Zero-Revelation RegTech: Detecting Risk through Corporate Emails

Sanjiv R. Das
Santa Clara University

Joint work with Seoyoung Kim (SCU) & Bhushan Kothari (Google)

Hyderabad
February 2018

Big Picture

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

Our Purpose

1. Our purpose is to explore the predictive power of information conveyed by employee emails.
2. Specifically, we are interested in:
   - The sentiment conveyed by email content.
   - The information conveyed by structural characteristics, such as email volume or length.
   - Other nonverbal 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.

The Enron Email Corpus

1. 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.
2. 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. Skillings 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.

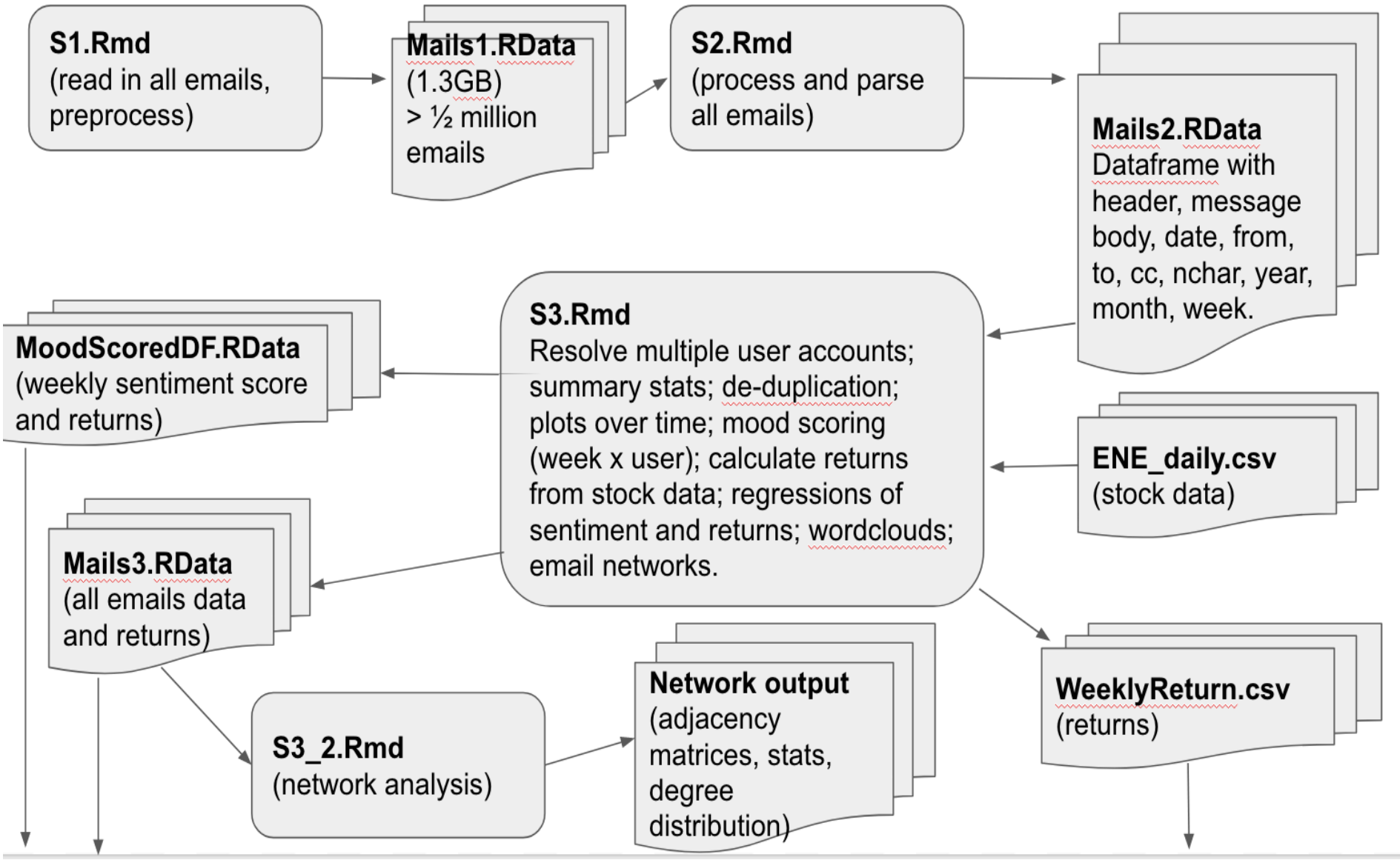
Curing the Data

- We focus on "sent" emails (rather than all emails) in order to ...
  1. Analyze content specifically written by Enron employees.
  2. 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:
  1. Emails greater than 3,000 characters in length.
  2. 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.

Program Flow - Part 1



Program Flow - Part 2

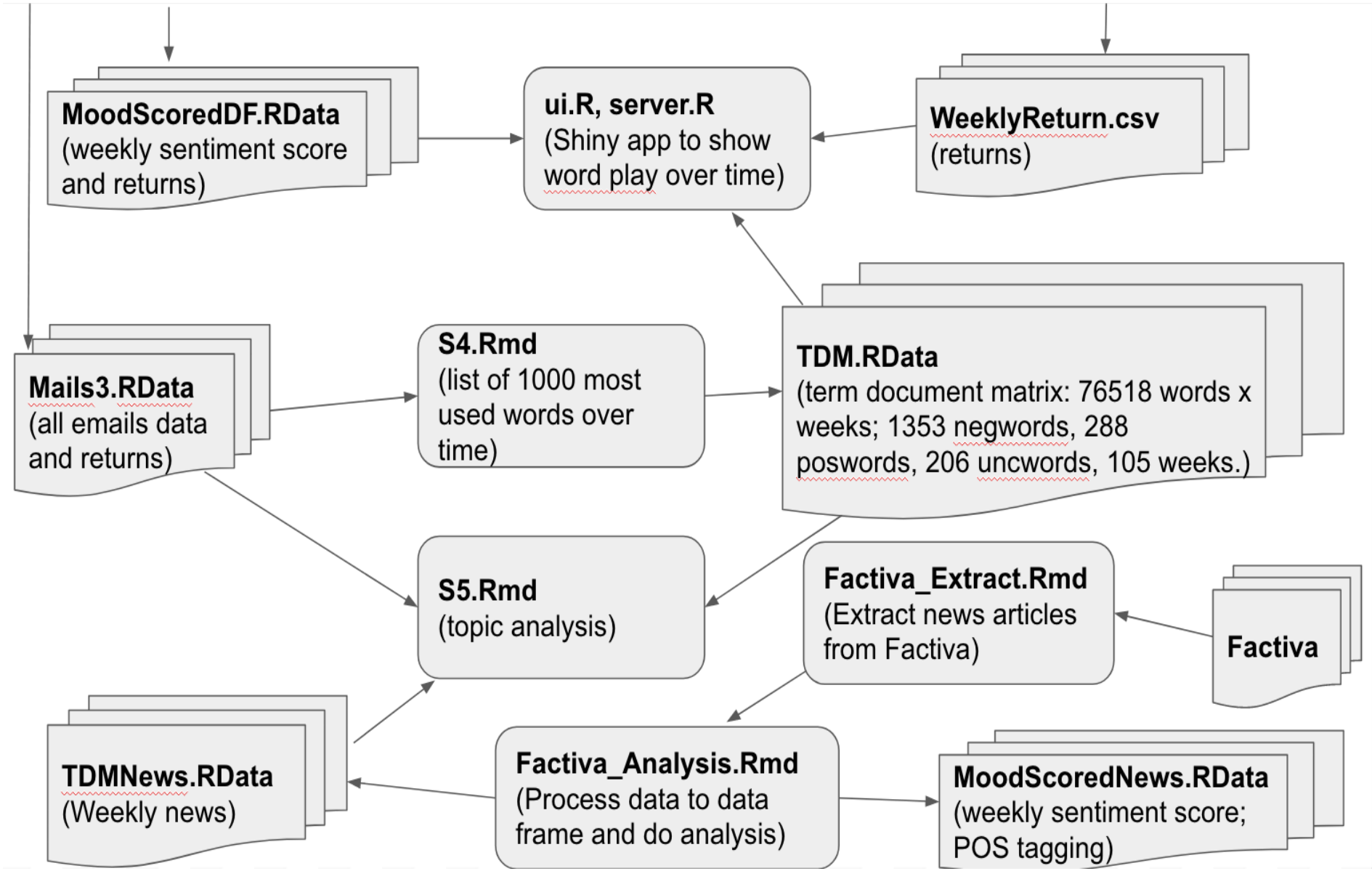


Table 1. Summary Statistics of Sent Mail

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

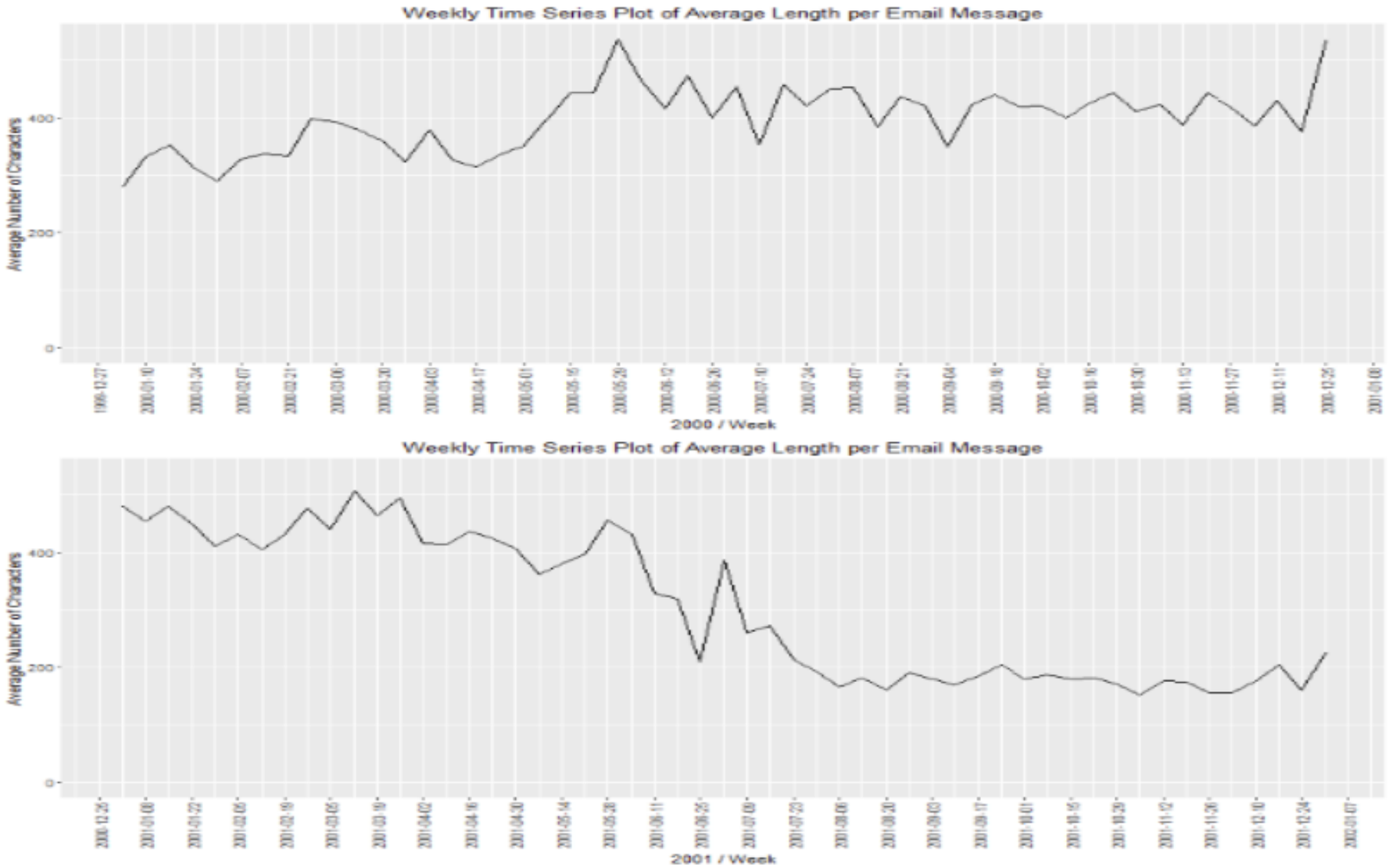


Figure 1. Average Email Length over Time

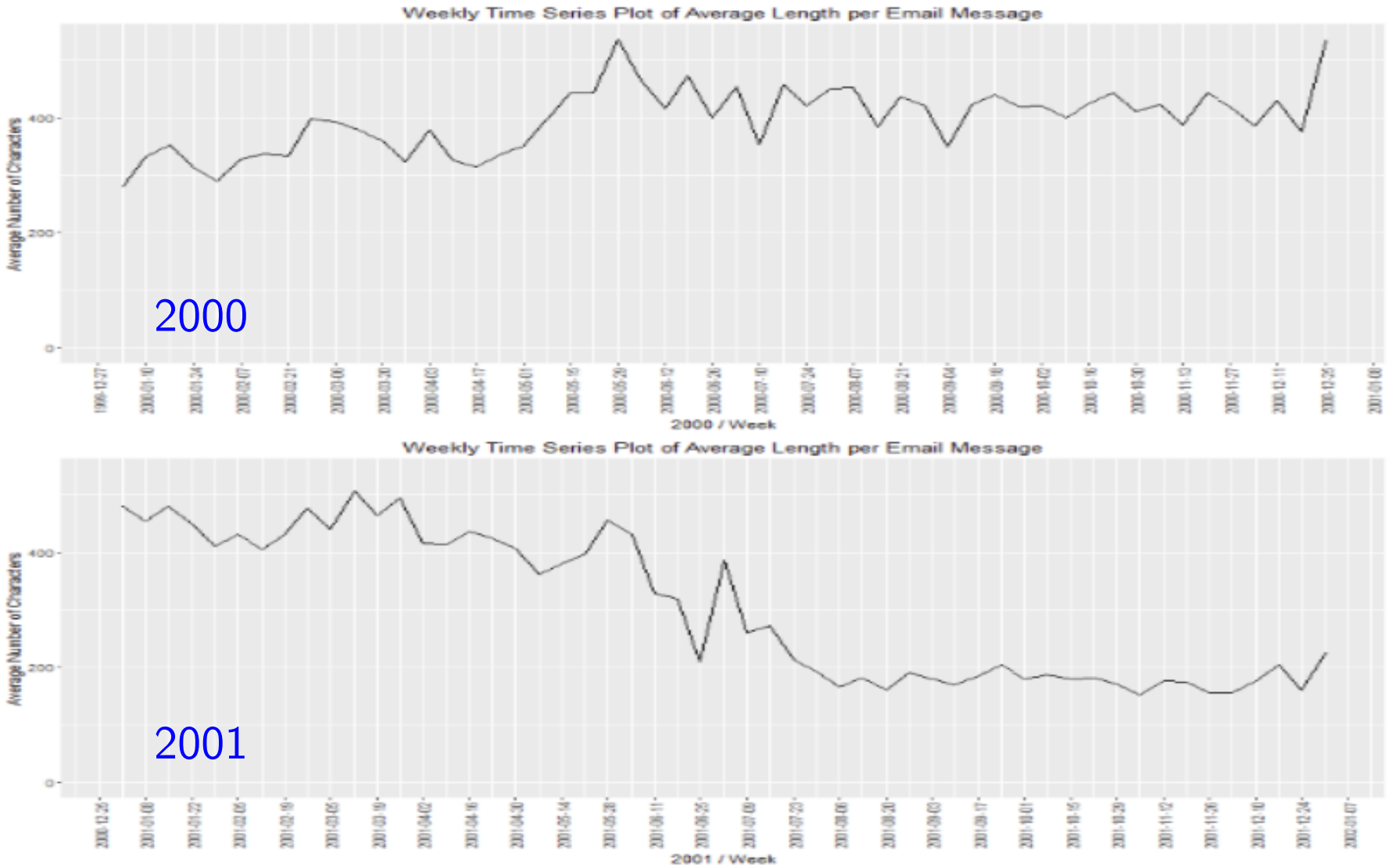


Figure 1. Average Email Length over Time

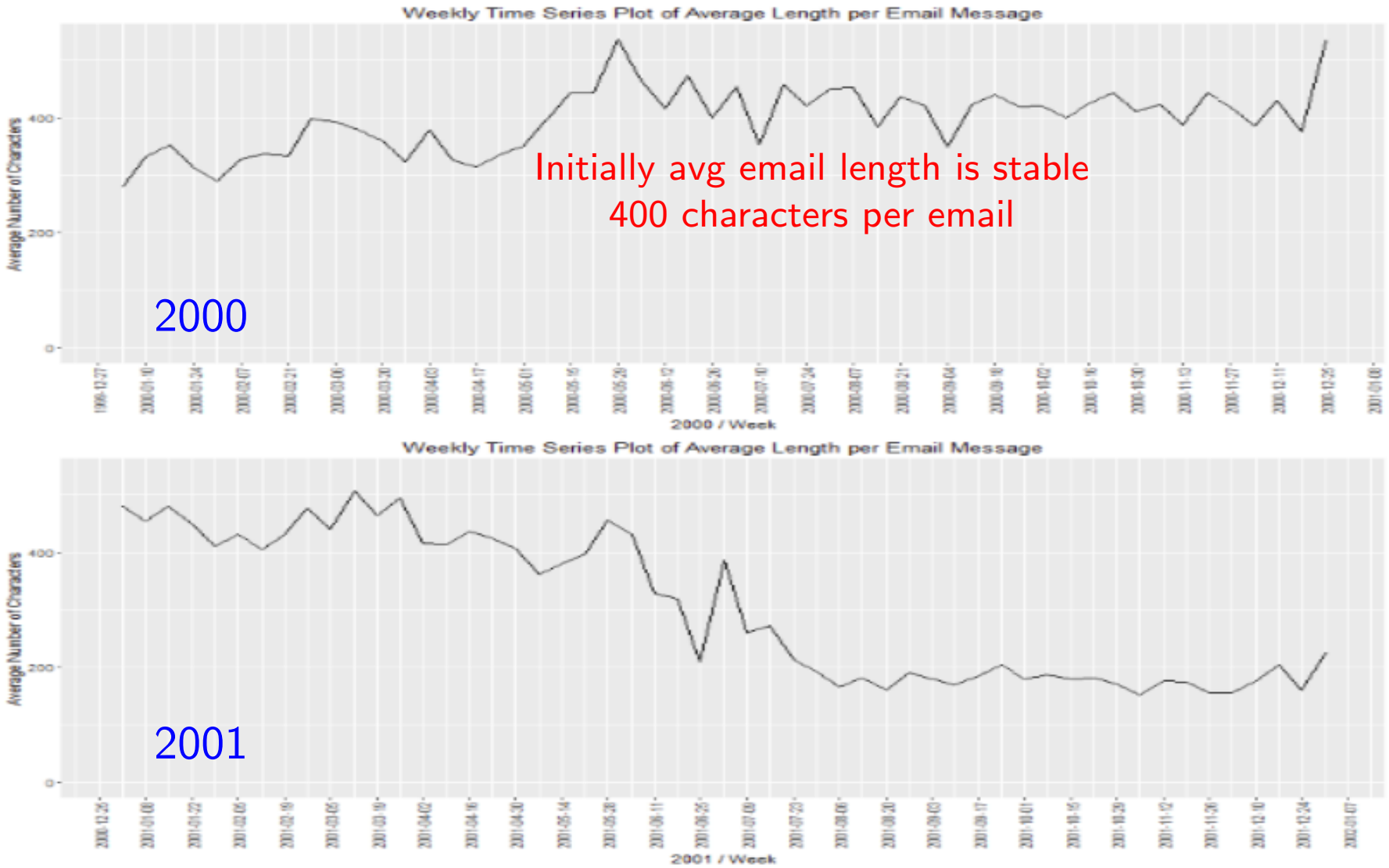


Figure 1. Average Email Length over Time

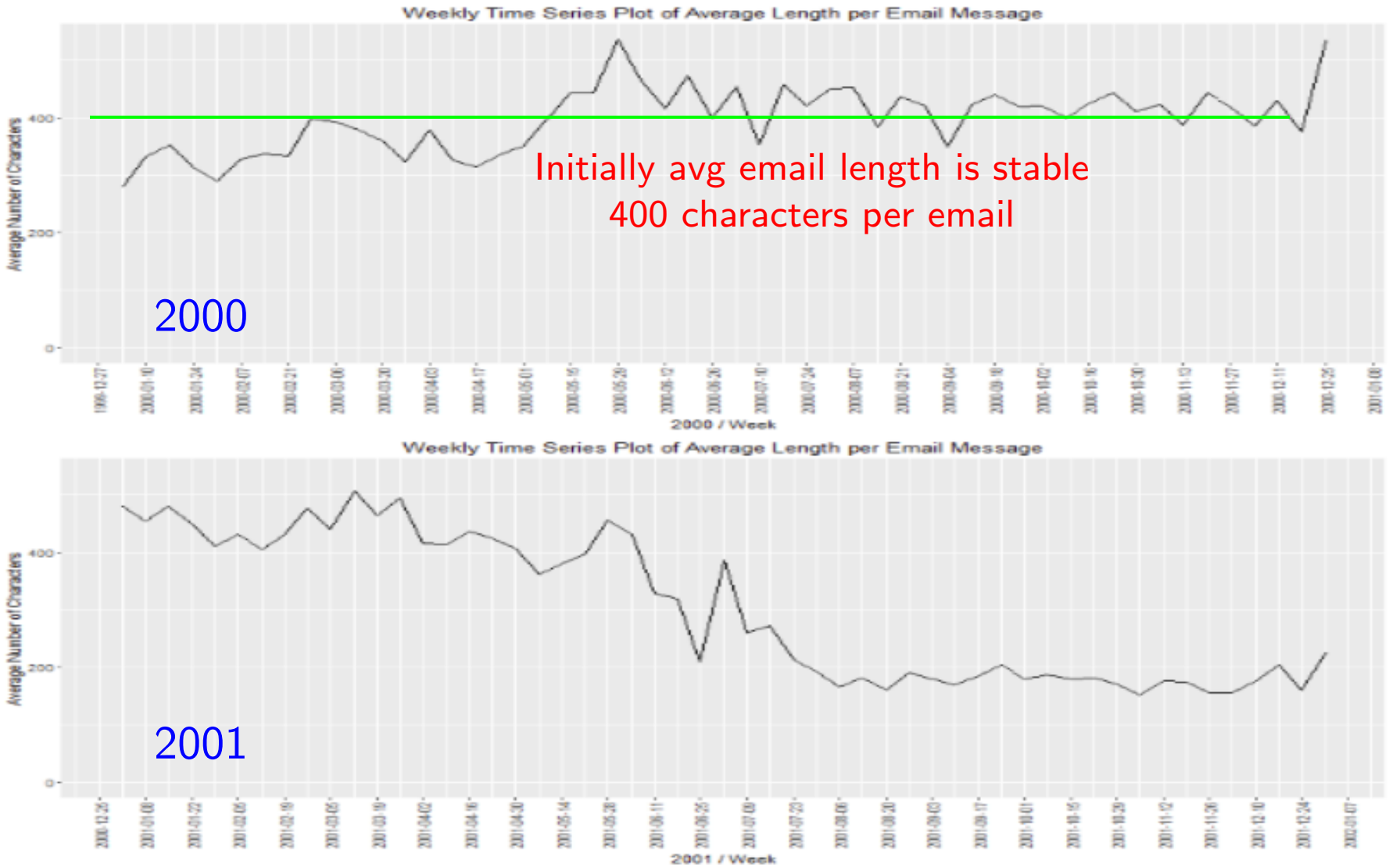


Figure 1. Average Email Length over Time

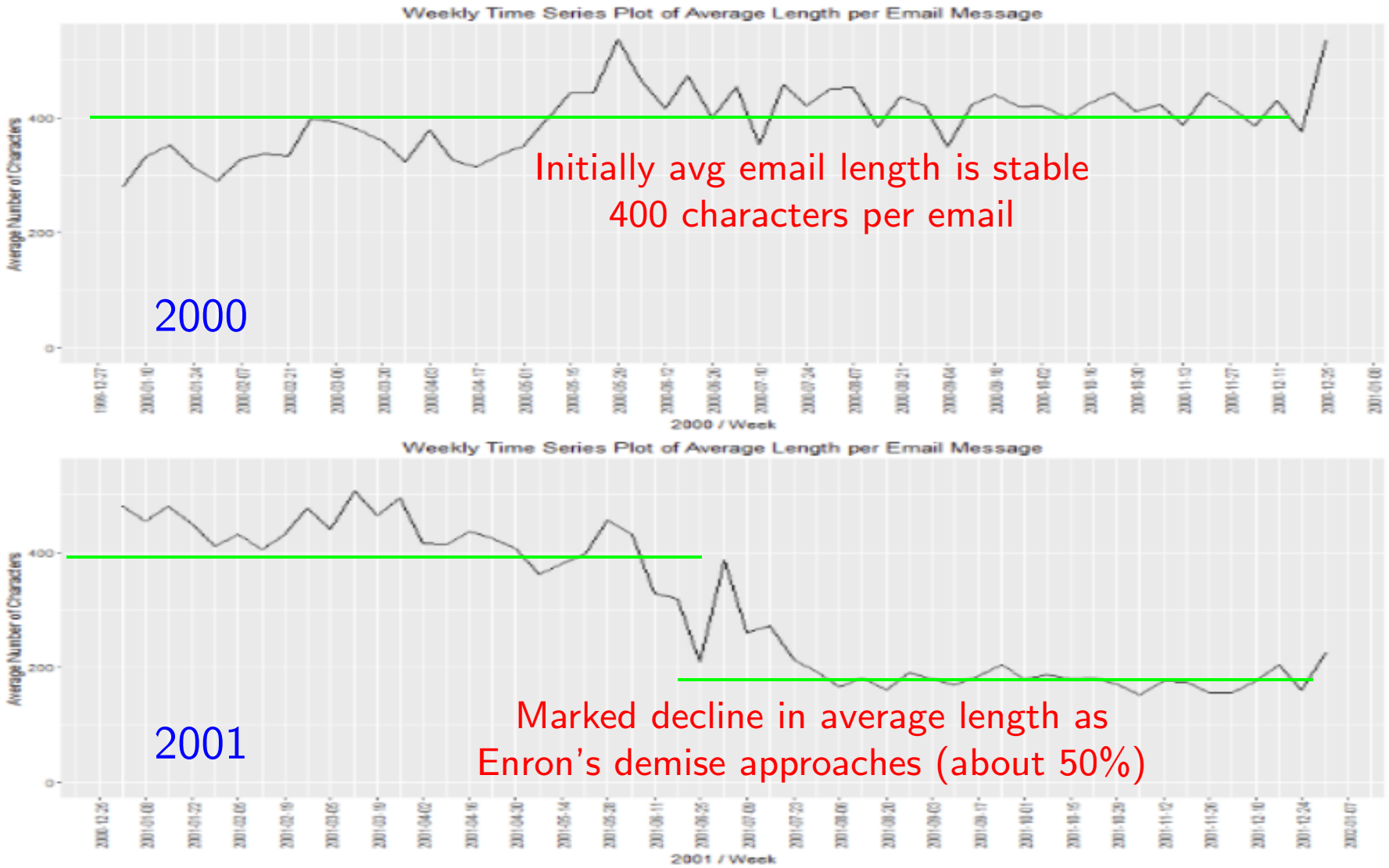


Figure 2. Email Sentiment and Disagreement over Time

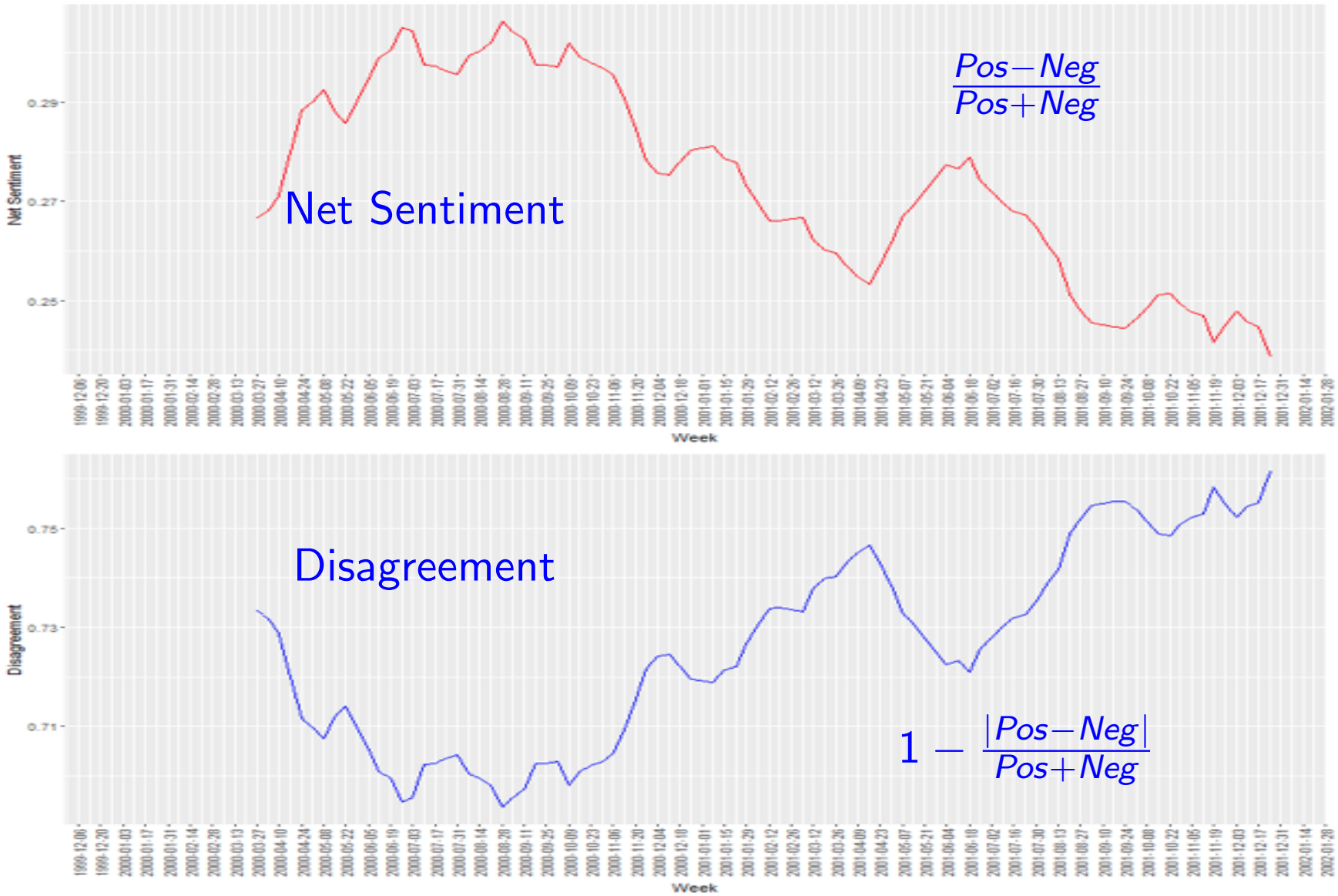


Figure 2. Email Sentiment and Disagreement over Time

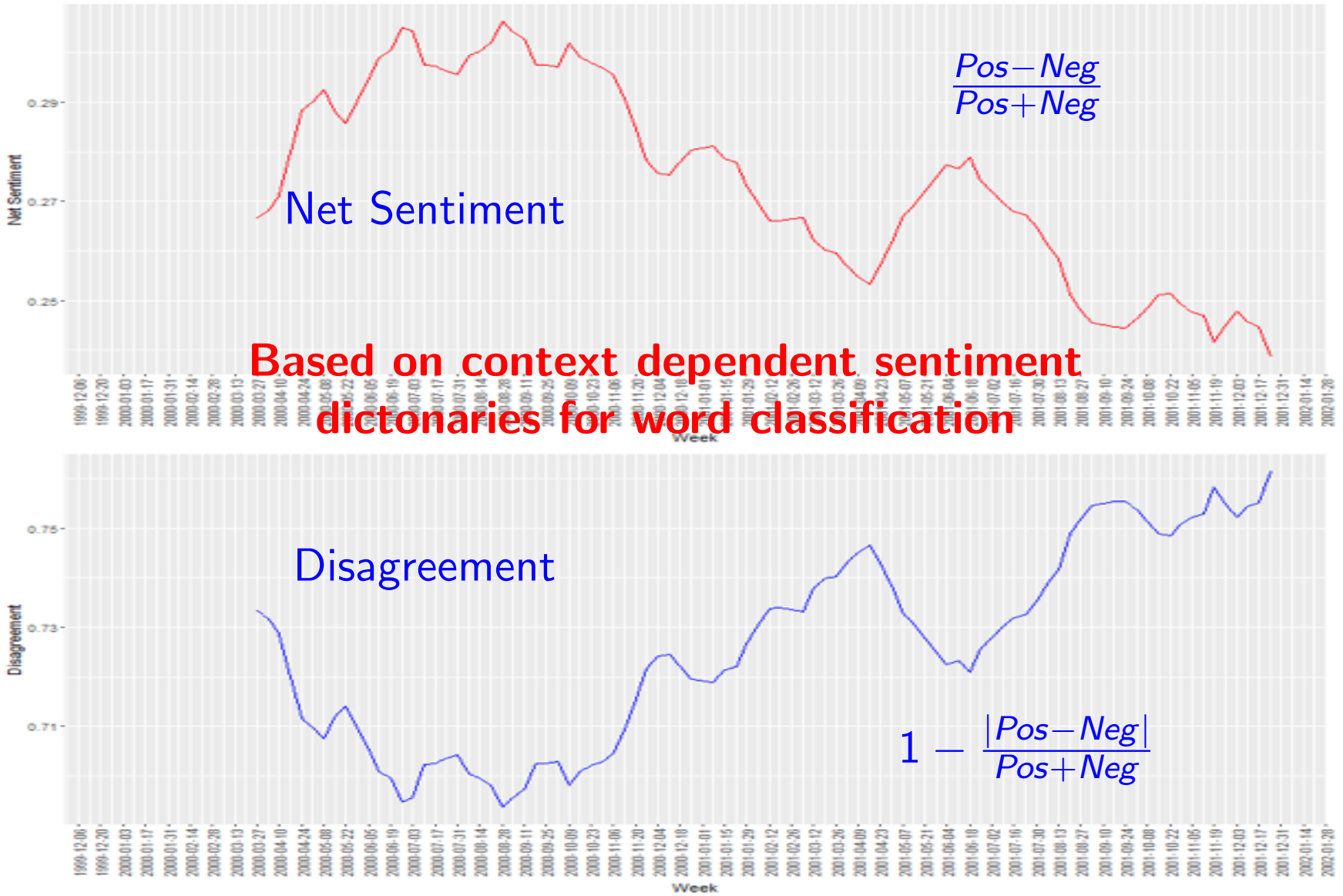


Figure 3. Factiva News Coverage over Time

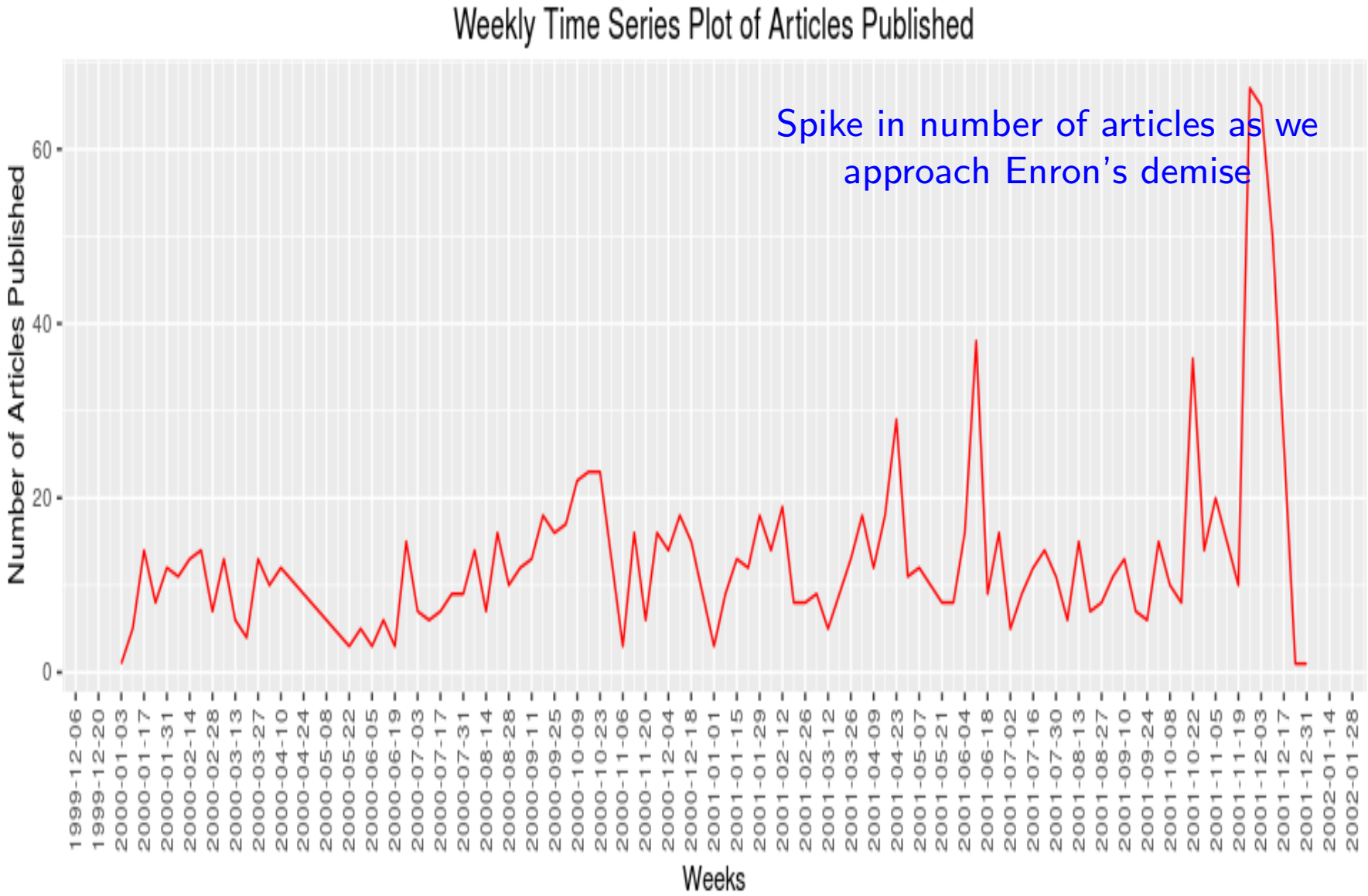


Figure 4. Factiva News Sentiment over Time

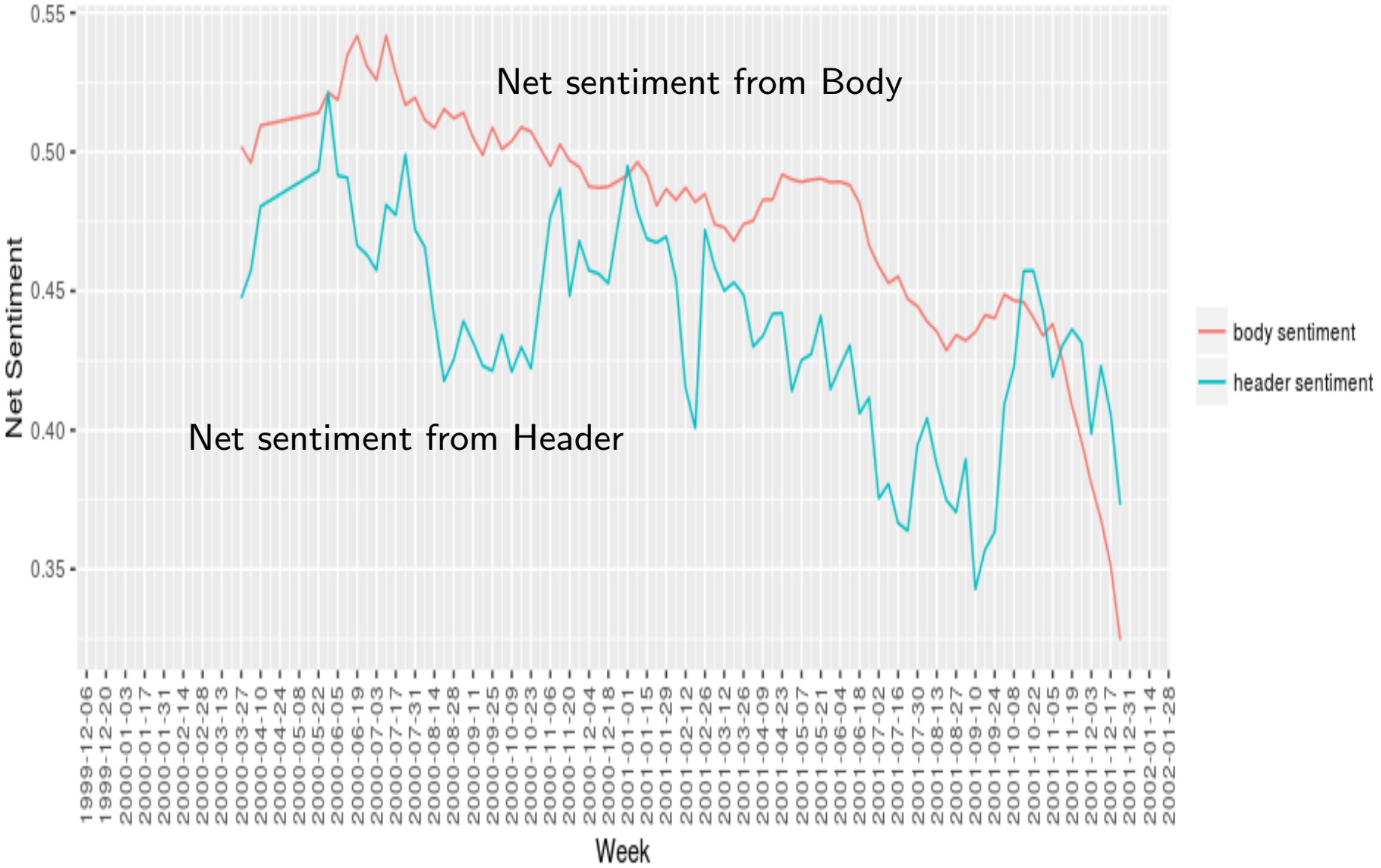


Figure 5. Stock Returns and Net Sentiment

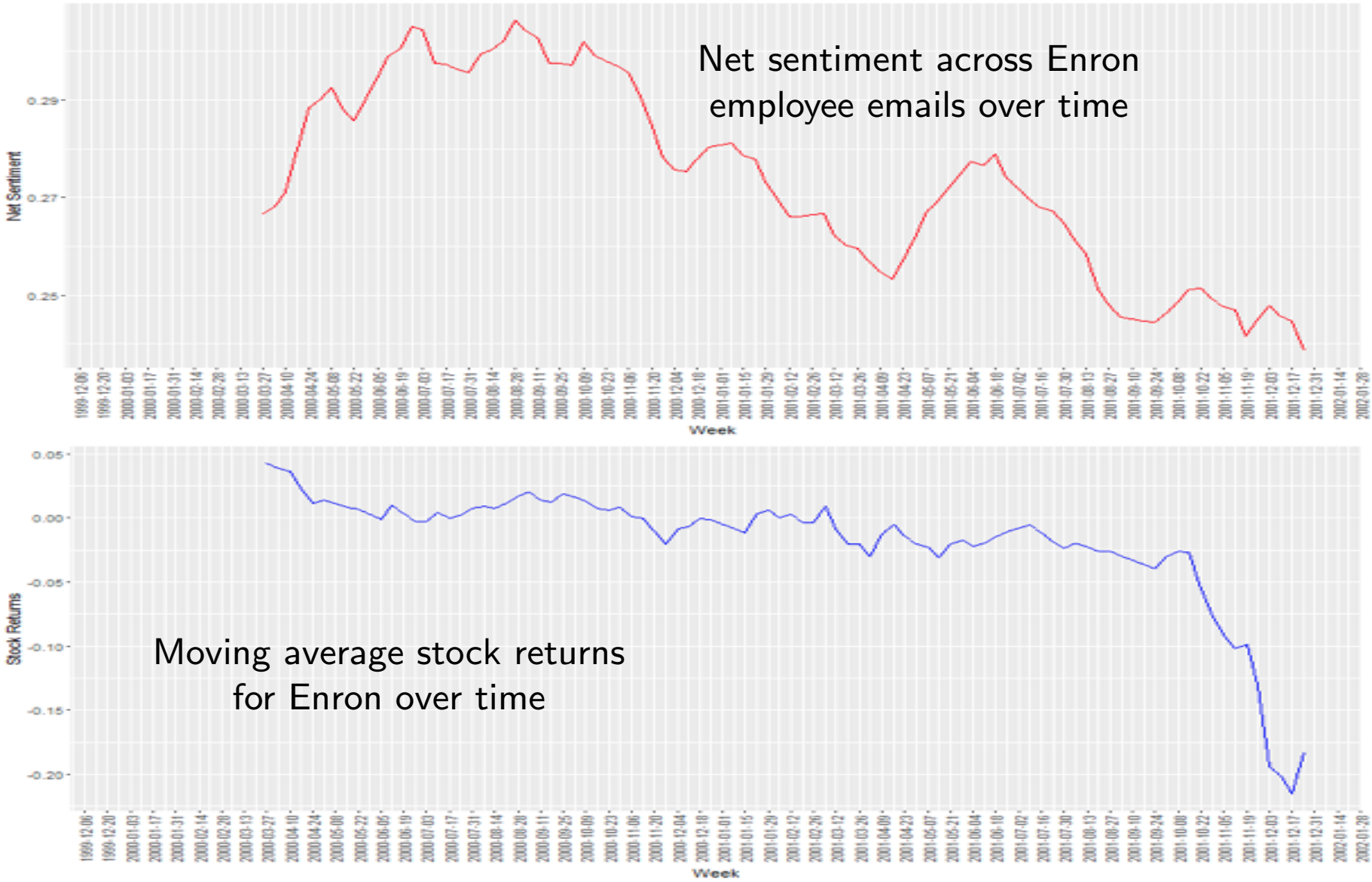


Figure 6. Stock Prices and Net Sentiment

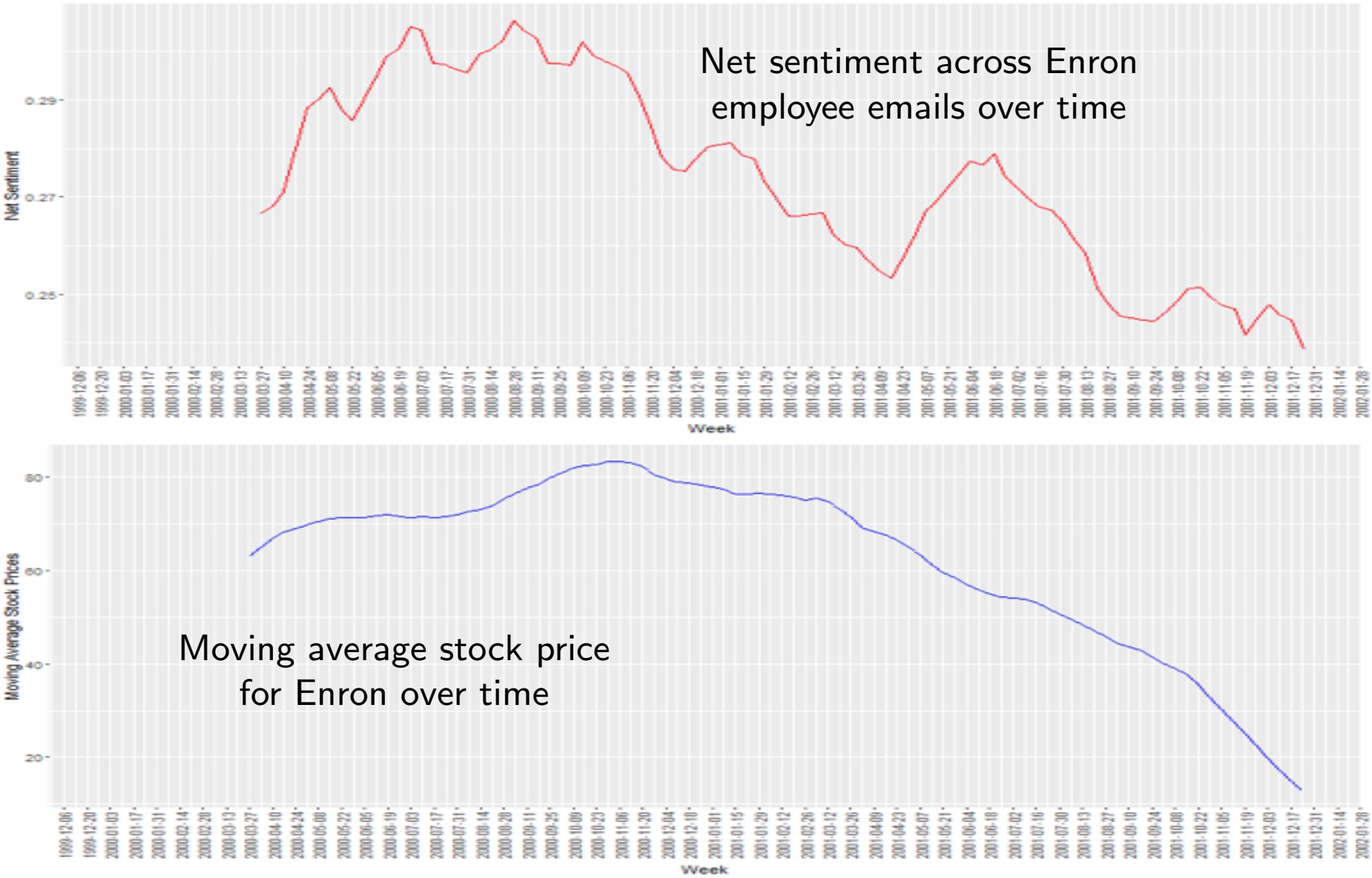


Figure 7. Stock Returns and Email Length

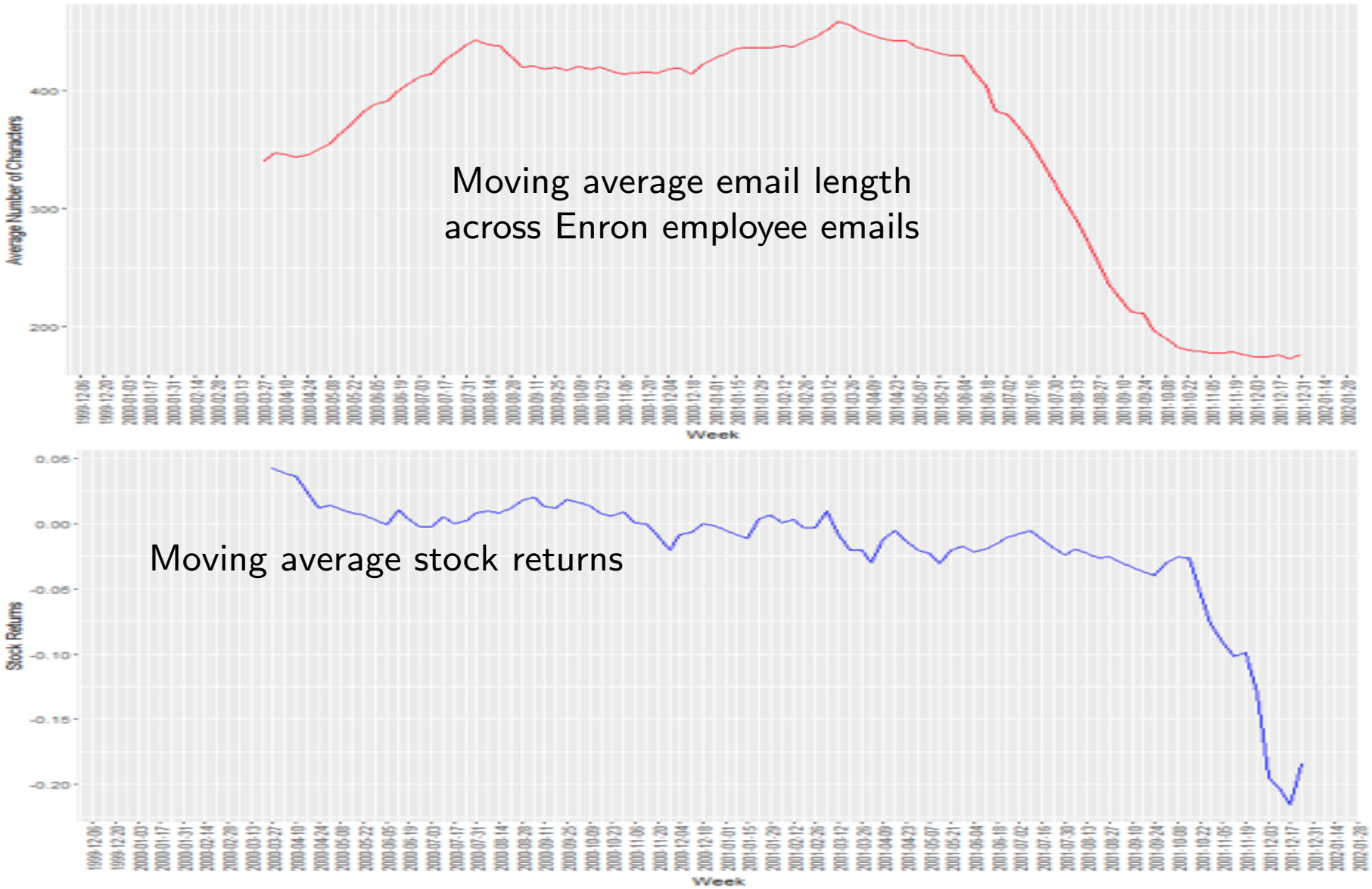


Figure 8. Stock Prices and Email Length

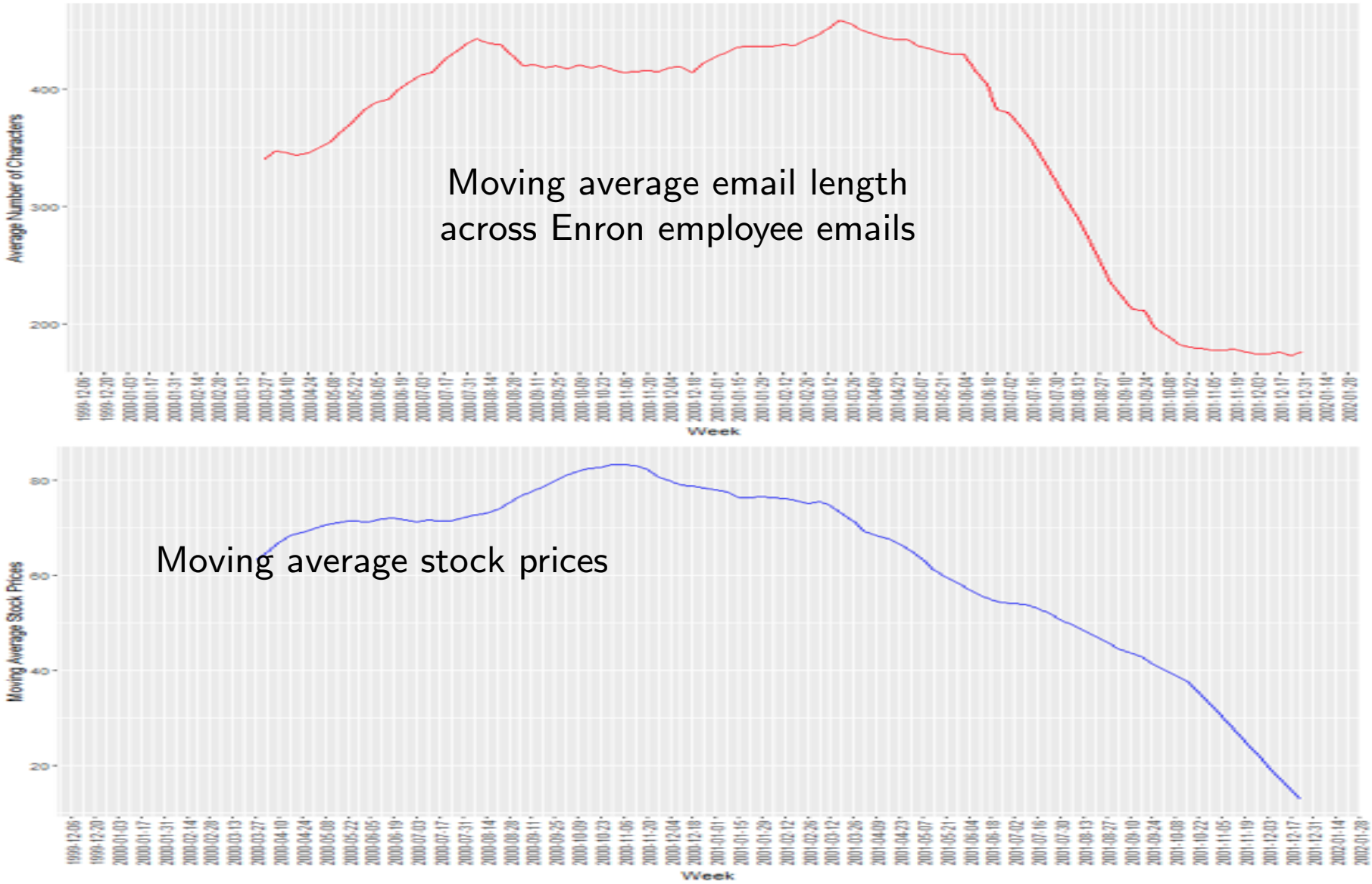


Figure 8. Stock Prices and Email Length

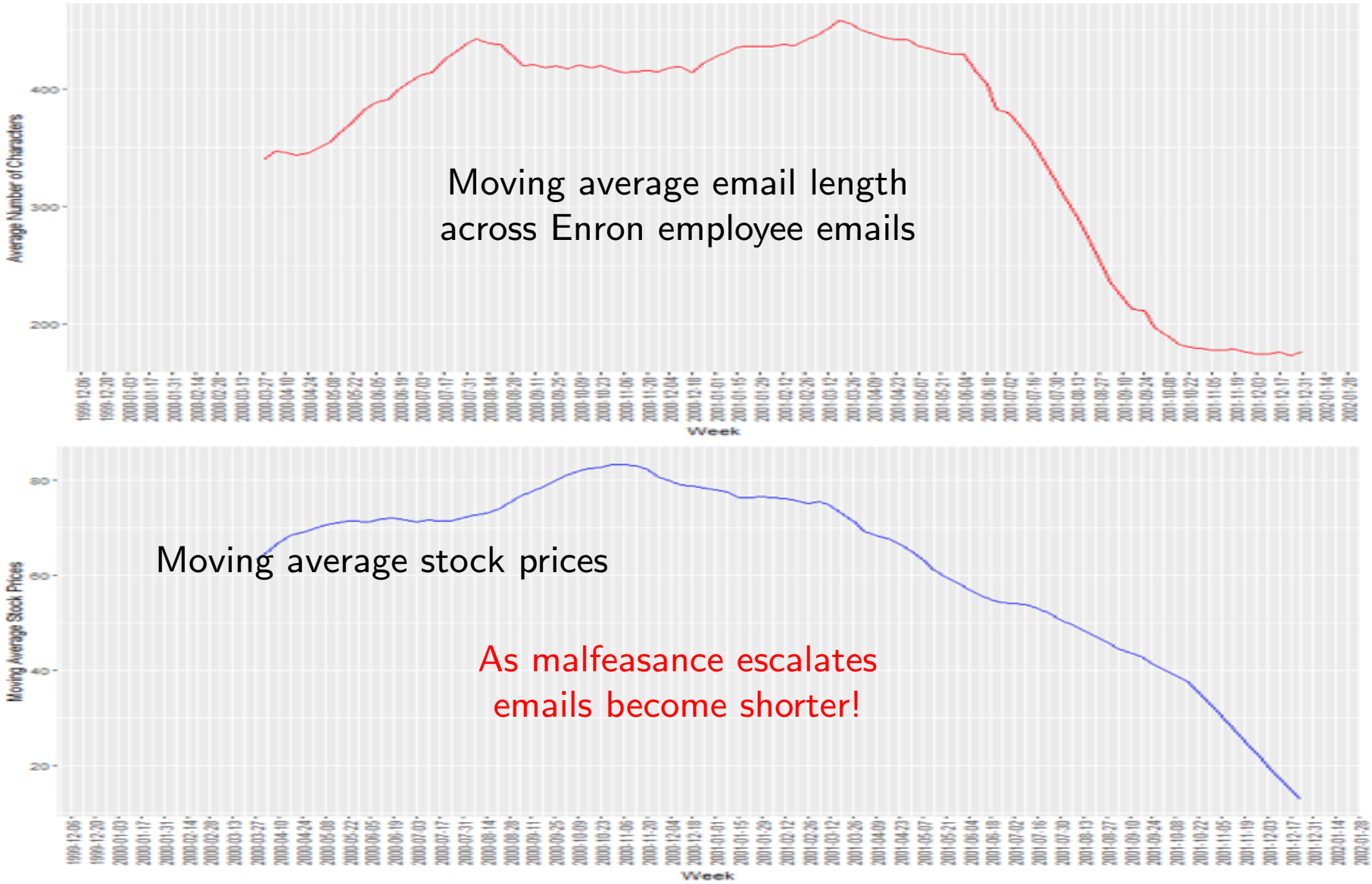


Table 2. Email content and Stock Returns

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

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 vs 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 vs 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 vs 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 vs 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 vs 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 vs 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.

Figure 9. Correlation between Email Sentiment and News Body Sentiment

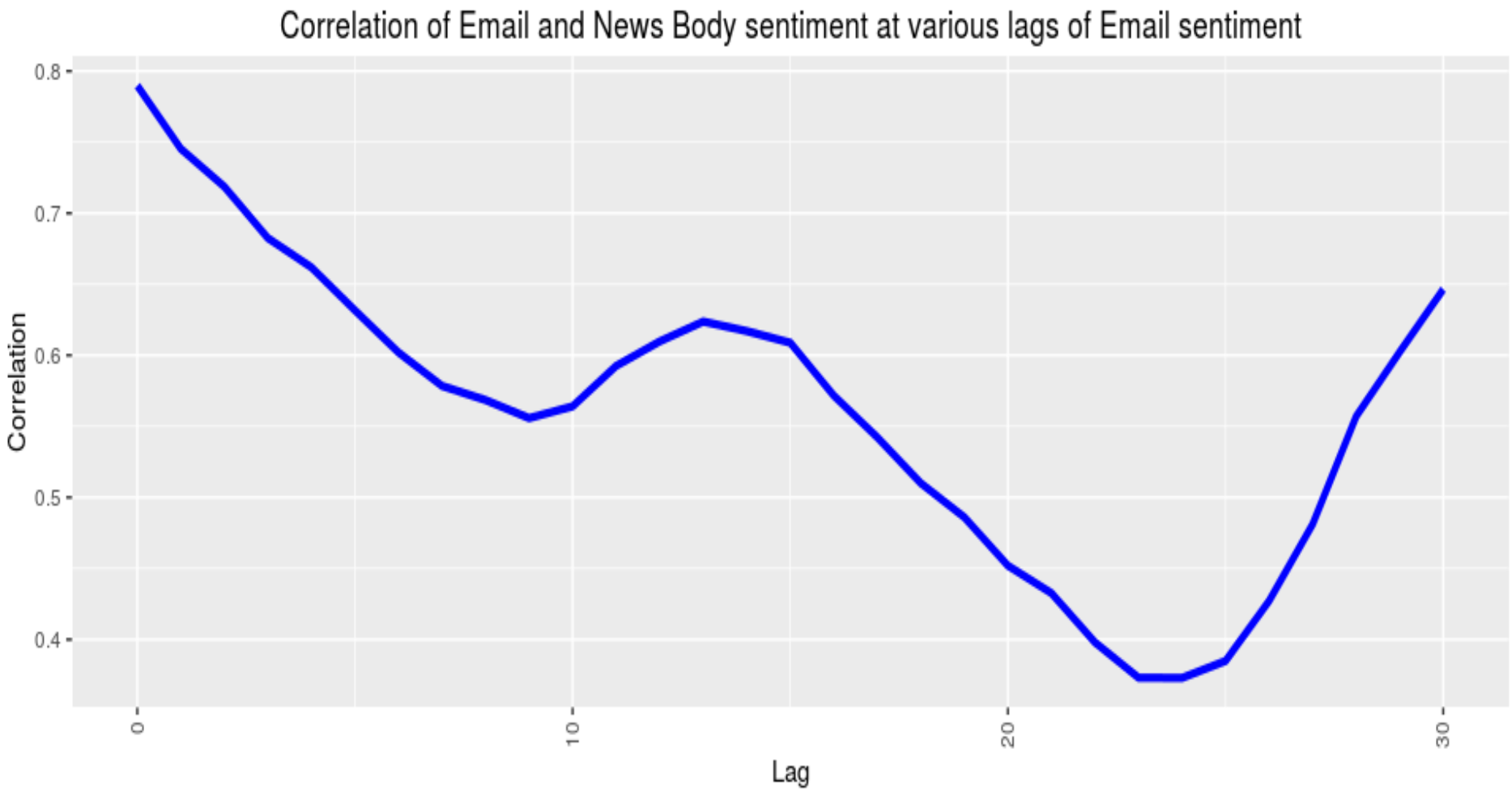
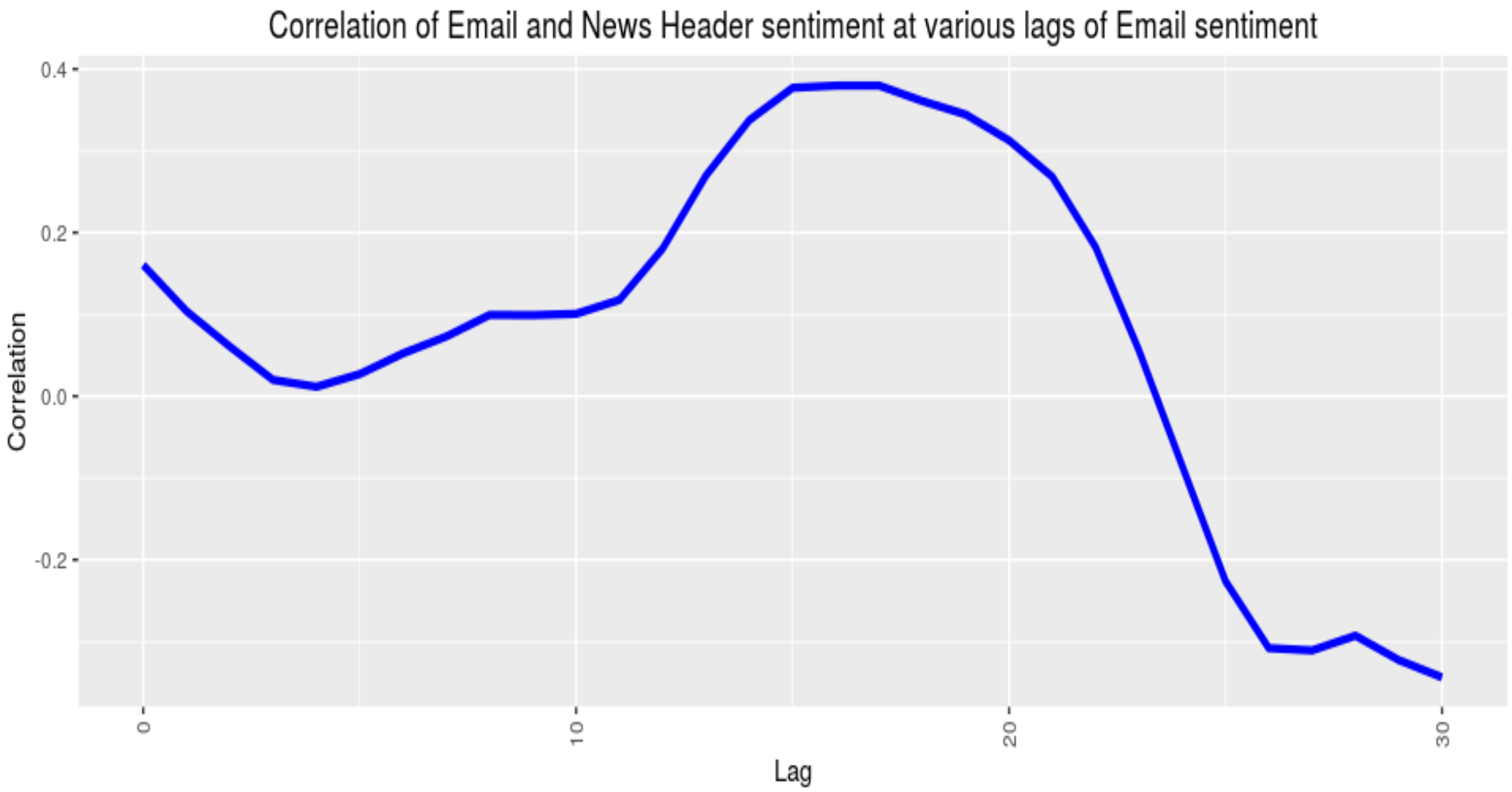


Figure 10. Correlation between Email Sentiment and News Header Sentiment



Summary so far ...

- 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 nonverbal, interaction- or network-based indicators of potential trouble.

Figure 11. Email network 2000-Q4

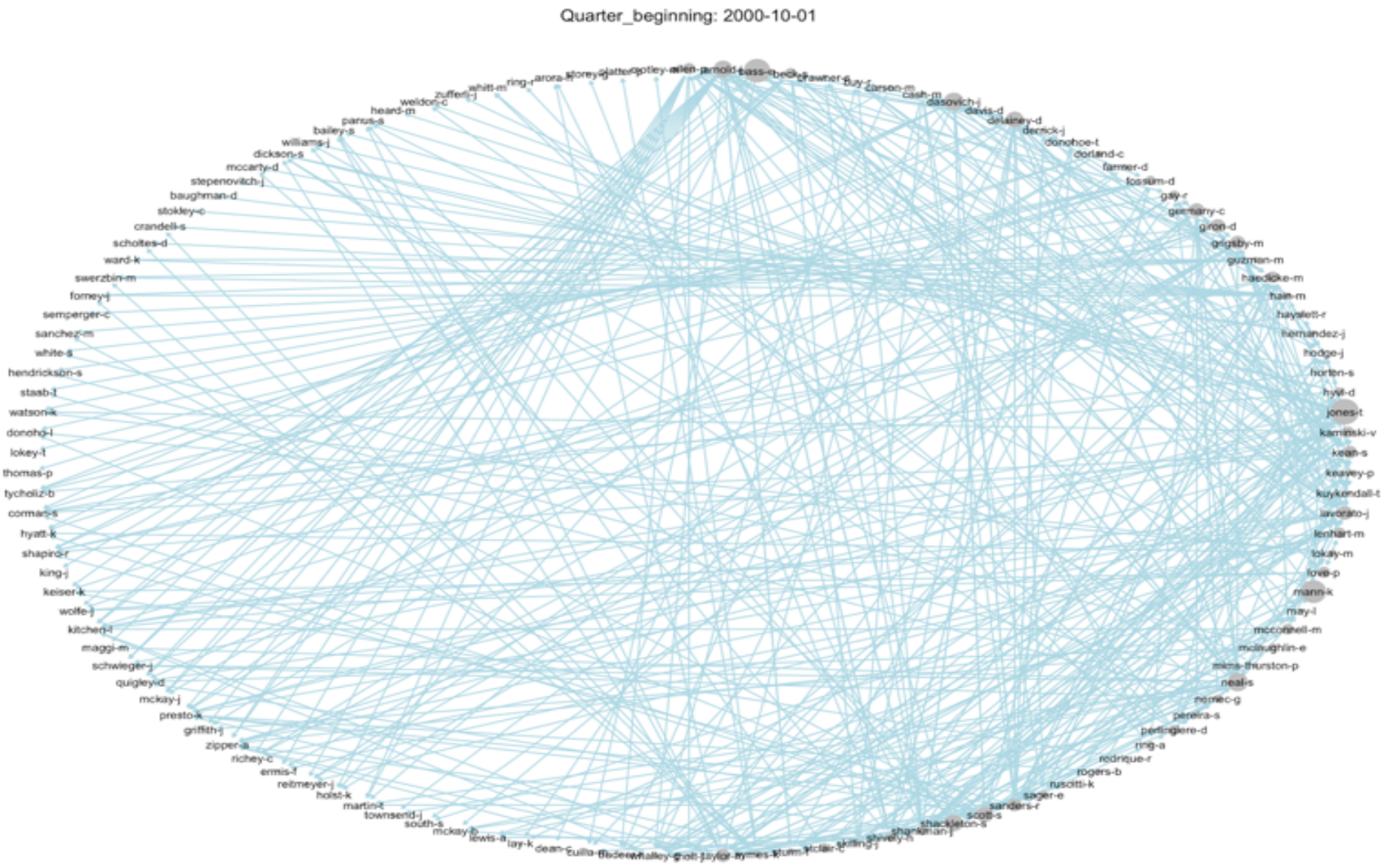


Figure 11. Email network 2001-Q4

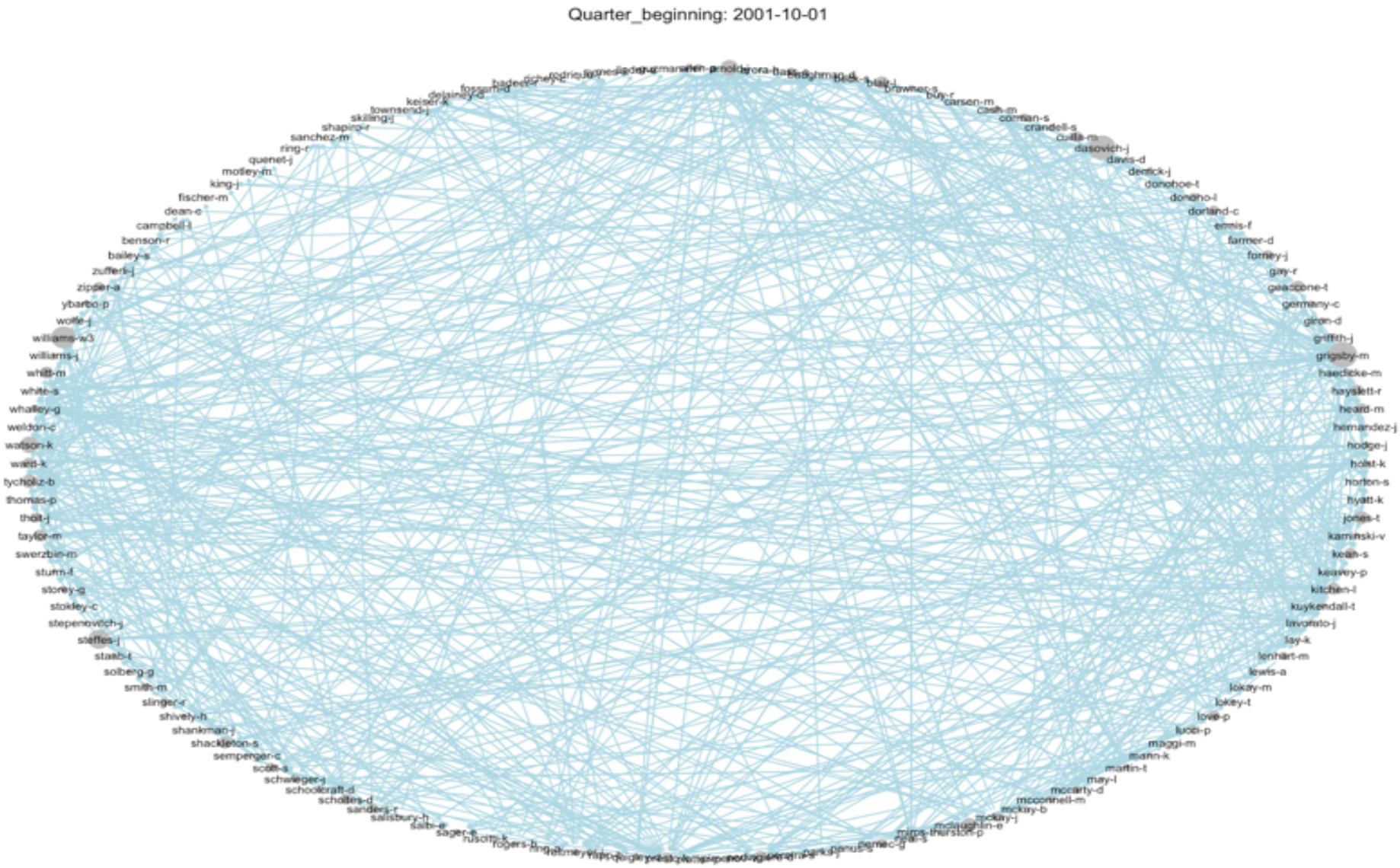


Figure 12. Degree Distribution 2000-Q4

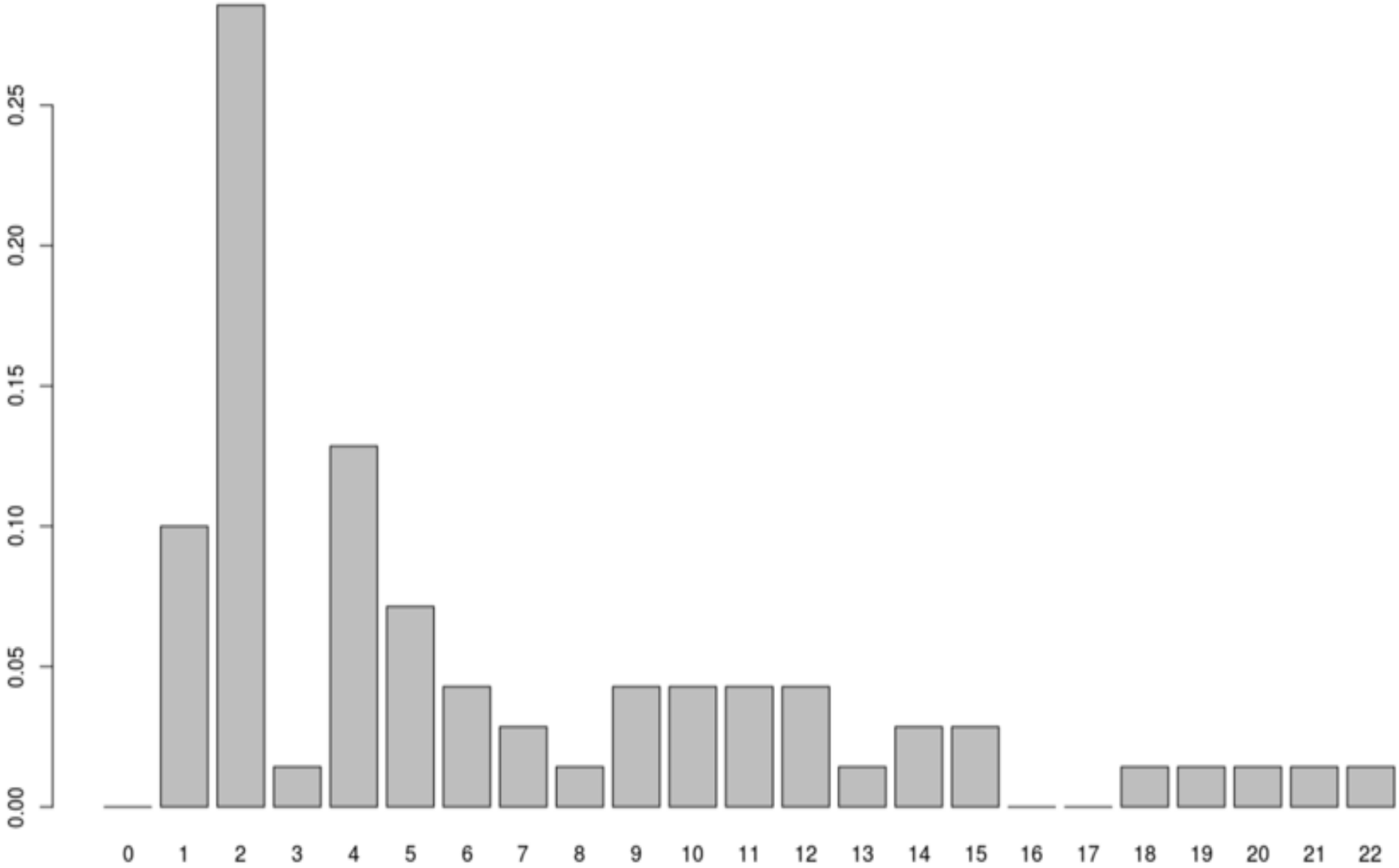


Figure 12. Degree Distribution 2001-Q4

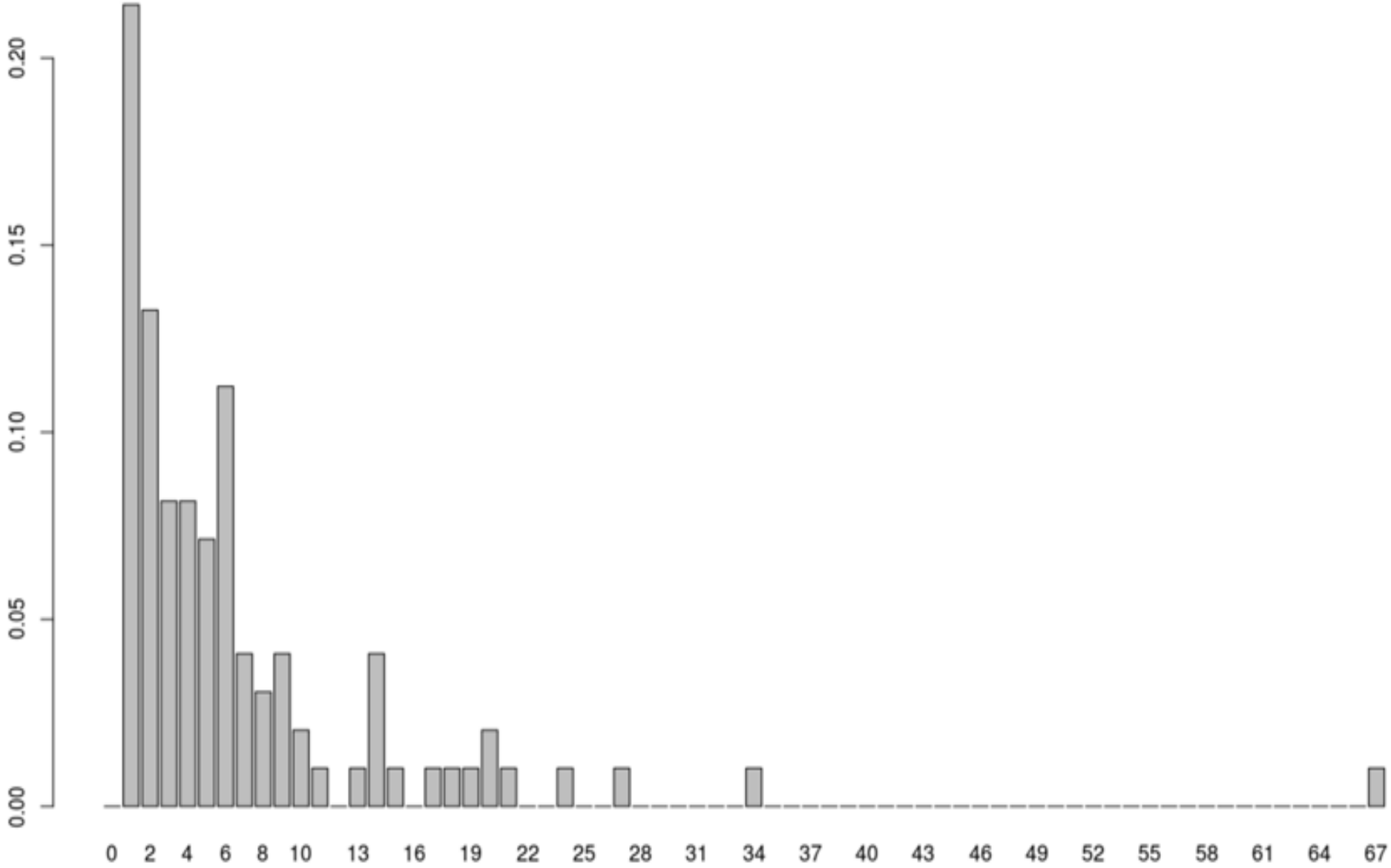


Figure 13. Vocabulary Trend Plots

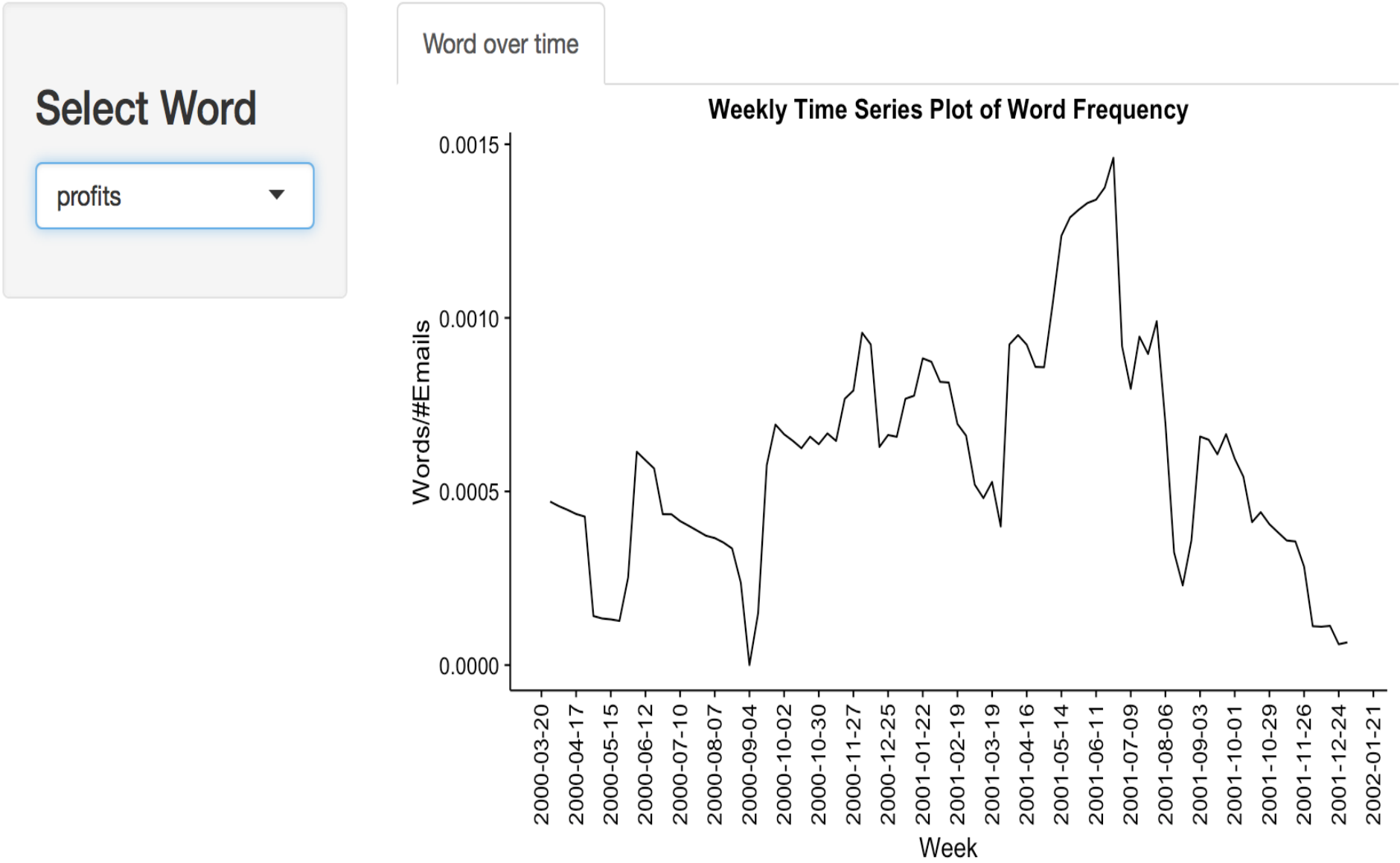
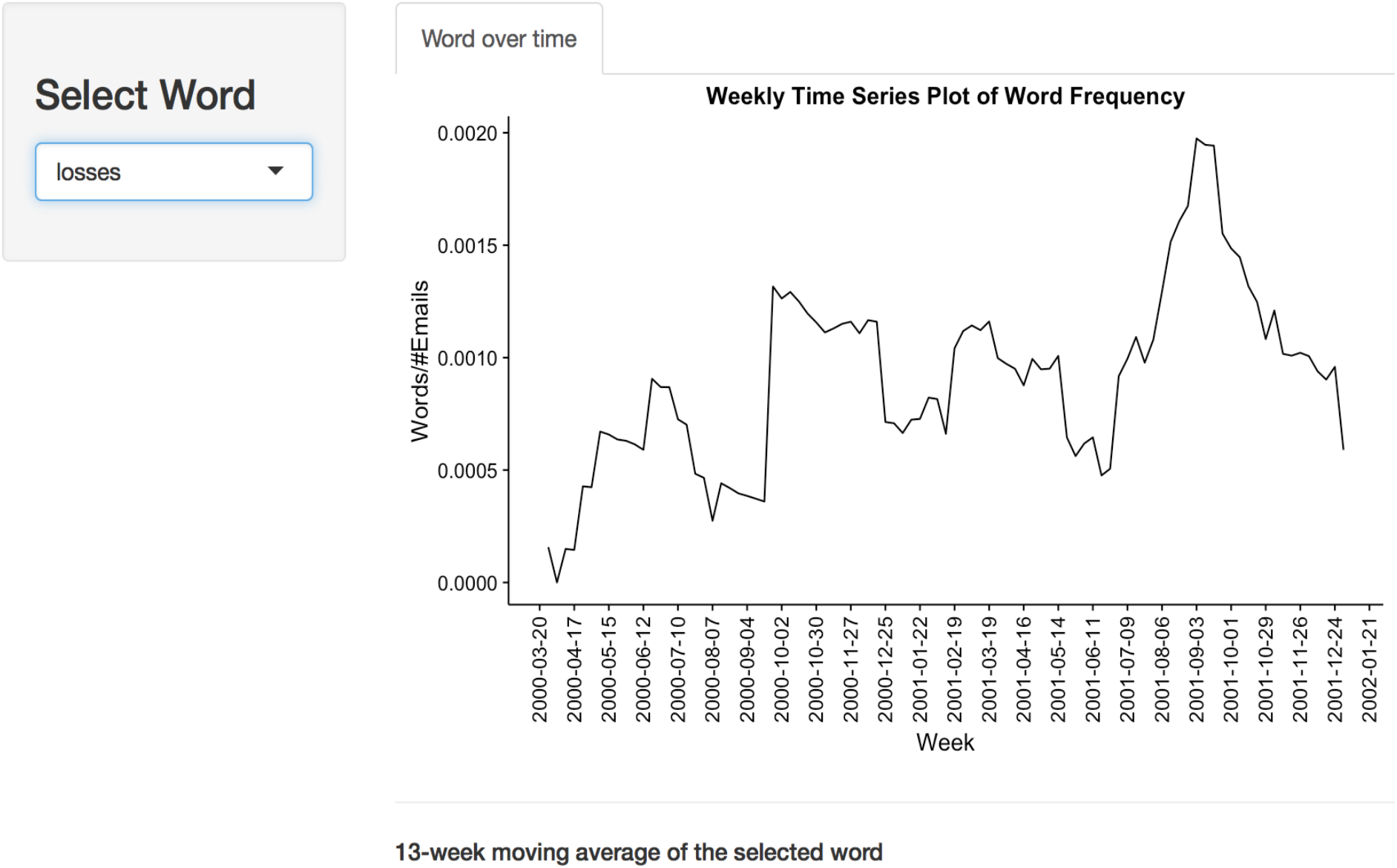
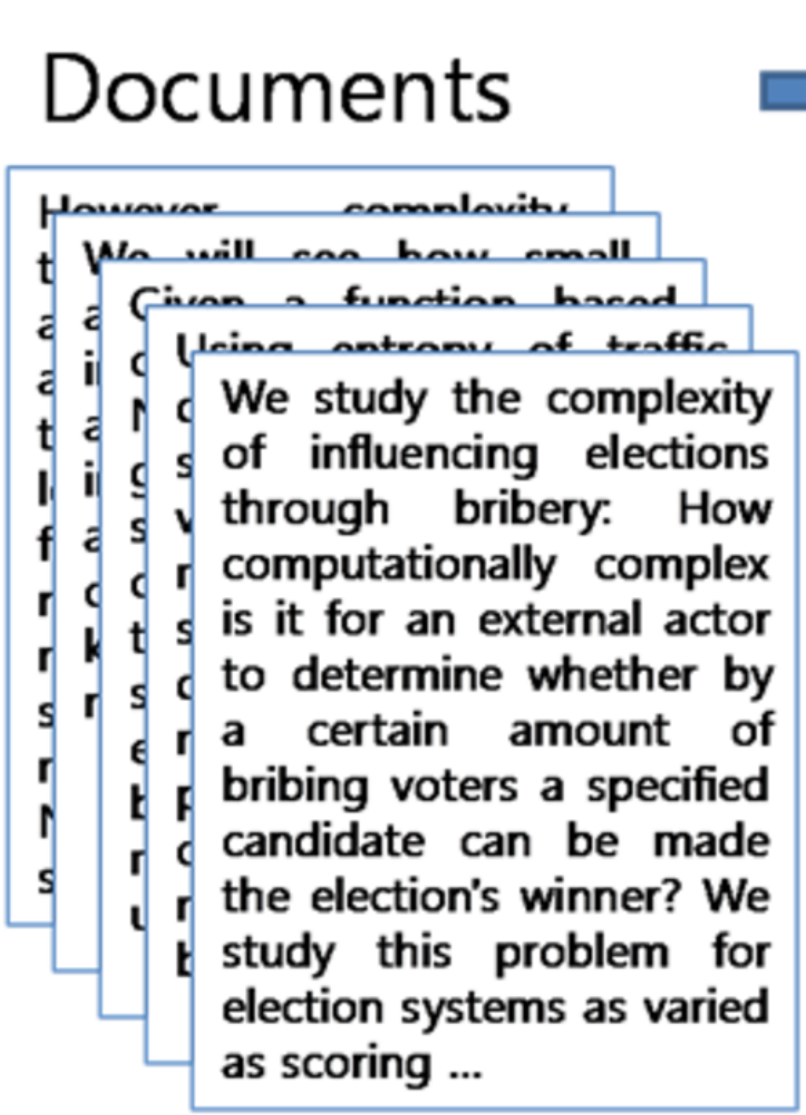


Figure 13. Vocabulary Trend Plots



Term Document Matrix



Vector-space representation

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Term-document matrix

Latent Semantic Analysis

Latent Semantic Analysis (LSA) reduces the dimension of the Term-Document Matrix (TDM, or DTM), which offers two benefits:

1. The DTM is usually a sparse matrix, so algorithms have to work harder on missing data, which is wasteful. Sparseness is attenuated by applying LSA to the TDM.
2. The problem of synonymy arises because many words have similar meanings, i.e., redundancy exists in the list of terms. LSA mitigates this redundancy.

- While not precisely the same thing, think of LSA in the text domain as analogous to PCA in the data domain.

How is LSA implemented using SVD?

- LSA is the application of Singular Value Decomposition (SVD) to the TDM, extracted from a text corpus. Define the TDM to be a matrix $M \in \mathcal{R}^{m \times n}$, where m is the number of terms and n is the number of documents.
- The SVD of matrix M is given by

$$M = T \cdot S \cdot D^{\top}$$

  where $T \in \mathcal{R}^{m \times n}$ and $D \in \mathcal{R}^{n \times n}$ are orthonormal to each other, and $S \in \mathcal{R}^{n \times n}$ is the "singular values" matrix, i.e., a diagonal matrix with singular values on the diagonal. These values denote the relative importance of the terms in the TDM.
- SVD tries to connect the correlation matrix of terms ($M \cdot M^{\top}$) with the correlation matrix of documents ($M^{\top} \cdot M$) through the singular matrix.
- To see this connection, note that matrix T contains the eigenvectors of the correlation matrix of terms. Likewise, the matrix D contains the eigenvectors of the correlation matrix of documents.

LDA Explained (Briefly)

- Latent Dirichlet Allocation (LDA) was created by David Blei, Andrew Ng, and Michael Jordan in 2003, see their paper titled "Latent Dirichlet Allocation" in the *Journal of Machine Learning Research*.
- The simplest way to think about LDA is as a probability model that connects documents with words and topics. The components are:
  - A Vocabulary of V words, i.e., $w_1, w_2, ..., w_i, ..., w_V$, each word indexed by i.
  - A Document is a vector of N words, i.e., $\mathbf{w}$.
  - A Corpus D is a collection of M documents, each document indexed by j, i.e. d_j.
- Next, we connect the above objects to K topics, indexed by l, i.e., t_l. We will see that LDA is encapsulated in two matrices: Matrix A and Matrix B.

Matrix A: Connecting Documents with Topics

- This matrix has documents on the rows, so there are M rows.
- The topics are on the columns, so there are K columns.
- Therefore $A \in \mathcal{R}^{M \times K}$.
- The row sums equal 1, i.e., for each document, we have a probability that it pertains to a given topic, i.e., $A_{jl} = Pr[t_l|d_j]$, and $\sum_{l=1}^{K} A_{jl} = 1$.

Matrix B: Connecting Words with Topics

- This matrix has topics on the rows, so there are K rows.
- The words are on the columns, so there are V columns.
- Therefore $B \in \mathcal{R}^{K \times V}$.
- The row sums equal 1, i.e., for each topic, we have a probability that it pertains to a given word, i.e., $B_{li} = Pr[w_i|t_l]$, and $\sum_{i=1}^{V} B_{li} = 1$.

Schematic of LDA

- Matrix $A = docs \times topics \in \mathcal{R}^{M \times K}$, $\sum Pr[t|d] = 1$.
- Matrix $B = topics \times words \in \mathcal{R}^{K \times V}$, $\sum Pr[w|t] = 1$.
- Document Term Matrix $DTM = docs \times words \in \mathcal{R}^{M \times V}$.
- θ is a topic vector; let $P(\theta|\alpha) \sim Dirichlet$, α is a parameter vector.
- A sampling of a topic t from mixture θ with probability $p(t|\theta)$.

Distribution of Topics in a Document

- Using Matrix A, we can sample a K-vector of probabilities of topics for a single document. Denote the probability of this vector as $p(\theta|\alpha)$, where $\theta, \alpha \in \mathcal{R}^K$, $\theta, \alpha \geq 0$, and $\sum_l \theta_l = 1$.
- The probability $p(\theta|\alpha)$ is governed by a Dirichlet distribution, with density function

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{l=1}^{K} \alpha_l)}{\prod_{l=1}^{K} \Gamma(\alpha_l)} \prod_{l=1}^{K} \theta_l^{\alpha_l - 1}$$

  where $\Gamma(\cdot)$ is the Gamma function.
- LDA thus gets its name from the use of the Dirichlet distribution, embodied in Matrix A. Since the topics are latent, it explains the rest of the nomenclature.
- Given θ, we sample topics from matrix A with probability $p(t|\theta)$.

Distribution of Words and Topics for a Document

- The number of words in a document is assumed to be distributed Poisson with parameter ξ.
- Matrix B gives the probability of a word appearing in a topic, $p(w|t)$.
- The topics mixture is given by θ.
- The joint distribution over K topics and V words for a topic mixture is given by

$$p(\theta, \mathbf{t}, \mathbf{w}) = p(\theta|\alpha) \prod_{l=1}^{K} p(t_l|\theta) p(w_l|t_l)$$

- The marginal distribution for a document's words comes from integrating out the topic mixture θ, and summing out the topics $\mathbf{t}$, i.e.,

$$p(\mathbf{w}) = \int p(\theta|\alpha) \left(\prod_{l=1}^{K} \sum_{t_l} p(t_l|\theta) p(w_l|t_l) \right) d\theta$$

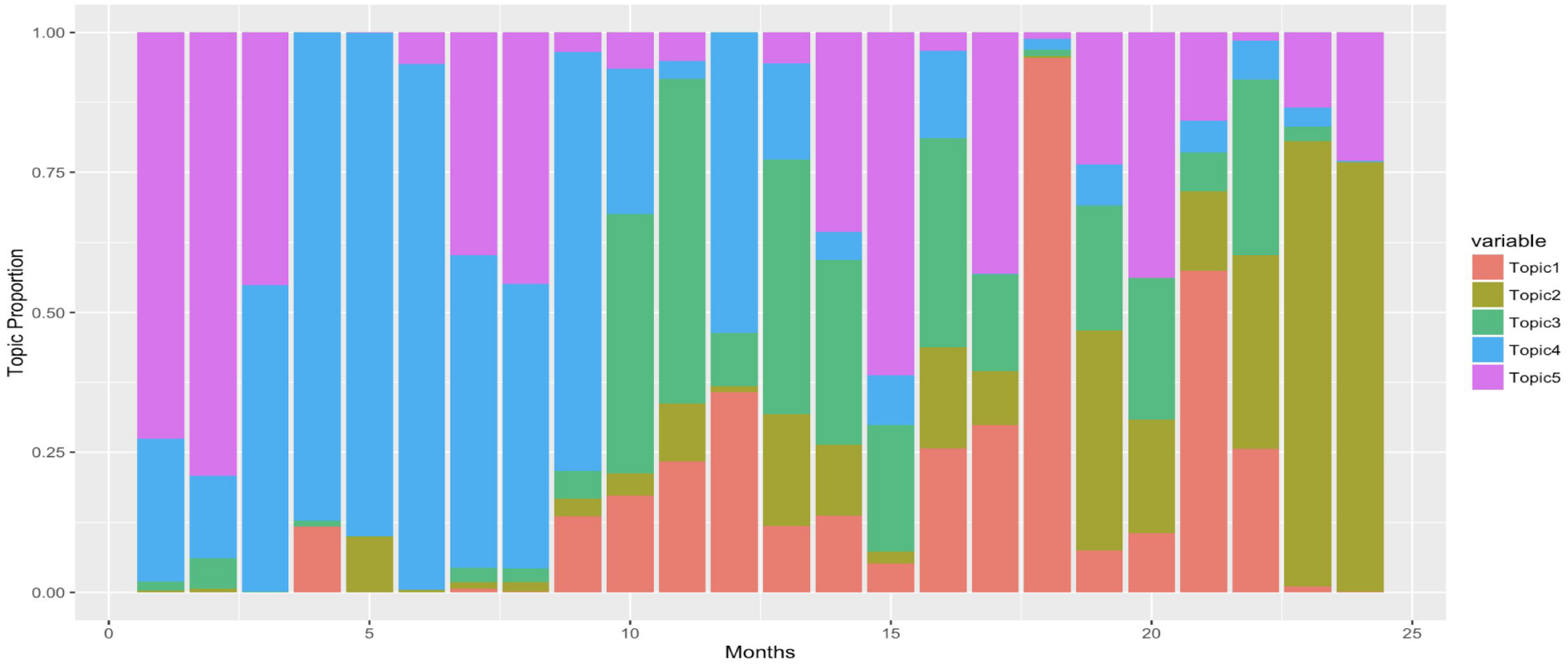
Likelihood of the entire Corpus

- This is given by:

$$p(D) = \prod_{j=1}^{M} \int p(\theta_j|\alpha) \left(\prod_{l=1}^{K} \sum_{t_{jl}} p(t_l|\theta_j) p(w_l|t_l) \right) d\theta_j$$

- The goal is to maximize this likelihood by picking the vector α and the probabilities in the matrix B. (Note that were a Dirichlet distribution not used, then we could directly pick values in Matrices A and B.)
- The computation is undertaken using MCMC with Gibbs sampling as shown in the example earlier.

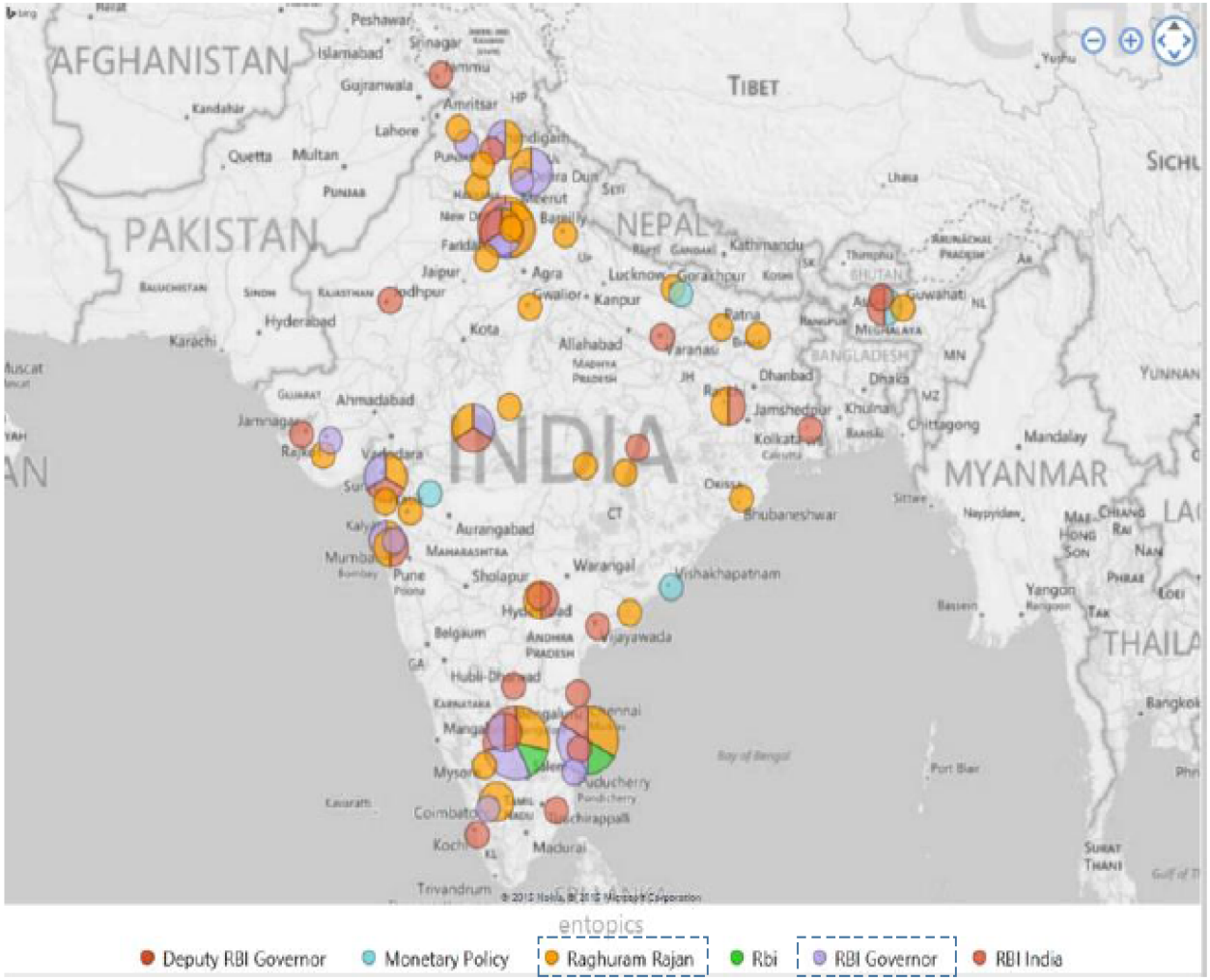
Figure 14. Topic Contributions by Month



Time distribution of top five topics from a distillation of 1,302 news articles on Factiva, from PR Newswire (US). The sentiment scores for the five topics are: $\{0.40, -0.47, 0.93, 0.66, -1.52\}$, i.e., topic 3 has the highest (most positive) sentiment, and topic 5 has the lowest (most negative) sentiment.

Examples in Finance - 1

Conversations across India and around RBI topycs



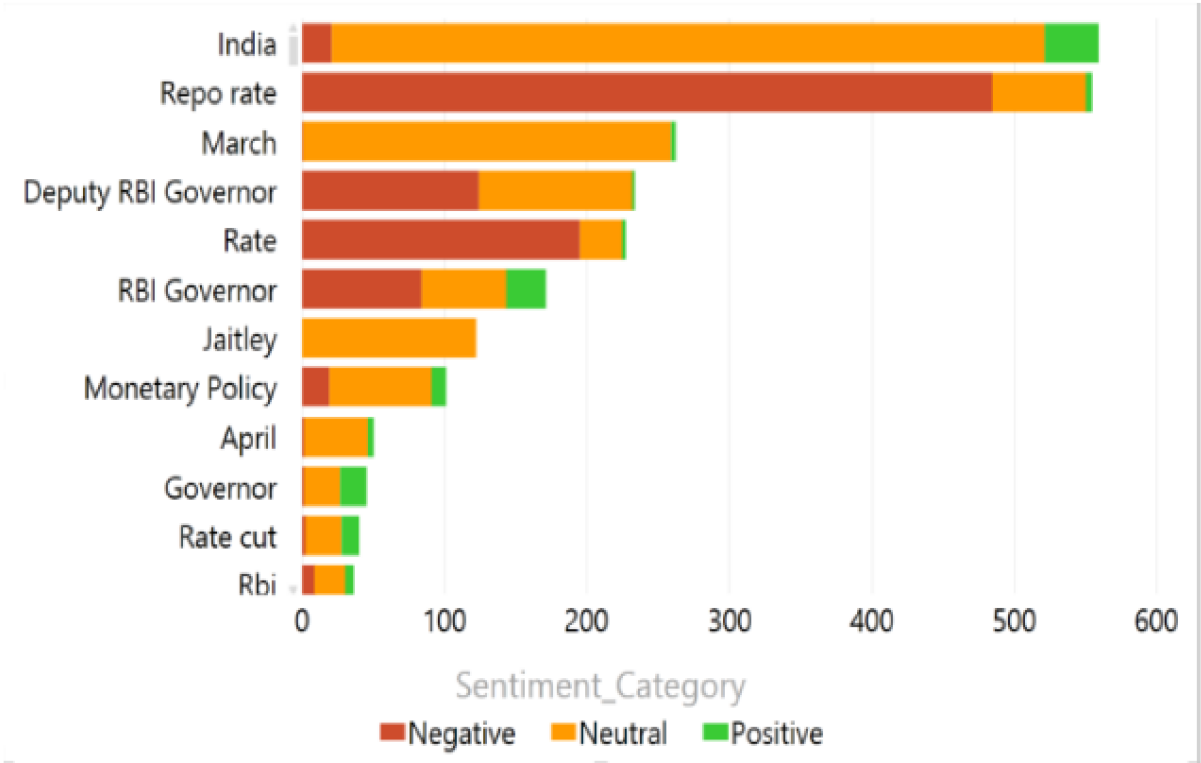
- Conversations across India on RBI, its people and the monetary policy
- Governor features in many conversations across both rural and urban areas
- Some conversations specifically around monetary policy

- Bubbles show split of conversations around Deputy RBI Governor, Monetary Policy, Raghuram Rajan, RBI and RBI Governor.
- Based on count of unique conversations
- Date Range: 1st – 14th April, 2015

Examples in Finance - 2

Top Topics along with RBI

topycs



text

"@NDTVProfit: RBI unlikely to change repo rate at policy review smlion

"Digging India's RBI Out of Morass of Debt" by on

"Financial stability is like Pornography. You can't define it but when you see it you know it" - D Subbarao (RBI Governor)

"I was disappointed by the fiscal relaxation." Ex-RBI Governor on India's budget and growth:

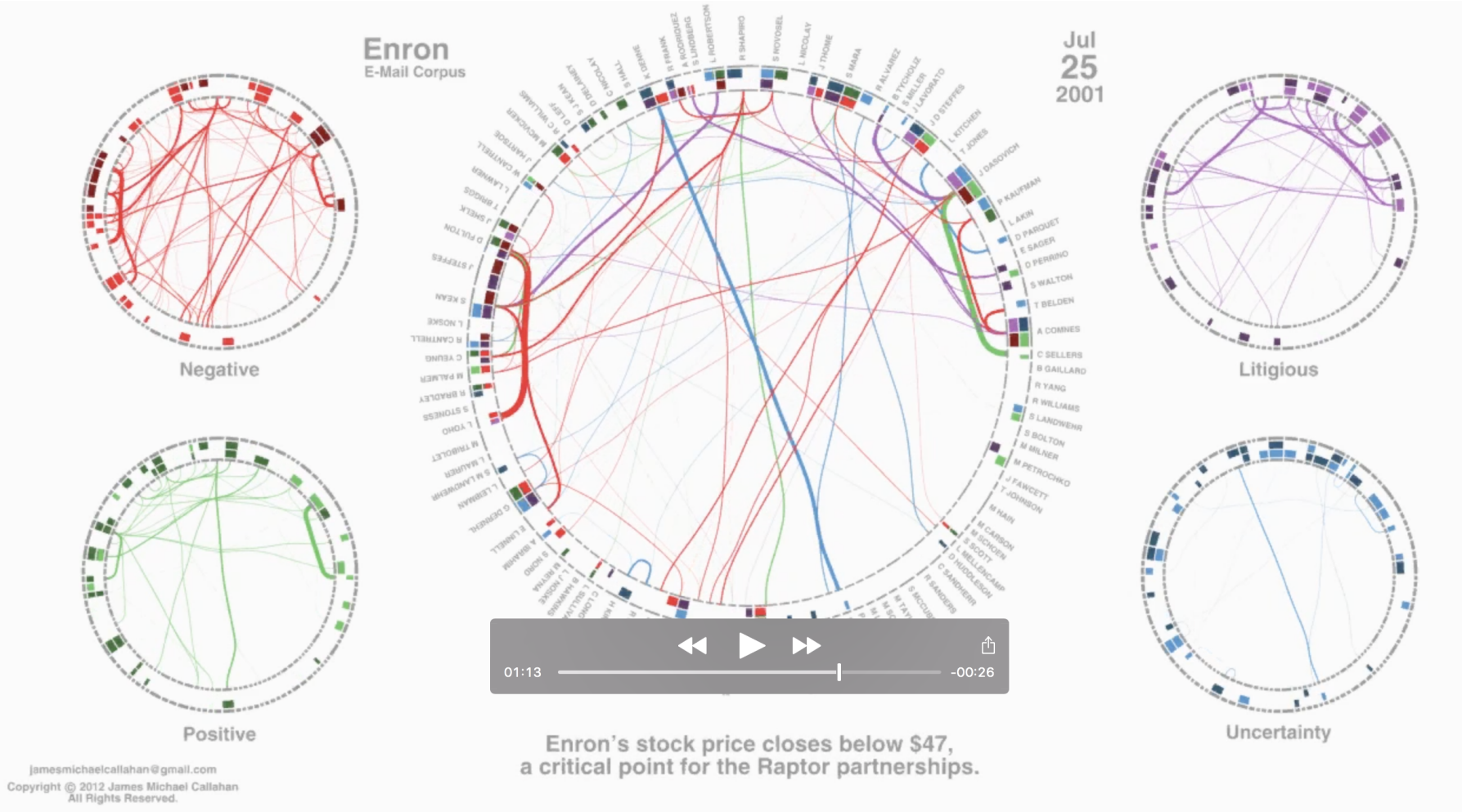
"Rajan is perfect, he explains complex economic," PM Modi on RBI governor.

"RBI Conference" shows up as trending topic in India at rank 10

- Repo rate evokes negative sentiment as people don't expect it to be changed
- Repo rate, rate cut and monetary policy are discussed frequently with RBI

- Vertical Axis – Topics of Discussion
- Horizontal Axis – Count of Unique Conversations
- Date – 25th March - 14th of April
- Colors represent sentiment for conversation, Negative – Red, Neutral – Orange, Positive - Green

Enron Movie by James Callahan



http://srdas.github.io/Presentations/JimCallahan_enron-sm.mov

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.