DATA 2: Financial Networks

Moo K. Chung

University of Wisconsin-Madison, USA

mkchung@wisc.edu

Abstract. This document describes a curated multivariate financial time series dataset for geometric, causal, and topological data analysis. The dataset consists of daily end-of-day OHLCV (open, high, low, close, volume) records for 18 representative U.S. equities and exchange-traded funds spanning multiple economic sectors over the period January 1, 2000 to December 31, 2025. The data are publicly available from the Stooq historical market data repository and were programmatically retrieved and organized into a structured MATLAB format to support reproducible analysis. Beyond serving as a standard time-series benchmark, the dataset is explicitly designed to illustrate *how multivariate time series data can be interpreted geometrically as trajectories, causally as feedback-driven systems, and topologically as filtrations that encode multiscale structure through extrema, persistence, and recurrence.* The dataset therefore provides a unified empirical foundation for exploring geometric modeling, cycle-aware causal inference, and topological data analysis of dynamic systems. This is downloaded from https://github.com/laplcebeltrami/BMI768/tree/main/laplacetransform.

1 Financial Time Series Data

We analyze a curated panel of daily U.S. market assets covering the period from January 1, 2000 to December 31, 2025, publicly available from the Stooq historical market data repository (https://stooq.com), which provides free end-of-day OHLCV (open, high, low, close, volume) records for U.S.-listed equities and exchange-traded funds. All series were programmatically downloaded directly from using standardized Stooq ticker queries inside MATLAB.

The file `stocks.mat` contains OHLCV data for a set of 18 representative U.S. market assets spanning multiple economic sectors. Specifically, the dataset includes the following ticker–company pairs: `AAPL` (Apple), `MSFT` (Microsoft), `AMZN` (Amazon), `NVDA` (NVIDIA), `GOOGL` (Alphabet), `TSLA` (Tesla), `INTC` (Intel), `CVX` (Chevron), `XOM` (Exxon Mobil), `JPM` (JPMorgan Chase), `C` (Citigroup), `PLTR` (Palantir), `PG` (Procter & Gamble), `KO` (Coca-Cola), `JNJ` (Johnson & Johnson), `COST` (Costco), `UPS` (UPS), and `VOO` (Vanguard S&P 500 ETF). This selection intentionally spans technology, consumer, energy, financial, health care, transportation, and broad-market exposure, enabling the study of both sector-specific and cross-sector dynamic interactions within a unified financial network framework.

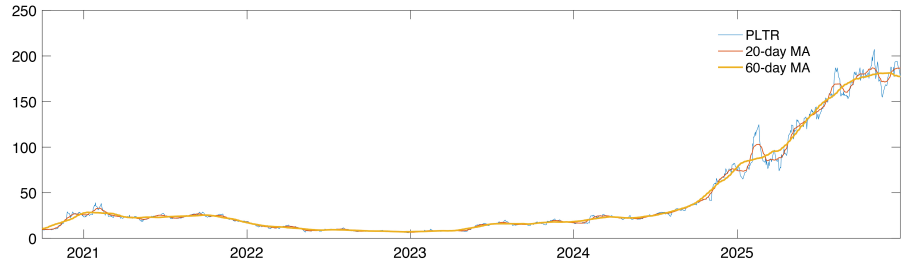

Fig. 1: Daily closing price of PLTR together with the 20-day and 60-day moving averages (MA), illustrating short-term and intermediate-term price smoothing.

The data are stored in `stocks.mat` as a structured array of MATLAB tables, where each element corresponds to a single traded asset and contains its complete daily trading history over the available period. Each table includes the following variables: `Date`, a `datetime` variable indicating the trading day; `Symbol`, a `string` specifying the stock ticker used in the U.S. market; `Company`, a `string` giving the company name; `Open`, the opening price of the trading day; `High`, the highest price reached during the trading day; `Low`, the lowest price reached during the trading day; `Close`, the closing price at the end of the trading day; and `Volume`, a `double` representing the number of shares traded on that day.

After loading `stocks.mat`, we concatenate all individual asset tables into a single long-format table using

```
T = vertcat(stocks{:});
```

This operation vertically stacks the daily records from all assets into one table `T`, where each row corresponds to a single asset on a single trading day and the columns retain the common variables `Date`, `Symbol`, `Company`, `Open`, `High`, `Low`, `Close`, and `Volume`. The resulting long-format representation is particularly convenient, as it allows straightforward filtering by asset or date, supports grouping and aggregation operations, and serves as a flexible intermediate format before reshaping the data into a fully aligned panel for network and time-lagged analyses.

The financial data are topological in nature in the sense that the daily `High` and `Low` prices explicitly encode local extrema of an underlying price trajectory, which is the topological information used by *Morse filtration* in topological data analysis (TDA) (Chung et al. 2009). If we regard the price (or log-price) as a real-valued function of time, $f(t)$, then local maxima and minima are the critical points of f in the classical one-dimensional Morse setting, and their ordering in time induces a natural filtration: as we sweep a threshold α through the range of f, the sublevel sets $\{t : f(t) \leq \alpha\}$ change topology only when α passes a critical value. In discrete time, the pair $(\texttt{Low}(t), \texttt{High}(t))$ provides an interval-valued observation of the latent function on day t, capturing the intraday excursion

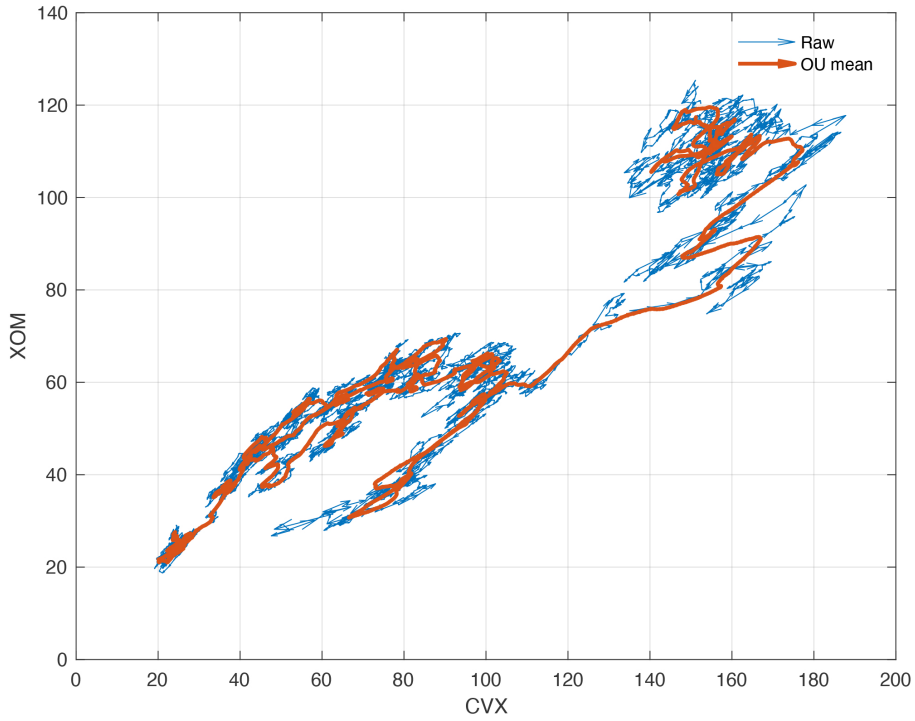

Fig. 2: Time–varying vector field of stock prices and its Ornstein–Uhlenbeck (OU) mean. Daily closing prices of Chevron (CVX) and Exxon Mobil (XOM) are embedded as trajectories in the (CVX, XOM) price plane. Arrows represent one–day price increments $(x_{t+1} - x_t,\ y_{t+1} - y_t)$, yielding a time–dependent velocity field. Thin arrows show the raw day–to–day price changes, which are dominated by high–frequency fluctuations. Thick arrows show the Ornstein–Uhlenbeck mean (with $\lambda = 20$ trading days), revealing persistent co–movement and the dominant drift structure shared by the two stocks.

and therefore offering direct evidence of where local extrema occur even when we do not observe the full intraday path. This makes the OHLCV record more than a standard time series: it can be interpreted as a sequence of daily *level-set intervals* whose endpoints represent local extreme values, and these extrema are the events that drive topological changes in filtrations.

Fig. 1 displays the daily price evolution of Palantir (`PLTR`). As the company began public trading only in late 2020, no data are available prior to this date, resulting in a comparatively short observation window relative to long-established firms. While single-stock analysis is useful for illustrating basic time-series behavior and smoothing effects, it is inherently limited because price movements are strongly influenced by market-wide and sector-level forces that cannot be dis-

entangled in isolation. Consequently, analyzing individual stocks alone provides an incomplete picture of underlying financial dynamics.

On the other hand, joint analysis of multiple assets enables the identification of co-movement, lead–lag structure, and collective behavior that are central to network-based and topological representations of financial markets. For instance, Fig. 2 visualizes a canonical example of strong within-sector co-movement by embedding two energy-sector assets, `CVX` (Chevron) and `XOM` (Exxon Mobil), into the two-dimensional price plane. This choice is deliberate: assets within the same sector are often driven by shared latent forces, so their joint evolution tends to exhibit coherent drift and structured dependence that can be interpreted meaningfully at the network level. In the figure, the raw day-to-day increments define a time-varying vector field along the observed trajectory, but this field is dominated by high-frequency fluctuations that obscure persistent structure. Applying the Ornstein–Uhlenbeck mean smoothing reveals a more stable drift pattern, highlighting the shared long-horizon motion of the two series. By contrasting such strongly coupled pairs with weaker cross-sector pairs, students can directly observe how the strength of coupling modulates geometric coherence, and why this coherence is essential when interpreting multivariate dynamics as networks with time-varying interactions.

Project 2: Causal Models without DAG

Motivating Example. Financial markets offer a stringent test bed for causal modeling (Pearl 2009, Schölkopf et al. 2021) because their core dynamics are inherently feedback-driven rather than feedforward. Price changes alter risk measures, updated risk measures trigger mechanical trading and margin responses, and these responses feed back into prices. A canonical illustration is the September 2008 financial crisis surrounding the collapse of Lehman Brothers. Crucially, Lehman's failure was not a single, isolated cause. As asset prices declined, many institutions were forced to liquidate positions to satisfy margin requirements and risk constraints. These forced sales further depressed prices, which in turn triggered additional liquidations by other institutions. Even a modest initial shock could therefore initiate a self-reinforcing cascade, often referred to as a *death spiral* in financial markets, in which price declines and forced selling continuously amplify one another.

Limitation of Existing Causal Models. Conventional causal models struggle to represent such dynamics. Regression-based approaches such as Granger causality (Friston et al. 2014) tend to decompose the episode into a sequence of external shocks, implicitly assuming that causes act unidirectionally over time. Time-varying or regime-switching models (Hamilton 1989) summarize the crisis through latent states or volatility shifts, but do not explicitly expose the feedback loop itself as a causal object. Such feedback mechanisms violate the assumptions of acyclic causal models based on directed acyclic graphs (DAGs) (Zheng et al. 2018), often leading to unstable estimates or ambiguous interpretations in standard time-series analysis.

There have been relatively few attempts to develop causal models that relax the DAG assumption and are explicitly cycle-aware. Early works permit algebraic or temporal feedback among variables in principle, but they typically rely on strong assumptions such as linearity, Gaussian noise, or equilibrium behavior; as a result, these models become difficult to identify, scale, or interpret in high-dimensional settings and often perform poorly in practice (Bollen 1989, Lohmöller 2013). Vector autoregressive (VAR) models and state-space formulations implicitly supposed to accommodate feedback through temporal recursion, yet their causal interpretation remains fragile because directionality is tied to lag choice and model specification rather than structural mechanisms (Geweke 1982, Hamilton 1989).

More recent work in causal discovery has begun to consider cyclic or equilibrium causal graphs, including linear non-Gaussian models with feedback and continuous-time dynamical systems. However, these approaches typically emphasize identifiability under *restrictive* assumptions, or interpret cycles as equilibrium artifacts rather than as persistent causal structures that actively govern system dynamics (Lacerda et al. 2008, Mooij et al. 2013). For example, Cyclic Causal Discovery (CCD) (Forré & Mooij 2018) explicitly allows directed cycles, but in practice its reliance on conditional independence makes it highly sensitive to strong multicollinearity among nodes participating in feedback loops.

In such settings, partial covariance inversions become ill-conditioned, leading to numerical instability and unreliable causal orientation in most applications. As a result, despite their conceptual relevance for feedback-dominated systems, cycle-aware causal models remain comparatively underdeveloped. This gap motivates alternative formulations that treat cyclic causality as a first-class structural object—rather than as a pathological violation of DAG-based assumptions—to be identified and characterized directly.

Goals. The objective of this project is to investigate existing causal models without DAG assumption and to identify their fundamental limitations through carefully designed simulated toy examples. The student will implement representative cycle-aware causal discovery methods such as in (Mooij et al. 2013, Forré & Mooij 2018) and examine their behavior under controlled settings that exhibit strong feedback, multicollinearity, and near-equilibrium dynamics. By systematically varying noise level, coupling strength, and dimensionality, the student will analyze when and why these models become ill-conditioned, unstable, or non-identifiable. Then use such models to identify cyclic structures in distributed multivariate time series data, and critically assess the extent to which the detected cycles are stable, interpretable, and robust to noise, collinearity, and model misspecification. The project then aims to motivate and explore alternative formulations that treat cyclic causality as a primary structural object rather than as a violation of DAG-based assumptions, highlighting the need for causal representations that remain well-defined in feedback-dominated systems.

Learning Outcomes. The student will be able to (i) articulate the conceptual and mathematical limitations of DAG-based causal models when applied to feedback-dominated systems, (ii) implement and critically evaluate representative cycle-aware causal discovery methods on multivariate time series data, and (iii) diagnose common failure modes of such methods, including ill-conditioning, numerical instability, and non-identifiability under strong feedback and multicollinearity. The student will also gain experience in interpreting cyclic causal estimates on real-world data with appropriate caution, and in communicating why persistent feedback structures require alternative causal representations beyond conventional regression- and independence-based frameworks.

Project 3: Geometric Modeling of Dynamic Data

The project will develop a joint geometric representation of multivariate time-varying data and use it to quantify, visualize, and compare system dynamics across time. The core goal is to move beyond analyzing individual time series in isolation, which often provides only limited insight into the underlying structure of complex data, and instead model the system as a time-evolving trajectory in a high-dimensional state space, where each time point corresponds to a single point whose coordinates are the multiple variables of interest. In this formulation, temporal evolution is encoded geometrically: the sequence of system states traces out a curve in the embedding space, and successive changes are represented by a discrete velocity field along this curve. Figure 2 illustrates the simplest nontrivial instance of this idea in two dimensions, where a Laplace-transform–inspired Ornstein–Uhlenbeck (OU) mean (Uhlenbeck & Ornstein 1930) is used to smooth a bivariate trajectory and reveal its dominant drift structure.

We will explore various ways to embed multivariate time series data into a low-dimensional manifold and to analyze the resulting trajectory using geometric methods. By treating the evolving system as a curve in an embedding space, the project emphasizes the global structure of temporal dynamics rather than isolated coordinate-wise behavior. To reveal meaningful structure across multiple temporal scales, it may be necessary to control or regularize high-frequency variability while preserving salient geometric features of the trajectory. Accordingly, the project will implement and compare filtering strategies for the geometric object, ranging from simple coordinate-wise approaches to trajectory-level operators motivated by linear dynamical systems or diffusion-based models. The proposed framework will emphasize geometric descriptors with domain-independent interpretation, such as angles between successive velocity vectors, measures of directional persistence, and curvature-like quantities that capture abrupt changes in the direction of system evolution. The student will examine how these geometric features evolve over time and relate their behavior to externally identified events, regime changes, or structural shifts in the underlying system.

It is expected that students will develop a statistical inference procedure to assess the significance, stability, and uncertainty of the geometric features, thereby connecting geometric modeling with principled statistical reasoning.

Learning Outcomes. The student will be able to formulate multivariate time-varying data as a geometric object evolving in an embedding space, implement and compare geometric modeling strategies at both coordinate and trajectory levels, and extract interpretable geometric descriptors of system dynamics. The student will also be able to demonstrate, through reproducible analysis and visualization, why joint geometric representations reveal collective structure and regime-level behavior that are not accessible through univariate time-series analysis.

Project 4: Topological Modeling of Dynamic Data

Goal. The goal of this project is to model multivariate time-varying data from a topological perspective and to extract qualitative, scale-invariant features of system dynamics that are robust to noise and temporal misalignment. In contrast to Project 3, which emphasizes geometric structure such as trajectories, velocities, and curvature in an embedding space, this project focuses on the topology of the evolving system, namely the presence of recurrent patterns, cycles, and persistent structures that cannot be captured by local geometric descriptors alone. The central idea is to treat dynamic data as inducing a filtration (Chung et al. 2009), a nested sequence of spaces indexed by an ordering parameter such as time, scale, or function value, and to study how topological features emerge, persist, and disappear as the system evolves with increasing values of this parameter. In this framework, the data are not analyzed at a single resolution or snapshot, but through a progressive view in which structures are created, merged, or eliminated, and only features that persist across a wide range of the filtration are interpreted as meaningful signals rather than noise. This perspective captures the essence of *topological data analysis* (TDA).

In this project, the student will explore how topological structure can be extracted from dynamic data through appropriate choices of filtrations induced by the data themselves. Beginning with multivariate time-varying observations, the student will examine how different ways of ordering, thresholding, or aggregating the data give rise to distinct topological viewpoints of system evolution. The emphasis is on interpretation rather than algorithmic detail: the student will investigate how topological summaries reflect global organization, recurrence, and structural change in the underlying dynamics, and how these summaries complement geometric and statistical analyses. A successful project will demonstrate that topological modeling reveals invariant, global aspects of dynamic systems that are not accessible through coordinate-wise or purely geometric approaches, and will clearly articulate the conceptual advantages of a topological perspective for understanding complex time-varying data.

It is expected that students will develop a statistical inference procedure to assess the significance, stability, and uncertainty of the topological features, thereby connecting topological modeling with principled statistical reasoning.

Learning Outcomes. The student will be able to articulate the principles of topological modeling for dynamic data, explain how filtrations encode multiscale structure in time-varying systems, and interpret topological summaries as descriptors of global organization and persistence. The student will also be able to critically assess how topological viewpoints complement geometric and statistical analyses, and to clearly communicate why topological features provide robust, scale-invariant insight into dynamic systems that is *not* accessible through local or coordinate-based geometric approaches.

Bibliography

Bollen, K. (1989), *Structural equations with latent variables*, John Wiley & Sons.

Chung, M., Bubenik, P. & Kim, P. (2009), 'Persistence diagrams of cortical surface data', *Proceedings of the 21st International Conference on Information Processing in Medical Imaging (IPMI), Lecture Notes in Computer Science (LNCS)* **5636**, 386–397.

Forré, P. & Mooij, J. (2018), Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders, *in* 'Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)', pp. 952–961.

Friston, K., Bastos, A., Oswal, A., Van Wijk, B., Richter, C. & Litvak, V. (2014), 'Granger causality revisited', *NeuroImage* **101**, 796–808.

Geweke, J. (1982), 'Measurement of linear dependence and feedback between multiple time series', *Journal of the American Statistical Association* **77**, 304–313.

Hamilton, J. D. (1989), 'A new approach to the economic analysis of nonstationary time series and the business cycle', *Econometrica: Journal of the econometric society* pp. 357–384.

Lacerda, G., S., P., Ramsey, J. & Hoyer, P. (2008), Discovering cyclic causal models by independent components analysis, *in* 'Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI)', pp. 366–374.

Lohmöller, J.-B. (2013), *Latent variable path modeling with partial least squares*, Springer Science & Business Media.

Mooij, J., Janzing, D., Heskes, T. & Schölkopf, B. (2013), From ordinary differential equations to structural causal models: The deterministic case, *in* 'Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)', pp. 440–448.

Pearl, J. (2009), *Causality: Models, Reasoning, and Inference*, 2nd edn, Cambridge University Press.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N., Kalchbrenner, N., Goyal, A. & Bengio, Y. (2021), 'Toward causal representation learning', *Proceedings of the IEEE* **109**, 612–634.

Uhlenbeck, G. & Ornstein, L. (1930), 'On the theory of the brownian motion', *Physical review* **36**, 823.

Zheng, X., Aragam, B., Ravikumar, P. & Xing, E. (2018), 'Dags with no tears: Continuous optimization for structure learning', *Advances in Neural Information Processing Systems* **31**.