Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards

and

Internet Stock Message Boards and Stock Returns

Werner Antweiler and Murray Frank
University of British Columbia

Presentations at the University of Calgary (October 4, 2002)
and the University of Alberta, Edmonton (October 7, 2002)

Papers available on the web at http://pacific.commerce.ubc.ca/antweiler/

Outline

1. **Introduction**
   - Background • Research Questions
   - Theory: Why do people trade? • Data Set
2. **Computer Linguistics and Text Classification**
   - Classification Algorithms
   - Bullishness Index • Agreement Index
3. **Financial Econometrics — Short Time Horizon**
   - Contemporaneous Effects
   - Time Sequencing Tests and Stock Returns
   - Predictive Ability: Wall Street Journal
   - Volatility • Trading Volume
4. **Conclusions — Short Time Horizon**
5. **Financial Econometrics — Longer Time Horizon**
   - Portfolio Simulations • Market Factor Analysis
6. **Conclusions — Longer Time Horizon**

Background

"Internet message boards have come of age. [...] even investment pros are watching the message boards closely and are profiting from it. With 'posts' running in the millions, Internet message boards have become an essential part of the savvy investor's arsenal. [...] Internet messages really do move markets, for better or worse."

— G. Weiss, Business Week, May 22, 2000

- Securities Exchange Commission has prosecuted people for internet messages
  - "Tokyo Joe" fined US$754,630 for "scalping"
  - 15-year old Jonathan Lebed
  - more than 250 cases filed by March 2001
- Recent case in British Columbia:

  The *Vancouver Sun* (Wednesday Sept. 25, 2002 "Stock scam brings ban") reports that Jesse Hogan (26) of Burnaby, BC had posted hundreds of false messages on internet stock message boards. He gained US$41,753 in illicit profits using a "pump and dump" scam involving high-tech penny stocks. The B.C. Securities commission has fined him C$25,000, and he must surrender the profits.

Problems for the SEC

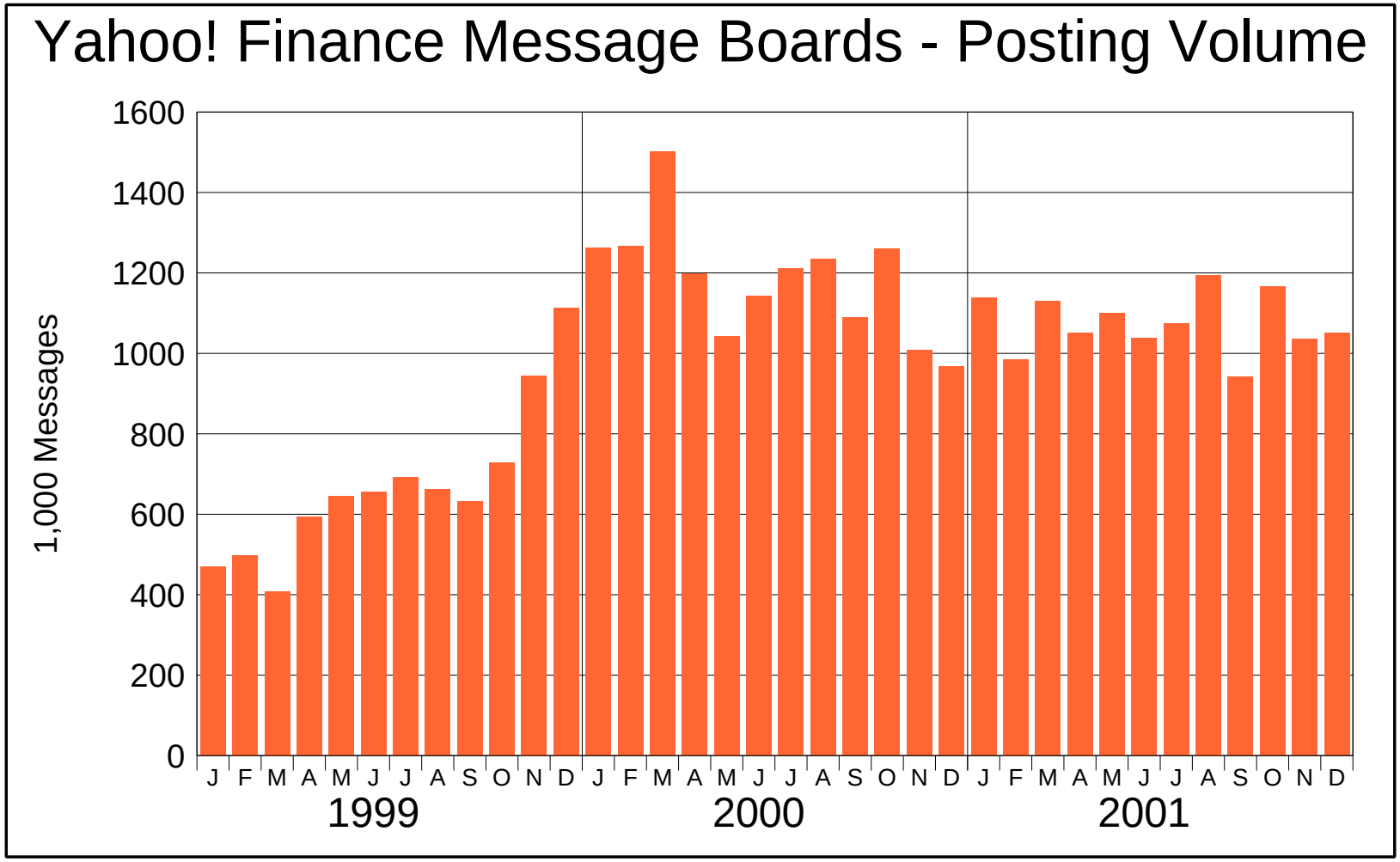
- Old scams in new disguise: "pump and dump"; Ponzi schemes
- Issues: Balancing investor protection against free speech
- **SEC Office of Internet Enforcement**
  - formed in July 1998
  - OIE coordinates "cyberforce" of over 200 commission attorneys, accountants and investigators nationwide working on internet surveillance
  - http://www.sec.gov/divisions/enforce/internetenforce.htm

stock message boards

Internet Information Channels for Investors

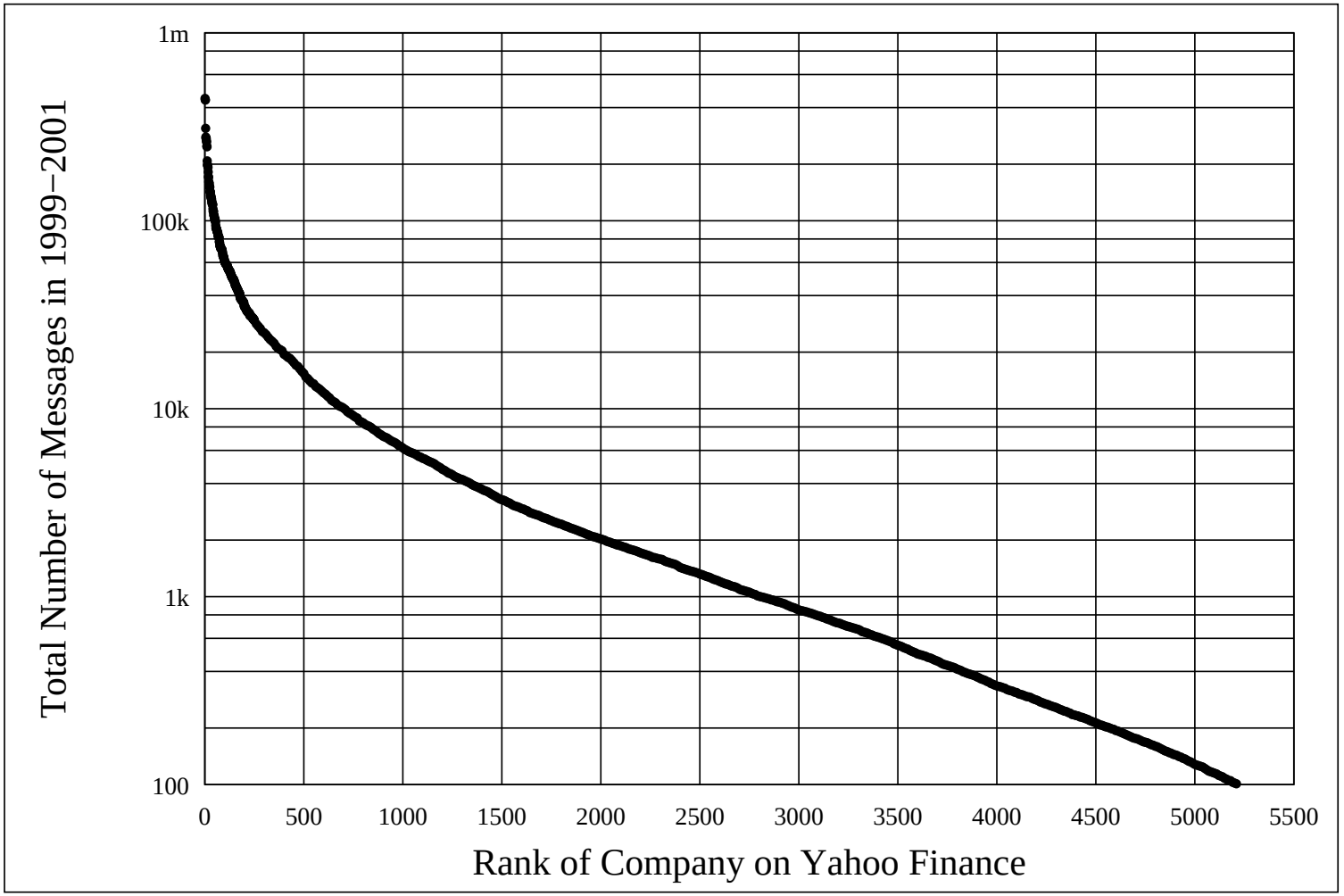
- **Internet Stock Message Boards**
  - Yahoo! Finance
  - Raging Bull
  - Silicon Investor
  - Motley Fool
- **Web Sites** — often company-specific
  - http://www.dotcomscoop.com (mostly rumours)
  - http://www.tedsturnovers.com (Ex-CNN staff; now shut)
  - http://www.netslaves.com
  - http://justsaynowey.com (Weyerhauser hostile tender for Williamette. Employees opposed. SEC shut down web site; WSJ May 7, 2001)
- **Chat Rooms**
- **Newsgroups** (e.g., alt.invest.penny-stocks)

Message Posting Activity



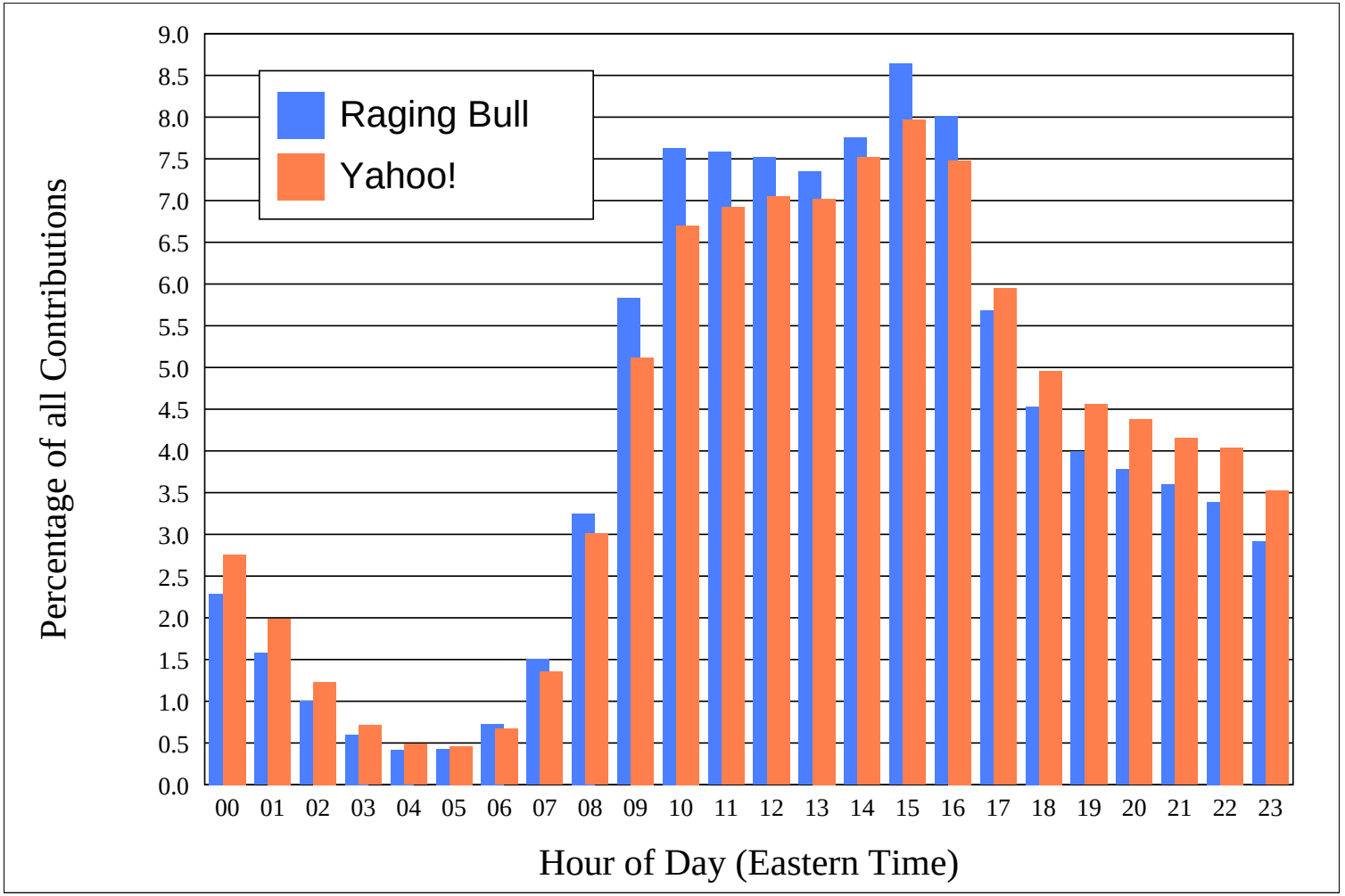
- 1 million messages per month on Yahoo Finance

Rank Distribution of Posting Volume



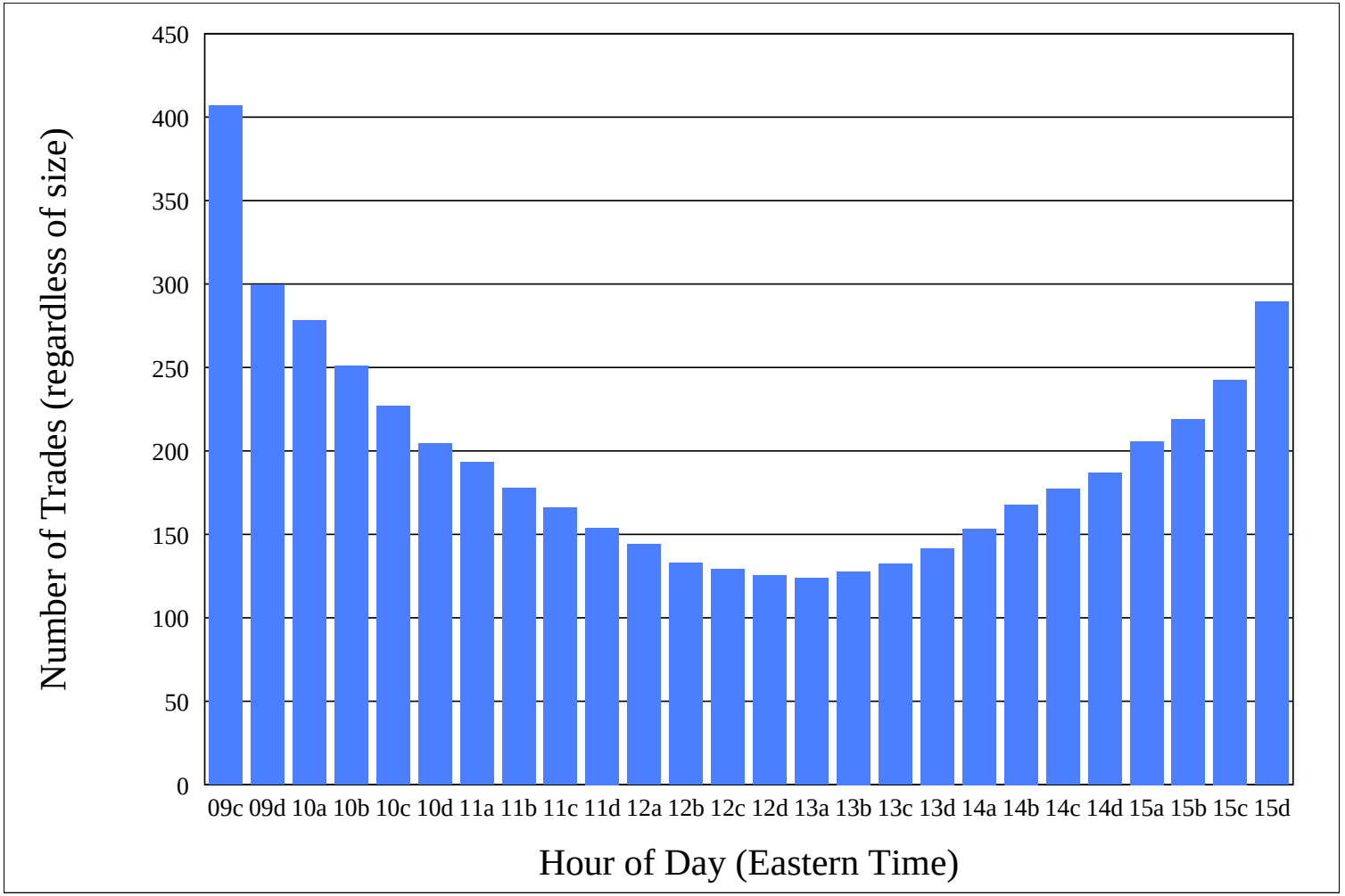
- another application of Zipf's law?

Intra-Day Distribution of Posting Activity



- most of the posting activity is *during* the business day; many people must be accessing the message boards from work
- who are the message posters? day traders?

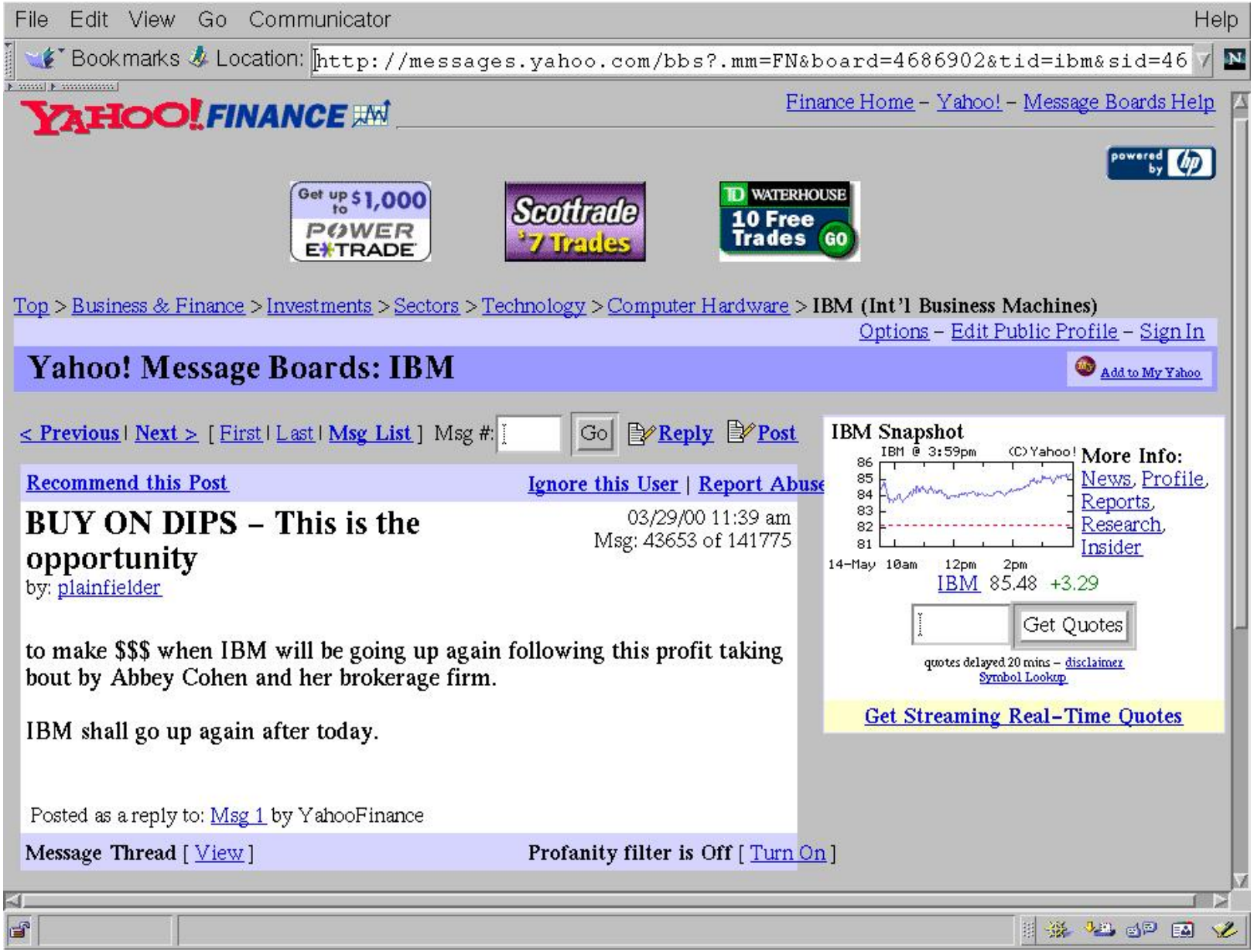
Intra-Day Distribution of Trading Activity



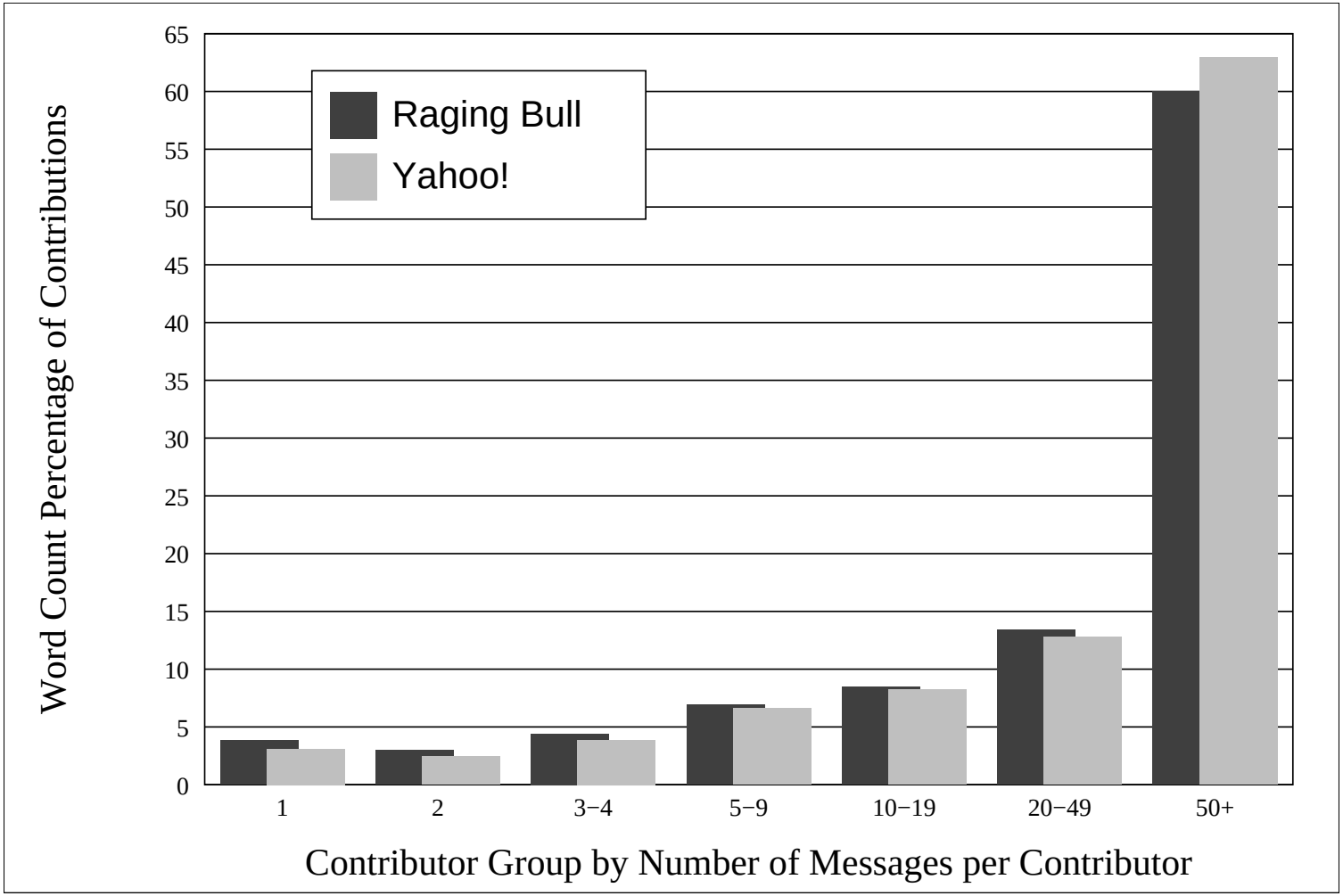
- trading activity is U-shaped between 09:30 and 16:00

stock message boards

An Example

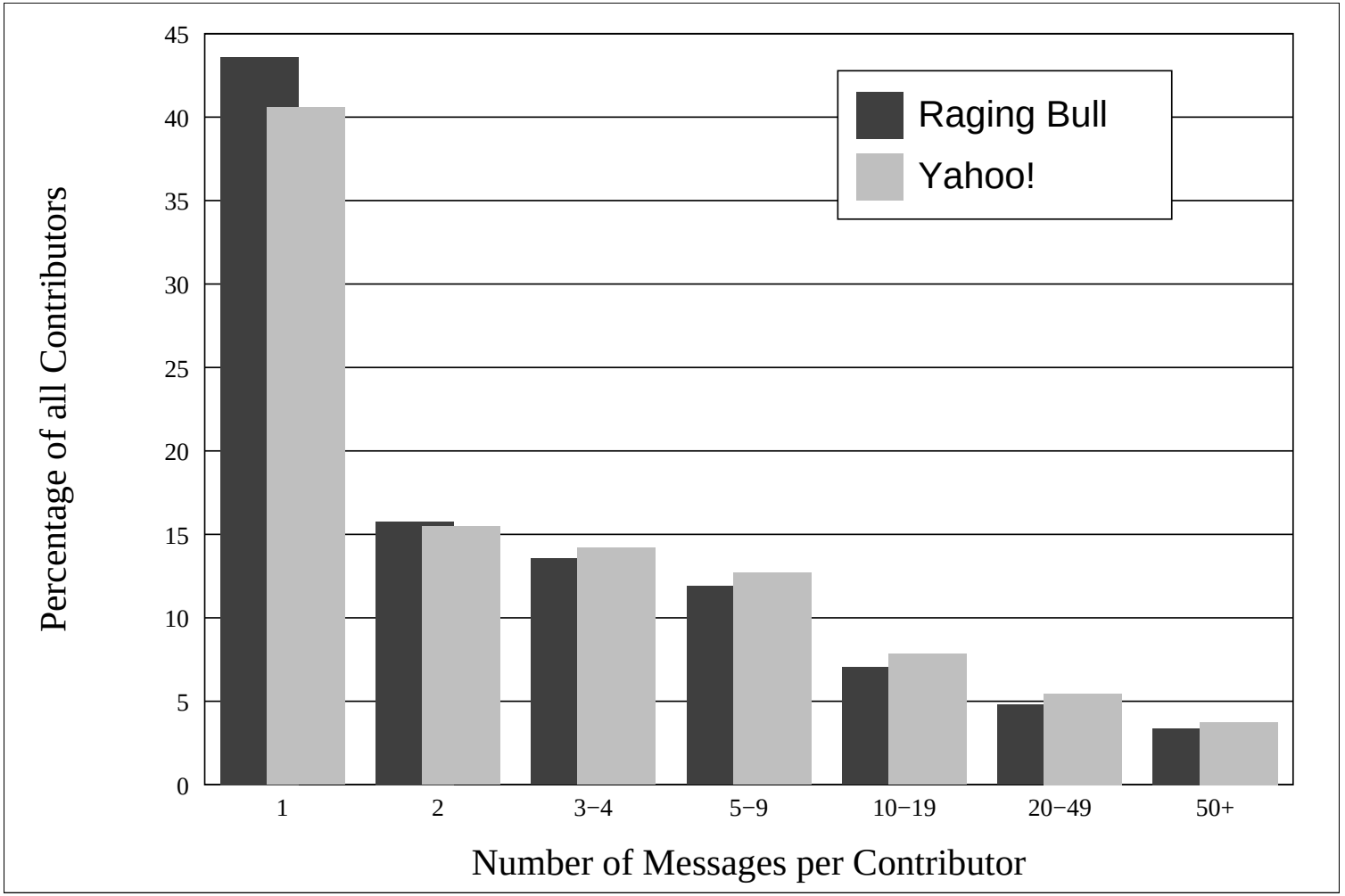


Who are the message posters?



- The bulk of messages come from repeat-posters...

Who are the message posters?



- ...but the repeat-posters are a small fraction of the entire population of message posters

Research Questions

1. **Do the messages help predict returns?**
2. **Is disagreement among the messages associated with more trades?**
   - Karpoff (1986) and Harris and Raviv (1993)
     Disagreement induces trading
   - No-trade theorem of Milgrom and Stokey (1982):
     Disagreement leads to revision of beliefs, not trades
3. **Do the messages help predict volatility [and trading volume]?**

Related Work

- Wysocki (1999) is first study of internet stock message boards.
- Tumarkin and Whitelaw (2001) find no predictive ability for returns (sample of 9 firms).
- Das and Chen (2001) develop new natural language algorithm to classify stock messages.
- Tufano, Das, and Martinez-Jerez (2001)

Research Data Set — Dimensions

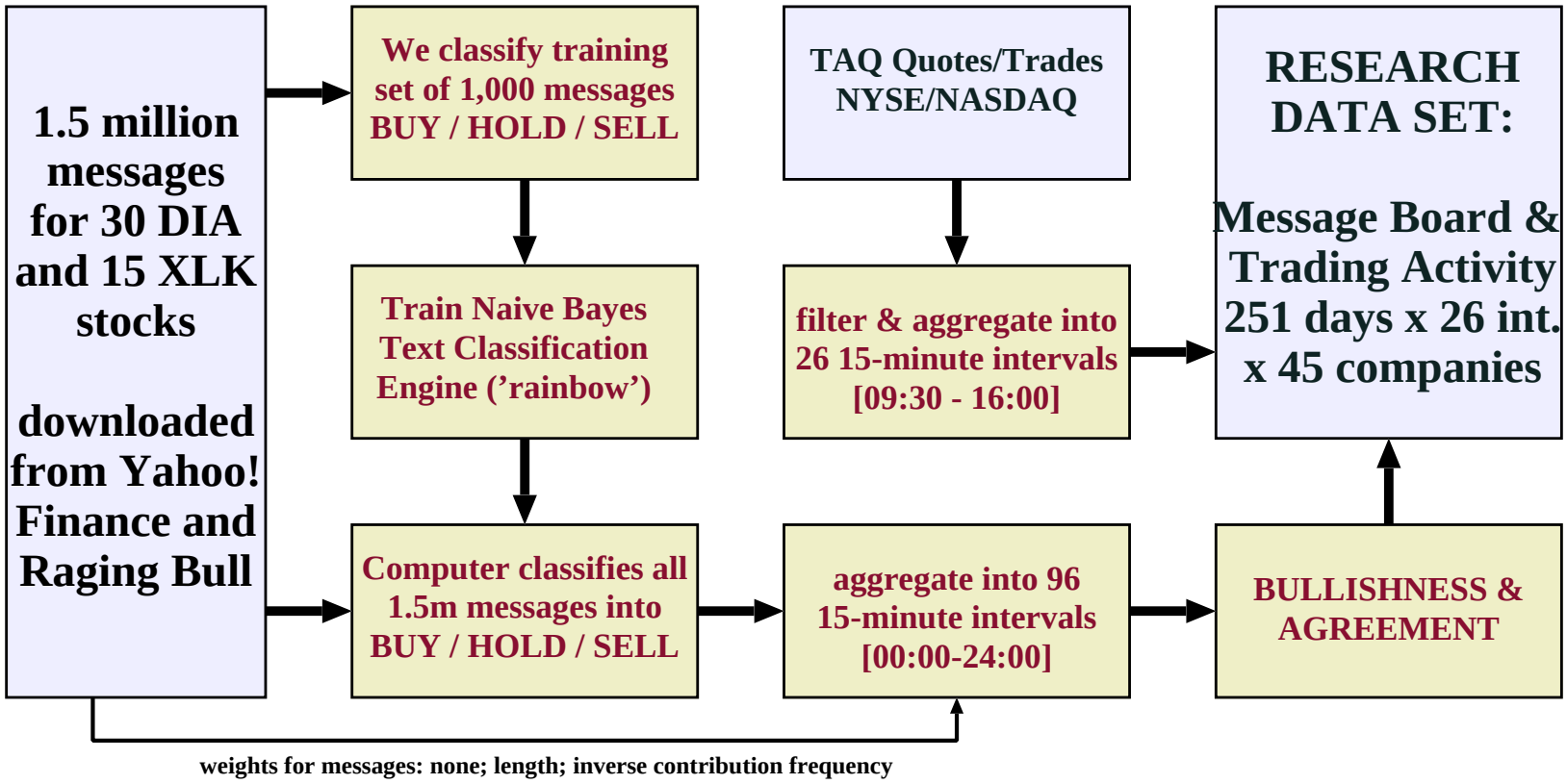
- 251 trading days on the NYSE, AMEX, NASDAQ
- 26 fifteen-minute intervals per day (09:30–16:00)
- 45 stocks in our sample:
  - **30 DIA companies** reflecting the "old economy"

    Alcoa; American Express; Boeing; Citigroup; Caterpillar; Du Pont; Walt Disney; Eastman Kodak; General Electric; General Motors; Home Depot; Honeywell; Hewlett Packard; IBM; Intel; Intn'l Paper; Johnson&Johnson; JP Morgan; Coca-Cola; McDonalds; Minnesota Mining; Philip Morris; Merck; Microsoft; Procter&Gamble; SBC Communications; AT&T; United Technologies; Wal Mart; Exxon.
  - **15 XLK companies** reflecting the "new economy"

    Amazon; Ameritrade; CNet; E*Trade; E-Bay; Etoys; Go2net; Healtheon; Lycos; MP3.com; Priceline; Ticketmaster; Verticalnet; Webvan Group; Yahoo!.

Preparing the Research Data Set

Extracting ⇒ Classifying ⇒ Aggregating ⇒ Merging
Data Sources • Computer Linguistics



Computer Linguistics

- How can one machine-classify large amounts of text?
  statistical methods *vs.* content analysis
- Statistical methods are crude but remarkably accurate
- We use Naive Bayes:
  - each word is associated with an odds ratio
    for being 'bullish' relative to 'bearish'
  - messages are examined word by word
  - algorithm can be fooled easily (e.g., negation) but is
    remarkably robust for longer messages.
- We also obtain results from another classification method:
  Support Vector Machine
- General problem:
  message boards have a low signal-to-noise ratio.

Naive Bayes in Detail (1)

- Start with prior $P(T)$ that a document is "buy", "sell", or "hold" based on observed frequency of documents in our training data set.
- Update odds-ratios by using conditional probabilities $P(W_i|T)$ and $P(W_i|\tilde{T})$ observed in the training set:

$$\frac{P(T|W_i)}{1-P(T|W_i)} = \frac{P(T|W_{i-1})}{1-P(T|W_{i-1})} \cdot \frac{P(W_i|T)}{P(W_i|\tilde{T})}$$

- Employ log of odds ratios for better accuracy and sum over all words in the document:

$$P(T|W_N) = P(T)\exp\left[\sum_{i=1}^{N}\log\left(\frac{P(W_i|T)}{P(W_i|\tilde{T})}\right)\right]$$

- Use *Laplace smoothing* when $P(W_i|T)P(W_i|\tilde{T}) = 0$.

Naive Bayes in Detail (2)

- Prune vocabulary to the 1000 words with highest information gain
  (occurrence frequency and largest abs(log(odds-ratio))).
- Calculate separate probabilities for signals "buy", "hold" and "sell", then choose

$$x_k = \arg\max_{T_c} P(T_c|W_n^k)$$

- Use software package "rainbow" from McCallum (1996)

Classification Accuracy

Classified:		By Algorithm		
By Us	%	Buy	Hold	Sell
Buy	25.2	18.1	7.1	0.0
Hold	69.3	3.4	65.9	0.0
Sell	5.5	0.2	1.2	4.1
1,000 messages[1]		21.7	74.2	4.1
All Messages[2]		20.0	78.8	1.3

- Naive Bayes has excellent in-sample classification accuracy
- but out-of sample accuracy difficult to measure
- we gauge accuracy by using the following method
  - train system with $N - M$ messages
  - use remaining M messages to validate learning
  - repeat by randomizing the selection of M messages
- Support Vector Machines tends to overfit

Bullishness Index

- Aggregate over 15-minute intervals.
- Have $M_t = M_t^{\text{BUY}} + M_t^{\text{SELL}}$ messages in time period t
- Which measure of bullishness?

$$B_t \equiv \frac{M_t^{\text{BUY}} - M_t^{\text{SELL}}}{M_t^{\text{BUY}} + M_t^{\text{SELL}}} \in [-1, +1]$$

$$B_t^* \equiv \ln\left[\frac{1 + M_t^{\text{BUY}}}{1 + M_t^{\text{SELL}}}\right] \approx B_t \ln(1 + M_t)$$

$$B_t^{**} \equiv M_t^{\text{BUY}} - M_t^{\text{SELL}} = B_t M_t$$

- We choose intermediate measure B^* which reflects direction and size of sentiment, but
- We ignore neutral/hold messages (mostly noise)
- Weights:
  - unweighted [preferred]
  - length of the message (number of words)
  - author's inverse contribution frequency

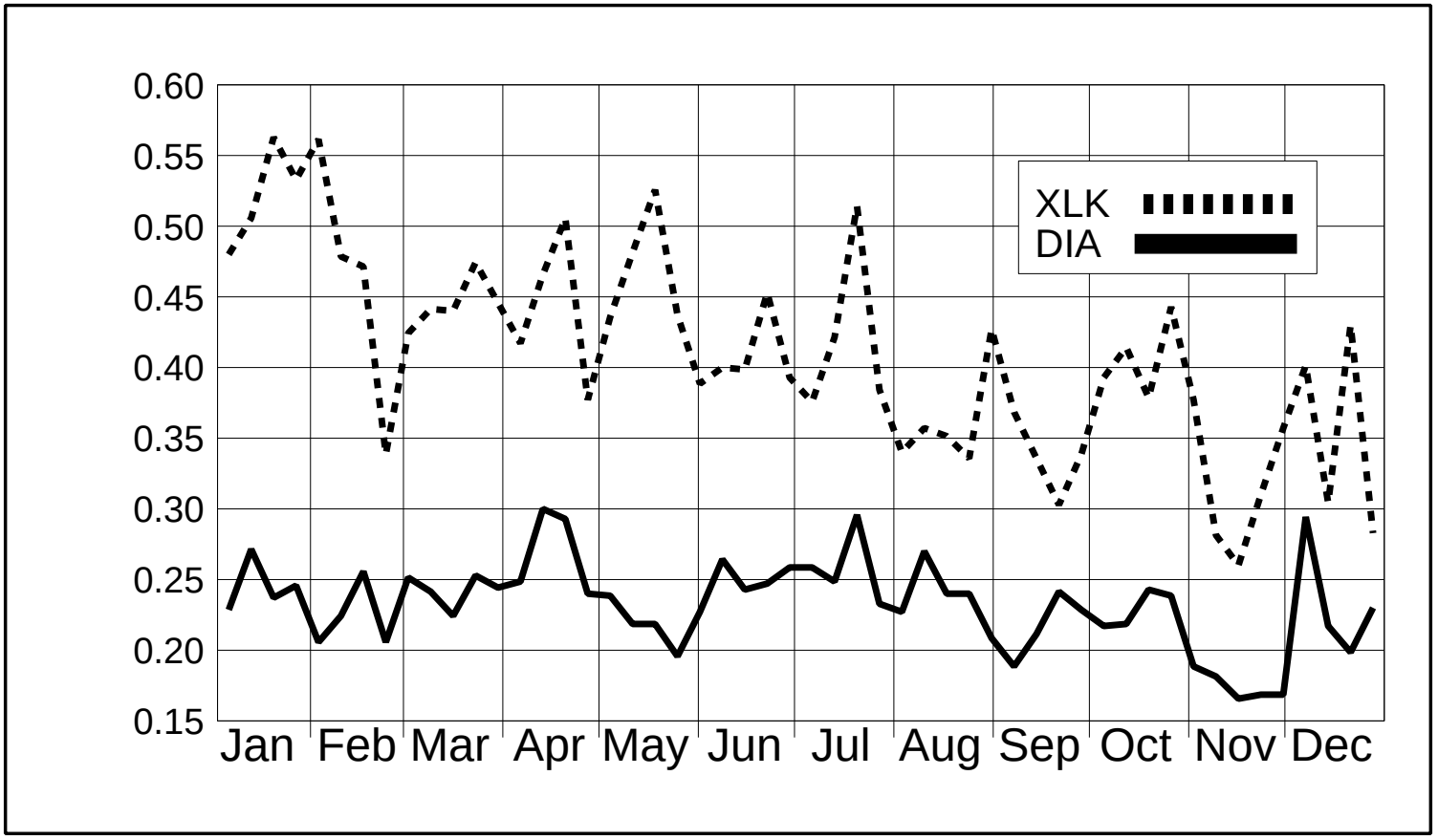
Agreement Index

- We derive a measure of disagreement from our bullishness measure $B_t \in [-1, +1]$
- It can be shown that

$$A_t \equiv 1 - \sqrt{1 - B_t^2} \in [0, 1]$$

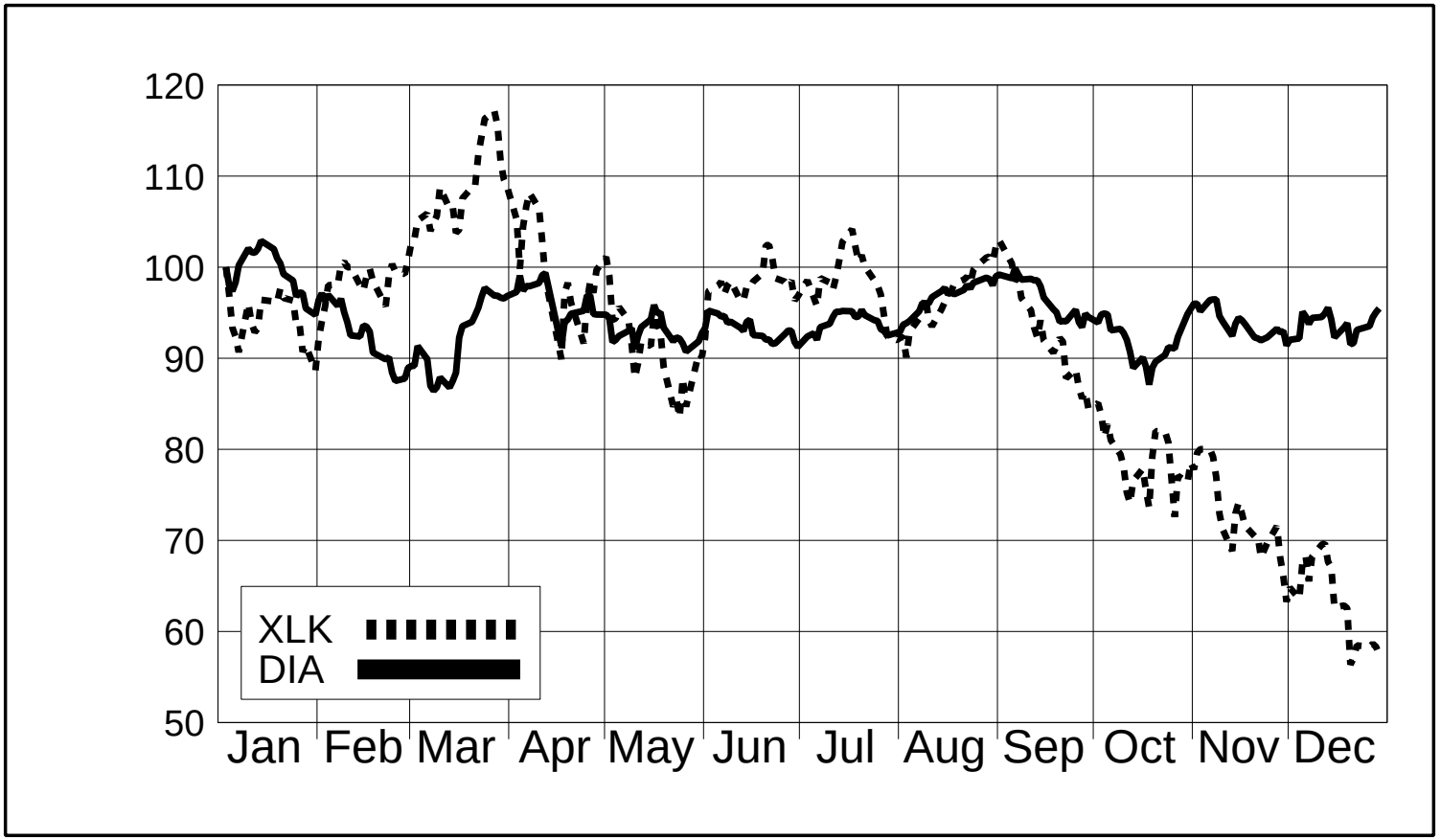
  is the standard deviation of bullishness index B_t.
- Examples:
  - 2 sell, 0 buy messages:
    $B_t = (0 - 2)/(0 + 2) = -1$ is bearish;
    $A_t = 1 - \sqrt{1 - (-1)^2} = 1$ indicates full agreement.
  - 2 sell, 1 buy message:
    $B_t = (1 - 2)/(1 + 2) = -0.5$ is half bearish;
    $A_t = 1 - \sqrt{1 - (-0.5)^2} = 0.13$ indicates weak agreement.
  - 2 sell, 2 buy messages:
    $B_t = (2 - 2)/(2 + 2) = 0$ is neutral;
    $A_t = 1 - \sqrt{1 - 0} = 0$ indicates no agreement.

Bullishness Index – XLK vs. DIA



- On Yahoo, "new economy" (XLK) stocks remain consistently more bullish during 2000 than "old economy" (DIA) stocks.

Stock Index Performance – XLK vs. DIA



- Internet bubble bursts in August 2000

Contemporaneous Effects: 15-minute intervals

	Log of Messages	Bullishness Index	Agreement Index	Market	R^2
Return	-0.331 (1.382)	1.747^b (3.208)	-0.240 (0.455)	0.716^c (120.7)	0.049
Volatility	0.041^c (35.7)	0.033^c (12.74)	-0.029^c (11.41)	-1.178^c (81.85)	0.538
Log Small Trades	0.225^c (102.1)	0.181^c (36)	-0.123^c (25.3)	-1.541^c (55.88)	0.984
Log Medium Trades	0.119^c (43.53)	0.161^c (25.82)	-0.096^c (15.84)	-0.464^c (13.55)	0.931
Log Large Trades	0.082^c (37.29)	0.052^c (10.39)	-0.021^c (4.382)	-0.222^c (8.073)	0.642
Log Trading Volume	0.259^c (82.37)	0.170^c (23.81)	-0.109^c (15.72)	-2.417^c (61.55)	0.995
Spread	0.001 (0.766)	0.009^b (2.861)	-0.004 (1.369)	-0.047^b (2.763)	0.245

Significance level indicators: 95%=^{a}; 99%=^{b}; 99.9%=^{c}. T-stats in parentheses.

- We regress financial variables on message board variables (panel data, fixed company effects)
- Key results:
  - Increased message posting activity is associated with higher trading volatility and higher trading volume.
  - Greater agreement is associated with reduced market volatility and small trades
  - There is a contemporaneous (non-predictive) link between bullishness and price returns.

Time Sequencing Tests: The Chicken-&-Egg Problem

- **Empirical Strategy**
  - Vector Auto Regression of pooled panel:
    (return, volatility, trades by size, volume, spread)
    against (messages, words, bullishness, agreement)
  - we include a market variable (price/return of SPY) and time-of-day dummies
  - Conduct Granger causality tests (with p lags)
    $(RSS_0 - RSS_1)/RSS_1 \sim \chi^2(p)$
  - Daily data regressions and 15-minute regressions.
- **Key Results**
  - Bullishness does not predict returns (but number of trades)
  - Greater agreement today predicts *more* trades tomorrow, in contrast to the contemporaneous effect
  - Predictive ability works both ways: message board activity predicts some financial variables, and financial variables predict some message board variables.

Granger Causality Tests: 15-minute intervals

X	Y	$Y = f(X_{-1}, X_{-2}, \text{NWK}, \text{Market})$					$X = f(Y_{-1}, Y_{-2}, \text{NWK}, \text{Market})$				
		X_{-1}	X_{-2}	NWK	Market	χ^2	Y_{-1}	Y_{-2}	NWK	Market	χ^2
messages	return	-0.002^a	0.002^a	-0.002	0.096^c	11.2^a	-0.083	0.106	-0.525^c	-0.332	1.34
messages	volatility	0.015^c	-0.010^b	-0.013^a	-0.557^c	22.0^c	-0.050	0.055	-0.527^c	-0.328	4.48
messages	small	0.074^c	-0.027^c	-0.043^c	-0.507^c	200.^{c}	-0.060^a	0.094^c	-0.535^c	-0.293	25.5^c
messages	medium	0.049^c	-0.051^c	-0.100^c	0.209	55.4^c	-0.029	0.011	-0.531^c	-0.318	7.59
messages	large	0.100^c	-0.067^c	-0.206^c	0.123	96.6^c	-0.021	0.010	-0.529^c	-0.409	5.50
messages	volume	0.111^c	-0.029^c	-0.156^c	-0.987^c	288.^{c}	-0.030	0.056^b	-0.532^c	-0.276	14.9^b
messages	spread	0.002	-0.002	-0.000	-0.042	2.05	-0.029	-0.034	-0.525^c	-0.330	0.86
words	return	-0.001	0.001^a	-0.002	0.096^c	7.44	-0.275	0.244	-0.597^c	-0.753^b	3.89
words	volatility	0.005	-0.003	-0.008	-0.558^c	5.62	-0.005	0.064	-0.598^c	-0.688^a	2.76
words	small	0.025^c	-0.006	-0.018	-0.489^c	64.4^c	0.012	0.123^c	-0.595^c	-0.601^a	34.8^c
words	medium	0.017^b	-0.018^c	-0.083^c	0.204	18.9^c	-0.016	0.017	-0.601^c	-0.762^b	0.89
words	large	0.036^c	-0.022^b	-0.176^c	-0.075	33.5^c	0.006	0.013	-0.598^c	-0.736	2.08
words	volume	0.043^c	-0.005	-0.125^c	-0.936^c	122.^{c}	0.040	0.082^b	-0.594^c	-0.498	35.4^c
words	spread	0.001	-0.000	0.001	-0.042	0.26	-0.009	-0.010	-0.598^c	-0.761^b	0.03
bullishness	return	-0.002	-0.003	-0.003	0.098^c	2.83	-0.036	0.018	-0.008	0.204^c	1.15
bullishness	volatility	0.038^b	-0.026^a	-0.002	-0.565^c	17.4^b	0.018	0.008	-0.008	0.238^c	9.51^a
bullishness	small	0.136^c	-0.064^c	0.006	-0.534^c	78.9^c	0.053^c	0.003	-0.005	0.299^c	137.^{c}
bullishness	medium	0.117^c	-0.144^c	-0.062^c	0.209	49.2^c	0.016^c	-0.001	-0.007	0.198^c	35.5^c
bullishness	large	0.209^c	-0.076	-0.139^c	-0.236	46.6^c	0.007	0.005	-0.009	0.044	17.2^b
bullishness	volume	0.185^c	-0.045	-0.086^c	-1.013^c	83.4^c	0.030^c	0.005	-0.008	0.296^c	73.3^c
bullishness	spread	0.016^a	0.000	0.002	-0.048	11.2^a	0.024	0.024	-0.009	0.202^c	4.75
agreement	return	-0.002	-0.008	-0.003	0.098^c	4.93	-0.010	0.012	0.003	0.162^c	0.39
agreement	volatility	0.026	-0.024	-0.003	-0.560^c	4.51	0.013	0.001	0.003	0.180^c	7.63
agreement	small	0.091^b	-0.054	0.004	-0.506^c	16.7^b	0.026^c	0.003	0.005	0.210^c	104.^{c}
agreement	medium	0.106^a	-0.138^b	-0.064^c	0.204	19.2^c	0.010^c	0.002	0.004	0.157^c	48.1^c
agreement	large	0.226^c	-0.050	-0.141^c	-0.231	21.9^c	0.004	0.004	0.003	0.132	22.3^c
agreement	volume	0.142^c	-0.057	-0.089^c	-0.953^c	21.5^c	0.013^c	0.004	0.004	0.206^c	47.0^c
agreement	spread	0.028^b	-0.001	0.002	-0.049	13.3^a	0.018	0.033^a	0.003	0.161^c	14.1^b

Predictive Ability: Wall Street Journal

- What do we do?
  - Try to predict whether message posting today predicts articles in the WSJ tomorrow (or day after tomorrow)
  - We use a logistic regression of daily data with company and day-of-week fixed effects
- What do we find?
  - Message boards predict news stories in WSJ
    - one day before publication
    - two days before publication: somewhat weaker but still significant

Volatility

- If message posters are 'noise traders,' then their actions may induce market volatility.
- Trading volume and volatility are correlated, and message posting volume is correlated with trading volume.
- We want to predict fractionally-integrated realized volatility $v_{i,t}$ as in Andersen et.al. (2001)
  - Estimate fractional integration parameter
  - Calculate a measure of realized volatility
  - Estimate realized volatility by linking it to message board activity
    - as a panel
    - for individual stocks, then average
- Use GARCH, EGARCH, and GJR-TGARCH as robustness checks; the latter two allow for asymmetric responses (the 'leverage effect')

Fractional Integration

- Fractional integration allows for 'long memory' of innovations.
- Estimate the periodogram $\psi_{i,k}$ for company i at frequency points $k = 0, 1, .., \lfloor T^{0.6} \rfloor$.
  (Most stats packages have a routine for this.)
- Determine the fractional integration parameter d through a log-periodogram panel regression

$$\ln(\psi_{i,k}) = \alpha - d \ln \left[4 \sin \left(\frac{k\pi}{T} \right) \right] + \mu_i + u_{i,k}$$

- We estimate $d = 0.297$ with standard error 0.021. This is close to the median value of 0.349 reported by Andersen et.al.

Realized Volatility

- Our dependent variable is the log standard deviation of log returns whose MA(1) component has been filtered out: $\ln(v_{i,t})$
- We estimate (individually and as a panel)

$$\begin{aligned}(1-L)^d \ln(v_{i,t}) \;=\; & \beta_i + \beta_{v,i} \ln(v_{i,t-1}) + \beta_{vr,i} \ln(v_{i,t-1}) I(r_{i,t-1} < 0) \\ & + \beta_{A,i} A_{i,t-1} + \beta_{M,i} \ln(1 + M_{i,t-1}) \\ & + \beta_{N,i} \ln(N_{i,t-1}) + u_{i,t}\end{aligned}$$

  - A_{it} is our agreement index
  - M_{it} is posting volume
  - N_{it} is trading volume
  - $I(\cdot)$ is the indicator function

Volatility: Results

- Individual regressions
  - Strong cross-sectional differences
  - Amazon, IBM, MP3.com and Microsoft all have significant effects from message posting to volatility.
  - Agreement appears to have insignificant effect.
- Panel regression: clear effect of message posting on volatility; smaller effect for DIA than XLK firms.

Panel	Intercept	$\ln(v_{t-1})$	$\ln(v_{i,t-1}) \times I(r_{i,t-1}<0)$	Messages $\ln(1+M_{i,t-1})$	Trades $\ln(N_{i,t-1})$	Agreement $A_{i,t-1}$	R^2
All	0.065 (1.16)	-0.007 (.750)	0.005^c (7.08)	0.039^c (8.82)		-0.010 (1.01)	0.118
DIA	-0.098 (1.38)	-0.039^c (3.48)	0.003^c (3.85)	0.023^c (4.63)		-0.011 (.990)	0.078
XLK	0.396^c (3.82)	0.039^a (2.35)	0.011^c (7.30)	0.025^b (2.65)		0.002 (.076)	0.149
All	-0.510^c (3.80)	-0.033^b (2.98)	0.005^c (7.67)		0.073^c (6.51)	-0.003 (.248)	0.052
DIA	-0.433^a (2.16)	-0.052^c (3.37)	0.003^c (4.15)		0.044^b (2.62)	-0.007 (.672)	0.074
XLK	0.486 (1.78)	0.050 (1.90)	0.011^c (7.59)		0.008 (.487)	0.004 (.579)	0.138
All	-0.386^b (2.85)	-0.033^b (2.92)	0.005^c (7.29)	0.033^c (6.83)	0.043^c (3.62)	-0.009 (.893)	0.086
DIA	-0.689 (1.80)	-0.069^c (3.44)	0.003^c (4.04)	0.022^c (3.34)	0.056 (1.56)	-0.011 (1.03)	0.076
XLK	0.515^c (4.04)	0.047^b (2.78)	0.011^c (7.21)	0.025^b (2.84)	-0.010 (1.20)	-0.004 (.187)	0.152

Trading Volume

- There is a strong *contemporaneous* link between trading volume and message board activity.
- But does message board activity *predict* trading volume?
- We estimate a volume model along the lines of Chordia, Roll, and Subrahmanyam (2001)
- We analyze three measures of daily change in volume
  - Δ \$-Volume: change in \$-value of traded shares
  - Δ Small: change in number of small (<\$100k) trades
  - Δ Large: change in number of large (>\$1m) trades
- We control for: *posting activity and disagreement* shortly before market opening; individual stock momentum; market momentum; 5-day volatility; day of week effects; an alternative news channel (Wall Street Journal)
- What do we find?
  - Message board posting volume 4 hours prior to market opening helps predict day's trading volume
  - No effect when considering longer aggregation periods.

Trading Volume: Results

Dependent Variable	Δ $-Volume	Δ Small	Δ Large
Stock Up Yesterday	-1.020^{c} (7.12)	-1.167^{c} (10.7)	-1.034^{c} (4.46)
Stock Down Yesterday	-0.468^{c} (4.20)	-0.335^{c} (3.96)	-0.419^{a} (2.32)
Stock Last 5 Days Up	-0.589^{c} (6.54)	-0.415^{c} (5.86)	-0.522^{c} (3.53)
Stock Last 5 Days Down	-0.188^{c} (3.60)	-0.104^{a} (2.23)	-0.191^{a} (2.11)
Stock 5 Day Volatility	-0.492^{b} (2.87)	-0.346 (1.27)	-0.109 (.254)
Market Up Yesterday	0.027 (.092)	-0.244 (1.13)	-0.007 (.015)
Market Down Yesterday	0.049 (.530)	0.071 (1.03)	0.082 (.555)
Market Last 5 Days Up	-0.255 (.850)	0.059 (.265)	-0.187 (.390)
Market Last 5 Days Down	-0.284^{c} (3.58)	-0.118^{a} (1.97)	-0.411^{b} (3.22)
Market 5 Day Volatility	0.190 (1.84)	0.089 (1.07)	0.060 (.352)
Federal Funds Rate	-41.15^{c} (3.73)	-45.97^{c} (5.59)	-53.74^{b} (3.04)
Term Spread	-19.88 (1.88)	-19.17^{a} (2.43)	-36.20^{a} (2.13)
Quality Spread	-17.76 (1.13)	-1.569 (.134)	-24.83 (.987)
Message Board Volume	0.012^{b} (3.05)	0.009^{b} (2.91)	0.016^{a} (2.55)
Agreement Index	-0.015 (1.48)	-0.010 (1.36)	-0.027 (1.71)
Intercept	0.005 (.432)	-0.016 (1.32)	-0.006 (.285)
Monday	0.010 (.719)	0.111^{c} (10.3)	-0.029 (1.23)
Tuesday	0.108^{c} (7.88)	0.062^{c} (6.02)	0.130^{c} (5.92)
Wednesday	0.096^{c} (7.14)	0.085^{c} (8.43)	0.115^{c} (5.30)
Thursday	0.026 (1.90)	0.034^{c} (3.29)	0.040 (1.80)
Holiday	0.017 (.694)	0.088^{c} (4.71)	-0.042 (1.04)
Articles in WSJ Today	0.016 (1.89)	0.013^{a} (2.00)	0.028^{a} (2.01)
Articles in WSJ Yesterday	-0.044^{c} (5.10)	-0.029^{c} (4.49)	-0.073^{c} (5.29)
Articles in WSJ 2 days ago	-0.013 (1.53)	-0.011 (1.71)	-0.000 (.009)
σ^2 (mean squared error)	0.196^{c} (74.2)	0.109^{c} (73.3)	0.504^{c} (73.8)
ρ (cross-section variance ratio)	0.000 (.000)	0.000 (.000)	0.000 (.001)
Observations	10,973	10,973	10,973
Companies	45	45	45
Pseudo-R^2	0.203	0.228	0.145

Conclusions - Short Horizon Analysis

- There is useful information present on the stock message boards. All that talk is *not* just noise.
- Message boards do not successfully predict stock returns at short time horizons.
- Support for Harris and Raviv (1993) that disagreement induces trading.
- Message posting helps predict volatility both at daily frequency and within the trading day.
- Stock messages reflect public information rapidly.
  - stock messages may be helpful in studies of insider trading and event studies
  - since the messages are time-stamped to the minute, they may also prove quite helpful in market microstructure studies.
- Caveat: automated text classification still problematic in an environment with low signal-to-noise ratio

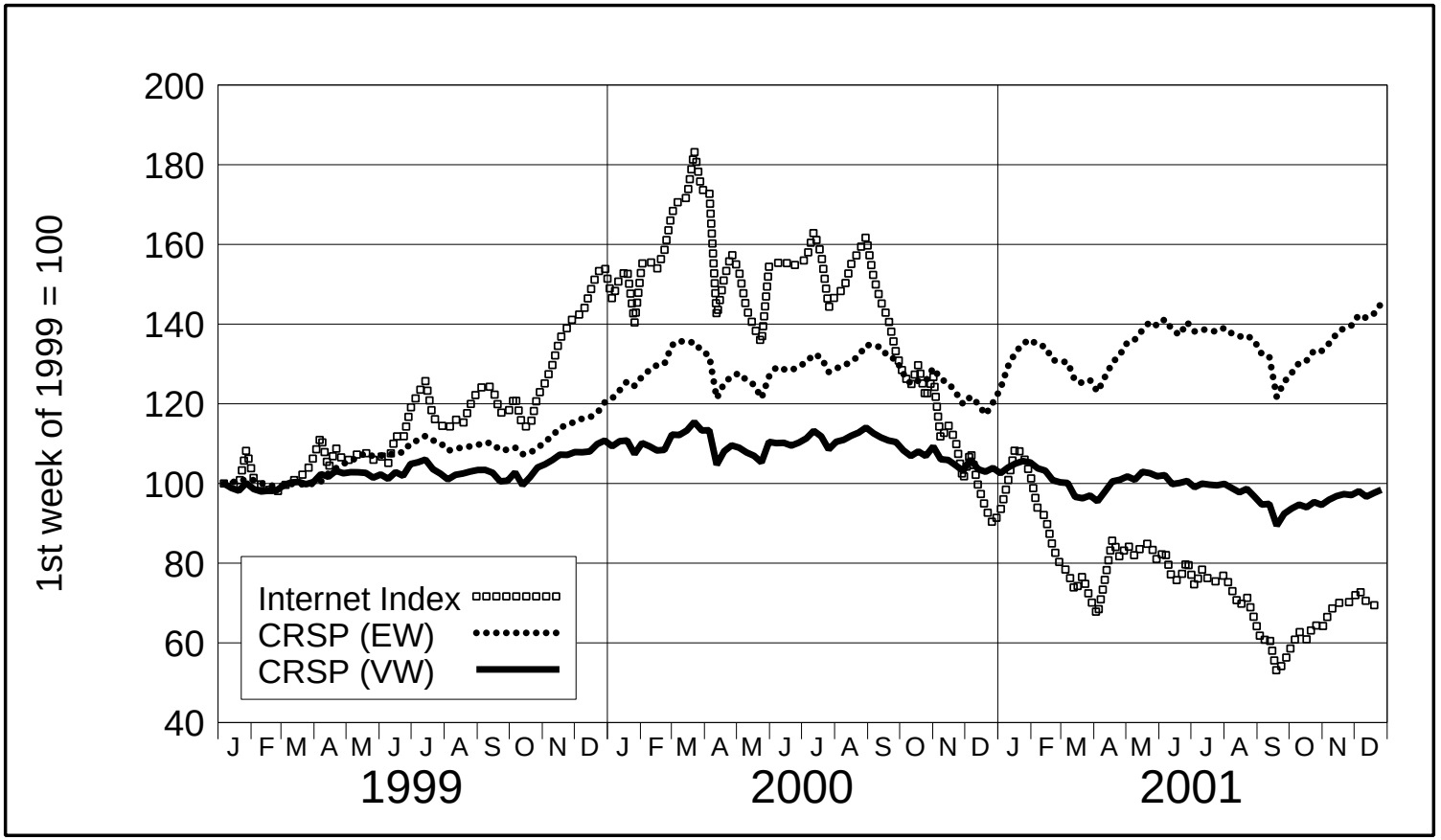
Stock Returns and Stock Message Boards

- What about using a longer time horizon?
- Why not look at all Yahoo-discussed companies?
- So we look at
  - 36 months of data between 1999-2001
  - more than 5,000 companies
  - 35 million messages from Yahoo
    (but posting volume only; no classification)
  - monthly CRSP data of returns, market caps, volatility
- Do stock message boards predict stock returns at monthly frequency?
- Do *noisy* companies do better than *quiet* companies?
- Is posting volume a measure for the riskiness of a stock or a (non-diversifiable) "market factor" along the lines of CAPM beta and Fama-French factors?

Research Strategy

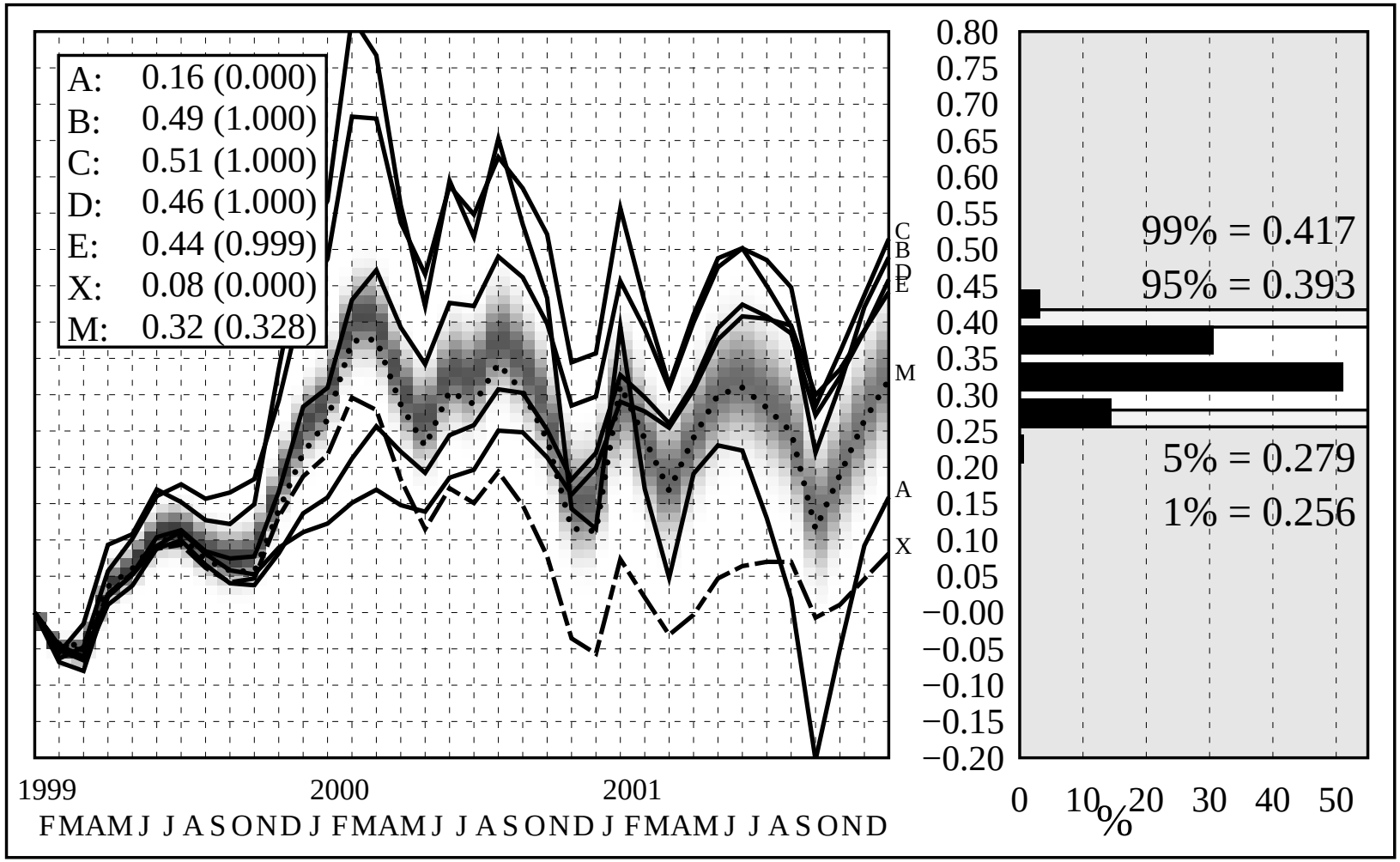
- Sort stocks into quintile portfolios based on their Yahoo posting volume
- Calculate returns of these portfolios
- Obtain P-values for these portfolio returns through bootstrap (10,000 replications of random portfolios of appropriate size)
- Look at double-sorted portfolios (posting volume, market cap)
- Financial econometrics: market factor analysis

Market Performance: XLK vs. CRSP (equal/value-weighted)



- Two phases: internet boom (until August 2000) and bust (after August 2000)
- Equal-weighted market portfolio has a markedly higher return compared

Equal-Weighted Quintile Portfolio Returns and P-Values



- Portfolios A-E ranked by Yahoo posting volume
- Portfolio X contains stocks not discussed on Yahoo
- Dotte M line is the CRSP value-weighted index

Double-Sorted Portfolio Returns: Posting Volume & Market Cap

Portfolio	Market Cap			all	Posting Volume
	high	medium	low		
A (high)	-0.04 (0.000)	-0.15 (0.000)	0.56 (1.000)	0.16 (0.003)	924.5
B	0.12 (0.000)	0.52 (0.998)	0.71 (1.000)	0.49 (0.990)	79.32
C	0.13 (0.001)	0.57 (1.000)	0.77 (1.000)	0.52 (0.997)	23.81
D	0.30 (0.282)	0.37 (0.711)	0.67 (1.000)	0.46 (0.971)	7.19
E (low)	0.36 (0.681)	0.29 (0.272)	0.63 (1.000)	0.44 (0.945)	1.19
X (none)	0.14 (0.001)	-0.08 (0.000)	0.16 (0.003)	0.08 (0.000)	0.00
all	0.16 (0.004)	0.35 (0.572)	0.47 (0.979)	0.34 (0.515)	
Market Cap	7.37	0.21	0.04		

- Large companies tend to have higher returns than small companies. This reflects the conventional risk-return trade-off as small companies are more volatile.
- Noisy stocks tend to underperfom the market
- Quiet stocks tend to outperform the market

Double-Sorted Portfolio Volatilities

Portfolio	Market Cap			all	Posting Volume
	high	medium	low		
A (high)	8.27 (1.000)	10.75 (1.000)	11.70 (1.000)	10.24 (1.000)	924.5
B	6.56 (0.000)	9.12 (1.000)	10.85 (1.000)	8.85 (1.000)	79.32
C	5.77 (0.000)	7.75 (0.673)	9.94 (1.000)	7.82 (0.982)	23.81
D	5.30 (0.000)	6.33 (0.000)	8.70 (1.000)	6.78 (0.000)	7.19
E (low)	5.34 (0.000)	5.50 (0.000)	7.83 (0.992)	6.22 (0.000)	1.19
X (none)	6.27 (0.000)	6.54 (0.000)	7.87 (0.999)	6.89 (0.000)	0.00
all	6.88 (0.000)	7.66 (0.050)	8.64 (1.000)	7.73 (0.502)	
Market Cap	7.37	0.21	0.04		

- Large companies tend to have lower volatility than small companies.
- Noisy stocks have high volatility
- Quiet stocks have low volatility

Market Factors

- CAPM Beta measure the correlation between excess individual return $r_{i,t} - r_t^f$ and excess market return $r_t^m - r_t^f$.
- Fama and French (1993) Factors measure differential portfolio returns and capture risk factors priced by the market.
  - **SMB** (small-minus-big): company size differential portfolio based on companies' market capitalization
  - **HML** (high-minus-low): value vs. growth differential portfolios based on companies' book-to-market ratio
  - **UMD** (up-minus-down): market momentum differential portfolios
- **LIQ** is the Pástor and Stambaugh (2002) measure of market liquidity risk that is captured by a measure of signed "order-flow"; reconstructed by us
- **NMQ** (noise-minus-quiet) measures the differential portfolio return of the top quintile minus the bottom quintile based on Yahoo posting volume; constructed by us

Estimating equation

$$\begin{aligned} r_{i,t} - r_t^f \;=\; & \alpha + \beta(r_t^m - r_t^f) + \gamma_1 \,\mathrm{SMB}_t + \gamma_2 \,\mathrm{HML}_t \\ & + \gamma_3 \,\mathrm{UMD}_t + \gamma_4 \,\mathrm{LIQ}_t + \gamma_5 \,\mathrm{NMQ}_t + \mu_i + \epsilon_{it} \end{aligned}$$

- Estimated as an $N \times T$ panel with company fixed effects μ_t and idiosyncratic effect ϵ_{it}

Results

Line	Model	α	β	γ_1 (SMB)	γ_2 (HML)	γ_3 (UMD)	γ_4 (LIQ)	γ_5 (NMQ)
				Panel A: Conventional Models				
1	4-factor	1.134^c	0.868^c	0.693^c	0.001	-0.303^c		
		(19.43)	(56.36)	(56.90)	(0.058)	(42.47)		
2	Pástor-Stambaugh	1.103^c	0.870^c	0.702^c	0.001	-0.306^c	-5.209^c	
		(18.70)	(55.73)	(56.29)	(0.055)	(42.65)	(10.04)	
				Panel B: Inclusion of NMQ Factor				
3	4-factor	0.975^c	0.569^c	0.540^c	0.457^c	-0.238^c		0.437^c
		(16.27)	(28.14)	(38.19)	(17.95)	(30.96)		(23.03)
4	Pástor-Stambaugh	0.966^c	0.555^c	0.537^c	0.464^c	-0.238^c	-6.124^c	0.455^c
		(15.96)	(27.12)	(37.23)	(18.08)	(30.80)	(11.45)	(23.79)
				Panel C: Robustness checks (baseline model: line 8)				
5	1999/01–2000/07	0.703^c	0.720^c	0.517^c	0.560^c	-0.156^c	-0.327	0.407^c
		(8.766)	(26.92)	(27.72)	(13.28)	(7.231)	(0.376)	(15.97)
6	2000/08–2001/12	0.727^c	0.250^c	0.843^c	0.568^c	-0.170^c	-11.04^c	0.641^c
		(6.855)	(7.018)	(28.89)	(15.25)	(11.58)	(14.72)	(18.72)
7	Top Yahoo Stocks	4.192^c	0.337^a	0.400^c	0.130	-0.226^c	-3.259	1.410^c
		(10.04)	(2.377)	(4.002)	(0.746)	(4.318)	(0.894)	(10.71)

- NMQ factor is positive and significant, and is robust to variations in sample period
- NMQ intensity is larger for top 250 stocks on Yahoo

Conclusions - Long Horizon Analysis

- Highly-discussed ("noisy") stocks have poor returns and high volatility
- Profitable trading strategy: go short on noisy stocks
  ... but such a portfolio is high risk
- Sorting out between competing hypotheses:
  - pump-and-dump scams: low returns would be concentrated in small-cap stocks — no
  - differences of opinion, plus institutional difficulties for short selling — yes
  - people post messages about risky stocks to alleviate investment anxiety — yes