Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Learned Attention or Mechanical Flattening? Diagnosing Index Gains in Spillover Networks

Ozan Nadirgil

onadirgil@gmail.com

Independent https://orcid.org/0009-0003-5584-0488

Research Article

Keywords: Systemic risk; Contagion; Frequency-domain spillovers; Graph attention; Shrinkage benchmark

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11353782/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Learned Attention or Mechanical Flattening? Diagnosing Index Gains in Spillover Networks

Ozan Nadirgil

Abstract

Spillover models are increasingly augmented with learned components, and when such a component operates within the connectedness measure itself, its most visible footprint is a shift in the index. This paper examines how much a shift of this kind can actually reveal. For this purpose, daily returns on thirteen assets, covering cryptocurrencies, commodities, equity indices for technology and for clean energy, and several macro-financial variables, are analysed over the period January 2018 through July 2026. A graph-based attention mechanism reweights each channel of transmission in the spectral variance decomposition obtained from a VAR re-estimated over rolling windows (GAT-FC-TVP-VAR). For the linear benchmark, system-wide connectedness has a mean of 55.06% and reaches its peak, 66.28%, during March 2020. Roughly 60% of all transmitted variance dissipates within five trading days. Transmission is clustered, since pairwise entries average 0.229 among the cryptocurrencies but only 0.014 among the commodities, and three equity indices, covering world stocks, technology and green energy, act as the dominant net senders of shocks. With attention, the index settles 3.72 percentage points above the benchmark, a gap that a moving-blocks bootstrap does not remove. Yet an ablation reveals that such a gain carries little information about the layer itself. Shrinking each decomposition towards equal weights through one parameter delivers an identical increase at a smaller validation error, while the objective used in training is minimised when no adjustment is made at all. Attention and shrinkage diverge only in how own shares of variance are reallocated across assets, and even this feature survives when attention is reduced to a scalar. This paper thus contributes a map of transmission among five asset groups together with an inexpensive benchmark that should accompany any connectedness study built around a learned element.

Keywords: Systemic risk; Contagion; Frequency-domain spillovers; Graph attention; Shrinkage benchmark.

1. Introduction

1.1. Motivation

Since the work of Diebold and Yilmaz (2009, 2012, 2014), variance-decomposition measures of connectedness have become the usual tool for tracing how shocks travel from one market to another. Their popularity owes much to practical convenience. A single VAR, that is, a vector autoregression, delivers a directional figure for each ordered pair of markets together with an aggregate index for the whole system, and rewriting that decomposition spectrally allows each of these figures to be separated by investment horizon (Baruník and Křehlík, 2018). Over the same period, the range of markets studied with these tools has widened. Clean-energy and technology stocks, cryptocurrencies and energy commodities each support a sizeable body of research on hedging and diversification, and each was exposed to the sell-off of March 2020, to the 2022 rate

hikes and, in that same year, to the failures of FTX and Terra-LUNA. Whether shocks of this kind propagate across these asset classes or remain confined to them matters for portfolio diversification no less than for systemic-risk surveillance.

Methodology has advanced in parallel. Connectedness frameworks with time-varying coefficients, frequency bands, quantiles or penalised estimation are by now standard, and a newer strand of research inserts learned elements, typically neural networks, into volatility and spillover models. This paper concentrates on one such element. Graph attention is stacked onto a connectedness model built on frequency bands, and the question posed is deliberately narrow. Does the decomposition reweighted by attention capture the structure of transmission more accurately than the purely linear version it supersedes? The evidence assembled in this paper suggests a largely negative answer, and much of the analysis is devoted to explaining why.

1.2. Related Literature

In the approach of Diebold and Yilmaz (2009, 2012), spillovers are quantified through a generalised decomposition, GFEVD, of the variance in a VAR's forecast errors (Koop et al., 1996; Pesaran and Shin, 1998), and Diebold and Yilmaz (2014) demonstrate that the matrix it produces can be viewed as a weighted, directed network. Time variation was originally handled with rolling windows. Antonakakis et al. (2020) replace rolling estimation with a VAR whose parameters vary over time (TVP-VAR) following the tradition of Primiceri (2005), estimated by Kalman filtering with the forgetting factors proposed by Koop and Korobilis (2014), which avoids losing observations at the beginning and removes the arbitrary choice of how long a window should be. Baruník and Křehlík (2018) break the same measures down by frequency, and both extensions are combined by Chatziantoniou et al. (2023) into a frequency connectedness model with time-varying parameters. Further refinements address high dimensionality with regularised estimators (Demirer et al., 2018) and tail dependence with quantile VARs (Ando et al., 2022), and Billio et al. (2012) propose Granger-causality networks as an alternative to variance decompositions. One feature common to all these variants is central to what follows. However much the coefficient matrices and the covariance of shocks are allowed to vary, each element of a connectedness table remains a fixed mapping of these two estimated objects. It is precisely this restriction that the layer studied in this paper seeks to relax.

Empirical work on the markets covered in this paper offers a reasonably coherent picture. An early contribution, Corbet et al. (2018), finds cryptocurrencies largely cut off from conventional financial and economic assets. Symitsi and Chalvatzis (2018), by contrast, document one-way spillovers in returns and volatility, together with two-way shock spillovers, between Bitcoin and firms in the energy and technology sectors, while Ji et al. (2019) observe growing integration of cryptocurrencies into a wide commodity network. Evidence on green and environmental assets is less uniform. According to Duan et al. (2023), cryptocurrencies pass only weak volatility to green and conventional assets alike, and clean coins are less connected than dirty ones. Hassan et al. (2022) detect no strong spillover from environmental attention towards cryptocurrencies to green bonds, commodities or equities linked to the environment, even though their quantile regressions

indicate that, under certain market conditions, most of these assets suffer from it. Using a TVP-VAR, Yousaf et al. (2024) uncover sizeable spillovers of both returns and volatility linking green cryptocurrencies with green equities, where the equity indices transmit and the coins receive on net, and Naeem and Karim (2021) show that green financial assets, clean-energy assets in particular, hedge Bitcoin effectively. Wang et al. (2024) report that energy tokens group with cryptocurrencies and become net receivers under extreme conditions, whereas renewable-energy equities are more tightly linked among themselves at median quantiles.

A further strand of research links cryptocurrencies with stock markets, commodities and policy uncertainty. Narayan and Kumar (2024), who pair several network techniques with a Bayesian graphical VAR, report that cryptocurrencies relate negatively to the remaining asset classes and that their ties to them were weak in the COVID-19 period. Bhanja et al. (2023) describe Bitcoin as fairly isolated from equities and from precious metals, with connectedness concentrated at short horizons and more pronounced in market downturns. Gaies et al. (2024) argue that cryptocurrencies can hedge equity exposure over short horizons when conditions are calm, and Younis et al. (2024) estimate a TVP-VAR where decentralised finance tokens are closely tied to Islamic equity indices and spillovers intensify in crises. Studies centred on policy and specific events add further nuance. Helmi et al. (2023) find that the way cryptocurrency markets react to announcements about digital currencies issued by central banks changes considerably through time, and Nasir et al. (2025), applying cross-quantilogram analysis together with quantile time-frequency connectedness, show that exchange tokens, green coins and energy tokens react unevenly to uncertainty surrounding crypto policy, depending on the quantile and the horizon. Esparcia et al. (2024) feed intraday volatility from a multiplicative component GARCH model into a TVP-VAR, identifying FTX's own FTT token as the principal net source of volatility shocks once the exchange had collapsed. Outside the cryptocurrency space, Mai et al. (2022) observe that spillovers across China's carbon markets were far larger while the pandemic lasted than either before or afterwards, and Wu and Hu (2024) rely on a quantile VAR to show how links between uncertainty about climate policy and Chinese financial and physical assets vary with the state of the market.

Three patterns recur across these studies, and each of them reappears in the evidence of Section 3. Cryptocurrencies usually form a distinct cluster that is only loosely tied to other markets, spillovers are concentrated at short horizons (Le et al., 2021; Bhanja et al., 2023), and connectedness increases when markets are under stress. Far less common is a design that brings together cryptocurrencies, commodities, clean-energy equities, technology equities and the principal macro-financial drivers in one system. Most of the papers cited cover only two or three of these groups, so a link such as that between clean-energy stocks and cryptocurrencies may be estimated without controlling for the world equity factor to which both are exposed.

Learned components have also entered volatility and spillover modelling through a different route. Araya et al. (2024) pass volatility estimates from generalised autoregressive conditional heteroscedasticity (GARCH) models (Bollerslev, 1986) to convolutional and recurrent networks, Liu et al. (2025) embed a long short-term memory (LSTM) network in the Realized GARCH specification, and Sahiner (2024) couples dynamic conditional correlations with an LSTM to construct an early-warning indicator of contagion. For problems in-

volving many assets, graph neural networks are a natural candidate (Wu et al., 2021). Graph attention, introduced by Veličković et al. (2018), has since been applied to infer the inter-firm relations that drive momentum spillovers (Cheng and Li, 2021) and to predict returns on stock indices across countries (Lee et al., 2025). Zhang et al. (2025) forecast realised volatility with graph neural networks that allow for spillovers. In their results, nonlinear spillover terms raise accuracy at horizons up to one week, information from neighbours several hops away adds little that is clearly useful, and a quasi-likelihood loss outperforms a squared-error loss. The closest antecedent of the present paper is Morshed (2025), who initialises a graph neural network from TVP-VAR estimates, compares it with the Diebold–Yilmaz index and interprets the learned edges as an explainable spillover network.

Embedding a learned element within the connectedness measure raises two difficulties. One concerns evaluation. Forecasts can be compared with what actually happens, whereas a reweighted variance decomposition has nothing observable to be compared with, so its effect is seen most clearly as a shift in the index itself. Taken by itself, such a movement is weak evidence. The index records nothing but the average off-diagonal share across rows, so it rises whenever a transformation moves weight towards off-diagonal cells, informative or not. The other difficulty concerns interpretation. Weights produced by attention are commonly interpreted as signalling which inputs matter, yet research in natural language processing has found that such an interpretation can fail (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). The connectedness studies surveyed above address neither difficulty, and it therefore remains unclear whether a layer learned at this stage improves measurement or merely alters it. This paper takes up both difficulties.

1.3. Contribution

At the core of the analysis lies a frequency connectedness VAR with time-varying parameters (FC-TVP-VAR), fitted to daily returns of thirteen series covering January 2018 through July 2026. An attention layer, defined over a graph of the markets, then reweights its variance decomposition in each window. The study compares the GAT-FC-TVP-VAR thus obtained with its linear counterpart in two ways. The conventional comparison tests whether index levels differ between the extended model and its benchmark. The second, less conventional, comparison is an ablation whereby the trained attention is replaced in turn by equal weights, a network that has not been trained, shuffled asset labels, a training sequence with its time order scrambled, shrinkage with one parameter towards equal weights, and an attention matrix reduced to a scalar. By conventional criteria, attention appears superior. Measured connectedness rises 3.72 percentage points (pp), positive gaps occur in 1,860 out of 1,884 windows, and a moving-blocks bootstrap leaves the difference intact. The ablation overturns this verdict. One shrinkage parameter is sufficient to replicate the gain precisely, and its validation error is below the fitted layer's. The training loss itself, moreover, is lowest when no change is made to the decomposition.

This design allows the paper to make three contributions. First, it maps return transmission among five asset groups within one system, documenting a dense block of cryptocurrencies, a commodity block that

stands apart and spillovers that are largely exhausted within five trading days, together with a core of net transmitters formed by the world equity index and both sector indices. Second, it proposes a simple benchmark for connectedness studies that rely on learned components. Any improvement in the index ought to be weighed against what a shrinkage towards equal weights, calibrated on that same index, would deliver, because such an improvement requires no learning at all. Third, it identifies what genuinely sets attention apart from shrinkage. Attention reverses the direction in which own shares are reallocated, drawing most heavily on markets that already depend least on their own shocks, but this behaviour follows from combining a bilinear score with a softmax, not from any estimated parameter. Specification weaknesses exposed by the ablation, such as redundant attention heads and a nearly trivial training target, are reported as findings in their own right rather than set aside.

The remaining sections are arranged as follows. Section 2 describes the data and both models, Section 3 reports the empirical findings and the ablation, and Section 4 draws the main conclusions.

2. Dataset and Methods

2.1. Sample Construction and Preliminary Tests

Investing.com provides all data, which consist of daily closes for fourteen series observed between 18 January 2018 and 30 July 2026. These series are organised in five groups. Two of them contain sector stock indices, for technology and for clean energy, a third contains cryptocurrencies, a fourth contains energy and commodity contracts, and the fifth gathers variables capturing macro-financial conditions and global risk. Energy and commodity exposure comes from three series. OIL and GAS are NYMEX futures on WTI crude and on natural gas respectively, while GOLD measures the dollar spot price of one troy ounce of gold. Sector indices enter as levels, whereas cryptocurrency prices are quoted against the US dollar. The macro-financial group includes MSCI's All Country World Investable Market Index (ACWI IMI), which serves as a wide-coverage world equity index, together with DXY, which tracks the dollar against a basket of major currencies, the price of futures on ten-year US Treasuries, and the VIX, CBOE's index of implied volatility. Using futures prices rather than yields for Treasuries expresses all fourteen series as log returns, which ensures a common unit across the whole decomposition. The VIX is likewise log-differenced for this reason.

While cryptocurrencies trade seven days a week, other markets close at weekends and observe different public holidays. Only trading days common to the fourteen markets are therefore retained. Of 2,338 dates when one or more markets outside the cryptocurrency group were open, 2,135 meet this requirement, which yields $T = 2{,}134$ return observations for each series. No value is interpolated and no price is carried forward, so that no zero return is generated artificially. Exact zeros are most common in Treasury futures (1.9%) and in DXY (1.6%), which both carry only two decimals, and stay below 0.5% elsewhere. Compared with filling non-trading days with the last available price, this rule is more demanding and avoids artificial co-movement of the kind that stale prices would produce at exactly those short horizons targeted by the spectral split.

The sample starts where the data allow, since the Ethereum series begins only on 18 January 2018, while rolling estimation calls for a panel without gaps. Even so, the period covers two full cycles in cryptocurrency markets and a sequence of distinct episodes, from the volatility spike of early February 2018 and the COVID-19 crash to the ample liquidity that lasted through 2020 and into 2021, the rate hikes of 2022 that coincided with the FTX and Terra-LUNA failures, and the rebound that came afterwards. Regime shifts of this kind are precisely what time-varying parameters are designed to accommodate.

Table 1 lists descriptive statistics, the full name of every series and the code used for it from here on. Table 2 presents the normality test of Jarque and Bera (1987), which rejects at 1% for all fourteen series. Gaussian innovations can therefore be excluded, although this result is by itself silent on nonlinearity in conditional means. Stationarity is checked with the Dickey–Fuller test in its augmented form, ADF (Dickey and Fuller, 1979), which includes an intercept and lets the Akaike criterion pick at most 26 lags (see Table 3). Rather than relying on asymptotic tables, p-values and critical values come from simulated null distributions for the present sample size. Every return series rejects a unit root at 1%. For prices in levels, the VIX alone rejects (ADF statistic −4.524, p-value 0.0001), as expected for an index that mean-reverts by design, and the other thirteen level series are I(1). The VIX is nonetheless log-differenced so that units stay homogeneous, a choice Section 4 revisits when limitations are discussed.

To detect conditional heteroscedasticity, Engle's (1982) Lagrange multiplier test for ARCH effects (ARCH-LM) is applied with lag orders of one, five and ten. It regresses squared residuals on lagged squared residuals, and rejecting constancy of the conditional variance points to volatility clustering.

Rejections dominate Table 4. With five or ten lags, rejection at 1% is universal. At one lag, the only exception is natural gas, which rejects at 10% but not at 5% ($p = 0.060$). Volatility clustering is therefore present in every market of the panel, a pattern in line with news reaching markets in bursts and with shocks spreading between them via covariances that change over time.

Jointly, the normality and ARCH-LM tests exclude a Gaussian model with constant variance and support letting the covariance of innovations evolve through time. Since neither of them tests whether conditional means are linear, neither by itself establishes that a nonlinear correction is warranted. This question is addressed directly with the BDS statistic (Brock et al., 1996), computed on residuals from the fitted system (Table 6).

Variable	Code	Mean	Median	Std. Dev.	Skewness	Excess kurtosis
TECHNOLOGY						
NASDAQ 100 Technology	NAST	0.000667	0.001295	0.018986	-0.2064	4.9287
MSCI World Information Technology	MSCIT †	0.000764	0.001586	0.015893	-0.2938	7.0664
CLEAN ENERGY						
NASDAQ Clean Edge Green Energy	NASC	0.000391	0.001196	0.024099	-0.2068	2.9371
S&P Global Clean Energy	SPC	0.000282	0.000271	0.016873	-0.4306	6.4021
CRYPTOCURRENCY						
Bitcoin	BTC	0.000829	0.000644	0.040281	-0.9569	12.8726
Ethereum	ETH	0.000310	0.000061	0.054138	-0.7270	10.6765
BNB	BNB	0.001750	0.001514	0.054002	-0.4133	15.7692
COMMODITIES AND ENERGY						
Crude oil futures (WTI, NYMEX)	OIL	0.000126	0.001518	0.033012	-0.4713	92.8040
Gold spot (USD per troy ounce)	GOLD	0.000529	0.000753	0.010644	-0.6821	7.1410
Natural gas futures (NYMEX)	GAS	-0.000039	0.000358	0.043446	-1.1859	28.4238
MACRO-FINANCIAL AND GLOBAL RISK						
MSCI ACWI IMI	ACWI	0.000328	0.000751	0.009874	-1.0757	15.5651
US Dollar Index	DXY	0.000046	0.000000	0.004172	-0.1745	1.6992
US 10-year Treasury note futures	ZN	-0.000056	0.000000	0.003603	0.1388	1.9904
CBOE Volatility Index	VIX	0.000157	-0.007917	0.082159	1.2950	8.1932

Table 1 Summary Statistics of Daily Log Returns

(Daily log returns, N = 2,134 per series, covering 19 January 2018 through 30 July 2026. Series are ordered by asset class, and the abbreviations listed under Code are used for each series throughout the text. The symbol † flags the series dropped because of its VIF (Table 5). That series still appears up to Table 5, allowing readers to verify the decision, but it plays no role in any later estimation.)

Code	JB statistic	p-value	Skewness	Excess kurtosis
NAST	2175.16	<0.0001	-0.2064	4.9287
MSCIT †	4470.71	<0.0001	-0.2938	7.0664
NASC	782.27	<0.0001	-0.2068	2.9371
SPC	3710.42	<0.0001	-0.4306	6.4021
BTC	15059.50	<0.0001	-0.9569	12.8726
ETH	10323.34	<0.0001	-0.7270	10.6765
BNB	22171.49	<0.0001	-0.4133	15.7692
OIL	765881.91	<0.0001	-0.4713	92.8040
GOLD	4699.68	<0.0001	-0.6821	7.1410
GAS	72336.90	<0.0001	-1.1859	28.4238
ACWI	21953.53	<0.0001	-1.0757	15.5651
DXY	267.55	<0.0001	-0.1745	1.6992
ZN	359.13	<0.0001	0.1388	1.9904
VIX	6565.25	<0.0001	1.2950	8.1932

Table 2 Normality Tests (Jarque–Bera)

(The statistic combines sample excess kurtosis and skewness. Kurtosis is reported net of three, which matches the convention of the descriptive statistics. Every series departs significantly from normality, as is typical of heavy-tailed, leptokurtic returns. Because conditional heteroscedasticity produces a non-normal unconditional distribution even when conditional means are linear, the test argues for a specification free of Gaussian innovations, not for a nonlinear conditional mean per se. † flags the series dropped because of its VIF (Table 5).)

Code	ADF (level)	p-value	ADF, log returns	p-value	Integration order
NAST	0.2639	0.9762	-51.9381	<0.0001	I(1)
MSCIT †	0.8266	0.9946	-50.1668	<0.0001	I(1)
NASC	-1.7989	0.3828	-47.8754	<0.0001	I(1)
SPC	-2.0099	0.2850	-43.0573	<0.0001	I(1)
BTC	-1.3090	0.6289	-48.7180	<0.0001	I(1)
ETH	-2.1510	0.2270	-49.1648	<0.0001	I(1)

Code	ADF (level)	p-value	ADF, log returns	p-value	Integration order
BNB	-1.4661	0.5520	-11.6662	<0.0001	I(1)
OIL	-2.1322	0.2341	-30.9232	<0.0001	I(1)
GOLD	0.4100	0.9835	-45.8958	<0.0001	I(1)
GAS	-2.3564	0.1564	-48.7287	<0.0001	I(1)
ACWI	0.7798	0.9939	-14.2893	<0.0001	I(1)
DXY	-2.2362	0.1952	-45.1082	<0.0001	I(1)
ZN	-0.6290	0.8618	-47.0906	<0.0001	I(1)
VIX	-4.5240	0.0001	-16.1285	<0.0001	I(0)

Table 3 Unit-Root Tests (ADF)

(Each regression includes an intercept, with lag length chosen according to Akaike's criterion. P-values and critical values are simulated, using 120,000 replications of a unit-root process at the actual sample size. At 1% the simulated critical values are −3.43 for returns and −3.44 for prices in levels, within 0.01 of the MacKinnon (1996) response surface. All return series are stationary, whereas among price levels the VIX alone is, a result of how that index is built and not of any property of asset prices. † flags the series dropped because of its VIF (Table 5).)

Code	LM, one lag	p-value	LM, five lags	p-value	LM, ten lags	p-value
NAST	243.75	<0.0001	374.28	<0.0001	386.22	<0.0001
MSCIT †	237.08	<0.0001	380.45	<0.0001	410.88	<0.0001
NASC	47.66	<0.0001	354.55	<0.0001	372.14	<0.0001
SPC	94.13	<0.0001	377.53	<0.0001	390.33	<0.0001
BTC	8.94	0.0028	34.97	<0.0001	36.21	<0.0001
ETH	25.29	<0.0001	56.77	<0.0001	58.48	<0.0001

Code	LM, one lag	p-value	LM, five lags	p-value	LM, ten lags	p-value
BNB	54.78	<0.0001	108.57	<0.0001	125.28	<0.0001
OIL	613.56	<0.0001	893.82	<0.0001	917.73	<0.0001
GOLD	51.25	<0.0001	212.61	<0.0001	223.72	<0.0001
GAS	3.53	0.0602	219.23	<0.0001	227.17	<0.0001
ACWI	171.26	<0.0001	590.62	<0.0001	661.81	<0.0001
DXY	131.52	<0.0001	227.83	<0.0001	267.18	<0.0001
ZN	93.32	<0.0001	238.66	<0.0001	243.61	<0.0001
VIX	70.90	<0.0001	97.15	<0.0001	104.97	<0.0001

Table 4 Tests for ARCH Effects

(Without ARCH effects, the statistic, n times R^2, follows a $\chi^2(q)$ distribution, q denoting how many lags are used. Every cell rejects at 1% apart from GAS with one lag, which rejects only at 10%. † flags the series dropped because of its VIF (Table 5).)

Multicollinearity is gauged through variance inflation factors (VIF) computed on log returns (Table 5). The two technology benchmarks, NAST and MSCIT, are nearly indistinguishable (r = 0.948), so retaining both would leave the estimates badly conditioned, with factors of 13.248 for MSCIT and 12.453 for NAST once all fourteen series are included. MSCIT, which has the higher factor, is consequently dropped from every subsequent step. There is also an economic rationale for this choice. The world equity exposure common to MSCIT and ACWI is captured by ACWI directly, whereas NAST gives a cleaner reading of technology as a sector. Two caveats apply. A VIF concerns how precisely the coefficients of a single regression equation are estimated, while in a variance decomposition the correlation among series is informative and not a problem. Removing MSCIT is therefore warranted by how well the rolling estimates are conditioned numerically, not by any bias in connectedness. As the rightmost column of that table confirms, the removal resolves the issue. NAST's factor falls to 5.010 (from 12.453) and ACWI's to 5.023 (from 6.158), whereas the other series are barely affected, NASC showing the next largest drop, to 5.766 from 5.872. That value remains the highest among the 13 series kept and is still comfortably under the usual cut-off of ten.

#	Variable	VIF, all 14 series	VIF, 13 series kept	Severity	Interpretation
1	MSCI World Information Technology (MSCIT)	13.248	excluded	Severe	Almost identical to NAST (r = +0.948), and therefore removed to avoid counting the same technology portfolio twice.
2	NASDAQ 100 Technology (NAST)	12.453	5.010	Severe	Inflated mainly through MSCIT, given a correlation of +0.948. Declines to 5.01 after that index is dropped.
3	MSCI ACWI IMI (ACWI)	6.158	5.023	High	Loads on the world equity factor shared with MSCIT, given a correlation of +0.872. Acceptable, as coefficients remain identified despite wider standard errors.
4	NASDAQ Clean Edge Green Energy (NASC)	5.872	5.766	High	Shares a clean-energy factor with SPC, given a correlation of +0.824. Acceptable, as coefficients remain identified despite wider standard errors.
5	S&P Global Clean Energy (SPC)	3.847	3.846	Moderate	Moderate correlation (r = +0.824) with NASC, reflecting a shared clean-energy factor. No action needed.
6	Ethereum (ETH)	3.598	3.597	Moderate	Moderate correlation (r = +0.823) with BTC, reflecting a shared cryptocurrency factor. No action needed.
7	Bitcoin (BTC)	3.336	3.336	Moderate	Moderate correlation (r = +0.823) with ETH, reflecting a shared cryptocurrency factor. No action needed.
8	CBOE Volatility Index (VIX)	2.143	2.123	Moderate	Negative correlation (r = -0.699) with MSCIT, the familiar inverse link between implied volatility and stock returns. No action needed.
9	BNB (BNB)	2.123	2.123	Moderate	Moderate correlation (r = +0.704) with ETH, reflecting a shared cryptocurrency factor. No action needed.
10	US Dollar Index (DXY)	1.490	1.478	Low	Weakly related to the other series (its strongest correlation, -0.436, is with GOLD).
11	Gold spot (USD per troy ounce) (GOLD)	1.325	1.325	Low	Weakly related to the other series (its strongest correlation, -0.436, is with DXY).
12	US 10-year Treasury note futures (ZN)	1.268	1.268	Low	Weakly related to the other series (its strongest correlation, -0.341, is with DXY).
13	Crude oil futures (WTI, NYMEX) (OIL)	1.094	1.093	Low	Weakly related to the other series (its strongest correlation, +0.224, is with ACWI).
14	Natural gas futures (NYMEX) (GAS)	1.013	1.013	Low	Weakly related to the other series (its strongest correlation, +0.091, is with OIL).

Table 5 Variance Inflation Factors

(Each factor is obtained from an auxiliary OLS regression with an intercept, VIF being the reciprocal of one minus R^2_i, with R^2_i denoting the fraction of series i's variance explained by the other series. The column for all 14 series covers the complete set, and the 13-series column describes the system after MSCIT has been removed.)

Adding attention is justified only if a linear model fails to absorb all systematic dependence, and this condition is testable. For this purpose, the BDS statistic (Brock et al., 1996) is calculated from what a full-sample VAR(1) leaves unexplained for the 13 return series kept after screening, the very specification underlying

every rolling decomposition. The null states that residuals are i.i.d., and because no particular alternative is specified, the test reacts alike to omitted linear dynamics, to heteroscedasticity in conditional variances and to dependence of a nonlinear kind. Standardised values are given for four embedding dimensions, $2 \leq m \leq 5$, with the distance parameter set at one standard deviation. Table 6 applies the test to the raw residuals in Panel A. In Panel B the residuals are first divided by their conditional volatility, taken from GARCH estimates with one ARCH and one GARCH lag (Bollerslev, 1986), which distinguishes dependence caused by volatility clustering from dependence of any other origin.

Series	m = 2	m = 3	m = 4	m = 5
Panel A, residuals before filtering				
NAST	7.67***	9.71***	11.22***	12.83***
NASC	5.54***	7.63***	8.79***	10.06***
SPC	4.37***	6.20***	7.07***	7.82***
BTC	5.01***	5.78***	6.22***	7.03***
ETH	4.69***	5.15***	5.61***	5.94***
BNB	9.12***	10.17***	10.08***	10.77***
OIL	8.95***	9.79***	10.34***	10.99***
GOLD	2.63***	4.74***	6.14***	7.54***
GAS	6.19***	8.52***	9.79***	11.15***
ACWI	9.10***	10.50***	11.28***	12.29***
DXY	2.25**	2.35**	3.22***	3.53***
ZN	5.82***	7.53***	8.46***	9.38***
VIX	7.00***	7.71***	8.09***	7.99***
Panel B, residuals after GARCH filtering				
NAST	1.90*	2.75***	2.81***	2.58***
NASC	0.63	0.78	0.58	0.53
SPC	-0.38	-0.08	-0.12	-0.16
BTC	1.15	1.16	1.16	1.48
ETH	0.89	0.88	1.21	1.43
BNB	2.19**	2.27**	1.66*	1.42

Series	m = 2	m = 3	m = 4	m = 5
Panel A, residuals before filtering				
OIL	1.67*	1.38	0.84	0.37
GOLD	-1.61	-1.08	-0.91	-0.68
GAS	0.91	1.69*	1.72*	1.97**
ACWI	1.10	0.95	0.88	0.78
DXY	-0.02	-0.73	-0.39	-0.47
ZN	0.63	0.41	-0.19	-0.75
VIX	1.03	1.14	1.38	1.15

Table 6 Independence Tests (BDS) on Linear VAR Residuals
*(Standardised BDS statistics, calculated from what a full-sample VAR(1) leaves unexplained for the 13 series that remain after screening, with embedding dimension m between 2 and 5 and a distance parameter equal to one standard deviation, estimated from the sample. Asterisks *, ** and *** denote significance against the i.i.d. null at 10%, 5% and 1%. Before testing in Panel B, each series of residuals is scaled by conditional volatility taken from GARCH estimates with one ARCH and one GARCH lag.)*

For the raw residuals, independence fails at 1% in 50 out of 52 pairs of series and dimension, the exceptions being DXY at the two lowest dimensions, where rejection occurs at 5%. The statistics are sizeable for all groups and increase with m, peaking for NAST (between 7.67 and 12.83), ACWI (between 9.10 and 12.29) and WTI crude (between 8.95 and 10.99). Structure thus remains in what the linear VAR leaves unexplained, and this residual structure provides the empirical motivation for a correction based on attention.

The second panel reveals the main source of this structure. Once the residuals are standardised by GARCH volatility, the statistics drop sharply. Only six combinations out of 52 are still significant at 5%, and ten series out of thirteen have no dimension that is significant at that level. Conditional heteroscedasticity therefore explains most of the departure from i.i.d. behaviour, in agreement with Table 4. Some dependence persists after filtering in three cases, namely NAST for m of three or more (between 2.58 and 2.81) and BNB, an exchange token, for dimensions two and three (2.19 at the lower dimension, 2.27 at the higher), while natural gas does so at dimension five (1.97). Dependence that outlives a filter for volatility suggests discrete, event-type behaviour more than smoothly evolving variance.

These results carry two implications for the empirical design. What dependence remains after filtering is genuine yet modest, confined to a handful of series and of moderate size. Under such conditions a learned correction might add new information, instead of simply re-fitting volatility dynamics that GARCH already captures. Moreover, since conditional heteroscedasticity explains most raw rejections, any improvement credited to attention must be carefully split between volatility clustering, which linear decompositions average out, and the nonlinear part that the filter leaves intact. Section 3.4 returns to this issue.

2.2. Modelling Strategy

Section 2.1 documents volatility clustering, heavy tails and a degree of residual nonlinearity, features that put pressure on any linear, constant-parameter description of how shocks move between markets. The approach adopted in this paper proceeds in two stages. The first estimates a frequency connectedness model over rolling windows so that both coefficients and covariances may vary through time. The second passes each window's variance decomposition to a graph-based attention layer, which reweights its elements by means of one learned matrix, allowing each transmission channel to count for more or less depending on the broader exposures of the markets at its two ends. Figure 1 outlines the procedure.

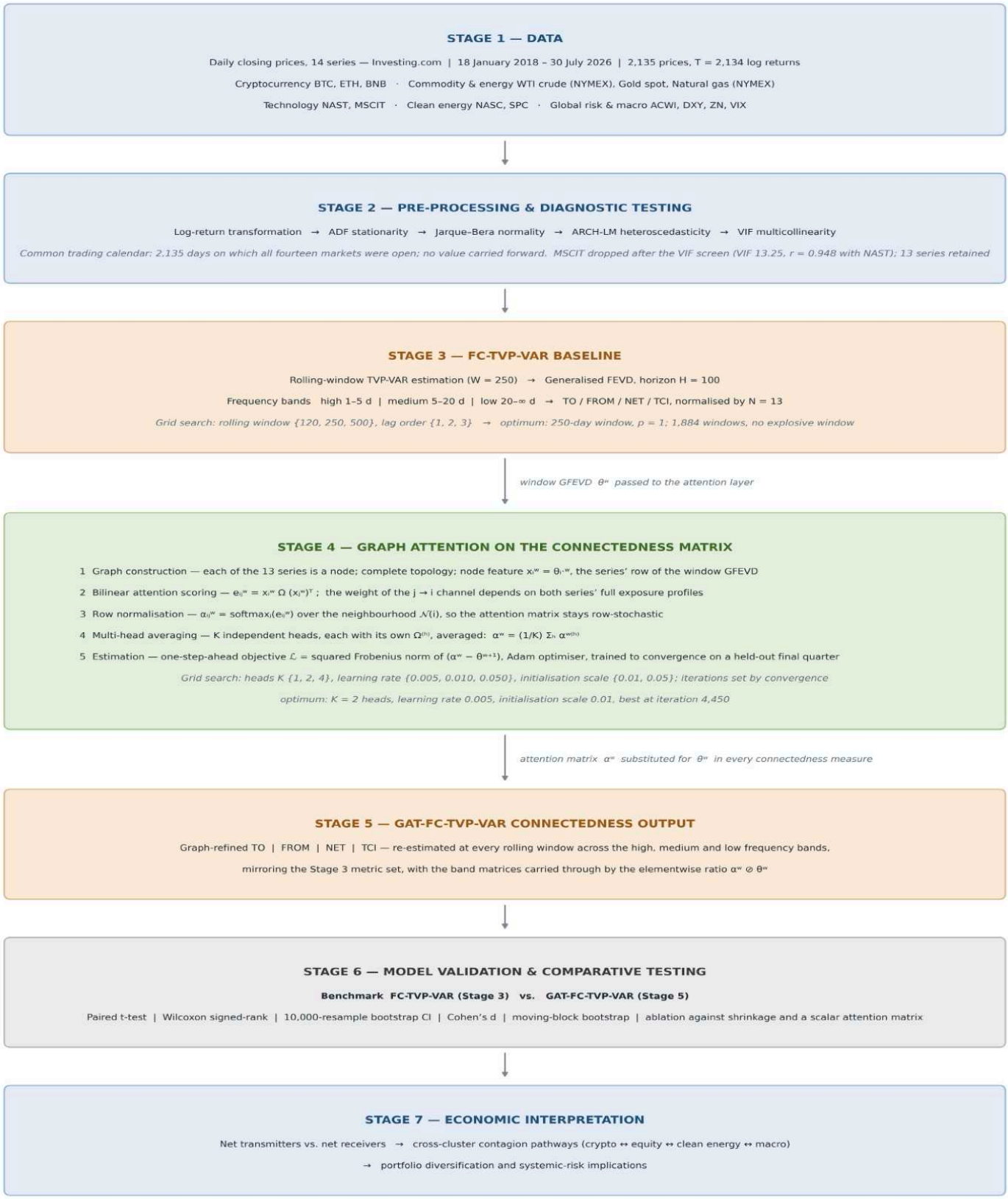


Figure 1 Overview of the Estimation Steps

2.2.1. Linear Frequency Connectedness Benchmark

In the FC-TVP-VAR, a VAR with time-varying parameters is paired with Baruník and Křehlík's (2018) spectral decomposition, which allows spillovers to be separated by horizon in addition to direction. Time variation comes from re-estimating coefficient and covariance matrices over windows of constant length that

roll forward through time, and TVP-VAR is used only in this wider sense. The Kalman-filter estimator that relies on forgetting factors (Koop and Korobilis, 2014; Antonakakis et al., 2020) is a different tool serving the same end. Because that estimator is not employed, there is no decay or forgetting parameter to report, and window length takes over its role. That length minimises one-step forecast errors over the candidate values (Table 9) and equals 250 days. This set-up accommodates gradual drift in the parameters, whereas abrupt structural breaks can only be approximated, since jumps are not modelled. As in Antonakakis et al. (2020), Equation (1) states the model, in which s_t and ϵ_t are vectors of dimension N × 1, α_t gathers the coefficients on lag n, for n from 1 to p, in a square matrix of order N that evolves through time, with p lags in total, and Σ, also indexed by time, is the covariance matrix of innovations. The normality assumed by Equation (1) is what Pesaran and Shin (1998) require to derive generalised impulse responses. Rolling least squares does not need this assumption. The normality rejections of Table 2 therefore bear on interpreting the decomposition, whereas coefficient estimates stay consistent.

$$s_t = \sum_{n=1}^{p} \alpha_{nt} s_{t-n} + \epsilon_t, \epsilon_t \sim N(0, \Sigma_t) \tag{1}$$

Equation (2) sets out the GFEVD of Koop et al. (1996) and Pesaran and Shin (1998). Here $\tilde{\omega}^{g}_{ijt}(H)$ measures how much a shock to variable j contributes to variable i's prediction error variance at horizon H. Normalising by the total over all N sources makes these contributions add up to one for every i and every date.

$$\tilde{\omega}^{g}_{ijt}(H) = \frac{\omega^{g}_{ijt}(H)}{\sum_{j=1}^{N} \omega^{g}_{ijt}(H)} \tag{2}$$

A spectral counterpart to the GFEVD follows when Equation (2) is evaluated across frequencies, as in Baruník and Křehlík (2018). Equation (3) restricts the decomposition to a band D of frequencies, where $\tilde{\omega}^{g}_{ijt}(D)$ is the fraction of i's prediction error variance in band D that shocks to j account for. Adding up the bands splits total connectedness into long-, medium- and short-horizon components.

$$\tilde{\omega}^{g}_{ijt}(D) = \frac{\omega^{g}_{ijt}(D)}{\sum_{j=1}^{N} \omega^{g}_{ijt}(D)} \tag{3}$$

Directional and aggregate measures follow Diebold and Yilmaz (2012). TO, defined in Equation (4), is the connectedness running from variable j towards all others, namely the part of their prediction error variances that stems from shocks to j. FROM, defined in Equation (5), is the part of i's prediction error variance due to shocks hitting any other variable, and NET in Equation (6) subtracts FROM from TO. A positive NET marks a series that transmits shocks on balance and a negative one a series that receives them. Equation (7) defines total connectedness, abbreviated TCI, in proportion form, while Equation (8) gives its rolling counterpart for window w in percent. As in Diebold and Yilmaz (2012), TCI values are divided by N, the count of variables, as opposed to the N − 1 used in the adjusted index proposed by Antonakakis et al. (2020). Because N equals 13, each reported TCI equals the total off-diagonal share over 13.

$$TO_{jt}(D) = \sum_{i=1, i \neq j}^{N} \tilde{\omega}_{ijt}^{g}(D) \quad (4)$$

$$FROM_{it}(D) = \sum_{j=1, j \neq i}^{N} \tilde{\omega}_{ijt}^{g}(D) \quad (5)$$

$$NET_{it}(D) = TO_{it}(D) - FROM_{it}(D) \quad (6)$$

$$TCI_{t}(H) = \sum_{D} TCI_{t}(D) = \sum_{D} \left(N^{-1} \sum_{i=1}^{N} \left(TO_{it}(D) \right) \right) = \sum_{D} \left(N^{-1} \sum_{i=1}^{N} \left(FROM_{it}(D) \right) \right) \quad (7)$$

$$TCI_{w}(H) = \left(N^{-1} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \tilde{\omega}_{ijw}^{g}(H) \right) \times 100 \quad (8)$$

2.2.2. Attention-Weighted Extension

Although the benchmark separates transmission by frequency, its foundation is still a linear moving-average form. Every element of its decomposition is mechanically pinned down by two estimated objects, the coefficients plus the covariance of innovations, which leaves the model unable to give extra weight to channels that dominate in some particular regime. Graph-based attention provides a means of doing so. Each market becomes a node, every pair of nodes is linked, the links are held fixed, and trainable coefficients decide how strongly each node attends to the others. Attention here works directly with the matrix of connectedness and not with returns or some auxiliary prediction. In each estimation window, the first stage supplies a spectral decomposition that can be regarded as a complete directed graph on thirteen markets, to whose transmission channels the layer then assigns fresh weights. What is learned is thus the very structure of transmission, precisely what the analysis seeks to measure. The intention is to equip the linear model with three additional features.

(i) weights on individual channels that respond to shifting market conditions;

(ii) weights for each channel shaped by the full exposure profile of both endpoint markets, so that indirect exposure via third markets feeds into how much any direct link counts; and

(iii) an objective that looks forward, whereby weights are judged on how well they anticipate the decomposition one window ahead, not on matching what is already observed. Whether any of this improves measurement is an empirical matter for Section 3.4. Protection against the heteroscedasticity and heavy tails revealed by the preliminary tests comes from re-estimating the covariance of innovations within every first-stage estimation window, not from attention, since Equation (14) defines the loss as a plain, unweighted average of squared errors.

The first stage returns, for every window w, an N × N GFEVD matrix θ^{w} with (i, j) element equal to that part of i's H-step prediction error variance which shocks to j explain, each row summing to one. θ^{w} is then regarded as a network with directed edges, G^{w}, whose node set V contains the thirteen markets kept after screening and whose edge set E links every ordered pair. Since every pair is connected in every window, E needs no time index, and only attention weights change from window to window. In window w, row i of θ^{w},

a vector x^w_i in $\mathbb{R}^N$, describes node i by giving a full breakdown of where i's prediction error variance originates. This representation has two consequences that matter later. First, node features carry no scale, since the entries of every feature vector are non-negative and add up to one. Standardisation is therefore unnecessary, and a node cannot dominate attention scores simply by having a large variance of its own. Second, the features are themselves system-wide estimates, which means attention operates on an estimated structure of transmission instead of on raw noise underlying it.

Attention scores. For each ordered couple of markets (i, j), Equation (9) gives the raw, unnormalised score as a bilinear form, with $\Omega \in \mathbb{R}^{N \times N}$ a trainable matrix common to every edge in the graph.

$$e^w_{ij} = x^w_i \Omega \left(x^w_j \right)^\top = \Sigma^N_{k=1} \Sigma^N_{l=1} \theta^w_{ik} \Omega_{kl} \theta^w_{jl} \tag{9}$$

Equation (9) does not use the additive form of attention employed by Veličković et al. (2018). It adopts instead the bilinear, or 'general', score of Luong et al. (2015). The choice has practical consequences. Brody et al. (2022) demonstrate that an additive score imposes one common ranking of neighbours on all nodes, while with a bilinear score the ranking varies with the node that receives attention. Since the score combines two full decomposition rows bilinearly, Ω_{kl} weights cases in which i is exposed to k while j is exposed to l. Consequently, the weight given to the channel j → i is determined not by that pair in isolation but by how these two exposure patterns interact across the whole system, so that markets sharing an exposure to some third market are weighted differently from pairs that do not. Indirect propagation through the system's dynamics is already embedded in the GFEVD element for j → i, yet only via a fixed linear mapping. Bilinearity adds a second route, this one learned, by which common exposures k → i and k → j, for instance to the world equity index, can make any direct link stronger or weaker.

Equation (10) normalises scores with a row-wise softmax over $\mathcal{N}$(i), the neighbourhood of i, which with full connectivity contains every market. The matrix that results is row-stochastic, as is the GFEVD it stands in for, so its entries can still be interpreted as variance shares.

$$\alpha^w_{ij} = softmax_j \left(e^w_{ij} \right) = \frac{exp\left(e^w_{ij} \right)}{\Sigma_{k \in \mathcal{N}(i)} exp\left(e^w_{ik} \right)} \tag{10}$$

Following the proposal of Veličković et al. (2018), K heads are run side by side, each head h with a separate $\Omega^{(h)}$. Equation (11) averages the normalised outputs of the heads. Using several heads is supposed to make estimation more stable and to allow individual heads to specialise in different patterns. The grid favours two heads (Table 10), although Section 3.4 shows that they are identical in practice.

$$\alpha^w = K^{-1} \Sigma^K_{h=1} \alpha^{w(h)} \tag{11}$$

Entering the spillover measures. Equation (12) replaces the GFEVD with α, after which every directional and aggregate index of Section 2.2.1 is computed from $\tilde{\theta}^w$ instead of θ^w.

$$\tilde{\theta}^w = \alpha^w \tag{12}$$

The band-specific matrices that the Baruník–Křehlík approach produces are adjusted through the ratio of Equation (13), in which ⊙ multiplies and ⊘ divides element by element.

$$\tilde{\theta}_D^w = \theta_D^w \odot \left(\alpha^w \oslash \theta^w\right) \tag{13}$$

Under Equation (13), each band inherits the channel level implied by attention, while the split of that channel across frequencies is kept as the linear model sets it. The construction is approximate, as it presumes that attention scales a channel equally at every frequency instead of shifting weight from one band to another. It follows that the near-identical band shares that Table 14 reports for both specifications are a by-product of that equation, and they cannot be taken as separate evidence about how transmission is distributed across frequencies.

Training objective. Training selects the learnable matrices Ω so as to minimise Equation (14), the error in forecasting next window's connectedness matrix, with M rolling windows in total and $\|\cdot\|$ a Frobenius norm.

$$L(\Omega) = (M-1)^{-1} \sum_{w=1}^{M-1} \left\| \alpha^w - \theta^{w+1} \right\|_F^2 \tag{14}$$

The objective is forward-looking by design. It asks what arrangement of the channels at time w best foreshadows the one observed at w + 1, and it therefore rewards weights for what they predict about transmission and not for reproducing it within a window. Adam (Kingma and Ba, 2015) is used for optimisation. Mirroring how the benchmark was tuned, three hyperparameters, namely K, initialisation scale and learning rate, are chosen on a grid, using prediction error on a validation set made of the last 25% of windows. Iterations are not tuned. Training continues until convergence, after which the iteration giving the lowest validation error is simply recorded. Table 10 gives the grid alongside the values finally selected. Two aspects of this objective call for comment. Being an unweighted sum of squared errors, it amounts to a Gaussian criterion with constant variance. A covariance-weighted version, or a quasi-likelihood loss like the one Zhang et al. (2025) find preferable for forecasting volatility, would match the preliminary evidence better. Furthermore, the objective is an undemanding one. As Section 3.4 documents, a naive forecast that leaves the decomposition unchanged between consecutive windows outperforms the fitted layer, which means Equation (14) serves to identify the weights of attention and offers no proof of superior predictive power.

In terms of design, three features distinguish the extended model from its linear benchmark. Its learned weights, collected in $\alpha^w = [\alpha^w{}_{ij}]$, aim to reveal, for each window separately, which links between markets carry transmission, whereas the benchmark fixes every element mechanically through estimated parameters and covariances. Equation (9), being bilinear, links each direct channel's weight to how both endpoints are ex-

posed. Finally, because a single estimated Ω acts on a new decomposition in every window, attention weights follow market conditions without being forced into a parametric form. Section 3.4 assesses the extent to which the data support this design.

A graph-based layer is a natural choice among neural architectures because connectedness is inherently a relational problem. Feed-forward and convolutional networks regard assets as interchangeable inputs, while recurrent networks capture dependence over time yet ignore relations among assets, which they could only approximate by stacking dense inputs from many assets. Here each head has one shared matrix Ω, giving N^2 parameters per head regardless of how many windows there are or how long the sample runs, and the computational burden per window is driven by two products of $N \times N$ matrices. With thirteen markets, neither consideration is binding. A considerably larger panel would require limiting how many parameters there are, for example by writing $\Omega = UV^T$, where U and V belong to $\mathbb{R}^{Nxr}$ with r small, which keeps the interaction bilinear while letting the cost rise only linearly in N once r is fixed. Figure 4 shows how large the weights learned by the layer are, on average across all windows. Whether edges with persistently high weights can be interpreted as evidence about transmission, as is often done with attention weights, is among the issues the later ablation (Section 3.4) should settle.

3. Empirical Findings

3.1. Linear Benchmark

Table 9 reports how the benchmark is configured. Window length (120, 250 or 500 days) is chosen jointly with lag order (between one and three) to minimise the mean of squared one-step forecast errors (MSFE). The selected model has one lag and a window of 250 days, giving 1,884 windows in all, the first of which ends on 17 January 2019, with the last one closing on the final day available. Every window has a stable companion matrix. The spectral radius has median 0.290 and maximum 0.576, so a generalised decomposition exists in all windows without any stabilising correction. This stability reflects the longer window. When 120-day windows are used, this system occasionally yields estimates close to explosive and forecasts less accurately.

Averaged across all windows, the TCI equals 55.06% (Table 8). Its low point, 47.76%, is reached during March 2024, and its maximum, 66.28%, is dated 17 March 2020, early in the COVID-19 crash, as all markets fell in unison and asset-class boundaries momentarily vanished. Indeed, the twelve largest values recorded anywhere in the sample are concentrated in the stretch from 13 March to 1 April 2020. Annual averages are highest in 2020 (61.8%) and 2023 (60.1%), and lowest in 2024 (50.5%) and 2021 (51.9%), which shows the system alternating between almost synchronised phases and phases during which the groups diverge.

By frequency, long cycles of 20 days or more account for 0.0554 of connectedness, cycles of between 5 and 20 days for 0.1657, and the shortest cycles, of five days or less, for 0.3259. Relative to their total, these figures are 10.1% for long cycles, 30.3% for medium ones, and short cycles take the remaining 59.6%. Trans-

mission here is thus predominantly short-lived, since roughly 60% of the variance passed between markets is absorbed inside five trading days. Le et al. (2021) report a comparable profile in volatility spillovers among fintech, cryptocurrency and green-bond markets, and Bhanja et al. (2023) in a system that combines Bitcoin with equity and precious-metal markets. Since the decomposition runs to a 100-day horizon, low frequencies are captured far beyond the 20-day cycle that bounds the band. The small share at long horizons therefore mirrors the data and is no artefact of truncating the spectrum. Across bands the sum is 0.5470, against 0.5506 overall, and there is no reason for the two numbers to coincide. The overall figure is a time-domain quantity that places equal weight on all forecast horizons, while the band figures come from integrating the normalised spectrum with the same weight on each frequency. The two agree only for white noise, and here the gap amounts to 0.36 pp, a relative gap of about 0.65%.

Table 7 reports net positions, which the left panel of Figure 2 also displays. Five series are net senders. ACWI leads (NET = +0.2618), NASC (+0.1406) and NAST (+0.1198) follow, with BTC (+0.0224) and ETH (+0.0494), the leading cryptocurrencies, positive too. All other series receive on balance. This ranking is economically plausible. ACWI embodies the common component behind all other series, a role that the sector indices for clean energy and technology partly share. At the opposite end are the commodities. Crude oil ranks as the biggest net recipient (−0.1404), while the least connected series, natural gas, takes from other markets just 0.127 of the variance of its prediction errors, compared with 0.873 that originates in shocks specific to it. DXY (−0.1096) and Treasury futures (−0.0949) also receive on net, in keeping with their function of pricing factors, not of independent origins of shocks during the sample. Le et al. (2021) find a similar division for volatility, with world equity indices on the transmitting side and the dollar, oil and gold on the receiving side.

Segmentation is pronounced. On average, pairwise entries reach 0.2285 among the cryptocurrencies and 0.1981 between NASC and SPC, but only 0.0144 among the commodities and in a range from 0.008 to 0.037 whenever a cryptocurrency is paired with a commodity or a macro-financial series. Of the sector groups, clean energy and technology show the tightest connection (0.1428 in one direction, 0.1315 in reverse), a result that owes at least as much to shared index constituents as to any separate channel of transmission. The cryptocurrencies thus make up a tight cluster with limited exchange with other markets, as Corbet et al. (2018), Duan et al. (2023) and Narayan and Kumar (2024) also find, while commodities stay the clearest source of diversification in this sample. This differs from Ji et al. (2019), according to whom cryptocurrencies are becoming more integrated into a wide commodity network. Their entropy-based approach, sample and choice of commodities are not those adopted here, and the contrast suggests that the integration they report does not carry over to daily return spillovers once the world equity factor enters.

Series	TO	FROM	NET	FROM (low)	FROM (med)	FROM (high)
NAST	0.8448	0.7250	0.1198	0.0731	0.2186	0.4328
NASC	0.8721	0.7315	0.1406	0.0724	0.2176	0.4397
SPC	0.6487	0.6666	-0.0179	0.0682	0.2028	0.3916
BTC	0.6557	0.6333	0.0224	0.0626	0.1885	0.3811
ETH	0.6901	0.6407	0.0494	0.0638	0.1916	0.3841
BNB	0.5399	0.5899	-0.0500	0.0565	0.1727	0.3589
OIL	0.1654	0.3058	-0.1404	0.0357	0.1015	0.1637
GOLD	0.3273	0.4084	-0.0812	0.0408	0.1217	0.2402
GAS	0.0536	0.1267	-0.0732	0.0137	0.0375	0.0736
ACWI	1.0148	0.7530	0.2618	0.0756	0.2266	0.4493
DXY	0.3901	0.4997	-0.1096	0.0540	0.1591	0.2738
ZN	0.3056	0.4005	-0.0949	0.0372	0.1147	0.2427
VIX	0.6495	0.6764	-0.0268	0.0665	0.2012	0.4053
TCI		0.5506		0.0554	0.1657	0.3259

Table 7 Connectedness Measures, Linear Benchmark

(Summing each GFEVD row net of its diagonal element gives FROM, the same operation on columns gives TO, and NET is TO minus FROM. The three rightmost columns break FROM down by frequency band, from low to high. Since the band values are spectral whereas FROM is a time-domain figure, they add up to FROM only approximately, with a discrepancy of 0.013 for DXY and no more than 0.006 for any other series, as Section 3.1 explains. The bottom row gives the TCI, which averages FROM across all 13 series.)

TCI statistic	Value
Average of rolling TCI, full sample	55.06%
Minimum TCI	47.76%
Maximum TCI	66.28%
Low frequency (20 days or more)	5.54%
Medium frequency (5 to 20 days)	16.57%
High frequency (5 days or less)	32.59%

TCI statistic	Value
Total across bands	54.70%

Table 8 Benchmark TCI by Statistic and Band

(The top three rows describe how the index behaves across windows, and the bottom four its split by frequency. Band contributions total 0.5470, whereas averaging the index over all windows gives 0.5506. The former is a frequency-domain and the latter a time-domain quantity, so they need not be equal (Section 3.1). Every TCI value uses N = 13 as divisor.)

Window (days)	Lag (p)	One-step MSFE	Result
120	1	1.226679	
120	2	1.449174	
120	3	1.766066	
250	**1**	**1.124963**	**selected**
250	2	1.216067	
250	3	1.331437	
500	1	1.155620	
500	2	1.206163	
500	3	1.269287	

Table 9 Selecting Window Length and VAR Order

(Each cell reports the MSFE of one-step forecasts from a VAR refitted in every window, for a given window length and lag order. The smallest error determines which configuration is kept.)

3.2. Attention-Weighted Estimates

Next, the 13 series are analysed with the extended model. As Section 2.2.2 explains, attention operates on each window's benchmark decomposition, not on returns, and its output stands in for the original decomposition throughout all rolling calculations of connectedness. Adam (Kingma and Ba, 2015) is again the optimiser. Table 10 indicates, dimension by dimension, whether each grid setting shifts the solution itself or only how many iterations it takes to get there. Window length and lag order, both picked by forecast error, matter. Three further settings do not. Varying the step size (learning rate) multiplies the required iterations by six yet shifts validation error by under one per cent, the scale used for initialisation does not matter, while the head

count makes no difference at all, for reasons explained later (Section 3.4). Optimisation terminates at convergence. The configuration finally selected attains its minimum validation error at iteration 4,450.

Parameter	Grid range	Selected value	Impact on the converged solution
VAR lag length (p)	{1, 2, 3}	1	Relevant, chosen by one-step MSFE
Window length in days	{120, 250, 500}	250	Relevant, chosen by one-step MSFE
Number of heads (K)	{1, 2, 4}	2	No effect, all heads converge on one solution (Section 3.4)
Learning rate	{0.005, 0.010, 0.050}	0.005	Affects only how fast the optimum is reached, not the optimum itself
Initialisation scale	{0.01, 0.05}	0.01	No effect, validation MSFE moves by just 5.8e-06 over the whole grid
Training iterations	until convergence (see note)	4,450	Determined by the run rather than set on the grid
Decomposition horizon (H)	fixed	100	-
Number of frequency grid points	fixed	200	-

Table 10 Hyperparameter Grid and Selection

(Training has no fixed iteration budget. It stops at convergence, defined as a relative improvement of the objective smaller than 10^{-6} across fifty successive iterations, or as a rise in validation error at six successive checks. How many iterations are run thus emerges from the run itself. The last column records, for each grid dimension, whether it alters where training ends up or merely how it gets there, with support from the ablation (Section 3.4).)

With attention weighting, integration is high, the mean TCI being 58.78% (Table 11). The lowest value, 49.29%, occurs during March 2022, whereas the top value, 76.12%, is dated 17 March 2020, also the benchmark's peak date (Table 12). Out of 1,884 windows, only 24 show an attention-based index no higher than the benchmark. Section 3.4 discusses what this gap can reveal and what it cannot. Taken alone, the gap indicates only that own shares, on average, drop to 0.4122 from 0.4494.

The frequency profile is almost unaltered. Under attention the bands contribute 0.3493 at high frequencies, 0.1787 at medium ones and 0.0598 at low ones, which corresponds to 59.4%, 30.4% and 10.2% of their total, against 59.6%, 30.3% and 10.1% under the benchmark. Such similarity is mechanical and therefore uninformative about the data. Under Equation (13), each attention-implied channel level is spread over frequencies using the split the benchmark model already gives it. Band shares thus pass through intact, and only the level they share moves. Whatever separates attention from the benchmark therefore concerns channel strength and not how a channel is spread over horizons.

Series	TO	FROM	NET	FROM (low)	FROM (med)	FROM (high)
NAST	0.9197	0.7889	0.1308	0.0796	0.2388	0.4705
NASC	0.8915	0.7914	0.1001	0.0789	0.2372	0.4754
SPC	0.7082	0.7088	-0.0006	0.0734	0.2185	0.4169
BTC	0.6992	0.6874	0.0118	0.0673	0.2031	0.4170
ETH	0.7104	0.7037	0.0067	0.0696	0.2095	0.4246
BNB	0.5802	0.6204	-0.0402	0.0594	0.1814	0.3796
OIL	0.2118	0.3019	-0.0902	0.0367	0.1030	0.1623
GOLD	0.3357	0.4213	-0.0856	0.0436	0.1297	0.2480
GAS	0.0910	0.1162	-0.0252	0.0123	0.0336	0.0703
ACWI	0.9361	0.8127	0.1234	0.0828	0.2476	0.4824
DXY	0.4472	0.5365	-0.0893	0.0599	0.1762	0.3004
ZN	0.3543	0.4090	-0.0547	0.0397	0.1208	0.2485
VIX	0.7559	0.7430	0.0128	0.0742	0.2233	0.4455
TCI		0.5878		0.0598	0.1787	0.3493

Table 11 Connectedness Measures with Attention Weighting
(FROM, TO and NET are defined as for Table 7. Here, though, the band columns sum exactly to FROM itself, and the bottom line gives the TCI, that is, mean FROM over the 13 series.)

Across all frequencies combined, Table 11, together with panel (b) of Figure 2, shows the same split as the benchmark, apart from one change. Net transmitters remain NAST, ACWI and NASC, with NET values of +0.1308, +0.1234 and +0.1001, along with ETH (+0.0067) and BTC (+0.0118). The VIX, at +0.0128, is now among them as well, having received marginally on net under the benchmark. WTI crude, DXY and gold continue to be the biggest net receivers. In the ranking by net transmission, twelve series move by no more than two places, the VIX being the exception with a rise of three places, while for twelve the NET sign is unchanged. Attention therefore lifts the system's level while leaving the ranking largely as it was.

TCI statistic	Value
Average of rolling TCI, full sample	58.78%
Minimum TCI	49.29%
Maximum TCI	76.12%
Low frequency (20 days or more)	5.98%

TCI statistic	Value
Medium frequency (5 to 20 days)	17.87%
High frequency (5 days or less)	34.93%
Total across bands	58.78%

Table 12 Attention-Weighted TCI by Statistic and Band

(With attention weighting, 58.78% is at once the mean across windows and the cross-band total, since Equation (13) apportions band contributions according to α. Table 8's benchmark instead derives its bands spectrally, which leaves the two figures 0.36 pp apart.)

Bilateral links are strongest among the cryptocurrencies. BTC and ETH exchange 0.2810 one way and 0.2791 the other, ETH and BNB 0.2154, BTC and BNB 0.2050. Next in strength come NASC with SPC (0.2100) and NAST with ACWI (0.1899 one way, 0.1888 the other). Commodity markets display uniformly smaller gross scores, with natural gas sending 0.091 and receiving 0.116, roughly ten times weaker than the links between cryptocurrencies.

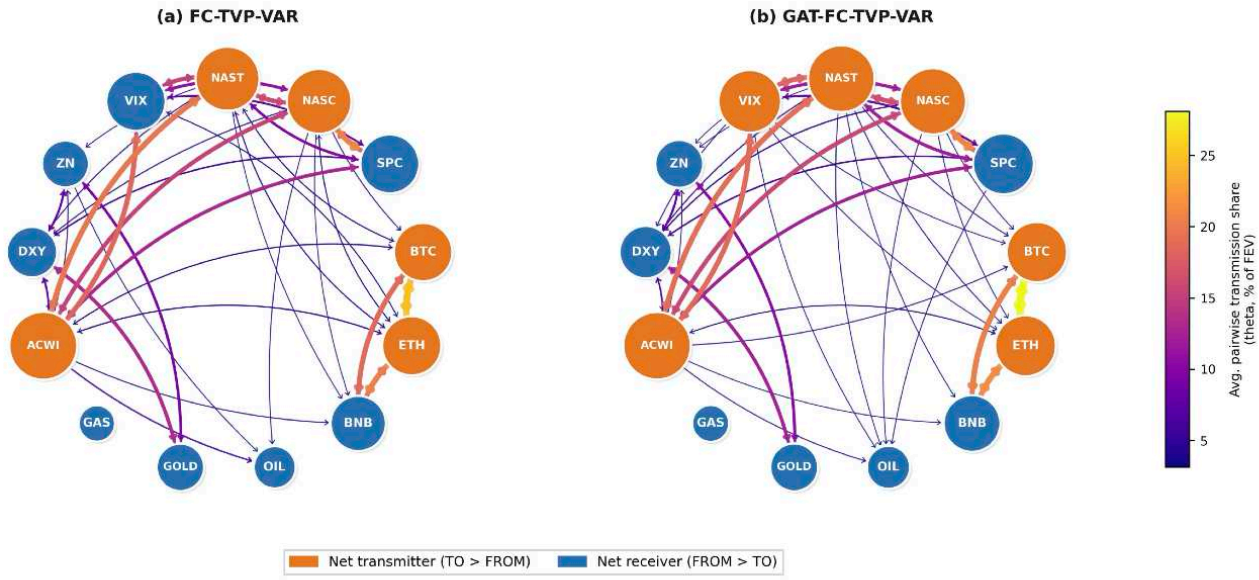


Figure 2 Average spillover networks, linear benchmark in panel (a) and attention-based extension in panel (b)

(Average networks over all rolling windows, showing only links in the top 38% by strength. Edges are coloured and sized by the average share transmitted, measured in percent of prediction error variance and drawn on one scale for both networks. The size of a node reflects gross involvement, TO + FROM, again on a shared scale. Orange marks net transmitters, with TO above FROM, and blue marks net receivers.)

That connectedness is concentrated at high frequencies holds for each series individually, so it characterises the entire sample and not one group. Table 11 shows no series with a FROM value above 0.10 for low frequencies, and every series apart from natural gas exceeds 0.15 at high frequencies.

3.3. Model Comparison

On average, attention weighting places the TCI 3.72 pp above the benchmark, which is a rise of 6.76%. In 1,860 out of 1,884 windows the difference has a positive sign, reaching its widest point of 10.17 pp during March 2020. Where it turns negative, in the remaining 24 windows, it is never lower than −0.32 pp. The chosen configuration has two heads, initial weights scaled by 0.01 and a step size of 0.005, but none of these settings alters the solution, a point recorded by Table 10 and demonstrated by the ablation (Section 3.4).

Statistic	FC-TVP-VAR	GAT-FC-TVP-VAR	Δ (GAT − FC)	% change
Average TCI, full sample	0.5506	0.5878	+0.0372	+6.76%
Minimum TCI	0.4776	0.4929	+0.0153	+3.21%
Maximum TCI	0.6628	0.7612	+0.0984	+14.85%
Low frequency	0.0554	0.0598	+0.0044	+7.95%
Medium frequency	0.1657	0.1787	+0.0130	+7.82%
High frequency	0.3259	0.3493	+0.0234	+7.19%

Table 13 Benchmark and Attention-Weighted TCI Compared
(Means, minima and maxima refer to all windows, and the means coincide with the all-band totals of Table 8 and Table 12. The % change column expresses Δ in per cent, relative to benchmark levels.)

Figure 3 traces both indices through time. Their paths are parallel, separated by a gap that stays positive and reasonably steady. Starting 2019 at roughly 53% (benchmark) and 56% (attention), they rise to their highest levels in the opening months of 2020, when the COVID-19 sell-off brought all markets into step, and retreat through 2021, when abundant post-COVID liquidity let the groups decouple. The sharp drop in March 2021 is partly mechanical, because it coincides with the observations of March 2020 dropping out of the estimation window, which spans 250 days. Connectedness is high again throughout 2023, averaging 65.7% under attention and 60.1% under the benchmark, a year marked by the banking turmoil of March 2023 and continued monetary tightening. Annual averages bottom out in 2024 (50.5% without attention, 53.3% with it), and both series climb again over the first half of 2026. The distance between them is greatest when connectedness is high, exactly as a multiplicative rescaling of off-diagonal weight would imply.

Table 13 sets the two specifications' summary statistics side by side. Going from the benchmark to attention, the mean TCI increases to 0.5878 from 0.5506, the minimum moves up to 0.4929 from 0.4776, while the maximum rises to 0.7612 from 0.6628. Table 14 shows each band's contribution growing in proportion, by 0.0044 for long cycles, 0.0130 for medium cycles and 0.0234 for short cycles, with band shares never more than 0.1 pp away from their benchmark values.

Band	FC-TVP-VAR	Share	GAT-FC-TVP-VAR	Share	Δ TCI
Low frequency (20 days or more)	0.0554	10.13%	0.0598	10.17%	+0.0044
Medium frequency (5 to 20 days)	0.1657	30.29%	0.1787	30.40%	+0.0130
High frequency (5 days or less)	0.3259	59.58%	0.3493	59.43%	+0.0234
Total	0.5470	100.00%	0.5878	100.00%	+0.0408

Table 14 Decomposition of Total Connectedness by Frequency Band
(Each share is computed against its own column's sum over bands, namely 0.5878 with attention and 0.5470 without it. The bottom-row Δ TCI is how much this sum changes, so it also equals the band-level changes added together.)

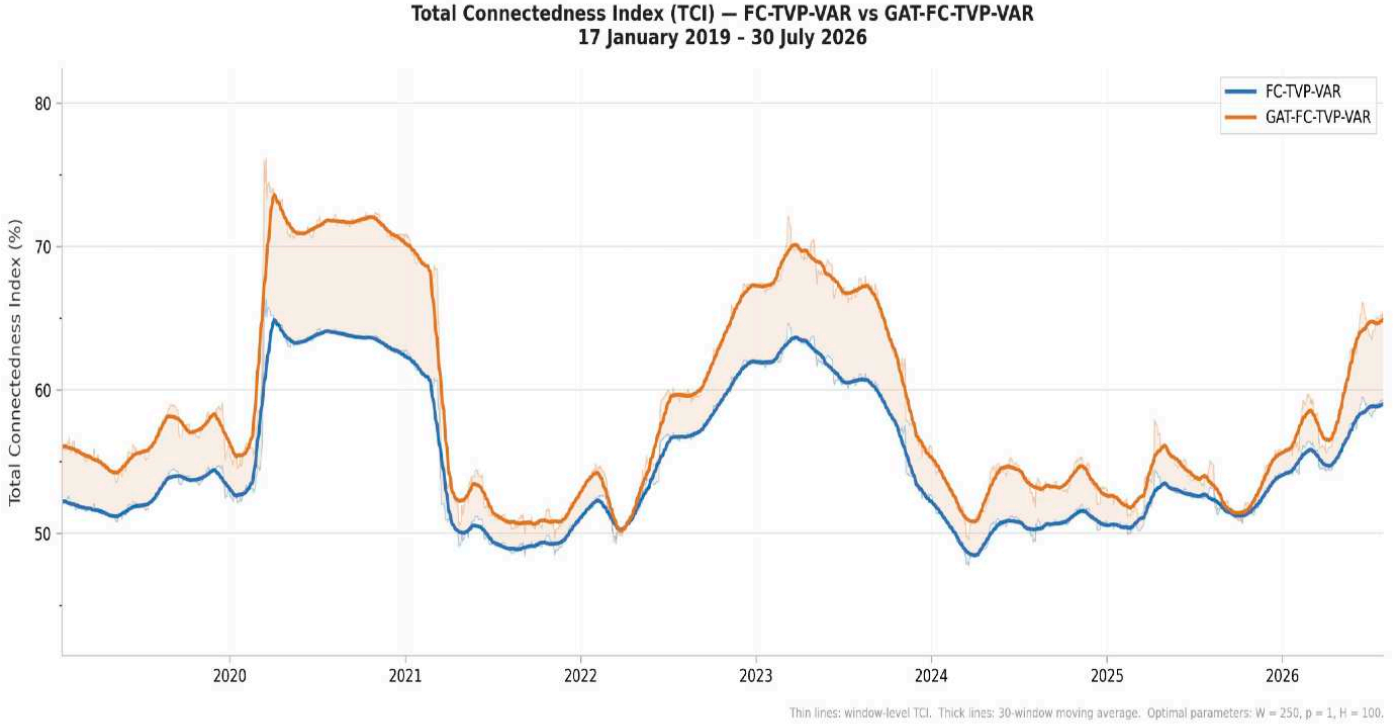


Figure 3 Rolling TCI under Both Specifications

Series by series, FROM grows most for markets whose benchmark connectedness was highest. The VIX releases 6.67 pp of its own share, and the next largest releases, in increasing order, are ACWI (5.97), NASC (5.99), ETH (6.30) and NAST (6.39). The two energy commodities behave differently, their own shares going up by 0.39 pp in crude oil and by 1.05 pp in natural gas. This correction is therefore distinctly uneven. It makes own shares more dispersed across series, so that series which were already well connected gain further links while isolated ones become more isolated still. Section 3.4 considers to what extent this reallocation distinguishes the trained attention from alternatives that use one parameter.

Three tests plus one measure of effect size evaluate the series-level differences, using paired total-band FROM values for the 13 series, and Table 15 collects the results. For paired data, t equals 4.899, with a p-value of 0.0004. The signed-rank test of Wilcoxon (1945) yields a statistic of W = 4, with a p-value of 0.0017, and Efron's (1979) bootstrap places the average difference within [0.0224, 0.0510], wholly above zero. The

effect size, measured by Cohen's d, equals 1.359, which counts as large under conventional thresholds (Cohen, 1988).

Test	Type	Statistic	p-value	95% interval for Δ	Result
t-test (paired)	Parametric, paired	t = 4.899 (12 df)	p = 0.0004	[0.0207, 0.0538]	Significant ***
Wilcoxon signed-rank	Paired, non-parametric	W = 4.000	p = 0.0017	-	Significant ***
Bootstrap CI	10,000 resamples	0.0372 (mean Δ)	-	[0.0224, 0.0510]	Excludes zero
Effect size (Cohen d)	Mean difference in SD units	d = 1.359	-	-	Large effect
Moving-blocks bootstrap (block length 10)	10,000 replications	3.71 pp (mean Δ)	p < 0.0001	[3.39, 4.04]	Excludes zero
Moving-blocks bootstrap (block length 20)	10,000 replications	3.70 pp (mean Δ)	p < 0.0001	[3.26, 4.16]	Excludes zero
Moving-blocks bootstrap (block length 40)	10,000 replications	3.70 pp (mean Δ)	p < 0.0001	[3.08, 4.34]	Excludes zero
Moving-blocks bootstrap (block length 60)	10,000 replications	3.68 pp (mean Δ)	p < 0.0001	[2.95, 4.45]	Excludes zero

Table 15 Inference on the Connectedness Gap

(Rows 1 to 4 use each series' total-band FROM value (n = 13), with the t-statistic computed as attention model minus benchmark. The moving-blocks bootstrap rows resample the per-window difference between both TCIs, so serial correlation in that gap is taken into account, and block lengths are measured in windows.)

Since the four statistics share one set of thirteen series, which by construction are interdependent across series, they cannot be counted as independent evidence. What they show is a systematic gap rather than one confined to a couple of series. Dependence over time needs a separate treatment. The per-window gap persists strongly, its lag-one autocorrelation equalling 0.997 over a run of 1,884 windows. Tests that assume independent windows would therefore be heavily oversized. Künsch's (1989) moving-blocks bootstrap, whose results fill the bottom half of the table, is designed for precisely this problem. It puts the mean gap at 3.72 pp, with a 95% interval that stays away from zero whatever block length is used, [3.39, 4.04] for the shortest blocks and [2.95, 4.45] for the longest. Serial dependence therefore cannot account for the margin by which attention lifts the index above the benchmark.

These tests do not, however, demonstrate that attention yields a better account of transmission. They confirm a systematic gap that sampling noise cannot explain, but they cannot tell whether its direction is informative. That question is left to Section 3.4.

3.4. Ablation Analysis and Learned Weights

Measured connectedness can rise without the layer having learned anything. The TCI reflects nothing but the average weight lying outside the diagonal, so that any reallocation of weight towards those cells raises it. A matrix carrying no information whatsoever, with all elements equal to 1/N, would produce (N − 1)/N, or 92.31%, well above both models. Table 16 reports an ablation intended to distinguish genuine learning from purely mechanical index gains. In every row, an alternative takes the place of the trained attention weights, after which connectedness is recomputed on identical decompositions.

Specification	TCI (%)	Change in TCI (pp)	Average own share	Validation MSFE (1-step)	NET correlation with benchmark	Stability of top-20 edges
Benchmark FC-TVP-VAR (attention removed)	55.06	+0.00	0.4494	0.000003	1.000	0.885
Trained attention, baseline specification	58.78	+3.72	0.4122	0.000558	0.937	0.885
Equal attention weights, no training	92.31	+37.25	0.0769	0.019194	-	1.000
GAT at random initial values, untrained	92.31	+37.25	0.0769	0.019192	-0.139	0.861
Learned weights with asset labels permuted	66.50	+11.44	0.3350	0.002518	0.876	0.866
Attention fitted to windows in shuffled time order	59.43	+4.37	0.4057	0.000750	-	-
Shrinkage with λ minimising the loss (λ = 0.000)	55.06	+0.00	0.4494	0.000003	1.000	0.885
Shrinkage, λ set to match the attention TCI (λ = 0.100)	58.78	+3.72	0.4122	0.000193	1.000	0.885
Scalar attention, c·I in place of Ω, with c = 10.03 matching the attention TCI	58.78	+3.72	0.4122	0.001722	0.936	0.882

Table 16 Alternative Attention Specifications

(Rows differ only in what stands in for trained attention, and all measures are recomputed on unchanged decompositions. Δ TCI is measured relative to row one, the benchmark. Validation MSFE measures the out-of-sample error made when forecasting the next decomposition. The column headed NET correlation with benchmark gives the correlation, across series, between each row's NET values and benchmark NET values, and stability of top-20 edges is the fraction of the twenty largest links in each window found also in the ranking computed over all windows. Finally, a multiple of the identity, c·I, takes the place of the trainable Ω, with c fixed to make the implied index equal to that of trained attention.)

Two observations follow immediately from the table. Equal weights and an untrained network turn out to be degenerate rather than neutral placebos. With a small initialisation scale, an untrained network produces scores close to zero, the softmax then returns almost equal weights, and both variants settle exactly at 92.31%, the value of the uninformative matrix. Their validation error is about 34 times the fitted layer's and over a thousand times the error obtained when no change is made to the decomposition. Applying the trained weights

to shuffled asset labels raises the TCI to 66.50%, clearly above the value with labels in place, while validation error becomes roughly four and a half times larger. Index and training objective thus disagree, so the index cannot by itself rank specifications.

The comparison with linear shrinkage proves decisive. Suppose each decomposition is pulled towards equal weights through a parameter λ, and λ is picked to minimise the very validation loss used to train attention. The optimum is λ = 0.000, which means that no shrinkage is best and that leaving the benchmark untouched gives the most accurate forecast for the coming decomposition, with a validation error of 0.000003 against 0.000558 for the fitted layer. If λ is instead tuned until the index implied by attention is reproduced, its value is 0.100, and the trained model's TCI is matched to four decimals. This is achieved with a validation error of 0.000193, about one-third of what the fitted model attains, with top-edge stability unchanged (0.885), and with NET values perfectly correlated (1.000) with benchmark ones, because shrinkage merely rescales them. One scalar thus equals the layer on the very index that attention was expected to raise and does better on the criterion used for training. Shrinkage of this type is familiar from covariance estimation (Ledoit and Wolf, 2004), where its merit lies in better conditioning. Here it supplies the benchmark that any learned reweighting must surpass.

Three additional checks shed light on what is actually being fitted. First, window order is irrelevant. After retraining with the windows in random order, the mean index is 59.43%, a little above 58.78% under the true sequence, so attention captures a regularity that holds across series on average and not a pattern that unfolds over time. Second, the seed used for initialisation has no bearing on the estimates, since five seeds produce indices within 0.001 pp of each other and validation errors identical to six decimals. Third, head count is irrelevant. On the grid, validation errors vary by 0.0000056 when the head count is held fixed, a spread due to the step size, yet by only 0.00000004 when head counts are compared, more than a hundred times less. The explanation lies in the architecture. Every head is fitted to an identical objective in isolation before averaging takes place, and nothing in the specification, such as head-specific subspaces or a learned rule for combining heads, would push them apart. All heads thus converge on one matrix, whose average is again that same matrix. In the current specification, additional heads are therefore redundant, and one head describes the estimate faithfully.

Figure 4 plots the weights that attention learns, which are markedly uneven. Single weights lie between 0.0008 and 0.975, while equal weighting would give 0.0769 to each. Mean row entropy equals 1.770 nats under attention, 2.565 nats when rows are uniform and 1.715 nats under the benchmark. Rows produced by attention are thus marginally more dispersed than GFEVD rows, though still heavily concentrated. The left panel displays a distinct block pattern. Among the cryptocurrencies, weights are dense and nearly symmetric, the largest belonging to BTC and ETH (0.281 one way, 0.279 the other). NASC and SPC make up a second block and the pair of technology and world equity indices a third, while gold, oil and gas send and receive almost nothing. Attention therefore recovers the same segmentation as the linear model. Net positions are largely preserved as well. NET vectors from attention and from the benchmark correlate at 0.937, and the VIX

alone changes sign, moving from receiving marginally on net (−0.0268) to transmitting marginally (+0.0128), whereas SPC goes from a small net intake to essentially zero.

The layer and shrinkage nonetheless remain distinct, and their divergence becomes apparent across series. By construction, shrinkage towards equal weights compresses the distribution of own shares, so the change it makes to a series' own share correlates at −1.000 with where that share started. Attention behaves in the reverse manner. Its adjustment and each series' starting own share correlate at +0.945, so variance is drawn mostly from series that depended least on their own shocks to begin with, and a small amount is returned to the pair of commodities that were most isolated. Across series, shrinkage and attention adjustments correlate at −0.945. Equal indices therefore conceal reallocations of contrary sign that coincide only in their average.

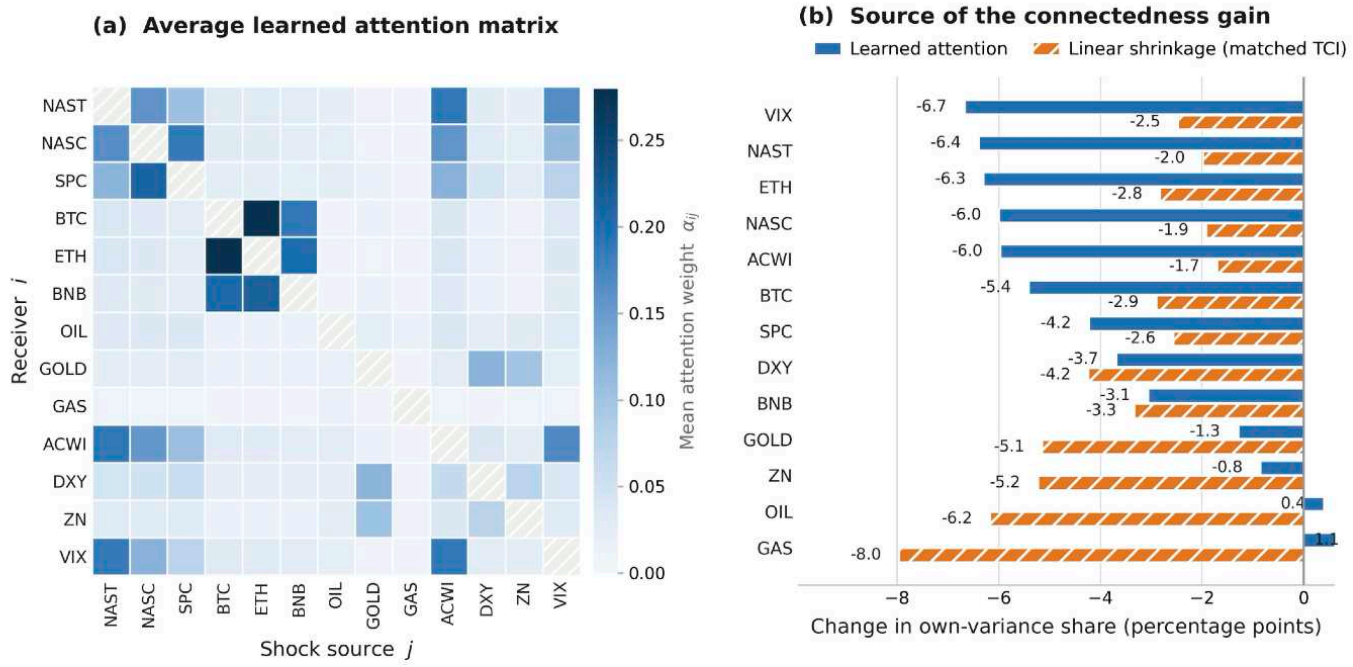


Figure 4 Estimated Attention Matrix and Sources of the TCI Gain
(The left panel plots window-averaged attention weights, leaving out and hatching the diagonal, where own shares sit, so that structure between series can be read. The right panel gives, series by series, how much each own share moves away from its benchmark value, under trained attention and under shrinkage with its single parameter set to deliver the same TCI. These profiles correlate at −0.945. Shrinkage removes own variance chiefly where there is most of it, attention where there is least.)

This direction, however, does not originate in estimation. With Ω replaced by c·I, a scalar times the identity, no cross-terms remain in the score, and the softmax operates on nothing but inner products between exposure profiles. Calibrating c until the index reaches its attention-based level gives c = 10.03, the value adopted for the last Table 16 specification. This restricted variant reproduces the TCI to two decimals, the weights it assigns between distinct markets correlate at 0.952 with those learned by attention, its own shares correlate at 0.995 with their learned counterparts, and even the reallocation points the same way (+0.950, versus +0.945 when Ω is learned). Whatever distinguishes attention and shrinkage thus comes from combining a bilinear score with a softmax, and the N^2 parameters that are estimated play no part in it. Their contribution is genuine,

though modest and measurable. Validation error drops to 0.000558 once Ω is learned, from 0.001722 with the matched scalar, roughly three times smaller, although both stay hundreds of times above the 0.000003 achieved by leaving the decomposition unchanged. Estimated weights therefore achieve what no scalar can, yet this achievement leaves no trace in the TCI, which one number alone can match in two ways, as a shrinkage weight or as a scale factor.

The ablation supports three conclusions, which temper the comparison of Section 3.3 without reversing it. First, a TCI that is 3.72 pp higher does not mean transmission is described more accurately. One shrinkage parameter delivers exactly the same rise while also lowering validation error. Moreover, the objective used in training is minimised by leaving the decomposition alone. Whenever gains in connectedness are credited to learned components, this comparison should be reported, since such a gain requires no learning. Second, attention and shrinkage shift own shares in contrary directions. This asymmetry is intrinsic to the mechanism, but it does not stem from estimation, because a scalar stand-in for Ω generates it too. Estimating N^2 weights reduces validation error to one-third of the matched scalar's, an improvement the TCI does not register. Third, the specification is more elaborate than its estimates warrant. Its heads duplicate one another, its initialisation has no influence, and its learning rate affects merely the speed of convergence. Nothing more elaborate than a bilinear reweighting with one head, trained until convergence, finds support in the evidence.

4. Concluding Remarks

With daily data for the period January 2018 through July 2026, this paper has traced return transmission among thirteen series covering cryptocurrencies, commodities, clean-energy and technology equities, together with macro-financial variables. It has also examined whether attention, learned on a frequency-band connectedness model, offers any gain over the purely linear decomposition.

Three empirical results merit emphasis. The first is strong and persistent connectedness. Under the benchmark, the TCI averages 55.06% with a floor of 47.76%, and around 60% of all transmitted variance dissipates within five trading days, so that transmission here is mostly a matter of short horizons. The second is a network divided into clearly defined blocks. Cryptocurrencies make up a dense cluster, with an average internal share of 0.229 as opposed to 0.014 among commodities, and the main net transmitters are the world equity index, NAST and NASC. The third is pronounced dependence on the regime, with the highest value, 66.28%, recorded during March 2020, elevated values throughout 2023 and the lowest in 2024.

The methodological result differs in nature and is arguably of greater value. Attention weighting raises the TCI by 3.72 pp, a gap that survives once the persistence of consecutive windows is accounted for. Shrinkage towards equal weights with one parameter, however, recovers all of that gain with less validation error. In addition, the training loss reaches its minimum when nothing is adjusted. Because the TCI rises mechanically whenever flatter decompositions are produced, an increase that is not assessed against this benchmark is uninformative. Attention differs from shrinkage only in how it reallocates own shares, dispersing them in cases where shrinkage can only compress them. Their adjustments correlate at −0.945. This asymmetry comes

from pairing a bilinear score with a softmax, not from estimated weights, given that a scalar stand-in for Ω produces it as well. Estimation merely reduces validation error threefold, with no effect on the TCI. The principal contribution of this work therefore lies in the diagnostic and not in the higher index.

Several practical implications follow. Given the segmentation of the network, investors can still use commodities, natural gas above all, to diversify against shocks that start in equity or cryptocurrency markets, whereas spreading a position over several cryptocurrencies achieves little, as shocks move rapidly within that dense block. Because transmission is concentrated within a five-day span and roughly nine-tenths of it is complete within a month, hedges have to be rebalanced at short horizons, since most of a spillover has played out before a monthly rebalancing date. For risk monitoring, the fact that ACWI and both sector indices lead as net transmitters singles out the equity factor as the most likely route by which a shock would spread to other markets. For researchers, the ablation offers a simple and inexpensive check. Before learned components are credited with better connectedness measurement, their effect on the TCI should be set against two simple references, namely a shrinkage tuned to hit that same TCI, and an otherwise identical functional form with only a scalar parameter.

These conclusions are subject to several limitations, which also suggest directions for future research. The layer learns to predict what the decomposition will be one window later, yet when each window holds 250 days, two successive decompositions share 249 of their 250 observations. Such a target changes very little, so that simply carrying the current decomposition forward is almost optimal. Defining the objective on something the decomposition should anticipate, such as next period's realised covariance, would put the mechanism to a harder test and is a natural extension. As currently specified, multiple heads add nothing, because each one minimises an identical objective separately before they are averaged, and assigning heads to distinct frequency bands or subsamples would give the parameter a meaning. Time variation here comes from rolling windows, whereas a TVP-VAR estimated with the Kalman filter and forgetting factors (Antonakakis et al., 2020) would neither lose the first window nor display the mechanical shifts caused by extreme observations dropping out of the estimation sample. Lastly, although the VIX is a stationary level series, it is log-differenced to keep units homogeneous, which differences a variable that is I(0) once too often.

ABBREVIATIONS			
Augmented Dickey–Fuller	ADF	Long Short-Term Memory	LSTM
Autoregressive conditional heteroscedasticity Lagrange multiplier test	ARCH-LM	Mean Squared Forecast Error	MSFE
Brock–Dechert–Scheinkman–LeBaron test	BDS	Net total directional connectedness	NET
Frequency Connectedness Time-Varying Parameter Vector Autoregression	FC-TVP-VAR	Ordinary Least Squares	OLS
Generalised Autoregressive Conditional Heteroscedasticity	GARCH	Percentage points	pp
Generalised Forecast Error Variance Decomposition	GFEVD	Time-Varying Parameter Vector Autoregression	TVP-VAR
Graph Attention Network	GAT	Total Connectedness Index	TCI
Graph Attention Network Frequency Connectedness Time-Varying Parameter Vector Autoregression	GAT-FC-TVP-VAR	Variance Inflation Factor	VIF
Jarque–Bera	JB	Vector Autoregression	VAR

REFERENCES

Ando, T., Greenwood-Nimmo, M., & Shin, Y. (2022). Quantile connectedness: Modeling tail behavior in the topology of financial networks. Management Science, 68(4), 2401–2431.

Antonakakis, N., Chatziantoniou, I., & Gabauer, D. (2020). Refined measures of dynamic connectedness based on time-varying parameter vector autoregressions. Journal of Risk and Financial Management, 13(4), 84.

Araya, H. T., Aduda, J., & Berhane, T. (2024). A hybrid GARCH and deep learning method for volatility prediction. Journal of Applied Mathematics, 2024, 6305525.

Baruník, J., & Křehlík, T. (2018). Measuring the frequency dynamics of financial connectedness and systemic risk. Journal of Financial Econometrics, 16(2), 271–296.

Bhanja, N., Shah, A. A., & Dar, A. B. (2023). Aggregate, asymmetric and frequency-based spillover among equity, precious metals, and cryptocurrency. Resources Policy, 80, 103145.

Billio, M., Getmansky, M., Lo, A. W., & Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. Journal of Financial Economics, 104(3), 535–559.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307–327.

Brock, W. A., Dechert, W. D., Scheinkman, J. A., & LeBaron, B. (1996). A test for independence based on the correlation dimension. Econometric Reviews, 15(3), 197–235.

Brody, S., Alon, U., & Yahav, E. (2022). How attentive are graph attention networks? In International Conference on Learning Representations (ICLR).

Chatziantoniou, I., Gabauer, D., & Gupta, R. (2023). Integration and risk transmission in the market for crude oil: New evidence from a time-varying parameter frequency connectedness approach. Resources Policy, 84, 103729.

Cheng, R., & Li, Q. (2021). Modeling the momentum spillover effect for stock prediction via attribute-driven graph attention networks. Proceedings of the AAAI Conference on Artificial Intelligence, 35(1), 55–62.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Corbet, S., Meegan, A., Larkin, C., Lucey, B., & Yarovaya, L. (2018). Exploring the dynamic relationships between cryptocurrencies and other financial assets. Economics Letters, 165, 28–34.

Demirer, M., Diebold, F. X., Liu, L., & Yilmaz, K. (2018). Estimating global bank network connectedness. Journal of Applied Econometrics, 33(1), 1–15.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74(366), 427–431.

Diebold, F. X., & Yilmaz, K. (2009). Measuring financial asset return and volatility spillovers, with application to global equity markets. The Economic Journal, 119(534), 158–171.

Diebold, F. X., & Yilmaz, K. (2012). Better to give than to receive: Predictive directional measurement of volatility spillovers. International Journal of Forecasting, 28(1), 57–66.

Diebold, F. X., & Yilmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. Journal of Econometrics, 182(1), 119–134.

Duan, K., Zhao, Y., Urquhart, A., & Huang, Y. (2023). Do clean and dirty cryptocurrencies connect with financial assets differently? The role of economic policy uncertainty. Energy Economics, 127(Part A), 107079.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7(1), 1–26.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50(4), 987–1007.

Esparcia, C., Escribano, A., & Jareño, F. (2024). Assessing the crypto market stability after the FTX collapse: A study of high frequency volatility and connectedness. International Review of Financial Analysis, 94, 103287.

Gaies, B., Chaâbane, N., & Bouzouita, N. (2024). Navigating the storm: Time-frequency quantile dependence and non-linear causality between crypto-currency market volatility and financial instability. The Quarterly Review of Economics and Finance, 93, 43–70.

Hassan, M. K., Hasan, M. B., Halim, Z. A., Maroney, N., & Rashid, M. M. (2022). Exploring the dynamic spillover of cryptocurrency environmental attention across the commodities, green bonds, and environment-related stocks. The North American Journal of Economics and Finance, 61, 101700.

Helmi, M. H., Çatık, A. N., & Akdeniz, C. (2023). The impact of central bank digital currency news on the stock and cryptocurrency markets: Evidence from the TVP-VAR model. Research in International Business and Finance, 65, 101968.

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 3543–3556). Association for Computational Linguistics.

Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. International Statistical Review, 55(2), 163–172.

Ji, Q., Bouri, E., Roubaud, D., & Kristoufek, L. (2019). Information interdependence among energy, cryptocurrency and major commodity markets. Energy Economics, 81, 1042–1055.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR).

Koop, G., & Korobilis, D. (2014). A new index of financial conditions. European Economic Review, 71, 101–116.

Koop, G., Pesaran, M. H., & Potter, S. M. (1996). Impulse response analysis in nonlinear multivariate models. Journal of Econometrics, 74(1), 119–147.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. The Annals of Statistics, 17(3), 1217–1241.

Le, T.-L., Abakah, E. J. A., & Tiwari, A. K. (2021). Time and frequency domain connectedness and spill-over among fintech, green bonds and cryptocurrencies in the age of the fourth industrial revolution. Technological Forecasting and Social Change, 162, 120382.

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis, 88(2), 365–411.

Lee, T. K., Choi, I., & Kim, W. C. (2025). Symmetry-aware graph neural approaches for data-efficient return prediction in international financial market indices. Symmetry, 17(9), 1372.

Liu, C., Wang, C., Tran, M.-N., & Kohn, R. (2025). A long short-term memory enhanced realized conditional heteroskedasticity model. Economic Modelling, 142, 106922.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1412–1421). Association for Computational Linguistics.

MacKinnon, J. G. (1996). Numerical distribution functions for unit root and cointegration tests. Journal of Applied Econometrics, 11(6), 601–618.

Mai, T., Foley, A. M., McAleer, M., & Chang, C. (2022). Impact of COVID-19 on returns-volatility spillovers in national and regional carbon markets in China. Renewable and Sustainable Energy Reviews, 169, 112861.

Morshed, A. (2025). Graph neural networks and explainable spillovers: Global monetary and oil shocks in GCC financial markets. Economies, 13(11), 308.

Naeem, M. A., & Karim, S. (2021). Tail dependence between bitcoin and green financial assets. Economics Letters, 208, 110068.

Narayan, S., & Kumar, D. (2024). Unveiling interconnectedness and risk spillover among cryptocurrencies and other asset classes. Global Finance Journal, 62, 101018.

Nasir, R. M., He, F., & Yousaf, I. (2025). Relationship of green cryptocurrencies, energy tokens, centralized and decentralized exchange tokens with crypto policy uncertainty. Research in International Business and Finance, 75, 102743.

Pesaran, H. H., & Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. Economics Letters, 58(1), 17–29.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. The Review of Economic Studies, 72(3), 821–852.

Sahiner, M. (2024). Volatility spillovers and contagion during major crises: An early warning approach based on a deep learning model. Computational Economics, 63(6), 2435–2499.

Symitsi, E., & Chalvatzis, K. J. (2018). Return, volatility and shock spillovers of Bitcoin with energy and technology companies. Economics Letters, 170, 127–130.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In International Conference on Learning Representations (ICLR).

Wang, X., Liu, J., & Xie, Q. (2024). Quantile frequency connectedness between energy tokens, crypto market, and renewable energy stock markets. Heliyon, 10(3), e25068.

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 11–20). Association for Computational Linguistics.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1(6), 80–83.

Wu, G., & Hu, G. (2024). Asymmetric spillovers and resilience in physical and financial assets amid climate policy uncertainties: Evidence from China. Technological Forecasting and Social Change, 208, 123701.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4–24.

Younis, I., Du, A. M., Gupta, H., & Shah, W. U. (2024). Dynamic spillover effects and interconnectedness of DeFi assets, commodities, and Islamic stock markets during crises. International Review of Financial Analysis, 96(Part A), 103563.

Yousaf, I., Cui, J., & Ali, S. (2024). Dynamic spillover between green cryptocurrencies and stocks: A portfolio implication. International Review of Economics & Finance, 96(Part B), 103661.

Zhang, C., Pu, X., Cucuringu, M., & Dong, X. (2025). Forecasting realized volatility with spillover effects: Perspectives from graph neural networks. International Journal of Forecasting, 41(1), 377–397.