

Using High-Frequency Transaction Data to Estimate the Probability of Informed Trading

Anthony Tay, Christopher Ting, Yiu Kuen Tse and Mitch Warachka
Singapore Management University

23 February, 2009

First Singapore Conference on Quantitative Finance

Abstract

- This paper applies the Asymmetric Autoregressive Conditional Duration (AACD) model to estimate the probability of informed trading (PIN).
- We model trade direction (buy versus sell orders) and the duration between trades jointly.
- Extending the Easley, Hvidkjaer and O'Hara (2002) approach, which uses the aggregate numbers of daily buy and sell orders to estimate PIN, we use transaction data.
- We allow for interactions between consecutive buy-sell orders.
- We account for the duration between trades and the volume of trade.

- We allow the probabilities of good news and bad news to vary each day.
- Our PIN estimates can be computed daily as well as over intraday intervals.

Outline

- Review of Probability of Informed Trading (PIN)
- Review of the ACD and AACD Model
- AACD model of buy and sell orders
- Model with time varying probability of news
- Data
- Empirical results

Probability of Informed Trading

- Easley, Hvidkjaer and O'Hara (2002) (EHO) uses the aggregate numbers of daily buy and sell orders to estimate PIN.
- Each trading day is characterized by good news (G), no news (N) and bad news (B) to form the set $S = \{G, N, B\}$.
- Let the probability of a day containing news be θ_E .
- Conditional on the arrival of news, we denote the probability of bad news by θ_B .
- The aggregate numbers of buy and sell orders on a trading day follow independent Poisson distributions, where the intensities of sell and

buy orders on a no-news day, denoted by λ_{-1} and λ_1 , respectively, are constant throughout the sample period.

- On a good-news day, the buy intensity increases by a positive amount δ , with no change in the sell intensity. Likewise, on a bad-news day, the sell intensity increases by δ while the buy intensity remains unchanged.
- With D days of data, the mixture-of-distributions assumption implies the likelihood function equals

$$\prod_{d=1}^D \left[(1 - \theta_E) \frac{\lambda_1^{B_d} e^{-\lambda_1}}{B_d!} \frac{\lambda_{-1}^{S_d} e^{-\lambda_{-1}}}{S_d!} + \theta_E \theta_B \frac{\lambda_1^{B_d} e^{-\lambda_1}}{B_d!} \frac{(\lambda_{-1} + \delta)^{S_d} e^{-(\lambda_{-1} + \delta)}}{S_d!} \right. \\ \left. + \theta_E (1 - \theta_B) \frac{(\lambda_1 + \delta)^{B_d} e^{-(\lambda_1 + \delta)}}{B_d!} \frac{\lambda_{-1}^{S_d} e^{-\lambda_{-1}}}{S_d!} \right]$$

where B_d and S_d are the respective aggregate number of buy and sell orders on day d .

- From this model, EHO estimate PIN as

$$\text{PIN} = \frac{\theta_E \delta}{\theta_E \delta + \lambda_{-1} + \lambda_1}.$$

- In other words,

$$\text{PIN} = \frac{\text{Expected number of trades per day initiated by informed traders}}{\text{Expected total number of trades per day}}.$$

- Some criticisms

1. PIN is assumed to be time invariant.
2. Volume is ignored in the estimation of PIN.

3. Applications typically estimate PIN (assuming to be constant) over several months. Recently, some studies highlight the importance of high-frequency PIN.
4. Easley, Engle, O'Hara and Wu (2008) propose a bivariate time series approach for the estimation of PIN daily.

Asymmetric Autoregressive Conditional Duration (AACD) Model

- AACD is an extension of the autoregressive conditional duration (ACD) model of Engle and Russell (1998) and Engle (2000).
- Let y_i denote the trade direction of the i th trade at time t_i , which may take values of $j = -1$ or 1 representing a sell-initiated and buy-initiated trade, respectively.
- Conditional on the information set Φ_{i-1} after the $(i-1)$ th trade, the inter-arrival time random variables of the latent processes follows an exponential distribution with mean (conditional expected duration) ψ_{ji} , where j denotes the trade direction and i denotes the trade at time t_i .

- We denote $\lambda_{ji} = 1/\psi_{ji}$, which is the *intensity* of the latent Poisson process.
- Denote $x_i = t_i - t_{i-1}$ as the duration of trade, the log-likelihood function may be written as

$$\sum_{i=1}^N \log p_i(x_i, y_i | \Phi_{i-1}) = - \sum_{i=1}^N \left[\sum_{j=-1,1} \frac{x_i}{\psi_{ji}} - \log \left(\sum_{j=-1,1} \frac{D_{y_i}(j)}{\psi_{ji}} \right) \right],$$

where $D_{y_i}(j) = 1$, if $j = y_i$ and 0 otherwise.

- The parameters of the model can be estimated using maximum likelihood estimation (MLE) once the functional forms of the conditional expected durations ψ_{ji} are specified.

- We adopt the ACD model of Engle and Russel (1998)

$$\log \psi_{ji} = \nu_{j,-1} D_{-1}(y_{i-1}) + \nu_{j1} D_1(y_{i-1}) + \alpha_j \log \psi_{j,i-1} + \beta_j \log x_{i-1}, \quad j = -1, 1.$$

- We now specify the conditional expected duration equation as being dependent on the state of good news, no news or bad news.
- First, we define the following function f_{ji}^s as the basis of the equations for the conditional expected duration in each of the three states in S ,

$$f_{ji}^s \equiv \nu_{j,-1} D_{-1}(y_{i-1}) + \nu_{j1} D_1(y_{i-1}) + \alpha_j \log \psi_{j,i-1}^s + \beta_j \log x_{i-1} + \varsigma_j y_{i-1} \log v_{i-1},$$

for $j = -1, 1$ and $s \in S$, where v_{i-1} is the volume of the trade at time t_{i-1} .

- For a no-news day ($s = N$), we assume the basic logarithmic conditional expected duration holds.
- For the buy-orders ($j = 1$) on a good-news day ($s = G$), we reduce f_{1i}^N by a positive constant μ to yield the following logarithmic conditional expected duration

$$\log \psi_{1i}^G = f_{1i}^G - \mu,$$

while the logarithmic conditional expected duration for a sell trade is the basis function $f_{-1,i}^G$,

$$\log \psi_{-1,i}^G = f_{-1,i}^G.$$

- Conversely, on a bad-news day ($s = B$), we have

$$\log \psi_{1i}^B = f_{1i}^B,$$

and

$$\log \psi_{-1,i}^B = f_{-1,i}^B - \mu.$$

- Let $N_d = S_d + B_d$ denote the number of trades on day d . The likelihood function is then given by

$$\prod_{d=1}^D \left[\sum_{s \in S} \pi_s \left(\prod_{i=1}^{N_d} p_{si}(x_i, y_i | \Phi_{i-1}) \right) \right].$$

- We compute PIN as the ratio of the total expected number of trades due to informed traders to the total expected number of all trades over all trading intervals, i.e.,

$$\text{PIN} = \frac{\sum_{d=1}^D \sum_{i=1}^{N_d} \left(\pi_G \lambda_{1i}^G + \pi_B \lambda_{-1,i}^B \right) x_i}{\sum_{d=1}^D \sum_{i=1}^{N_d} \left(\lambda_{-1,i}^N + \lambda_{1i}^N + \pi_G \lambda_{1i}^G + \pi_B \lambda_{-1,i}^B \right) x_i},$$

where the index d for the intensities and the data is suppressed.

- Denoting PIN_d as PIN on day d , we can estimate PIN_d by

$$\text{PIN}_d = \frac{\sum_{i=1}^{N_d} \left(\pi_G \lambda_{1i}^G + \pi_B \lambda_{-1,i}^B \right) x_i}{\sum_{i=1}^{N_d} \left(\lambda_{-1,i}^N + \lambda_{1i}^N + \pi_G \lambda_{1i}^G + \pi_B \lambda_{-1,i}^B \right) x_i},$$

where the data (x_i, y_i) and the estimated parameters $\lambda_{-1,i}^N$, λ_{1i}^N , λ_{1i}^G and $\lambda_{-1,i}^B$ pertain to day d .

Time-Varying Probabilities of News

- We assume a logistic model in which the arrival of good news, no news and bad news on day d depends on the aggregate *volume* of buy and sell orders.
- We denote \bar{V}^B as the average number of lots traded per day initiated by buy orders, and \bar{V}^S as the average number of lots traded per day initiated by sell orders.
- The number of lots traded on day d initiated by buy and sell orders are denoted V_d^B and V_d^S , respectively. We then assume the probability of no news on day d to be

$$\pi_{Nd} = 1 - \theta_{Ed} = \frac{1}{1 + \exp \left\{ \delta_1 + \delta_2 \left[\log(V_d^B + V_d^S) - \log(\bar{V}^B + \bar{V}^S) \right] \right\}}.$$

- We expect $\delta_2 > 0$, so that when the aggregate volume on day d , $V_d^B + V_d^S$, increases relative to the daily average volume $\bar{V}^B + \bar{V}^S$, the probability of no news decreases.

- Given news on day d , the probability of good news is assumed to be

$$\theta_{Gd} = \frac{1}{1 + \exp \left[\delta_3 (\log V_d^S - \log \bar{V}^S) - \delta_4 (\log V_d^B - \log \bar{V}^B) \right]}.$$

- We expect δ_3 and δ_4 to be positive, so that $V_d^S > \bar{V}^S$ or $V_d^B < \bar{V}^B$ imply a decreased probability of good news.
- If $V_d^S = \bar{V}^S$ and $V_d^B = \bar{V}^B$, then the probability of good news and bad news, given there is news, equals one-half.

- To compute PIN on day d , we use the formula

$$\text{PIN}_d = \frac{\sum_{i=1}^{N_d} \left(\pi_{Gd} \lambda_{1i}^G + \pi_{Bd} \lambda_{-1,i}^B \right) x_i}{\sum_{i=1}^{N_d} \left(\lambda_{-1,i}^N + \lambda_{1i}^N + \pi_{Gd} \lambda_{1i}^G + \pi_{Bd} \lambda_{-1,i}^B \right) x_i},$$

with time varying π_G and π_B .

- This formula can be used to compute PIN over intraday intervals.

Data and Empirical Results

- We use intraday data of five NYSE companies: Boeing (BA), General Electric (GE), International Business Machines (IBM), Altria Group (formerly Philip Morris) (MO), and AT&T (T).
- The data are obtained from the TAQ database for July 1, 1994 to June 30, 1995.
- The average number of trades per day varies from a low of 243.3 (BA) to a high of 677.9 (GE).

Table 1. Summary Statistics of Duration and Trade Direction

Statistics	Ticker Symbols				
	BA	GE	IBM	MO	T
	Average Diurnally Adjusted Duration (in seconds)				
All Trades \bar{x}	88.78	31.83	41.42	48.88	39.29
Buy-initiated Trades $\hat{\psi}_1$	197.86	55.23	86.31	110.29	72.29
Sell-initiated Trades $\hat{\psi}_{-1}$	161.04	75.12	79.64	87.79	86.07
	Order-Flow Statistics (volume in lots)				
Frequency of Buys (%)	44.87	57.63	47.99	44.32	54.35
Frequency of Sells (%)	55.13	42.37	52.01	55.68	45.65
Serial Correlation of Trade Direction	0.35	0.32	0.52	0.32	0.40
Runs Test of Trade Direction	-81.32	-132.56	-186.27	-105.77	-146.61
Average Volume (lot size)	27.80	19.91	30.83	31.48	25.31
Average Log Volume	1.97	1.70	2.36	2.13	1.61
Average Daily Number of Trades	243.30	677.90	521.10	442.30	549.10
Average Daily Number of Buy-Trades	109.17	390.67	250.08	196.03	298.44
Average Daily Number of Sell-Trades	134.13	287.23	271.02	246.27	250.66
Number of Observations in Sample	54,500	170,157	129,239	110,120	135,087

Table 2. Estimates of PIN-EHO Model

Variables	Parameters	Ticker Symbols				
		BA	GE	IBM	MO	T
Intensity for Sell-Initiated Trade	λ_{-1}	98.7929 (3.2689)	345.8723 (6.5457)	200.8416 (6.4490)	175.2851 (5.1569)	235.5643 (4.8911)
Intensity for Buy-Initiated Trade	λ_1	110.9957 (3.6987)	269.7381 (4.7169)	251.5821 (6.6138)	227.1735 (5.9444)	241.7717 (5.6469)
Adjustment for Information	δ	92.6196 (6.8988)	134.4572 (8.6544)	148.5920 (8.9076)	144.5634 (15.4491)	200.0465 (17.5005)
Probability of News	θ_E	0.3511 (0.0366)	0.4560 (0.0426)	0.4556 (0.0347)	0.2683 (0.0350)	0.3539 (0.0358)
Given News, Probability of Bad News	θ_B	0.6859 (0.0758)	0.2870 (0.0611)	0.2858 (0.0684)	0.4443 (0.1010)	0.1433 (0.0555)
PIN-EHO		0.1342	0.0906	0.1302	0.0879	0.1292

Figures in parentheses are standard errors.

Table 3. Estimates of the PIN-AACD Model with constant probabilities of good news, no news and bad news

Trade Variables	Parameters	Ticker Symbols				
		BA	GE	IBM	MO	T
Sale after Sale	$v_{-1,-1}$	1.8791 (0.2475)	2.3126 (0.0835)	2.9647 (0.1096)	1.7877 (0.1611)	1.6118 (0.0902)
Sale after Buy	$v_{-1,1}$	2.1772 (0.2638)	2.6272 (0.0885)	3.8651 (0.1306)	1.9956 (0.1747)	2.0551 (0.1067)
Buy after Sale	$v_{1,-1}$	1.7056 (0.2711)	1.7832 (0.0622)	2.8227 (0.1701)	1.6843 (0.1685)	2.1341 (0.0890)
Buy after Buy	v_{11}	1.4295 (0.2430)	1.6645 (0.0618)	2.1598 (0.1432)	1.4336 (0.1564)	1.6809 (0.0771)
Conditional Duration for Sales	α_{-1}	0.5597 (0.0445)	0.3946 (0.0171)	0.1860 (0.0195)	0.5079 (0.0328)	0.5080 (0.0207)
Lagged Duration for Sales	β_{-1}	0.0875 (0.0054)	0.0722 (0.0040)	0.1064 (0.0067)	0.1181 (0.0058)	0.1302 (0.0040)
Conditional Duration for Buys	α_1	0.6181 (0.0455)	0.5409 (0.0137)	0.3776 (0.0252)	0.5722 (0.0291)	0.5048 (0.0171)
Lagged Duration for Buys	β_1	0.1233 (0.0059)	0.0698 (0.0031)	0.1400 (0.0072)	0.1473 (0.0069)	0.1169 (0.0040)
Adjustment for Information	μ	0.5167 (0.0275)	0.2692 (0.0137)	0.3854 (0.0205)	0.4351 (0.0340)	0.4703 (0.0300)
Probability of News	θ_E	0.3223 (0.0388)	0.3848 (0.0383)	0.3400 (0.0386)	0.2070 (0.0304)	0.2730 (0.0425)
Given News, Probability of Bad News	θ_B	0.9565 (0.0411)	0.3893 (0.1213)	0.3754 (0.2395)	0.7276 (0.1147)	0.0288 (0.0252)
Volume - Direction for Sales	ς_{-1}	0.0363 (0.0051)	0.0951 (0.0036)	0.0201 (0.0036)	0.0498 (0.0037)	0.0397 (0.0032)
Volume - Direction for Buys	ς_1	-0.0466 (0.0055)	-0.0959 (0.0037)	-0.0421 (0.0035)	-0.0629 (0.0044)	-0.0373 (0.0032)
PIN-AACD (from equation (19))		0.1485	0.0413	0.0816	0.1011	0.1241
Daily PIN_d (from equation (20))						
Minimum		0.1336	0.0393	0.0731	0.0887	0.1082
Maximum		0.1617	0.0423	0.0890	0.1095	0.1372
Mean		0.1483	0.0411	0.0816	0.1009	0.1240

Figures in parentheses are standard errors.

Table 4. Estimates of the PIN-AACD Model with varying probabilities of good news, no news and bad news

Trade Variables	Parameters	Ticker Symbols				
		BA	GE	IBM	MO	T
Sale after Sale	$v_{-1,-1}$	1.7528	2.3227	2.9234	1.6651	1.6234
Sale after Buy	$v_{-1,1}$	2.0474	2.6395	3.8171	1.8652	2.0689
Buy after Sale	$v_{1,-1}$	1.8132	1.7785	2.7992	1.7993	2.1314
Buy after Buy	v_{11}	1.5293	1.6597	2.1413	1.5425	1.6770
Conditional Duration for Sales	α_{-1}	0.5794	0.3930	0.1953	0.5329	0.5059
Lagged Duration for Sales	β_{-1}	0.0903	0.0716	0.1074	0.1198	0.1296
Conditional Duration for Buys	α_1	0.6007	0.5416	0.3817	0.5505	0.5042
Lagged Duration for Buys	β_1	0.1221	0.0702	0.1406	0.1466	0.1175
Adjustment for Information	μ	0.4934	0.2677	0.3804	0.4251	0.4686
Volume - Direction for Sales	ς_{-1}	0.0338	0.0948	0.0199	0.0477	0.0395
Volume - Direction for Buys	ς_1	-0.0483	-0.0959	-0.0420	-0.0655	-0.0372
Prob equation, coefficient 1	δ_1	-0.6764	-0.5004	-0.0217	-1.0131	-0.7969
Prob equation, coefficient 2	δ_2	2.3823	1.3632	1.9046	0.1505	0.7540
Prob equation, coefficient 3	δ_3	1.4445	0.5441	0.0052	0.0650	-0.0067
Prob equation, coefficient 4	δ_4	0.0234	0.5631	1.0908	0.0547	0.0016
Probability of good news						
Minimum		0.0129	0.0759	0.0148	0.1137	0.0708
Maximum		0.3655	0.3558	0.7735	0.1585	0.2930
Mean		0.1422	0.1800	0.2274	0.1315	0.1463
Std Dev		0.0698	0.0546	0.1339	0.0063	0.0378
Probability of no news						
Minimum		0.1093	0.3352	0.0806	0.6842	0.4171
Maximum		0.9855	0.8313	0.9181	0.7705	0.8577
Mean		0.6925	0.6401	0.5478	0.7369	0.7074
Std Dev		0.2053	0.1017	0.1653	0.0123	0.0751
Probability of bad news						
Minimum		0.0007	0.0764	0.0671	0.1158	0.0715
Maximum		0.7273	0.3770	0.3760	0.1573	0.2899
Mean		0.1653	0.1799	0.2249	0.1315	0.1464
Std Dev		0.1591	0.0584	0.0506	0.0064	0.0373
Daily PIN_d (from equation (24))						
Minimum		0.0109	0.0989	0.0650	0.1496	0.1021
Maximum		0.4261	0.3020	0.4258	0.1945	0.3179
Mean		0.1883	0.1889	0.2452	0.1675	0.1876
Std Dev		0.1022	0.0428	0.0676	0.0065	0.0386

Table 5. Daily correlations between PIN and return volatility

Correlation	Ticker Symbols				
	BA	GE	IBM	MO	T
PIN with varying probability of news					
Corr(PIN _{d-1} , V _d)	0.1938	0.1746	0.1175	0.0856	0.3564
Corr(PIN _d , V _d)	0.6001	0.5158	0.5248	0.4448	0.6722
PIN with constant probability of news					
Corr(PIN _{d-1} , V _d)	-0.0772	0.0759	0.0492	-0.0936	-0.0114
Corr(PIN _d , V _d)	0.2362	-0.0042	0.2077	0.1143	0.0517

V_d denotes the integrated volatility on day d .

Fig 1: Prob of Good News, No News, Bad News and PIN of BA Data

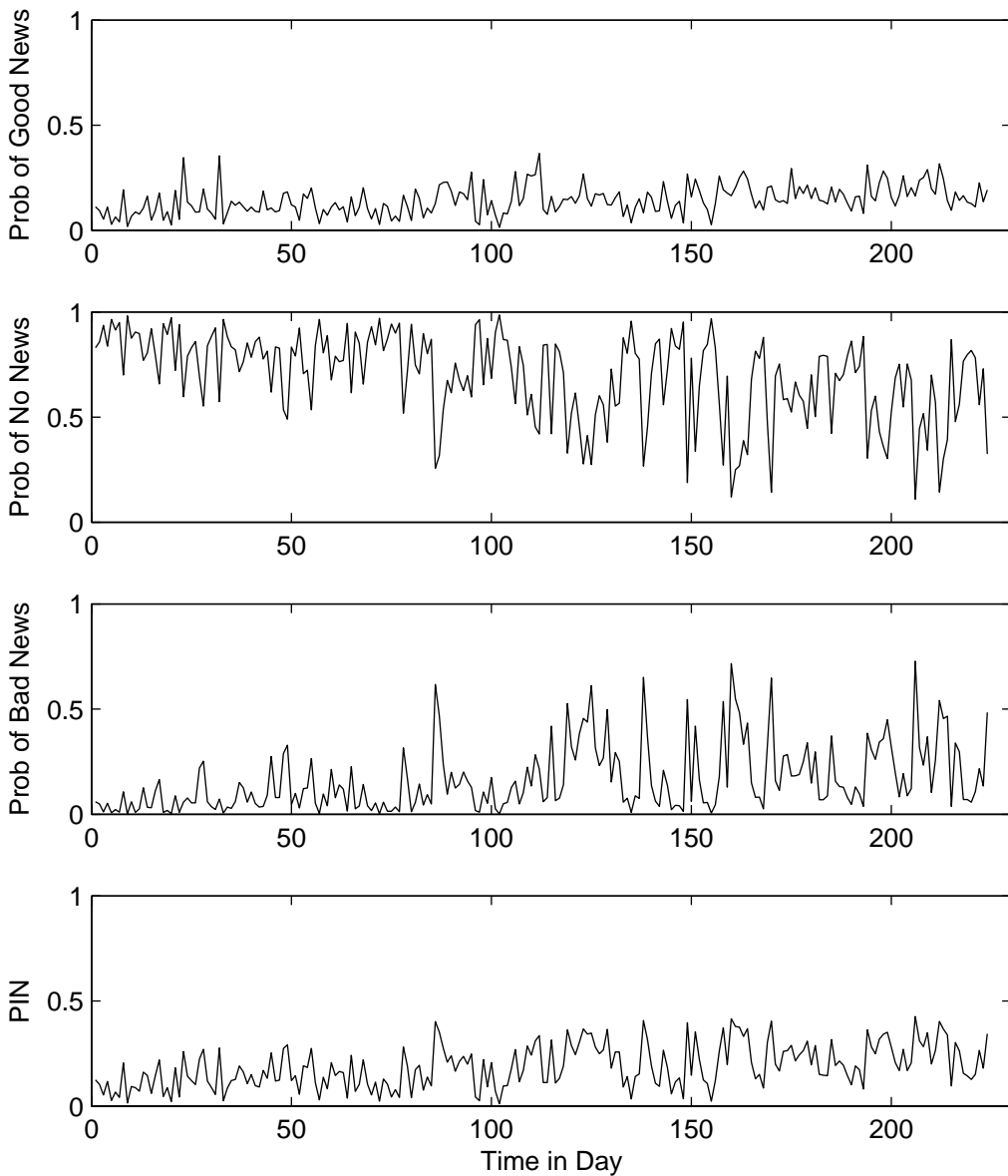


Fig 2: Prob of Good News, No News, Bad News and PIN of GE Data

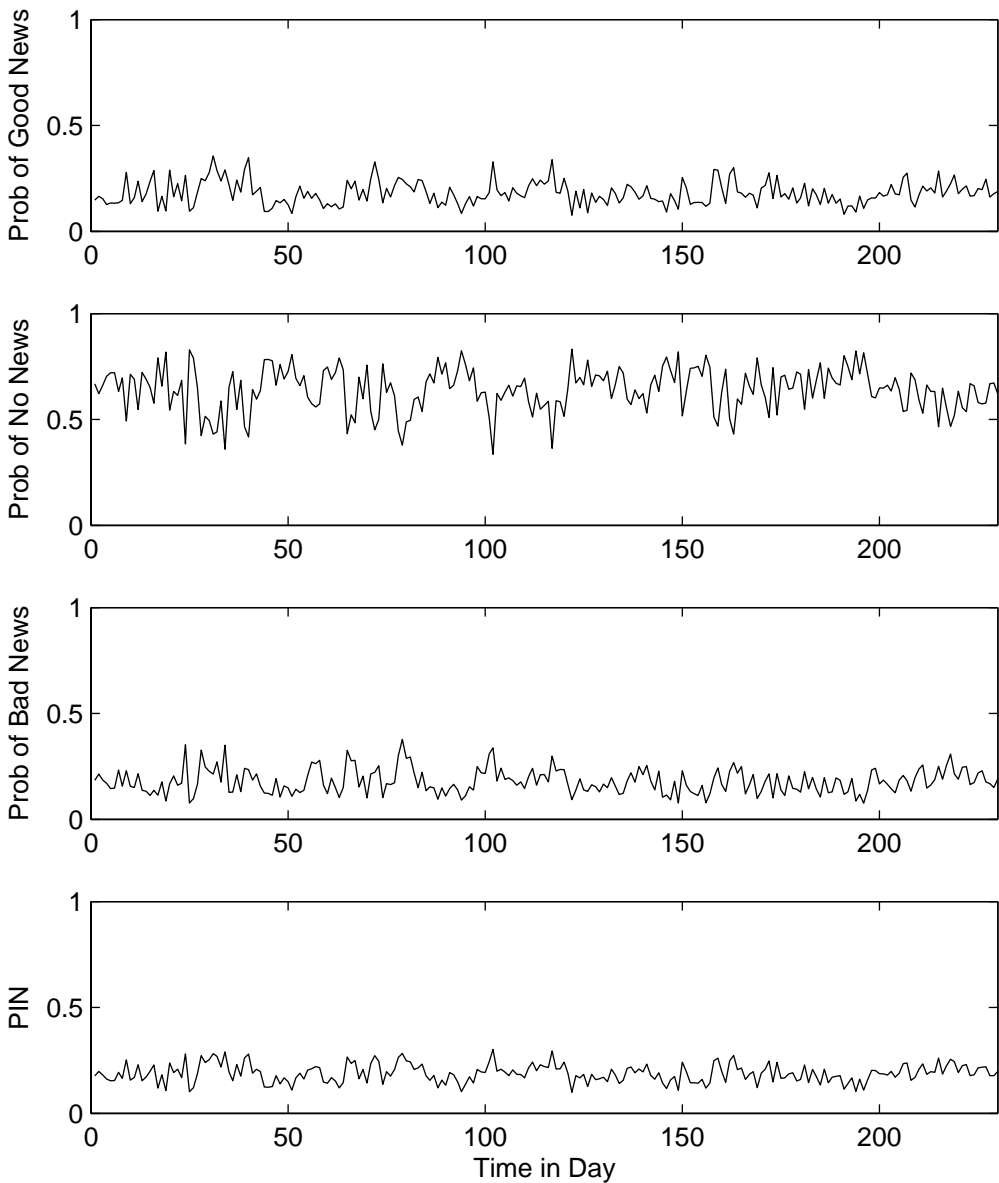
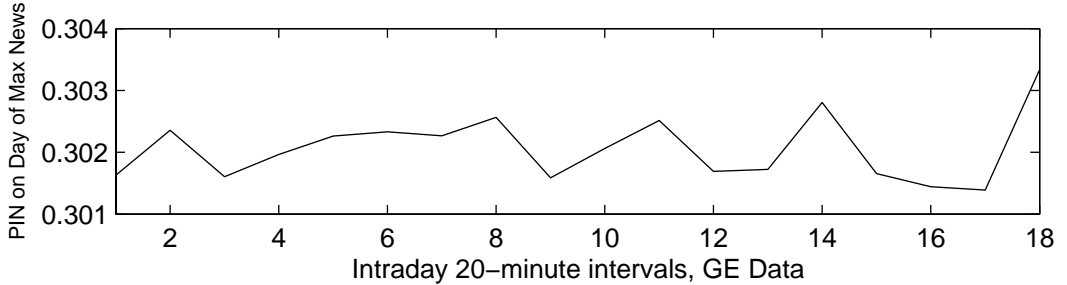
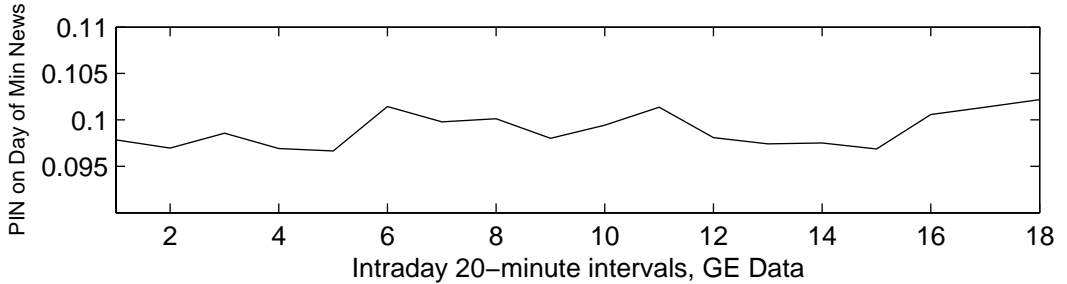
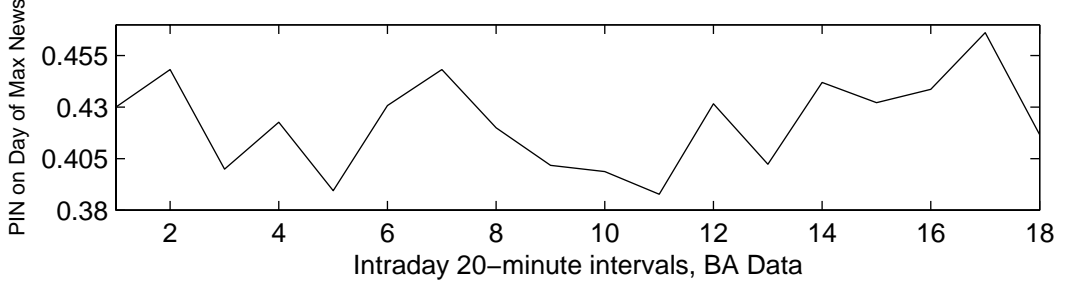
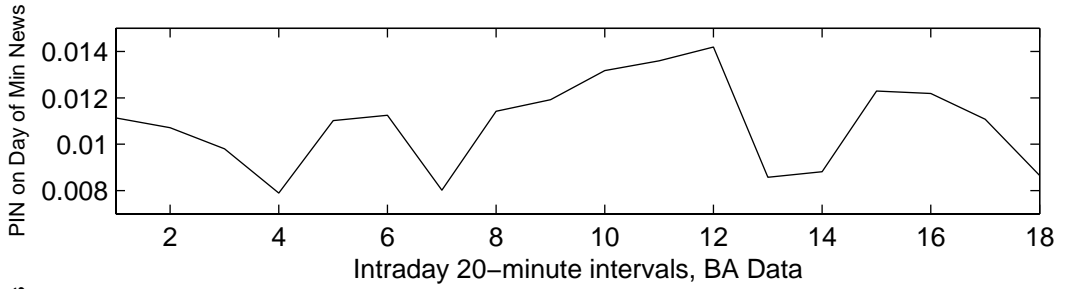


Fig 3: Intraday PIN of BA and GE Data



Conclusions

- We extend the AACD to estimate the probability of informed trading.
- The use of transaction data allows us to relax the assumption of independent aggregate buy and sell orders in the EHO framework.
- Our methodology yields daily estimates for the probability of informed trading, and allows the underlying probabilities of good news and bad news to be time varying.
- Future research may utilize our daily estimates for the probability of informed trading to study the impact of various events such as earnings announcements or merger activity on the level of informed trading.