Mapping Intraday Volatility Transmission Across 300 Stocks Using Attention-Based Causal Inference

Siddhant Sukhani Alexandre Roger Mahmood Alhusseini

1. Introduction and Motivation

Modern equity markets are deeply interconnected. A shock in one stock—such as a large sell order in a major index constituent—can ripple across related firms, sectors, or the broader market within minutes. These interactions occur rapidly due to algorithmic trading, order-book liquidity propagation, and correlated information flows. Traditional econometric models, typically operating on daily or hourly data, fail to capture these *intraday causal dynamics* that govern volatility transmission in real time.

Understanding these micro-temporal dependencies is critical for several reasons:

- **Intraday risk management:** anticipating cross-stock volatility spikes.
- **High-frequency portfolio allocation:** reducing exposure to correlated shocks.
- **Market microstructure research:** uncovering sectoral and systemic information flow channels.

We propose an attention-based causal inference framework that learns how volatility shocks propagate among 300 large-cap U.S. equities sampled every 5 minutes. The model identifies which stocks influence others and quantifies both the *direction* and *lag structure* of those influences.

2. Problem Description

Given high-frequency log-returns

$$r_{i,t} = \log\left(\frac{P_{i,t}}{P_{i,t-1}}\right),$$

for $N = 300$ stocks over 5-minute intervals, the goal is to predict short-term realized volatility:

$$\hat{\sigma}_{i,t+1} = f\left(\{r_{j,t-\tau_j:t}\}_{j=1}^N\right),$$

where τ_j is a learned lag denoting how long it takes for movements in stock j to influence stock i.

The objectives are:

1. Forecast next-interval (5-minute-ahead or 1-hour-ahead) realized volatility for each stock.
2. Recover a directed causal graph of volatility transmission among stocks with adaptive time lags.

3. Challenges

- **High-frequency noise:** microstructure effects such as bid-ask bounce and order clustering.
- **Extreme dimensionality:** 300 equities × thousands of 5-minute intervals.
- **Lag heterogeneity:** some pairs react instantaneously while others show delayed feedback.
- **Nonstationarity:** relationships evolve with regime shifts (e.g., earnings, macro news).
- **Interpretability:** standard neural models fail to reveal causal or sectoral pathways.

4. Dataset

4.1. Data Sources

- **Price & volume:** 5-minute returns data from Professor Markus Pelger's website.
- **Universe:** S&P 300 constituents (large-cap and mid-cap U.S. equities).
- **Period:** 2004–2016 (approx. 500k intervals per stock).

4.2. Target Variable

Realized volatility is computed over a rolling window of $n = 12$ intervals (approximately 1 hour):

$$\sigma_{i,t} = \sqrt{\frac{1}{n}\sum_{k=1}^n r_{i,t-k}^2}.$$

5. Preprocessing

- Compute log-returns and realized volatility at 5-minute granularity.
- Apply z-score normalization per stock.
- Remove overnight gaps and low-liquidity periods.
- Handle missing ticks via Kalman or spline interpolation.
- Introduce learnable lag parameters τ_j for temporal alignment.
- Chronological split (70% / 15% / 15%) with walk-forward validation.

6. Learning Method

6.1. Model Architecture

For each stock i, the model predicts its next-interval realized volatility using its own recent 5-minute data and attention-weighted information from other stocks:

$$\hat{\sigma}_{i,t+1} = \mathrm{MLP}\big([c_{i,t},\ r_{i,t}]\big),$$

where $r_{i,t}$ denotes the recent log returns of stock i, and $c_{i,t}$ represents the aggregated influence from other stocks via the attention mechanism. The multi-layer perceptron (MLP) learns to combine these components nonlinearly to forecast volatility.

Intuitively, the model for each stock learns how much of its future volatility can be explained by its own past versus the lagged signals arriving from other correlated equities.

6.2. Lagged Attention Mechanism

The attention mechanism identifies which other stocks influence a given stock's volatility, with what delay, and to what magnitude.

Lagged Histories. For each stock j, we construct a lagged sequence of past returns:

$$H_{j,t} = [r_{j,t-1}, r_{j,t-2}, \ldots, r_{j,t-L}],$$

where L is the lookback window length (e.g., 12 intervals ≈ 1 hour of 5-minute data).

Attention Weights. To determine which stocks and time lags are most relevant to predicting stock i's volatility, the model computes attention weights:

$$\alpha_{i \leftarrow j,t} = \text{softmax}\left(\frac{q_i^\top k_{j,t-\tau_j}}{\sqrt{d_k}}\right),$$

where q_i and $k_{j,t-\tau_j}$ are the query and key vectors for stocks i and j, respectively, and τ_j is a learned lag (in 5-minute units) that represents how long it takes for changes in stock j to influence stock i. The softmax operation normalizes the weights across all potential sources.

Aggregation. The attention-weighted information from all other stocks is aggregated into a context vector:

$$c_{i,t} = \sum_{j,\ell} g_j\, \alpha_{i \leftarrow j,t}\, v_{j,\ell,t-\tau_j},$$

where $v_{j,\ell,t-\tau_j}$ are value vectors containing encoded features of stock j's lagged returns, $\alpha_{i \leftarrow j,t}$ are the attention weights, and g_j is a causal gate that controls whether stock j exerts influence on others. The gates promote sparsity so that only a small subset of stocks drive cross-volatility effects.

This design enables the model to learn a directed and lagged network of volatility transmission among equities, uncovering causal structure dynamically as market conditions evolve.

6.3. Causal Regularization

To prevent overfitting and ensure interpretability, the total loss function includes several regularization components:

$$\mathcal{L} = \text{MSE}(\hat{\sigma}, \sigma) + \lambda \|g\|_{2,1} + \gamma\, \text{TV}(\alpha) + \eta\, \text{IRM},$$

where:

- $\text{MSE}(\hat{\sigma}, \sigma)$: mean squared error ensuring predictive accuracy.
- $\lambda \|g\|_{2,1}$: group-lasso penalty encouraging a sparse set of active causal gates g_j.
- $\gamma\, \text{TV}(\alpha)$: total variation penalty promoting temporal smoothness in attention weights.
- $\eta\, \text{IRM}$: invariant risk minimization term ensuring stability of learned relations across regimes.

This formulation yields sparse, interpretable, and stable causal dependencies between stocks, producing lag structures that remain consistent across varying market conditions.

7.   Evaluation

7.1.   Baselines

- High-dimensional Vector Autoregression (VAR) with Granger causality.
- Long Short-Term Memory (LSTM) without explicit lag modeling.
- Standard Transformer without causal regularization.

7.2.   Metrics

- RMSE, MAE, and directional accuracy.
- Diebold–Mariano test for statistical significance.

7.3.   Causal Validation

- Masking test: remove influential stock and assess impact on prediction.
- Permutation test and rolling-window stability analysis.

7.4.   Economic Evaluation

- Volatility-targeted intraday portfolio backtesting.
- Stress-period propagation analysis (e.g., sector-wide shock diffusion).
- Visualization of dynamic causal graphs between correlated sectors (e.g., technology, energy).

8.   Expected Results

- 5–10% improvement in intraday volatility forecasting relative to benchmarks.
- Interpretable causal graphs showing lagged dependencies (e.g., AAPL $\rightarrow$ NVDA with 10-min lag).
- Regime-consistent propagation patterns between sectors.
- Scalable framework for high-frequency volatility prediction.

9. Conclusion

This project extends attention-based causal inference to high-frequency equity markets, modeling how volatility shocks propagate across hundreds of stocks in near real time. By learning both the *direction* and *timing* of influence, the approach enables interpretable, data-driven insight into market microstructure and supports risk-aware intraday trading strategies.