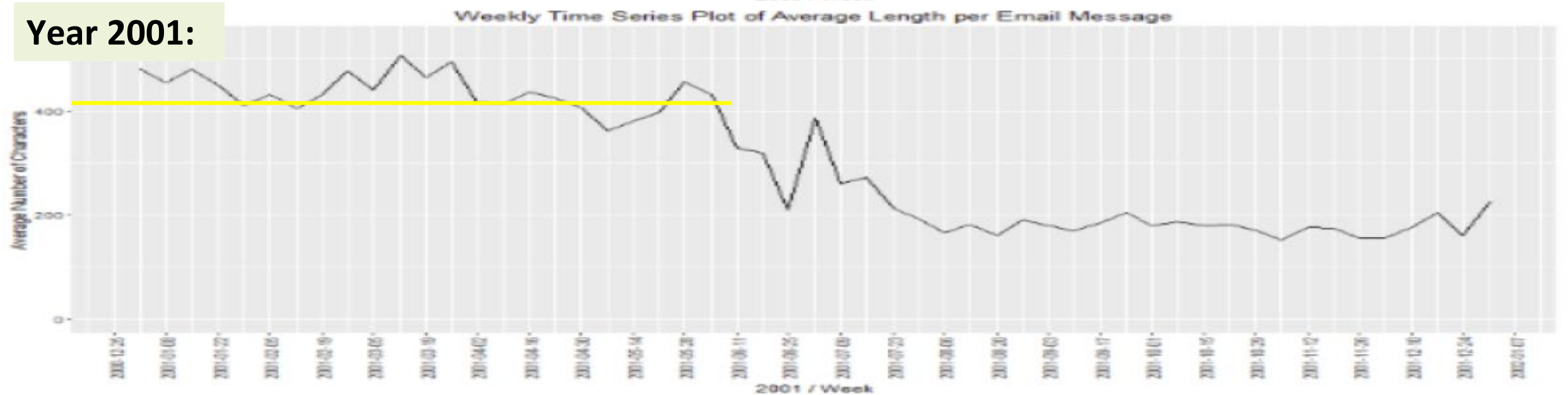
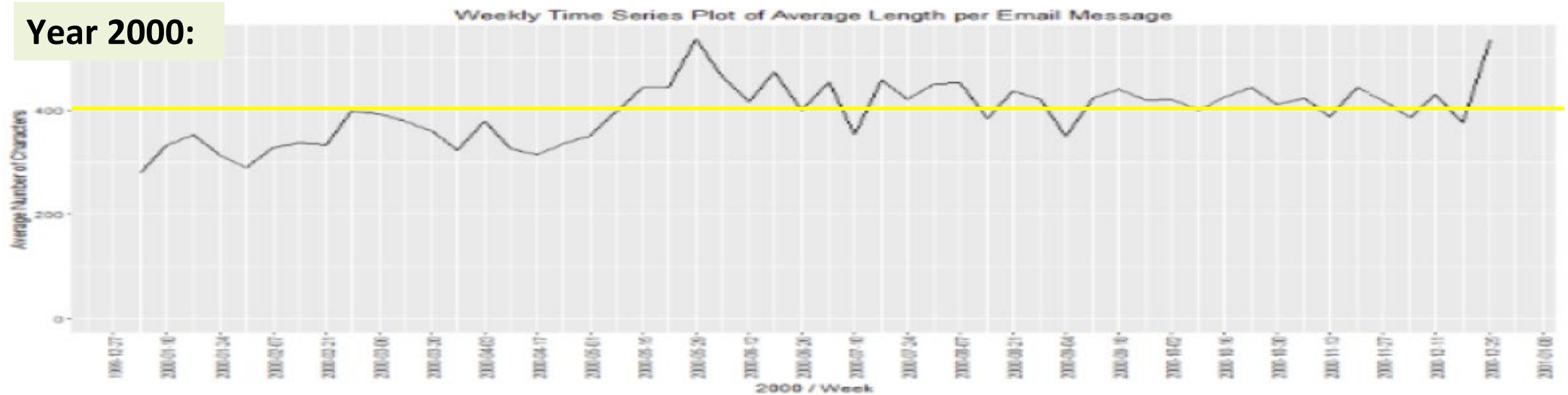
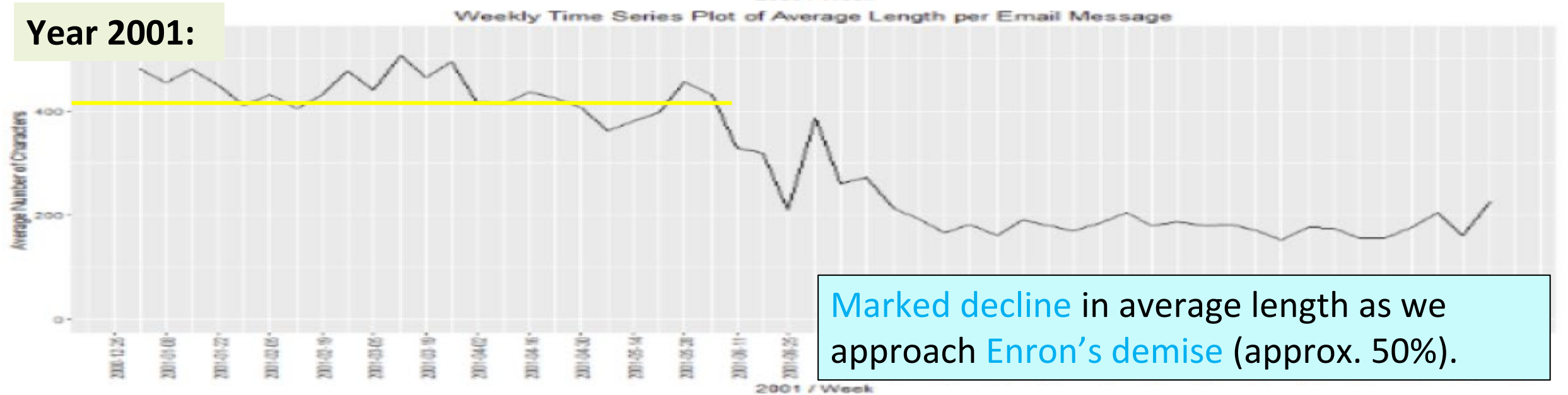


Figure 1. Average Email Length over Time

Year 2000:



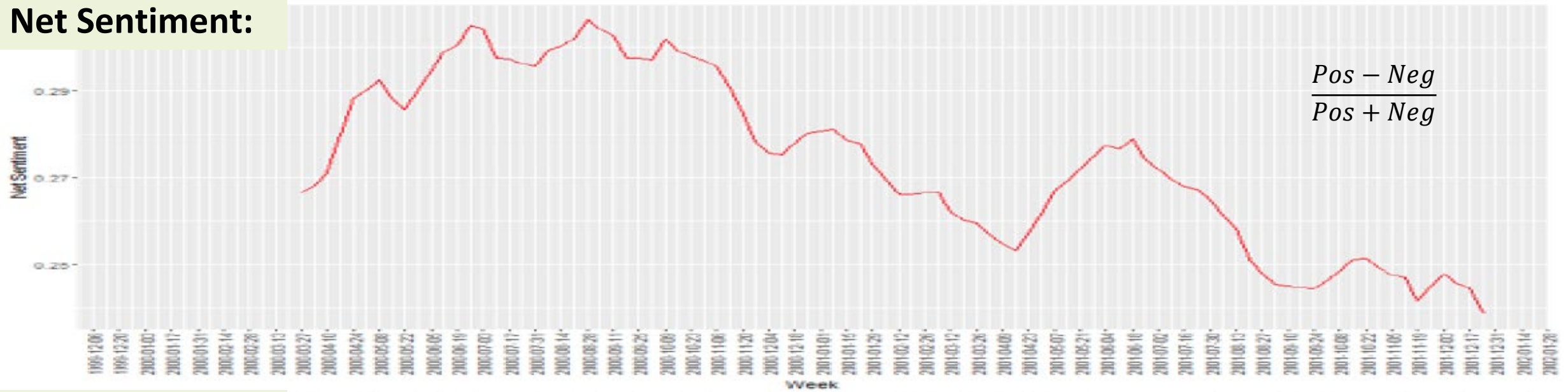
Year 2001:



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

Figure 2. Email Sentiment and Disagreement over Time

Net Sentiment:



Disagreement:

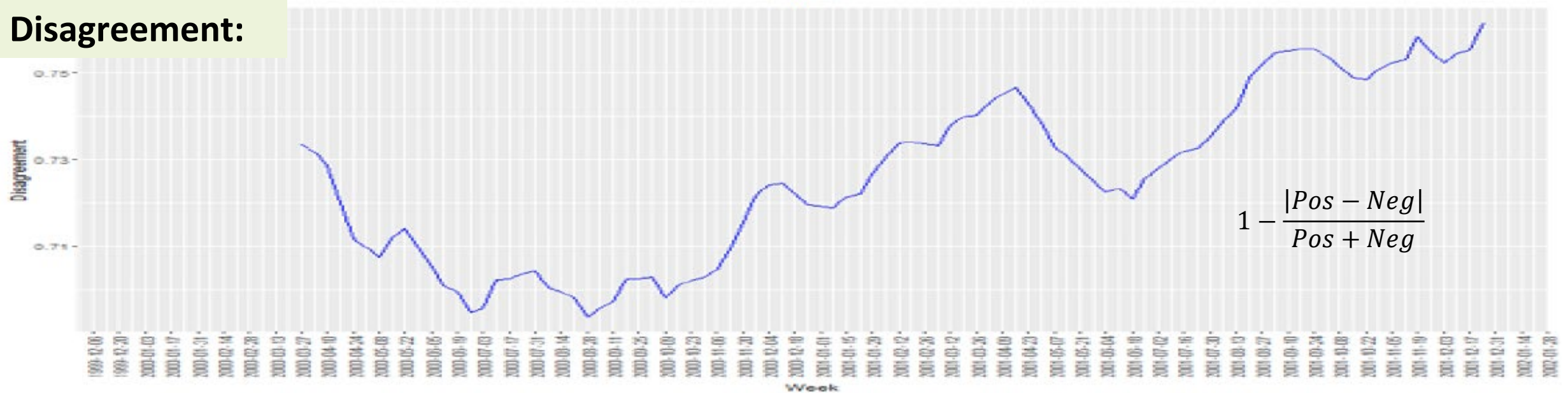
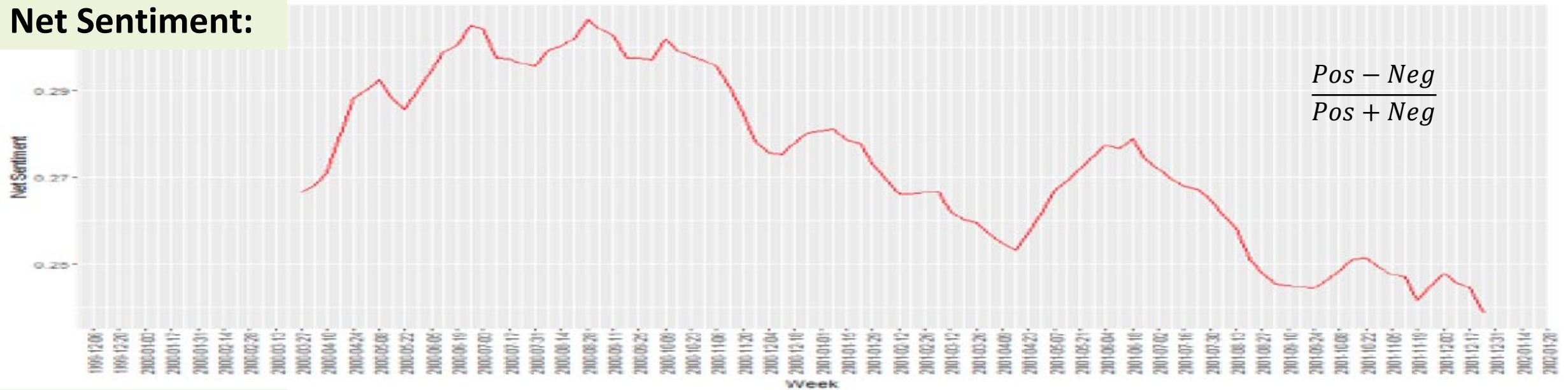
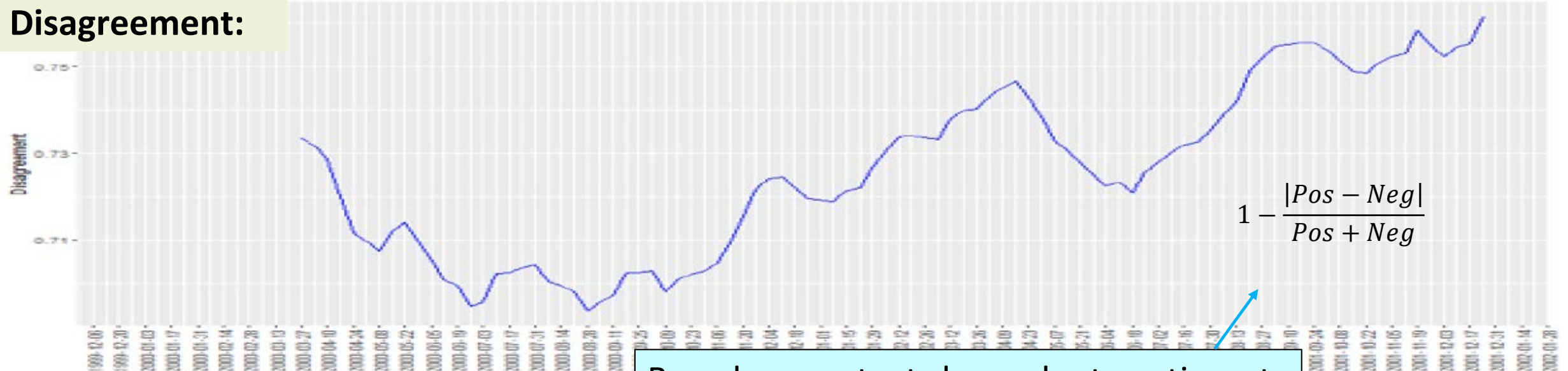


Figure 2. Email Sentiment and Disagreement over Time

Net Sentiment:



Disagreement:



Based on context-dependent sentiment dictionaries for word classification

Figure 3. Factiva News Coverage over Time

Weekly Time Series Plot of Articles Published

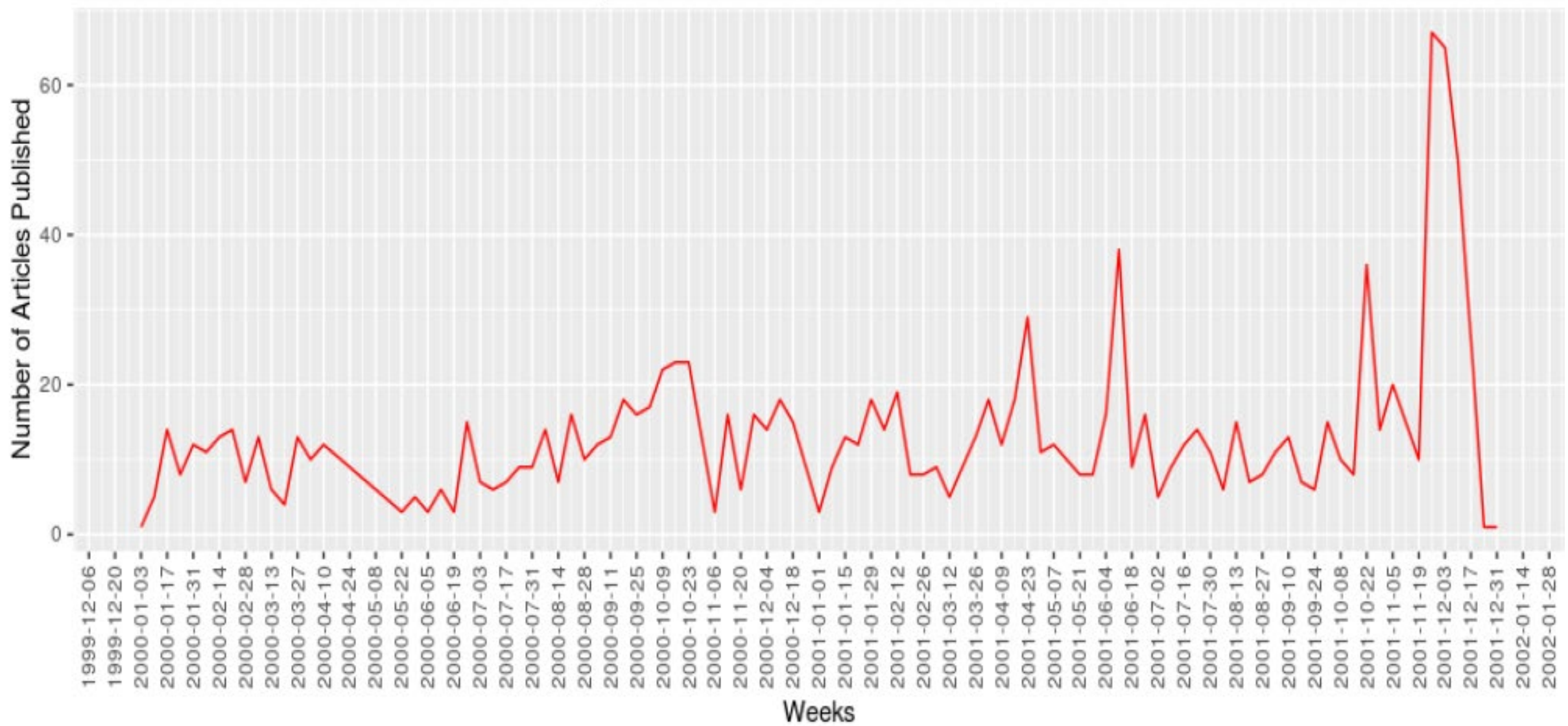


Figure 3. Factiva News Coverage over Time

Weekly Time Series Plot of Articles Published

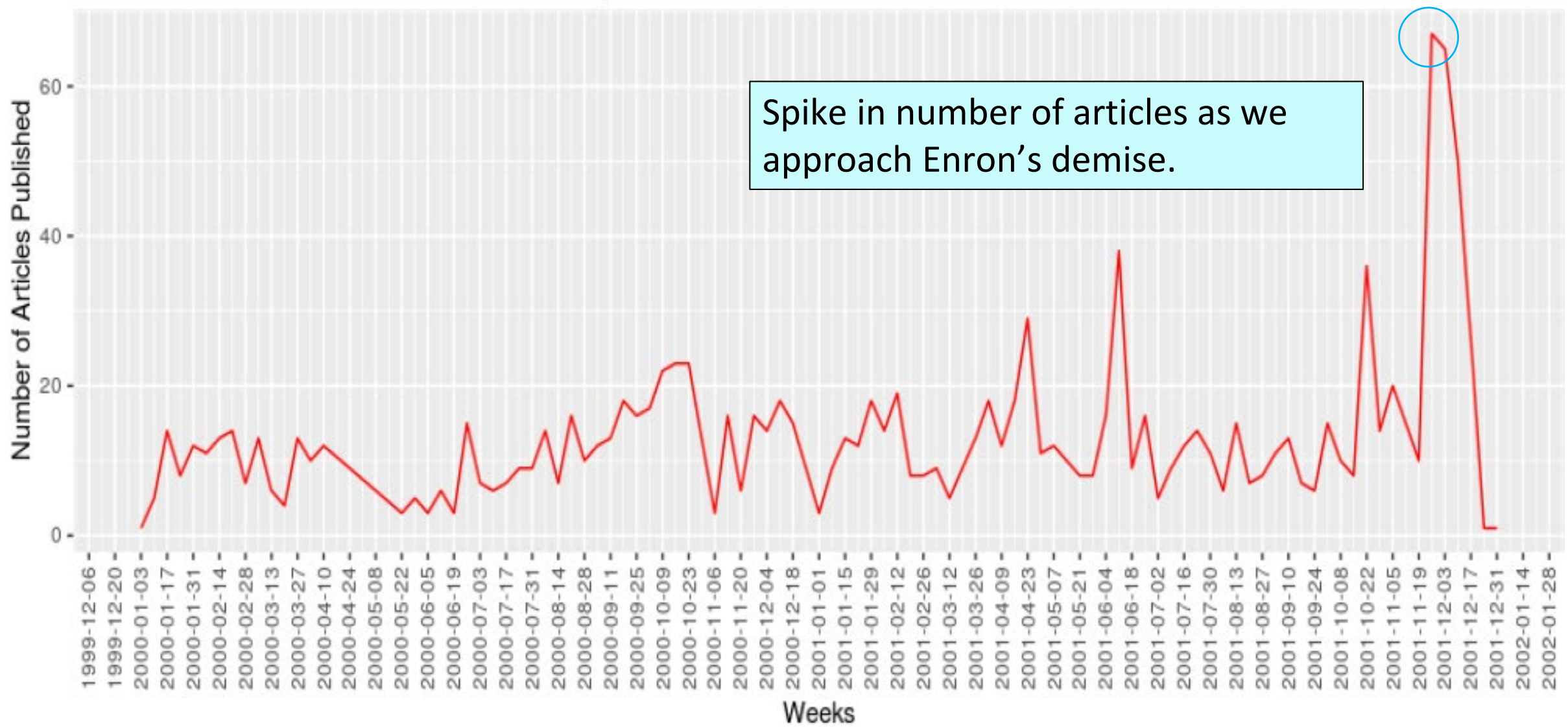


Figure 4. Factiva News Sentiment over Time

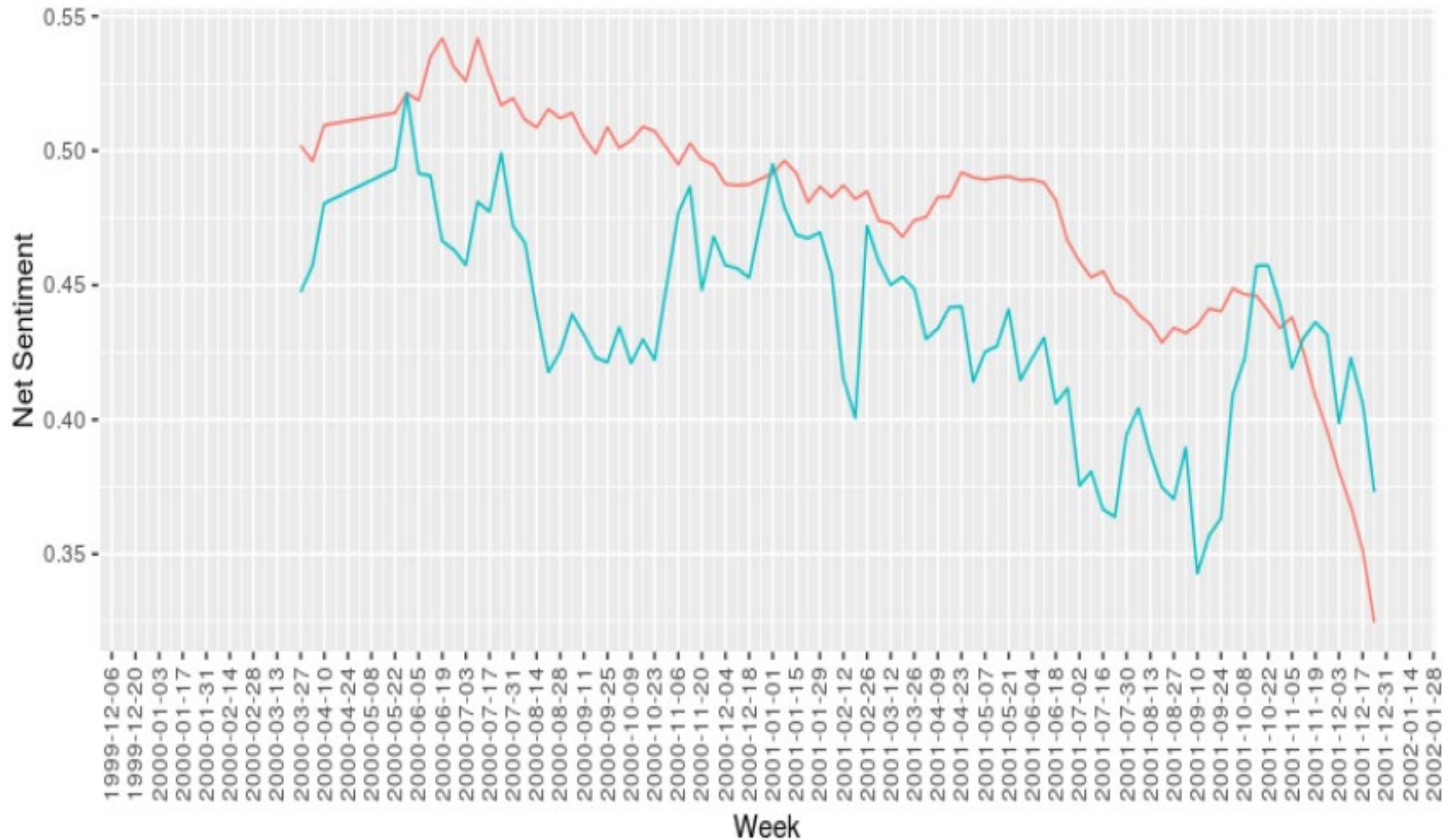


Figure 4. Factiva News Sentiment over Time

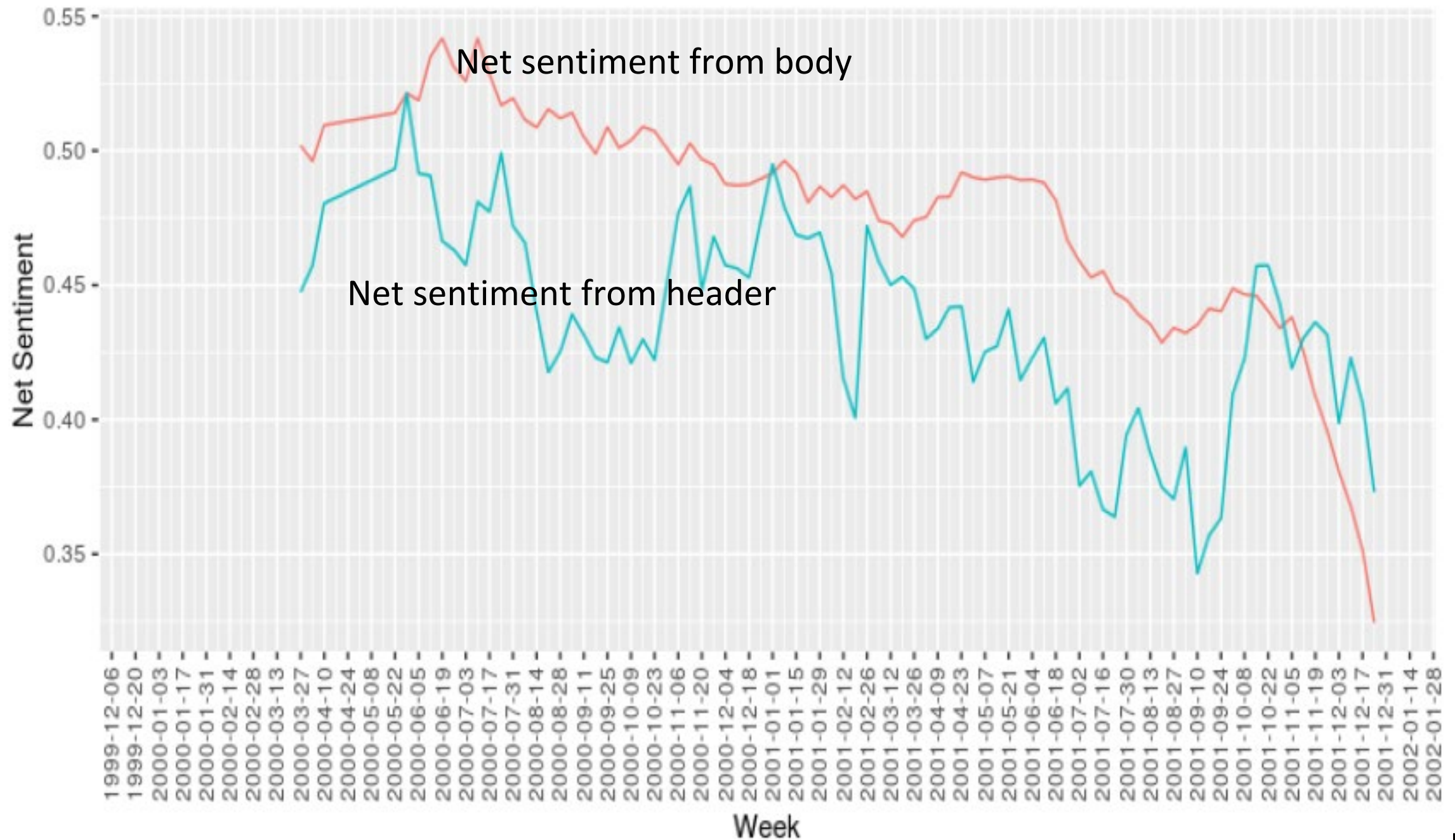
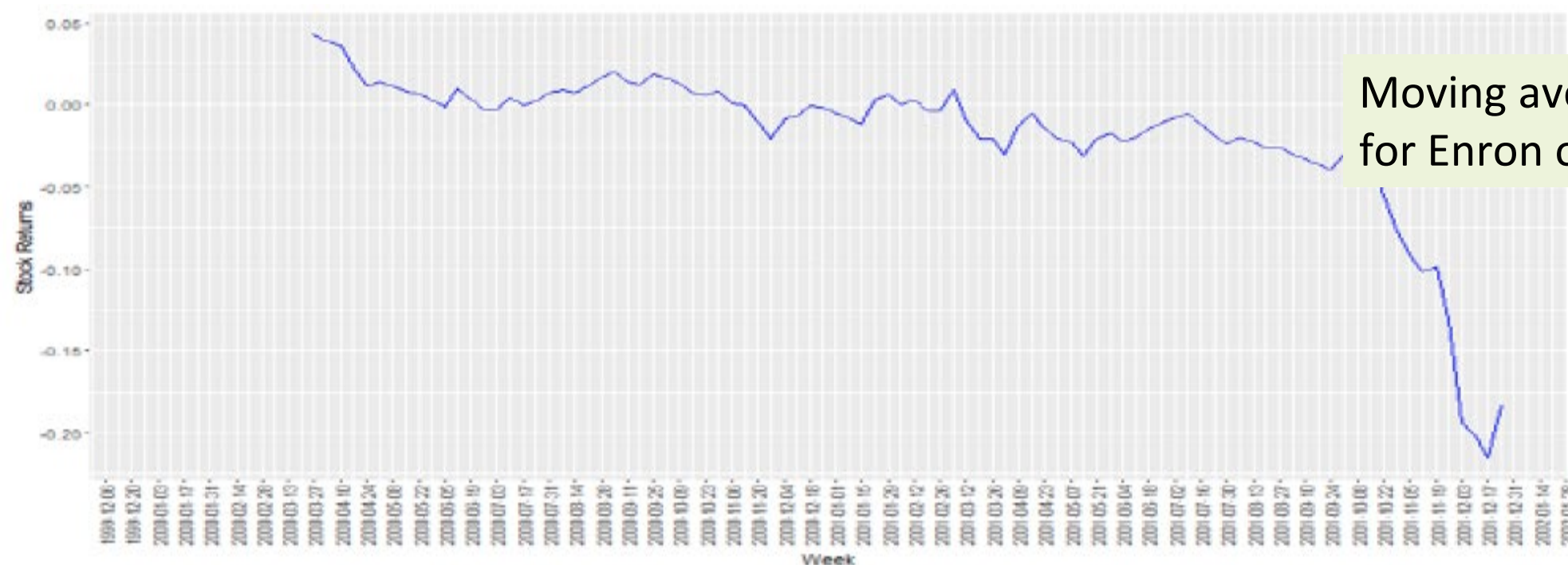


Figure 5. Stock Returns and Net Sentiment over Time

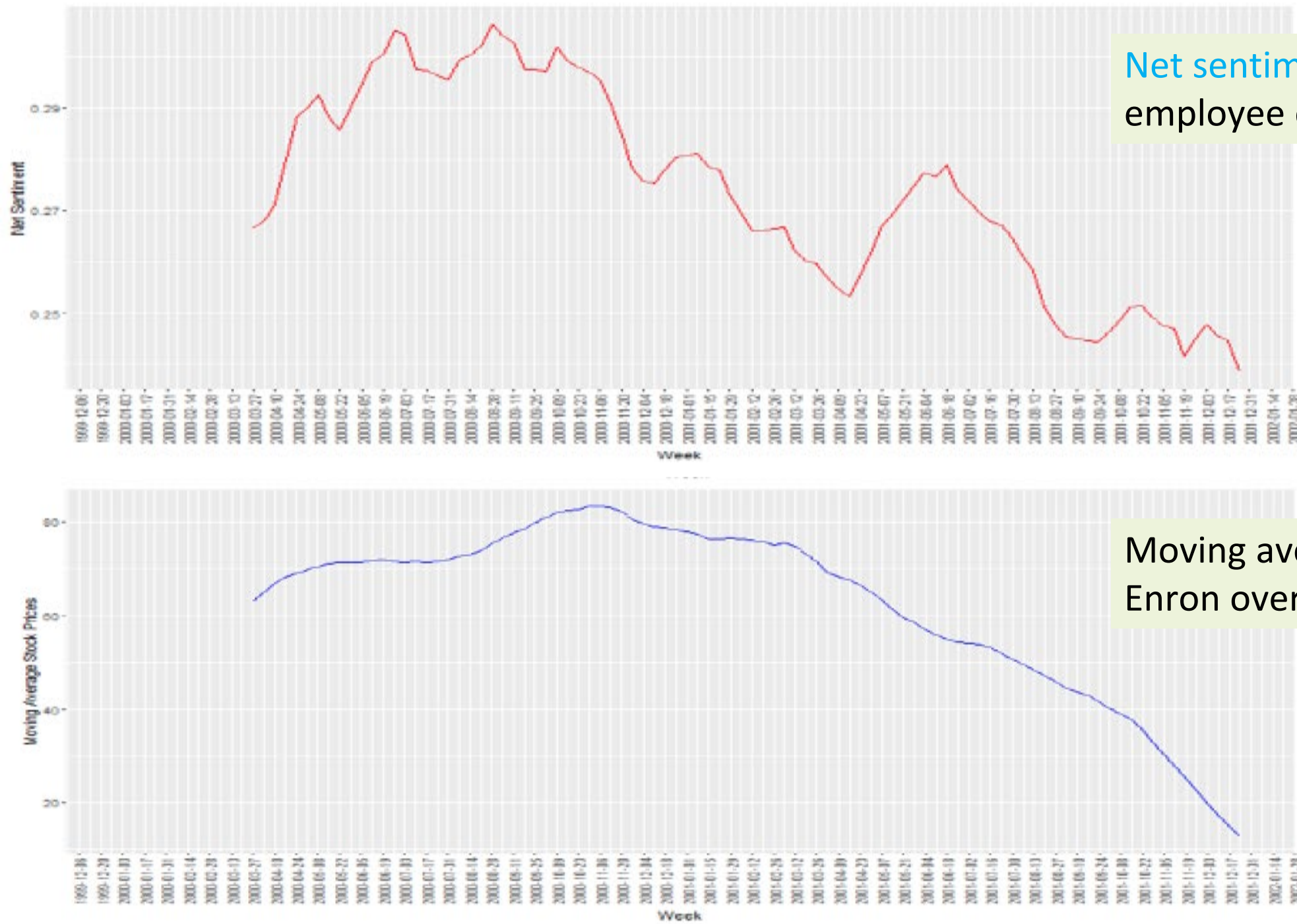


Net sentiment across Enron employee emails over time



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

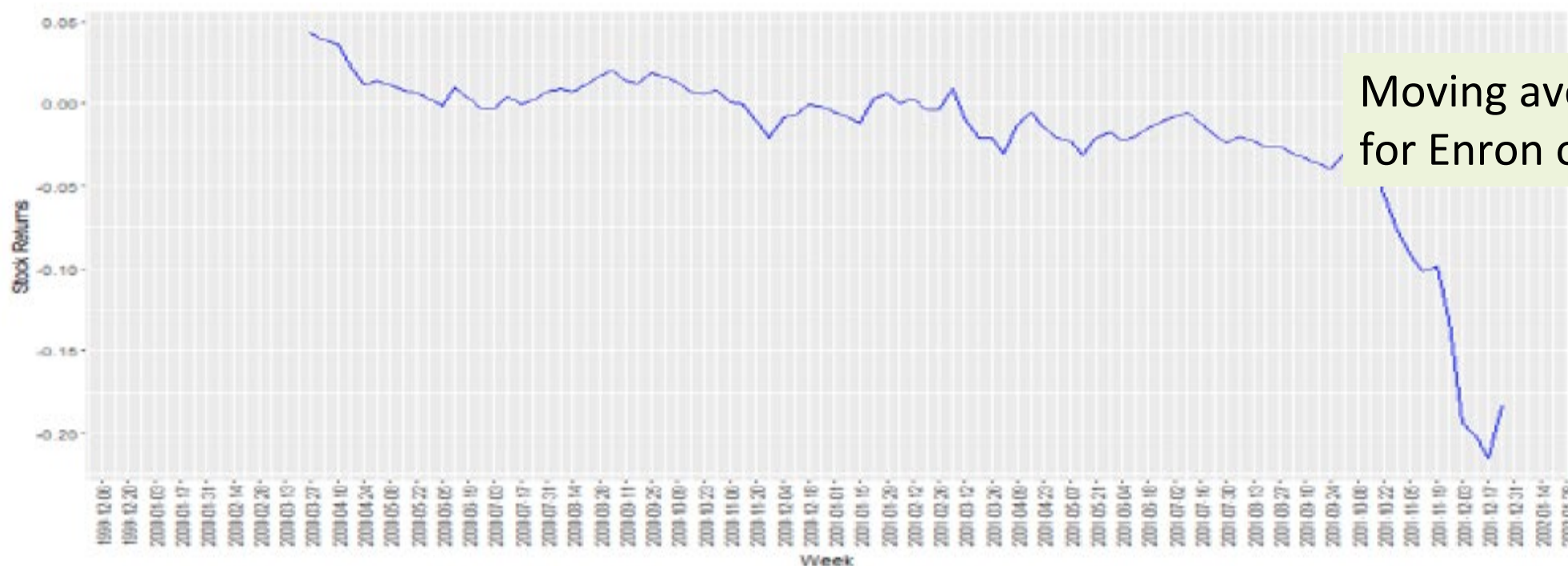
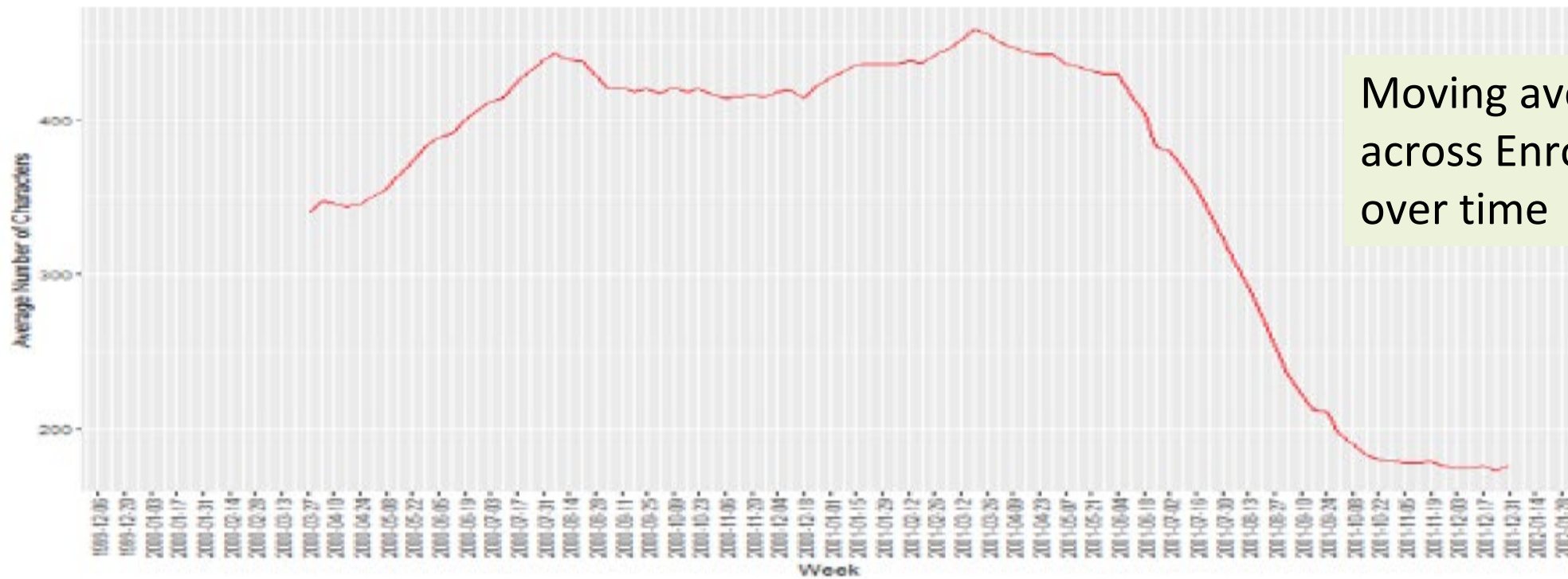
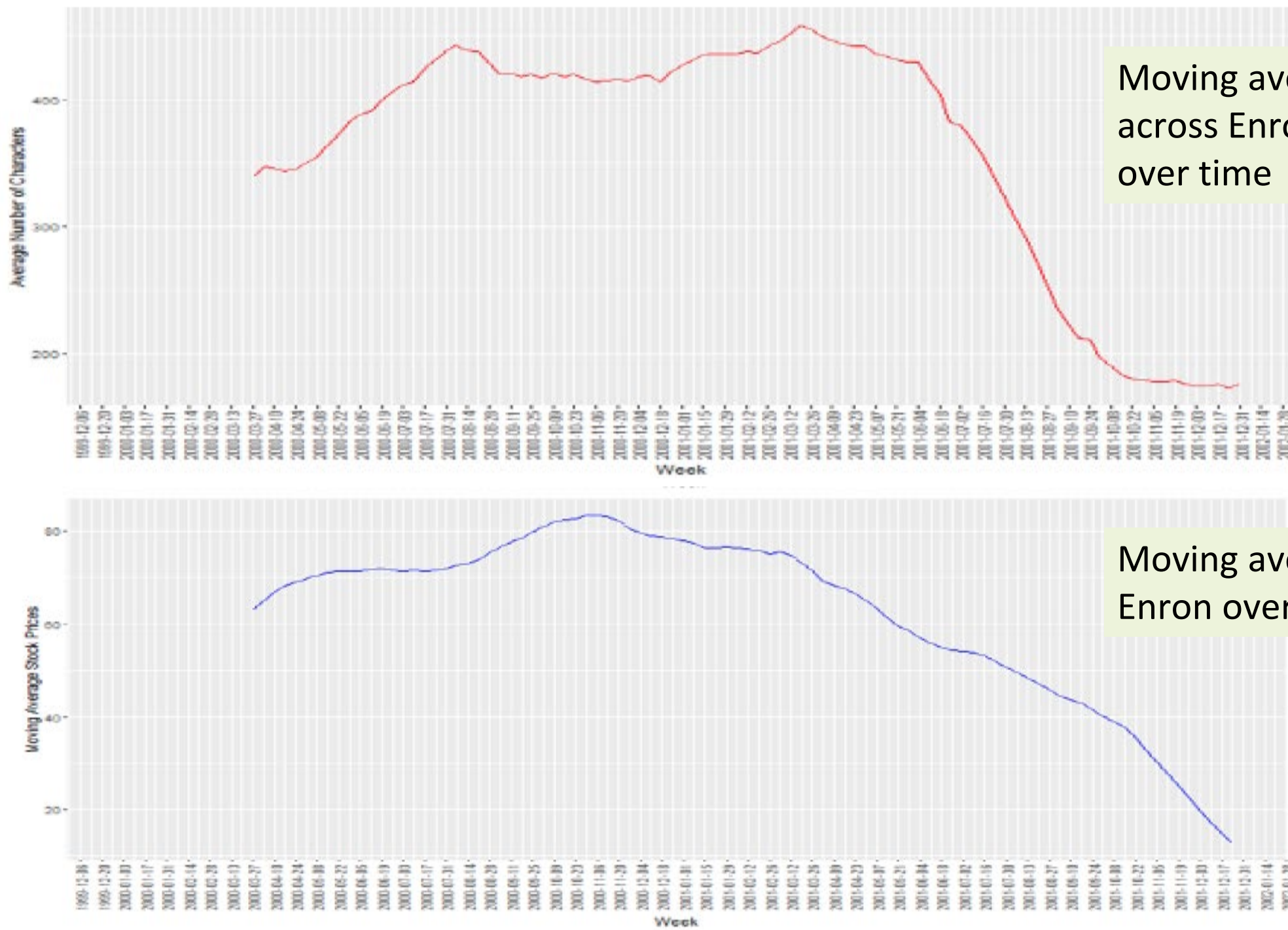


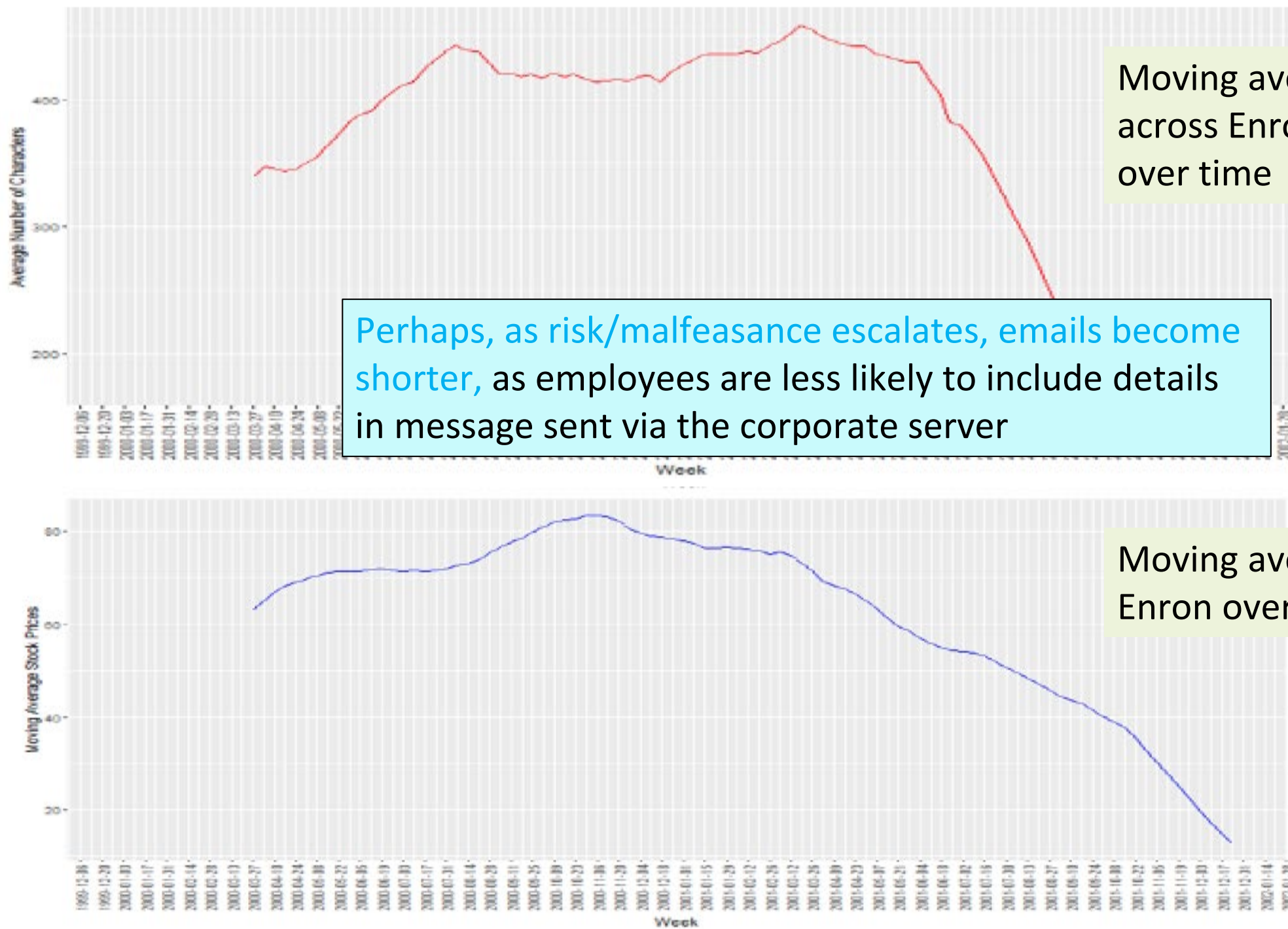
Figure 8. Stock Prices and Email Length over Time



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*** (3.27)	0.575 (0.63)	2.330*** (3.14)	-1.397 (-1.25)
MA Email Length $_t$		0.584*** (2.97)		1.046*** (4.19)
MA Total Emails $_t$			-0.004 (-0.10)	-0.131*** (-2.83)
Intercept	-0.680*** (-3.45)	-0.406* (-1.93)	-0.671*** (-3.08)	0.117 (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 _{<i>t</i>}	2.347*** (3.27)	0.575 (0.63)	2.330*** (3.14)	-1.397 (-1.25)
MA Email Length _{<i>t</i>}				
MA Total Emails _{<i>t</i>}			-0.004 (-0.10)	-0.131*** (-2.83)
Intercept	-0.680*** (-3.45)	-0.406* (-1.93)	-0.671*** (-3.08)	0.117 (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*** (3.27)	0.575 (0.63)	2.330*** (3.14)	-1.397 (-1.25)
MA Email Length _t		0.584*** (2.97)		1.046*** (4.19)
MA Total Emails _t			-0.004 (-0.10)	-0.131*** (-2.83)
Intercept	-0.680*** (-3.45)	-0.406* (-1.93)	-0.671*** (-3.08)	0.117 (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*** (3.27)	0.575 (0.63)	2.330*** (3.14)	-1.397 (-1.25)
MA Email Length $_t$		0.584*** (2.97)		1.046*** (4.19)
MA Total Emails $_t$				
Intercept	-0.680*** (-3.45)	-0.406* (-1.93)	-0.671*** (-3.08)	0.117 (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 (-1.31)	-1.136* (-1.96)	-0.772 (-1.34)	-1.210** (-2.03)	-0.893 (-1.61)
MA Email Sentiment _t		2.628*** (3.30)	0.705 (0.66)	2.566*** (3.18)	-1.254 (-1.03)
MA Email Length _t			0.560** (2.59)		1.026*** (3.93)
MA Total Emails _t				-0.024 (-0.59)	-0.138*** (-2.91)
Intercept	0.307 (1.15)	-0.256 (-0.84)	-0.096 (0.75)	-0.178 (-0.54)	0.485 (1.39)
Adjusted <i>R</i> 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 (-1.31)	-1.136* (-1.96)	-0.772 (-1.34)	-1.210** (-2.03)	-0.893 (-1.61)
MA Email Sentiment _t		2.628*** (3.30)	0.705 (0.66)	2.566*** (3.18)	-1.254 (-1.03)
MA Email Length _t			0.560**		1.026***
MA Total Emails _t				-0.024 (-0.59)	-0.138*** (-2.91)
Intercept	0.307 (1.15)	-0.256 (-0.84)	-0.096 (0.75)	-0.178 (-0.54)	0.485 (1.39)
Adjusted <i>R</i> 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 (-1.31)	-1.136* (-1.96)	-0.772 (-1.34)	-1.210** (-2.03)	-0.893 (-1.61)
MA Email Sentiment _t		2.628*** (3.30)	0.705 (0.66)	2.566*** (3.18)	-1.254 (-1.03)
MA Email Length _t			0.560** (2.59)		1.026*** (3.93)
MA Total Emails _t				0.024	0.128***
Intercept	0.307 (1.15)	-0.256 (-0.84)	-0.096 (0.75)	-0.178 (-0.54)	0.485 (1.39)
Adjusted <i>R</i> 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*** (3.95)	1.501** (2.49)	0.657 (0.87)	1.505** (2.48)	-0.827 (-0.92)
MA Email Sentiment _t		-0.245 (-0.19)	0.377 (-0.29)	-0.284 (-0.22)	-1.293 (-1.02)
MA Email Length _t			0.486* (1.81)		1.380*** (3.34)
MA Total Emails _t				-0.009 (-0.24)	-0.164*** (-2.77)
Intercept	-0.711*** (-4.18)	-0.688*** (-3.27)	-0.426* (-1.69)	-0.668*** (-2.94)	0.399 (1.04)
Adjusted <i>R</i> 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*** (3.95)	1.501** (2.49)	0.657 (0.87)	1.505** (2.48)	-0.827 (-0.92)
MA Email Sentiment _t		-0.245 (-0.19)	0.377 (-0.29)	-0.284 (-0.22)	-1.293 (-1.02)
MA Email Length _t					
MA Total Emails _t					
Intercept	-0.711*** (-4.18)	-0.688*** (-3.27)	-0.426* (-1.69)	-0.668*** (-2.94)	0.399 (1.04)
Adjusted <i>R</i> 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*** (3.95)	1.501** (2.49)	0.657 (0.87)	1.505** (2.48)	-0.827 (-0.92)
MA Email Sentiment _t		-0.245 (-0.19)	0.377 (-0.29)	-0.284 (-0.22)	-1.293 (-1.02)
MA Email Length _t			0.486* (1.81)		1.380*** (3.34)
MA Total Emails _t					
Intercept	-0.711*** (-4.18)	-0.688*** (-3.27)	-0.426* (-1.69)	-0.668*** (-2.94)	0.399 (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:

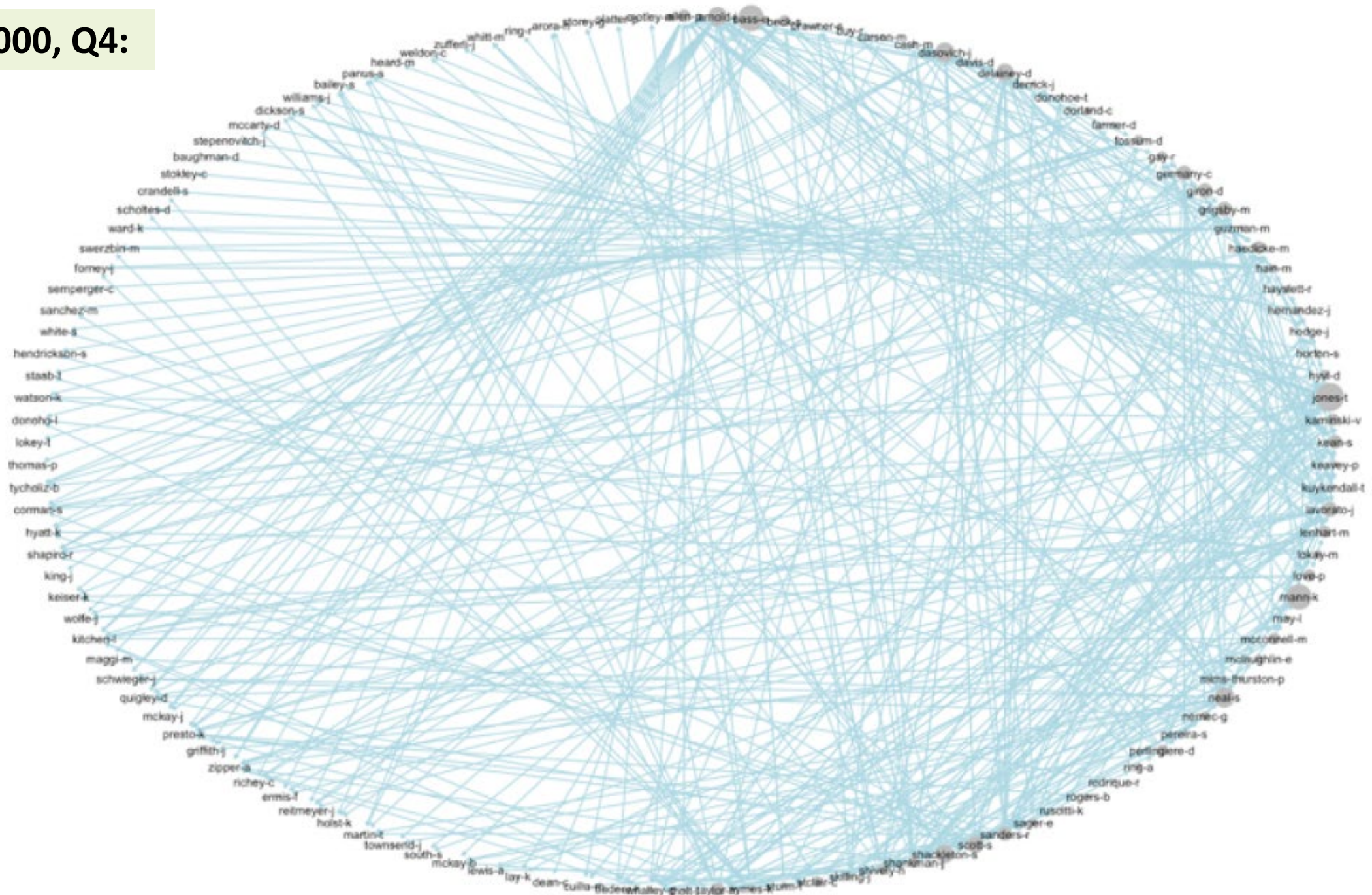


Figure 11. Email Networks

Year 2001, Q4:

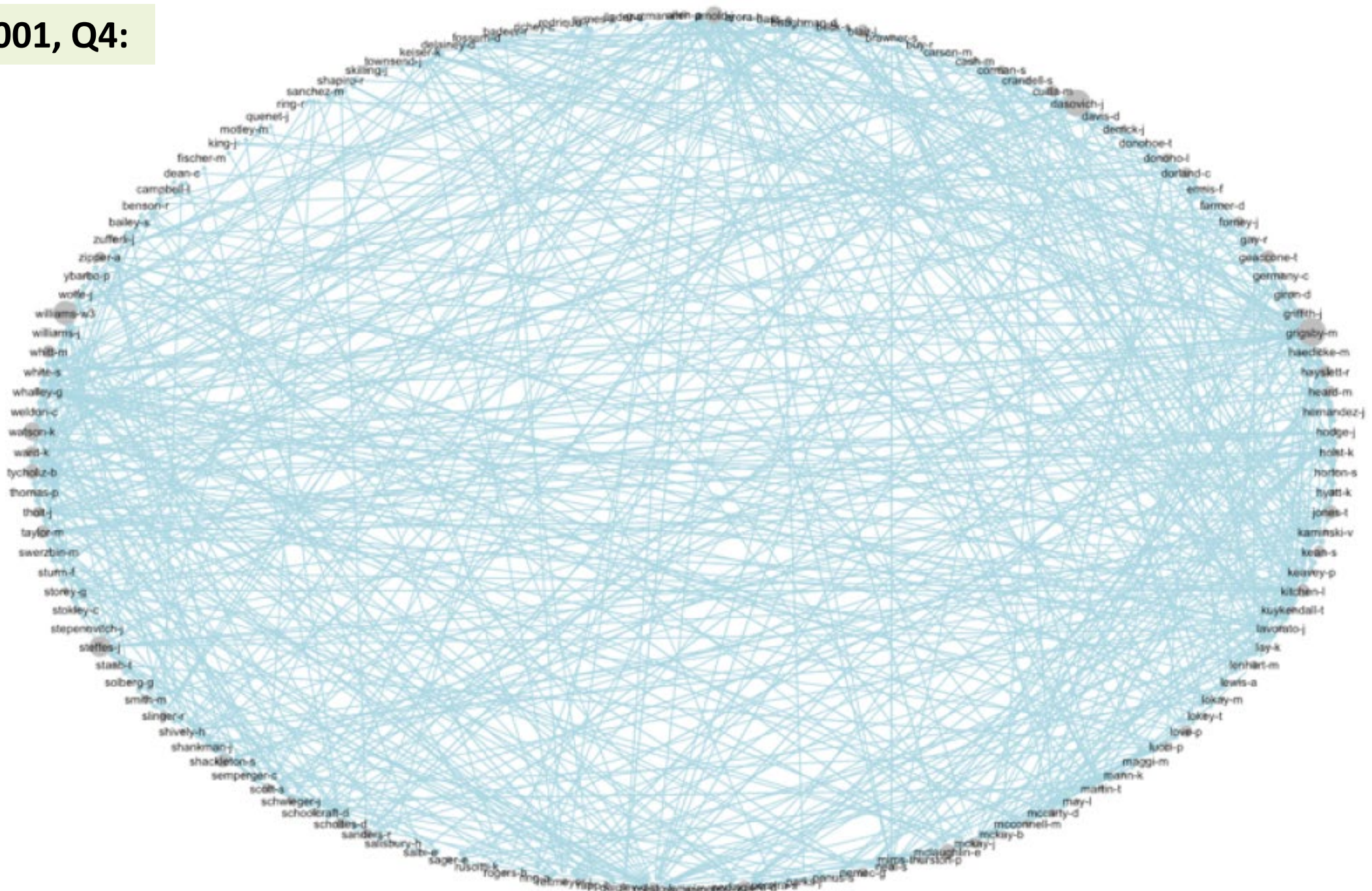


Figure 13. Vocabulary Trends

Select Word

profits

Word over time

Weekly Time Series Plot of Word Frequency

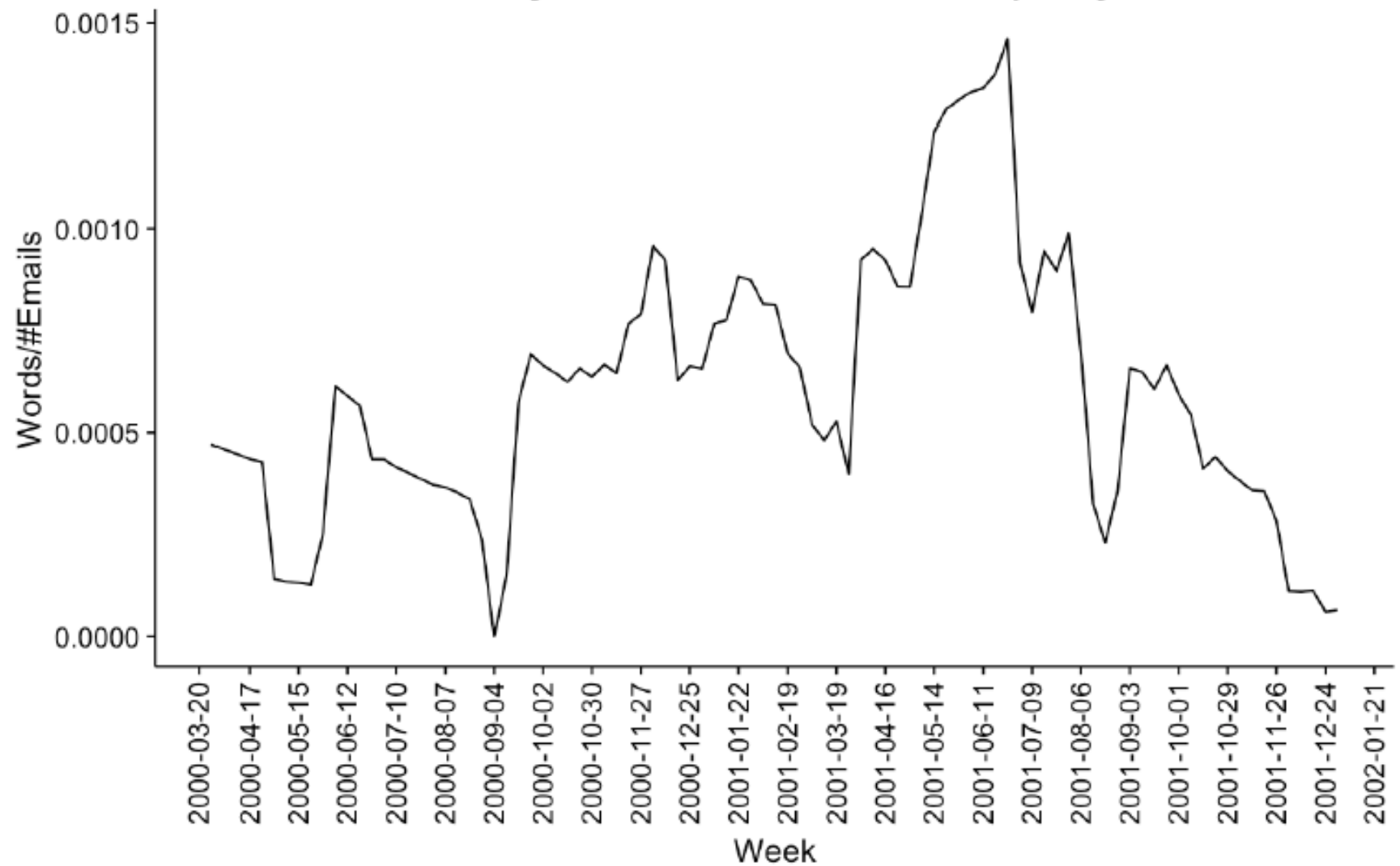


Figure 13. Vocabulary Trends

Select Word

profits

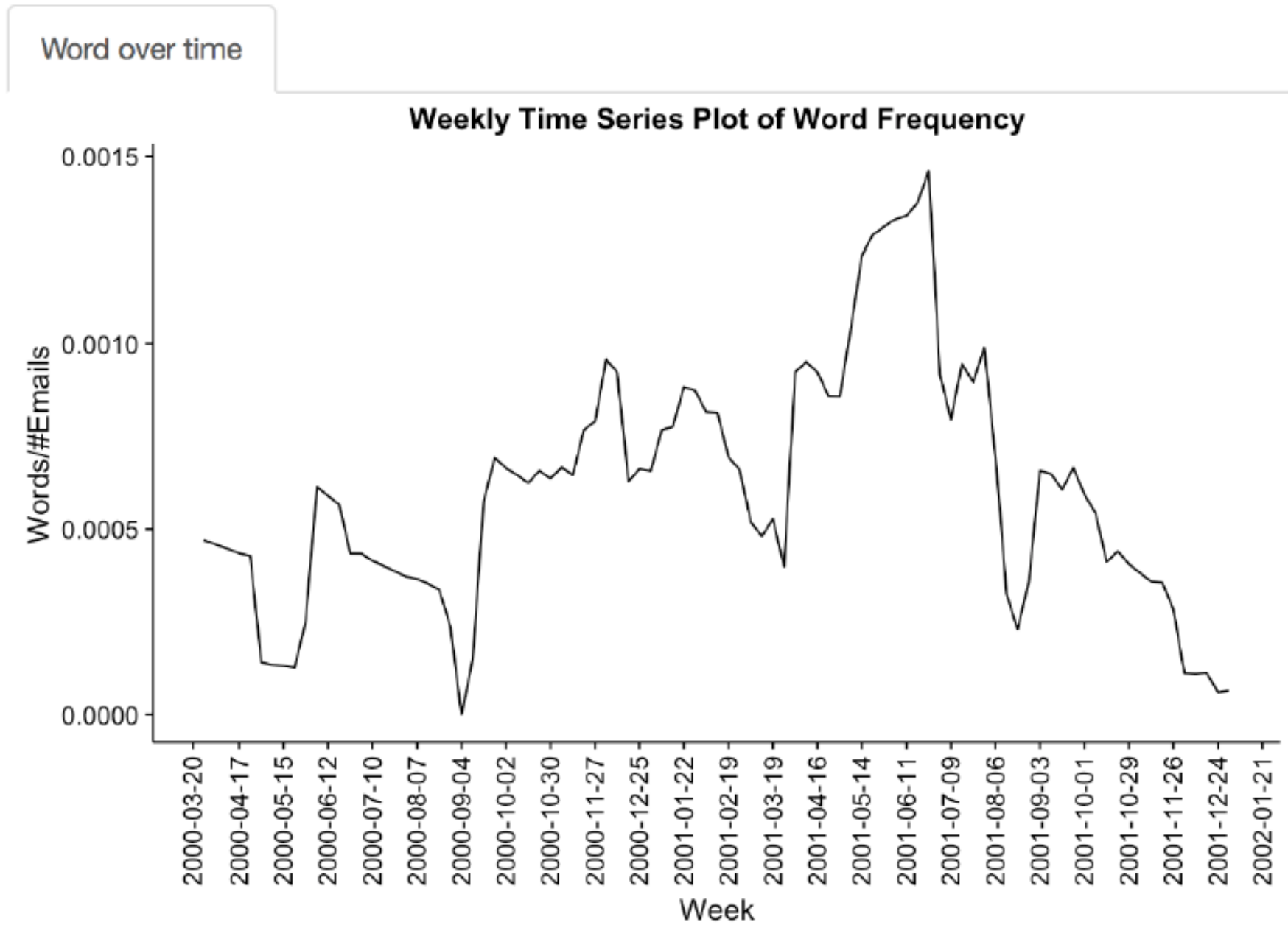


Figure 13. Vocabulary Trends

Select Word

losses

Word over time

Weekly Time Series Plot of Word Frequency

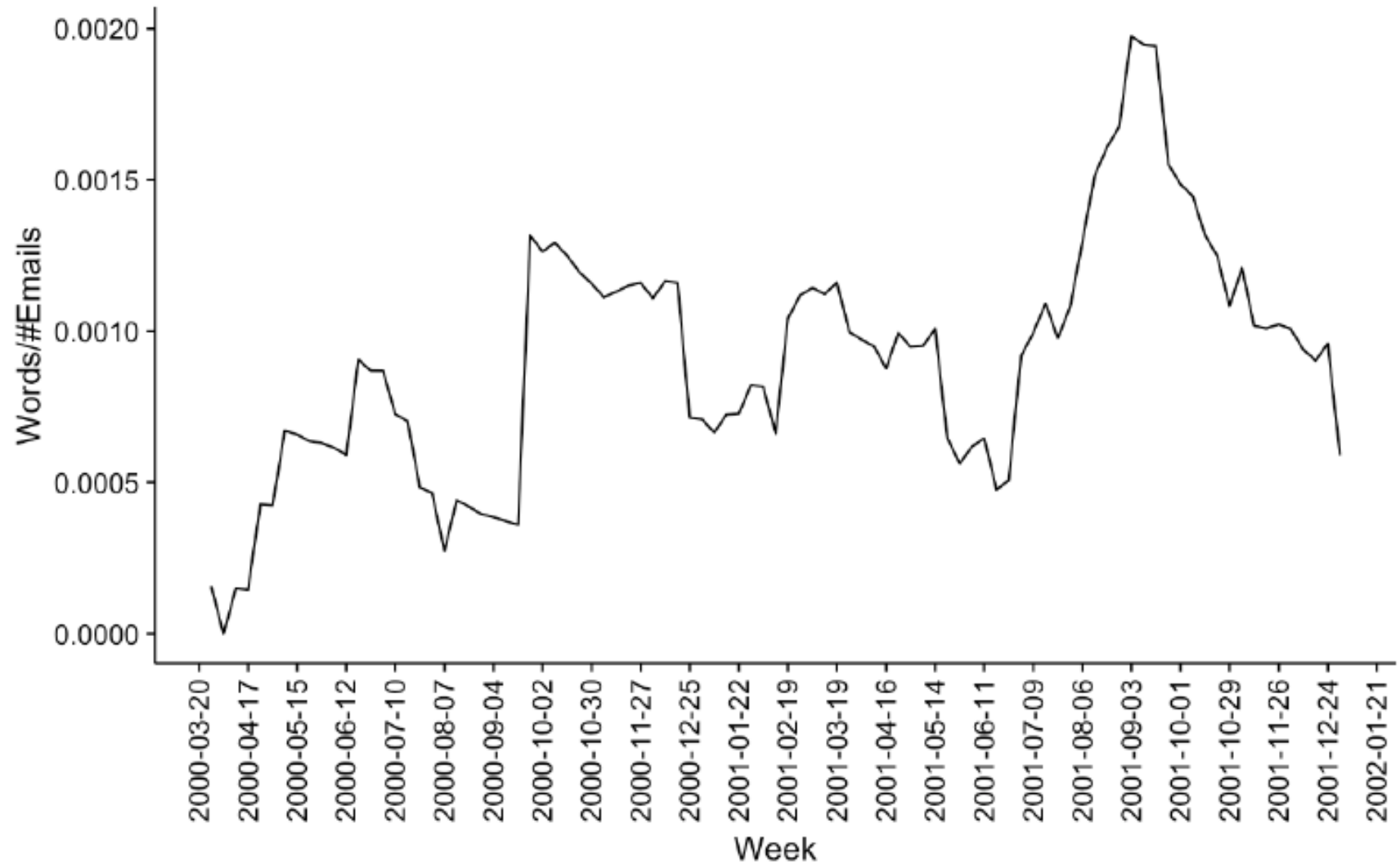


Figure 14. Topic Analysis over Time

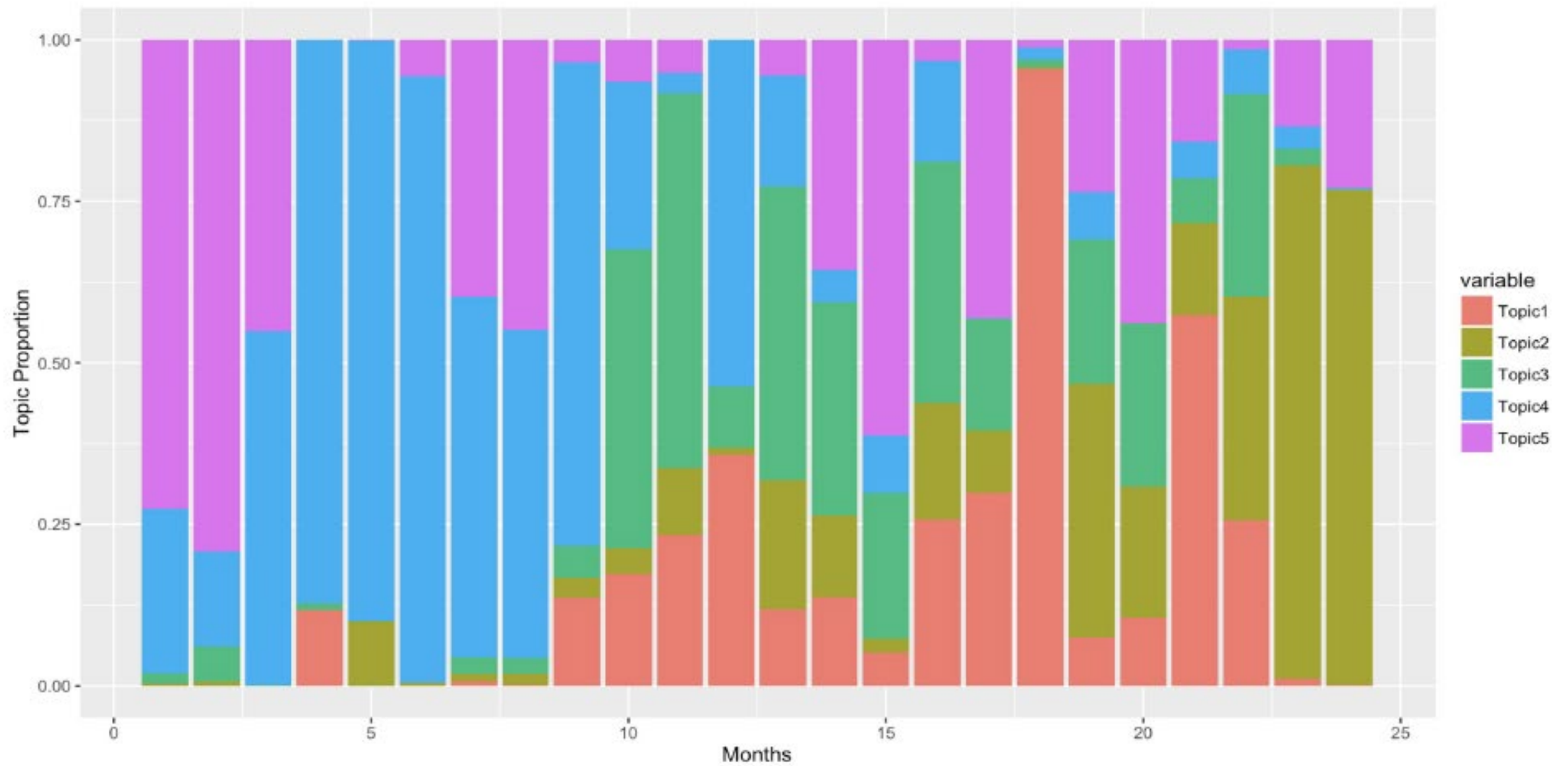
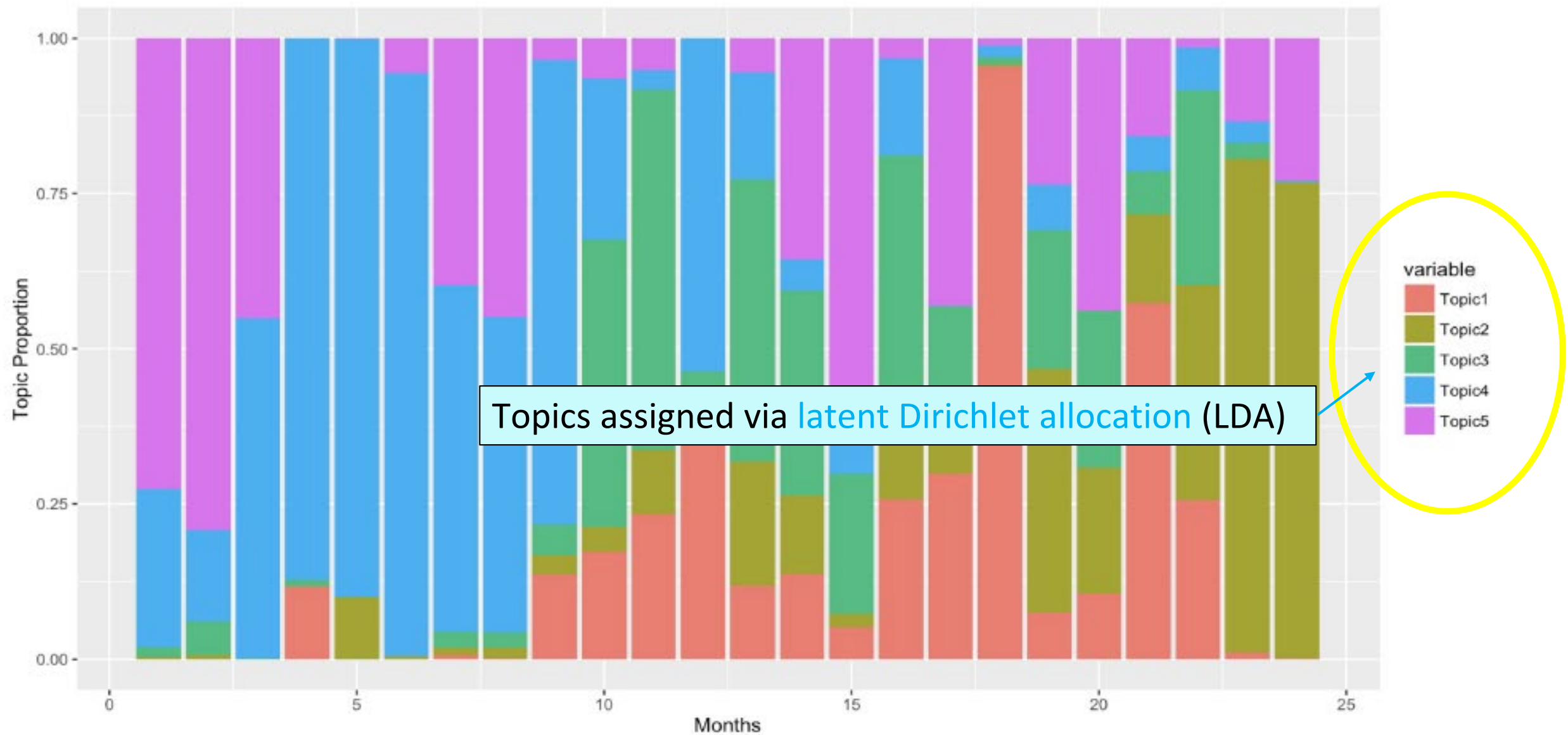


Figure 14. Topic Analysis over Time



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.

FDP EXAM TOPICS



1. Introduction to Data Science & Big Data

2. DM & ML: Introduction

3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models

4. DM & ML: Classification & Clustering

5. DM & ML: Performance Evaluation, Backtesting & False Discoveries

6. DM & ML: Representing & Mining Text

7. Big Data, DM & ML: Ethical & Privacy Issues

8. Big Data and Machine Learning in the Financial Industry

Reading(s):

- Guida, T. (2019). *Big Data and Machine Learning in Quantitative Investments*. West Sussex, UK: John Wiley & Sons Ltd. Chapter 10.
- Das, S., S. Kim and B. Kothari. (2019). **Zero-Revelation RegTech: Detecting Risk through Linguistic Analysis of Corporate Emails and News.** *The Journal of Financial Data Science*, 1(2), 8-34. DOI: <https://doi.org/10.3905/jfds.2019.1.2.008>

Sample Keywords:

Big data (p. 4),
Artificial intelligence (p. 4),
Natural language processing (p. 5),
Machine learning (ML) (p. 4),
Supervised learning (p. 5),
Unsupervised learning (p. 5),
Deep learning (p. 5),
Reinforcement learning (p. 5),
Sentiment indicators (p. 10),
Trading signals (p. 11),
Fraud detection (p. 11),
RegTech (p. 11),
InsurTech (p. 13),

Chatbots (p. 14),
Know your customer (KYC) (p. 20),
SupTech (p. 21),
Auditability (p. 33),
Fintech (p. 35),
Robo-advisors (p.35),
Tonality analysis (p.36)
Fintech (p. 79),
Robo-advisor (p. 80),
Work-flow (p. 83),
D2C platforms (p. 86),
Hybrid (p. 86),
B2B platforms (p. 86)...



FDP EXAM TOPICS



1. Introduction to Data Science & Big Data

2. DM & ML: Introduction

3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models

4. DM & ML: Classification & Clustering

5. DM & ML: Performance Evaluation, Backtesting & False Discoveries

6. DM & ML: Representing & Mining Text

7. Big Data, DM & ML: Ethical & Privacy Issues

8. Big Data and Machine Learning in the Financial Industry

Sample Learning Objectives:

8.10.1 Using linguistic analysis to perform risk analysis of investments.

A. Explain the difficulties associated with manual parsing of unstructured text.

B. Describe the concept of RegTech.

C. Describe how the content and structure of emails could be used for risk analysis.

...

Sample Question:

According to the article “Zero-Revelation RegTech: Detecting Risk through Linguistic Analysis of Corporate Emails and News,” what does the decomposition of the ‘document term matrix’ facilitate?

Topic sentiment*

Network activity

Vocabulary trends

Source. LO 8.10.1. p. 28-29



Q & A

Kind reminders of upcoming webinars as we go through the Q & A.
Add you questions in the chat room please.



WEBINAR SERIES
A Conversation With...

Joe Simonian

Co-editor, Journal of
Financial Data Science

February 27, 2020
1pm EST



WEBINAR SERIES
A Conversation With...

Tony Guida

Executive Director, Sr. Quant
Research
Ram Active Investment

March 5, 2020
10 am EST

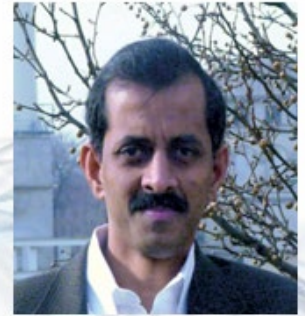


WEBINAR SERIES
A Conversation With...

Ganesh Mani

Adj. Faculty Carnegie Mellon
"Data Supply Chain Mgmt."

March 10, 2020
1pm EST



WEBINAR SERIES
A Conversation With...

Rick Roche, CAIA

Man. Dir. Little Harbor Advisors

March 17, 2020
11 am EST



WEBINAR SERIES
A Conversation With...



Mike Chen, Ph.D.
PanAgora

George Mussalli, CFA
PanAgora



**"An integrative Approach to
Quantitative ESG Investing"**

Date to be
determined

In Closing

- The Next FDP Exam: March 16 –April 4, 2020
- Registration is open
- For a recent candidate webinar go to www.fdpinstitute.org/webinars

Learn more about the FDP Institute at
www.fdpinstitute.org