Sample Research Proposal

Debiased Causal Inference in High-Dimensional (Financial) Networks

Abstract

Many modern and complex data structures, for instance, financial markets, constitute complex, high-dimensional networks where the behavior of one entity can influence others, leading to systemic risk and contagion. Understanding the **causal** magnitude of these spillovers is critical for robust risk management and financial policy, yet most network analysis in finance remains correlational. Current causal inference methodologies are primarily tailored for social science experiments, traditional econometrics, or biological networks, and do not adequately address the unique challenges of high-dimensional, observational financial data, which is rife with confounding and endogeneity. This proposal outlines a novel framework to bridge this gap by integrating Double/Debiased Machine Learning (DML) with modern network econometric models. I propose to adapt the DML framework, which leverages Neyman-orthogonal moments and cross-fitting, to estimate causal peer effects within a Network Vector Autoregression (NAR) structure. This approach allows for the estimation of low-dimensional causal parameters (e.g., the strength of network spillovers, β_1) while non-parametrically controlling for high-dimensional nuisance parameters. The expected contributions are: (1) a novel, statistically rigorous estimator for causal peer effects in observational financial networks; (2) a method for constructing valid, network-robust confidence intervals for these effects; and (3) an empirical quantification of causal shock propagation and systemic risk in a real-world financial dataset. This research aims to provide a vital tool for regulators and financial institutions to move from correlation to causation in their risk models.

Keywords: Causal Inference, Double/Debiased Machine Learning, Network Vector Autoregression, High-Dimensional Econometrics, Financial Networks, Systemic Risk.

1 Introduction and Purpose

The 2008 financial crisis and subsequent periods of extreme market volatility serve as stark reminders of the interconnectedness of the global economic system. A shock originating in one corner of the market can propagate through a complex web of counterparty obligations and shared information channels, leading to systemic failure. Consequently, financial systems are increasingly modeled as networks (Sadeghi et al., 2023). A critical challenge, however, is distinguishing causality from mere correlation. For instance, if two banks' stock prices fall simultaneously, is it because one's distress caused the other's (influence), or because both are exposed to the same external risk factors (confounding)? This is the classic reflection problem in network science (Manski, 1993; An et al., 2022). Misattributing correlation to causation leads to flawed risk models and ineffective policy.

While recent advances have pushed financial network analysis towards causality, a significant gap remains. Current research often focuses on causal discovery (i.e., learning the network structure) or prediction in specific verticals like credit risk. My research, in contrast, focuses on causal inference: rigorously estimating the magnitude and statistical significance of causal links in an observational setting, which is essential for policy and

what-if scenario analysis.

Econometric models for network data, such as the Network Vector Autoregression (NAR) model (Zhu et al., 2017), offer a promising foundation. The NAR model parsimoniously captures the dynamics of a node's response (Y_{it}) as a function of its past value, its covariates, and the average of its neighbors' past values (the network effect, β_1). This β_1 parameter is precisely what we want to estimate. However, in modern finance, the set of potential confounders (e.g., firm fundamentals) is high-dimensional, rendering standard OLS estimates of β_1 biased and inconsistent.

This research proposes to solve this specific problem by integrating the NAR framework with Double/Debiased Machine Learning (DML) (Chernozhukov et al., 2018). DML is a state-of-the-art method designed to estimate a low-dimensional parameter of interest (like β_1) in the presence of high-dimensional nuisance parameters. It uses Neyman-orthogonal moments to become robust to estimation errors in the nuisance functions and employs cross-fitting to eliminate bias from overfitting (Gao and Ding, 2025). By adapting DML to the financial network context, I aim to deliver the first robust, $\sqrt{n}$-consistent estimates of causal network effects from high-dimensional, observational financial data. My prior research experience in causal machine learning, specifically DML and Neyman Orthogonality, provides the exact theoretical toolkit required for this challenge.

2 Research Questions

This proposal aims to develop and validate a new methodology for causal inference in financial networks. This leads to the following specific research questions:

1. **Methodological:** How can the Double/Debiased Machine Learning (DML) framework be formally adapted to estimate the causal network effect (e.g., β_1 in a Network Vector Autoregression model) in a setting where high-dimensional nodal covariates confound financial asset returns?

2. **Inferential:** Building on this, how to construct valid, network-robust confidence intervals for (a) the average causal network effect across the system and (b) a specific node's (e.g., a systemically important bank's) causal influence on its neighbors, accounting for the network-induced dependence in the data?

3. **Empirical:** What is the empirical magnitude, sign, and statistical significance of these causal spillover effects in the US equity market? How does a shock to a central node causally propagate, and can this framework provide a more accurate, causation-based measure of systemic risk?

3 Review of Related Literature and Techniques

3.1 From Correlation to Causality in Financial Networks

The modeling of financial markets as complex networks has provided deep insights, but most studies rely on correlation-based metrics. Recently, the field has begun to embrace causality (Giamouridou et al., 2024). Sadeghi et al. (2023) introduces CD-NOTS, a framework for causal discovery in financial time series, which is a significant step forward as it explicitly handles the non-stationary nature of the data. Similarly, Liu et al. (2024)

applies Bayesian networks to the specific problem of credit risk prediction. While these studies validate the pursuit of causality in finance, their focus remains distinct from this proposal. Causal discovery aims to find the existence of an edge (Jiang et al., 2025), and credit risk models are for prediction. My research focuses on causal inference: estimating the magnitude and statistical significance of a specific causal link (e.g., peer effect) in an observational setting, which is essential for policy and what-if scenario analysis.

3.2 Causal Inference in Network Settings

Estimating causal effects in networks is notoriously difficult due to three primary challenges: confounding, homophily versus influence (the reflection problem), and interference (An et al., 2022; Manski, 1993). Several approaches have been proposed to tackle these issues.

First, Network Experiments, where treatment is randomized, represent the gold standard. Gao and Ding (2025) provides a rigorous analysis of such experiments, using regression-based Hájek estimators and network-robust HAC-type standard errors to account for interference. However, large-scale, randomized experiments are generally infeasible in macro-finance; thus, we must rely on observational methods. Second, Longitudinal Models like Stochastic Actor-Oriented Models (SAOMs) attempt to disentangle selection and influence by jointly modeling the co-evolution of the network structure and node behavior over time (An et al., 2022). While powerful, they are computationally intensive and make strong assumptions about actors' decision processes. Third, Econometric Network Models offer a more scalable alternative. The **Network Vector Autoregression (NAR) model** (Zhu et al., 2017) is a parsimonious framework. It models the response of node i at time t (Y_{it}) as:

$$Y_{it} = \beta_0 + Z_i^\top \gamma + \beta_1 \underbrace{(n_i^{-1} \sum_{j=1}^{N} a_{ij} Y_{j(t-1)})}_{\text{Network Effect}} + \beta_2 \underbrace{Y_{i(t-1)}}_{\text{Momentum Effect}} + \epsilon_{it}$$

This model elegantly captures peer effects (β_1), momentum (β_2), and heterogeneity (γ). The central problem, which this proposal addresses, is that this model is only a standard regression if the covariates Z_i are low-dimensional. In modern finance, Z_i is high-dimensional. Estimating β_1 via OLS will suffer from omitted variable bias, and using regularized regression (e.g., Lasso) on the entire equation will yield a biased and inconsistent estimate of β_1.

3.3 Double/Debiased Machine Learning (DML) Framework for Causal Inference

This proposal's core methodological solution stems from Chernozhukov et al. (2018), the foundational paper on Double/Debiased Machine Learning (DML). DML is designed for semi-parametric models of the form $Y = \theta_0 D + g(Z) + \epsilon$, where we want to estimate a low-dimensional causal parameter θ_0 (our network effect β_1) in the presence of a high-dimensional, unknown nuisance function $g(Z)$ (our nodal effect $g_0(Z_i)$).

A naive plug-in approach, where one first uses an ML algorithm to estimate $\hat{g}(Z)$ and then regresses the residual $Y - \hat{g}(Z)$ on D, fails. This estimator is contaminated by regularization bias and overfitting, and it is not $\sqrt{n}$-consistent or asymptotically normal, invalidating standard inference.

The DML solution introduces two key ingredients. First, it uses a **Neyman-orthogonal score**. In the context of the partially linear model (PLR), this involves partialling out Z from both Y and D. We estimate the nuisance functions $E[Y|Z] = g(Z)$ and $E[D|Z] = m(Z)$. The orthogonal score is based on the residuals $V = D - m(Z)$ and $U = Y - g(Z)$. The parameter θ_0 is then estimated from $E[VU] = E[V^2]\theta_0$. Because the moment condition is insensitive to first-order errors in estimating g and m, the resulting estimator is robust (Foster and Syrgkanis, 2019).

Second, to remove the bias from overfitting, DML employs **cross-fitting**. The data is split into K folds. For each fold k, the nuisance functions $\hat{g}_k$ and $\hat{m}_k$ are trained on the other $K-1$ folds. The residuals and the final estimate of θ_0 are then computed on fold k. This ensures the nuisance function estimates are statistically independent of the data used for the final estimation, which provably removes the bias and restores $\sqrt{n}$-consistency. My research plan is to apply this robust DML-PLR framework to the NAR model.

4 Proposed Methodology

4.1 Model Formulation

I begin with a generalized version of the Network Vector Autoregression (NAR) model (Zhu et al., 2017). Let Y_{it} be the outcome of interest (e.g., log-return) for firm i at time t. Let Z_i be a high-dimensional vector of time-invariant or slowly-varying covariates (e.g., sector, firm fundamentals). Let X_t be a high-dimensional vector of time-varying market-wide confounders (e.g., interest rates, sentiment indicators). I propose the following semi-parametric network model:

$$Y_{it} = \theta_0 D_{it} + g_0(Z_i, X_t) + \beta_2 Y_{i(t-1)} + \epsilon_{it}$$

Here, θ_0 is the low-dimensional parameter of interest: the average causal network effect. $D_{it} = n_i^{-1} \sum_{j=1}^{N} a_{ij} Y_{j(t-1)}$ is the treatment variable, representing the lagged influence from neighbors. The term $g_0(Z_i, X_t)$ is the high-dimensional nuisance function. I can absorb the momentum term $\beta_2 Y_{i(t-1)}$ into g_0, as $Y_{i(t-1)}$ can be treated as just another (lagged) covariate. This transforms the problem into the canonical DML setup, the Partially Linear Model (PLR) (Chernozhukov et al., 2018):

$$Y_{it} = \theta_0 D_{it} + g_0(W_{it}) + \epsilon_{it}$$

where $W_{it} = (Z_i, X_t, Y_{i(t-1)})$ is the complete high-dimensional vector of confounders.

4.2 Estimation via Double-Debiased Machine Learning

To estimate θ_0 while non-parametrically controlling for $g_0(W_{it})$, I will implement the DML-PLR algorithm:

1. **Data Splitting (Cross-Fitting):** Randomly partition the full sample of $N \times T$ observations into K folds (e.g., $K = 5$).
2. **Nuisance Function Estimation:** For each fold $k \in \{1, ..., K\}$:
   - Use the data from all folds except k to train two machine learning models (e.g., Lasso, Random Forests, or Boosted Trees):

  - An outcome model: $\hat{g}_{0,k}(w) \approx E[Y_{it}|W_{it} = w]$.
  - A treatment model: $\hat{m}_{0,k}(w) \approx E[D_{it}|W_{it} = w]$.

3. **Orthogonalized Residuals:** Using only the data in fold k, compute the Neyman-orthogonalized residuals:
   - Outcome residual: $\hat{U}_{it} = Y_{it} - \hat{g}_{0,k}(W_{it})$
   - Treatment residual: $\hat{V}_{it} = D_{it} - \hat{m}_{0,k}(W_{it})$
4. **Final Estimator:** The DML estimator for the causal network effect $\tilde{\theta}_0$ is the pooled OLS coefficient:

$$\tilde{\theta}_0 = \left(\sum_{i,t} \hat{V}_{it}^2 \right)^{-1} \left(\sum_{i,t} \hat{V}_{it} \hat{U}_{it} \right)$$

As established by Chernozhukov et al. (2018), this estimator $\tilde{\theta}_0$ is asymptotically normal and $\sqrt{NT}$-consistent, provided the nuisance models (ML) converge at a sufficient rate (e.g., $o_p(n^{-1/4})$ in MSE).

4.3 Statistical Inference and Network-Robust Variance

A key contribution will be to develop the correct variance estimator for $\tilde{\theta}_0$ in this network context. The observations (U_{it}, V_{it}) are not IID; they are correlated across i (contemporaneous network dependence) and t (time-series dependence). The standard DML variance formula $\hat{\sigma}^2 = \hat{E}[V^2]^{-2}\hat{E}[V^2U^2]$ must be adapted.

I will derive a network-robust (HAC-type) variance estimator, drawing from the literature on spatial and network econometrics (Gao and Ding, 2025; An et al., 2022; Zhu et al., 2017). The asymptotic variance will have the form $Var(\tilde{\theta}_0) = \Omega^{-1}\Psi\Omega^{-1}$, where $\Omega = E[V_{it}^2]$ and Ψ is the long-run covariance matrix of the orthogonalized product $V_{it}U_{it}$, which accounts for network and time dependencies. This will allow for the construction of valid confidence intervals and hypothesis tests.

4.4 Research Significance and Impact

This research will be one of the first to provide a rigorous, general-purpose framework for debiased causal inference in high-dimensional observational financial networks. It synthesizes state-of-the-art econometric machine learning (DML) with modern network econometrics (NAR). Methodologically, this is a significant contribution. Practically, this methodology will empower financial institutions and regulators to move beyond simple correlation-based risk models. It will allow them to quantify the causal magnitude of spillovers, identify systemically important firms based on their causal influence, and conduct what-if scenario analysis (Liu et al., 2024) on the propagation of shocks. This provides a data-driven tool for macroprudential policy and systemic risk management.

5 Expected Results

First, on a theoretical level, I will formally prove the asymptotic normality and $\sqrt{NT}$-consistency of the proposed DML-NAR estimator $\tilde{\theta}_0$ under suitable regularity conditions for financial network data (e.g., bounds on network sparsity, weak dependence).

Second, through simulation, I will conduct extensive Monte Carlo studies. These will demonstrate that: (a) Naive plug-in ML estimators are severely biased and their CIs have poor coverage, and (b) The proposed DML-NAR estimator is approximately unbiased and its CIs achieve the nominal 95% coverage rate (Chernozhukov et al., 2018). This will validate the theoretical claims.

Third, I will apply the framework empirically to a large dataset of S&P 500 firms. The network A will be constructed from market-based data, for instance, using **regularized partial correlations (e.g., Graphical Lasso) or Granger causality** to estimate the conditional dependency structure, or based on known supply-chain linkages. Y_{it} will be daily returns, and W_{it} will be a high-dimensional set of firm fundamentals, technical indicators, and text-based news sentiment data. I expect to find a statistically significant positive causal network effect ($\tilde{\theta}_0 > 0$), quantifying the magnitude of contagion. This will allow me to rank firms by their causal contribution to systemic risk.

6 Research Plan

This research is designed to be completed within a 5-year PhD program.

Phase	**Year 1**	**Year 2**	**Year 3**	**Year 4-5**
Coursework	Advanced Causal Inference, Financial Econometrics, ML Theory, Advanced Probability		Thesis Research	
Research	Extensive Lit. Review (DML, Networks)	Theoretical Derivations (Asymptotics, Variance)	Data Acquisition (S&P 500), Data Cleaning	Thesis Compilation & Defense
Milestone 1	Refine theoretical framework	Complete simulation framework	Complete empirical analysis	Submit Paper 2 / Extension
Milestone 2	Develop initial simulation code	Draft Paper 1 (Methodology)	Draft Paper 2 (Empirical Applications)	Defend Thesis
Output	Research proposal finalized	Submit Paper 1 to journal (e.g., JASA, JoE)	Present at international conference	Final Dissertation

7 Expected Impact and Discussion

This research is positioned at the fertile intersection of causal ML, econometrics, and finance. By developing a rigorous, interpretable, and high-dimensional causal model, this work will provide a critical toolkit for understanding one of the most fundamental questions in modern finance: how does risk really spread?

For academia, it bridges a significant gap between econometric theory and financial practice, offering a tool that respects the complex, high-dimensional nature of economic data. For policy and society, the impact is more direct. Regulators are tasked with maintaining financial stability, yet their tools for measuring systemic risk are often based

on size or correlation, not causal influence. A successful outcome of this research would provide a data-driven, causal measure of systemic importance. This would allow for more effective and targeted macroprudential policies, helping to identify and mitigate the risks posed by too-connected-to-fail institutions before they trigger a systemic crisis.

References

Weihua An, Roberson Beauville, and Benjamin Rosche. Causal network analysis. *Annual Review of Sociology*, 48(1):23–41, 2022. doi: 10.1146/annurev-soc-030320-102100. URL https://doi.org/10.1146/annurev-soc-030320-102100.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. In *Conference on Learning Theory*, pages 1290–1321. PMLR, 2019. URL http://proceedings.mlr.press/v99/foster19a.html.

Mengsi Gao and Peng Ding. Causal inference in network experiments: regression-based analysis and design-based properties. *Journal of Econometrics*, 252:106119, 2025. URL https://arxiv.org/abs/2309.07476.

Myrsini Giamouridou, Spyridon Vrontos, Panagiotis Galakis, and Ioannis Vlachos. Causal machine learning for finance: A practical overview. *Expert Systems with Applications*, 254:124555, 2024. doi: 10.1016/j.eswa.2024.124555. URL https://doi.org/10.1016/j.eswa.2024.124555.

Hongyang Jiang, Yuezhu Wang, Ke Feng, Chaoyi Yin, Yi Chang, and Huiyan Sun. Biological regulatory network inference through circular causal structure learning. *arXiv preprint arXiv:2511.02332*, 2025. doi: 10.48550/arXiv.2511.02332. URL https://arxiv.org/abs/2511.02332.

Jiaming Liu, Xuemei Zhang, and Haitao Xiong. Credit risk prediction based on causal machine learning: Bayesian network learning, default inference, and interpretation. *Journal of Forecasting*, 43(5):1625–1660, 2024. doi: 10.1002/for.3080. URL https://doi.org/10.1002/for.3080.

Charles F Manski. Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542, 1993. doi: 10.2307/2298123. URL https://doi.org/10.2307/2298123.

Agathe Sadeghi, Achintya Gopal, and Mohammad Fesanghary. Causal discovery in financial markets: A framework for nonstationary time-series data. *arXiv preprint arXiv:2312.17375*, 2023. URL https://arxiv.org/abs/2312.17375.

Xuening Zhu, Rui Pan, Guodong Li, Yuewen Liu, and Hansheng Wang. Network vector autoregression. *The Annals of Statistics*, 45(3):1096–1123, 2017. doi: 10.1214/16-AOS1476. URL https://doi.org/10.1214/16-AOS1476.