Significance Without Substance: How Overlapping Observations Inflate an Insider-Trading Signal

Fabio E. Juranek[1]

This version: June 5, 2026

Abstract

Cluster buying by corporate insiders—multiple insiders purchasing the same stock within a short window—is widely regarded as a strong predictor of future returns. Using point-in-time S&P 500 membership and EDGAR Form-4 filings from 2015 to 2024, we document a 12-week forward return spread between cluster-buy and no-activity stocks that is significant at $p \approx 1.6 \times 10^{-9}$ under a conventional two-sample test. This significance does not survive correction for the overlapping, cross-correlated structure of the underlying observations. Under Newey–West HAC standard errors, a stationary block-bootstrap, and non-overlapping subsampling, the p-value rises to approximately 0.6–0.7, and the mean weekly spread is statistically indistinguishable from zero (-0.17%, 95% CI $[-1.19\%, +0.85\%]$). We isolate two inflation channels and show that the dominant one is not autocorrelation per se but the pooling of overlapping within-week observations as if independent; the entire visible effect concentrates in the 2020 regime. The signal that appears overwhelmingly significant under naive testing is, under correct inference, not distinguishable from noise.

Keywords: insider trading, cluster buying, return predictability, overlapping observations, robust inference, market efficiency
JEL Classification: G11, G12, G14

[1]Independent research, Vienna, Austria. Contact: fabio.juranek@bhakwien13.at.

1 Introduction

The trading behavior of corporate insiders has long been studied as a source of return predictability. A particularly popular variant is *cluster buying*: the observation that when several insiders at the same firm buy stock within a short window, subsequent returns are unusually high. The intuition is appealing—coordinated purchases by informed agents should carry more signal than isolated trades—and the empirical pattern is easy to reproduce. A naive test on a decade of large-cap data yields a forward-return spread significant at $p \approx 10^{-9}$, a number that would seem to place the effect beyond reasonable doubt.

This paper argues that such confidence is misplaced, and that the apparent significance is largely a measurement artifact. The central question is narrow and concrete: *does a clustered insider-buying signal generate a return spread that is statistically distinguishable from zero once the dependence structure of the observations is correctly accounted for?* We find that it does not.

The problem lies in how the test observations are constructed. Forward returns measured at weekly frequency over a 12-week horizon overlap by eleven weeks, and cluster events occurring in the same week share common market movements. A conventional two-sample test treats each cluster observation as independent, so a single turbulent week—2020 being the extreme case—in which dozens of firms simultaneously register clustered buying is counted as dozens of independent pieces of evidence rather than as one correlated market event. Correcting this collapses the result.

This study connects directly to a companion analysis (Juranek, 2026), which showed that an apparent technical-trading edge disappears once survivorship bias is removed. The two findings share a structure: a result that looks compelling under naive measurement dissolves under a single methodological correction. Here the correction is not to the data universe but to the treatment of dependence among observations.

We make three contributions. First, we re-examine a widely cited predictability signal under inference methods that respect the overlapping, cross-correlated structure of the data—Newey–West HAC, a stationary block-bootstrap, and non-overlapping subsampling—and show that its significance does not survive. Second, we separate two distinct sources of p-value inflation, demonstrating that the dominant channel is the pooling of non-independent observations rather than serial autocorrelation alone. Third, we show that the residual visible effect is confined to a single market regime, with no significance outside it.

The remainder of the paper proceeds as follows. Section 2 reviews the relevant literature. Section 3 describes the data. Section 4 sets out the signal definition and the inference methods. Section 5 presents results. Section 6 discusses interpretation and limitations, and Section 7 concludes.

2 Related Literature

Three strands of work bear on this analysis.

The first concerns insider trading as a predictor of returns. A substantial literature documents that insider purchases, and clustered purchases in particular, precede positive abnormal returns (Lakonishok and Lee, 2001); the standard reading is that insiders trade on private information and that coordinated buying carries a stronger signal than isolated transactions. We take this literature as the source of the hypothesis under examination rather than as settled fact, and ask whether one widely reproduced version of the pattern withstands robust inference.

The second strand concerns the econometrics of overlapping observations. When returns are measured over horizons longer than the sampling interval, successive observations share data and are mechanically autocorrelated. Hansen and Hodrick (1980) first formalized the resulting bias in long-horizon return tests, and Newey and West (1987) provided the heteroskedasticity- and autocorrelation-consistent covariance estimator now standard for such settings. A parallel concern is cross-sectional dependence: observations drawn from the same period share common shocks, so treating them as independent overstates the effective sample size. Both problems are present here, and we show the second dominates.

The third strand concerns the broader credibility of empirical asset-pricing findings. Harvey et al. (2016) argue that much of the documented cross-section of returns fails to clear an appropriate significance hurdle once multiple testing is accounted for, and a wider replication literature has stressed how sensitive published effects are to specification and to the treatment of dependence. Our contribution is narrower but in the same spirit: a single signal, examined in detail, whose headline significance is an artifact of how its observations are pooled.

3 Data

3.1 Insider Transactions

Insider transactions are drawn from SEC EDGAR Form-4 filings, sourced from the SEC DERA quarterly flat-files covering 2014Q4 through 2025Q1. We retain open-market purchases by officers and directors, applying the following filters: transaction code P (open-market purchase), acquisition (disposition flag A), a per-share price above $1, and a positive share count. After filtering, the raw dataset comprises 202,047 qualifying purchases across 7,250 distinct tickers.

To avoid look-ahead bias, each transaction is dated by its filing date rather than its trade date, and the signal becomes active one business day after filing. Observations are then

aggregated to a weekly (Friday) frequency.

3.2 Universe and Prices

The test universe is the S&P 500, reconstructed on a point-in-time basis so that each stock enters the sample only while it was an actual index constituent. This is the same survivorship-free membership database used in the companion study (Juranek, 2026) and is essential here for the same reason: applying current index membership retroactively would silently drop delisted and removed firms. The analysis window is 2015 through 2024, bounded by the point-in-time membership data rather than by insider-data availability.

Prices are daily, adjusted for splits and dividends. Of the point-in-time constituents, 620 have price histories of sufficient length (at least 60 weeks) to support 12-week forward-return computation and are retained.

3.3 Signal Sparsity

Table 1: Sample composition and signal sparsity.

Item	Value
Qualifying insider buys	202,047
Distinct tickers	7,250
PIT S&P 500 tickers (usable prices)	620
Analysis window	2015–2024
Avg. cluster names per week	5.3
Cluster names as % of members	1.16%

Notes. Qualifying buys are open-market purchases (Form-4 code P, disposition A) by officers or directors with price above $1 and positive share count. The universe is point-in-time S&P 500 membership over 2015–2024.

The clustered-buying signal is sparse. Across the sample, an average of 5.3 stocks per week qualify as cluster-buy names, representing roughly 1.16% of index members in a given week. The distribution is highly uneven: the great majority of member-weeks register no clustered buying at all, while a small number of weeks—concentrated in market-stress episodes—carry many simultaneous cluster names. This unevenness is not incidental to the result; as Section 5 shows, it is the mechanism behind it.

4 Methodology

4.1 Signal Definition

For each stock and each week, we define a clustered-buying event as the presence of at least two distinct insiders purchasing the stock within the trailing four-week window. A stock-week with no qualifying purchases in that window is classified as a no-activity observation. The outcome variable is the 12-week forward return, measured from the week the signal becomes active.

The choice of a two-insider threshold over a four-week window follows the conventional definition of cluster buying; we examine sensitivity to the threshold in Section 5. Returns are computed on a point-in-time basis, so a stock contributes forward-return observations only for weeks in which it was an index constituent with sufficient price history.

4.2 The Inference Problem

The naive comparison of mean forward returns between cluster and no-activity observations is a two-sample test, and in its conventional form it treats every stock-week observation as an independent draw. Two features of the data violate this assumption severely.

First, *overlap*. A 12-week forward return measured each week overlaps its neighbor by eleven weeks. Successive observations therefore share most of their underlying return data, inducing strong positive autocorrelation in the series. This is the classic long-horizon problem of Hansen and Hodrick (1980): the effective number of independent observations is far smaller than the nominal count.

Second, and more importantly here, *cross-sectional dependence*. Cluster events are not spread evenly through time; they concentrate in market-stress weeks. When thirty firms register clustered buying in the same week and then share a common forward path driven by a market-wide rebound, those thirty observations are not thirty independent pieces of evidence. They are closer to a single observation of one market episode. A pooled test that counts them thirtyfold inflates the effective sample size—and hence the apparent significance—accordingly.

These two channels are conceptually distinct, and a central aim of this section is to separate them.

4.3 Robust Inference Methods

We assess the cluster-buy spread under four specifications, the first naive and the remaining three robust to the dependence structure described above. To keep the inference consistent with the cross-sectional model used in the companion work, the robust estimators are the

same implementations applied there. The spread is estimated by regressing the forward return on a cluster dummy,

$$r_{i,\,t\to t+12} = \alpha + \beta\,\mathrm{Cluster}_{i,t} + \varepsilon_{i,t}, \tag{1}$$

where β is the spread of interest.

Naive two-sample test. A Welch t-test on the pooled stock-week observations, equal variances not assumed. This reproduces the conventional approach and serves as the baseline.

Newey–West HAC. Equation (1) is estimated with heteroskedasticity- and autocorrelation-consistent standard errors. The lag length is set to 12, matching the forward horizon; this is mildly conservative relative to the data-driven Politis–White optimal block length, which selects 9 to 10 weeks.

Stationary block-bootstrap. We resample the weekly spread series in blocks to preserve its autocorrelation structure, using a block length equal to the 12-week horizon and 5,000 resamples, and construct the empirical confidence interval and bootstrap p-value from the resampled distribution.

Non-overlapping subsampling. As the most conservative check, we retain only every twelfth week, yielding genuinely non-overlapping 12-week observations on which a naive test is valid. Because this depends on the starting week, we repeat it across all twelve phase offsets and report the distribution of outcomes.

4.4 Isolating the Two Channels

To separate the overlap channel from the observation-unit channel, we additionally apply a naive i.i.d. t-test to the *weekly aggregated* spread series—that is, one spread observation per week rather than one per stock-week, with no HAC or bootstrap adjustment. If significance were driven primarily by serial autocorrelation, this aggregation alone would leave a strong result that only the HAC and bootstrap corrections remove. If instead it is driven by the pooling of non-independent within-week observations, significance should largely disappear at the aggregation step itself, before any autocorrelation correction is applied. Section 5 shows the latter.

5 Results

5.1 The Naive Result

Under the conventional two-sample test, clustered insider buying appears strongly predictive. Over 2015–2024, cluster-buy stocks earn a mean 12-week forward return of 4.77%, against 2.77% for no-activity stocks—a spread of +1.99% (95% CI [+1.35%, +2.64%]), significant at $p \approx 1.6 \times 10^{-9}$ (Welch $t = 6.05$, $n = 2{,}723$ cluster observations). Taken at face value, this is overwhelming evidence of a predictive effect.

5.2 Significance After Correction

Table 2: Cluster-buy spread: naive versus overlap-robust significance.

Method	Spread (%)	p (full)	p (ex-2020)	Verdict
Welch (naive, pooled)	+1.99	1.6×10^{-9}	0.239	2020-only
Newey–West HAC (lag 12)	−0.17	0.744	0.696	n.s.
Block-bootstrap (block 12)	−0.17	0.711	0.659	n.s.
Non-overlapping (1/12w)	−0.63	0.613	0.898	n.s.

Notes. 12-week forward return spread, cluster minus no-activity. The naive Welch test pools 2,723 overlapping, within-week cross-correlated observations as if independent, overstating significance by roughly eight orders of magnitude. The robust methods are the same implementations used for the cross-sectional model in the companion study. "n.s." denotes not significant at conventional levels. The ex-2020 column re-estimates each spread excluding calendar year 2020.

That significance does not survive correction for the dependence structure of the observations. Table 2 reports the spread under all four specifications. Under Newey–West HAC standard errors, the stationary block-bootstrap, and non-overlapping subsampling alike, the p-value rises to approximately 0.6–0.7. The point estimate of the mean weekly spread is −0.17% (95% CI [−1.19%, +0.85%])—not merely insignificant but slightly negative, and statistically indistinguishable from zero. The naive test overstates significance by roughly eight orders of magnitude.

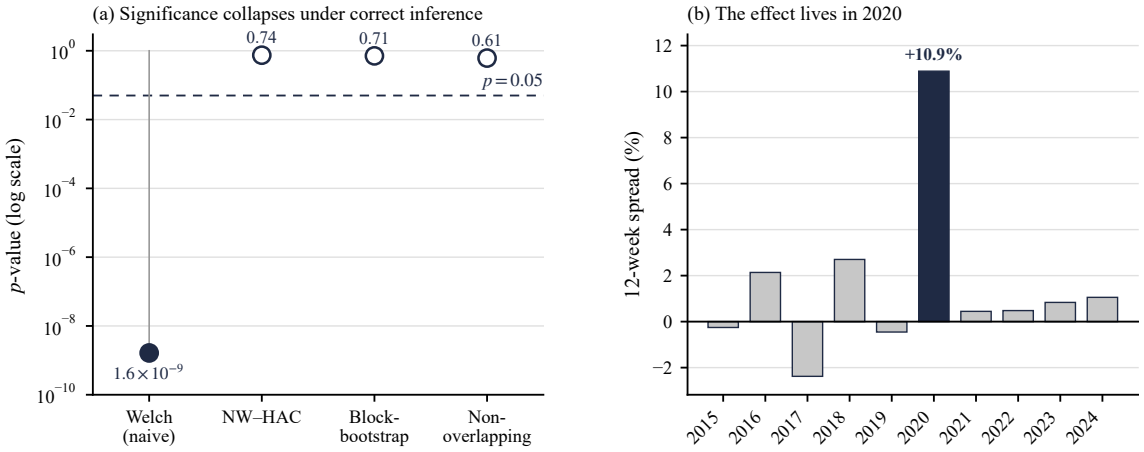


Figure 1: Significance collapse and regime concentration. Panel (a) plots the p-value of the cluster-buy spread under the naive Welch test and the three overlap-robust methods, on a logarithmic scale; the dashed line marks the 0.05 threshold. Panel (b) shows the annual 12-week cluster-minus-no-activity spread, with 2020 highlighted.

5.3 Which Channel Drives the Inflation

The two inflation channels of Section 4 can now be separated, and the result is clear: the dominant channel is the pooling of non-independent observations, not serial autocorrelation.

The decisive evidence is the aggregation step. A naive i.i.d. t-test applied to the weekly spread series—one observation per week, with no HAC or bootstrap adjustment—already yields $p = 0.585$. Significance is therefore lost at the moment observations are aggregated to the level at which they are plausibly independent, before any autocorrelation correction is applied. The subsequent move from i.i.d. $p = 0.585$ to HAC $p = 0.744$ is comparatively small, because the mean weekly spread is already near zero.

In other words, the headline $p \approx 10^{-9}$ does not arise because the weekly effect is strong and merely autocorrelated. It arises because the pooled test treats 2,723 overlapping, within-week cross-correlated stock-weeks as 2,723 independent draws. The slight sign reversal of the point estimate, from +1.99% pooled to −0.17% week-weighted, is itself a consequence of this reweighting: the positive pooled mean is carried by a few dense weeks that week-level weighting no longer overcounts. A single stress week with many simultaneous cluster names—2020 being the extreme—enters the pooled test as many independent confirmations of an effect, when it is closer to one observation of one market episode.

5.4 The Effect Is Confined to 2020

Consistent with this mechanism, the visible spread is concentrated almost entirely in the 2020 regime. Excluding 2020, even the naive Welch test loses significance ($p = 0.239$), and the mean of the annual cluster spreads excluding 2020 is +0.51%, compared with +10.9%

in 2020 alone. The robust methods remain insignificant whether or not 2020 is included (Table 2, ex-2020 column). Across all twelve phase offsets of the non-overlapping test, none produces $p < 0.05$, with a mean p-value of 0.43. No choice of sampling phase recovers the effect.

5.5 Robustness

The result is not an artifact of the signal definition. Raising the clustering threshold from two to three or more insiders does not strengthen the spread—if anything it weakens it, with the highest-cluster bucket underperforming the two-insider bucket—so the finding is not driven by an unfavorable choice of threshold. The conservative block length of 12 weeks exceeds the data-driven Politis–White selection (9–10 weeks), so the loss of significance is not an artifact of an overly long block.

5.6 Tradeability

Given that the spread is statistically indistinguishable from zero, the question of whether it can be traded is largely settled in advance; we confirm it directly for completeness. A long-only weekly tilt toward current cluster names, benchmarked against an equal-weight buy-and-hold of the same universe and charged 5 basis points per side of turnover, earns a Sharpe ratio of 0.52 against the benchmark's 0.70. The signal underperforms a passive hold of the same stocks. Beyond this, the signal's sparsity—an average of 5.3 names per week—would impose severe capacity and diversification constraints on any standalone implementation, but these are secondary: a spread that is not distinguishable from zero offers nothing to trade in the first place.

6 Discussion

6.1 Interpretation

The result admits a simple reading: at least on liquid large-cap stocks over the past decade, clustered insider buying carries no forward-return predictability that survives correct inference. The effect that appears in naive testing is not a weak-but-real signal degraded by noise; it is an artifact of how the test observations are counted. Once observations are aggregated to a level at which independence is plausible, the spread is indistinguishable from zero, and what visible effect remains is a single market regime.

This does not imply that insiders are uninformed, nor that insider transactions are uninformative in general. It speaks to one specific construction—a two-insider, four-week clustering rule applied to S&P 500 constituents, evaluated at a 12-week horizon—and to the way its

significance has been measured. Markets in which this signal is tested are deep and heavily scrutinized; the absence of a surviving edge there is consistent with, rather than surprising under, informational efficiency in large-cap names. The finding may not extend to small-cap or less liquid universes, where insider information plausibly diffuses more slowly, and we make no claim about those settings.

6.2 The Methodological Point

The broader lesson generalizes beyond insider trading. Wherever a predictive signal is evaluated on overlapping return horizons, and wherever its events cluster in time, a pooled test that treats observations as independent will overstate significance—often by many orders of magnitude. The inflation here was not subtle: eight orders of magnitude separate the naive and corrected p-values. Crucially, the larger share of that inflation came not from serial autocorrelation, the channel most commonly corrected with HAC standard errors, but from cross-sectional pooling of contemporaneous observations. A researcher who applied HAC alone, without first asking at what level observations are genuinely independent, would have corrected the smaller of the two problems and retained a result that is still essentially an artifact.

This mirrors the finding of the companion study (Juranek, 2026), in which an apparent technical-trading edge proved to be survivorship bias rather than skill. In both cases a single, identifiable methodological flaw—there in the construction of the universe, here in the counting of observations—was sufficient to manufacture a result that looked compelling. The recurring theme is that the credibility of a backtested signal depends less on the size of its headline statistic than on whether the procedure that produced it respects the structure of the data.

6.3 Limitations

Several limitations qualify the analysis. The universe is restricted to large-cap S&P 500 constituents, and the conclusions may not transfer to smaller or less liquid stocks. The horizon and clustering rule are fixed by convention rather than optimized, by design; we test a standard construction rather than search for the specification under which the signal performs best. Form-4 filing data, while comprehensive, may contain classification errors in the underlying flat-files. Finally, the tradeability assessment uses a simple weekly tilt with a flat transaction-cost assumption and does not model market impact in detail—an acceptable simplification here only because the underlying spread is already indistinguishable from zero, but one that would require refinement for any signal that did survive.

7 Conclusion

A clustered insider-buying signal that appears significant at $p \approx 10^{-9}$ under conventional testing is, under inference that respects the overlapping and cross-correlated structure of its observations, statistically indistinguishable from zero. We trace the inflation to its dominant source—the pooling of non-independent within-week observations rather than serial autocorrelation alone—and show that the residual effect is confined to the 2020 regime and survives no robustness check.

The contribution is deliberately narrow: one signal, examined carefully, whose headline significance does not survive. But the pattern it illustrates is general. Statistical significance and economic substance are not the same thing, and the gap between them is often a matter of how observations are counted rather than how strong the underlying effect is. For signals evaluated on overlapping horizons with time-clustered events, the level at which observations are genuinely independent is the first question to ask—before, not after, computing a p-value. Future work might extend the analysis to less liquid universes, where the conclusion may differ, or examine whether combining insider signals with orthogonal information recovers predictability that the standalone signal lacks.

References

Hansen, L. P. and Hodrick, R. J. (1980). Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis. *Journal of Political Economy*, 88(5), 829–853.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). . . . and the Cross-Section of Expected Returns. *Review of Financial Studies*, 29(1), 5–68.

Juranek, F. E. (2026). *Survivorship Bias, Not Skill: A Forensic Case Study of a Retail Equity Strategy.* SSRN Working Paper. `https://ssrn.com/abstract=6878588`.

Lakonishok, J. and Lee, I. (2001). Are Insider Trades Informative? *Review of Financial Studies*, 14(1), 79–111.

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3), 703–708.