Online Appendix: Trend and Reversal of Idiosyncratic Volatility Revisited

Markus Leippold, *

Michal Svatoň, *

*University of Zurich, Department of Banking and Finance, Plattenstrasse 14, 8032 Zurich, Switzerland

A Important CRSP features

Below we summarize several important features of the CRSP database and discuss their relevance for volatility estimation. The information is based on the Data Description Guide for CRSP US Stock & US Index Databases (update as of March 31, 2017).

A.1 Exchange addition

CRSP stock files for NYSE, Amex, and NASDAQ stocks start in December 31, 1925, July 2, 1962 and December 14, 1972 respectively. Each new inclusion substantially increases the size of the available stock universe. Given that stocks listed on the latter two exchanges are typically smaller and considered more risky, the aggregate volatility is likely to increase around on the date of their inclusion.

A.2 Availability of trading prices

In the CRSP database, prices are recorded in two different ways: either as a closing price, or as a bid-ask midpoint. Midpoint is used whenever the closing price is unavailable, which is the case for stocks with zero trading volume, and for stocks without trading data, i.e., all NASDAQ National

Date	Event	Effects
02-Jul-1962	Amex added to the database	Increase in number of firms, possibly riskier on average.
14-Dec-1972	NASDAQ added to the database	Huge increase in number of firms, riskier on average. Initially, no trading prices available, so recorded price are quote midpoints. Midpoints have no bid-ask bounce effects, but zero-return percentage spikes up to 50%.
July 1980	NASDAQ switched to inside quotation	With inside quotation the bid-ask spread is more volatile. Spread volatility can potentially spillover to volatility of quote midpoints.
01-Nov-1982	First day of availability of trading prices on NASDAQ National	Closing prices introduce bid-ask bounce to returns, thus biases the sample variances up. Share of zero daily returns decreases.
15-Jun-1992	First day of availability of trading prices on NASDAQ SmallCap	Closing prices introduce bid-ask bounce to returns, thus biases the sample variances up. Share of zero daily returns decreases. Effects possibly larger than on NASDAQ National due to lower prices and price discreteness.
03-Sep-1992	Tick size reduction on Amex	Lower severity of microstructure effects in daily returns.
29-Dec-1992	Quotes available for NYSE on daily basis (again)	Possibility to use quote data, measure bid-ask spreads.
May-June 1997	Tick size reduction on NYSE (June 24), NASDAQ (June 2) and Amex (May 7)	Lower severity of microstructure effects in daily returns.
09-Apr-2001	Completion of quote decimalization	The microstructure effects in daily returns become quantitatively negligible.

Table 1: List of important data-related events.

Description: Major events in the CRSP database and their effects. We summarize several important features of the CRSP database and discuss their relevance for volatility estimation. The information is based on Data Description Guide for CRSP US Stock & US Index Databases (update as of March 31, 2017).

Interpretation: Since 1962, there are many events in the CRSP database related to data availability that affect the measurement of returns and variances.

Market securities before November 1, 1982, and NASDAQ SmallCap Market securities before June 15, 1992.

Closing prices are affected by bid-ask bounce, i.e., oscillation between closing ask and bid depending on the direction of the last trade. As a consequence, sample variance will overstate the true volatility (see, for example, **Roll**). When the price type changes from midpoint to closing price, substantial differences in the behavior of the price processes and resulting returns occur, with closing prices generating substantial bid-ask bounce. Moreover, the bid-ask bounce results in less frequent zero returns. Because the frequency of zero returns depends on whether the price is a closing price or a midpoint, its use as a liquidity proxy (proposed by **Lesmond1999**) using CRSP data is problematic.

A.3 Availability of quotes

While trading prices are available for NYSE securities throughout the period under consideration, quote data are generally not. Between February 24, 1942 and December 27, 1992 quotes are available only when the trading price is unknown, i.e., when the price is recorded as a midpoint. Therefore, between 1972 and 1992 there are both stocks for which only trading price is available and stocks for which only quotes are.

A.4 Change to inside quotation

The information content of the quote data also differs across exchanges. On NYSE/Amex, and on NASDAQ before 1980 the price corresponds to the last representative[1] quote before market close. The switch to inside quotation has a direct effect on daily spread volatility. For example, even if all dealers have the same bid-ask spread, but their quotes are "shifted" (e.g. $9.98 - 10.02\$$ and $9.99 - 10.03\$$), the spread resulting from the inside quotes ($9.98 - 10.03\$$) may shrink. Figure 1 demonstrates the effect of the

[1]The description of representative quote for NYSE in the description guide reads: "This unrepresentative quote showed very large spreads, frequently a bid of a penny and an ask of approximately double the price. These were usually posted by a market marker not on the primary listed exchange, who was required to post a quote but not interested in making a trade."

change to inside quotes in July 1980 on spread volatility. On both stock and market level the volatility increases.[2] The volatility may generate artificial volatility of quote midpoints.

A.5 Tick size reductions

Price discreteness, stemming from minimum tick size requirements, is an amplifying force of most microstructure effects. Therefore, historical tick-size reductions mitigate part of the microstructure biases. Amex reduced tick size for stocks in price range 1$−5$ to 1/16$ in September 1992.[3] The tick size of 1/16$ was extended to all stocks on NYSE, NASDAQ and Amex in May-June 1997. "Decimalization" of quotes was completed in April 2001. On the contrary, the feature of NASDAQ market, documented by **CS1994** that the market makers on NASDAQ avoided quotes ending with odd eighths in historical period, likely upscales microstructure problems for NASDAQ. Figure 2 confirms that this phenomenon holds over the entire pre-decimalization era.

One of the consequences of price discreteness is bias in correlations of observed returns towards zero. When the tick size is large relative to stock price, discrete prices generate flat price segments, or equivalently zero returns. Zero returns have zero contribution in covariance computation, regardless whether they stem from a lack of trading activity or from price discreteness. Price discreteness is most important for low-price stocks in historical periods when tick size amounted to 1/8$. Furthermore, price discreteness may generate additional upward bias in sample variance (**Harris**).

B Critical Assessment of the Approach of LPSZ

The closest work related to ours is that of **HL2017** (LPSZ), who argue that the trend in value-weighted idiosyncratic volatility (IV) before 2000 documented by **CLMX2001** is a consequence of the bid-ask bounce, which

[2]The jump in NYSE/Amex series stems from the fact that bid and ask quotes were available for NYSE only after 1997.

[3]The range was extended up to 10$ in February 1, 1995 (**Angel1997**).

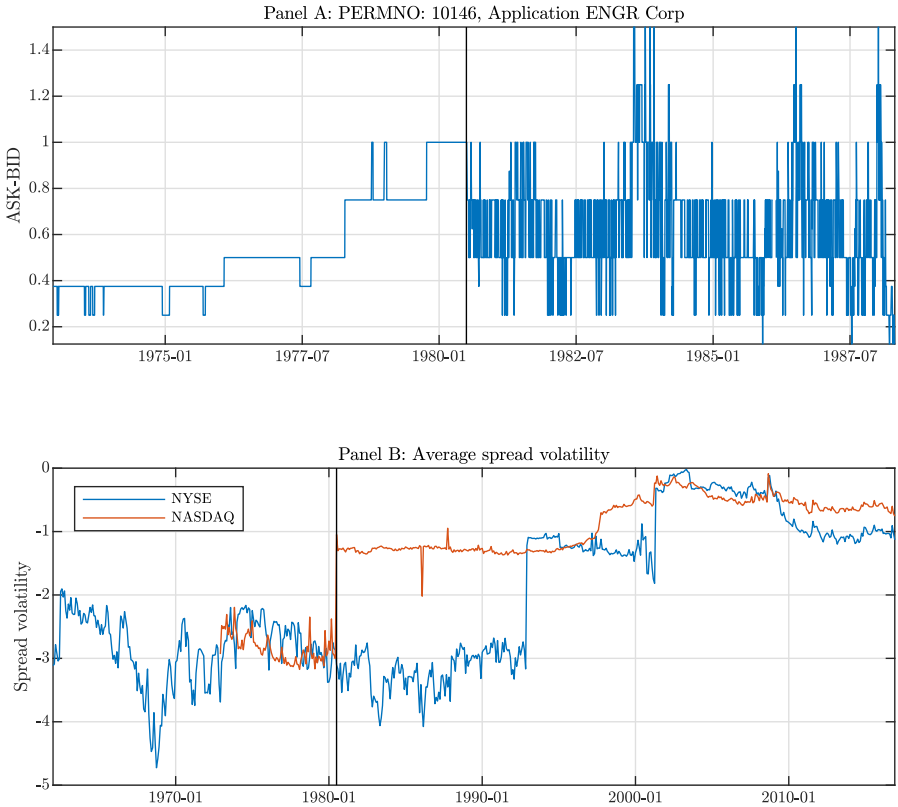

Figure 1: Inside quotation effect

Description: Panel A: Width of a spread of a NASDAQ stock. Panel B: Average volatility of spread ($\log\log(A/B)$) across stocks, by exchange. NYSE denotes the sample composed of NYSE and Amex. Vertical line indicates the date of switch to inside quotes.

Interpretation: Change to inside quotation on NASDAQ increased volatility of bid-ask spreads.

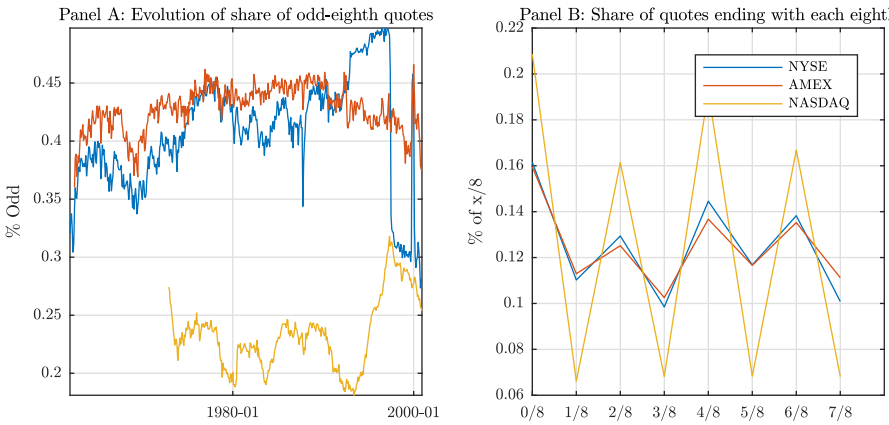

Figure 2: Odd vs. even eighths.

Description: Panel A: Percentage of odd eighths in quotes by exchange. Panel B: Percentage of bid quotes ending with given eighth before 2000.

Interpretation: In line with **CHRISTIE1999409** the NASDAQ market makers avoided odd-eights in their quotes, magnifying the price-discreteness effect.

biases realized volatility estimates up.[4] We briefly review their methodology and raise some nontrivial concerns about the validity and interpretation of their results.

As we have shown, the measurement of IV is driven by three components, industry concentration, average variance, and average correlation. In contrast, LPSZ remains silent about the role of these components. In particular, they do not discuss the role of variation in correlations, the bias in their sample estimates, and the evolution of industry concentration. Given the insights from our theoretical analysis, we can empirically disentangle the effects of these channels from that of the bid-ask bounce. Therefore, we fill an important gap in understanding the drivers in IV measurement. Furthermore, we highlight the different roles of individual channels not only for the value-weighted (VW) but also for the equal-weighted (EW) IV. Because the EW average loads heavily on small stocks, it provides a stronger "test" of the microstructure explanation, since these stocks are affected the most.

[4]LPSZ only consider value-weighted IV but not equal-weighted IV, in which microstructure biases would show up more prominently.

B.1 Main arguments of LPSZ

LPSZ support their conclusion along three different lines of arguments. As a first argument, they study the effect of the bid-ask bounce on IV by comparing trend estimates for IV measures based on closing prices from CRSP and quote midpoints from ISSM (International Study of Security Markets) and TAQ (Trade and Quote) databases. Unfortunately, the midpoint data are available only since 1983 which prevents a complete analysis of the CLMX period. In particular, the analysis cannot explain why the IV did not spike right after the inclusion of NASDAQ in 1972, despite the common perception that NASDAQ stocks are riskier. Also, the stock coverage between the databases may differ, especially for less liquid stocks, which usually have higher variances and are also more prone to microstructure effects. Using quote midpoints also does not resolve price discreteness and asynchronicity biases in correlations which, as we show in our paper, have some substantial impact on IV measurement.

As a second argument, LPSZ study the effect of exogenous shocks to bid-ask spreads, which led to a reduction of the realized IV estimates. LPSZ use as shocks the publication of **CS1994** which shows that NASDAQ market makers avoided odd-eighth quotes in the early 1990s and started to abandon this practice afterwards (**CHRISTIE1999409**), and the introduction of decimalization. Even though the reduction of spreads and an increase in liquidity indeed eliminated most of the bid-ask bounce, the same applies to price-discreteness and asynchronicity biases in correlation estimates. Therefore, the analysis of the IV itself provides little guidance on the relative importance of biases in correlation and variance estimates. Furthermore, by focusing on those specific events, one still cannot provide an answer to the question of what led to the increase in the IV (or the bias in its estimates) in the first place.

As the third argument, LPSZ study whether the time-series behavior of the IV can be explained by the variation in microstructure effects, which they measure by the (cross-sectional) average of the bid-ask spread. They estimate the spreads from the CRSP daily high and low price data using the estimator of **CS2012** (CS). Since this analysis relies on the CRSP database, it spans the entire CLMX sample starting in 1962. They assess the viability of the bid-ask bounce explanation by running regressions of the form

$$\mathrm{IV}_t = \alpha + \beta_t t + \beta_s s_t + (\beta_x^\top X_t) + \varepsilon_t, \tag{1}$$

where IV_t is the idiosyncratic variance measure, t is the time (trend) variable, s_t is the spread variable, X_t is a vector of control variables, and ε_t is the error term.

We identify two pitfalls related to such a regression analysis. The first one relates to an omitted variable bias. The significant loading on the spread might be just a consequence of the correlation of the spread with other variables. In Figure 6, we show that industry concentration, which is negatively related to the IV through diversification of the industry portfolios, also has a trend in the CLMX period and reverts afterward. Moreover, in Figure 5, we show that correlations, even when corrected for asynchronicity, are at their sample low in the 1990s. Therefore, it is difficult to disentangle these effects unless measures of correlation and industry concentration are included in the regression, or raw variances are used instead of the IV.

An even more pressing concern is endogeneity. If the spread is an increasing function of (true) volatility,[5] then regressing a volatility measure on the spread reduces almost to a tautology. To safeguard against this endogeneity problem, LPSZ use a lagged value of the spread in their regression analysis. Furthermore, they use the share of the NASDAQ capitalization as an instrument to control for the potential endogeneity bias. However, in a realistic setting, instrumenting the bid-ask spread with the percentage of (value-weighted) market value of NASDAQ listed firms versus the (value-weight) market of all firms does not solve the problem. To see this, consider the following toy example with one NASDAQ and one NYSE stock with $\sigma_{NYSE} < \sigma_{NASDAQ}$, and the squared spread s^2 being affine in variance,

$$s_j^2 = a + b\sigma_j^2, \tag{2}$$

with $a > 0$, $b > 0$, for $j \in \{\text{NYSE, NASDAQ}\}$. We further assume that the capitalization share of the NASDAQ stock follows a random walk, i.e.,

$$w_t = w_{t-1} + u_t. \tag{3}$$

Then, the value-weighted average variance is affine in w_{t-1}

$$\bar{\sigma}^2 = w_t\sigma^2_{NASDAQ} + (1-w_t)\sigma^2_{NYSE} = \sigma^2_{NYSE} + w_t(\sigma^2_{NASDAQ} - \sigma^2_{NYSE}) \tag{4}$$

$$= \sigma^2_{NYSE} + w_{t-1}(\sigma^2_{NASDAQ} - \sigma^2_{NYSE}) + u_t(\sigma^2_{NASDAQ} - \sigma^2_{NYSE}). \tag{5}$$

[5] See, e.g., the early study of **bollerslev1994bid** for the foreign exchange market.

However, the same applies to the lagged value of average squared spread,

$$\bar{s}^2_{t-1} = w_{t-1}s^2_{NASDAQ} + (1-w_{t-1})s^2_{NYSE} = s^2_{NYSE} + w_{t-1}(s^2_{NASDAQ} - s^2_{NYSE}) \tag{6}$$

$$= a + b\sigma^2_{NYSE} + w_{t-1}b(\sigma^2_{NASDAQ} - \sigma^2_{NYSE}). \tag{7}$$

As a consequence, regressing average variance on the share of NASDAQ stocks or the lagged value of average spread yields a positive coefficient, and with plausible parameter values also very high R-squared. We stress out that in this example, the variance is assumed to be observed directly, hence there is no microstructure bias. Therefore, if the regression of potentially biased variance estimates on spread finds a positive relation, it is not clear whether it indicates the relation of the spread with the true volatility or rather with the bias in its estimates. As argued above, proposed instruments do not resolve this issue. It shall be noted that the assumptions of the toy example, while simplistic, are reasonable. The positive relation between spread and volatility is consistent, e.g., with a model of **BSW2004** Capitalization share of NASDAQ is indeed very persistent and perhaps even trending in the pre-decimalization period. Finally, in the example above we consider a regression of average variance on the spread, not of the IV as the dependent variable. Even though the relations described above would not hold exactly for the IV, the problem would remain. Our direct approach, when we test for the presence of a trend in a "clean" measure of variance, is free of endogeneity concerns, and is therefore capable of separating the variation in bias from the variation in the variance itself.

B.2 The CS spread measure and microstructure noise

The regression approach in Equation (1) using the CS spread estimator raises two fundamental questions. First, we need to explore the quality of the CS spread measure. Second, one should critically scrutinize the suitability of the CS spread measure as a proxy for microstructure noise. In what follows, we argue that the CS spread measure is biased and ill-suited as a proxy for microstructure noise. Besides, we show that when we perform an exchange-specific analysis, the CS measure applied for NASDAQ stocks indicates that the spread is consistently high and non-decreasing until 1992, an observation that runs counter to the argumentation of LPSZ.

B.2.1 Quality of the CS spread measure

The CS spread estimator is based on a comparison of high-low ranges on two consecutive days. Unfortunately, its implementation with CRSP data is challenging for several reasons. The CRSP database does not contain trading prices (including daily high and low) for all NASDAQ stocks before 1982 and for NASDAQ SmallCap stocks before 1992. In those cases, as well as on days with zero trading volume, the CRSP database reports closing quotes (BIDLO and ASKHI) instead of daily high and low. Inspecting the LPSZ spread series (their Figure 1), we find essentially no jump in the spread measure on the inclusion of NASDAQ firms, which are commonly perceived to have a higher spread. In contrast, our smoothed spread measures exhibit a clear jump in December 1972, regardless of whether we weight the individual spreads equally or by their market capitalization. Hence, this absence of a jump in LPSZ indicates potential deficiencies of the CS estimator, which we elaborate on next.

Unfortunately, it is not clear from their paper how LPSZ handle the above problem. A straightforward implementation of the CS estimator would use directly the BIDLO and ASKHI variables, neglecting the fact that they contain closing quotes when the trading data are unavailable. A better treatment is to use available closing quote data directly, and use the CS estimator only when quotes are not available, which is the case for NYSE stocks before 1992. Alternatively, the spread could be measured by the CS estimator whenever high and low prices are available, while using the reported quotes whenever the trading prices are not available, as was the case for NASDAQ stocks before 1982 and SmallCap stocks before 1992. One concern is whether the spreads estimated using the CS estimator and the ones based on CRSP data are broadly consistent.

Figure 3 suggests that there is indeed an inconsistency between CS and CRSP spreads. In Panel A, we consider NASDAQ stocks for which initially only quotes were available, but trading prices were not. For negative values on the x-axis, the spread is computed using CRSP quote data. For the positive half-line, where trading data are available, the spread is estimated by the CS estimator. The switch from CRSP to CS induces a large jump in spread levels, leading to a doubling of the spread. In Panel B of Figure 3, we showcase the change in the opposite direction, i.e., from the CS estimator to spreads reported by CRSP. In this case, the sample comprises

stocks for which initially only the low and high prices were available, but not the quotes, as was the case for NYSE stocks in 1992. Here, the jump is in the opposite direction, but again large. Average spreads reduce by more than 50%. From this analysis, we conclude that the transition between the spread-estimation methods generates an artificial jump in the estimates, so the practical application of the CS estimator to study the effect of microstructure bias on the IV series with CRSP data is highly questionable and may lead to wrong conclusions.

A possible explanation of the inconsistency of the CS estimates and the spread values reported in the CRSP database may lie in the violation of the assumptions behind the CS estimator. First, the CS estimator assumes that stocks are monitored continuously, while in practice the price process is observed over a discrete number of trade times. Therefore, the low and high prices are measured with an error that depends on the stock's liquidity. Second, CS estimator relies on the assumption that the spread is constant over a two-day horizon. In Figure 1, we show that after the change to inside quotation on NASDAQ the observed daily bid-ask spread became highly volatile. Hence, also this second assumption becomes questionable under these circumstances.

To gain further insight into the accuracy of the CS estimator, we compare the spreads of NASDAQ stocks reported in the CRSP database with the CS estimates, where we naively apply the estimator on BIDLO and ASKHI variables, even though they may represent closing quotes rather than the daily extremes. In the early period, when the spread is estimated using spread data, the CS estimator slightly underestimates the true spread. When the CS estimator uses high and low prices, the bias is relatively larger and changes in sign over time. This behavior raises concerns about the interpretation of the regression results using the CS estimates. Moreover, it seems that the spread on NASDAQ stocks is steady, and from the 1980s to 1992 even increasing. Hence, it is difficult to bring this observation in line with the argument of LPSZ that the trend in IV is caused by a decrease in spreads.

B.3 Suitability of the bid-ask spread as a measure of microstructure bias

We argue that, even if we can measure the spread accurately, it is not an entirely appropriate measure of the severity of microstructure biases.

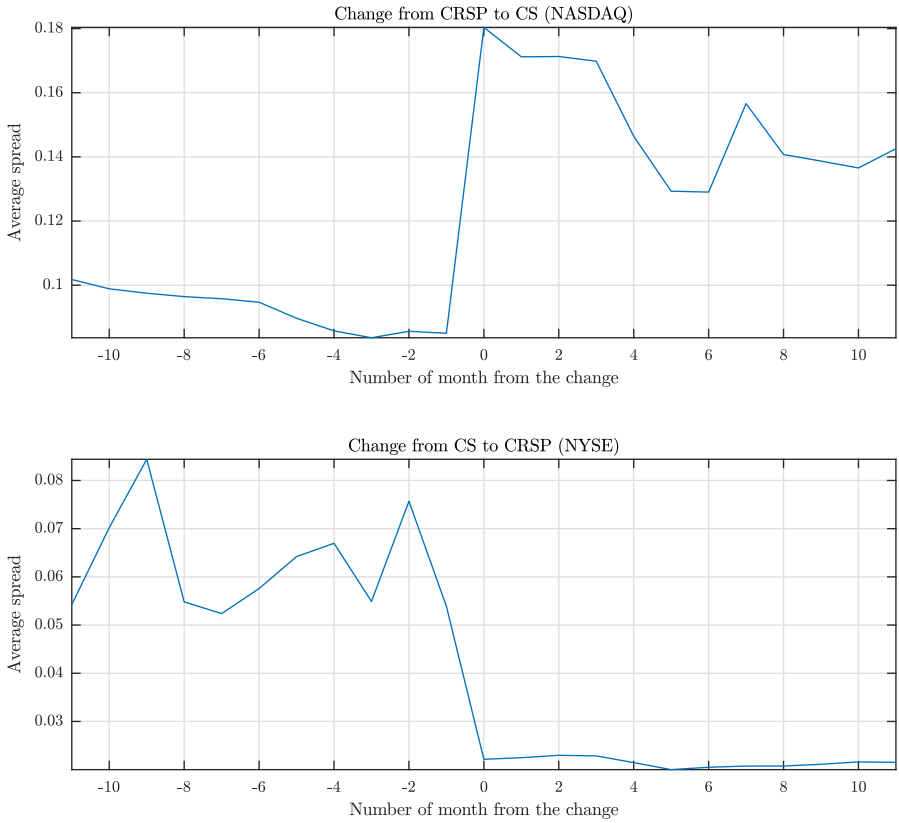

Figure 3: CS vs. CRSP spreads.

Description: Average equal-weighted percentage spread around dates of change in the availability of trading prices (high and low) and quote data. Panel A: Change of spread computation from spreads as reported in the CRSP database to the CS estimator when trading prices became available. The sample contains NASDAQ stocks for which trading prices were initially unavailable. Panel B: Change of spread computation from the CS estimator to spreads as reported in the CRSP database when quote data became available. The sample contains NYSE stocks for which quote data were initially unavailable.

Interpretation: The spread estimator of **CS2012** generates higher spreads than those in the data. Change in data availability induces jumps in spreads.

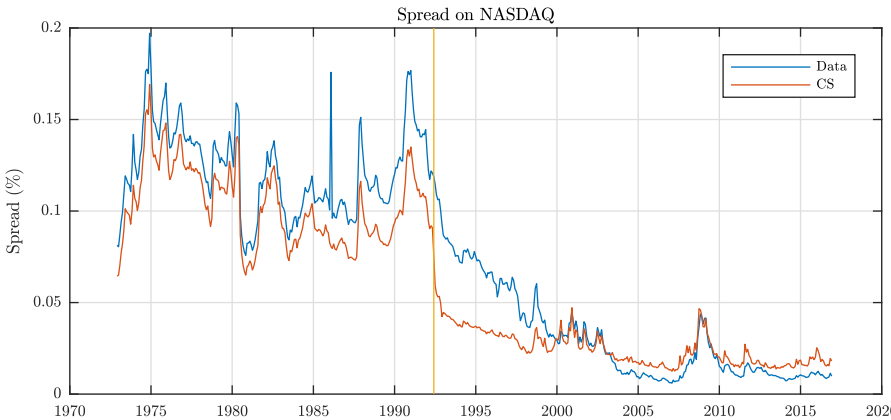

Figure 4: NASDAQ spread comparison.

Description: Comparison of average equal-weighted spread estimates. Vertical line indicates the date when trading prices became available for NASDAQ SmallCap market.

Interpretation: Average spread on NASDAQ is non-increasing in the CLMX sample, despite the increasing severity of microstructure biases. Variation in spreads is not sufficient to explain the trend in the IV on NASDAQ over the CLMX sample.

The reason is that due to the lack of trading data, the prices of NASDAQ stocks before 1982 and those of SmallCap stocks before 1992 correspond to a midpoint. Hence, the realized variance estimator is unaffected by the bid-ask bounce. The increasing availability of the trading prices for NASDAQ stocks after 1982 is crucial to explain why the IV started to trend upwards only in the 1980s, and not immediately after the inclusion of NASDAQ in 1972.

In Figure 1, Panel A in the main text, we demonstrate the effect of the transition from the midpoint quotation to closing prices on a single stock level. Clearly, the stock becomes much more volatile. While Panel A gives an idea about how a single stock may be affected, the change in notation has also a significant impact on an aggregated level. In Panel B, we show how the realized variance increases after changing to trading prices. The increase in aggregated (equal weighted) realized volatility is substantial. It jumps from levels below 20% to almost 40%. This spike is mainly due to the inclusion of the bid-ask bounce after the change.

From these results, it becomes clear that, instead of using the spread

directly, it is more sensible to "count" the spread only in case the price corresponds to a close, and not when it represents a midpoint. In Figure 5, we compare the average spread measure with the average of the spread multiplied by the indicator taking a value of one for the closing price. Here we use the smoothed spread from our model because, unlike CRSP spreads and CS spreads, we can use it for the entire sample. Since the smoothed spread is net of discretization effects, it still understates the bid-ask bounce effects in realized variance. However, the purpose of Figure 5 is to demonstrate the importance of accounting for the availability of trading prices.

As expected, the average spread jumps up on the inclusion of NASDAQ, but the series that controls for the availability of trading prices does not. In the more recent period starting with the end of the CLMX sample, the series almost coincide, because trading prices are missing only in case of zero trading volume.

The jump in the spread in 1972, when NASDAQ stocks were added to the sample, is inconsistent with a lack of one in the IV. The spread measure corrected for the availability of closing prices does not jump up (like IV) in 1972 and starts to increase only in the early 1980s when trading prices became available for NASDAQ stocks. The same pattern is visible in the average variance series (Figure 5 in the main text). Therefore, spread itself is not a suitable measure of the strength of the microstructure biases. The role of the availability of trading prices becomes even more apparent when we look at individual exchanges. On NASDAQ, where **BCL2008** find that the IV trend concentrates, the average spread is flat or even decreasing before decimalization (Figure 4). We confirm this observation by a univariate regression of the IV on NASDAQ on the spread, which yields a negative coefficient (not disclosed) for both the CS estimator and the smoothed measure, and both EW and VW case. When we use the "Spread x Closing" measure instead of spread, the coefficients turn positive, as predicted by the microstructure theory. However, all other regression-related problems discussed above likely remain.

B.4 Summary

LPSZ assess the validity of the bid-ask bounce story in three ways. Two of them – the comparison of TAQ midpoints and CRSP data, and event

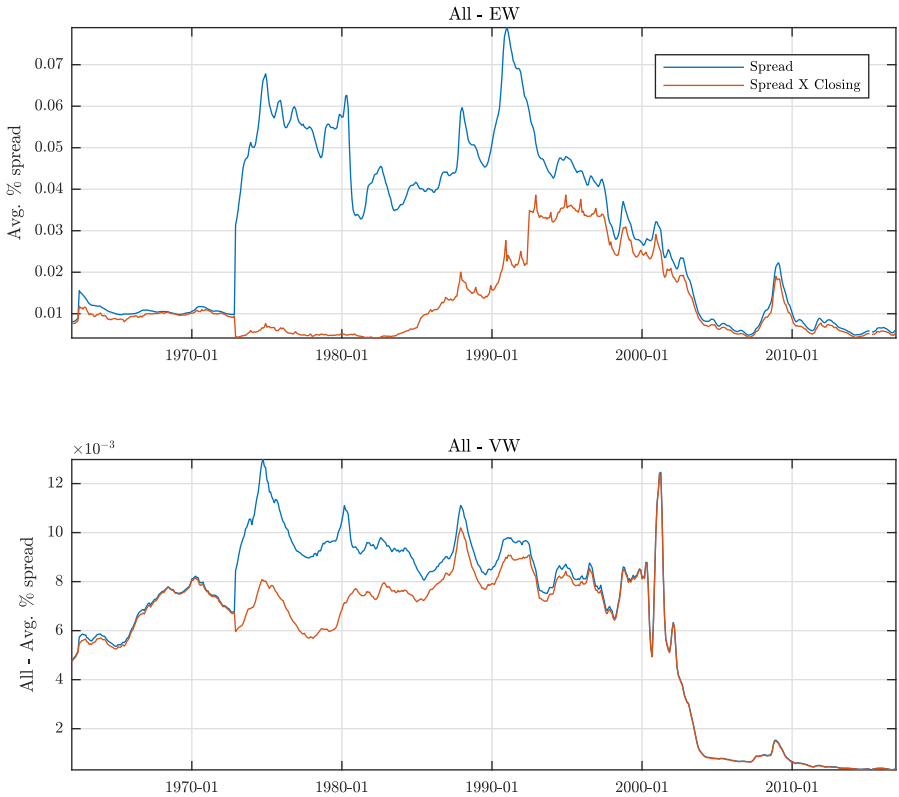

Figure 5: Spreads vs. spreads times closing price indicator.

Description: Time series of average percentage spread obtained from the model described in Section 4.2 in the main text. The series labeled "Spread x Closing" multiply the spread by the indicator variable that takes a value of one if trading price is available on a given day.

Interpretation: The interaction of spread width with the availability of trading prices matches the trends in average variance, while raw spreads do not.

studies (decimalization, publication of **CS1994**) – are restricted to a small subsample of the original CLMX study. On the full sample, LPSZ assess the validity of the bid-ask bounce story on the full sample by regressing the IV on the average spread, which has, as discussed above, nontrivial pitfalls. The first one is endogeneity stemming from a possible relation of spreads and true volatility together with the persistence of the spread. The regression approach is also prone to an omitted variable bias due to variation in correlation and industry concentration LPSZ do not control for in the regression. We also demonstrate that the spread is not an appropriate measure of severity of microstructure biases, because it neglects the fact that trading prices (which are affected by the bid-ask bounce) are not always available.

Our approach, where we directly use a "clean" measure of variance to test for a trend avoids the problems associated with LPSZ regressions. Our filtering procedure explicitly handles different cases of data availability. Hence, we can consistently capture the severity of the microstructure biases throughout all periods and across different stock exchanges, particularly for NASDAQ stocks which play a prominent role in the IV story. Because we also decompose the IV into variance, correlation, and industry concentration channels and study the channels separately, we provide additional insights on the role of those channels, and biases in variances and correlations respectively. Since we study the evolution of not only for value-weighted IV but also for its equal-weighted counterpart, our results are a stronger test for a microstructure explanation, because these effects tend to be more dominant for small stocks.

C Replication of CLMX at lower frequencies

Table 2 shows our replication of CLMX results for equal-weighted (EW) IV. Tables 3 and 4 present the results for IV series based on lower frequencies of return data. The results confirm the observations of CLMX that the trend in the idiosyncratic component is stronger for EW IV, and it is weaker at lower frequencies, but still remains significant.

	Raw data			Downweighted crash		
	MKT	IND	FRM	MKT	IND	FRM
Daily, EW (replicated), (N = 426)						
Mean × 100	1.02	0.55	35.91	0.97	0.55	35.90
Std. dev. × 100	2.26	0.27	23.82	1.53	0.26	23.78
Std. dev. × 100 detrended	2.25	0.27	15.37	1.52	0.26	15.32
Linear trend $\times 10^5$	−0.09	0.01	14.78	−0.12	0.01	14.77
PS-statistic	−0.06	−0.02	10.70	−0.10	−0.02	10.69
Confidence interval	(-0.21, 0.09)	(-0.09, 0.06)	(0.40, 20.99)	(-0.25, 0.04)	(-0.09, 0.06)	(0.36, 21.01)
Daily, EW (original), (N = 426)						
Mean × 100	1.21	1.25	33.90	1.15	1.25	33.90
Std. dev. × 100	2.62	0.55	23.11	1.72	0.41	23.11
Std. dev. × 100 detrended	2.61	0.55	14.12	1.70	0.55	14.12
Linear trend $\times 10^5$	−0.11	0.02	12.39	−0.14	0.02	12.39
PS-statistic	−0.08	−0.00	11.23	−0.13	−0.00	11.22
Confidence interval	(-0.33, 0.17)	(-0.15, 0.14)	(5.29, 17.17)	(-0.38, 0.11)	(-0.15, 0.14)	(5.30, 17.14)

Table 2: Replication of main CLMX results.

Description: Replication of **CLMX2001** on the CLMX sample (July 1962 - December 1997) for equal-weighted IV computed with daily returns. The means and standard deviations are annualized. The standard deviation detrended correspond to a standard deviation of residuals in a regression of variance component on linear trend. Linear trend statistic is based on OLS of the variance measure on a linear trend. PS statistic is computed as in **V1998** the corresponding implied confidence interval is computed on 90% level. Panel labeled "replicated" shows the values computed by us, panel labeled "original" shows the corresponding numbers as presented in **CLMX2001** Raw data columns compute the statistics directly from estimates of individual variance components. For downweighted crash columns, we replace the largest observation by the second largest value over the CLMX (1962-1997) sample.

Interpretation: Our replication is reasonably accurate and confirms the results of **CLMX2001** Idiosyncratic volatility trended in period 1962-1997 more strongly when the stocks are weighted equally.

	Raw data			Downweighted crash		
	MKT	IND	FRM	MKT	IND	FRM
Weekly, VW (replicated), (N = 426)						
Mean × 100	1.48	0.88	4.98	1.47	0.88	4.98
Std. dev. × 100	2.18	0.54	1.83	2.17	0.53	1.82
Std. dev. × 100 detrended	2.18	0.53	1.64	2.17	0.51	1.63
Linear trend $\times 10^5$	0.00	0.09	0.66	0.00	0.09	0.66
PS-statistic	0.10	0.08	0.62	0.10	0.08	0.62
Confidence interval	(-0.22, 0.43)	(0.01, 0.15)	(0.20, 1.03)	(-0.22, 0.43)	(0.01, 0.15)	(0.21, 1.03)
Weekly, VW (original), (N = 426)						
Mean × 100	1.90	1.22	5.84	1.86	1.22	5.84
Std. dev. × 100	2.52	0.73	2.21	2.16	0.73	2.21
Std. dev. × 100 detrended	2.52	0.72	1.92	2.16	0.72	1.92
Linear trend $\times 10^5$	0.00	0.05	0.74	−0.02	0.05	0.74
PS-statistic	0.12	0.10	0.41	0.08	0.10	0.41
Confidence interval	(-0.33, 0.56)	(-0.13, 0.32)	(0.13, 0.69)	(-0.36, 0.52)	(-0.13, 0.32)	(0.13, 0.69)

Table 3: Replication of main CLMX results.

Description: Replication of **CLMX2001** on the CLMX sample (July 1962 - December 1997) for value-weighted IV computed with weekly returns. The means and standard deviations are annualized. The standard deviation detrended correspond to a standard deviation of residuals in a regression of variance component on linear trend. Linear trend statistic is based on OLS of the variance measure on a linear trend. PS statistic is computed as in **V1998** the corresponding implied confidence interval is computed on 90% level. Panel labeled "replicated" shows the values computed by us, panel labeled "original" shows the corresponding numbers as presented in **CLMX2001** Raw data columns compute the statistics directly from estimates of individual variance components. For downweighted crash columns, we replace the largest observation by the second largest value over the CLMX (1962-1997) sample.

Interpretation: Our replication is reasonably accurate and confirms the results of **CLMX2001** The trend in idiosyncratic volatility in period 1962-1997 is weaker when weekly returns are used instead of daily returns.

	Raw data			Downweighted crash		
	MKT	IND	FRM	MKT	IND	FRM
Monthly, VW (replicated), (N = 426)						
Mean × 100	2.39	1.29	5.08	2.32	1.29	5.08
Std. dev. × 100	4.70	0.93	2.14	4.01	0.91	2.12
Std. dev. × 100 detrended	4.70	0.93	1.89	4.01	0.91	1.87
Linear trend $\times 10^5$	0.03	0.06	0.80	−0.01	0.06	0.80
PS-statistic	0.20	0.09	0.70	0.14	0.09	0.70
Confidence interval	(-0.22, 0.61)	(-0.07, 0.25)	(0.31, 1.09)	(-0.27, 0.55)	(-0.07, 0.25)	(0.31, 1.09)
Monthly, VW (original), (N = 426)						
Mean × 100	()	1.27	5.04	()	1.27	5.04
Std. dev. × 100	()	1.03	2.20	()	1.03	2.20
Std. dev. × 100 detrended	()	1.03	1.93	()	1.03	1.93
Linear trend $\times 10^5$	()	0.03	0.72	()	0.03	0.72
PS-statistic	()	0.09	0.78	()	0.09	0.78
Confidence interval	()	(-0.20, 0.39)	(0.28, 1.28)	()	(-0.20, 0.39)	(0.28, 1.28)

Table 4: Replication of CLMX results at lower frequencies.

Description: Replication of **CLMX2001** on the CLMX sample (July 1962 - December 1997) for value-weighted IV computed with monthly returns. The means and standard deviations are annualized. The standard deviation detrended correspond to a standard deviation of residuals in a regression of variance component on linear trend. Linear trend statistic is based on OLS of the variance measure on a linear trend. PS statistic is computed as in **V1998** the corresponding implied confidence interval is computed on 90% level. Panel labeled "replicated" shows the values computed by us, panel labeled "original" shows the corresponding numbers as presented in **CLMX2001** Raw data columns compute the statistics directly from estimates of individual variance components. For downweighted crash columns, we replace the largest observation by the second largest value over the CLMX (1962-1997) sample.

Interpretation: Our replication is reasonably accurate and confirms the results of **CLMX2001** The trend in idiosyncratic volatility in period 1962-1997 is the weakest when monthly returns are used instead of daily returns.

D Derivation of Equation (8) of the paper

In the derivation we suppress the time subscripts for brevity. We first rewrite the IV as

$$\sigma_\eta^2 = \sum_i w_i \sum_{j\in i} w_{ji} \operatorname{Var}(R_{ji} - R_i) \tag{8}$$

$$= \sum_i w_i \sum_{j\in i} w_{ji} (\operatorname{Var}(R_{ji}) + \operatorname{Var}(R_i) - 2\operatorname{Cov}(R_{ji}, R_i)). \tag{9}$$

Next, we note that

$$\sum_i w_i \sum_{j\in i} w_{ji} \operatorname{Var}(R_i) = \sum_i w_i \operatorname{Var}(R_i) \tag{10}$$

$$\sum_i w_i \sum_{j\in i} w_{ji} \operatorname{Cov}(R_{ji}, R_i) = \sum_i w_i \operatorname{Cov}\left(\sum_{j\in i} w_{ji} R_{ji}, R_i\right) \tag{11}$$

$$= \sum_i w_i \operatorname{Cov}(R_i, R_i) = \sum_i w_i \operatorname{Var}(R_i). \tag{12}$$

Plugging in the above in Equation (9) gives the desired result.

E Derivation of Equation (9) of the paper

In the derivation we suppress the time subscripts for brevity. We need to only compute average firm and industry variances. Let $\bar{w}$ denote the vector of weights within an industry. Direct computation gives

$$\sigma_\eta^2 = \sum_i w_i \sum_{j\in i} w_{ji} \operatorname{Var}(R_{ji}) - \sum_i w_i \operatorname{Var}(R_i) \tag{13}$$

$$\sum_i w_i \sum_{j\in i} w_{ji} \operatorname{Var}(R_{ji}) = \sum_i w_i \sum_{j\in i} w_{ji} \bar{\sigma}_i^2 = \sum_i w_i \bar{\sigma}_i^2 \tag{14}$$

$$\sum_i w_i \operatorname{Var}(R_i) = \sum_i w_i \bar{w}^\top \bar{\sigma}_i^2 ((1-\rho) I + \rho \mathbf{1}\mathbf{1}^\top) \bar{w} \tag{15}$$

$$= \sum_i w_i \bar{\sigma}_i^2 \left((1-\rho) \sum_j \bar{w}_j^2 + \rho \right). \tag{16}$$

Combining the average variance of firms and industries verifies Equation (**??**).

F Filtering procedure

To estimate our nonlinear state-space model outlined in Equations (15) to (18) of the paper, we implement a particle filter. We distinguish three cases: when only closing prices are available (CP), when bid and ask quotes are available (BA), and when price information is unavailable (NA). The state vector f is four-dimensional, $f = \{s^*, \log(\sigma), c, \chi\}$. We first sample $\log(\sigma_0)$ from a standard Gaussian distribution with a mean equal to a logarithm of Roll's estimate based on the first 22 return observation as measured by CRSP return series. We bound Roll's standard deviation by 0.001 from below. For χ_0, we use a Gaussian distribution with standard deviation of 0.2. We set the mean to $\log(\log(A_1/B_1)/4)$ in the BA case. In the CP case, we use the log-spread implied by the model of Roll, equal to $\log(\sqrt{-\mathrm{Cov}(r_t, r_{t-1})})$, and estimate the autocorrelation based on the first 22 returns. In case the autocorrelation is non-negative, we use $\log(\log(\min\{(S_1+d_1)/(S_1-d_1), 5\})/4)$ instead. We draw s_0^* from a Gaussian distribution with standard deviation equal to $2e^{\chi_0}$, centered either around the log of the price on the first day (CP), or around the average of the log-bid and log-ask (BA).

We denote observations at time t by Y_t, which consists either of the bid-ask pair or the closing price. The filtering procedure sequentially approximates $p(f_{1:t}|Y_{1:t}) \propto p(Y_t|f_t)p(f_t|f_{t-1})p(f_{1:t-1}|Y_{1:t-1})$ by a distribution over a discrete set of particles. The transition densities are specified by Equations (**??**) to (**??**), and $p(f_{1:t-1}|Y_{1:t-1})$ is approximated in the previous step of the filter. In the BA case, when Y_t consists of the observed bid and ask quote B_t and A_t, Equation (14) of the paper implies

$$p(Y_t|f_t) \propto \mathbb{1}_{\{\log(A_t-d_t)\le s_t^*+C_t\le\log(A_t)\wedge\log(B_t)\le s_t^*-C_t\le\log(B_t+d_t)\}}. \tag{17}$$

Therefore, the observation equation only imposes constraints on the price-spread pair (rectangular for $\{s^*, C\}$). Because $c = \log(C)$ as well as $\log(S^*F + X)$ are Gaussian, conditionally on current volatility and previous states, the target distribution is bivariate Gaussian constrained to a set with nonlinear boundaries. Due to these restrictions, it is desirable to avoid sampling outside of the feasible set of state values. Next, we denote the normal distribution truncated to the interval $[l, u]$ by $tN(\mu, \sigma^2, l, u)$ and by $tN(\mu, \sigma^2, I)$ when truncated to the set I. Further denoting the i-th

particle by the superscript (i), we sample the states as follows to ensure that the draws are within the desired region.

1. We draw log-volatilities from the Gaussian transition equation (**??**).
2. We sample $c_t^{(i)}$ from $tN(\xi_{t-1}^{(i)}, \gamma_u^2 + \gamma_\chi^2, k_l, k_u)$. The bounds of the truncation region are equal to

$$k_l = \log(\max((\log(A_t - d_t) - \log(B_t + d_t))/2, 0)), \tag{18}$$
$$k_u = \log((\log(A_t) - \log(B_t))/2). \tag{19}$$

3. Conditional on draws $\sigma^{(i)}$ and $c_t^{(i)}$, we sample $\log(S^*F + X)^{(i)}$ from $tN(s_{t-1}^{(i)}, (\sigma^2)^{(i)}, j_l, j_u)$, with

$$\tilde{j}_l = \max(\log(A_t - d_t) - C_t^{(i)}, C_t^{(i)} + \log(B_t)), \tag{20}$$
$$\tilde{j}_u = \min(\log(A_t) - C_t^{(i)}, C_t^{(i)} + \log(B_t + d_t)), \tag{21}$$
$$j_l = \log(\exp(\tilde{j}_l)F_t + X_t), \tag{22}$$
$$j_u = \log(\exp(\tilde{j}_u)F_t + X_t). \tag{23}$$

   The $\tilde{j}$'s correspond to bounds on the log-price, the j's account for dividend payments.

4. Conditional on $c_t^{(i)}$, we sample $\chi_t^{(i)}$ from $N(\mu_\chi, \sigma_\chi^2)$, where

$$\mu_\chi = \bar{\chi} + \varphi_\chi(\chi_t - \bar{\chi}) + \frac{\gamma_\chi^2}{\gamma_c^2 + \gamma_\chi^2}(c_t^{(i)} - (\bar{\chi} + \varphi_\chi(\chi_t - \bar{\chi}))), \tag{24}$$

$$\sigma_\chi = \frac{\gamma_c^2 \gamma_\chi^2}{\gamma_c^2 + \gamma_\chi^2}. \tag{25}$$

In the CP case, we observe only the closing price, which can be either of ask, so that $s_t^* + C_t \in [S_t - d_t, S_t]$, or bid which implies $s_t^* - C_t \in [S_t, S_t + d_t]$. Thus,

$$p(Y_t|f_t) \propto 0.5\mathbb{1}_{\{\log(P_t - d_t) \le s_t^* + C_t \le \log(P_t)\}} + 0.5\mathbb{1}_{\{\log(P_t) \le s_t^* - C_t \le \log(P_t + d_t)\}}. \tag{26}$$

For the CP case, the following sampling scheme guarantees that we draw in the feasible region.

1. We draw log-volatilities from the Gaussian transition equation (16) of the paper.

2. We sample $\{c_t^{(i)}, \chi_t^{(i)}\}$ from their transition equations (17)-(18) of the paper.

3. Conditional on draws $\sigma^{(i)}$ and $c_t^{(i)}$, we sample $\log(S_t^* F_t + X_t)^{(i)}$ from $tN(s_{t-1}^{(i)}, \sigma^{2(i)}, I_t)$. The truncation region is a union of two intervals, $I_t = [m_l, m_u] \cup [n_l, n_u]$, corresponding to the bid and ask cases. The interval boundaries for the log-price are defined as

$$\tilde{m}_l = \log(S_t) + C_t, \quad \tilde{m}_u = \log(S_t + d_t) + C_t, \tag{27}$$
$$\tilde{n}_l = \log(S_t - d_t) - C_t, \quad \tilde{n}_u = \log(S_t) - C_t, \tag{28}$$

   and the boundaries for $\log(S_t^* F_t + X_t)^{(i)}$ are obtained analogously to Equation (22).

When an observation is missing (NA), we directly sample from the transition equations, with zero incremental weights. After each step, we resample the particles if the effective (relative) sample size drops below one half. [6] In addition, we make the number of particles N_t time-varying, increasing their count in situations when future observations are weakly informative about the states. We let the number of particles vary as $N_t = M_t \bar{N}$, where the baseline particle count $\bar{N}$ is set to 500, and the dynamic multiplier is a smoothed version of raw multipliers defined below,

$$M_t = \max(\tilde{M}_{\max(t-99,1):t}). \tag{29}$$

The computation of raw multiplier $\tilde{M}$ differs for the BA and CP case. In the BA case, we first flag BA observations as informative if either the bid or the ask changes ($INFT_t = 1$). Then, we compute its forward-looking moving average, $\overline{INFT}_t = \frac{1}{100} \sum_{i=0}^{99} INFT_{t+i}$, and define

$$\tilde{M}_t^{BA} = \max\left(\min\left(\left\lceil \overline{INFT}_t^{-1} \right\rceil, 10\right), 1\right), \tag{30}$$

[6] We use binomial resampling. We also experimented with systematic and residual resampling, but the choice of resampling method has a negligible impact on variance estimates. Therefore, we opt for the simplest scheme.

having more particles if price changes are less frequent in future periods. For the CP case, we first compute

$$Z_t = \max\left(\frac{20}{S_t}\frac{d_t}{1/8}, 1\right), \tag{31}$$

then we take its forward-looking average,

$$\bar{Z}_t = \frac{1}{100}\sum_{i=0}^{99} Z_{t+i}, \tag{32}$$

and finally define

$$\tilde{M}_t^{CP} = \min(\lceil Z_t \rceil, 10). \tag{33}$$

This criterion increases the number of particles in cases the price is low relative to the prevailing tick size. By taking the rolling maximum in Equation (29), we avoid frequent oscillation in the particle count when the BA and CP cases switch due to lack of trading activity and imply a different raw multiple.

The next ingredient in the implementation of the filter is specification of the tick-size d_t. We extract the values for each stock individually. For each date we take the set of all quotes and trading prices (not midpoints) that occurred up to date, and find the smallest difference among the prices in the set. For the first 22 observations, we use the window of first 22 days of prices. An advantage of this approach is its simplicity and that it is able to capture aspects such as gradual implementation of quote decimalization, or different rules of exchanges applying to cross-listed stocks. One disadvantage of this procedure is that the tick reduction is permanent. In particular, since for stocks priced below 1$ the tick sizes are lower, recovery of the price above this threshold will result in an underestimation of d_t. The same issue may arise in case of quote errors, or unusual quotes (ticks) on NASDAQ, where the size of price increments was restricted by customs rather than formal rules. If the tick size is too low, the observed and the latent quotes almost coincide and the filtered price is close to the quote midpoint, due to the assumption of a symmetric spread.

We keep the parameters of the transition dynamics constant and identical for all stocks, which avoids the computational burden that would stem from a formal estimation. We set the mean-reversion parameters

φ_σ and φ_χ to 0.999, implying a (prior) half-life of 623 days.[7] We fix the reversion levels to $\log(\bar{\sigma} = \log(0.5/252))$ and $\bar{\chi} = \log(\log(1.25)) - \log(2)$. However, the exact choice of the constant is not crucial due to relatively long half-life of the processes. For the volatility in the mean spread, we choose $\gamma_\chi = 0.02$, motivated by slow variation in long-term spread levels. To capture the spread oscillation around its local level, we let $\gamma_c = 0.5$. This choice becomes vital to capture the spread volatility following the switch to inside quotes for NASDAQ stocks in 1980. Finally, we set the volatility-of-volatility to $\gamma_\sigma = 0.1$. This choice allows for fast changes in volatility. A two-standard deviation increase over ten-day horizon corresponds to drop of volatility by -53% or an increase by 88%.

Before applying the filtering procedure outlined above, we eliminate outliers from the data, which often stem from data errors (unrepresentative quotes). Panels A and B in Figure 6 show two such observations, one corresponding to a closing price and the other to a bid-ask midpoint.

We identify suspicious quotes in the data by combining multiple criteria. First, we flag as a quote error observations with extreme or unusually wide bid-ask spread, identified as

$$C_{ES} = \frac{A}{B} > 5, \tag{34}$$

$$C_{US} = \left| \text{med}\left(\log\left(\frac{A}{B}\right), 9 \right) - \log\left(\frac{A}{B}\right) \right| > 1, \tag{35}$$

$$C_{Q1} = C_{ES} \vee C_{US}. \tag{36}$$

Then, we add a criterion to capture extreme quotes, which still possibly exhibit a reasonable (or less extreme) spread. First, we check whether either the spread is wide (in absolute or relative terms) or the recorded price is outside of prevailing quotes. Formally,

$$C_{REL} = \max\left(1, \frac{(A-1/8)}{(B+1/8)}\right) - 1 > 1, \tag{37}$$

$$C_{ABS} = \left(\max\left(1, \frac{(A-1/8)}{(B+1/8)}\right) - 1 > \frac{1}{2}\right) \wedge (A - B > 3), \tag{38}$$

$$C_{OUT} = (P > A) \vee (P < B). \tag{39}$$

[7] We also experimented with a unit root specification instead of mean-reverting process. Overall, the estimates are barely affected by this choice, but the mean-reverting specification is more robust to long, uninformative periods.

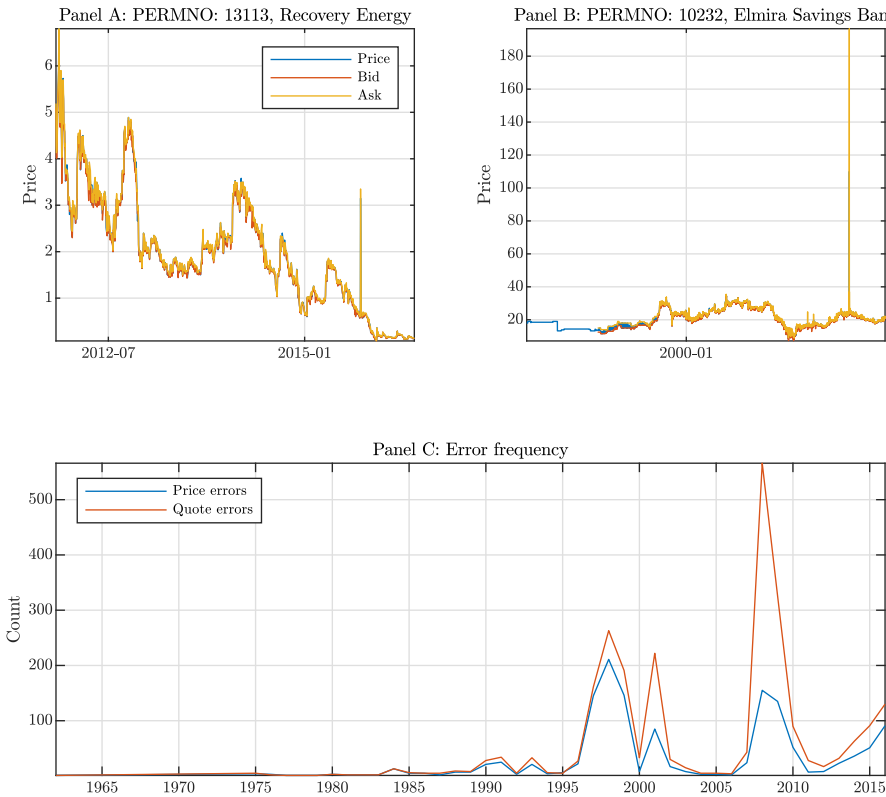

Figure 6: CRSP problematic quotes

Description: Panel A: Example of a stock with outlier price observation, corresponding to closing price. Panel B: Example of a stock with outlier price observation, corresponding to quote midpoint. Panel C: Time series of outlier occurrences.

Interpretation: CRSP data contain numerous outliers, some of which correspond to data errors or unrepresentative quotes. Data cleaning is therefore important.

and combine them to $C_{QCS} = C_{REL} \vee C_{ABS} \vee C_{OUT}$. Then, we compare the log-bid and log-ask, adjusted for dividends and stock splits, with their moving medians over a window of 9 days. We check the same criterion on bid (ask) level as well, and consider the quote to be problematic if both the level and logarithmic criteria hold. We combine the spread criterion C_{QCS} with the time-series criterion C_{QTS} to obtain the final criterion for erroneous quotes C_Q,

$$\begin{aligned} C_{QTS} =& \left(\left(|b - \text{med}(b,9)| > \frac{1}{2}\right) \wedge \left(|B - \text{med}(B,9)| > \frac{1}{2}\right)\right) \vee \\ & \left(\left(|a - \text{med}(a,9)| > \frac{1}{2}\right) \wedge \left(|A - \text{med}(A,9)| > \frac{1}{2}\right)\right), \end{aligned} \tag{40}$$

$$C_{Q2} = C_{QTS} \wedge C_{QCS}, \tag{41}$$

$$C_Q = C_{Q1} \vee C_{Q2}. \tag{42}$$

For the price series, we label the observation as suspicious if the log-price is far away from its moving median over a window of five observations. To avoid false positives for low-priced stocks, we also require the same criterion to hold for price levels. Formally, we let

$$C_{PREL} = |p - \text{med}(p,5)| > 1, \quad C_{PABS} = |P - \text{med}(P,5)| > 1. \tag{43}$$

and $C_{PTS} = C_{PREL} \wedge C_{PABS}$. In addition, we also treat all prices computed as quote midpoint, when the quotes are erroneous (C_Q holds) as an error, C_P: $C_{PTS} \vee (C_Q \wedge \text{Midpoint})$. In both cases we keep the last observation in the sample, whether the criteria described above hold or not, so that we do not exclude pre-delisting information. Table 5 provides examples of detected errors by our cleaning procedure.[8] Panel C in Figure 6 shows that most of the outliers are detected around the dot-com bubble, the financial crisis, and in the most recent period. The surge in the former two

[8] In most cases, the presence of a quote "error" is clear. For example, in the second line of Table 5, the observation is flagged as problematic because the ask/bid is too large. With a wide bid-ask spread, the discrepancy between midpoints under assumptions of a symmetric spread in levels and logs becomes large. As a consequence, the resulting filtered returns might be large if the adjacent observations have either narrower spread or do not have available quotes. As another example, the observation in the sixth block is possibly a false positive, flagged because of its relatively low bid compared to neighboring observations, and the fact that the price is outside of the quotes. As a consequence, the filtering procedure uses the price instead of the quotes.

periods likely reflects a combination of low liquidity and false positives. In shallow markets, a huge price change may occur due to the price impact of a block trade. Extreme quotes often occur due to the unwillingness of the market maker to trade, motivating them to post extreme quotes. The increase in the recent period is more puzzling, perhaps indicating that the unrepresentative quotes are only gradually assessed and eliminated by CRSP.

Using the criteria above we distinguish several cases. If both C_P and C_Q hold (or quotes are unavailable) then we treat the observation as missing, i.e., the NA case. If only C_Q holds, we eliminate the quotes and use the closing price only (CP). When the quotes are error-free, but the price is not, the effect on our filtering procedure is limited, as in such cases we use the quotes. Still, the price error might cause problems for the initialization of the filter, so we convert an erroneous price to its previous value and recompute the resulting returns.[9]

From the particle filter we compute two sets of estimates, a filtered series and a smoothed series. The latter is obtained from a fixed-lag approximation using $L = 100$. While other smoothing methods might be preferable, we opt for the fixed-lag approximation for its simplicity and low computational cost, which is of practical relevance given that we apply the filter for approximately 24,000 stocks.

[9]Furthermore, price information might be missing for other reasons than our exclusion of outliers, e.g., due to suspension from trading. We treat those missing values analogously, i.e., as the NA case, unless there are missing values for more than 22 consecutive observation. In such cases, the variance of the filter would be too large. Instead, we split the full data range of the stock into connected segments, where no such gaps occur, and estimates the states on each segment separately.

ID	Date	PRC	BID	ASK	BIDLO	ASKHI	RET	Mid
10001	03-Aug-2009	8.050	8.050	8.150	7.750	8.340	-0.012	false
10001	04-Aug-2009	8.600	7.750	999.990	8.400	8.600	0.068	false
10001	05-Aug-2009	8.578	8.350	8.600	7.960	8.590	-0.003	false
10042	05-Aug-1999	0.188	-	-	0.156	0.188	0.200	false
10042	06-Aug-1999	0.172	0.031	0.313	0.031	0.313	-0.083	true
10042	09-Aug-1999	0.156	-	-	0.156	0.188	-0.091	false
10100	11-Apr-2008	1.050	0.820	1.280	0.820	1.280	-0.014	true
10100	14-Apr-2008	2.595	0.710	4.480	0.710	4.480	1.471	true
10100	15-Apr-2008	1.035	0.820	1.250	0.820	1.250	-0.601	true
10100	09-May-2008	1.025	0.850	1.200	0.850	1.200	0.000	true
10100	12-May-2008	0.850	0.350	1.780	0.850	0.850	-0.171	false
10100	13-May-2008	1.065	0.850	1.280	0.850	1.280	0.253	true
10100	26-Sep-2008	0.775	0.670	0.880	0.670	0.880	-0.119	true
10100	29-Sep-2008	0.880	0.250	1.250	0.880	0.880	0.136	false
10100	30-Sep-2008	0.670	0.790	0.880	0.670	0.790	-0.239	false
10100	18-Dec-2008	0.520	0.490	0.550	0.490	0.550	-0.096	true
10100	19-Dec-2008	0.310	0.390	0.440	0.310	0.530	-0.404	false
10100	22-Dec-2008	0.310	0.310	0.820	0.310	0.310	0.000	false
10100	13-Jan-2009	0.600	0.710	0.790	0.600	0.680	-0.143	false
10100	14-Jan-2009	1.390	0.040	2.740	0.040	2.740	1.317	true
10100	15-Jan-2009	0.725	0.680	0.770	0.680	0.770	-0.478	true
10205	26-Sep-2008	12.000	11.900	12.010	12.000	12.970	-0.016	false
10205	29-Sep-2008	10.800	6.000	12.100	9.050	12.100	-0.100	false
10205	30-Sep-2008	10.650	10.650	10.660	10.550	12.440	-0.014	false
10232	09-Dec-2013	25.200	23.350	29.750	23.201	25.200	0.019	false
10232	10-Dec-2013	110.035	23.400	196.670	23.400	196.670	3.366	true
10232	11-Dec-2013	24.010	23.520	25.000	23.310	25.250	-0.782	false
10256	28-Jun-2001	0.600	0.530	0.600	0.550	0.600	0.091	false
10256	29-Jun-2001	0.550	0.110	3.000	0.550	0.570	-0.083	false
10256	02-Jul-2001	0.550	0.500	0.550	0.490	0.550	0.000	false

Table 5: Examples of flagged observations by error detection procedure.

Description: Table of first ten detected errors. Each block shows period $t-1$ to $t+1$ for an error detected for time t. ID column contains PERMNO, i.e., the CRSP security identifier. PRC is the price series, i.e., an absolute value of price series from CRSP database. BID and ASK are the closing or inside quotes, depending on exchange under consideration. BIDLO and ASKHI are the closing (inside) quotes when the trading price is not available (Midpoint), and the daily low and high price otherwise. Mid indicates whether the price corresponds to quote midpoint.

Interpretation: The error identification procedure successfully identifies many problematic quotes that often generate huge artificial returns.