Pricing Offshore Services: Evidence from the Paradise Papers

Marcus Gawronsky

Supervisors: *Associate Professors Tim Gebbie and Kanshukan Rajaratnam*

University of Cape Town
Minor Dissertation presented in
partial fulfilment of the requirements for the degree of
Masters in Advanced Analytics and Decision Sciences (STA5004W)

Abstract

The Paradise Papers represent one of the largest public data leaks comprising 13.4 million confidential electronic documents. A dominant theory presented by Neal (2014) and Griffith, Miller and O'Connell (2014) concerns the use of these offshore services in the relocation of intellectual property for the purposes of compliance, privacy and tax avoidance. Building on the work of Fernandez (2011), Billio et al. (2016) and Kou, Peng and Zhong (2018) in Spatial Arbitrage Pricing Theory (s-APT) and work by Kelly, Lustig and Van Nieuwerburgh (2013), Ahern (2013), Herskovic (2018) and Procházková (2020) on the impacts of network centrality on firm pricing, we use market response, discussed in O'Donovan, Wagner and Zeume (2019), to characterise the role of offshore services in securities pricing and the transmission of price risk. Following the spatial modelling selection procedure proposed in Mur and Angulo (2009), we identify Profit Margin and Price-to-Research as firm-characteristics describing market response over this event window. Using a social network lag explanatory model, we provide evidence for social exogenous effects, as described in Manski (1993), which may characterise the licensing or exchange of intellectual property between connected firms found in the Paradise Papers. From these findings, we hope to provide insight to policymakers on the role and impact of offshore services on securities pricing.

Keywords: Paradise Papers, Social Network Econometrics, Spatial Arbitrage Pricing Theory

Acknowledgements

I would like to extend our gratitude to Associate Professors Tim Gebbie and Kanshukan Rajaratnam for their warm and valued mentorship through my time at the University of Cape town. Their counsel has been extraordinarily generous offering profound insight into our role as scholars on society. Their astute guidance and technical contributions have provided this research with the rich and diverse perspectives on which this analysis is built and has provided an incredible journey though this fascinating literature. I would also like to thank the contributions of both my peers and senior academics in the University of Cape Town Statistical Finance Research Group and Statistical Science Department, whose perspectives have strengthened this work greatly.

Declaration

1. I am presenting this dissertation in partial fulfilment of the requirements for my degree.
2. I know the meaning of plagiarism and declare that all of the work in the dissertation, save for that which is properly acknowledged, is my own.
3. I hereby grant the University of Cape Town free licence to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever of the above dissertation.
4. I have used the Author Year convention for citation and referencing. Each contribution to, and quotation in, this dissertation from the work(s) of other authors has been attributed, and has been cited and referenced.

Approved: _____

Marcus Gawronsky
Master of Science Candidate

Accessibility

This work has made a number of unorthodox decisions regarding its choice of formatting and color maps which may be unfamiliar to some readers. These decisions have been made based on research across peer reviewed user studies on the accessibility of particular fonts and palettes to users with certain vision impairment or learning disorders (Mikhailov, 2019; Rello and Baeza-Yates, 2013; Wery and Diliberto, 2017). Where readers face particular challenge not presented in these studies, they are most welcome to contact the authors for assistance.

Reproduciblity

Code for our analysis was written in Python 3.8.2 using GCC release 7.3.0. All software in this research is version controlled and has been made available on GitHub in Gawronsky, Gebbie and Rajaratnam (2020*b*) with reference to a specific commit hash. The data used in our research has been make available on ZivaHub in Gawronsky, Gebbie and Rajaratnam (2020*a*) along with all intermediate analysis. For readers looking to review and reuse our software documentation has been made available in Gawronsky, Gebbie and Rajaratnam (2020*c*). A detailed list of dependencies has been made available through this repository for those looking to reproduce our analysis and has been tested using version 4.8.3 of the Anaconda Package Manager (Anaconda Inc., 2020). While our research has made broad use of the PyData Ecosystem, we have detailed specific requirements on PySal 2.1.0 for its trusted and benchmarked implementations in spatial regression, Scipy 1.5.2 and ARPACK for efficient linear algebra over sparse matrices, NetworkX 2.5 for its efficient implementation of common graph algorithms, and PyGSP 0.5.1 for its tested implementations in Graph Signal Processing (Virtanen et al., 2020; Lehoucq et al., 1977; Rey and Anselin, 2007; Hagberg et al., 2008). All intermediate data has been versioned throughout our analysis using Kedro 0.16.5, which has provided our work with a best-in-class framework for data pipelining (QuantumBlack, 2020). While certain exploratory work was performed on commodity cloud computing hardware, the majority of our analysis has relied on consumer hardware with 16GB of Micron MT53E1G32D4NQ-046 memory clocked at 4267MHz and a Quadcore Intel® Core™i7-1065G7 processor clocked at 1.30GHz. At the release of our work, we are unaware of any known issues in the software and hardware used by our analysis which may affect the reliability of our results.

Notation

Across our discussions particular markup is used to clarify our use of mathematical notation. $\vec{x}$, indicated in lower-case and in bold will be used to denote a vector, X provided in upper-case and in bold will be used to indicate a matrix and x offered in neither in upper-case nor a bold font will indicate some scalar value. Certain constant values like N, are used to indicated the number of observations in a particular statistic is in upper-case but not bold. This work has looked to follow notation offered by original authors where possible and should offer clarity to readers in the captions accompanying our equations.

Contents

List of Figures

List of Tables

List of Exhibits

1 Introduction

On the 7 November 2017, the International Consortium of Investigative Journalists (ICIJ) released one of history's largest public financial data leaks. This leak, dubbed the Paradise Papers, describes a corpus of some 13.4 million confidential electronic documents relating to offshore investments managed by legal firm Appleby and spans over 120000 persons and companies across 19 different tax jurisdictions (International Consortium of Investigative Journalists, n.d.*b*).

Inside of academia, data leaks held by the ICIJ have seen interest across disciplines from research in Information Systems, to research in Tax Law and Multidimensional Visualization (Zhuhadar and Ciampa, 2019; Wiedemann et al., 2018). While the ICIJ release only metadata on the leaks, this metadata has been used to describe rich graphs of entities, their relationships and jurisdictions, which provide insight into the role and structure offshore special purpose vehicles in ensuring company privacy, compliance and tax avoidance. Research by O'Donovan et al. (2019) has used this information to price the impact of these leaks based on market reactions. Work by Hajek and Henriques (2017) and Garcia Alvarado and Mandel (2019) has looked at this metadata in bad entity detection to explore optimal privacy and avoidance strategies for markets on graphs.

In Financial Econometrics, graphs have played an important role in the study of capital flow, direct contagion and market position, stemming from microeconomic foundations on graph centrality (Chen, Zenou and Zhou, 2018). These works, presented in Sections 2.5 and 2.6, have either learned, constructed or sampled graph structures in order to explore centrality, social network effects or spatial correlation in asset prices (Procházková, 2020; Kelly et al., 2013; Billio et al., 2016; Ahern, 2013; Fernandez, 2011). Many of these techniques make important assumptions when sampling or constructing their graphs concerning the local structure of their network and the observation of price across all entities. In real-world graphs, prices may not be directly observable along all nodes in the graph and may feature local structure relating to requirements around compliance, privacy or tax avoidance which may obfuscate the substance of counter-party risk.

In work by Griffith et al. (2014), discussed in Section 2.7, researchers explore the impact of regulatory changes on the location of new intellectual property and its impact on tax revenue under a mixed (or random coefficients) logit model. This provides insight into the role and response of regulators to the establishment of IP holding and development companies. Using metadata from the Paradise Papers, researchers are given a natural experiment through which to understand the impact of sentiment and regulatory changes on market pricing, as well the impact or presence of social network effects which propagate risk along the graph.

In this work, we look to characterise the use of offshore services based on market reaction. Using work by Kelly et al. (2013), O'Donovan et al. (2019), Neal (2014) and Griffith et al. (2014), we investigate a range of firm characteristics which we believe may characterize market response to this leak and extend this work using methods in spatial econometrics and Spatial Arbitrage Pricing Theory (s-APT) to investigate the social network effects which may impact price over our window. To work by Billio et al. (2016), Fernandez (2011) and Kou et al. (2018) our research serves as a fascinating application of methods to social network data, to work by O'Donovan et al. (2019), Neal (2014) and Griffith et al. (2014), our work presents findings concerning the social network effects of Price-to-Research and Profit Margin, presented in Table 10 of Section 6.2.

This thesis will look to introduce important concepts in spatial statistics, financial econometrics and graph theory, which are used to motivate our application of spatial econometrics and graph-based approaches to the ICIJ leaks. In Section 2, we introduce important findings concerning the Paradise Papers and ICIJ leaks, which frame our work in Section 4.2 characterizing the structural components of our graph. In Section 4.4, we explore intermediate structures found between matched listed entities along our graph using Graph Gourier Analysis and common shortest-path algorithms to detail path lengths spanning 4 or 6 nodes which reach through Bermuda across US-based companies. Using these findings, in Section 4.3 we construct a spatial weighting matrix through which we explore and identify a spatial lag explanatory model comprising Price-to-Research and Profit Margins as firm characteristics using spatial selection procedure proposed in Mur and Angulo (2009). From this work, presented in Table 10 and figures 24 and 25, we determine the presence of social effects which we believe evidence the use of offshore services in relocating intellectual property, as discussed in Griffith et al. (2014) and O'Donovan et al. (2019). Using this finding, we hope to have provided both investors and policy-makers insight into the role and importance of offshore services which we believe may impact future policy decisions.

2 Literature Review

In this literature review, we aim to provide the necessary foundations in Graph Theory, Social Network and Spatial Econometrics and Financial Economics through which to explore their interest in Microeconomics, Game Theory and Spatial Arbitrage Pricing Theory. This work will present existing findings from these domains in Financial Markets and detail existing research on Financial Data Breaches. This review pieces together a broad literature and so focuses its efforts to particular areas of Spatial Statistics and Graph Theory unfamiliar to readers from Financial Economics and the Statistical Sciences. We provide the reader with this broad context through which to motivate the interest and impact of our study in motivating and exploring particular methodologies for our unique data.

2.1 Financial Economics

In Modern Financial Economics asset prices are considered a function of expected discounted future cash flows and market risk-preference. While the concept of Market Efficiency is strongly debated, the idea that prices adjust to reflect the expected impact of information on discounted future cash flows forms a common premise to empirical research on market behaviour.

According to Arbitrage Pricing Theory (APT), given a large enough a universe of investable assets and adequate access to liquidity, information should be priced instantaneously by the market. By construction, Ross (1976) demonstrates this property of market pricing at the limit to yield a linear pricing rule for investable portfolios of risk factors, which has been used by many authors to identify sources of risk priced by the market (Eugene and French, 1992; Asness, 1997).

Using APT, many authors have explored a broad range of stylized facts through the application of factor models with which researchers have identified attributes such as momentum, size, value, investment and profitability as characteristics priced by the market through the construction of factor mimicking portfolios (Fama and French, 1992; Banz, 1981; Asness, 1997).

In Haugen, Baker et al. (1996), the authors provide an alternative framework through which to describe market pricing kernels using characteristic-based modelling (CBM). This approach offers insight into the non-linear nature through which information may be assimilated into market prices through firm characteristics. In Wilcox and Gebbie (2015), authors compare the out-of-sample, ex-ante risk and returns for models based on APT under risk-factor based, zero-cost portfolios with characteristic-based, zero-cost portfolios using lagged risk factors. Under this approach, Wilcox and Gebbie (2015) demonstrate that while cross-sectional characteristic-based models yield portfolios with higher excess monthly returns and lower risk, CBM can still yield pricing rules which are consistent with linear pricing kernels where parameter estimates are a function of characteristics.

An important note is that while information may be assimilated instantaneously by the market, serial correlation and statistical arbitrages are permitted on the short term based on market microstructure, liquidity and transaction costs, though they may not be exploitable by investors Asness (1997).

2.2 Spatial and Social Network Econometrics

Inside of Spatial Econometrics, the term spatial auto-correlation is used to describe the tendency for samples close in space to take on similar (positive spatial correlated) or dissimilar (negative spatial correlated) values. In Radil (2011), a diagram is provided, common across published works, which uses grids of coloured squares to illustrate this concept visually for readers, shown in Figure 1.

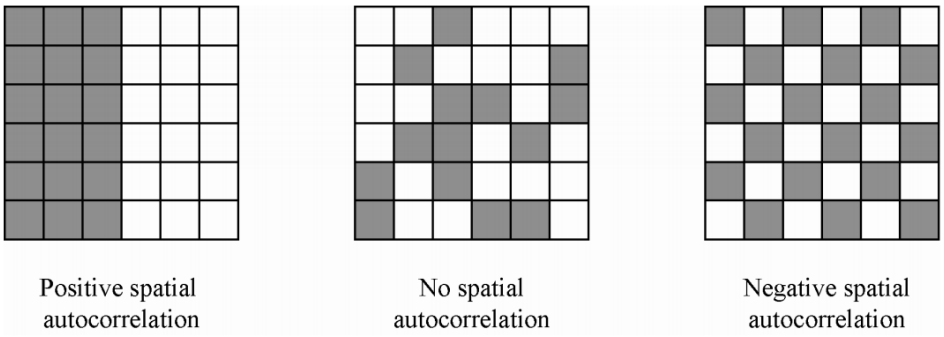


Figure 1: Diagram of spatial correlation as provided in Radil (2011). In the left-most facet we observe a tendency of black squares to neighbour other black squares and white squares to neighbour other white squares. This propensity of similarly shaded squares to lie adjacent to one another represents positive spatial correlation. In the right-most facet the tendency of white squares to lie adjacent to black squares and white squares to lie adjacent to black squares under von Neumann adjacency, represents negative spatial correlation. In the centre facet, the placement of shaded squares appears random and suggests no spatial autocorrelation among shaded squares is present.

While the diagram in Figure 1 shows spatial correlation under von Neumann adjacency, researchers rely on a number of approaches upon which to build either fixed or adaptive spatial weighting matrices, with the common application of either Bisquare or Gaussian kernels. Under fixed kernels, kernel bandwidth parameters are fixed according to the spatial distribution of the data, domain expertise or guiding hypothesis of the study; under adaptive kernels, bandwidth is chosen under optimisation using either Bayesian methods, meta-heuristics or grid-search under k-fold cross-validation. Importantly, the key requirement in selecting a spatial weighting matrix is the ability to establish some quasimetric to construct a row-normalised matrix (Miller and Wentz, 2003). As studies in spatial regression typically investigate spatial correlation across some manifold or vector-space, Euclidean distances form a common metric upon which to build the spatial weighting matrix. However, given the many overlaps of spatial regression with auto-regressive time-series models, symmetry can easily be broken based on some assumed structure in the data. Using a Spatial Auto-regression Model, an AR1 process could be modelled using an off-diagonal spatial weighting matrix, while a moving average (MA) model could be modelled based on the structure of lagged errors. As with the use of adjacency matrices to represent adjoining suburbs, countries or continents, graphs can be used to construct a spatial weighting matrix for use in spatial analysis. This matrix could be symmetric, as would be the case for an undirected graph, or asymmetric, as we would see in a directed graph. This formulation of spatial weighting matrices is common in Social Network Econometrics in which researchers typically look to investigate the correlations of participant responses across peer groups.

Social Endogenous Social Correlated

$$Y = \boxed{\rho WY} + \alpha \mathbf{1_N} + X\boldsymbol{\beta} + \boxed{WX\boldsymbol{\theta}} + \boldsymbol{u}; \boldsymbol{u} = \boxed{\lambda W \boldsymbol{u}} + \boldsymbol{\epsilon} \tag{1.1}$$

Social Exogenous

Exhibit 1: Manski (1993) Model of Social Effects, where in Y is our response variable, X our exogenous variables, W a row-normalized spatial weighting matrix describing the distance or relationship between participants or sampled observations and ϵ our error term. β and θ represent vectors of regression coefficients for non-spatial and spatially lagged exogenous variables. With scalars ρ and λ presenting the effect of our spatially lagged endogenous and spatially lagged error terms. This forms the basis for equation 2.1 in Exhibit 2.

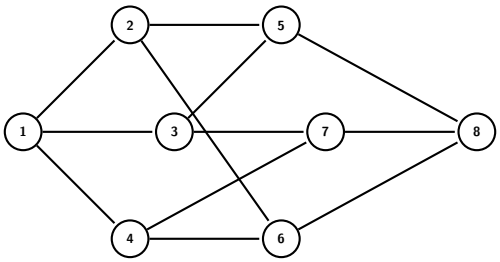
While Social Network Econometrics shares many techniques in the field of Spatial Econometrics, work by Manski (1993) describes important challenges in model identification, which come as a result of the unique spatial effects economists look to explore. These effects comprise some combination of Endogenous effects, wherein individual behaviour varies against its reference group, Exogenous (contextual) effects, which describes the propensity of individuals behaviour to vary due to the characteristics of its neighbours or reference group, and Correlated effects, whereby individual responses are described

by the propensity of neighbours or a reference group to share common characteristics, as described in Figure 1.

By formalizing the linear-in-means model, shown in the equation in Exhibit 1, Manski (1993) demonstrates key limitations in distinguishing real social effects (endogenous + exogenous) from correlated effects in low-rank networks which describe non-overlapping communities with isolated individuals. In Bramoullé, Djebbari and Fortin (2009), this challenge lies in connected students sharing similar socioeconomic attributes, such as age, gender or parental income, or involvement in recreational services. Under such a condition, modelling approaches may be unable to distinguish the effects of social interaction from the effects of students' shared socioeconomic attributes at the point at which these are highly correlated in the data.

In Bramoullé et al. (2009), researchers demonstrate the important conditions under which endogenous and exogenous effects can be distinguished, as related to the linear dependence of the network. Bramoullé et al. (2009) lean on the two-stage least-squares regression approach presented by Lee (2003) and Kelejian and Prucha (1998) to identify appropriate instruments to control for multicollinearity between exogenous and endogenous social effects, which has been seen in numerous other studies, including Angrist (2014) and De Giorgi et al. (2010) in exploring the impact of peer effects in education.

To deal with these challenges in identification, researchers in spatial and social network econometrics have experimented with many approaches in testing unsaturated models under various statistical paradigms. A diagram in Elhorst (2010) details the structural assumptions which describe this hierarchy of these unsaturated models, shown in Exhibit 2.



1. **Manski (1993) Model:** which assumes the presence of all real and correlated social effects

$$\begin{aligned} \boldsymbol{y} &= \rho W\boldsymbol{y} + \alpha\mathbf{1_N} + X\beta + WX\theta + \boldsymbol{u}; \\ \boldsymbol{u} &= \lambda W\boldsymbol{u} + \epsilon \end{aligned} \tag{2.1}$$

2. **Kelejian-Pruscha Model (KP):** which assumes only correlated and endogenous social effects

$$\begin{aligned} \boldsymbol{y} &= \rho W\boldsymbol{y} + \alpha\mathbf{1_N} + X\boldsymbol{\beta} + \boldsymbol{u}; \\ \boldsymbol{u} &= \lambda W\boldsymbol{u} + \boldsymbol{\epsilon} \end{aligned} \tag{2.2}$$

3. **Spatial Durban Model (SDM):** which assumes only real social endogenous and exogenous social effects

$$\boldsymbol{y} = \rho W\boldsymbol{y} + \alpha\mathbf{1_N} + X\boldsymbol{\beta} + WX\theta + \boldsymbol{\epsilon}; \tag{2.3}$$

4. **Spatial Durban Error Model (SDEM):** which assumes only exogenous and correlated social effects

$$\begin{aligned} \boldsymbol{y} &= \alpha\mathbf{1_N} + X\boldsymbol{\beta} + WX\theta + \boldsymbol{u}; \\ \boldsymbol{u} &= \lambda W\boldsymbol{u} + \boldsymbol{\epsilon} \end{aligned} \tag{2.4}$$

5. **Simultaneous Autoregressive Model (SAR) or Spatial Lag Model (SLM):** which assumes only endogenous social effects

$$\boldsymbol{y} = \rho W\boldsymbol{y} + \alpha\mathbf{1_N} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}; \tag{2.5}$$

Simultaneous Autoregressive Moving Average Model (SARMA):

$$\begin{aligned} \boldsymbol{y} &= \rho W\boldsymbol{y} + \alpha\mathbf{1_N} + X\boldsymbol{\beta} + \boldsymbol{u}; \\ \boldsymbol{u} &= (I - \boldsymbol{\theta}W)\boldsymbol{\epsilon} \end{aligned} \tag{2.6}$$

6. **Spatial Error Model (SEM):** which assumes only correlated social effects, and

$$\begin{aligned} \boldsymbol{y} &= \alpha\mathbf{1_N} + X\boldsymbol{\beta} + \boldsymbol{u}; \\ \boldsymbol{u} &= \lambda W\boldsymbol{u} + \boldsymbol{\epsilon} \end{aligned} \tag{2.7}$$

7. **Spatial Lag Explanatory Model (SLX):** which is not detailed in the Elhorst (2010) diagram but assumes only exogenous social effects

$$\boldsymbol{y} = \alpha\mathbf{1_N} + X\boldsymbol{\beta} + WX\boldsymbol{\theta} + \boldsymbol{\epsilon}; \tag{2.8}$$

8. **(OLS) Linear Model:** which assumes no spatial or social network effects

$$\boldsymbol{y} = \alpha\mathbf{1_N} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}; \tag{2.9}$$

Exhibit 2: Hierarchy of Unsaturated Model Spatial Models presented in Elhorst (2010). In this diagram and accompanying equations, we shown the relationships between different spatial models, moving from the Manski Model to non-spatial regression model as we constrain spatial or social network effects. In the equations above $\boldsymbol{y}$ represents our endogenous variables, X our exogenous variables and W our spatial weighting matrix, with $\boldsymbol{\epsilon}$ our error term. This is discussed further in in the context of estimation and model selection in Section 2.2.

These unsaturated models are used across domains and are often chosen based on their fit against some hypothesis test or likelihood criteria, domain knowledge or ease of interpretation in a given application.

Identifying appropriate unsaturated models across this wide family of competing hypotheses remains a major challenge in academic work. Many authors have proposed stepwise procedures supported by domain expertise and Monte Carlo simulation to accurately identify the most appropriate model given generated data. In Floch and Le Saout (2018), authors review the progress and application of these procedures across the academic literature, shown in Section A of the appendix. These procedures look to identify and attribute the presence of particular spatial or social network effects in a systematic manner to identify the most appropriate model for some given data.

Floch and Le Saout (2018) break these procedures into two categories: top-down approaches, as suggested by Arnold (2011) and Mur and Angulo (2009), and bottom-up approaches, as detailed in Florax, Folmer and Rey (2003) and Elhorst (2010). Under the top-down procedures, complex SDM, SDEM or KP models are estimated with which to perform likelihood ratio and t-tests on parameter estimates to restrict the model to either SLX, spatial error or spatial lag or auto-regressive models. In bottom-up approaches, a constrained linear model is estimated under assumptions concerning the distribution and homoskedasticity of errors, on which spatial methods are applied to determine whether errors exhibit various forms of spatial or social network effects. Based on these tests, authors provide recommendations on the use of spatial error, spatial lag explanatory and spatial auto-correlation models, as well as the progression to more complicated saturated models.

While model identification remains critical, model estimation plays an equally important role in the success of a given study. Research into estimation is and has remained an area of important innovation in the field, owing to increasingly robust methods reliable across small and large datasets. Popular approaches across spatial techniques rely on the use of least-square, (quasi) maximum likelihood and generalised moments estimators, which have been explored in numerous Monte Carlo studies. For SAR models, Kelejian and Prucha (1999) argues the use of either (quasi) maximum likelihood and generalised moments estimators, chosen according to the computational resources available to the study, over least-squares estimation. Another motivation for generalized moments estimators has been through innovation by Kelejian and Prucha (1999), Arraiz et al. (2010) and Anselin (2011) to robustly control for the heteroskedastic innovations found in many spatial studies. This work has provided tooling for SEM and KP models to offer increasingly robust estimation and testing in cases where these models are most appropriate to the domain and data collected.

As the breadth of spatial modelling techniques has grown, researchers have come to increasingly rely on a number of common software implementations to aid in the consistency and reproducible of their work. While there exist minor differences in certain defaults based on competing findings in research, these differences are well documented with powerful abstractions to suit the requirements in particular studies. These implementations in Stata®, MATLAB®, R and Python have been benchmarked by researchers and have been shown to be consistent both across implementations and against original work (Bivand and Piras, 2015). Bivand and Piras (2015) do flag some exceptions, many of which have been corrected in subsequent software releases of these packages.

These methods provide rich and abstract tools through which to explore the effects of interaction in real-world systems, allowing researchers the ability to test and control for a broad range of effects which propagate through networks and geographies.

2.3 Spatial Arbitrage Pricing Theory

Inside of financial markets, research by Fernandez (2011) has looked to extend the CAPM to incorporate the effects of spatial interaction on market returns. This spatial interaction is explored under various models, but is derived by Kou et al. (2018) according to the mean-variance criteria of Sharpe (1963) and Markowitz (1952) under a Spatial Auto-regressive Model. In Figure 2, Kou et al. (2018) illustrate graphically the effects of spatial interaction on a market's efficient front. In this simulation, authors structure the asset return covariance matrix as some function of a spatial weighting matrix to vary the effects of spatial correlation on asset prices through ρ. In Fernandez (2011) and Kou et al. (2018), this matrix is defined against a variety of features and metrics but still abides by the common properties described in subSection 2.2.

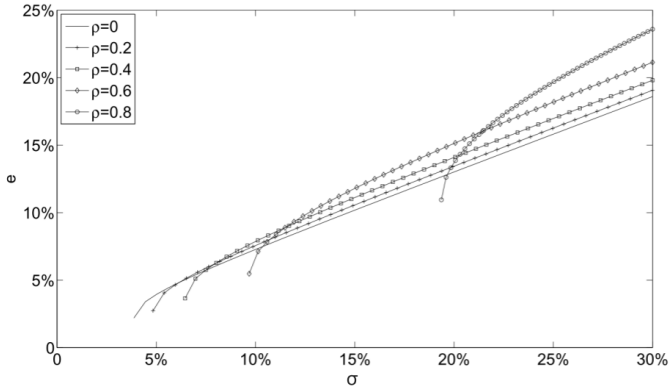


Figure 2: Diagram of simulated efficiency frontier under ρ varying spatial correlation as provided in Kou et al. (2018). In this graph, the standard deviation in returns of a mean-variance efficient portfolio is given on the x-axis with expected return given on the y-axis. As ρ increases, the efficient frontiers are seen to shift to the left and upwards, becoming increasingly convex to the origin. These contributions are further detailed in our discussions in Section 2.3 of this work.

In Kou et al. (2018), the ability to frame asset covariance and portfolio risk against some spatial weighting matrix has no impact on the ability to formulate a capital market line against some risk-free asset, but rather serves to rationalise a covariance structure observed in the data. This allows authors to trivially justify their spatial approach under CAPM, and by defining an investible spatial portfolio of assets specific to some region in space, to extend a spatial interpretation of APT. While this result is somewhat general, this result is constrained under CAPM, to symmetric and positive definite spatial weight matrices in order to satisfy the KTT conditions of the portfolio optimisation problem proposed in Markowitz (1952). In Buraschi and Porchia (2012) this two-fund separation constraint is intuited as a limitation in diversifying firm-specific risk in a directed star network governed by some central node. In such a graph, even as the number of firms or portfolio size grows very large in this scale-free network, investors would be unable to diversify the firm-specific risks of this central node by taking stakes in other nodes along the graph.

$$\min_{\boldsymbol{w},\Sigma} \quad \frac{1}{2}\boldsymbol{w}^T\Sigma\boldsymbol{w}; \qquad \text{s.t.} \quad \boldsymbol{e}^T\boldsymbol{w} = 1; \boldsymbol{m}^T\boldsymbol{w} \geq \boldsymbol{\mu_b} \tag{3.1}$$

Exhibit 3: Portfolio optimization problem proposed by Markowitz (1952) in which w represents a vector of portfolio weights, Σ our asset variance-covariance matrix, m is the vector represented the expected return on our universe of investible assets, e is a vector representing our budget constraints and μ_b our required rate of return on our investment which is greater or equal to the risk-free rate in the market.

Using this s-APT framework, Kou et al. (2018) explore the impact of spatial auto-correlation on US SP/Case-Shiller Home Price indices and Eurozone stock indices, when controlling for factors such as size, momentum and credit-risk. In this work authors found, using a symmetric spatial weight matrix of driving distances, evidence for spatial correlation in both these markets; which we may posit as a function of shared-geographies, social influence or direct economic interaction across countries or regions.

While Kou et al. (2018) critique the theoretical justification of their approach, Fernandez (2011) offer a fascinating contribution to the literature on Arbitrage Pricing Theory and spatial econometrics on their choice of spatial weight matrix which differentiates their research from common and later applications of S-CAPM. Fernandez (2011) argue that, in the context global financial markets, the vast global footprint of listed brands renders geographic spatial weight matrices inappropriate due to the level of diversification which companies have across country-specific latent-risk factors. While this argument may vary greatly depending on one's investible universe, Fernandez (2011) look to the vast ocean of research into factor and characteristics-based models to describe a metric for spatial modelling across firm-specific characteristics (Banz, 1981; Fama and French, 1992). Using the Spearman correlation across company market capitalisation, market-to-book,

EV/EBITDA and debt ratios, Fernandez (2011) discover strong evidence for spatial correlation across characteristics, supporting similar findings under APT for these characteristics.

While these results from Fernandez (2011) are comparable to many APT studies, the generality of spatial regression methods provide limited guidance to investors concerning the impact of latent risk factors. This generality is useful in cases in which factors bear non-linear and interacting effects on market returns, but can describe under a given estimate either a linear pricing rule for a subset of factors or non-linearly separable clusters of correlated returns. By using factors to compute a spatial weighting matrix, not only do we forego any definitive statement around these factors, but also pose limitations in our ability to explore social exogenous-type effects which provide actionable investment opportunities.

$$(r_{it} - r_{rf,t}) = \alpha + \beta(r_{mt} - r_{rf,t}) + \rho \sum_{j=1}^{N} w_{i,j}(r_{jt} - r_{rf,t}) + \epsilon_i; \sum_{j=1}^{N} w_{i,j} = 1 \tag{4.1}$$

Exhibit 4: Spatial (Lag) Capital Asset Pricing Model as formulated by Fernandez (2011). In this equation, r_{it} represents the returns of asset, i, and time, t, r_{mt} the return of the market, m, at time, t, and $r_{rf,t}$, the risk-free rate at time, t. $w_{i,j}$ is used to present the spatial weighting between asset i and spatially neighbouring asset, j and ϵ_i, our error term. This equation is further introduced to readers in Section 2.3 where it is compared against other spatial modelling approaches.

$$(r_{it} - r_{rf,t}) = \alpha + \beta(r_{mt} - r_{rf,t}) + u_i; \tag{5.1}$$

$$u_i = \lambda \sum_{j=1}^{N} w_{i,j} u_i + \epsilon_i; \tag{5.2}$$

$$\sum_{j=1}^{N} w_{i,j} = 1 \tag{5.3}$$

Exhibit 5: Spatial (Error) Capital Asset Pricing Model as formulated by Fernandez (2011). In this equation, r_{it} represents the returns of asset, i, and time, t, r_{mt} the return of the market, m, at time, t, and $r_{rf,t}$, the risk-free rate at time, t. $w_{i,j}$ is used to present the spatial weighting between asset i and some spatially neighbouring asset, j, and ϵ_i, our error term. This use of the spatial error model in factor modelling is further discussed in Section 2.3.

While Fernandez (2011) and Kou et al. (2018) explore only SAR and SEM s-APT and s-CAPM interpretations, shown in equations 4 and 5, these models offer vastly different insights to investors on market behaviour. Under SAR and SLX models, s-APT provides insight into the investment of spatially diversified characteristic portfolios, defining obvious strategies for investors to diversify their portfolios across firm-characteristics and some region of space. Under SEM models, researchers look to improve non-spatial estimates by correcting for spatial correlations in market excess return. While there is obvious overlap in these approaches, in SDEM, SDM and KP models, interpretation remains key in justifying these methods inside financial economics as studies extend their applications.

These methods provide researchers with rich tools through which to assess the impact of firms' interactions in propagating risk-return within markets. These interactions can be conceived based on geography or firm characteristics and provide a valuable abstraction through which to understand shared risk.

2.4 Graph Theory

Across the many disciplines of economics, sociology, chemistry, computer science and epidemiology, the study of networks has offered a rich perspective on the evolution, structure and interaction of various systems and processes (Newman, 2010). In mathematics, and in particular graph theory, a graph or network is seen as a mathematical object comprised of nodes (also referred to as vertices or points) and edges which connect them. In various domains, these nodes may represent collections of participants in a study, pixels in an image, atoms in a molecule or servers on a network; while the edges of a graph may represent the interaction between participants, adjacency between pixels, bonds between atoms or network connections between servers. Graphs can be represented in many different ways, either graphically, as in Exhibit 2, more

formally as a series of vertices and edges, as in the equation in Exhibit 6, or as an adjacency matrix, shown in the equation in Exhibit 7.

$$G = (V, E); \tag{6.1}$$
$$V = 1, 2, 3, 4, 5, 6, 7, 8; \tag{6.2}$$
$$E = 1, 2, 1, 3, 1, 4, 3, 5, 3, 7, 4, 7, 4, 6, 5, 8, 7, 8, 6, 8 \tag{6.3}$$

Exhibit 6: Graph, G, comprising vertices, V, and edges, E, representing the graph in Exhibit 2 given in Elhorst (2010). In these equations, each number represents a unique node in our graph or edge between nodes.

$$W = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \tag{7.1}$$

Exhibit 7: Adjacency matrix representing the graph in Exhibit 2 where in each entry in the matrix represents the weight given to each edge in the graph. Weights of zero, suggest no edge or relationship is present between nodes. This equation along with the equation in Exhibit 6 are further developed in Section 2.4 on the representation of graphs.

These graphs can bear many attributes over their nodes and edges, representing the type of interaction or the distances between nodes. Graphs may also form directed or undirected graphs in the case of a one-way road or asymmetric travel times in which it may take longer to get to a destination that return from it. In a directed graph, nodes are connected by edges which have a direction associated with them. These directed graphs are represented mathematically by adjacency matrices which are asymmetric and graphically using edges with pointed arrows. Undirected graphs are typically represented with edges without any arrow and adjacency matrices which are symmetric.

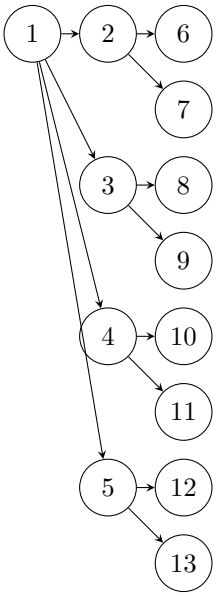


Figure 3: Diagram of scale-free network whose degree distribution follows a power law. In this diagram, numbers indicate unique nodes the graph.

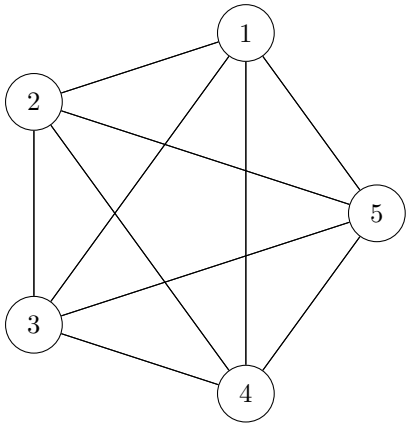


Figure 4: Diagram of complete network whose nodes are all connected. In this diagram, numbers indicate unique nodes the graph.

In the study of graphs, researchers look to a number of graph types or graph generating functions, described through a variety of properties. Complete graphs are graphs in which all nodes are connected via edges, shown in Figure 3; scale-free graphs are described according to degree distribution, shown in Figure 4. These graphs can be constructed or defined according to graph generating functions, which may be deterministic, as in the case of a chordal graph, or random, wherein edges are drawn from a probability distribution (Erdős and Rényi, 1960; Lubotzky, 1994). Graphs can also represent probabilistic relationships as in the case of fuzzy or stochastic graphs in which the certainty of particular edges is captured by a probability or unique probability distribution.

Across modelling applications, graphs can often be used to describe and explore the interaction between variables or form new features on which to explore the particular effects of graph structure. In statistical sciences, graphs can be used to describe a range of graphical models from Markov random fields and Factor Graphs, to Bayesian Networks. In Deep Learning, graphs can form or formalise both the data and learning apparatus itself (Hamilton, Ying and Leskovec, 2017).

In Dong, Thanou, Rabbat and Frossard (2019), authors provide an overview and history of approaches in graph or structure learning. Through these techniques, researchers look to use mathematical models and optimisation techniques to uncover a network based on observations of data generated from interactions across a graph. These approaches can rely on Vector Auto-correlation Models (VARMs), to exploit regular times at which effects propagate along edges to uncover a directed graph, or methods in Covariance Selection and Graph Signal Processing (GSP) to try to uncover this structure from some data.

Across applications, researchers look to uncover and describe particular properties of graph structures. These efforts may look to classify graphs into known types to relate these graphs to known properties, or describe attributes of nodes based on their position and connections along the graph (Giannakis, Shen and Karanikolas, 2018). In graph theory, degree is used to describe the number of edges of a given node. Using the distribution of this property across a graph, researchers can identify both the density of the graph and whether a graph mirrors known graph generating functions, such as those described by scale-free networks or Erdős–Rényi random graphs.

An important area of interest for researchers in the Social Sciences remains clique or community detection. While cliques detection follows a formal definition in identifying complete subgraphs in the network, community detection describes of a wide range of algorithms through which researchers identify either overlapping or non-overlapping subgraphs whose nodes form some dense region in the graph or who share similar properties. These techniques share many parallels in unsupervised clustering applications where-in researchers look to identify dense regions of customers, assets or pixels in some vector space in order to perform customer segmentation, portfolio optimisation or identify types of tissues in medical images (Javed, Younis, Latif, Qadir and Baig, 2018).

Among the various properties of graph nodes, centrality remains an important measure with which to explore and identify relevant web-pages, super-spreaders of disease or the key infrastructure across computer networks. Across applications, authors have various abstractions and approaches to centrality which fall into various classes of betweenness-like, closeness-like and degree-like measures, which rely upon either reachability, path length between nodes or degree of a particular node and closest neighbours in order to determine characteristics of that node in the structure of the graph (Vigna, 2016; Borgatti and Everett, 2006).

$$L = D - A \tag{8.1}$$

$$L = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 \end{bmatrix} - \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 3 & 0 & 0 & -1 & 0 & -1 & 0 \\ -1 & 0 & 3 & 0 & -1 & -1 & 0 & 0 \\ -1 & 0 & 0 & 3 & 0 & -1 & -1 & 0 \\ 0 & -1 & -1 & 0 & 3 & 0 & 0 & -1 \\ 0 & -1 & 0 & -1 & 0 & 3 & 0 & -1 \\ 0 & 0 & -1 & -1 & 0 & 0 & 3 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & -1 & 3 \end{bmatrix} \tag{8.2}$$

$$L = \chi \Lambda \chi^T \tag{8.3}$$

$$\mathcal{F}(G) = \chi^T f(G) \tag{8.4}$$

Exhibit 8: Graph Fourier Components, F(G), computed using function representing measurements over our vertices, f(G), and the eigenvectors, χ and ordered eigenvalue matrix, Λ, decomposed from the Graph Laplacian, L, of the undirected graph shown in Exhibit 2, using Adjacency Matrix, A, diagonal Degree Matrix, D, with entries representing the number of direct neighbours of each node.

While many of these measures and methods discussed describe graph structure, an important discussion lies in how to explore measurements which are observed on that graph structure and emanate as some function of the structure itself. To explore such a property Graph Signal Processing has extended many approaches in traditional signal processing, such as Fourier Transforms and Wavelets, into a broad set of methods through which to explore how signals vary across a graph (Shuman, Narang, Frossard, Ortega and Vandergheynst, 2013; Ricaud, Borgnat, Tremblay, Gonçalves and Vandergheynst, 2019). In the case of the Graph Fourier Transform, this is computed by computing the eigendecomposition of the graph Laplacian and vertex measurements, shown in the equation in Exhibit 8. While these methods often allow us to decompose oscillations across the graph, they also allow for methods through which to filter and smooth noisy measurements taken across a graph (Ramakrishna, Wai and Scaglione, 2020). This has formed an important basis for Graph Convolution, Graph Coarsening and Graph Neural Networks, which have been used extensively across domains of computer vision, drug-discovery and town-planning (Hamilton et al., 2017).

While the use of GSP has seen use through graph learning in research into financial markets (Cardoso and Palomar, 2020; Ramezani-Mayiami and Skretting, 2019), it has also been used to explore asset allocation decisions using community detection to identify optimal cuts in the graph over which to allocate resources (Dees, Stanković, Constantinides and Mandic, 2020). This work furthers a graph-theoretic capital allocation scheme based on measures of connectivity, which builds on work by De Prado (2016) on Hierarchical Risk Parity which motivates asset allocation based on methods in hierarchical clustering. While these methods use graph theory in order to motivate new approaches to common problems in empirical finance; many papers look to explore and imagine observations in financial markets as some function of a graph with particular characteristics (Billio et al., 2016; Ahern, 2013). These many techniques in graph theory and computation provide useful approaches through which to understand risk and how it is shared across parts of the market.

2.5 Network Models in Economics

Across the literature on networks in economics, many authors have explored the important interaction between market structure, economic rent and network topology (Jackson and Wolinsky, 1996; Jackson and Zenou, 2015; Bloch, 2016). In Chen et al. (2018), building on the work of Candogan, Bimpikis and Ozdaglar (2012), Bloch and Quérou (2013) and Fainmesser and Galeotti (2016), authors consider optimal firm strategies in markets of substitute goods in which an oligopoly of firms are able to set prices based on a consumer's position along a graph. In such a market, in which goods accrue externalities based on neighbour adoption, Chen et al. (2018) find an optimal strategy for firms under which

firms choose to price discriminate based on Katz-Bonacich Centrality, shown in the equation in Exhibit 9. Given varying centrality, firms are able to charge higher prices to more consumers by subsidising only a small subset of central consumers due to the efficient positive externalities they accrue to other consumers. Under such conditions, Chen et al. (2018) demonstrated the negative impact of network density on economic rents as changing network structure limits the efficacy of this price discrimination strategy.

$$b(G, \alpha, \beta) = \alpha(I - \beta G)^{-1}\tilde{\boldsymbol{e}} \tag{9.1}$$

Exhibit 9: Formulation of Katz-Bonacich Centrality under Spectral Ranking as discussed in Vigna (2016) in which $G \in_{+}^{N \times N}$ represents the adjacency matrix of our graph, $\beta \in [0, 1]$ is some discount factor, α is some constant and $\tilde{e}$ is the vector $[1, ..., 1] \in R_{+}^{N}$ which may represent a vector of initial weights (Katz, 1953; Bonacich, 1987).

Building on this work, Chen, Zenou, Zhou et al. (2020) consider a two-stage oligopolistic network competition model in which, given a random graph, firms consider consumption along the graph following a simultaneous pricing decision. From this work, authors show the negative impact of network size and density on market prices. By formalising players ability for substitution, Chen et al. (2020) demonstrate how network structure preferences of firms change depending on size and the homogeneity of the market. This echos findings in Roson and van den Bergh (2000), in which authors demonstrate total welfare maximisation in markets described by star networks under firm monopolies. These findings do not hold true for complete graphs, where Roson and van den Bergh (2000) demonstrate contradictory findings which suggest that total welfare is maximised under a market dominated by a duopoly. Chen et al. (2020) show the impact of Katz-Bonacich Centrality in influencing prices along the network, demonstrating the impact of product homogeneity and competition on firm profits as we vary network size and density.

A finding from Corbo et al. (2006) serves to justify these rents accrued to players' central to the market's network, as authors demonstrate undr markets with unique Nash equilibrium (NE), players aggregate contribution and total social welfare can be expressed through the use of the Bonacich index vector of the graph. In their peer-to-peer network model, Corbo et al. (2006) demonstrate both the maximisation of total social welfare under star networks and, under simulation, the impact of graph density in maximising total social welfare across the graph. This suggests that firms not only gain market dominance from their centrality, but also under particular network configuration this structure may play a crucial role in maximising the total social welfare for all players across the graph as, under different configurations, goods and services may flow more or less freely.

While the impact of central nodes on social welfare remains clear, the role of centrality in the contagion of risk or product adoption in markets characterised by heterogeneous goods may remain unclear. Within the Social Sciences, work by Watts and Dodds (2007) has explored extensively, under simulation, the impact of influential individuals in the process of public opinion formation. In this work, authors define a model under which a random population of individuals hold and exert influence on others in order to explore how opinions change and propagate. In this population, certain influential individuals are endowed with a greater number of connections and propensity to influence others. While authors find limited evidence for the unique importance of central individuals, Watts and Dodds (2007) note the important impact which small numbers of highly susceptible individuals play in propagating ideas. Using an analogy of wildfires, simulations by Watts and Dodds (2007) suggest the important roles of kindling in fire propagation over the location of some initial spark. In the context of financial markets, this would suggest that while firm centrality may have an important impact on market performance, it is in-fact a small number of distressed companies, which given some initial trigger, serve to propagate contagion across the market. This insight on susceptibility suggests the critical role which exogenous network effects play in node responses, as firms connections to distressed companies may play a critical role on their risk and return.

$$
\begin{aligned}
y &= \rho W y + X\beta + \epsilon \\
(I - \rho W)y &= X\beta + \epsilon \\
y &= (I - \rho W)^{-1}(X\beta + \epsilon) \\
y &= (I + \rho W + \rho^2 W^2 + \rho^3 W^3 + ...)(X\beta + \epsilon) \\
y &= (I + \rho W + \rho^2 W^2 + \rho^3 W^3 + ...)X\beta + (I + \rho W + \rho^2 W^2 + \rho^3 W^3 + ...)\epsilon \\
y &= (X\beta + \rho W X\beta + \rho^2 W^2 X\beta + \rho^3 W^3 X\beta + ...) + (\epsilon + \rho W\epsilon + \rho^2 W^2\epsilon + \rho^3 W^3\epsilon + ...)
\end{aligned} \tag{10.1}
$$

Exhibit 10: Centrality interpretation of SAR offered in LeSage and Pace (2009) using expansion given by Debreu and Herstein (1953). In this expansion, y, is our endogenous variable, X a matrix representing our exogenous variables, ϵ, our error vector, ρ, a spatial weighting term and, W, our spatial weighting matrix. Through this expansion authors demonstrate the interpretation of a spatial lag model as a Spatial Durbin Error Model at infinite lags, and its parallels to Katz-Bonacich Centrality in the equation in Exhibit 9, through our typical linear regression formula, $(X\beta + \epsilon)$. This is discussed in Section 2.2 to describe spatial models as some linear model which undergoes a linear transformation based on centrality to express spatial or social network interaction.

These joint findings in Chen et al. (2020), Chen et al. (2018) and Garcia Alvarado and Mandel (2019) concerning the impact of centrality on market position provide an insight into our observation of social network econometric models. In work by LeSage and Pace (2009) and Liu and Lee (2010), the authors point out the obvious centrality interpretation of spatial models, shown in equation 10. Under such an interpretation, LeSage and Pace (2009) suggest Katz-Bonacich Centrality, shown in the equation in Exhibit 9, be used as a mechanism through which to interpret the effects of spatial interactions as some function of centrality through which exogenous factors are exaggerated.

Under our centrality interpretation of spatial models, shown in equations 10, Bonacich (1987) provide fascinating insight with which to explain different spatial coefficients in a network based on the equation in Exhibit 9. In their work, authors use an analogy of power with which to frame centrality. A common interpretation of positive spatial coefficients, provided through this work, offers a probabilistic definition of spatial coefficients. Under this interpretation of spatial coefficients, weights bounded between 0 and 1, are intuited as the probability with which an originator influences its neighbours. Bonacich (1987) suggest that for decision-makers, magnitudes of our spatial coefficients can be perceived as some radii around which effects propagate at some confidence. While the probabilistic interpretation of centrality remains intuitive for positive spatial coefficients, Bonacich (1987) explore scenarios in which negative spatial coefficients best capture power in a network. To answer this question, Bonacich (1987) look to bargaining networks wherein employers best exert power over isolated employees with limited options. In such scenarios, negative spatial coefficients indicate to domain experts and researchers very different insights into the power of some agent based on their position in some graph. In financial markets, this may provide insight into specific market dynamics expressed through network structure. For positive spatial weights, a particular network structure may reward firms connected to other highly connected nodes, as local density allows goods and services to flow more efficiently across the graph; for negative spatial weights, similar network structures may reward firms connected to isolated firms as companies find themselves in more a dominant market-making position.

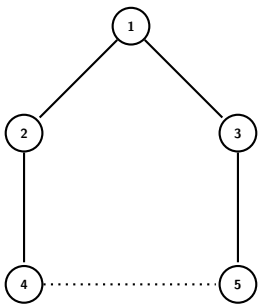


Figure 5: Illustrative graph from from Cook, Emerson, Gillmore and Yamagishi (1983). Here, the dotted line represents an edge of weaker value or lower weight. In this diagram nodes are numbered for the purposes of unique reference.

$$G = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0.5 \\ 0 & 0 & 1 & 0.5 & 0 \end{bmatrix} \tag{11.1}$$

Exhibit 11: Adjacency matrix from Cook et al. (1983) representing the adjacency matrix from Figure 5. In this matrix, the weak connection between nodes 4 and 5 are given a weight of 0.5, half that of other edge in the graph. This matrix is used below, in Table 1 to illustrate changing definitions of centrality.

In Table 1, we see under varying β coefficients, centrality shifts from node 1 or nodes 2 and 3, shown in Figure 5. This shift takes place under decreasing β as nodes 2 and 3 find themselves connected to weakly connected nodes 4 and 5. Under high values of β, node 1 finds itself most central due to its strong edges to highly connected nodes 2 and 3. Using least square, maximum likelihood or method of moments estimators, we may estimate a value for β or ρ, which provides insight into the market dynamics taking place over the graph. These dynamics may be local and depend on the markets which take place in particular cliques or across subgraphs, or they may be global, suggesting the returns or risk associated with particular estimates of graph centrality.

	Node				
Beta	**1**	**2**	3	4	5
1	**0.62**	0.496	0.496	0.248	0.248
$\frac{2}{3}$	**0.548**	0.487	0.487	0.335	0.335
$\frac{1}{3}$	**0.489**	0.470	0.470	0.398	0.398
0	**0.447**	**0.447**	**0.447**	**0.447**	**0.447**
$-\frac{1}{3}$	0.445	0.390	0.390	**0.490**	**0.490**
$-\frac{2}{3}$	2.340	**0.632**	**0.632**	0.316	0.316
-1	0.242	**0.485**	**0.485**	**0.485**	**0.485**

Table 1: Table illustrating the impact of β on Katz-Bonacich Centralities, shown in Exhibit 9, computed over Figure 5 and the adjacency matrix shown in the equation in Exhibit . In this table, α is kept constant at 1. Our most central nodes have been displated in bold to aid in comparison.

Looking to our study of financial markets, this work in economics exploring the impact of graph structure on markets may be critical in understanding the process through which firms generate future discounted cash-flows as a function of their position in and the structure of the market graph. This may allow us to better understand how transactions and relationships between firms should be priced by investors and contribute to changing systematic and firm-specific risk. Looking at the results of Chen et al. (2018), Bonacich (1987) and Corbo, Calvó-Armengol and Parkes (2006) investors may look to base an investment hypothesis on firm centrality, using it as a measure through which to to guide portfolio allocations decisions based on how the underlying market graph evolves through time. As firms engage in new strategic partnerships with which to position their firms as central to the market, investors may use this information as a signal of emerging market dominance with which to reevaluate their forecasts of future discounted cash-flows to account for this new market position.

2.6 Network Models in Finance

Research into network modelling has received obvious attention in empirical finance. Inside this body of literature, researchers have explored many approaches in graph learning and simulation to explore the effects of direct contagion in financial markets, in which one firm's default on its contractual obligations triggers distress in neighbouring counter-parties. While many efforts exist, a major challenge in this work lies in identifying or constructing a graph from data on which to perform various modelling techniques. In response to this challenge, work by Barigozzi and Brownlees (2019) and Diebold and Yılmaz (2014) investigates learning-based approaches to try to uncover sparse evolving weighted directed graphs from stock market data. These graphs aim to capture changing counter-party risk across their samples of banking and blue-chip stocks as priced by the market. These approaches rely heavily on methods in Vector Auto-correlation (VAR), shown in the equation in Exhibit 12, using either penalisation or decomposition to manage the robustness or sparsity of their parameter estimates validated under simulation.

$$y_{i,t} = \sum_{k=1}^{p} \sum_{j=1}^{n} a_{i,j,k} y_{i,t-k} + \epsilon_{i,t} \tag{12.1}$$

Exhibit 12: Vector Auto-correlation Model (VAR) as formulated in Barigozzi and Brownlees (2019), in which $y_{i,t}$ represents the returns of stock i at time t against j stocks at k lags under the assumption of normality in $\epsilon_{i,t}$. This is discussed in Section 2.6 as an approach to graph learning.

An interesting approach taken by Procházková (2020) to incorporate graph learning in traditional APT factor models, uses model weights to define a measure of connectedness through which to estimate the impacts of network exposure. Under this approach, authors posit that nodes with some latent connection should exhibit strong vector auto-correlation; relying on the decomposition method developed by Diebold and Yilmaz (2012) to estimate their VAR models. Unlike many spatial econometric methods, which explicitly row-normalise the spatial weight matrix to exclude impacts of centrality or connectedness, Procházková (2020) defines six different types of company connectedness based on both return and volatility vector autocorrelation and the presence and direction of node edges. These edges are then used to compute either `TO`, `FROM` or overall connectedness according to the number of non-zero edges either leading to a particular node or stemming from a particular node. In this work, Procházková (2020) identify strong positive effects from overall and `FROM` connectedness, with mixed findings on the impact of `TO` connectedness across sectors.

A major challenge faced by Procházková (2020) remains in how best to interpret these findings, given the lack of a widely accepted definition, measure or interpretation of connectedness (Diebold and Yılmaz, 2014). While VAR methods may accurately identify effects propagated at some lag across a graph, these edges may represent a variety of supplier, customer or investor relationships. Given this ambiguity, it may be challenging for investors to easily formulate valid and cohesive market hypotheses on which investors easily rationalise these risks as connectedness may point to a variety of phenomenon sensitive to market conditions.

In modern corporate finance, companies look to manage risk by diversifying investments, customers and suppliers across risk factors in order to limit the effects of latent risk factors and direct contagion on firm liquidity. While highly connected companies may diversify many firm-specific risk factors through their supply-chain decisions; market and supply chain complexity in a given industry may deeply impact whether companies can diversify the impacts of direct contagion through connectedness.

Although regionally diverse, supply chains in the automotive industry find themselves highly sensitive to single supplier insolvency as safety regulations often limit or delay a manufacturers ability to substitute parts when a supplier goes insolvent. This lies in stark contrast to fast-moving consumer goods (FMCG) in which products are easily substituted, and companies share and compete across a number of crowded product categories. In Procházková (2020), researchers identify consistent negative coefficients on return and volatility `TO` connectedness in the Health sector. These findings lie in contrast to findings in the energy sector, in which positive coefficients on return `TO` connectedness were identified. These differences may relate to supply chain and market complexity, in which single drug, drug-precursor or equipment suppliers or insurers can present major and identifiable liquidity risk. Under such conditions, high levels of connectedness may expose firms across many points of failure, rather than diversifying non-systematic risk factors. Alternatively, these findings may also suggest that highly connected companies in the healthcare sector present less value to investors in diversifying particular risk across the business cycle, compared to highly connected companies in the energy sector. These ambiguities concerning network connectedness, add challenge to investors looking to use this research to construct some consistent investment hypothesis with which to make capital allocation decisions.

A major criticism we may raise in Procházková (2020) lies in the impacts of omitted variable bias on the reliability of one's estimate. By using VAR to learn the weighted graph on which to compute our connectedness factors, authors risk under-estimating the impact of other risk factors in cases in which companies select firms with similar size or value, as shown in the equation in Exhibit 13. While Procházková (2020) may identify significant impacts of connectedness in their model, authors fail to address the possibility of obvious correlations between some Factor Mimicking Portfolio (FMP) such as Small-minus-Big (SMB) and company connectedness, which may explain the insignificance of SMB across their estimates.

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y \tag{13.1}$$
$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'(X\beta + Z\boldsymbol{\delta} + U) \tag{13.2}$$
$$= (X'X)^{-1}X'X\boldsymbol{\beta} + (X'X)^{-1}X'Z\boldsymbol{\delta} + (X'X)^{-1}X'U \tag{13.3}$$
$$= \boldsymbol{\beta} + (X'X)^{-1}X'Z\boldsymbol{\delta} + (X'X)^{-1}X'U \tag{13.4}$$
$$E[\hat{\boldsymbol{\beta}}|X] = \boldsymbol{\beta} + (X'X)^{-1}\,E[X'Z|X]\boldsymbol{\delta} \tag{13.5}$$
$$= \boldsymbol{\beta} + \text{bias} \tag{13.6}$$

Exhibit 13: Here, in equations 12.1, 12.2 and 12.3 we show the effects of omitted variable bias where y represents our response variable, X our design matrix and Z our omitted variable with effect δ. $\hat{\boldsymbol{\beta}}$ represents the estimate of our regression coefficients and $\boldsymbol{\beta}$, their true value. We can see based on equation 12.3 that the omission of Z causes bias in our estimates of β, based on $E[X'Z|X]\boldsymbol{\delta}$.

While these VAR approaches have provided much progress in exploring network effects in markets, these methods make a number of assumptions which lie in stark contrast to a number of dominant theories on market behaviour. Unlike under common structure learning approaches, given by Zheng, Aragam, Ravikumar and Xing (2018) and Aragam and Zhou (2015), VAR methods assume some lag at which network effects propagate as first passage effects across the graph. These first passage effects reflect the initial price impact which a firm may induce among connected firms, as opposed to some secondary effects which may reverberate following contagion. While the presence of serial correlation in markets is well understood, the time over which effects are observable my vary, distorting the estimates of our graph (Carhart, 1997; Asness, 1997). Additionally, depending on our assumptions around market efficiency, VAR approaches may form a different looking graph, either censoring the impacts of private transactions or presenting these effects over different time horizons.

Beyond VAR methods, contributions by Billio et al. (2016) have extended work on network effects in theoretical work and under simulation using the s-CAPM and s-APT frameworks proposed by Fernandez (2011) and Kou et al. (2018), shown in the equation in Exhibit 14. Unlike in graph learning approaches in which researchers look to uncover a graph from some lag structure in returns, Billio et al. (2016) assumes some graph determined exogenously to the market in order to explore contagion-type risks stemming from firm interconnectedness.

$$\boldsymbol{r}_t = E[\boldsymbol{r}_t] + \underbrace{F_t\widetilde{\boldsymbol{\beta_i}}}_{\text{Structural Exposure}} + \underbrace{\sum_{k=1}^{\infty}\rho^k W^k F_t\widetilde{\boldsymbol{\beta_i}}}_{\text{Network Exposure}} + \underbrace{\eta_t}_{\text{Idiosyncratic Shocks}} + \underbrace{\sum_{k=1}^{\infty}\rho^k W^k \eta_t}_{\text{Idiosyncratic Network Shocks}} \tag{14.1}$$

$$E[\boldsymbol{r_t}] = r_f + \underbrace{F_t\widetilde{\boldsymbol{\beta}}}_{\text{Structural Exposure}} + \underbrace{\sum_{k=1}^{\infty}\rho^k W^k F_t\widetilde{\boldsymbol{\beta_i}}}_{\text{Network Exposure}} \tag{14.2}$$

$$V[r_{p,t}] = \underbrace{\boldsymbol{\omega'}\boldsymbol{\beta}\Sigma_F\widetilde{\boldsymbol{\beta}'}\boldsymbol{\omega}}_{\text{Exogenous systemic effect}} + \underbrace{(\boldsymbol{\omega'}A\widetilde{\boldsymbol{\beta}}\Sigma_F\widetilde{\boldsymbol{\beta}'}A\boldsymbol{\omega} - \boldsymbol{\omega'}\widetilde{\boldsymbol{\beta}}\Sigma_F\widetilde{\boldsymbol{\beta}'}\boldsymbol{\omega})}_{\text{Endogenous systemic effect}} + \underbrace{\boldsymbol{\omega'}\Omega_\eta\boldsymbol{\omega}}_{\text{Structural idiosyncratic component}} + \underbrace{(\boldsymbol{\omega}'A\Omega_\eta A\boldsymbol{\omega} - \boldsymbol{\omega}'\Omega_\eta\boldsymbol{\omega})}_{\text{Endogenous idiosyncratic component}} \tag{14.3}$$

Exhibit 14: Decomposition of s-APT simultaneous auto-correlation model as formulated under simulation in Billio et al. (2016), in which $r_{i,t}$ represents the returns of stock i at time t, F_t some latent risk factor at time t, with $E[R_{i,t}]$ and $V[\boldsymbol{r_t}]$ our expected return and return variance of a particular portfolio. In these formulations $\eta_{i,t}$ a normally distributed error-term, p, with Ω representing the covariance of our idiosyncratic shocks and ω some portfolio vector, with $r_{p,t}$ representing the returns on a particular portfolio, p, at time, t. Under these formulations, W, given in equations 13.1 and 13.2, is a N by N matrix is used to represent edges between asset i and neighbour, j, where N is the size of our investable universe under investigation through out model. These equations are discussed in the context of further experiments by Billio et al. (2016) in Section 2.6.

Billio et al. (2016) use this framework of simultaneous auto-correlation to decompose risks stemming from Structural Exposure, Network Exposure, Idiosyncratic Shocks and the Network impact of Idiosyncratic Shocks. Under simulation, Billio et al. (2016) shows the growing impact of network-related risks relative to systematic risks as we grow our portfolio sampled from equally-weighted stocks inside our investible universe, shown in Figure 6. Billio et al. (2016) explore these results under a variety of distributional assumptions, varying their spatial effects, ρ, factor risks, β, and the density of their Erdős–Rényi graphs sample under a Bernoulli distribution. From this search, Billio et al. (2016) find consistent results concerning the growing relative risk stemming from network-related effects.

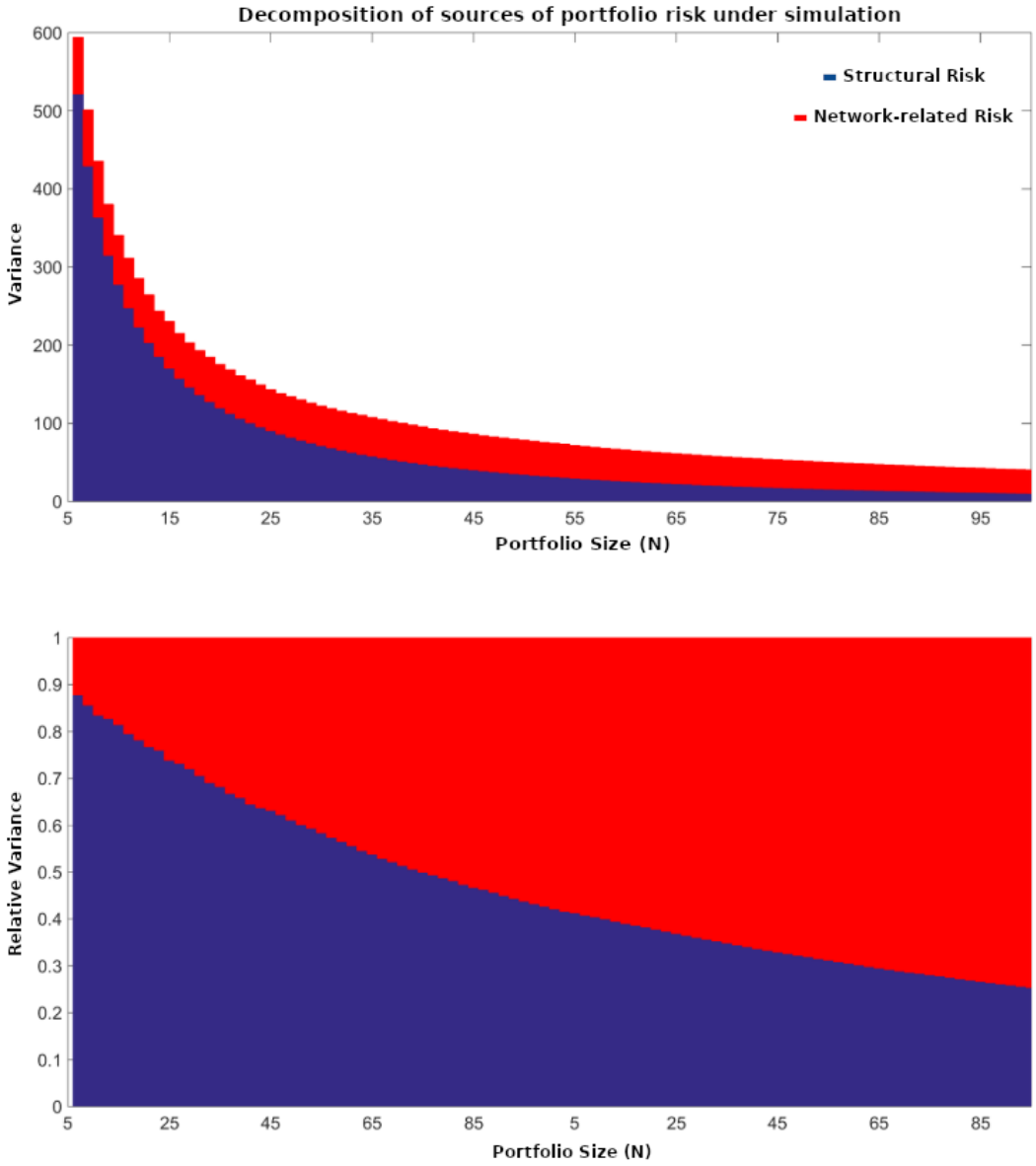


Figure 6: Diagram from Billio et al. (2016) illustrating the decomposition of simulated portfolio risk, as in shown in the equation in Exhibit 14, across different portfolio sizes with structural risk shown in blue and network-related risk shown in red. An absolute decomposition of the portfolio risks are shown in the upper facet, a relative decomposition of portfolio risks shown in the lower facet. Across both diagrams, as the portfolio sizes increase along the x-axis, the volatility of the portfolios decreases and become increasingly impacted by network-related risks as a portion of total risk.

A fascinating discovery in Billio et al. (2016) relates to the impact of spatial correlation in distorting estimates of market excess return under model misspecification, shown in Figure 7. To capture this model misspecification, Billio et al. (2016) simulate market returns generated with idiosyncratic network shocks and network exposure. Billio et al. (2016) then estimate a traditional factor model using ordinary least-squares regression and later generalised least-squares regression under the assumption that $\rho = 0$. In this simulation, Billio et al. (2016) shows that without controlling for this spatial correlation, models identify evidence for positive market excess returns incorrectly. This suggests that in the presence of spatial correlation, traditional approaches in portfolio performance evaluation may underestimate the risks associated with a given market position attributed to endogenous systematic and idiosyncratic effects. These impacts may arise when investors take positions in a diversified portfolio of stocks with dense inter-dependencies across the graph; as may be the case with investments across a diversified group of listed companies. Under such position, financial distress across a single company in the group may force a parent to sell their interest in other companies in order to meet certain liquidity requirements.

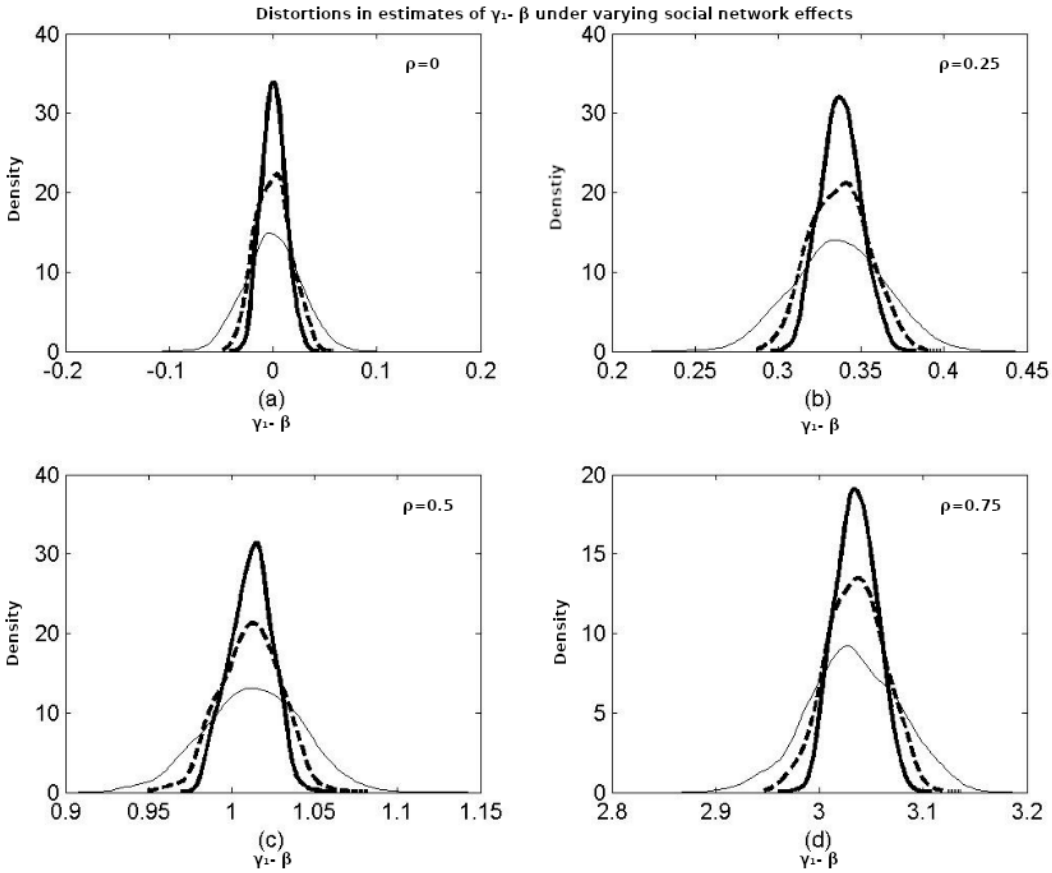


Figure 7: Results exploring the distortion in estimates of the true market excess return under simulation of increasing spatial interaction across a graph as given in Billio et al. (2016), with (a) $\rho = 0$, (b) $\rho = 0.25$, (c) $\rho = 0.5$ and (d) $\rho = 0.75$, and sample sizes of $T = 200$ in thin grey line, $T = 500$ as a dashed line and $T = 1000$ as a thick black line. Across the facets, spatial correlation is shown to result in increasing distortion in the estimate of beta, shown along the x-axis.

In Arnold, Stahlberg and Wied (2013), authors explore these findings in Billio et al. (2016) constructing a spatial weight matrix based on market capitalisation, sector adjacency and country to explore Value-at-Risk (VaR) estimates in securities markets. In this study, Arnold et al. (2013) find spatial models to provide a far more stable and faithful estimate of VaR compared to covariance and factor modelling approaches.

$$(I - HW)(\boldsymbol{r_t} - E[\boldsymbol{r_t}]) = F_t\bar{\beta} + \boldsymbol{\eta_t}; \tag{15.1}$$

$$H = \begin{bmatrix} \rho_1 & & \cdots & & 0 \\ & \ddots & & & \\ \vdots & & \rho_{N-n} & & \vdots \\ & & & \ddots & \\ 0 & & \cdots & & \rho_N \end{bmatrix} \tag{15.2}$$

Exhibit 15: A heterogeneous network impact model as described by Billio et al. (2016). In which I is the identity matrix, H is a diagonal matrix of unique heterogeneous spatial weighting coefficients ρ_i, $\boldsymbol{r_t}$ is a vector of returns at time t, F_t is a matrix of factors at time t and η_t is a vector of noise terms at time t.

An interesting discussion raised in Billio et al. (2016) concerns the interpretation of negative spatial effects across networks of financial assets. In this work, authors argue that this phenomenon should be interpreted as a 'flight-to-safety' on the part of investors, moving to cash-rich companies under the constraints of their investor mandate. While these impacts may depend greatly on dynamics modelled by the graph, Billio et al. (2016) argue under a sector rotation market hypothesis, economic shocks transmitted from financial assets to the industrial sector may elicit negative spatial correlation due to anti-cyclical trades by investors. While these explanations are indeed plausible, negative spatial correlation may also capture a 'hot-potato' effect in miss-priced transfers of risk between companies, as insurers, bondholders or lessors take undiversified stakes in assets with latent defects.

$$r_{i,t} = \sum_{s=1}^{T} \rho_s r_{i,t-s} + \sum_{s=1}^{T} \theta_s r_{\text{cl},i,t-s} + \sum_{s=1}^{T} \phi_s r_{\text{dist.},i,t-s} + \gamma_t + \epsilon_{i,t} \tag{16.1}$$

Exhibit 16: Model use in Ahern (2013) to explore the lag at which network effects propagate. Here, $r_{i,t}$ denotes the returns of industry, i, at time, t, $r_{cl,i,t-s}$ represents the average returns of industries close to industry, i, at time, and $r_{dist,i,t-s}$ represents the average returns of industries distant to industry, i, at time. In this work, Ahern (2013) compute close and *distant* industries based on the 25th and 75th percentiles of the shortest-path distance from Dijskra's Algorithm (Dijkstra et al., 1959). In their estimates γ_t is used to define some constant effect at time, t, and $\epsilon_{i,t}$ a normally distributed error term for asset, i, at time, t.

Based on the centrality interpretation of social network econometric models, shown in the equation in Exhibit 10, Billio et al. (2016) contrast their discussion on idiosyncratic network shocks and endogenous idiosyncratic risk with research by Ahern (2013) in which authors looked to identify higher expected excess returns for assets central to the network. Unlike with Katz-Bonacich Centrality, authors make use of Eigenvector Centrality calculated as the principal eigenvector of the network's adjacency matrix. This measure holds many parallels to Katz-Bonacich Centrality under Spectral Ranking, with β equally the inverse of the largest eigenvalue of the adjacency matrix (Vigna, 2016). This adjacency matrix was constructed using a modified sectoral Social Accounting Matrix, used to identify inter-sectoral dependence. While these approaches differ in some ways from what we observe in the equation in Exhibit 10, authors find Eigenvector Centrality to have a positive impact on market excess returns over their sample of 385 sectors between 1993 to 2002. This result echoes similar findings in Buraschi and Porchia (2012) for directed graphs, in which Dynamic Centrality is found to contribute to lower Price-to-Dividend ratios and higher expected returns both empirically, in their sample of NYSE, AMEX and NASDAQ listed companies between 1963 to 2007, and in their simulation of a Lucas Asset Pricing Model (Lucas Jr, 1978).

	Intercept	RM-RF	SMB	HML	UMD	CMP	Adj. R2
Panel A: Size and Book-to-Market Portfolios							
Coef.	0.900	-0.153	0.007	0.439*	2.726**		0.482
t-stat.	1.406	-0.211	0.030	1.969	2.780		
Coef.	1.141*	-0.411	0.021	0.405*	2.363***	-0.337	0.534
t-stat.	2.087	-0.659	0.093	1.920	3.030	-1.005	
Panel B: Industry Porfolios							
Coef.	0.365	0.972**	0.626**	-0.486*	1.165*		0.139
t-stat.	1.513	2.571	2.262	-1.862	1.954		
Coef.	0.236	1.057**	0.679**	-0.300	1.254**	0.507**	0.157
t-stat.	0.947	2.705	2.407	-1.060	2.104	2.093	
Panel C: Firm-level Returns							
Coef.	0.666***	0.598**	0.525***	-0.357*	0.336		0.086
t-stat.	7.975	2.303	2.599	-1.847	1.225		
Coef.	0.664***	0.612**	0.518**	-0.361*	0.345	0.301*	0.099
t-stat.	7.953	2.349	2.577	-1.883	1.245	1.746	

Table 2: Results from Ahern (2013) demonstrating the impacts of Eigenvector Centrality, given in Central Minus Peripheral (CMP), using factors, High Minus Low (HML), Up Minus Down (UMD) and Market Excess Return (RM-RF), from Carhart (1997). In Panels B and C, Ahern (2013) demonstrate positive and statistically significant estimates on CMP against an unsaturated model. On a firm-level, the reduction in SMB estimates is used to explore the impacts of variable inflation through the omission of CMP. In this table * represents coefficient estimates with corresponding p-value less that 0.1, ** with with corresponding p-value less that 0.05 and *** with corresponding p-value less that 0.01.

In a fascinating extension of this work in Ahern (2013), authors explore the impact and time-lag over which network effects propagate using a model defined in the equation in Exhibit 16. From this work, authors find shocks in close industries to propagate with almost immediate effects, with shocks in distant industries propagating at some lag. Using the model detailed in the equation in Exhibit 16, authors go on to identify statistically significant positive and negative correlations between their Central Minus Peripheral (CMP) factor and the Small Minus Big (SMB) and High Minus Low (HML) factors use by Fama and French (1992). While it is difficult to identify causality from these results, these findings may suggest either that size somehow may be intrinsically linked to centrality or that centrality may have some impact on the efficiency with which firms generate future discounted cash-flows for shareholders from some fixed number of assets. In their regression results presented in Table 2, we see the impact which the inclusion of their CMP factor in their model has on the significance of their HML and SMB estimates across their industry- and firm-level models. In this work,

authors suggest coefficient estimates on CMP may provide investors with a valuable ex ante estimate of beta, or a firm or industry's exposure to systematic risk. This supports findings from Aobdia, Caskey and Ozel (2014) who identify central firms to have a stronger correlation with systematic risk in the market.

$$\boldsymbol{g} = \Gamma W \boldsymbol{g} + \boldsymbol{\epsilon} = (I - \Gamma)^{-1} \boldsymbol{\epsilon} \tag{17.1}$$

Exhibit 17: Heterogeneous network model presented in Kelly et al. (2013) with $\boldsymbol{g}$ representing a vector of company growth rates, $\boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2 I)$ representing idiosyncratic shocks, W the adjacency matrix of a random directed graph shown in equation 17.2, I is the identity matrix and Γ a diagonal matrix of elements γ_i representing the heterogenous network effects used in equation 17.1. This model presents discussion on the role of centrality and connectedness on sectoral flows in financial markets presented in Section 2.6.

In work by Kelly et al. (2013), authors investigate the impact of firm size on return volatility. In their analysis, Kelly et al. (2013) argue that by diversifying their customer bases, large firms manage risk to their discounted future cash flows and reducing return volatility as priced by the market. Kelly et al. (2013) make use of a network model of supplier and customer relationships sampled from log-normal firm size distribution in order to sample their directed graph, shown in the equation in Exhibit 17. Under this approach, authors rely on a simulated method of moments technique to estimate their model using a customer-supplier network information from the Compustat segment and CRSP stock market datasets between 1980-2012. Despite controlling for internal firm diversification, authors find network effects to be significant in their model and find these effects to provide a crucial mechanism through which to capture the over-dispersion of return volatility. In addition, authors find evidence to suggest that firm volatility depend crucially on both on firm size and on its out-Herfindahl. This out-Herfindahl, or Herfindahl-Hirschman Index (HHI), measure is closely related to many centrality measures in Spectral Ranking and is used to determine its customer network concentration, as shown in the equation in Exhibit 18. This work supports findings from Ahern (2013) who argue that returns to centrality emerge as a premium for holding greater risk.

$$V(g_i) = \sigma_\epsilon^2 (1 + \gamma_i^2 H_i^{out}) \tag{18.1}$$

$$(I - \Gamma W)^{-1} = I + \Gamma W + (\Gamma W)^2 + ... \approx I + \Gamma W; w_{i,j} \leq 0; \text{s.t.} \gamma_i \leq 0 \tag{18.2}$$

$$H_i^{out} = \sum_{j=1}^{N} w_{i,j}^2 \tag{18.3}$$

Exhibit 18: Heterogeneous network model presented in Kelly et al. (2013) with $\boldsymbol{g}$ representing a vector of company growth rates and $\epsilon \sim N(0, \sigma_\epsilon^2 I)$ representing idiosyncratic shocks, W the adjacency matrix of a random directed graph and Γ a diagonal matrix of elements γ_i representing the heterogenous network effects.

In Herskovic (2018), authors rely on the U.S. Bureau of Economic Analysis' (BEA) Input-Output Tables on sectoral linkages between 1997 to 2012 in order construct a directed graph illustrating input flow between sectors with which to investigate the impact of structural changes on systematic risk and equilibrium asset pricing. Using a Cobb-Douglas production model, authors motivate the impact of technological progress on network structural changes to identify network sparsity and concentration as key measures of input and output firm specialisation. Using their model, authors were able to frame firm input weightings according to the elasticity of their investment decision with respect to possible suppliers. Using this approach, Herskovic (2018) again identify Katz-Bonacich Centrality as the determinant of firms' share of total output, with β in the equation in Exhibit 9 capturing the decreasing marginal returns to input investments.

$$\log \mathcal{C}_{t+1} - \log \mathcal{C}_t = \frac{1}{1-\eta}[\eta \Delta \mathcal{N}^{\mathcal{S}}_{t+1} - (1-\eta)\mathcal{N}^{\mathcal{C}}_{t+1} + \Delta e_{t+1}] \tag{19.1}$$

$$\mathcal{N}^{\mathcal{S}}_t = \sum_i \delta_{i,t} \sum_j w_{ij,t} \log w_{ij,t} \tag{19.2}$$

$$\mathcal{N}^{\mathcal{C}}_t = \sum_i \delta_{i,t} \log \delta_{i,t} \tag{19.3}$$

$$e_t = \sum_i \delta_{i,t} \log \epsilon_{i,t} \tag{19.4}$$

Exhibit 19: Measures network sparsity (inputs) and concentration (outputs) derived from a Cobb-Douglas Production Function in Herskovic (2018) in which $\log \mathcal{C}_{t+1} - \log \mathcal{C}_t$ represents equilibrium consumption expenditure growth at time t, based on consumption expenditure $\mathcal{C}$. $\delta_{i,t}$ represents firm Katz-Benocich Centrality, η represents the decreasing marginal returns to input investment, $w_{i,j}$ represents the directed weight between firm i and firm j along the graph, $\mathcal{N}^{\mathcal{S}}_{t+1}$ represents network sparsity, $\mathcal{N}^{\mathcal{C}}_{t+1}$ network concentration and e_t represents residual Total Factor Productivity (TFP) (Herskovic, 2018).

Looking to Equilibrium Consumption Expenditure Growth, Herskovic (2018) use computed Katz-Bonacich Centrality and firm input weights to compute their network measures of sparsity, concentration and residual Total Factor Productivity (TFP) for use in factor modelling. From these definitions, we can see the obvious extension their measures of sparsity and concentration provide over similar experiments in Buraschi and Porchia (2012). Looking at these measures, we see the implied penalty incurred by sectors relying on only a handful of central inputs or outputs based on decreasing returns to scale and substitutability. From this model Buraschi and Porchia (2012) expect both positive shocks to sparsity and negative shocks to concentration to lead to higher firm consumption and lower firm marginal utility. Using a sample of CRSP stocks over the period, Buraschi and Porchia (2012) estimate a sparsity-beta portfolio return spread of 6% and concentration-beta portfolio spread of -4% per year from their model, shown in the equation in Exhibit 20, supporting their findings. From this work, we gain a microeconomic perspective through which to support and understand the many findings of centrality provided in Ahern (2013), Aobdia et al. (2014) and Buraschi and Porchia (2012). These findings complement many of the models provided by Chen et al. (2018), Chen et al. (2018) and Roson and van den Bergh (2000), to build a nuanced view of centrality focused on input-output flow in the economy.

$$r_{i,t} = \alpha_i + \beta_{i,S}\delta \mathcal{N}^{\mathcal{S}}_t + \beta_{i,C}\delta \mathcal{N}^{\mathcal{C}}_t + \text{Controls} + \xi_{i,t} \tag{20.1}$$

Exhibit 20: Factor model from Herskovic (2018) in which α_i is the market excess return, $r_{i,t}$ represents the arithmetic returns of asset i in discrete time t, $\mathcal{N}^{\mathcal{S}}_{t+1}$ represents network sparsity, $\mathcal{N}^{\mathcal{C}}_{t+1}$ network concentration and $\xi_{i,t}$ some normally distributed error term.

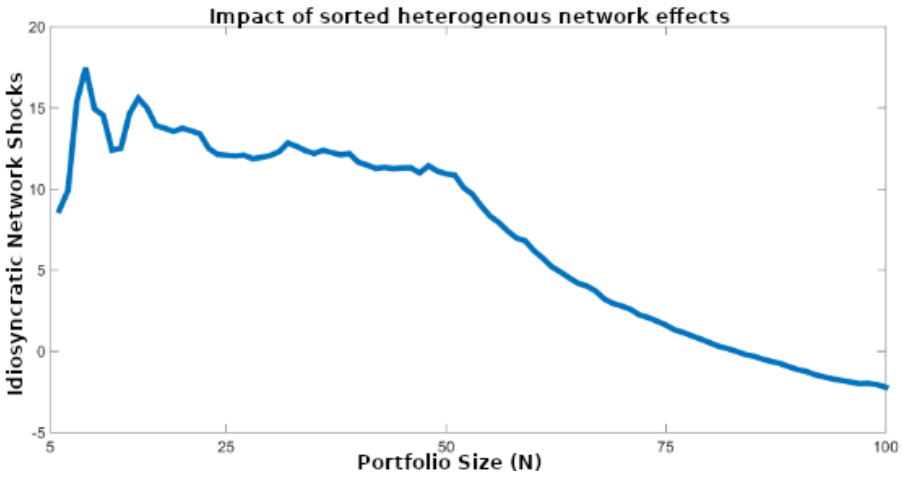


Figure 8: Diagram in Billio et al. (2016) illustrating the endogenous idiosyncratic risk absorption effects of the negatively spatially correlated stocks in a portfolio sampled from a universe of assets charcterised by heterogenous network effects. In this diagram, Billio et al. (2016) illustrate the decrease in portfolio risk shown along the y-axis, as the number of negatively spatially correlated assets increases along the x-axis. This relationship is discussed in further in Section 2.6.

While centrality remains an obvious and critical measure through which to assess firm exposure and impact on risk and return across the market, degree appears another critical factor through which to assess the spread of direct contagion. Similar to Ahern (2013), who concern their work with the lag at which effects propagate across a graph and impact of distance along the graph, Gai and Kapadia (2019) investigate the degree across which contagion propagates. Gai and Kapadia (2019) contrast this work to efforts in epidemiology wherein network models serve a critical approach under which to understand the impacts of social contact on how far diseases can spread. In this work, authors explore a contagion window to understand this spread, as illustrated in Figure 8.

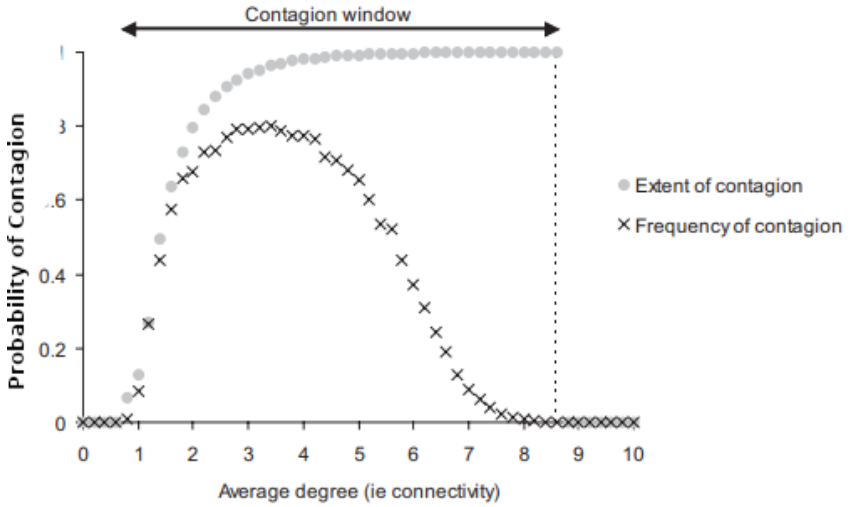


Figure 9: Diagram from Gai et al. (2004) illustrating average spread of contagion to nodes of a particular degree along a graph. In this diagram, the frequency at which a particular node is affected by contagion events increases with average degree for small average degrees and decreased with average degree for average high degrees. This figure is further discussed in the context of a simulation by Gai et al. (2004) in Section 2.6.

In Gai, Haldane and Kapadia (2011), authors explore under simulation the spread of distress across a graph. This network model looked to explore systemic liquidity crises in interbank lending. Using this approach, Gai et al. (2011) identify under behaviours across a contagion windows which inform regulatory decisions, shown in Figure 10.

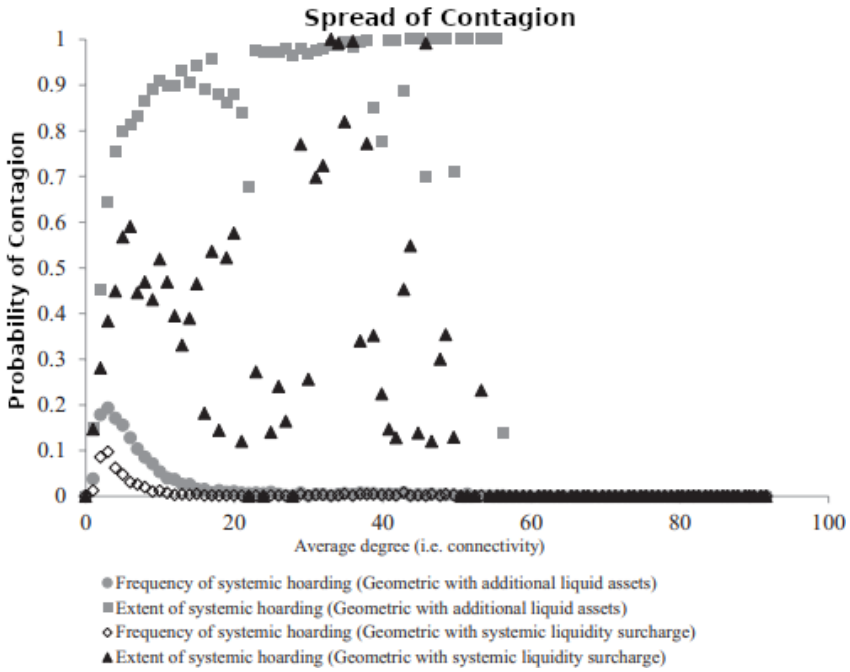


Figure 10: Diagram in Gai et al. (2011) illustrating the average spread of contagion to firms at a particular degree. This follows

Through these efforts, Gai et al. (2011) demonstrate the efficiency of policies targeting tougher liquidity requirements at a systemically important bank or banks central to the network in limiting systemic liquidity hoarding, as opposed to policies which set requirements across all lenders. This supports findings by Herskovic (2018) on the importance and impact of graph centrality on how risk propagates.

In an experiment exploring heterogeneous price impacts in which assets are allowed to take on unique values for ρ, shown in the equation in Exhibit 15. Billio et al. (2016) go on to explore the endogenous idiosyncratic risk absorption effects of negatively spatially correlated stocks in a portfolio. In order to illustrate these impacts, shown in Figure 8, Billio et al. (2016) order asset-specific network impacts sampled from a normal distribution, $\rho_i \sim N(0.5, 0.01)$, in descending order. By successively adding each stock with decreasing network impact to their equally weighted portfolio, they analyse how negative network impacts, starting at asset fifty, impact endogenous idiosyncratic risk. From this simulation, Billio et al. (2016) demonstrate the value of stocks with negative network effects in managing portfolio risk in a market characterised by heterogenous network effects. Billio et al. (2016) note how at asset eighty they observe a negative endogenous idiosyncratic risk component, suggesting the impact careful asset selection can have on portfolio's composition of risk in such a market. Comparing this work with findings in Bonacich (1987) and Roson and van den Bergh (2000), heterogeneous network effects may be a natural means through which to reconcile the modelling complexities in which firms exploit neighbouring firm density or isolation depending on the local network and market structure.

$$W^{\star} = \begin{bmatrix} W_1 & & \cdots & & 0 \\ & \ddots & & & \\ \vdots & & W_{T-k} & & \vdots \\ & & & \ddots & \\ 0 & & \cdots & & W_T \end{bmatrix} \tag{21.1}$$

Exhibit 21: A time-varying spatial weight matrix as discussed in Billio et al. (2016). Where $W^{\star}$ represents a matrix of T, N by N spatial weight matrices in which N represents our number of participants or assets and T are our time-steps.

An important hypothesis raised by Billio et al. (2016) concerns the impact of structural breaks under time-varying graphs, in which liquidation, acquisition or supply-chains disruption causes disconnected cliques to emerge across the graph when edges disappear. These changes are immediate under spatial interaction but appear slowly in rolling historical price covariances as we move over the event window in which these cliques emerged. Without controlling for network exposure, estimates on factor exposure may be highly uncertain as to the true underlying exposure of individual securities as estimates adjust to some halfway-house between structural breaks.

A major advantage of these spatial and social network approaches to factor modelling is that these structural breaks can be controlled for through changes in the underlying graph leading to estimates in factor exposures which may appear more stable. A major challenge to these methods, however, is around the complexity introduced as we construct low-rank, square, time-varying spatial weight matrices, as shown in the equation in Exhibit 21. This may not be practical in many applications, both computationally as methods often rely on storing and inverting these large matrices, and practically, as studies try to collect data on these graphs. While Billio et al. (2016) discuss these impacts, Blasques, Koopman, Lucas and Schaumburg (2016) go on to explore time-varying spatial correlation both empirically and under simulation by treating their spatial weight matrix of cross-border debt linkages as static and letting the parameter ρ vary through time in their Spatial Error Model. In this work, Blasques et al. (2016) motivate their choice of spatial weight matrix, arguing varying notions of economic and geographic dependency in global financial markets. In their study of European Credit Default Swaps (CDSs), Blasques et al. (2016) uncover a break in spatial dependence following regulatory changes introduced in 2012. Following 2008, this spatial approach to studying CDS and sovereign debt has provided valuable insights into the changing interdependence of markets through time, as explored through Eder and Keiler (2012) and Debarsy, Dossougoin, Ertur and Gnabo (2018). In Blasques et al. (2016), the sudden increase in spatial dependence following 2012 regulatory changes provided valuable insight to researchers and regulators into the impact of policy creation on systematic risk.

While many approaches consider a single space or graph on which to estimate spatial dependence, multiple spaces or graphs can be incorporated into spatial models in order to model spatial dependence across a range of factors. In Blasques et al. (2016), spatial dependence is extended to a notion of economic dependence using cross-border debt linkages, Asgharian, Hess and Liu (2013) rely on methods in spatial panel data analysis to explore and test for various measures of spatial or economic dependence across stock market indices using exchange rate volatility, absolute differences in inflation expectations, bilateral trade, bilateral FDI as well as geographical distances to construct spatial weight matrices. Here, using a sample of 41 equity markets between January 1995 and October 2010, Asgharian et al. (2013) were able to identify the critical simultaneous impact of bilateral trade and exchange rate volatility on index returns across time.

From this work, we gain a rich sense of the impact which market network structure has on asset and portfolio risk and return. Using these methods, we find insight into how to price risk and how financial distress spreads along a graph, as well as the influence and impact which centrality has in defining systematic risk in the market. These impacts are motivated across a variety of theoretical work in microeconomics, macroeconomics, portfolio optimisation and game theory, and can be easily decomposed and simulated across a variety of models and assumptions. Borrowing from work in Spatial and Social Network Econometrics, Graph Theory, Arbitrage Pricing Theory, these works outline a broad literature aimed as re-contextualising existing findings in Empirical Finance, as in Ahern (2013) and Kelly et al. (2013), as well as developing new approaches to financial regulation and investor decision-making, as in Gai et al. (2011) and Blasques et al. (2016).

2.7 Financial Data Breaches

In January of 2020, the world witnessed a data leak of some 715,000 emails, charts, contracts, audits and accounts which exposing a long history of corruption, nepotism and embezzlement between Isabel dos Santos, Africa's richest women, and the Angolan Government and State-owned Enterprises (Barr, 2020). This data breach dubbed the Luanda Leaks follows a series of similar ongoing breaches in recent years which have exposed large scale corruption around the world. Over the last decade, data from the Offshore Leaks (2013), Bahamas Leaks (2016), Panama Papers (2016), Paradise Papers (2017 2018) has provided the opportunity for teams of software engineers and journalist to index and report on the countless cases of corruption and fraud found in these petabyte-scale data dumps marking a significant step in digital journalism (International Consortium of Investigative Journalists, n.d.*b*).

While many of the previously leaked documents have been kept private for security, privacy and ethical reasons, the International Consortium of Investigative Journalists (ICIJ), who serve as custodians for the leaked data, do provide metadata extracted from these documents for use by the general public. This metadata is provided to the public through both an online portal for search and visualisation and a graph database which can be downloaded for use in further research. This database contains information on the named-entities mentioned in the leaked documents, as well as tags describing the relationships between these entities. While the ICIJ does not make explicit mention of how these entities are extracted for each of the data leaks, their software does point to several open-source tools designed for named-entity recognition and conference resolution, such as Standford's Core NLP Named Entity Recognizer and the Apache OpenNLP software (Lafferty, McCallum and Pereira, 2001; International Consortium of Investigative Journalists, n.d.*b*).

These leaks are of various sizes and have seen varied attention by the international community. The largest of the leaks, the Paradise Papers, spans over 13.4 million documents and 120 000 people and companies. The details of this leak were first released on 5 November 2017 triggering global inquiry by the ICIJ's network of some 380 journalists. While this

leak is the largest, many have critiqued public outcry citing the extraordinary response to the Panama Papers sparked worldwide White (2017). While technology and practises in journalism have grown to accommodate leaks of this scale, another criticism raised by journalists has been the obvious lag between public opinion, response and the law. While many journalists have uncovered cases of corruption, embezzlement in these leaks, much of the schemes captured in these leaked documents represent legal approaches to ensuring personal privacy and improving the tax efficiency for an individual or legal entity. Through time, these schemes have evolved to account for changes in business requirement, law and legal opinion; and have required various intermediaries through which to respond to regulatory changes (Brothers, 2014). In 2010, changes to international tax law spurred many US technology companies to restructure as changing Irish tax code eliminated the requirement of a Dutch intermediary from the common 'Double Irish and Dutch Sandwich' base erosion and profit shifting (BEPS) corporate tax avoidance tool (van der Does de Willebois, Halter, Harrison, Park and Sharman, 2011*b*).

Inside of academia, data leaks held by ICIJ has seen interest across disciplines from research in Information Systems, to research in Tax Law and Multidimensional Visualization (Zhuhadar and Ciampa, 2019; Wiedemann, Yimam and Biemann, 2018). While these leaks present many interesting opportunities for researchers across disciplines; to date, much of this research has been limited by the scope and quality of the metadata provided by the ICIJ. It is for this reason that many researchers have relied on the methods of Network Analysis to identify relationships and clusters in the data. A major draw for researchers exploring the ICIJ graphs lies in rich availability of information on private ownership and transactions. This provides researchers with the ability to realistically capture network effects not typically available through public data sources. In work by Hajek and Henriques (2017) and Joaristi, Serra and Spezzano (2018), this metadata provided the opportunity to compare and develop methods for identifying bad entities using extensions on the popular Node2Vec and PageRank algorithm for dimensionality reduction. In Caruana-Galizia and Caruana-Galizia (2016), this involved a panel study of leaks to explore the impact of regulatory change on ownership substitution between EU and non-EU jurisdictions by firms.

In Garcia Alvarado and Mandel (2019), researches draw on recent advances in Graph and Game Theory to explore the impacts of graph structure on the ability and propensity of firms to evade taxes. In their work, Garcia Alvarado and Mandel (2019) focus their attention on firm jurisdictions to arrive at an optimal deterrence strategy for social-planners in a Stackelberg competition against a strategic tax-evader. In their findings, Garcia Alvarado and Mandel (2019) recommend to social-planners, under this Stackelberg competition, to pursue tax information exchange agreements between countries based on the Bonacich centrality of particular nodes, shown in the equation in Exhibit 9. Garcia Alvarado and Mandel (2019) arrive at this conclusion based on (weak) approximations of the optimal policy through a greedy algorithm which investigates a utility function for tax-evaders based on detection probabilities across the edges of a graph. These detection probabilities represent the probability that a tax authority identities an evading company based on its relationship to other evading firms. An interesting conclusion from Garcia Alvarado and Mandel (2019) concerned over-dispersion identified in the degree distribution of the graph. In this finding, Garcia Alvarado and Mandel (2019) posit the graph to exhibit many properties atypical of Poisson Random Graphs, suggesting a number of common graph structures likely under their game-theoretic model. Work in Garcia Alvarado and Mandel (2019) suggests that these graphs likely organise in response to regulatory power, raising important concerns over approaches which ignore the impacts of transaction intermediaries and exo-havens in obfuscating transactions (Dharmapala and Hines Jr, 2009). In this work, Garcia Alvarado and Mandel (2019) posit the optimal formation of quasi-star or quasi-complete networks in response to regulatory deterrence. Under such conditions, while graphs exhibit important global and local geometry, all nodes may share the same number of edges. This suggests that under conditions in which firms arrange optimally based on regulation or market structure to form scale-free networks, methods used by Procházková (2020) exploring connectedness on a learned graph may form unreliable estimates on the impacts of direct contagion on market prices as all nodes form an equal number of connections.

In O'Donovan et al. (2019), researchers investigate the impacts of graph membership on market returns using factor modelling. Under this approach, researchers compare companies mentioned in the leaks to some sample on their respective exchanges. From this study, O'Donovan et al. (2019) estimate a $174 billion loss in market capitalisation for the 388 companies mentioned in the leak based on the estimated coefficient on their graph membership binary variable, used to indicate whether a company was found to be present in the leaks or not. Unlike approaches in graph learning which attempt to identify contagion-type effects based on features from their graph, this approach looks to limit its assumptions by ignoring graph structure or network effects. By focussing on membership, O'Donovan et al. (2019) may distort estimates by ignoring contagion effects which spill outside the ICIJ graph to the rest of the market as well as economic rents which may have accrued to companies based on their dominant position within the graph.

While this work by O'Donovan et al. (2019) serves as an important benchmark for studies inside of academic finance, its use of graph membership does little to explain the real social effects which arise when companies transact across the social graph described by these leaked documents. Intuitively, we may expect company exposure to be influenced by both their level of connectedness or centrality, as in Procházková (2020), or their relationship to bad entities as identified in Hajek and Henriques (2017) and Joaristi et al. (2018). These relationships may express themselves across the unique structure

of the graph as companies optimal organise to manage risk and capture excess profit.

While O'Donovan et al. (2019) explore the impact of these leaks on company market capitalisation narrowly, their work raises a number of important questions concerning the long-term price impact of shelf companies and special purpose vehicles (SPVs) used in tax avoidance across jurisdictions. While O'Donovan et al. (2019) estimate a negative price impact in graph membership over their event window, membership may capture varying effects across different time horizons. While membership may capture a number of regulatory risks, as presented by O'Donovan et al. (2019), membership in the ICIJ leaks may also signal a long-term strategic commitment to and investment in tax avoidance. While O'Donovan et al. (2019) point to the many cases of fraud and corruption exposed by the leaks, researchers may also observe a variety of concerns by investors over prosecution, counter-party risk, public backlash or regulatory changes which limit their ability to fully realise commitments in tax avoidance.

Looking across this existing research, room exists for work building on the progress and insights of Billio et al. (2016) and Fernandez (2011). These developments in Spatial-CAPM and Spatial-APT may serve as important frameworks through which to understand and explore relationships in the ICIJ metadata and their impact on market prices. Using these ICIJ leaks, we may look to find novel extensions and application of these approaches in investigating real-world market graphs and value tools in exploring market events. Looking to real-world data, these leaks may allow us to test the application and insights from Billio et al. (2016), while extending Ahern (2013), Buraschi and Porchia (2012) and Kelly et al. (2013) using data on firm-level networks with easily explainable and interpretable edges.

Unlike existing work exploring risk across market graphs, the metadata provided by the ICIJ from these leaks provides not only insight into direct relationships across public companies but insights into the structure and impact of private relationships which may be impossible to observe in other datasets. As in Garcia Alvarado and Mandel (2019), this 'dark subgraph' of private intermediaries provides not only insight on the many indirect connections between firms but also the important impact and role which regulation plays in structuring transactions and managing risk between firms through transaction intermediaries. Using this data, we may investigate methods in Graph Signal Processing from a pricing perspective, to understand how market signals oscillate across the structure of a market graph. This may build on the many findings of Gai et al. (2011), Ahern (2013), Gai and Kapadia (2019) and Herskovic (2018) to understand the spread of direct contagion in the market from emanating sources and provide insight into how network position effects ones exposure to risk across different subgraph structures. Using the methods of Spatial and Social Network Econometrics may also come to understand, under our Centrality interpretations, the impact of graph position in defining market dominance and risk-return through estimation of our model.

3 Data

The ICIJ have served as the custodians of over 29,4 million leaked financial documents, spanning five major leaks:

Leak	Date	Documents	Nodes	Edges
Offshore Leaks	2013-6-14	2.5 million	280000	561394
Panama Papers	2016-4-3	11.5 million	559604	674103
Bahamas Leaks	2016-8-21	1.3 million	202246	249191
Paradise Papers	2017-11-5	13.4 million	867936	1657841
Luanda Leaks	2020-01-20	0.7 million		

Figure 11: Table showing the size of different leaks held and published by the ICIJ. In this table, the Paradise Papers is shown to be the largest leak with 13.4 million documents.

While many of the activities described in these leaked documents are perfectly legal; reporting by ICIJ and its collaborators has revealed numerous cases of money laundering, tax evasion, fraud and other crimes. As many leaked documents reference personal bank accounts, email exchanges and financial transactions, the ICIJ have limited public access, providing only metadata on named-entities and their relationships extracted through a series of robust open-source softwares (Datashare, n.d.; International Consortium of Investigative Journalists, 2020; Stanford NLP, 2020; Manning et al., 2014). This software and its specific algorithmic approaches have been documented extensively across publications by the ICIJ and their partners, providing necessary transparency into the challenges and successes of these tools (Hunger and Lyon, 2016; Fitzgibbon, 20169). While the reliability of these approaches require consideration by those studying these leaks, existing research must acknowledge these uncertainties in their methodologies and findings, given the limited access to source material and important privacy concerns.

Given the size and critical research interest in the Paradise Papers Leaks, our study has opted to limit its efforts to this most recent graph. This allows our work to be more comparative in its findings while leveraging a more complete graph of documents exposed from the legal firm Appleby.

A major challenge in using this data for financial research has lied in the challenge of appropriately matching this data to lists of publicly traded companies. While in O'Donovan et al. (2019), this task was accomplished using fuzzy string-matching over cleaned and standardized names and manual verification, we apply a stricter matching process, requiring perfect matches over cleaned and standardized names with manual verification. While this yields fewer observations for our study, this approach provides greater assurances over the quality of our observations. In Appendix B we provide a list of one-hundred and forty-five matched entities along with their stock market ticker and OpenFIGI identifier.

In order to study the market reaction to these leaks, we must assume an appropriate event window for collecting financial data. For O'Donovan et al. (2019) working on the Paradise Papers, a 5-day event window was chosen starting on 3 November 2017 and ending on the 8 November 2018. While we would like to assume information regarding these leaks was instantaneously prices once publicly released, leaning on the event window of O'Donovan et al. (2019) allows or us to compare our results with their findings.

Information on adjusted share price, annual financial statements, market capitalization, market indexes and interest rates on 3-month AAA-rated US Treasury Bills was collected from the IEX Cloud Database for use in our analysis. Using this data, returns, price-to-earning, price-to-research and profit margins were computed for these companies for use in later modelling (*IEX Cloud API*, n.d.).

$$j = \frac{n}{6}\left(s^2 + \frac{1}{4}(k-3)^2\right); \tag{22.1}$$

$$s = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}; \tag{22.2}$$

$$k = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{2}} \tag{22.3}$$

Exhibit 22: Formula for the Jaruqe-Bera test statistic (j) in which n represents the number of observations and degrees of freedom in our sample, $\bar{x}$ our sample mean and $\hat{\sigma}^2$ our sample variance. Here, our test statistic is compared to the χ^2 with two degrees of freedom to test against our joint null hypothesis that our sample exhibits zero expected skewness and zero excess kurtosis (Jarque and Bera, 1980).

While no efforts were made to impute or smooth data, the data was conservatively winsorized at the lower 2.8% and upper 6.2%. This value was chosen by analyzing its effect on the moments of returns against our expectations for skewness and kurtosis from a normal distribution and validated against the Jarque-Bera test for normality, shown in the equations in Exhibit 22. While many studies in Empirical Finance concern their work with challenges in survivorship bias, our formulation as an event study hopes to repudiate these concerns as we focus our interest on the impacts of particular companies following the public disclosure of ICIJ data within a closely defined event window.

4 Exploratory Analysis

Our research aims to build on a number of common Graph Signal Processing and Social Network Econometric models and approaches aimed at testing and characterizing the presence of social effects in market response to the public release of information contained in these leaks. To arrive at such an approach, this work will rely heavily on exploratory analysis with which to draw this bridge through our data and hypothesis. Using this exploratory work and our theoretical foundations, we aim to strongly motivate the richness in our methodology in exploring such data.

In this exploratory work, we aim to investigate artefacts in directionality in our graph, the impacts of graph censorship on important summary statistics, expectations concerning path length, regularities in the position of matched listed companies and experienced price impact and early evidence for spatial correlation in between returns and firm characteristics. The term, graph censorship, will be used to describe the property of our graph whereby only price measurements on the impact of disclosure and graph structure on expected future discounted cash-flows are observed only for matched listed companies and not for all entities along our graph. Much of our exploratory work will look to explore the properties of this censorship on which to build techniques in spatial modelling. This will rely on important findings and methods in Graph Signal Processing and expectations concerning the properties of various Random Graphs and Graph Generating Functions. From these findings, we will motivate appropriate projections of our graph based on discussions in Natural Language Processing, Tax Law and Microeconomics exploring the ICIJ Leaks.

4.1 Foundations for Symmetry

As metadata from the ICIJ is built on text documents, the edge information provided to researchers has looked to capture the important semantic relationship between the entities based on text parsing techniques in Natural Language Processing. To convey this information the ICIJ detail edge information using a "START_ID", representing the unique identifier for each entity, a "TYPE" representing semantic information about the relationship, and an "END_ID". While this formatting of the data may easily be interpreted as defining a directed graph of entities, it is important to discern both artefacts of this approach, the intent of the publisher, as well as the key differences between semantic and economic notions of directedness uncovered in these documents. The edge types "connected_to" and "same_name_as" used by the ICIJ present obvious examples of how processing techniques capture artefacts in directedness in cases where "same_name_as" is used to suggest undirected or bidirectional relationships between identifiers. For the "officer_of" and "intemediary_of" types presented in the edge data, these attributes may present some legal or semantic relationship identified in these documents which may differ from the economic substance of how risk is shared and transferred. To uncover this substance, we require some general notion of how these relationships generate risk and return, as well as our uncertainties, through which to assert some structure for our analysis. To motivate such structure, we will look to a general notion of how transactions form in microeconomic theory to place assumptions on how these relationships may transmit risk and return.

While transactions may bear diverse strategic benefit between parties, under general (competitive) equilibrium, we must assume that both price and quantity are set to be Pareto optimal according to the Marginal Rate of Substitution of both parties (Arrow et al., 1951). Under such a model of transactions, we must assume, though not equal, that both parties benefit from and are harmed by counter-party risk. In insurance agreements, while an obvious transfer of risk dictates the underlying substance of the transaction, default by either party may trigger distress in the counter-party as insured parties find themselves exposed to undue risk and insurers find themselves either unable to collect premiums or under-diversified to firm-specific risks. This model of transaction places bidirectionality as an important property required in graph construction for graphs modelling transactions, through which to define the mechanism across which risk is transmitted across our graph.

This assumption may not be as well-motivated in other studies exploring market microstructure or interaction between sectors in the economy and may present challenges in model identifiability under certain structure learning approaches. However, under our foundation, this model for transactions provides our study with a motivation with which to assume symmetry. This assumption may bear an important impact on our findings but provides a critical safety net when compared to approaches that ignore or place no such constraint on bidirectionality. This symmetry serves as an important assumption required under particular methods in Graph Signal Processing and Spatial Regression on which to build our analysis and offers a framework under which to simplify our analysis. Given the application of VAR methods in Structure Learning approaches, it is critical to understand how directionality may be observed in measurements across an undirected graph as waves emanating from regions exposed to particular prominent exogenous risk factors. This provides some justification or bridge between structure learning models which assume graphs are either directed or acyclic and the approaches taken in our study.

Practically, assuming directionality in the edge data provided by the ICIJ, few companies are connected even across

long path distances. This suggests, that under directed relationships sparsity may render projections and analysis impossible. We will begin our analysis by assuming that edges relayed in the ICIJ data describe a symmetric graph, on which we will build further methodology.

4.2 Evidence for Structure

While methods in Simultaneous Auto-correlation and Graph Signal Processing provide rich tools with which to explore measurements along our graph, the presence of private firms, individuals and transaction intermediaries render application of these techniques to market data challenging as many of these methods assume measurements can be taken across all nodes of the network. For private firms and individuals, we cannot observe market pricing with which to estimate our models and so must look to either impute this missing information or project our graph in such a way as to preserve its important characteristics while ensuring measurements can be observed at all nodes in the new graph. This must not only be motivated theoretically but also from our domain knowledge of this graph and the economic and regulatory environment which drives its graph generating process.

As we subset our network on matched entities, we discover an important property; the absence of direct edges between our sample of listed companies. While it may be difficult to determine whether this property exists due to random variation or our matching procedure; companies are commonly known to incorporate and transact through multiple off-shore intermediaries for the purpose of tax avoidance and anonymity (van der Does de Willebois et al., 2011*b*). Given that these structures vary due to business requirement, law and legal opinion; identifying relationships between entities can be challenging as companies intentionally look to mask the economic substance of their transaction through the use of intermediaries (Brothers, 2014). Given the use of these intermediaries in optimally facilitating particular transactions under changing regulations, the local structure of this network may not serve to insulate or propagate risk, but rather exploit particular legal codes or aid in regulatory compliance. This may mean that unlike the graphs sampled in Billio et al. (2016), the local structure may mislead us as to the transfer and propagation of risk across the network as risk propagates along a latent graph which mirrors the economic substance of particular transactions rather than the structure which facilitates them. This then presents an important question, as to how to uncover or model this latent graph given our observations over the patent ICIJ metadata; or more simply, how to reduce our graph to model share price measurements over listed companies in order to assess the impact of these leaks and the network effects which emanate.

To study and motivate this property, exploring the intermediate structure between listed companies, we not only look to the numerous reports in van der Does de Willebois et al. (2011*a*), but to the regularity in listed companies across the graph. To identify this regularity, we look to Graph Fourier Analysis, shown in the equation in Exhibit 8, applying this approach to a signal marking companies as matched listed companies across the graph, with matched listed companies marked with a 1 and unlisted private firms, individuals and unmatched entities marked with a 0. Typically, researchers analyze the lowest eigenvalues (or frequencies) from the Graph Fourier Transform to motivate techniques in Community Detection or Graph Cutting. For our research, we pose interest in high-frequency components as we expect these components to signal regular structure between listed companies. Using the Lanczos Algorithm implemented in Lehoucq, Sorensen and Yang (1977), we compute the magnitudes of the 750 highest frequency components of the Paradise Papers Graph Fourier Series, shown in Figure 12 (Lanczos, 1950). From this analysis, we observe two frequencies with prominent magnitudes of 0.721 and 0.685 at eigenvalues 177.13 and 178.12. While it may be difficult to relate these frequencies to a particular structure or path length, we compare these results to a Monte Carlo simulation in which companies are placed randomly along a graph to identify how the largest magnitudes and eigenvalues vary against our observed results.

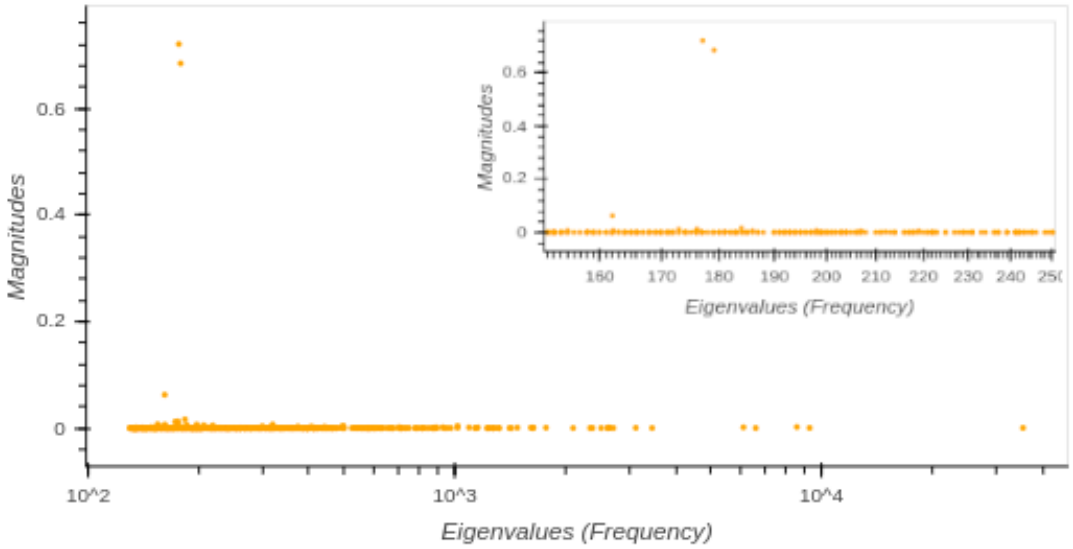


Figure 12: Application of Graph Fourier Analysis to ICIJ Graph with vertex signal marking matched listed companies, showing the magnitude of highest frequency (eigenvalue) Fourier Components indicating local structure. Eigenvalues of the Graph Fourier Transform are given along the x-axis, with magnitudes given along the y-axis. In this graph we observe our highest magnitude components contained in the highest frequencies of our graph. This suggests that listed companies find themselves tightly packed with some regular intermediate node structure. We use this finding in Section 4.2 in discussing certain observed stylized fact of our graph and market data. This figure can be recovered from the `gft_simulation` function provided in our project documentation (Gawronsky et al., 2020*c*).

To perform such a study, we perform 1000 shuffles of our listed companies across the graph. In Figure 13, we compare these results graphically to our observed frequencies to see a large prevalence of high magnitude lower frequency components compared to those observed in our data. Where we observe in the data largest magnitudes of 0.721 and 0.685 at eigenvalues 177.13 and 178.12; across simulations, we observe a median maximum magnitude and eigenvalue of 1.0 at median eigenvalue 130.99, with a standard deviation across our maximum magnitude eigenvalues of 26.35. This places our observed largest magnitude eigenvalues greater than 99.98% of our simulated largest magnitude eigenvalues. This prevalence of random signals to yield high magnitude components at larger eigenvalues we believe is suggestive of unique packing of our Matched Listed Companies in some small, central region of our graph.

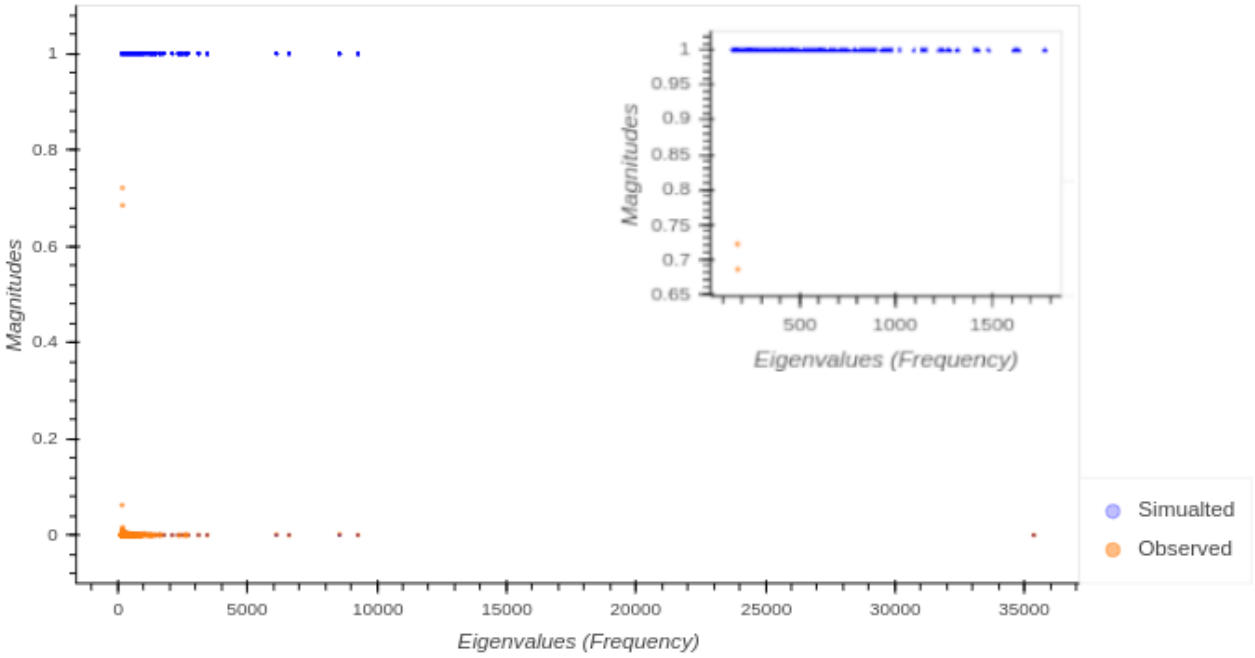


Figure 13: Graph Fourier Transform magnitudes and eigenvalues of one-thousand simulated signals of matched listed company on our ICIJ Graph, shown in blue, against our observed signals, shown in orange. Eigenvalues of the Graph Fourier Transform are given along the x-axis, with magnitudes given along the y-axis. Looking at our figure insert, in the top right-hand corner of our graph, we show against our simulation that our highest observed magnitudes are much smaller and fall much further to the left that our simulated positions of listed companies. We believe this may be an artifact which demonstrates listed companies concentrated in some region of our graph, with some mixed intermediate structure separating them. Under simulation, the graph tends to produce only one or two components with magnitudes close to one with much larger eigenvalues. This graph is discussed further alongside observations on shortest paths in Section 4.2. This figure can be recovered from the `gft_simulation` function provided in our project documentation (Gawronsky et al., 2020*c*).

In Table 3, we explore this finding in Figure 13 further by randomly, with a probability of 0.5, moving Matched Listed Company signals across to a nearest neighbour in our graph. While our experiment in Figure 13 affected global structure, this experiment looks only to perform small changes to local structure along our graph, moving entities only slightly closer or further along our graph. By making such changes we hope to observe sudden changes in magnitude which suggest regularities observed through our Graph Fourier Analysis have been disturbed. We perform this short random walk operation 1000 times and after each shuffle identify the eigenvalue which underwent the largest decrease in magnitude from our original Matched Listed Company signal to our perturbed signal. Across these simulations we compute the average magnitude difference for this eigenvector which undergoes the largest change in magnitude, along with the number of times it is recorded across our runs. From this experiment we observe an overwhelming and large drop-off at frequencies 184.025137 and 198.122279. We argue this result is suggestive of unique regularities in our matched listed companies which may serve as artefacts of common tax avoidance or compliance structures and may motivate an approach through which to use this regularity in summarizing our graph to better model the economic substance of firm interaction.

	$\max((\lvert\boldsymbol{\mathcal{F}}(\boldsymbol{G})_{\text{observed}}\rvert - \lvert\boldsymbol{\mathcal{F}}(\boldsymbol{G})_{\text{perturbed}}\rvert)_+)$	
Frequency	Average	Count
184.025137	0.010769	447
186.032203	0.006007	43
198.122279	0.005540	502
207.017494	0.005098	5
319.031244	0.006600	3
Total runs		1000

Table 3: To extend our simulation in Figure 13, we explore the impacts a short random walk of Matched Listed Company signals along our graph. To perform this experiment we allow, with a probability of 0.5, each Matched Listed Company signal to move to a neighbour. Across 1000 runs, we compare component magnitudes between our initial Matched Listed Company signal, denoted $\lvert\boldsymbol{\mathcal{F}}(\boldsymbol{G})_{\text{observed}}\rvert$, and our signal perturbed by these short random walks, denoted $\lvert\boldsymbol{\mathcal{F}}(\boldsymbol{G})_{\text{perturbed}}\rvert$. At each run of our simulation we identify the frequency with the largest decreases in magnitude from our original signal. From the results in the table above, we see an large drop-off at frequencies 184.025137 and 198.122279. We believe this is indicative of the presence of breaks in regular intermediate structure separating Matched Listed Companies along our graph. These findings are discussed in Section 4.2. Readers may recover this table using the `walks` function provided in our project documentation (Gawronsky et al., 2020*c*).

4.3 Graph Projections

To model these listed companies, we must then look to intentionally obfuscate this regular structure from the data to capture the economic substance of transactions which influence contagion. This involves not just identifying communities in our graph, but in projecting our graph onto some new adjacency matrix which ignores this structure. While this presents a major challenge to our analysis; this problem is common in certain research domains where researchers regularly look to model relationships across particular sets of nodes, as opposed to an entire graph.

Many real-world networks can be described as bipartite graphs of authors and papers, ingredients and recipes, or employers and their companies. These graphs describe disjoint and independent sets of nodes, separating all authors through their papers or all ingredients through recipes. To model these networks of authors or ingredients, researchers typically rely on simple, hyperbolic or resource allocation weighting to compress these graphs, through one-mode projection, onto a new graph representing a single independent set of nodes (Newman, 2001; Zhou, Ren, Medo and Zhang, 2007; Fan, Li, Zhang, Wu and Di, 2007). These methods present many trade-offs in the properties researchers look to preserve. In resource allocation bipartite projection, node degree is preserved in the projection, while in simple bipartite projections, symmetry is maintained in the projected unipartite graph (Coscia and Rossi, 2019).

Across our ICIJ metadata, while listed companies are not connected, private companies and individuals are connected constraining our ability to reliably apply these methods in bipartite network projection as our graph is not bipartite. The approach taken in this study explores projecting our graph onto a subgraph of listed companies by replacing intermediate nodes and edges with edges that capture the shortest path length between listed companies. This shortest-path approach ensures robustness to the highly local structure between companies, explored in our Graph Fourier Analysis, which captures the structure of transactions rather than the economic substance of those transactions which serve to propagate risk.

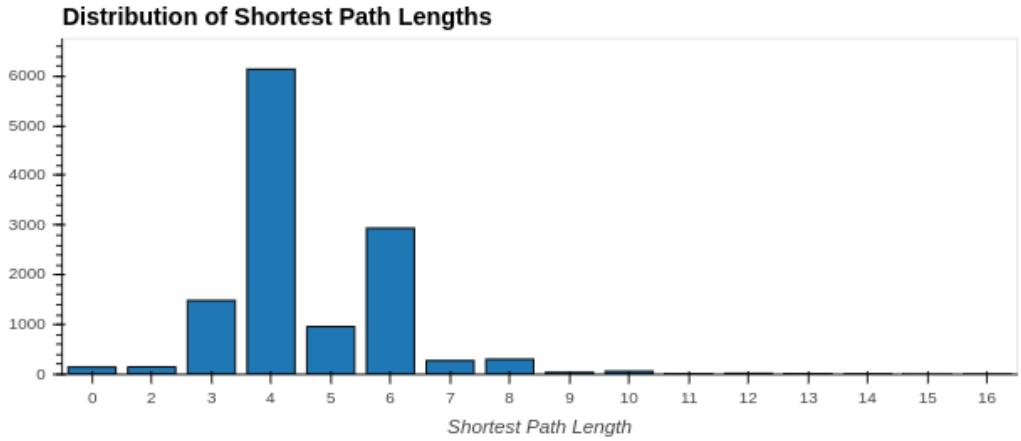


Figure 14: Histogram of the shortest path lengths between all matched listed companies in the ICIJ Graph. From this diagram, we see our most frequent shortest path length at a shortest path length of four. This we believe may be some properties of how matched listed companies have arranged themselves, which may suggest match listed companies separate themselves by some common intermediate structure which comprises non-matched listed companies with particular attributes. This is discussed further in alongside our Graph Fourier Analysis results in Section 4.2.

Using Dijkstra's algorithm, our approach ensures our projected graph is robust to this irrelevant local structure which may be falsely perceived as adding redundancy in the path between listed firms. This leads to a denser projected graph, as many nodes are reachable through other listed companies. While this is a challenge to our analysis, this is easily controlled for through either thresholding, the use of some kernel weighting function or the application of a backboning technique with which to extract the most significant edges, as has been explored in the literature (Coscia and Neffke, 2017; Neal, 2014; Serrano, Boguná and Vespignani, 2009).

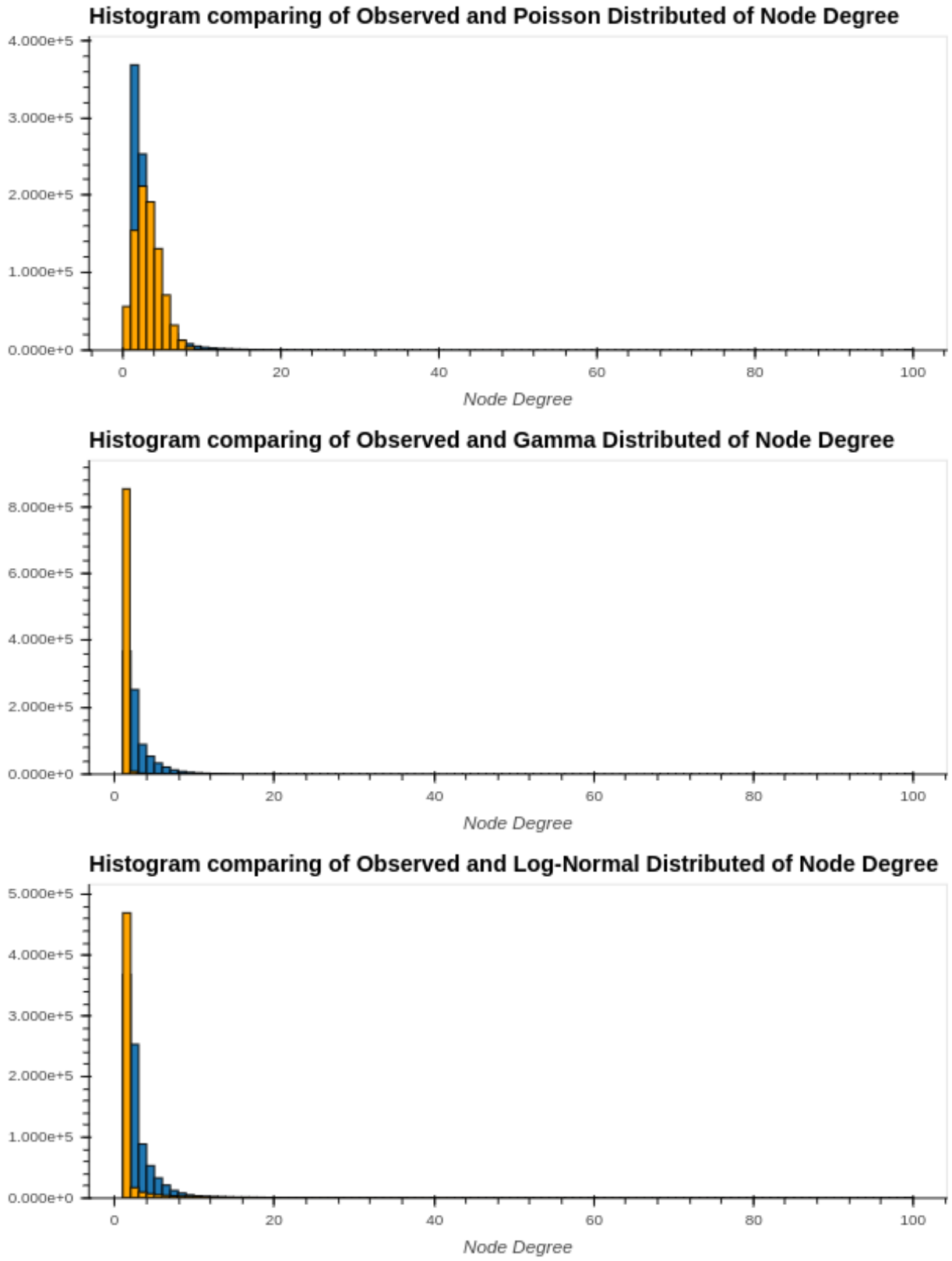


Figure 15: Comparison of possible node degree distributions estimated using maximum likelihood, shown in orange, shown against our observed degree distribution in blue. These estimates are provided in the top-most facet for the Poisson distribution, middle facet for the Gamma distribution and Log-normal distribution for the bottom-most facet. These graphs look to explore certain theoretical expectations of node degree in random graphs. From these comparison, while the Poisson distribution appears the most appropriate fit to our data, we observe many more nodes in our data with a degree less that 4. This suggests our graph may poorly mirror many properties of Erdős-Rényi and other random graphs and bear important local and global structure which describes the avoidance strategies explored in Garcia Alvarado and Mandel (2019). These plots are provided further discussion in Section 4.3 alongside expectations concerning the distribution of shortest path distances.

To explore our shortest path length approach, we analyze the distribution of these path lengths, shown in Figure 16, against some expectation to analyze possible structure in our graph. In Bauckhage, Kersting and Rastegarpanah (2013), authors explore the suitability of different distributions in modelling these shortest path lengths across Erdős-Rényi, Barabasi-Albert, power-law and log-normally distributed random graphs. In their work, authors derive and contrast proposals from various authors on the Gamma and Log-normal distributions with the Weibull distribution (Vazquez, 2006; Capocelli and Ricciardi, 1972). Under simulation, using Kullback–Leibler divergence, authors demonstrate the obvious suitability of the Weibull distribution when modelling shortest path lengths on Erdős-Rényi, Barabasi-Albert and log-normally distributed random graphs, with a strong argument for the application of the Gamma distribution to power-law graphs. An important caveat in this analysis offered by authors was use of the continuous Weibull distribution in modelling path lengths. Bauckhage et al. (2013) support this decision based on the results of their simulation as well as a remark on convenience of this distribution in statistical reasoning and inference. While we believe local structure exists in our graph, looking to our degree distribution, shown in Figure 15, we see strong suitability in the Poisson distribution when modelling the degree distribution of our graph. Comparing these findings to the graphs explore in Bauckhage et al. (2013), this result suggests the suitability of the Weibull distribution in modelling our graphs shortest path lengths, given an expectation that our graph may closely resemble globally properties of the Erdős-Rényi random graph.

Under such an assumption, in Figure 16, we illustrate the shape of our Weibull Distribution fit, using maximum likelihood, to our computed shortest paths between our matched listed companies. From this distribution, we find an expected shortest path length of 4.52, a median of 3.13 and a standard deviation of 4.52. This expected path length compares with a mean shortest path length of 4.66 observed among our match listed companies, which we believe serves as motivation for our distributional assumptions.

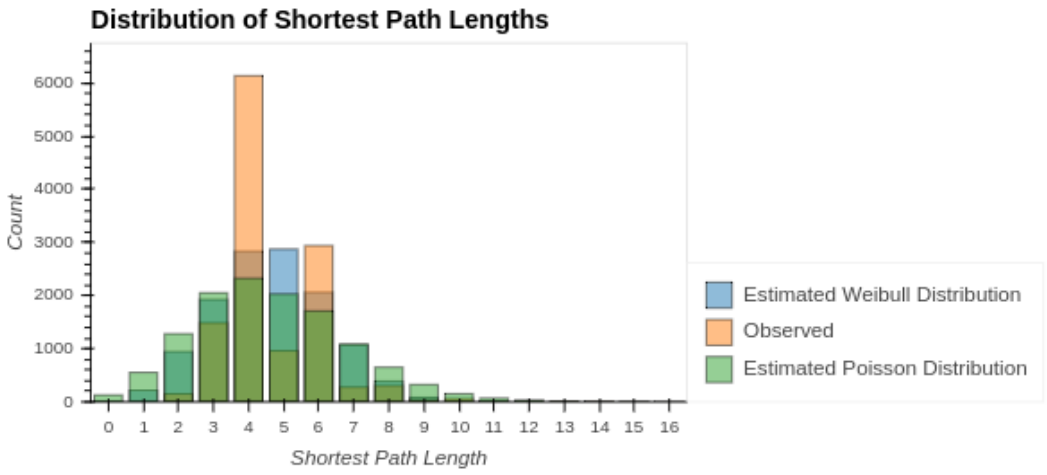


Figure 16: Probably Density and Mass Functions of the Weibull and Poisson Distributions estimated against the shortest path lengths between all matched listed companies using Maximum Likelihood. In the diagram, the estimated Weibull distribution is shown in blue and estimated Poisson distribution is shown in green, against the observed frequencies shown in orange. In this diagram we show the uncharacteristically common propensities in observing shortest path lengths of four and fix compared to the expectations presented by Bauckhage et al. (2013). This, we believe, is an important stylized fact which we believe characteristic of some common intermediate structure of firms separating matched listed companies in our graph. This finding is discussed further in Section 4.4.

Contrasting our shortest path lengths, in Figure 16, we observe unique and regular gaps and spikes at path lengths four and six. These spikes do not align with our expectations for Weibull distributed shortest path lengths on random graphs or the Poisson distribution used to model random path lengths. We believe this unique artefact of bimodality may again be indicative of the common tax avoidance or compliance structures explored in our Graph Fourier Analysis, which may suggest across our 183 matched listed companies a regular separating pattern exists in our data. While this approach is imperfect, it does provide an interesting insight into the structures surrounding our matched listed companies.

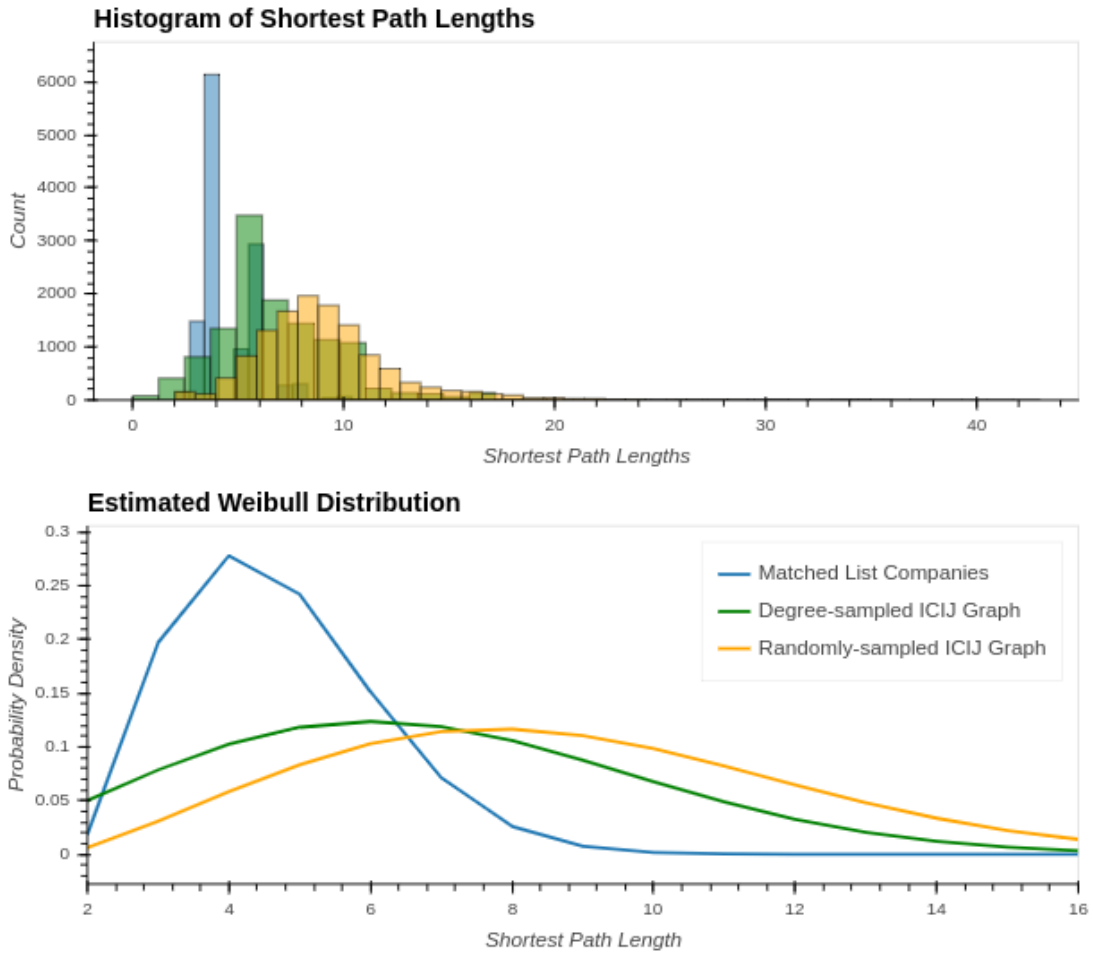


Figure 17: Comparison of shortest path lengths between our sample of matched listed companies, in orange, when compared to the those found across the complete ICIJ graph, in blue. In the top facet, we show a histograms of these shortest path lengths, with the probability in the lower facet. From these plots we observe that Matched Listed Companies exhibit shorter average shortest path lengths compared to those found randomly across the graph. This provides further argument to the discussions in Sections 4.2 and 4.3 that Matched Listed Company may occupy some close central region of the graph.

To assess the impact of company selection on shortest path lengths, we contrast the distribution of these shortest path lengths between matched listed companies and randomly sampled nodes across the graph. In Figure 17, we show a histogram of the shortest path lengths comparing these shortest lengths between matched listed companies, in blue, to degree proportionate and uniformly sampled random nodes, in orange and green. Alongside this diagram, we show the probability density function of the Weibull distribution estimated through maximum likelihood. Based on this diagram we see a far lower shortest path-length for matched listed companies when compared to the population of nodes in our graph. Based on the leftward shift in our average shortest path length between our uniformly sampled nodes and degree-weighted sampling, in orange and green, we can assume degree-based centrality may have an impact on shortest path lengths. This may suggest that against certain measures of centrality, our matched listed companies are more central than the average node in our graph which may impact how risk propagates along our graph and our capacity to impute or infer prices for non-listed companies, assuming centrality or path length has an impact on return in our graph. This may also imply that our matched listed companies may in-fact occupy some specific region of the graph in which they may find themselves more densely packed.

	Matched Listed Companies	non-Matched Listed Companies
count	183.000000	867598.000000
mean	6.180328	2.982588
std	15.603439	43.945201
min	1.000000	1.000000
25%	2.000000	1.000000
50%	2.000000	2.000000
75%	3.000000	3.000000
max	177.000000	35359.000000

Table 4: Summary statistics comparing node degree between matches listed companies and non-matched listed companies.From this table, we observe that Matched Listed Companies have a far higher node degree than non-Matched Listed Companies. This suggests that Matched Listed Companies are connected to more entities in the graph and provide further evidence to Figure 17, discussed in Section 4.3, on the centrality and position of Matched Listed Companies in the Paradise Papers graph.

Looking to the summary statistics computed in tables 4 and 5, we can see compared to non-matched listed companies, matched listed companies appear to have higher average degree counts with many orders of magnitude higher median eigenvector centrality. This suggests that matched listed companies are likely more central to the graph than most observed nodes. This provides valuable insights on the market dynamics which may be observed on our graph, with matched listed companies occupying positions of greater power or influence in markets (Bonacich, 1987).

	Matched Listed Companies	non-Matched Listed Companies
count	1.830000e+02	8.675980e+05
mean	1.433792e-04	1.535513e-04
median	0	0
std	7.209139e-04	1.062505e-03
min	0	0
25%	0	0
50%	0	0
75%	0	0
max	3.748343e-03	7.093683e-01

Table 5: Summary statistics comparing eigenvector centrality between matches listed companies and non-matched listed companies. This table builds on the evidence from Table 4 and Figure 17, to suggest based on the median eigenvector centrality of our Matched Listed Companies that Matched Listed Companies occupy a denser and more central position in our graph. This may suggest, based on our centrality-based interprettation of spatial models that these Matched Listed Companies may bear great influence and exposure to social network effect in the graph and based on our discussion Section 2.5, bear an important position of power in the market define by our graph.

Across the data leaks, the ICIJ describes these named-entities as either transaction intermediaries, such as law firms or bank, addresses or incorporated entities and their legal officers. Using the OpenCorperates database, the ICIJ provide rich information on the jurisdiction and incorporation date of these named-entities, along with data sources (*OpenCorporates: The Open Database Of The Corporate World*, n.d.; International Consortium of Investigative Journalists, n.d.*a*).

In order to investigate these path lengths in more detail, we characterize these shortest paths between listed companies according to the nationality of intermediate entities. We explore our common shortest path lengths of path length four and six, shown in figures 18 and 19. From these figures, we identify common paths between United States matched listed companies through Bermuda and Cayman Island intermediaries. This finding closely follows international business corporation (IBC) structures discussed in Krys (2016) and publicized in Johnston (2002), which we believe provides a further argument for regularity in local structure between matched listed companies.

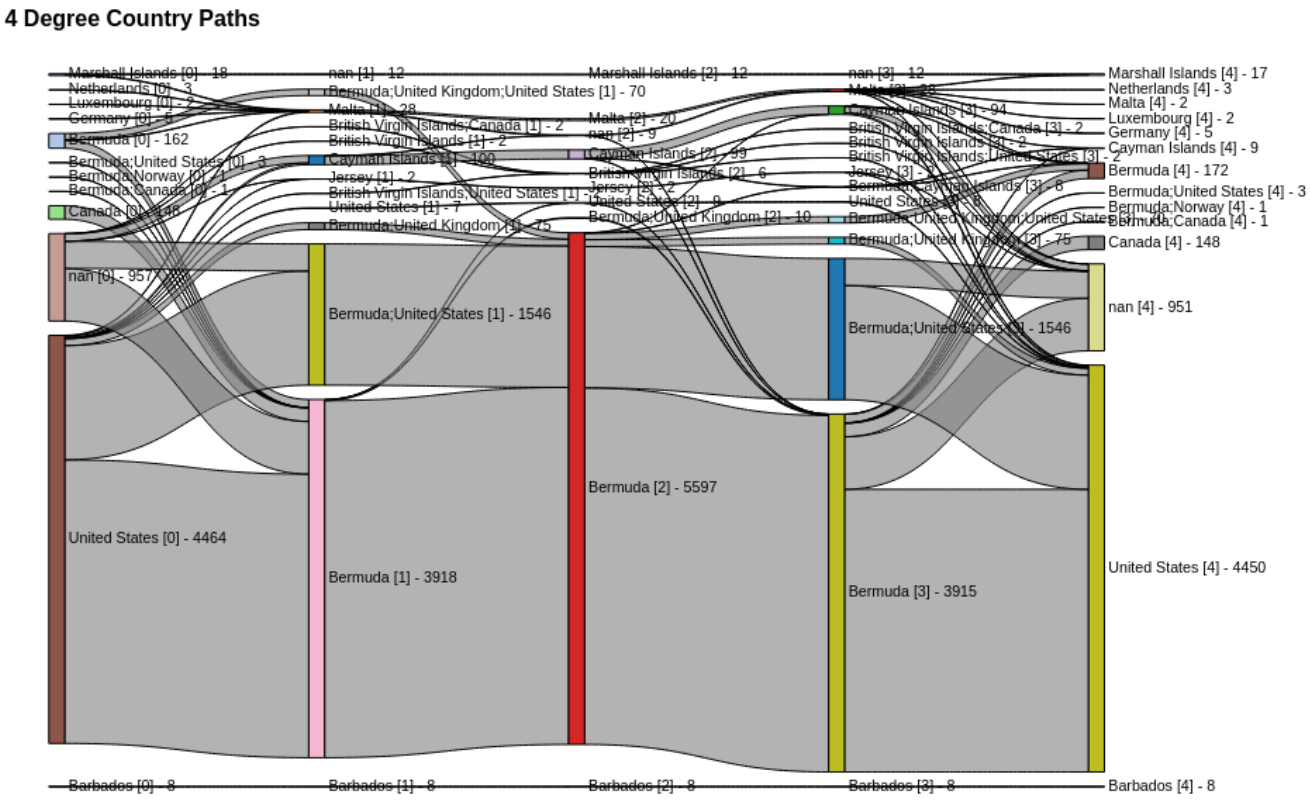


Figure 18: Sankey Diagram characterizing shortest path lengths between listed companies by entity nationality for path lengths of four edges. This diagram shows the most frequent passage of paths based on their shortest path length of four passing from the United States through Bermuda to the United States. This finding mirrors discussions by Krys (2016) and publicized in Johnston (2002) on common structures and legislative changes which have driven the adoption of particular international business corporation (IBC) structures for compliance, tax avoidance and privacy purposes. These findings are discussed alongside our results on centrality, node degree and Graph Fourier Analysis in Section 4.3 and mirror expectation we share concerning common intermediate structure between Matched Listed Companies.

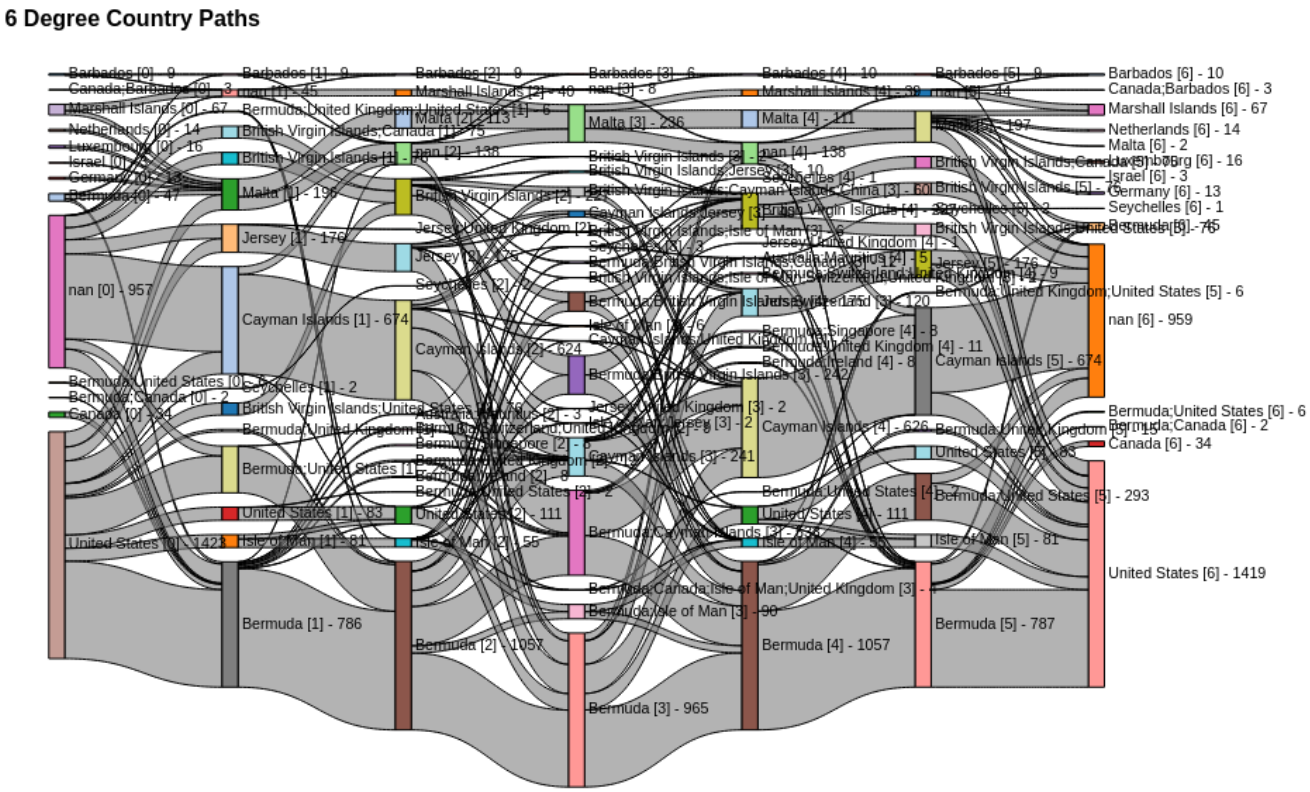


Figure 19: Sankey Diagram characterizing shortest path lengths between listed companies by entity nationality for path lengths of six edges. This diagram shows the most frequent passage based on their shortest path length comprising six nodes passing from the United States through Bermuda to the United States. This diagram follows the results of Figure 17 which provides readers a distribution observed shortest path lengths. Building on our findings in Figure 18, we again observe common paths through Bermuda which mirror common structures discussed in Krys (2016) and publicized in Johnston (2002). We believe this finding to support the idea that common intermediate structures are used by company for compliance, tax avoidance and privacy purposes, which lie in contradiction to our expectations for shortest path lengths under many common random graph generating functions (Bauckhage et al., 2013).

4.4 Properties of Path Lengths

Using a new dense graph of matched listed companies separated by edge attributes of shortest path lengths, we must now represent these edge attributes as weighted edges through which to describe a weighted graph representing the relationships between our matched listed companies. To identify these weighted edges, we must define a kernel weighting function which expresses the probability that two matched listed companies engaged in a transaction of economic substance, as a function of the shortest path length between these companies. This ensures that our graph has sufficient sparsity for our analysis while appropriately modelling the transfer of risk between entities across our graph.

$$f(x, c) = cx^{c-1}e^{-x^c} \tag{23.1}$$

Exhibit 23: The probability density function for the Weibull Minimum Extreme Value distribution with shape paramter c (Weibull, 1951). This probability density function is used as a spatial weighting kernel based on motivation from Bauckhage et al. (2013) and Katzav, Nitzan, ben Avraham, Krapivsky, Kühn, Ross and Biham (2015) on its use in modelling shortest path lengths across graphs. This function serves purpose to the Gaussian and bi-square weighting kernels used in spatial econometrics and spatial statistics and the observation by Hammersley (1950) concerning the distribution of pairwise Euclidean distances on a hypersphere. This distribution is used in Sections 4.3 and 4.4 in exploring and modelling observed shortest path lengths in our data.

While the application of Gaussian and bi-square weighting kernels remain common approaches in spatial statistics motivated by Hammersley (1950) with which to move from a distance metric between observations to a valid weighting matrix; work by Bauckhage et al. (2013) and Katzav et al. (2015) motivates the use of a Weibull weighting kernel, shown in the equation in Exhibit 23, through which to express the probability that two companies are engaged in a transaction of economic substance at some distance given our chosen data and analysis of shortest path length. We set the parameters

of this kernel through maximum likelihood estimation of our sample of shortest path lengths between matched listed companies. Using this discussion, we establish a graph comprising our matched listed companies whose weighted edges describe a Weibull kernel weighting of the shortest path length found in our original ICIJ graph.

In order to further our understanding of how price impact propagates on our Weibull-weighted shortest path length graph, we look to Graph Fourier Analysis to identify and characterize dominant frequencies in our markets response. Looking to our graph in Figure 20, we observe a definite cluster of eigenvalues around 9 on the x-axis, suggesting tightly packed frequencies informed by our graph structure. By analyzing the magnitude of these corresponding eigenvalues, we fail to observe any dominant frequencies as seen before in Figure 12. Looking to the position of our highest magnitude frequencies at eigenvalues 6.439 and 9.72, we may expect to find important intermediate structure in our model with price impact correlated among particular clusters of tightly connected companies Ortega, Frossard, Kovačević, Moura and Vandergheynst (2018). This may present a unique finding that price impact is felt along particular supply chains rather than globally across our graph.

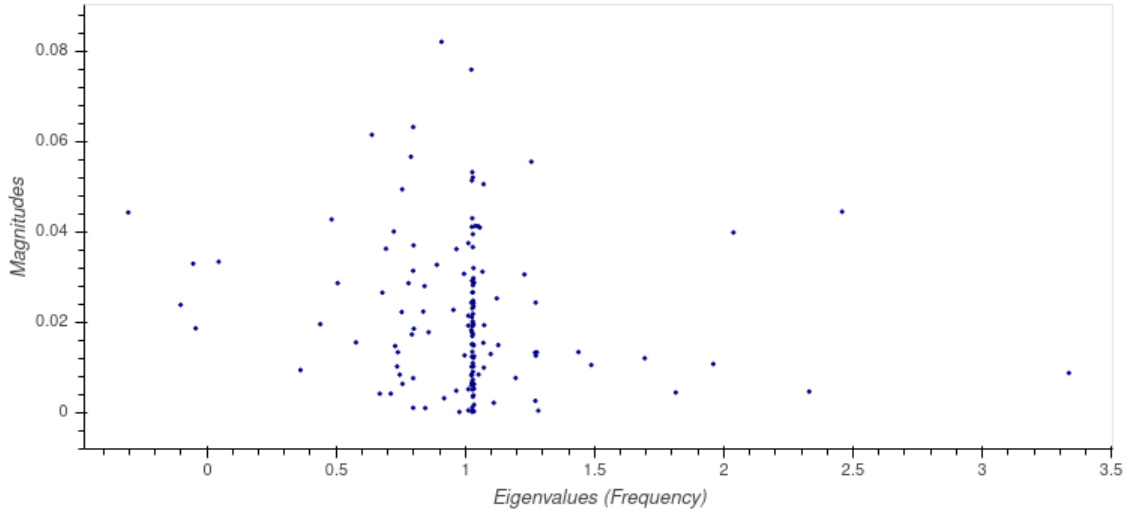


Figure 20: Graph Fourier Transform of excess returns computed from our shortest path distances scaled using a weibull weighting kernel with parameters estimated using maximum likelihood. Compared with figures 13 and 12 we see a strong concentration of frequencies at around 1. This concentration may be indicative of many small-world properties of our new graph, comprising a core dense structure, surrounded by loosely connected clusters of entities with small shortest path lengths between entities (Watts and Strogatz, 1998). This is discussed further against the properties of these path lengths in Section 4.4. This figure can be recovered using the `returns_weibull_gft` function provided in our project documentation (Gawronsky et al., 2020*c*).

In Figure 21, we plot the top six highest magnitude eigenvalues of our Graph Fourier Transform over our Kamada–Kawai graph layout. From this graph we observe important structural variation along our graph for our component at $\lambda = 0.046$, shown in the bottom-left facet. This represents one of our lowest frequencies which appear to visually separate our core dense clique, highlighted in this component in red, from other neighbourhoods, highlighted in this component in blue, with connecting entities highlighted in green.

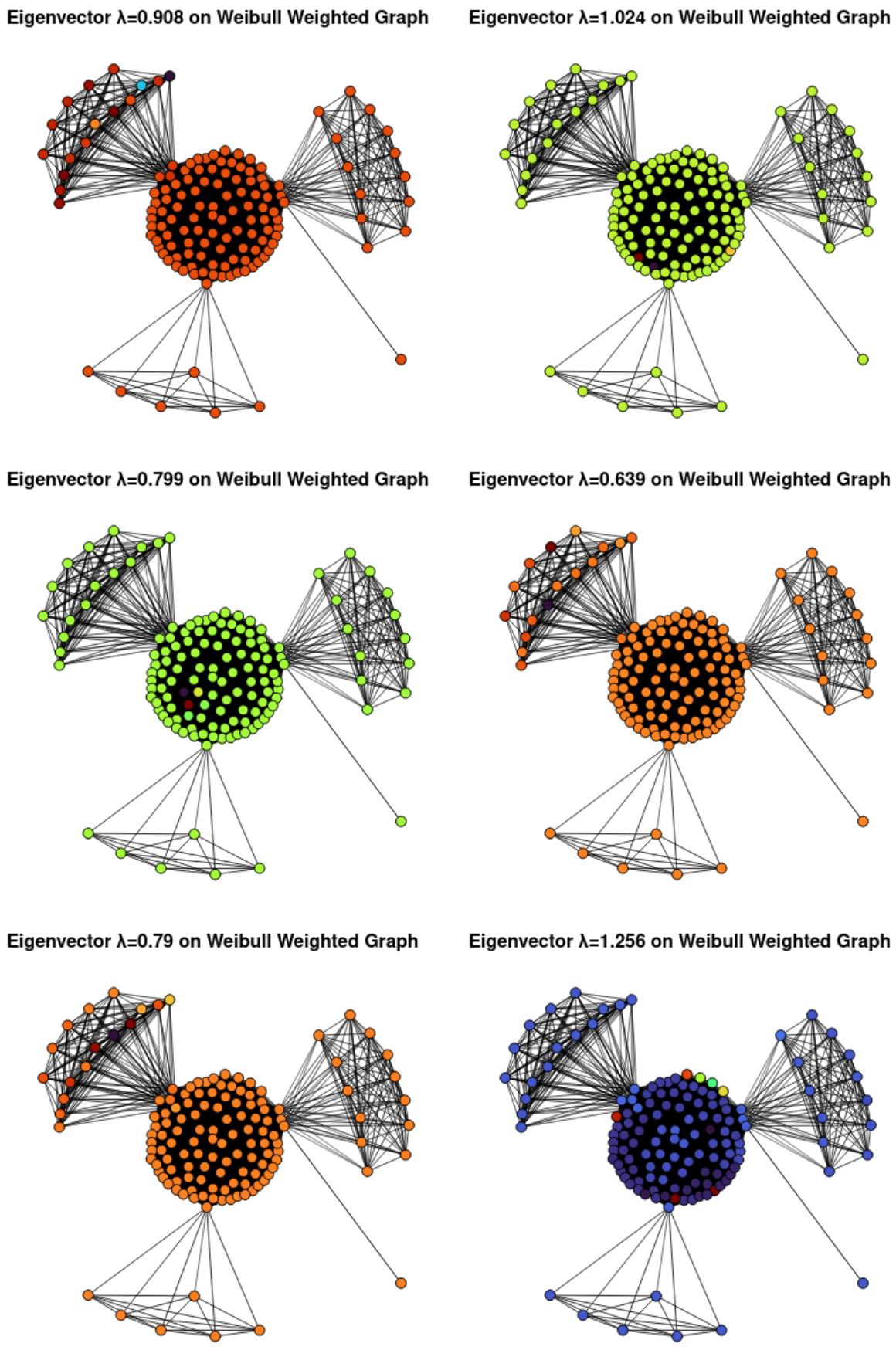


Figure 21: Layout of Weibull Weighted Graph using the Kamada–Kawai algorithm on which we plot the Inverse Graph Fourier Transform of the top six highest magnitude eigenvalues. Using the Turbo color mapping, we represent negative component values in red, with positive component values in blue (Mikhailov, 2019). Across these six facets the dominance of large eigenvectors among our largest magnitude components suggests price movements may propagate within local neighbourhoods of our graph. This observation is discussed further under our exploration on the properties of weighted shortest path lengths in Section 4.4. These figures can be recovered using the `returns_weibull_gft` function provided in our project documentation (Gawronsky et al., 2020*c*).

4.5 Discussion

From our exploratory discussion, we have looked to build grounding assumptions concerning the structure and representation of our graph. By analyzing the local structure of matched listed companies, alongside summary statistics on degree, eigenvector centrality and shortest path length, we have looked to detail early evidence for spatial correlation across our graph, as well as methods through which to assess and explore graph censorship. From this work, we have identified early evidence for selection bias in Profit Margins along our graph, along with evidence concerning the impact of centrality on market response. From this work, we look to explore further properties of spatial correlation along our graph, as they relate to returns and known stylized facts. Using this analysis, we look to explore the application and extension of Social Network Econometrics into Spatial-APT. Using this extension, we aim to provide common methodologies through which to explore similar such market events and datasets which describe possible contagions over censored graphs.

5 Methodology

5.1 Spatial Model Selection Procedure

Our methodology will look to establish a framework through which to navigate the universe of factors, spatial weighting matrices, models and estimating procedures to identify a plausible description of our data generating processes as characterized by some spatial or non-spatial linear model. The preference of our study will be to rely on the bottom-up selection approach offered in Mur and Angulo (2009). Given the limited guidance from this procedure on the selection of Kelejian-Pruscha and SDEM, we will explore these models following evidence from SEM and SLX models. We rely on the Robust Lagrangian Multiplier Tests offered by Penghui, Lihu and Zhengming (2015) to guide our analysis on model selection, as is common across LeSage and Pace (2009), Elhorst (2010) and Florax et al. (2003). To divorce our exploration of spatial correlation from the particular forms of the SAR and SEM models, we will further test for spatial correlation under Moran's I statistic, provided in Exhibit 24.

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{24.1}$$

Exhibit 24: Formula for Moran's I spatial correlation coefficient as derived in Moran (1950) with spatial weight, w_{ij}, and variable, x_i, and sample mean, $\bar{x}$. W is a sum over all w_{ij} and N represents the count of all i. This statistic forms part of our testing procedure for spatial dependence applied across Sections 6.1 and 6.2 in tables 7 and 10.

In order to explore the firm characteristics included in our models, we have opted to present the reader with a cross-section of models with which to compare the impacts of backwards feature selection based on the statistical significance of our chosen features. While this greedy approach is naive given published comparisons in genetic search, we believe this decision serves to best constrain our analysis given our universe of spatial models and feature sets while ensuring comparability across unsaturated models in our study (Vafaie and Imam, 1994).

5.2 Characteristics

Our chosen characteristics are partly inspired by Ahern (2013), to which we include Price-to-Research as a possible factor. Following discussions by Karkinsky and Riedel (2012) on the use of offshore structures in managing the location of company patents for tax avoidance purposes, we expect companies whose competitiveness is heavily reliant on royalties to be both disproportionately affected by the leaks and uniquely positioned in our graph. To capture the effects of firm centrality, we include the eigenvector centrality of our matched listed companies in our full ICIJ graph. This may control for our assumptions in graph projection and provide insight into the structure and role of firms in our original graph. Given our intention to fit spatial error models against our chosen spatial matrix, we believe measures of centrality against our projected graph is redundant under our centrality interpretations of spatial lag models.

While Ahern (2013) rely on a double-sorting procedure with which to construct factor-mimicking portfolios for their analysis, we simplify our approach given the size and scope of our cross-sectional event study to model our firm-characteristics directly as explanatory variables (Daniel and Titman, 1997). We believe this most appropriate given the reliability with which we would be able to establish these factor-mimicking portfolios and their variability across our event window.

5.3 Diagnostic Tests

While obvious concerns may exist regarding the impacts of multicollinearity on identification in our models, we investigate these effects by exploring both the statistical significance of pairwise correlation using Pearson's correlation coefficient and the Variable Inflation Factor computed against our explanatory variables, shown in the equation in Exhibit 25 (Akinwande, Dikko, Samson et al., 2015).

$$\boldsymbol{x}_i = \alpha_0 + \sum_{k \neq i}^{K} \alpha_k \boldsymbol{x}_k + \boldsymbol{\epsilon}; \tag{27.1}$$

$$\text{VIF}_i = \frac{1}{1 - R_i^2} \tag{27.2}$$

Exhibit 25: Formula for computing the Variable Inflation Factor on a variable with R_i^2 representing the coefficient of determination in the first step of the estimating procedure. This forms part of our testing procedure in identifying the impacts of multicollinearity applied across Sections 6.1 and 6.2 in tables 7 and 10.

To validate our assumptions in ordinary least squares multiple regression analysis, we will rely on the Jarque-Bera test provided in Bera and Jarque (1981) and Jarque and Bera (1980), shown in the equations in Exhibit 22, to test for normality in our residuals. While Thadewald and Büning (2007) share concerns over the power of the Jarque-Bera test for distributions with short tails, we find other tests like the Shapiro–Wilk and Kolmogorov-Smirnov tests highly sensitive to our choice of winsorization and hence preclude their application in our study. To test for heteroskedasticity in our residuals, we rely on the Breusch-Pagan, Koenker-Basset and White tests, shown in the equations in Exhibit 26. While the White test remains popular in testing non-linear heteroskedasticity, the concerns raised in MacKinnon and White (1985) over statistical power in small in high dimensional samples strongly motivates comparisons through linear tests for heteroskedasticity. While the Breusch-Pagan test remains popular in many econometric studies, concerns over the sensitivity of this test to departures from normality warrants comparison to the Koenker-Basset test to best conclude on the homoskedasticity of our errors (Koenker and Bassett Jr, 1982).

$$\hat{\boldsymbol{\epsilon}_{\boldsymbol{i}}}^{\boldsymbol{2}} = \alpha_0 + \alpha_1 \boldsymbol{x}_{1,p} + ... + \alpha_p \boldsymbol{x}_{i,p} + \alpha_{p+1} \boldsymbol{x}_{i,p}^2 + ... + \alpha_{2p+1} \boldsymbol{x}_{i,1} \boldsymbol{x}_{i,2} + ... + \boldsymbol{u_i} \tag{28.1}$$

$$LM = nR^2 \tag{28.2}$$

Exhibit 26: Formula for computing the Lagrange multiplier (LM) test statistic for White's test with n our sample size and R_i^2 representing the coefficient of determination in the first step of the estimating procedure in which we estimate our errors against the original regressors of our model along with their squares and cross-products. This statistics tests the null hypothesis that no non-linear heteroskedasticity is present and is evaluated against the χ^2 distribution, with $P - 1$ degrees of freedom equal, where P is the number of parameters in our auxiliary regression model. This test is similar to the Breusch-Pagan test, where squares and cross-products terms are excluded from the first step of our analysis.

5.4 Spatial Weighting Matrix

Our analysis will rely on a fixed spatial weighting kernel for use in spatial modelling, based on the interpretability of this approach, the comparability of this method across models and its impact on statistical power (Fotheringham, Brunsdon and Charlton, 2003). Following our discussion in Section 4.4, our study relies on the Weibull Minimum Extreme Value distribution through which to model our spatial interaction. In order to determine the shape parameter of this distribution, we rely on maximum likelihood estimation. This estimating procedure is performed against our shortest paths between matched listed companies to produce a spatial weighting matrix to represent the probability density of a particular path length between firms. This spatial weighting matrix is then row-normalized for use in our models, as is common spatial regression analysis.

5.5 Estimating Procedure for Spatial Modelling

While our study bears no particular preference to model estimating procedures in spatial regression, evidence from Kelejian and Prucha (1999), Anwar, Djuraidah and Wigena (2020) and Egger, Larch, Pfaffermayr and Walde (2009) on the robustness and efficiency of Generalized Method of Moment Estimators (GMM) presents strong evidence for their application in spatial regression model estimation. Given their many extensions in modelling spatial heterogeneity in Anselin (2011), Arraiz, Drukker, Kelejian and Prucha (2010) and Drukker, Egger and Prucha (2013), and simulation results in Kelejian and Prucha (1999), we will look to GMM estimators as required by our selection procedure. Given the variety of these estimating procedures, we will look to detail their methods though out our results so as to limit the breadth of techniques we are require from the reader.

6 Findings

6.1 Non-spatial Models

We will start our exploration of non-spatial models by analyzing the pairwise Pearson correlations, shown in Table 6, between our chosen characteristics of Price-to-Earnings, Market Capitalization, Profit-Margin, Price-to-Research and ICIJ (Eigenvector) Centrality across our graph. While we find many characteristics in our study statistically uncorrelated at some cutoff, we do find Profit-Margin positively statistically correlated with Price-to-Earnings and Market Capitalization at the 5%-level of significance. While all these characteristics may be important to our model, this correlation may present challenges in identifiability which we must be aware of. These findings may echo work by Kelly et al. (2013) exploring the relationships between size, centrality and market position in network models. Interestingly our ICIJ Centrality characteristic seems negatively correlated with many of our features, though statistically insignificant. This may suggest orthogonality to our other features, though we note a positive correlation of 0.0032 with Profit-Margin, which may be expected based on Chen et al. (2018) and Corbo et al. (2006).

Pearson Correlation	Price-to-Earnings	Market Capitalization	Profit-Margin	Price-to-Research	ICIJ Centrality	$r_m - r_f$
Price-to-Earnings						
Market Capitalization	0.0922 (0.2683)					
Profit-Margin	0.3431 (0.0)	0.1895 (0.0219)				
Price-to-Research	0.0394 (0.6372)	0.0067 (0.9356)	0.012 (0.8854)			
ICIJ Centrality	-0.0111 (0.8941)	-0.0662 (0.4276)	0.0032 (0.9694)	-0.0198 (0.8124)		
$r_m - r_f$	0.046 (0.5813)	0.0706 (0.3974)	0.0563 (0.4996)	-0.0117 (0.8888)	0.0343 (0.6807)	

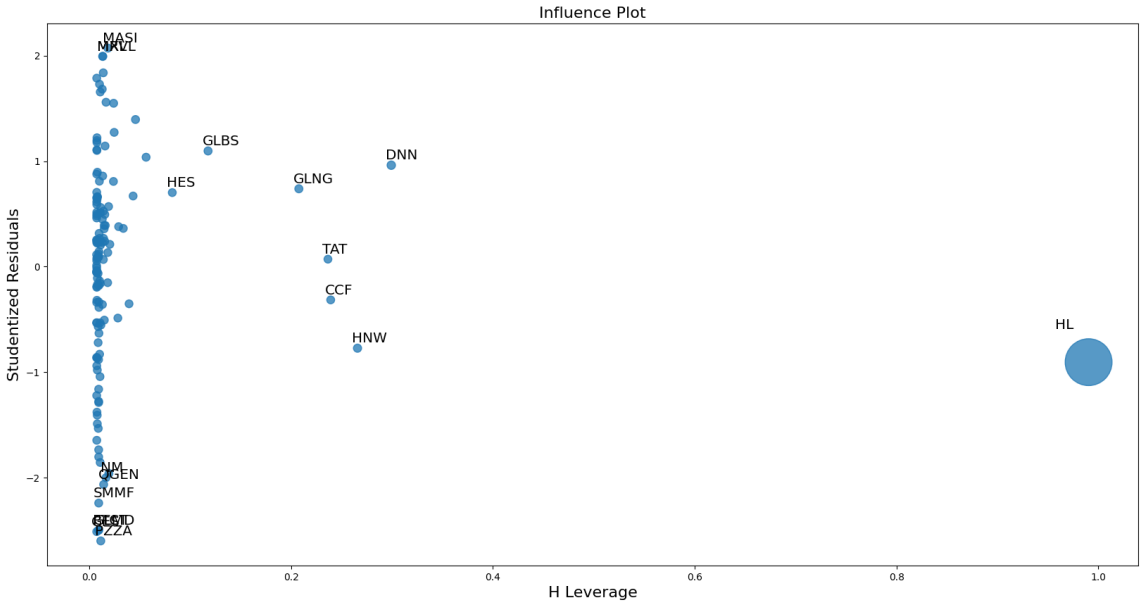
Table 6: Table of Pearson Correlation Coefficients with statistical p-values in brackets below. From this table we can observe the sign of the correlation between the variables alongside the statistical significance of these observed correlations. Looking across our exogenous variables, we observe our significant correlation between Profit-Margin, Price-to-Earning and Market Capitalization. These correlations are positive and may suggest with large Price-to-Earnings and Market Capitalizations are more profitable. This may bear concern for our analysis in including all these variable due to concern over multicollinearity which may skew our estimates. This we note in our analysis in Section 6.1, where we explore possible variable inflation in our results. This table can be recovered using the `pearson_corr` function provided in our project documentation (Gawronsky et al., 2020*c*).

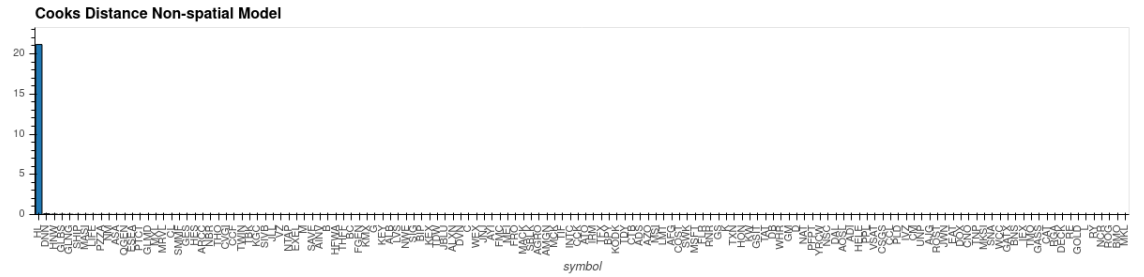
To explore our space of non-spatial models, we compare parameter estimates across six least squares regression models. At each new model, shown in Table 7, we eliminate our least significant non-zero parameter estimates, leaving market risk premia across our comparisons. To verify our assumptions across models, we compute the Variable Inflation Factors across our chosen characteristics to analyse the possible impacts of multicollinearity on parameter identification. To verify characteristics in our residuals to present Jarque-Bera, Breusch-Pagan, Koenker-Bassett and White test statistics to diagnose possible skewness, overdispersion and heteroskedasticity in our residuals. Alongside these tests, we compute Moran's I spatial correlation coefficients and perform Robust Lagrangian Multiplier tests to identify evidence for spatial correlation as per our procedure in Mur and Angulo (2009). These statistics presented alongside computed p-values estimated against their respective parameterised distributions. Across models, we test the joint significance of our parameter estimates using Fisher's F-statistic, which we will use to compare models.

Variable Inflation Factor	(1)	(2)	(3)	(4)*	(5)	(6)
Price-to-Earnings	1.1366					
Market Capitalization	1.0433	1.0426				
Centrality	1.0061	1.0061	1.0016			
Price-to-Research	1.0005	1.0005	1.0003	1.0002		
Profit-Margin	1.1673	1.0366	1.0036	1.0034	1.0034	
$r_m - r_f$	1.0091	1.0086	1.0049	1.0036	1.0034	1.0000
Coefficient Estimates						
Price-to-Earnings	6.789e-04					
	(0.641)					
Market Capitalization	1.847e-14	1.886e-14				
	(0.566)	(0.556)				
Centrality	-3.338e+00	-3.343e+00	-3.459e+00			
	(0.262)	(0.259)	(0.241)			
Price-to-Research	4.988e-06	4.983e-06	4.951e-06	4.994e-06		
	(0.170)	(0.169)	(0.171)	(0.168)		
Profit-Margin	-2.092e-02	-1.959e-02	-1.876e-02	-1.863e-02	-1.861e-02	
	(0.015)	(0.015)	(0.018)	(0.019)	(0.019)	
$r_m - r_f$	-2.814e-01	-2.767e-01	-2.612e-01	-2.796e-01	-2.88e-01	-3.481e-01
	(0.520)	(0.526)	(0.548)	(0.520)	(0.509)	(0.432)
α	6.570e-03	7.051e-03	7.312e-03	6.723e-03	7.197e-03	1.074e-03
	(0.076)	(0.047)	(0.038)	(0.054)	(0.039)	(0.643)
Overall significance						
F-statistic	1.678	1.982	2.402	2.733	3.117	0.621
	(0.131)	(0.085)	(0.053)	(0.046)	(0.047)	(0.432)
Test for Normality of Errors						
Jarque-Bera	6.265	5.338	6.078	5.117	5.253	2.399
	(0.044)	(0.069)	(0.048)	(0.077)	(0.072)	(0.301)
Diagnostic Tests for Heteroskedasticity						
Breusch-Pagan	6.396	4.243	3.381	2.387	2.007	0.037
	(0.380)	(0.515)	(0.496)	(0.496)	(0.367)	(0.848)
Koenker-Basset	5.69	3.865	3.07	2.249	1.914	0.038
	(0.459)	(0.569)	(0.546)	(0.522)	(0.384)	(0.845)
White	42.873	26.536	12.733	12.096	10.331	7.959
	(0.027)	(0.149)	(0.548)	(0.208)	(0.066)	(0.019)
Diagnostic Tests for Spatial Dependence						
Moran's I (error)	1.155	1.221	1.275	1.309	1.363	1.309
	(0.248)	(0.222)	(0.202)	(0.191)	(0.173)	(0.190)
Robust LM (lag)	0.041	0.209	0.331	0.044	0.263	0.115
	(0.839)	(0.647)	(0.565)	(0.834)	(0.608)	(0.734)
Robust LM (error)	0.031	0.183	0.295	0.012	0.160	0.131
	(0.859)	(0.669)	(0.587)	(0.912)	(0.689)	(0.717)

Table 7: Table illustrating estimates of least-squares models without social network effects. Here, we represent the results of our statistical tests in brackets with our test statistics or model coefficients above. The table contrasts six models computed against backwards features selection chosen based on the statistical significance of particular variables in our saturated model. A star (*) is placed against our chosen numbered models to indicate to readers the model which exhibits the greatest joint significance under Fisher's F-statistic. This test explores the null hypothesis that all coefficients are jointly zero and provides an indication to what authors believe to be the most reasonable description of our data generating process across these chosen models. Looking at model (4), while we observe limited statistical significance in the impacts of Price-to-Research in the model, we believe the overall significance of this model and the consistency of its findings with observations from Griffith et al. (2014) and Nabben (2017) on the use of offshore entities in relocating intellectual property warrant its inclusion in further spatial analysis. What we find curious with our estimates is the sign of this exogenous variable. Based on our expectations concerning market reaction, we would expect, given a negative market response, companies with greater relative investments in intellectual properties to have more negative exposure due to possible backlash which alters ones ability to strategically relocate the intangible asset. This finding may suggest more a realisation by the market of rents accruing to firms with large Price-to-Research ratios stemming from their use of offshore services. This table is discussed further and contextualized with our model diagnostics in Section 6.1. This table can be recovered using the `backwards_selection` function provided in our project documentation (Gawronsky et al., 2020*c*).

Looking at Table 7, we observe limited evidence for multicollinearity based on the observation of our Variable Inflation Factors. Typically cutoffs for Variable Inflation Factors are set between 5 and 10, depending on the study (Akinwande et al., 2015). In our analysis, we observe Variable Inflation Factors between 1.0000 and 1.1673, adding confidence to the robustness of our estimates. Correlated exogenous variables with large Variable Inflation Factors may lead to issues in identifiability in particular parameter estimates. Looking at our Jarque-Bera tests, while winsorization plays an important role in the robustness of this statistic, we cannot reject the hypothesis that estimates are not normally distributed. While models **(5)** and **(6)** show signs of non-linear heteroskedasticity, we believe this may present itself plausibly as an artefact in omitting either Paradise Graph Centrality and Price-to-Research from our analysis. An interesting observation in our diagnostic tests for spatial dependence suggests limited evidence for spatial autocorrelation according to our specifications in the spatial lag or spatial error models. While our market risk premia remain statistically insignificant across our specification, we note important dependence across firm Profit Margins. Looking at our joint tests for significance, we observe greatest join significance across models **(3)** and **(4)**. This suggests, given our assumptions, that Price-to-Research and Profit Margin may be critical firm characteristics which have served to influence market response over our event window. Looking at the estimated sign of these characteristics, we observe a negative correlation between our market risk premium and Profit Margins, *cet. par.* This may suggest a reversion by investors to estimates of future discounted free cash-flows based on the leaks given more typical estimates on the coefficients of Profit Margin across other studies (Fama and French, 1993, 2015). While statistically insignificant in our study, the negative sign on Paradise Graph Centrality provides comparisons to O'Donovan et al. (2019) where graph membership was found to have a negative impact on returns over this same event window. This result may suggest that market response is conditionally dependent on eigenvector centrality, rather than simple membership to our graph, and may suggest that estimates on the impact of these leaks may be biased when ignoring the degree of membership captured by centrality.





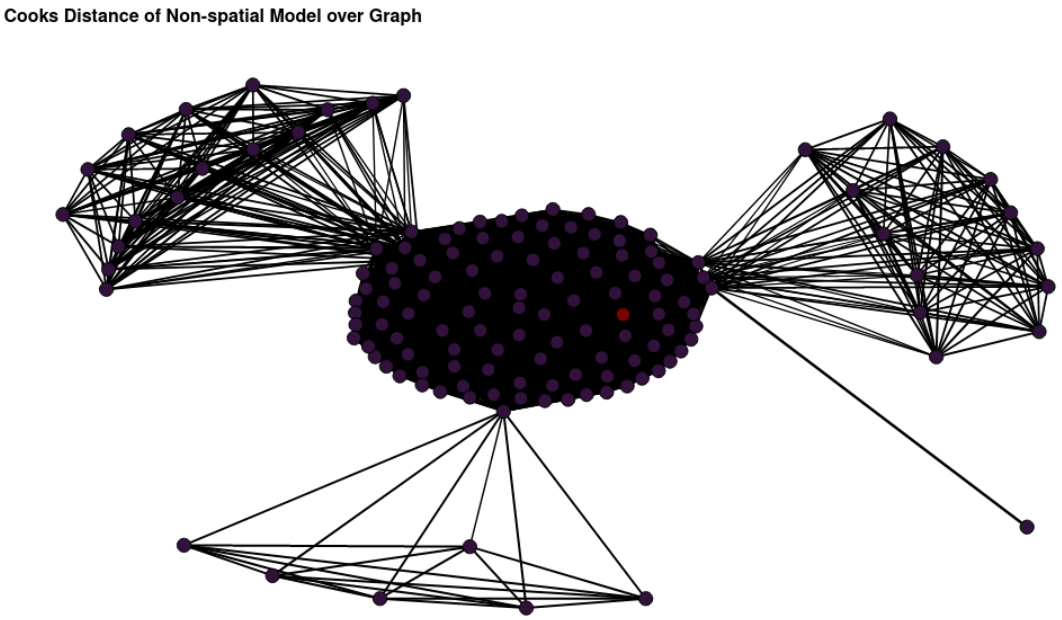


Figure 22: Analysis of leverage of particular data points over residuals and graph layout for non-spatial model. In the top most facet we scatter plot indicating the H leverage and Studentized Residuals found in our SLX regression model. Size is used to indicate the influence of particular observations. In the facet below we provide a bar chart of the Cooks Distances computed across our observation. From this bar chart we can see the large influence of five of our observation of our estimates. In the bottom-most facet, we provide the reader with a layout of our graph to analyze possible structural factors in our graph which may explain this influence. These figures can be recovered using the `get_regression_diagnostics` function provided in our project documentation (Gawronsky et al., 2020*c*).

While our diagnostic tests raise little alarm over our analysis, we note the extreme presence of outliers and their leverage upon our model estimates. While these points exert high leverage, we see little impact from their exclusion on our overall estimates in out analysis and opt not to exclude them in our study. In figures 22, we show three plots detailing the relationship between our standardized residuals, leverage and influence, alongside the distribution of our Cook's Distances, shown in the equation in Exhibit 27, as presented over a Kamada–Kawai layout of our graph. These plots are designed to illustrate and identify structural phenomenon across our points of high leverage. Looking at our graph layout, we observe little evidence, graphically, that points with large Cook's Distances occupy some unique neighbourhood or position in our graph. Looking to the three firms with highest leverage upon our analysis, Denison Mines Corp. (DNN), Globus Maritime Limited (GLBS) and Galmed Pharmaceuticals Ltd (GLMD), we see little reporting over the period of our event study to suggest some event which may have impacted these companies, nor do we observe these companies to bear unique centrality to our ICIJ or projected graphs. While these points bear high leverage over our parameters, we see little change in our spatial estimates after removing even the top ten data-points with the largest cooks distances from our model. Where we see significant changes is after removing the first seventeen data-points after which we observe certain evidence of spatial auto-correlation in our residuals which may suggest the application of the spatial lag or spatial error models- this evidence is limited and should not affect our findings or approach to statistical modelling.

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{y}_{\,j} - \widehat{y}_{\,j(i)} \right)^2}{ps^2}; \tag{29.1}$$

$$s^2 = \frac{\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}}{n-p} \tag{29.2}$$

Exhibit 27: Formula for computing Cook's Distance (D_i), where p the number of feature in our regression model, $\widehat{y}_{\,j(i)}$ represents the fitted response value obtained when excluding i, and s^2 the mean squared errors of our regression model, $\boldsymbol{\epsilon}$. The use of Cook's Distance forms part of our diagnostics testing procedure used in identifying observations with high leverage over our estimates and is applied across Sections 6.1 and 6.2 in figures 22 and 24.

Based on our findings in Table 7, we will continue our spatial model selection procedure based our findings in model **(4)**. This model includes the Price-to-Research, Profit Margin and firm market risk premia as characteristics in our analysis. Based on the results of our Robust Lagrangian Multiplier Tests we will explore a family of spatial lag explanatory models building on the guidance of Mur and Angulo (2009) wherein we investigate the inclusion of Price-to-Research, Profit Margin and firm market risk premia at a first-order spatial lag. This is computed based on the dot product between our spatial weighting matrix and our explanatory variables and will be estimated using the method of least-squares.

6.2 Spatial Lag Explanatory Models

We will begin our exploration of spatial lag explanatory models denoted,

$$\boldsymbol{y} = \alpha \mathbf{1}_{\boldsymbol{N}} + X\boldsymbol{\beta} + WX\boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{2.8}$$

with endogenous variable, $\boldsymbol{y}$, exogenous variables, X, spatial weighting matrix, W, and bias, $\mathbf{1}_{\boldsymbol{N}}$, by analyzing the pairwise Pearson correlations between our explanatory variables and their spatial lags, shown in Table 8. Using this analysis, we aim to identify early evidence of multicollinearity between our explanatory variables which may impact model identification. Compared to the analysis in Section 6.1, these correlation coefficients provide us interesting insight into the local structure of these variables along our projected graph. Comparing Market Capitalization, precluded from our analysis, with its spatial lag we see evidence for a positive spatial correlation, suggesting firms with similar sizes may be closely connected. Interestingly, we observe statistically significant positive correlations between ICIJ Centrality, lagged Market Capitalization, lagged Profit Margin and lagged market risk premia, which may suggest that highly central firms may be connected to companies with high Profit Margins and large Market Capitalizations. While this correlation is strong, more interesting are the comparisons to Table 6, where these effects are deemed statistically insignificant by our analysis when analyzing these non-spatial explanatory variables.

Pearson Correlation	W Price-to-Earnings	W Market Capitalization	W Profit-Margin	W Price-to-Research	W ICIJ Centrality	$W(r_m - r_f)$
Price-to-Earnings	-0.0159 (0.8486)	0.0417 (0.6173)	0.0336 (0.6872)	0.0395 (0.6358)	-0.0063 (0.9395)	0.0061 (0.9416)
Market Capitalization	0.0354 (0.6711)	-0.2527 (0.0021)	-0.0508 (0.543)	0.0084 (0.92)	-0.065 (0.4355)	0.0088 (0.9164)
Profit-Margin	0.0601 (0.4713)	-0.0555 (0.5058)	0.1409 (0.0897)	0.1441 (0.0827)	-0.0068 (0.9353)	-0.1717 (0.0382)
Price-to-Research	0.038 (0.6485)	0.0087 (0.9172)	0.0566 (0.4975)	-0.0768 (0.357)	-0.0179 (0.8298)	-0.0561 (0.5015)
ICIJ Centrality	-0.0816 (0.3272)	-0.6204 (0.0)	0.2002 (0.0154)	-0.3053 (0.0002)	0.9942 (0.0)	0.2412 (0.0034)
$r_m - r_f$	0.0818 (0.326)	0.0509 (0.542)	-0.0506 (0.5442)	-0.0346 (0.6786)	0.0345 (0.6794)	-0.0431 (0.6056)

Table 8: Table of Pearson Correlation Coefficients against spatially lagged features with statistical p-values in brackets. From this table we can observe the sign of the correlation between our variables, alongside the statistical significance of the correlations observed. Of particular interest in this table are cases of strong correlations across the diagonal. This diagonal suggests companies which may be connected to similarly characterized firms. This we expect from ICIJ Centrality, based on its definition, but observe for Market Capitalization and, arguably, Profit Margin. This may that similarly sized companies tend to transact, and may offer concerns over identifiability in cases where both variables are included in our model. This table can be recovered using the `pearson_corr` function provided in our project documentation (Gawronsky et al., 2020*c*).

Looking to the pairwise Pearson correlations between our spatially lagged variables, shown in Table 9, we observe a statistically significant correlation between many of our spatially correlated variables. We believe this a function of the density of our graph in which many firms occupy some main dense central neighbourhood. Interesting is the sign and statistical significance between our spatially lagged ICIJ Centrality and Price-to-Earnings explanatory variables, which is expected given the strong spatial auto-correlation in ICIJ Centrality. This strong spatial autocorrelation in ICIJ Centrality suggests that our shortest path projection of our graph has preserved important properties concerning centrality. By definition, we expect spatial lags of eigenvector centrality to have near-perfect spatial autocorrelation across lags given its definition against neighbour centrality.

Pearson Correlation	W Price-to-Earnings	W Market Capitalization	WProfit-Margin	W Price-to-Research	W ICIJ Centrality	$W(r_m - r_f)$
W Price-to-Earnings						
W Market Capitalization	0.4546 (0.0)					
W Profit-Margin	0.5371 (0.0)	-0.1894 (0.022)				
W Price-to-Research	0.7047 (0.000)	0.2344 (0.004)	0.8037 (0.000)			
W ICIJ Centrality	-0.0796 (0.340)	-0.6238 (0.0)	0.2067 (0.012)	-0.304 (0.002)		
$W(r_m - r_f)$	-0.1726 (0.037)	0.0587 (0.482)	-0.6993 (0.000)	-0.7314 (0.000)	0.2402 (0.004)	

Table 9: Table of Pearson Correlation Coefficients between spatially lagged features with statistical p-values in brackets. From this table, we can observe the sign of the correlation between our variables, alongside the statistical significance of the correlations observed. This table can be recovered using the `pearson_corr` function provided in our project documentation (Gawronsky et al., 2020*c*).

In Figure 23, below, we illustrate the relationship between these explanatory and spatially lagged explanatory variables using a Biplot computed via the Singular Value Decomposition (SVD) of our standardized spatial lag explanatory variance-covariance matrix. Here, we see, visually, the positive correlation between Price-to-Research and our market risk premia and our Profit Margin and spatially lagged Market Capitalization, discussed in tables 6, 8 and 9. This we contrast against the variance explained ratio of our Principal Component Analysis, where we observe a steady decay across our components. Using the Kaiser criteria, we would place a cut-off at eleven of our twelve principal components, with a natural 'elbow' at two or eight components.

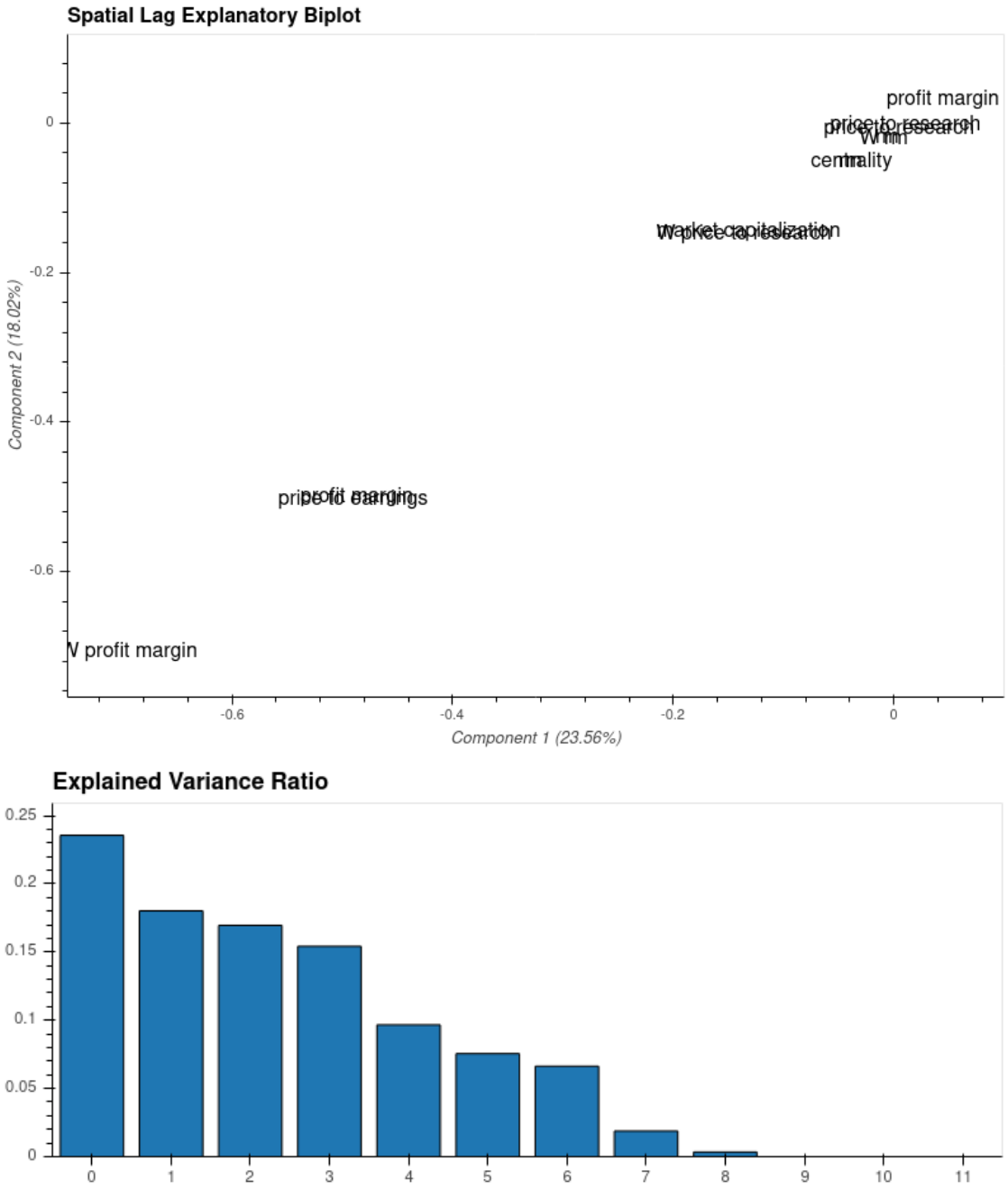


Figure 23: Biplot from the application of Principle Component Analysis, illustrating the relationship between Spatial Lag Explanatory Variables (Gabriel, 1971). In brackets, we show the variance explained ratio of the two Principle Components corresponding to our largest eigenvalues, computed as the ratio of our squared eigenvalues to our sum of squared eigenvalues. Using this plot we can explore, visually, linear correlation between our variables based on these proximity across the chosen dimensions provided in our scatter plot. In the bottom facet, we show the variance explain ratio of our Principal Component Analysis to provide context to the component displayed in our scatter plot above. While our first two components explain close to 42% of our observed variance, components three through eleven still explain a large remaining portion of the in our data variance. These graphs can be recovered using the `biplots` function provided in our project documentation (Gawronsky et al., 2020*c*).

In Table 10, we show the estimates of three spatial lag explanatory models, with lagged Profit-Margin, Price-to-Research and market risk premia explanatory variables. These models are based off model 4 of our non-spatial models, following our conclusions and evidence for spatial auto-correlation under the Mur and Angulo (2009) spatial model selection procedure, shown in appendix A. We begin our analysis by contrasting Variable Inflation Factors across our variables to provide in initial insight into the robustness of our model identification. Looking at these results, we observe considerable evidence for multicollinearity across our spatially lagged variables, as a function of the density in our graph. Again, we rely on a backwards selection procedure, iteratively removing spatially lagged market risk premia and Profit Margin from our

analysis based on their multicollinearity and statistical significance. Looking at our regression diagnostics, we observe little evidence to violate our assumptions given our analysis of the normality and homoskedasticity of our residuals. While we observe evidence for multicollinearity in model **(2)**, under the typical cutoff applied across studies, the joint significance of estimates, suggests the importance of the orthogonal components of these explanatory variables in model **(2)**. An important property to note in our models is the consistency across spatial lags, with Profit-Margin and its lag exhibiting negative estimates and Price-to-Research and its lag both exhibiting positive estimates. This is, in fact, an expected result which suggests what we believe is characteristic of the consistent pricing of latent risk factors across connected firms. We note how the statistical significance of our excess return, α, varies between our non-spatial models in Table 7 and SLX models in Table 10. Comparing these tables, we may determine that non-zero excess returns, α, across our non-spatial models may be a function of latent spatial risk factors, mirroring discussions in Billio, Caporin, Panzica and Pelizzon (2016) on the impact of spatial correlation on estimates of excess return. This may suggest that particular risk factors are in fact transferred through counter-party risk and may serve as a mechanism through which direct contagion may spread between firms in a graph.

Variable Inflation Factor	(1)	(2)*	(3)	(4)	(5)	(6)
W $r_m - r_f$	2.8716					
W Profit-Margin	3.3642	3.3045				
W Price-to-Research	4.7424	3.4044	1.0585			
Price-to-Research	1.2063	1.1408	1.0294	1.0001		
Profit-Margin	1.0372	1.0317	1.0313	1.0034	1.0034	
$r_m - r_f$	1.0253	1.0074	1.0062	1.0036	1.0034	1.0000
Coefficient Estimates						
W $r_m - r_f$	4.730e+00					
	(0.485)					
W Profit-Margin	-1.125e-01	-1.192e-01				
	(0.119)	(0.096)				
W Price-to-Research	3.803e-04	3.193e-04	1.264e-04			
	(0.022)	(0.023)	(0.106)			
Price-to-Research	8.622e-06	7.983e-06	5.991e-06	4.994e-06		
	(0.029)	(0.037)	(0.101)	(0.168)		
Profit-Margin	-2.005e-02	-2.045e-02	-2.074e-02	-1.863e-02	-1.861e-02	
	(0.012)	(0.01)	(0.01)	(0.019)	0.019)	
$r_m - r_f$	-2.277e-01	-2.679e-01	-2.434e-01	-2.796e-01	-2.88e-01	-3.481e-01
	(0.6)	(0.534)	(0.574)	(0.52)	(0.509)	(0.432)
α	3.312e-03	1.565e-02	-4.645e-03	6.723e-03	7.197e-03	1.074e-03
	(0.884)	(0.277)	(0.552)	(0.054)	(0.039)	((0.643))
Overall significance						
F-statistic	2.39	2.781	2.735	2.733	3.117	0.621
	(0.032)	(0.020)	(0.031)	(0.046)	(0.047)	(0.432)
Test for Normality of Errors						
Jarque-Bera	7.186	6.586	4.758	5.117	5.253	2.399
	(0.028)	(0.037)	(0.093)	(0.077)	(0.072)	(0.301)
Diagnostic Tests for Heteroskedasticity						
Breusch-Pagan	5.274	4.223	3.349	2.387	2.007	0.037
	(0.509)	(0.518)	(0.501)	(0.496)	(0.367)	(0.848)
Koenker-Basset	4.511	3.609	2.989	2.249	1.914	0.038
	(0.608)	(0.607)	(0.56)	(0.522)	(0.384)	(0.845)
White	22.968	14.622	14.379	12.096	10.331	7.959
	(0.687)	(0.798)	(0.422)	(0.208)	(0.066)	(0.019)
Diagnostic Tests for Spatial Dependence						
Moran's I (error)	0.831	0.539	0.998	1.309	1.363	1.309
	(0.406)	(0.590)	(0.318)	(0.191)	(0.173)	(0.190)
Robust LM (lag)	0.001	0.048	0.617	0.044	0.263	0.115
	(0.975)	(0.827)	(0.432)	(0.834)	(0.608)	(0.734)
Robust LM (error)	0.024	0.139	0.67	0.012	0.16	0.131
	(0.876)	(0.71)	(0.413)	(0.912)	(0.689)	(0.717)

Table 10: Table illustrating estimates of Spatial Lag Explanatory models. Below all estimated p-value corresponding to appropriate statistical tests are given in brackets. The table contrasts three models computed against backwards features selection chosen against the statistical significance of particular variables. A star (*) is placed against our chosen numbered models to indicate to readers the model which exhibits the greatest joint significance under Fisher's F-statistic. This test explore the null hypothesis that all coefficients are jointly zero and provides an indication to what authors believe to be the most reasonable description of our data generating process across these chosen models. From this table we observe under model (2), the firm-specific and social network impacts of Price-to-Research and Profit Margin. Unlike in Table 7, we observe when controlling for spatial lag explanatory social effects, greater statistical significance in our estimates concerning the effects of Price-to-Research. We believe the significance and consistency of social network lagged Price-to-Research suggests the effects of Price-to-Research to emerge from transactions among firms which may be indicative of the relocation and transaction of intellectual property, as discussed in Griffith et al. (2014). This table can be recovered using the `backwards_selection` function provided in our project documentation (Gawronsky et al., 2020*c*).

Taking inspiration from the spatial model selection procedure in Elhorst (2010), we apply a likelihood ratio test against our chosen spatial lag explanatory models to determine the suitability of a Spatial Durbin Error model to our estimates

in Table 10. Based on this analysis, we observe a likelihood ratio of 1.0579 corresponding to a p-value of 30.36% against our χ^2 distribution with one degree of freedom. When coupled with our evidence from our Robust Lagrangian Multiplier Tests, we determine the Spatial Durbin Error Model ill-suited to our analysis. We note the potential non-normality of our errors across models as presented in our Jarque-Bera test statistics. This may suggest bias across certain estimates of our model though may be difficult to identify. Given our selection procedure, we believe our choice of spatial models well placed to interpretation by researchers and investors, given the simple intuitions of spatially lagged explanatory variables, which provide strong motivation for their application in our work. From our estimates, we are able to determine clear sources of risk which can be used to describe our market response. These characteristics are clearly motivated by the rationale many firms hold when relying on international business corporations (IBCs) for compliance and tax avoidance purposes when licensing intellectual property across companies and provide insight into the consistency and rationale of investors in pricing social effects through these transactions (de Cooperación y Desarrollo Económico et al., 2001).

Building on our discussions of Table 10, we continue our analysis based on our estimates in model **(2)** through which we explore the impacts of our market risk premia, Profit-Margin, Price-to-Research, spatially lagged Profit-Margin and spatially lagged Price-to-Research on returns across our event window. Similar to our findings in Section 6.1, we observe observations Hecla Mining Company (HL), Denison Mines Corp. (DNN), Globus Maritime Limited (GLBS) and Galmed Pharmaceuticals Ltd (GLMD) bear significant influence upon estimates in our analysis. These are not found to change the sign or significance of our estimates but may affect the size of our observed coefficients. Looking to our Kamada–Kawai layout, these points of high leverage bear no obvious structural role in our graph, connecting particular neighbourhoods or holding particular centrality.

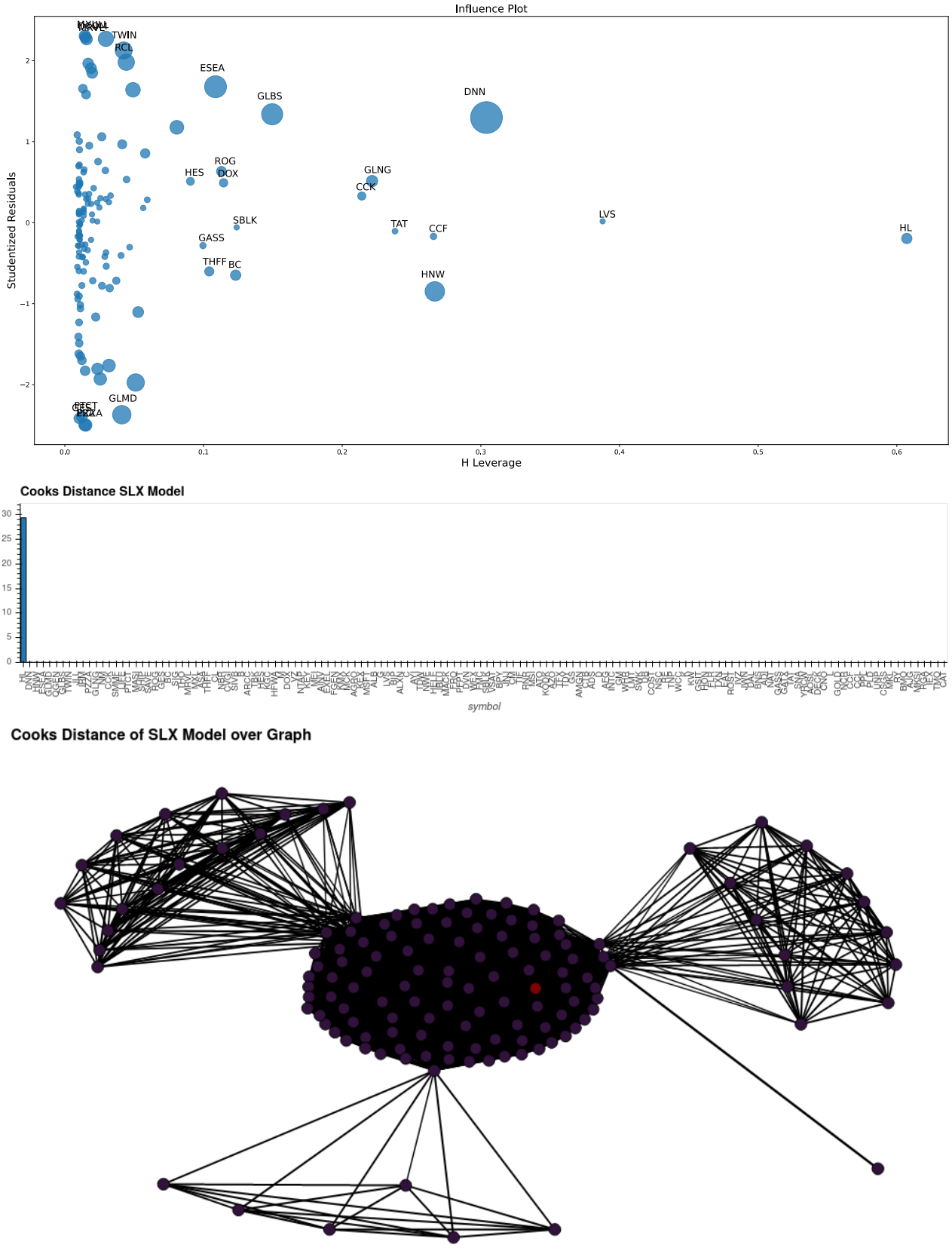


Figure 24: Analysis of leverage of particular data points over residuals and graph layout for SLX model. In the top most facet we scatter plot indicating the H leverage and Studentized Residuals found in our SLX regression model. Size is used to indicate the influence of particular observations. In the facet below we provide a bar chart of the Cooks Distances computed across our observation. From this bar chart we can see the large influence of five of our observation of our estimates. In the bottom-most facet, we provide the reader with a layout of our graph to analyze possible structural factors in our graph which may explain this influence. These figures can be recovered using the `get_regression_diagnostics` function provided in our project documentation (Gawronsky et al., 2020*c*).

Looking to the first two principal components of our SLX explanatory variables, shown in Figure 25, we can observe based on our overlayed Cooks distances, that Hecla Mining Company (HL), Denison Mines Corp. (DNN), Globus Maritime Limited (GLBS) and Galmed Pharmaceuticals Ltd (GLMD) serve as multivariate outliers in this space, placed at the periphery of our data. While we can characterize these points of high influence in this way, these particular outliers remain challenging to explain given available evidence inside our event window but may provide valuable insight for further study.

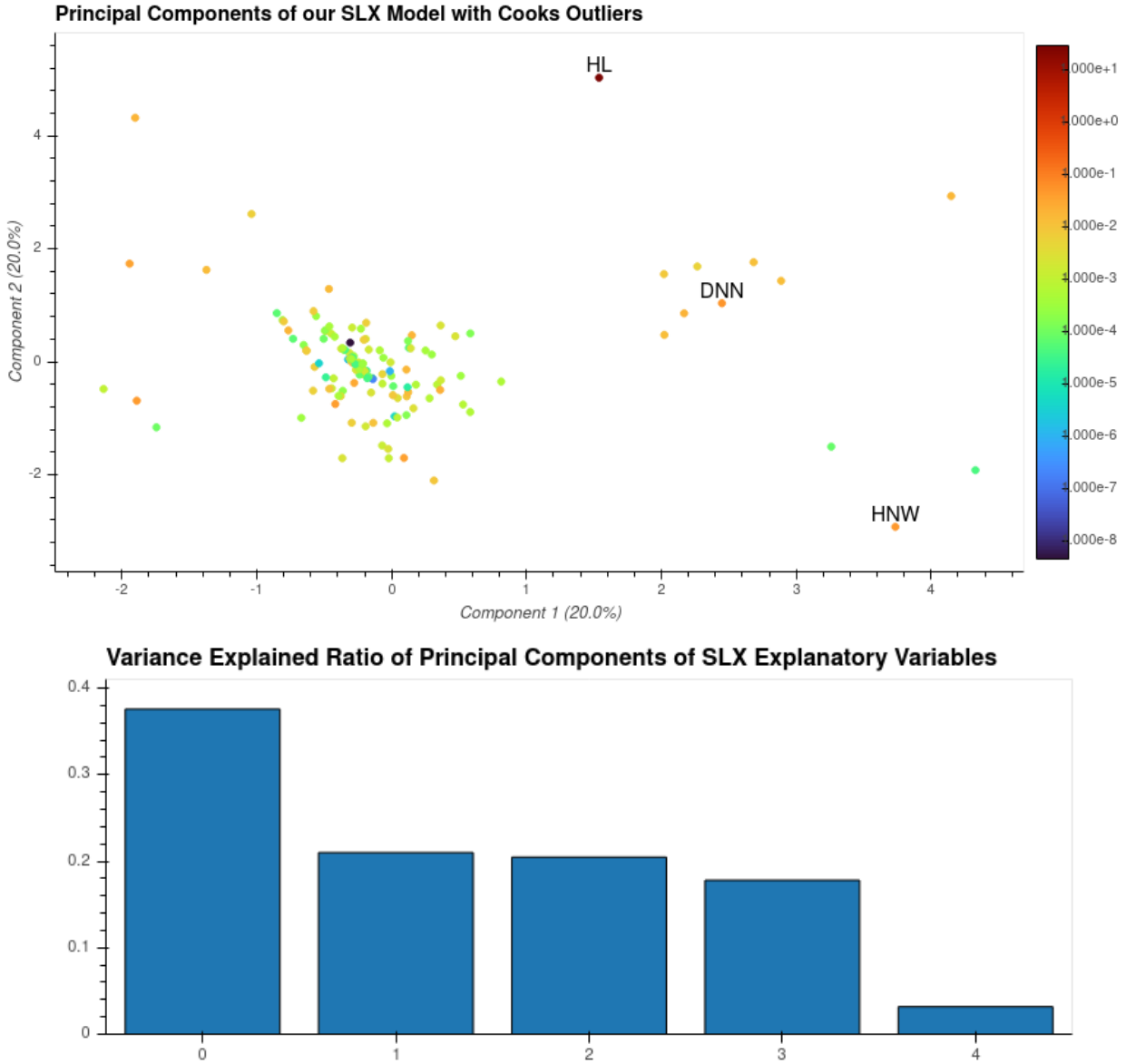


Figure 25: In our top most facet we provide the first two Principal Components of our SLX model visualized against our points of high influence. In this graph we observe our points of high influence on the periphery of data. We provide the variance explained ratios of our Principal Component Analysis in our bottom most facet to provide readers context to the first two components visualized in our scatter plot. These figures can be recovered using the `get_regression_diagnostics` function provided in our project documentation (Gawronsky et al., 2020*c*).

In order to investigate further the relationship between our SLX explanatory variables and possible structure in our graph, we apply Mean Shift Clustering to the first four Principal Components of our explanatory variables using automatic bandwidth selection set at the 30th quantile of our pairwise distances (Comaniciu and Meer, 2002). We choose Mean Shift Clustering based on its constraints in cluster shape, which we believe may aid in interpretability, its robustness to varying density and its automated selection procedure in determining our most appropriate number of clusters. From this analysis, shown in Figure 26, we note clusters 0 and 1 found in well-defined neighbourhoods in our graph. This is a close function of both our spatially lagged variables and the dense connections within these neighbourhoods and is not observed when

clustering over non-spatial features. This may signal artefacts of our spatial features in expressing particular structures in our graph given the particular existence of dense regions. We do not believe this to raise major concerns over our analysis and is determined an expected result.

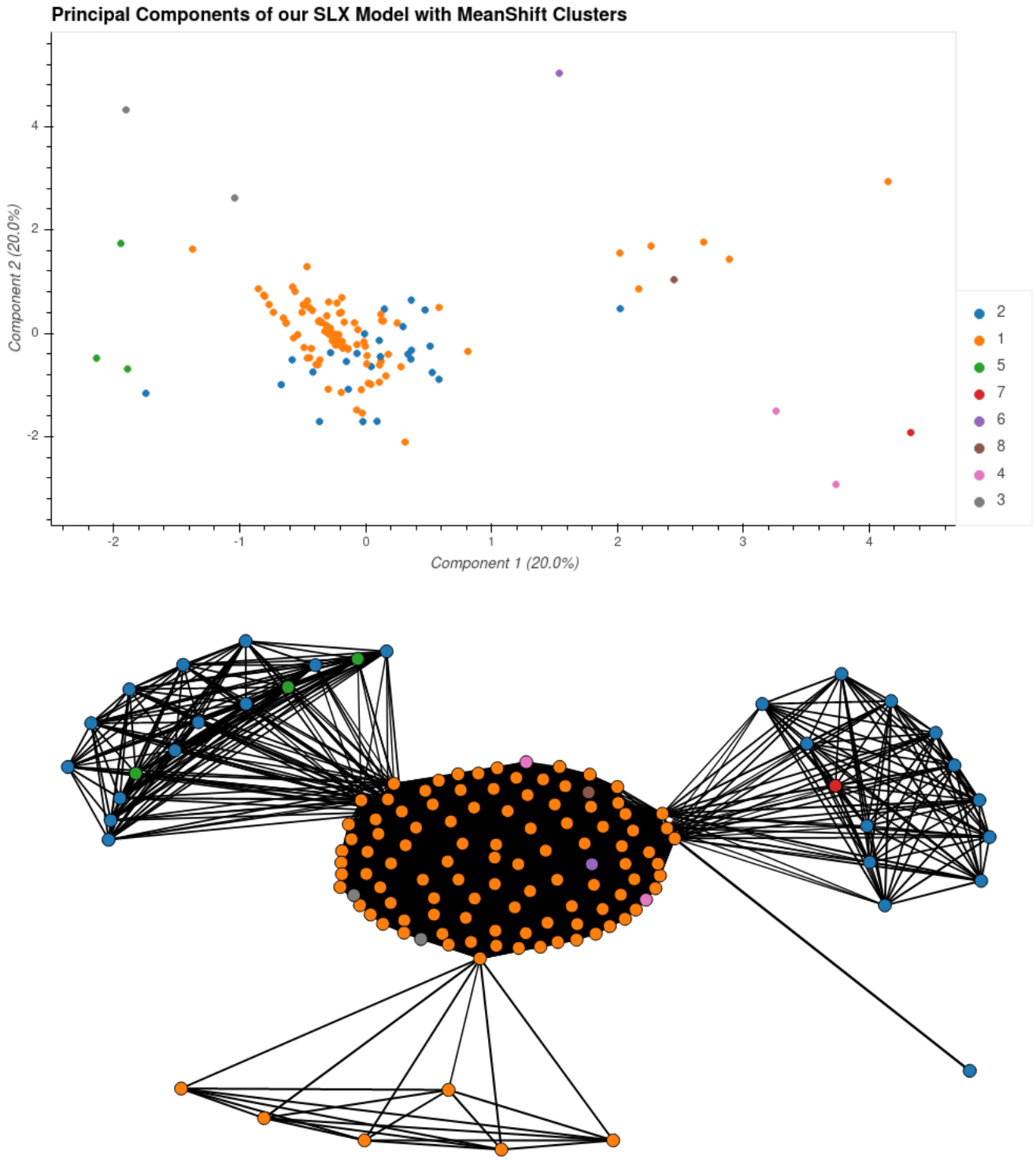


Figure 26: Analysing the results of a Mean Shift Clustering Algorithm applied to the first four Principal Components of our SLX explanatory variables visualized against a Kamada–Kawai layout of our graph (Comaniciu and Meer, 2002). In the bottom facet we observe cluster 0 populating a large dense region in our graph, with clusters 1 populating two dense region on the pheriphery. This Mean Shift Clustering was performed under a Radial Basis Kernel Function using automatic bandwidth selection chosen at the 30th percentile of pairwise distances. In the top facet we observe the first two Principal Components of our SLX features with the variance explained in each of these Principal components provided in brackets. These figures can be recovered using the `get_regression_diagnostics` function provided in our project documentation (Gawronsky et al., 2020*c*).

7 Conclusion

This work identifies and characterizes the price impact of economic transactions across the Paradise Papers. In order to investigate these effects it has looked comparatively across disciplines to the effects of local and global graph structure on financial distress, firm dominance and market efficiency, presented in Sections 2.5 and 2.6 (Chen, Zenou and Zhou, 2018; Corbo, Calvó-Armengol and Parkes, 2007). Building on the social network extensions to Spatial Arbitrage Pricing Theory presented by Billio et al. (2016), discussed in Section 2.6, this work has considered methods in spatial statistics (introduced and applied in Sections 2.2 and 6.2), spectral graph theory (found in Sections 2.4 and 4.3) and graph signal processing (as introduced and explored in Sections 2.4 and 4.2) to characterize important regularity and structure in our market graph. These findings support work by journalists, economists and taxation professionals on the role and function of offshore transaction intermediaries and special purpose vehicles in compliance, privacy and tax avoidance by analyzing both the the regular presence of common intermediate structure between listed companies in the graph, presented in figures 12, 13, 17, 18and 19, and the presence of social effects which evidence dependence based on transactions and relate to identifiable firm characteristics, presented in tables 7 and 10 (Krys, 2016; Johnston, 2002; van der Does de Willebois et al., 2011*b*; Brothers, 2014).

In Sections 6.1 and 6.2, following Ahern (2013) and O'Donovan et al. (2019), we explore the presence of social effects across our an event window following the publication of findings from the Paradise Papers. Using techniques in spatial statistics we characterize our market reaction over this window using a spatial lag explanatory model. From tables 7 and 10 we observe evidence for the social and firm-specific exogenous impacts of price-to-research and company profit margins on market pricing. These findings support works by Nabben (2017) and Griffith et al. (2014) on the role of offshore services in locating intellectual property and the impact of policy changes on the viability and effectiveness of tax avoidance.

In contrast to papers by Ahern (2013), Herskovic (2018), Kelly et al. (2013) and Procházková (2020), our research further exposes the role and impact of social network effects on market pricing when controlling for centrality-related measures (as presented in our spatial diagnostics in Table 7). When compared to Fernandez (2011), Kou et al. (2018) and Billio et al. (2016), we provide insights from our arrival at and interpretations of a spatial lag explanatory model under common spatial modelling selection procedures (discussed in Sections 5.1 and 6.2). While many studies have either learned, constructed or sampled graph structure in which price is measured across all firms, countries or sectors along their graphs, our work considers real-world networks of transactions covering many private entities to provide a methodology on which to model realistic economies in which measurements may be conditioned on firm characteristics. Within the literature in financial data breaches, our work sits strongly between studies by O'Donovan et al. (2019), analyzing the impacts of leak membership, and analysis by Joaristi et al. (2018) and Garcia Alvarado and Mandel (2019), on graph structure and its role in inferring bad entities. This bridge between impact and structure provides valuable insights for work by Griffith et al. (2014) exploring the impacts of regulatory changes on company decision-making, by offering a perspective on market response and a framework through which to test the reaction and impact of these changes.

While this work explores the Paradise Papers, further studies may extend our approach across similar datasets. These datasets may comprise similar publications by the ICIJ on the Panama Papers and Offshore Leaks, or public records concerning the incorporation and management of firms across jurisdictions. While studies on social interaction may interest themselves with the effects of direct contagion, we believe fascinating extensions exist in the application of s-APT to a wide variety of research questions. Using spatial econometric and machine learning techniques, future work could look to incorporate and explore the impact of unstructured data sources using embedding techniques to establish a vector on which to compute some quasimetric for spatial modelling.

Using this work, we hoped to have provided for future research an interdisciplinary approach through which to understand the impact of social network effects on portfolio performance evaluation and investor decision-making, as well as a tool through which to incorporate a range of spatial, social network, time-varying and unstructured data sources inside a more traditional and interpretable approach to financial modelling. We hope this work will have an impact on our regulatory understanding of offshore entities and the risks and role of firm connectedness on market contagion inside pricing events.

8 Bibliography

Ahern, K. R. (2013), 'Network centrality and the cross section of stock returns', *Available at SSRN 2197370* .

Akinwande, M. O., Dikko, H. G., Samson, A. et al. (2015), 'Variance inflation factor: as a condition for the inclusion of suppressor variable (s) in regression analysis', *Open Journal of Statistics* **5**(07), 754.

Anaconda Inc. (2020), 'Anaconda Software Distribution', https://docs.anaconda.com/.
URL: *https://docs.anaconda.com/*

Angrist, J. D. (2014), 'The perils of peer effects', *Labour Economics* **30**, 98–108.

Anselin, L. (2011), 'Gmm estimation of spatial error autocorrelation with and without heteroskedasticity', *Note (GeoDa Center, Arizona State University, 2011)* .

Anwar, R., Djuraidah, A. and Wigena, A. H. (2020), 'Comparison of maximum likelihood and generalized method of moments in spatial autoregressive model with heteroskedasticity'.

Aobdia, D., Caskey, J. and Ozel, N. B. (2014), 'Inter-industry network structure and the cross-predictability of earnings and stock returns', *Review of Accounting Studies* **19**(3), 1191–1224.

Aragam, B. and Zhou, Q. (2015), 'Concave penalized estimation of sparse gaussian bayesian networks', *The Journal of Machine Learning Research* **16**(1), 2273–2328.

Arnold, M. (2011), 'James le sage, robert k. pace: Introduction to spatial econometrics', *Statistical Papers* **52**(2), 493.

Arnold, M., Stahlberg, S. and Wied, D. (2013), 'Modeling different kinds of spatial dependence in stock returns', *Empirical Economics* **44**(2), 761–774.

Arraiz, I., Drukker, D. M., Kelejian, H. H. and Prucha, I. R. (2010), 'A spatial cliff-ord-type model with heteroskedastic innovations: Small and large sample results', *Journal of Regional Science* **50**(2), 592–614.

Arrow, K. J. et al. (1951), An extension of the basic theorems of classical welfare economics, *in* 'Proceedings of the second Berkeley symposium on mathematical statistics and probability', The Regents of the University of California.

Asgharian, H., Hess, W. and Liu, L. (2013), 'A spatial analysis of international stock market linkages', *Journal of Banking & Finance* **37**(12), 4738–4754.

Asness, C. S. (1997), 'The interaction of value and momentum strategies', *Financial Analysts Journal* **53**(2), 29–36.

Banz, R. W. (1981), 'The relationship between return and market value of common stocks', *Journal of financial economics* **9**(1), 3–18.

Barigozzi, M. and Brownlees, C. (2019), 'Nets: Network estimation for time series', *Journal of Applied Econometrics* **34**(3), 347–364.

Barr, C. (2020), 'Revealed: How Angolan ruler's daughter used her status to build $2bn empire — World news — The Guardian', https://www.theguardian.com/world/2020/jan/19/isabel-dos-santos-revealed-africa-richest-woman-2bn-empire-luanda-leaks-angola.

Bauckhage, C., Kersting, K. and Rastegarpanah, B. (2013), The weibull as a model of shortest path distributions in random networks, *in* 'Proc. Int. Workshop on Mining and Learning with Graphs, Chicago, IL, USA'.

Bera, A. K. and Jarque, C. M. (1981), 'Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte carlo evidence', *Economics letters* **7**(4), 313–318.

Billio, M., Caporin, M., Panzica, R. and Pelizzon, L. (2016), 'The impact of network connectivity on factor exposures, asset pricing and portfolio diversification'.

Bivand, R. and Piras, G. (2015), Comparing implementations of estimation methods for spatial econometrics, American Statistical Association.

Blasques, F., Koopman, S. J., Lucas, A. and Schaumburg, J. (2016), 'Spillover dynamics for systemic risk measurement using spatial financial time series models', *Journal of Econometrics* **195**(2), 211–223.

Bloch, F. (2016), Targeting and pricing in social networks, *in* 'The Oxford Handbook of the Economics of Networks'.

Bloch, F. and Quérou, N. (2013), 'Pricing in social networks', *Games and economic behavior* **80**, 243–261.

Bonacich, P. (1987), 'Power and centrality: A family of measures', *American journal of sociology* **92**(5), 1170–1182.

Borgatti, S. P. and Everett, M. G. (2006), 'A graph-theoretic perspective on centrality', *Social networks* **28**(4), 466–484.

Bramoullé, Y., Djebbari, H. and Fortin, B. (2009), 'Identification of peer effects through social networks', *Journal of econometrics* **150**(1), 41–55.

Brothers, J. P. (2014), 'From the double irish to the bermuda triangle', *Tax Analysts* **24**, 687–695.

Buraschi, A. and Porchia, P. (2012), Dynamic networks and asset pricing, *in* 'AFA 2013 San Diego Meetings Paper'.

Candogan, O., Bimpikis, K. and Ozdaglar, A. (2012), 'Optimal pricing in networks with externalities', *Operations Research* **60**(4), 883–905.

Capocelli, R. and Ricciardi, L. (1972), 'On the inverse of the first passage time probability problem', *Journal of Applied Probability* **9**(2), 270–287.

Cardoso, J. V. d. M. and Palomar, D. P. (2020), 'Learning undirected graphs in financial markets', *arXiv preprint arXiv:2005.09958* .

Carhart, M. M. (1997), 'On persistence in mutual fund performance', *The Journal of finance* **52**(1), 57–82.

Caruana-Galizia, P. and Caruana-Galizia, M. (2016), 'Offshore financial activity and tax policy: evidence from a leaked data set', *Journal of Public Policy* **36**(3), 457–488.

Chen, Y.-J., Zenou, Y. and Zhou, J. (2018), 'Competitive pricing strategies in social networks', *The RAND Journal of Economics* **49**(3), 672–705.

Chen, Y.-J., Zenou, Y., Zhou, J. et al. (2020), *Network topology and market structure*, Centre for Economic Policy Research.

Comaniciu, D. and Meer, P. (2002), 'Mean shift: A robust approach toward feature space analysis', *IEEE Transactions on pattern analysis and machine intelligence* **24**(5), 603–619.

Cook, K. S., Emerson, R. M., Gillmore, M. R. and Yamagishi, T. (1983), 'The distribution of power in exchange networks: Theory and experimental results', *American journal of sociology* **89**(2), 275–305.

Corbo, J., Calvó-Armengol, A. and Parkes, D. (2006), 'A study of nash equilibrium in contribution games for peer-to-peer networks', *ACM SIGOPS Operating Systems Review* **40**(3), 61–66.

Corbo, J., Calvó-Armengol, A. and Parkes, D. (2007), Network effects in local contribution economies: identification and regulation, Technical report, Mimeo, Universitat Autonoma de Barcelona.

Coscia, M. and Neffke, F. M. (2017), Network backboning with noisy data, *in* '2017 IEEE 33rd International Conference on Data Engineering (ICDE)', IEEE, pp. 425–436.

Coscia, M. and Rossi, L. (2019), The impact of projection and backboning on network topologies, *in* '2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)', IEEE, pp. 286–293.

Daniel, K. and Titman, S. (1997), 'Evidence on the characteristics of cross sectional variation in stock returns', *the Journal of Finance* **52**(1), 1–33.

Datashare (n.d.), 'Analyze documents - Datashare', https://icij.gitbook.io/datashare/all/analyze-documents. **URL:** *https://icij.gitbook.io/datashare/all/analyze-documents*

de Cooperación y Desarrollo Económico, O. O., for Economic Cooperation, O., OCSE., Staff, O., Staff, D. O., on Corporate Governance, D. S. G., i Rozwoju, O. W. G. et al. (2001), *Behind the corporate veil: Using corporate entities for illicit purposes*, Organisation for Economic Co-operation and Development.

De Giorgi, G., Pellizzari, M. and Redaelli, S. (2010), 'Identification of social interactions through partially overlapping peer groups', *American Economic Journal: Applied Economics* **2**(2), 241–75.

De Prado, M. L. (2016), 'Building diversified portfolios that outperform out of sample', *The Journal of Portfolio Management* **42**(4), 59–69.

Debarsy, N., Dossougoin, C., Ertur, C. and Gnabo, J.-Y. (2018), 'Measuring sovereign risk spillovers and assessing the role of transmission channels: A spatial econometrics approach', *Journal of Economic Dynamics and Control* **87**, 21–45.

Debreu, G. and Herstein, I. N. (1953), 'Nonnegative square matrices', *Econometrica: Journal of the Econometric Society* pp. 597–607.

Dees, B. S., Stanković, L., Constantinides, A. G. and Mandic, D. P. (2020), Portfolio cuts: A graph-theoretic framework to diversification, *in* 'ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 8454–8458.

Dharmapala, D. and Hines Jr, J. R. (2009), 'Which countries become tax havens?', *Journal of Public Economics* **93**(9-10), 1058–1068.

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

Dijkstra, E. W. et al. (1959), 'A note on two problems in connection with graphs', *Numerische mathematik* **1**(1), 269–271.

Dong, X., Thanou, D., Rabbat, M. and Frossard, P. (2019), 'Learning graphs from data: A signal representation perspective', *IEEE Signal Processing Magazine* **36**(3), 44–63.

Drukker, D. M., Egger, P. and Prucha, I. R. (2013), 'On two-step estimation of a spatial autoregressive model with autoregressive disturbances and endogenous regressors', *Econometric Reviews* **32**(5-6), 686–733.

Eder, A. and Keiler, S. (2012), 'Cds spreads and systemic risk-a spatial econometric approach'.

Egger, P., Larch, M., Pfaffermayr, M. and Walde, J. (2009), 'Small sample properties of maximum likelihood versus generalized method of moments based tests for spatially autocorrelated errors', *Regional Science and Urban Economics* **39**(6), 670–678.

Elhorst, J. P. (2010), 'Applied spatial econometrics: raising the bar', *Spatial economic analysis* **5**(1), 9–28.

Erdős, P. and Rényi, A. (1960), 'On the evolution of random graphs', *Publ. Math. Inst. Hung. Acad. Sci* **5**(1), 17–60.

Eugene, F. and French, K. (1992), 'The cross-section of expected stock returns', *Journal of Finance* **47**(2), 427–465.

Fainmesser, I. P. and Galeotti, A. (2016), 'Pricing network effects', *The Review of Economic Studies* **83**(1), 165–198.

Fama, E. F. and French, K. R. (1992), 'The cross-section of expected stock returns', *the Journal of Finance* **47**(2), 427–465.

Fama, E. F. and French, K. R. (1993), 'Common risk factors in the returns on stocks and bonds', *Journal of* .

Fama, E. F. and French, K. R. (2015), 'A five-factor asset pricing model', *Journal of financial economics* **116**(1), 1–22.

Fan, Y., Li, M., Zhang, P., Wu, J. and Di, Z. (2007), 'The effect of weight on community structure of networks', *Physica A: Statistical Mechanics and its Applications* **378**(2), 583–590.

Fernandez, V. (2011), 'Spatial linkages in international financial markets', *Quantitative Finance* **11**(2), 237–245.

Fitzgibbon, W. (20169), 'Datashare: Help test and improve our latest journalism tool'.
URL: *https://www.icij.org/inside-icij/2019/02/datashare-help-test-and-improve-our-latest-journalism-tool/*

Floch, J.-M. and Le Saout, R. (2018), 'Common models in spatial econometrics', *Handbook of Spatial Analysis: Theory and Practical Application with R. Insee-Eurostat, Luxembourg* pp. 149–177.

Florax, R. J., Folmer, H. and Rey, S. J. (2003), 'Specification searches in spatial econometrics: the relevance of hendry's methodology', *Regional Science and Urban Economics* **33**(5), 557–579.

Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2003), *Geographically weighted regression: the analysis of spatially varying relationships*, John Wiley & Sons.

Gabriel, K. R. (1971), 'The biplot graphic display of matrices with application to principal component analysis', *Biometrika* **58**(3), 453–467.

Gai, P., Haldane, A. and Kapadia, S. (2011), 'Complexity, concentration and contagion', *Journal of Monetary Economics* **58**(5), 453–470.

Gai, P., Hayes, S. and Shin, H. S. (2004), 'Crisis costs and debtor discipline: the efficacy of public policy in sovereign debt crises', *Journal of International Economics* **62**(2), 245–262.

Gai, P. and Kapadia, S. (2019), 'Networks and systemic risk in the financial system', *Oxford Review of Economic Policy* **35**(4), 586–613.

Garcia Alvarado, F. and Mandel, A. (2019), 'The worldwide network of tax evasion: Evidence from the panama papers', *Available at SSRN 3527765* .

Gawronsky, M., Gebbie, T. and Rajaratnam, K. (2020*a*), 'Markets on Networks: Evidence from the Paradise Papers Data', https://figshare.com/s/aadf75970f5aaf89a13e. doi:10.25375/uct.13168160.
URL: *https://figshare.com/s/aadf75970f5aaf89a13e*

Gawronsky, M., Gebbie, T. and Rajaratnam, K. (2020*b*), 'Precarious Papers', https://github.com/marcusinthesky/precarious-papers. doi:10.25375/uct.13168115 commit:19875ad15df461a43e542f4bd425f8831e32b091.
URL: *https://github.com/marcusinthesky/precarious-papers*

Gawronsky, M., Gebbie, T. and Rajaratnam, K. (2020*c*), 'Precarious Papers Documentation', https://marcusinthesky.github.io/precarious-papers.
URL: *https://marcusinthesky.github.io/precarious-papers/*

Giannakis, G. B., Shen, Y. and Karanikolas, G. V. (2018), 'Topology identification and learning over graphs: Accounting for nonlinearities and dynamics', *Proceedings of the IEEE* **106**(5), 787–807.

GNU General Public License (n.d.), http://www.gnu.org/licenses/gpl.html.
URL: *http://www.gnu.org/licenses/gpl.html*

Griffith, R., Miller, H. and O'Connell, M. (2014), 'Ownership of intellectual property and corporate taxation', *Journal of Public Economics* **112**, 12–23.

Hagberg, A. A., Schult, D. A. and Swart, P. J. (2008), Exploring network structure, dynamics, and function using networkx, *in* G. Varoquaux, T. Vaught and J. Millman, eds, 'Proceedings of the 7th Python in Science Conference', Pasadena, CA USA, pp. 11 – 15.

Hajek, P. and Henriques, R. (2017), 'Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods', *Knowledge-Based Systems* **128**, 139–152.

Hamilton, W. L., Ying, R. and Leskovec, J. (2017), 'Representation Learning on Graphs: Methods and Applications', pp. 1–24.
URL: *http://arxiv.org/abs/1709.05584*

Hammersley, J. M. (1950), 'The Distribution of Distance in a Hypersphere', *The Annals of Mathematical Statistics* **21**(3), 447–452.

Haugen, R. A., Baker, N. L. et al. (1996), 'Commonality in the determinants of expected stock returns', *Journal of Financial Economics* **41**(3), 401–439.

Herskovic, B. (2018), 'Networks in production: Asset pricing implications', *The Journal of Finance* **73**(4), 1785–1818.

Hunger, M. and Lyon, W. (2016), 'Analyzing the Panama Papers with Neo4j: Data Models, Queries More'.
URL: *https://neo4j.com/blog/analyzing-panama-papers-neo4j/*

IEX Cloud API (n.d.), https://iexcloud.io/docs/api/{#}historical-prices.
URL: *https://iexcloud.io/docs/api/#historical-prices*

International Consortium of Investigative Journalists (2020), 'DataShare', https://github.com/ICIJ/datashare.

International Consortium of Investigative Journalists (n.d.*a*), 'About — ICIJ Offshore Leaks Database', https://offshoreleaks.icij.org/pages/about.
URL: *https://offshoreleaks.icij.org/pages/about*

International Consortium of Investigative Journalists (n.d.*b*), 'About the ICIJ Offshore Leaks Database'.
URL: *https://offshoreleaks.icij.org/pages/about*

Jackson, M. O. and Wolinsky, A. (1996), 'A strategic model of social and economic networks', *Journal of economic theory* **71**(1), 44–74.

Jackson, M. O. and Zenou, Y. (2015), Games on networks, *in* 'Handbook of game theory with economic applications', Vol. 4, Elsevier, pp. 95–163.

Jarque, C. M. and Bera, A. K. (1980), 'Efficient tests for normality, homoscedasticity and serial independence of regression residuals', *Economics letters* **6**(3), 255–259.

Javed, M. A., Younis, M. S., Latif, S., Qadir, J. and Baig, A. (2018), 'Community detection in networks', *Journal of Network and Computer Applications* **108**(C), 87–111.

Joaristi, M., Serra, E. and Spezzano, F. (2018), 'Inferring bad entities through the Panama papers network', *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018* pp. 767–773.

Johnston, D. C. (2002), 'U.S. corperations are using Bermuda to slash tax bills'.
URL: *https://www.nytimes.com/2002/02/18/business/us-corporations-are-using-bermuda-to-slash-tax-bills.html*

Karkinsky, T. and Riedel, N. (2012), 'Corporate taxation and the choice of patent location within multinational firms', *Journal of international Economics* **88**(1), 176–185.

Katz, L. (1953), 'A new status index derived from sociometric analysis', *Psychometrika* **18**(1), 39–43.

Katzav, E., Nitzan, M., ben Avraham, D., Krapivsky, P., Kühn, R., Ross, N. and Biham, O. (2015), 'Analytical results for the distribution of shortest path lengths in random networks', *EPL (Europhysics Letters)* **111**(2), 26006.

Kelejian, H. H. and Prucha, I. R. (1998), 'A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances', *The Journal of Real Estate Finance and Economics* **17**(1), 99–121.

Kelejian, H. H. and Prucha, I. R. (1999), 'A generalized moments estimator for the autoregressive parameter in a spatial model', *International economic review* **40**(2), 509–533.

Kelly, B., Lustig, H. and Van Nieuwerburgh, S. (2013), Firm volatility in granular networks, Technical report, National Bureau of Economic Research.

Koenker, R. and Bassett Jr, G. (1982), 'Robust tests for heteroscedasticity based on regression quantiles', *Econometrica: Journal of the Econometric Society* pp. 43–61.

Kou, S., Peng, X. and Zhong, H. (2018), 'Asset pricing with spatial interaction', *Management Science* **64**(5), 2083–2101.

Krys, K. (2016), 'Recovering illicit assets offshore: Demystifying the black hole'.
URL: *http://kenneth.krys@krys-global.com/*

Lafferty, J., McCallum, A. and Pereira, F. C. (2001), 'Conditional random fields: Probabilistic models for segmenting and labeling sequence data'.

Lanczos, C. (1950), *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, United States Governm. Press Office Los Angeles, CA.

Lee, L.-f. (2003), 'Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances', *Econometric Reviews* **22**(4), 307–335.

Lehoucq, R., Sorensen, D. and Yang, C. (1977), 'Arpack users' guide: Solution of large-scale eigenvalue problems by implicitely restarted arnoldi methods, siam, philadelphia, pa, 1998', *The software and this manual are available at URL http://www. caam. rice. edu/software/ARPACK* .

LeSage, J. and Pace, R. K. (2009), 'Introduction to spatial econometrics crc press', *Boca Raton, FL* .

Liu, X. and Lee, L.-f. (2010), 'Gmm estimation of social interaction models with centrality', *Journal of Econometrics* **159**(1), 99–115.

Lubotzky, A. (1994), 'Discrete groups, expanding graphs and invariant measures. with an appendix by jonathan d. rogawski', *Progress in Mathematics* **125**.

Lucas Jr, R. E. (1978), 'Asset prices in an exchange economy', *Econometrica: journal of the Econometric Society* pp. 1429–1445.

MacKinnon, J. G. and White, H. (1985), 'Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties', *Journal of econometrics* **29**(3), 305–325.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S. and McClosky, D. (2014), The stanford corenlp natural language processing toolkit, *in* 'Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations', pp. 55–60.

Manski, C. F. (1993), 'Identification of endogenous social effects: The reflection problem', *The review of economic studies* **60**(3), 531–542.

Markowitz, H. (1952), 'Portfolio selection, journal of finance', *Markowitz HM—1952.—№* pp. 77–91.

Mikhailov, A. (2019), 'Turbo, an improved rainbow colormap for visualization', *Google AI Blog* .

Miller, H. J. and Wentz, E. A. (2003), 'Representation and spatial analysis in geographic information systems', *Annals of the Association of American Geographers* **93**(3), 574–594.

Moran, P. A. (1950), 'Notes on continuous stochastic phenomena', *Biometrika* **37**(1/2), 17–23.

Mur, J. and Angulo, A. (2009), 'Model selection strategies in a spatial setting: Some additional results', *Regional Science and Urban Economics* **39**(2), 200–213.

Nabben, R. (2017), 'Intellectual property tax planning in the light of base erosion and profit shifting'.

Neal, Z. (2014), 'The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors', *Social Networks* **39**, 84–97.

Newman, M. E. (2001), 'Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality', *Physical review E* **64**(1), 016132.

Newman, M. E. J. (2010), *Networks: An Introduction*, Oxford University Press Inc.

O'Donovan, J., Wagner, H. F. and Zeume, S. (2019), 'The Value of Offshore Secrets: Evidence from the Panama Papers', *Review of Financial Studies* **32**(11), 4117–4155.

OpenCorporates: The Open Database Of The Corporate World (n.d.), https://opencorporates.com/.
URL: *https://opencorporates.com/*

Ortega, A., Frossard, P., Kovačević, J., Moura, J. M. and Vandergheynst, P. (2018), 'Graph signal processing: Overview, challenges, and applications', *Proceedings of the IEEE* **106**(5), 808–828.

Penghui, G., Lihu, L. and Zhengming, Q. (2015), 'Robust test for spatial error model: Considering changes of spatial layouts and distribution misspecification', *Communications in Statistics-Simulation and Computation* **44**(2), 402–416.

Procházková, V. (2020), 'Asset prices, network connectedness, and risk premium'.

QuantumBlack (2020), 'Kedro', https://github.com/quantumblacklabs/kedro.

Radil, S. M. (2011), Spatializing social networks: making space for theory in spatial analysis, PhD thesis, University of Illinois at Urbana-Champaign.

Ramakrishna, R., Wai, H.-T. and Scaglione, A. (2020), 'A user guide to low-pass graph signal processing and its applications', *arXiv preprint arXiv:2008.01305* .

Ramezani-Mayiami, M. and Skretting, K. (2019), Robust graph topology learning and application in stock market inference, *in* '2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)', IEEE, pp. 240–244.

Rello, L. and Baeza-Yates, R. (2013), Good fonts for dyslexia, *in* 'Proceedings of the 15th international ACM SIGACCESS conference on computers and accessibility', pp. 1–8.

Rey, S. J. and Anselin, L. (2007), 'PySAL: A Python Library of Spatial Analytical Methods', *The Review of Regional Studies* **37**(1), 5–27.

Ricaud, B., Borgnat, P., Tremblay, N., Gonçalves, P. and Vandergheynst, P. (2019), 'Fourier could be a data scientist: From graph fourier transform to signal processing on graphs', *Comptes Rendus Physique* **20**(5), 474–488.

Roson, R. and van den Bergh, J. C. (2000), 'Network markets and the structure of networks', *The Annals of Regional Science* **34**(2), 197–211.

Ross, S. (1976), 'The arbitrage theory of capital asset pricing', *Journal of Economic Theory* **13**(3), 341–360.

Serrano, M. Á., Boguná, M. and Vespignani, A. (2009), 'Extracting the multiscale backbone of complex weighted networks', *Proceedings of the national academy of sciences* **106**(16), 6483–6488.

Sharpe, W. F. (1963), 'A simplified model for portfolio analysis', *Management science* **9**(2), 277–293.

Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A. and Vandergheynst, P. (2013), 'The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains', *IEEE signal processing magazine* **30**(3), 83–98.

Stanford NLP (2020), 'CoreNLP', https://github.com/stanfordnlp/CoreNLP.

Thadewald, T. and Büning, H. (2007), 'Jarque–bera test and its competitors for testing normality–a power comparison', *Journal of applied statistics* **34**(1), 87–105.

Vafaie, H. and Imam, I. F. (1994), Feature selection methods: genetic algorithms vs. greedy-like search, *in* 'Proceedings of the international conference on fuzzy and intelligent control systems', Vol. 51, p. 28.

van der Does de Willebois, E., Halter, E. M., Harrison, R. A., Park, J. W. and Sharman, J. (2011*a*), The misuse of corporate vehicles.

van der Does de Willebois, E., Halter, E. M., Harrison, R. A., Park, J. W. and Sharman, J. (2011*b*), *The Puppet Masters*.

Vazquez, A. (2006), 'Polynomial growth in branching processes with diverging reproductive number', *Physical review letters* **96**(3), 038702.

Vigna, S. (2016), 'Spectral ranking', *Network Science* **4**(4), 433–445.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. and SciPy 1.0 Contributors (2020), 'SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python', *Nature Methods* **17**, 261–272.

Watts, D. J. and Dodds, P. S. (2007), 'Influentials, networks, and public opinion formation', *Journal of consumer research* **34**(4), 441–458.

Watts, D. J. and Strogatz, S. H. (1998), 'Collective dynamics of 'small-world'networks', *nature* **393**(6684), 440–442.

Weibull, W. (1951), 'A statistical distribution function of wide applicability. journal of applied mechanics 18: 293-297.', *Statistical and Computational Analysis* **291**.

Wery, J. J. and Diliberto, J. A. (2017), 'The effect of a specialized dyslexia font, opendyslexic, on reading rate and accuracy', *Annals of dyslexia* **67**(2), 114–127.

White, M. (2017), 'Why aren't the streets full of protest about the Paradise Papers? — Micah White — Opinion — The Guardian'.
URL: *https://www.theguardian.com/commentisfree/2017/nov/10/protest-paradise-papers-micah-white*

Wiedemann, G., Yimam, S. M. and Biemann, C. (2018), 'New/s/leak 2.0 – Multilingual information extraction and visualization for investigative journalism', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11186 LNCS**, 313–322.

Wilcox, D. L. and Gebbie, T. J. (2015), 'On pricing kernels, information and risk', *Investment Analysts Journal* **44**(1), 1–19.

Zheng, X., Aragam, B., Ravikumar, P. K. and Xing, E. P. (2018), Dags with no tears: Continuous optimization for structure learning, *in* 'Advances in Neural Information Processing Systems', pp. 9472–9483.

Zhou, T., Ren, J., Medo, M. and Zhang, Y.-C. (2007), 'Bipartite network projection and personal recommendation', *Physical review E* **76**(4), 046115.

Zhuhadar, L. and Ciampa, M. (2019), 'Leveraging learning innovations in cognitive computing with massive data sets: Using the offshore Panama papers leak to discover patterns', *Computers in Human Behavior* **92**, 507–518.
URL: *https://doi.org/10.1016/j.chb.2017.12.013*

A Unsaturated Spatial Model Step-wise Selection Procedures

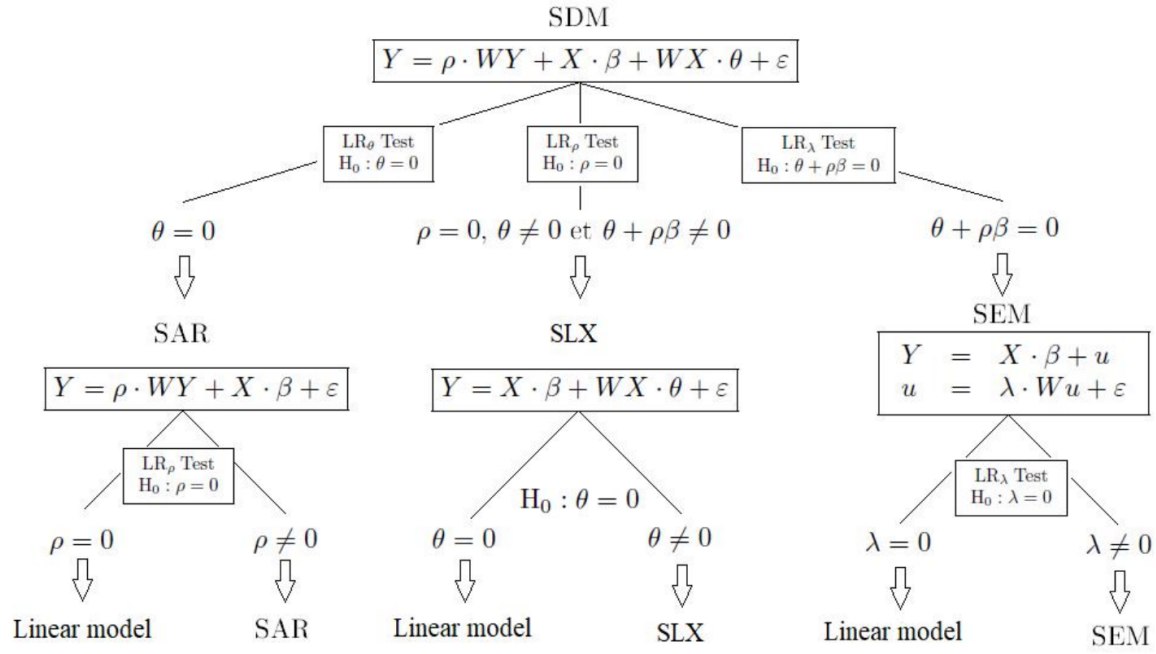


Figure 27: Step-wise procedure for unsaturated models identification provided in Mur and Angulo (2009).

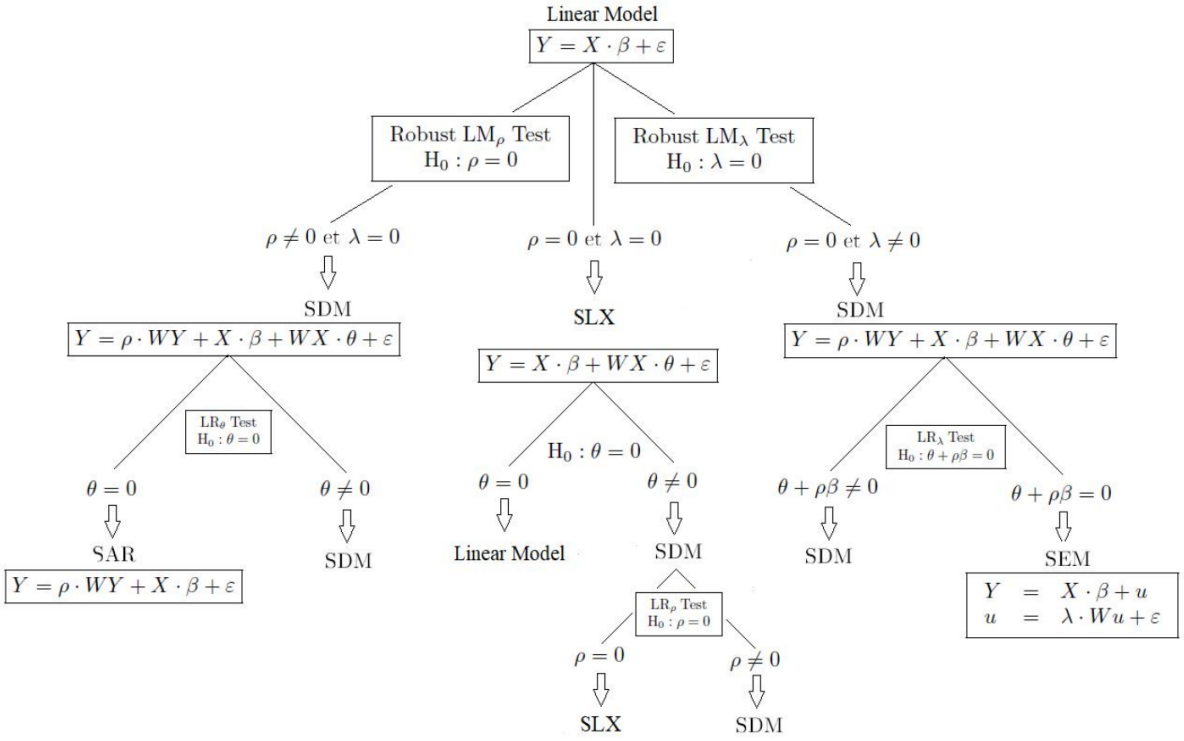


Figure 28: Step-wise procedure for unsaturated models identification provided in Elhorst (2010).

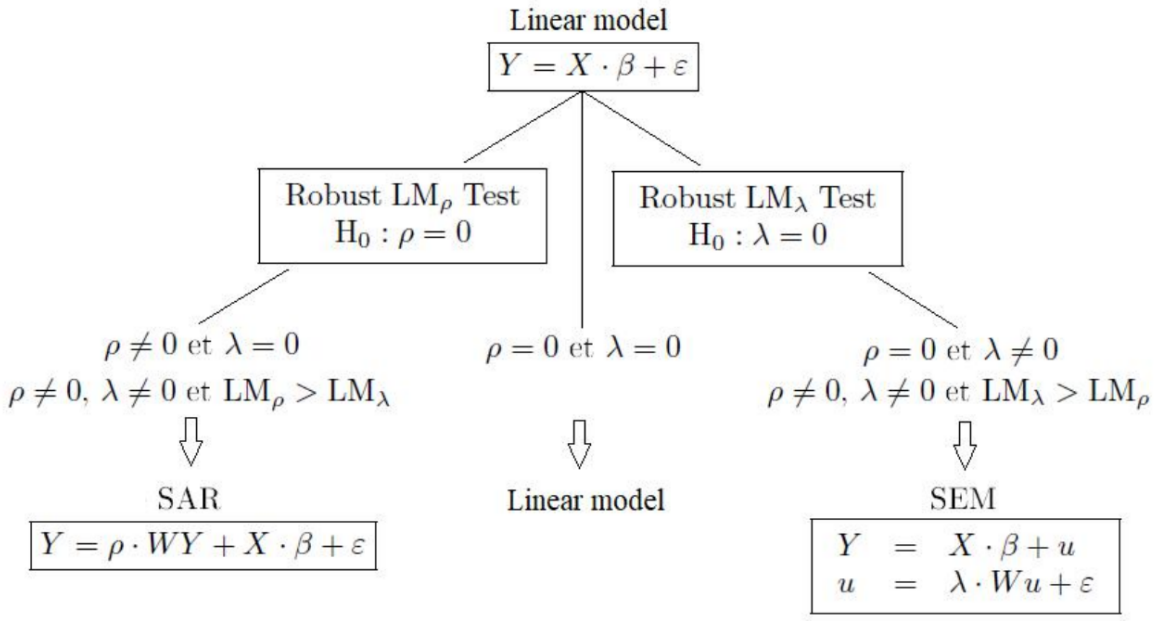


Figure 29: Step-wise procedure for unsaturated models identification provided in Florax et al. (2003).

B Matched Entities

	ICIJ Entry	Listed Company	Ticker	OpenFIGI
0	Goldman Sachs Group Inc	Goldman Sachs Group Inc.	GS	BBG000C6CFJ5
1	Goldman Sachs Group, Inc.	Goldman Sachs Group Inc.	GS	BBG000C6CFJ5
3	Frontline Ltd.	Frontline Ltd.	FRO	BBG000BTQNQ6
4	FRONTLINE LTD.	Frontline Ltd.	FRO	BBG000BTQNQ6
6	American Financial Group, Inc	American Financial Group Inc.	AFG	BBG000DPKKK0
8	Invesco Ltd.	Invesco Ltd.	IVZ	BBG000BY2Y78
9	INVESCO LTD.	Invesco Ltd.	IVZ	BBG000BY2Y78
11	TransAtlantic Petroleum Ltd.	TransAtlantic Petroleum Ltd.	TAT	BBG000C4NR20
13	Genpact Limited	Genpact Limited	G	BBG000RQBFV2
15	ASA Gold and Precious Metals Limited	ASA Gold and Precious Metals Limited	ASA	BBG000BCDWN0
16	NORDIC AMERICAN TANKERS LIMITED	Nordic American Tankers Limited	NAT	BBG000JK57G6
18	Marvell Technology Group Ltd.	Marvell Technology Group Ltd.	MRVL	BBG000BYWTX7
20	Alpha and Omega Semiconductor Limited	Alpha and Omega Semiconductor Limited	AOSL	BBG000QLW222
21	Nabors Industries Ltd.	Nabors Industries Ltd.	NBR	BBG000BZTW70
22	NABORS INDUSTRIES LTD.	Nabors Industries Ltd.	NBR	BBG000BZTW70
24	Golar LNG Limited	Golar LNG Limited	GLNG	BBG000K14TT5
25	Brookfield Infrastructure Partners L.P.	Brookfield Infrastructure Partners L.P.	BIP	BBG000B9Y8D8
27	Brookfield Property Partners L.P.	Brookfield Property Partners LP	BPY	BBG004BMM0N0
28	Delta Air Lines, Inc.	Delta Air Lines Inc.	DAL	BBG000R7Z112
29	DELTA AIR LINES, INC.	Delta Air Lines Inc.	DAL	BBG000R7Z112
32	JetBlue Airways Corporation	JetBlue Airways Corporation	JBLU	BBG000BRQ6L2
33	JETBLUE AIRWAYS CORPORATION	JetBlue Airways Corporation	JBLU	BBG000BRQ6L2
36	SPIRIT AIRLINES, INC.	Spirit Airlines Inc.	SAVE	BBG000BF6RQ9
37	CANADIAN IMPERIAL BANK OF COMMERCE	Canadian Imperial Bank of Commerce	CM	BBG000FKTHQ1
39	ROYAL CARIBBEAN CRUISES LTD.	Royal Caribbean Cruises Ltd.	RCL	BBG000BB5792
40	CARMAX, INC.	CarMax Inc.	KMX	BBG000BLMZK6
41	AUTOZONE INC.	AutoZone Inc.	AZO	BBG000C7LMS8
42	CHASE CORPORATION	Chase Corporation	CCF	BBG000BL9JB7
43	SUMMIT FINANCIAL GROUP INC.	Summit Financial Group Inc.	SMMF	BBG000DLWWK0
44	ROYAL BANK OF CANADA	Royal Bank of Canada	RY	BBG000BSSC44
47	MOTOROLA SOLUTIONS, INC.	Motorola Solutions Inc.	MSI	BBG000BP8Z50
48	EVEREST RE GROUP, LTD	Everest Re Group Ltd.	RE	BBG000C1XVK6
49	WESCO INTERNATIONAL INC.	WESCO International Inc.	WCC	BBG000D0FNV3
50	HELEN OF TROY LIMITED	Helen of Troy Limited	HELE	BBG000BL21Z7
51	HERITAGE FINANCIAL CORPORATION	Heritage Financial Corporation	HFWA	BBG000BY3302
52	TIDEWATER INC.	Tidewater Inc	TDW	BBG00HBQ35R8
53	BRUNSWICK CORPORATION	Brunswick Corporation	BC	BBG000BCWSS3
54	CROWN HOLDINGS INC.	Crown Holdings Inc.	CCK	BBG000BF6756
55	FIRST FINANCIAL CORPORATION	First Financial Corporation	THFF	BBG000CDH213
56	ROGERS CORPORATION	Rogers Corporation	ROG	BBG000BS9HN3
57	TEAM INC.	Team Inc.	TISI	BBG000BVCV15
58	TAPESTRY INC	Tapestry Inc.	TPR	BBG000BY29C7
59	AMDOCS LIMITED	Amdocs Limited	DOX	BBG000C3MXG5
60	CIT GROUP INC.	CIT Group Inc.	CIT	BBG000Q0BPZ4
61	TRANSOCEAN LTD	Transocean Ltd.	RIG	BBG000BH5LT6
62	BCE INC.	BCE Inc.	BCE	BBG000BCXNS3
63	IRON MOUNTAIN INC.	Iron Mountain Inc.	IRM	BBG000KCZPC3
64	DORMAN PRODUCTS INC.	Dorman Products Inc.	DORM	BBG000BM22F5
65	JOHNSON & JOHNSON	Johnson & Johnson	JNJ	BBG000BMHYD1
66	Johnson & Johnson	Johnson & Johnson	JNJ	BBG000BMHYD1
67	Analog Devices, Inc.	Analog Devices Inc.	ADI	BBG000BB6G37
68	IDEX Corporation	IDEX Corporation	IEX	BBG000C1HN22
69	Kennedy-Wilson Holdings Inc	Kennedy-Wilson Holdings Inc.	KW	BBG000CTY4J6
70	Las Vegas Sands Corp.	Las Vegas Sands Corp.	LVS	BBG000JWD753
71	Salesforce.com, Inc	salesforce.com inc.	CRM	BBG000BN2DC2
72	Teleflex Incorporated	Teleflex Incorporated	TFX	BBG000BV59Y6
73	TELEFLEX INCORPORATED	Teleflex Incorporated	TFX	BBG000BV59Y6
74	WEX, Inc.	WEX Inc.	WEX	BBG000BVZP59
75	Albemarle Corporation	Albemarle Corporation	ALB	BBG000BJ26K7
76	Alexion Pharmaceuticals, Inc.	Alexion Pharmaceuticals Inc.	ALXN	BBG000G30YX4
77	Alliance Data Systems Corporation	Alliance Data Systems Corporation	ADS	BBG000BFNR17
78	Amgen Inc.	Amgen Inc.	AMGN	BBG000BBS2Y0
79	Apollo Investment Corporation	Apollo Investment Corporation	AINV	BBG000CBNX94
80	Ares Capital Corporation	Ares Capital Corporation	ARCC	BBG000PD6X77
81	Arthur J. Gallagher & Co.	Arthur J. Gallagher & Co.	AJG	BBG000BBHXQ3
82	Atmos Energy Corporation	Atmos Energy Corporation	ATO	BBG000BRNGM2
83	aTyr Pharma, Inc.	aTyr Pharma Inc.	LIFE	BBG001J2P692
84	Barnes Group Inc.	Barnes Group Inc.	B	BBG000BCSCB1
85	Barrick Gold Corporation	Barrick Gold Corporation	GOLD	BBG000BB07P9
86	RenaissanceRe Holdings Ltd.	RenaissanceRe Holdings Ltd.	RNR	BBG000BFVZ83
87	Brinker International, Inc.	Brinker International Inc.	EAT	BBG000BK28N7

	ICIJ Entry	Listed Company	Ticker	OpenFIGI
88	Carnival Corporation	Carnival Corporation	CCL	BBG000BF6LY3
89	Caterpillar Inc.	Caterpillar Inc.	CAT	BBG000BF0K17
90	CenturyLink Inc.	CenturyLink Inc.	CTL	BBG000BGLRN3
91	Citigroup, Inc.	Citigroup Inc.	C	BBG000FY4S11
92	CNO Financial Group, Inc	CNO Financial Group Inc.	CNO	BBG000Q1GK24
93	Colgate-Palmolive Company	Colgate-Palmolive Company	CL	BBG000BFQYY3
94	Commercial Vehicle Group, Inc	Commercial Vehicle Group Inc.	CVGI	BBG000PZ0SW7
95	Cooper Tire & Rubber Company	Cooper Tire & Rubber Company	CTB	BBG000BGKXV2
96	Costco Wholesale Corporation	Costco Wholesale Corporation	COST	BBG000F6H8W8
97	CSG Systems International, Inc.	CSG Systems International Inc.	CSGS	BBG000G3TQV2
98	Deckers Outdoor Corporation	Deckers Outdoor Corporation	DECK	BBG000BKXYX5
99	DENISON MINES CORP.	Denison Mines Corp.	DNN	BBG000CX6DQ0
100	Deutsche Bank AG	Deutsche Bank AG	DB	BBG000BR1W32
101	DEUTSCHE BANK AG	Deutsche Bank AG	DB	BBG000BR1W32
102	Devon Energy Corporation	Devon Energy Corporation	DVN	BBG000BBVJZ8
103	Dominion Energy Inc.	Dominion Energy Inc	D	BBG000BGVW60
104	Eastman Kodak Company	Eastman Kodak Company	KODK	BBG0057GTG80
105	Era Group Inc.	Era Group Inc.	ERA	BBG001YH8PR9
106	Exelixis, Inc.	Exelixis Inc.	EXEL	BBG000BQ4WF8
107	Macy's, Inc.	Macy's Inc	M	BBG000C46HM9
108	FibroGen, Inc.	FibroGen Inc.	FGEN	BBG000FW5ZL6
109	FMC Corporation	FMC Corporation	FMC	BBG000BJP882
110	GATX Corporation	GATX Corporation	GATX	BBG000BKGXQ4
111	General Motors Company	General Motors Company	GM	BBG000NDYB67
112	GSI Technology, Inc.	GSI Technology Inc.	GSIT	BBG000D0BQK2
113	Guess? Inc.	Guess? Inc.	GES	BBG000BC26P7
114	Hecla Mining Company	Hecla Mining Company	HL	BBG000BL5W86
115	Hess Corporation	Hess Corporation	HES	BBG000BBD070
116	Honeywell International Inc.	Honeywell International Inc.	HON	BBG000H556T9
117	Intel Corporation	Intel Corporation	INTC	BBG000C0G1D1
118	KeyCorp	KeyCorp	KEY	BBG000BMQPL1
119	Kinross Gold Corporation	Kinross Gold Corporation	KGC	BBG000BB2DM7
120	Kirby Corporation	Kirby Corporation	KEX	BBG000BMQCP6
121	Acuity Brands, Inc	Acuity Brands Inc.	AYI	BBG000BJ5HK0
122	Lockheed Martin Corporation	Lockheed Martin Corporation	LMT	BBG000C1BW00
123	Loews Corporation	Loews Corporation	L	BBG000C45984
124	Markel Corporation	Markel Corporation	MKL	BBG000FC7366
125	Masimo Corporation	Masimo Corporation	MASI	BBG000C3W281
126	MaxLinear, Inc.	MaxLinear inc	MXL	BBG000BB6R33
127	Merrimack Pharmaceuticals, Inc.	Merrimack Pharmaceuticals Inc.	MACK	BBG000CQXS37
128	MKS Instruments, Inc.	MKS Instruments Inc.	MKSI	BBG000BVMG26
129	NCR Corporation	NCR Corporation	NCR	BBG000BMXK89
130	NetApp, Inc.	NetApp Inc.	NTAP	BBG000FP1N32
131	Nordstrom Inc.	Nordstrom Inc.	JWN	BBG000G8N9C6
132	Norfolk Southern Corporation	Norfolk Southern Corporation	NSC	BBG000BQ5DS5
133	NorthWestern Corporation	NorthWestern Corporation	NWE	BBG000Q1NMJ4
134	Papa John's International, Inc.	Papa John's International Inc.	PZZA	BBG000BFWF13
135	Pioneer Diversified High Income Trust	Pioneer Diversified High Income Trust	HNW	BBG000R571V4
136	Pioneer High Income Trust	Pioneer High Income Trust	PHT	BBG000DY3VB6
137	PPL Corporation	PPL Corporation	PPL	BBG000BRJL00
138	Prologis, Inc.	Prologis Inc.	PLD	BBG000B9Z0J8
139	PTC Therapeutics, Inc.	PTC Therapeutics Inc.	PTCT	BBG000QT15P7
140	QUALCOMM Incorporated	QUALCOMM Incorporated	QCOM	BBG000CGC1X8
141	Reinsurance Group of America Incorporated	Reinsurance Group of America Incorporated	RGA	BBG000BDLCQ0
142	Ross Stores, Inc.	Ross Stores Inc.	ROST	BBG000BSBZH7
143	Seanergy Maritime Holdings Corp.	Seanergy Maritime Holdings Corp.	SHIP	BBG000RNPD67
144	SEANERGY MARITIME HOLDINGS CORP.	Seanergy Maritime Holdings Corp.	SHIP	BBG000RNPD67
145	Snap-on Incorporated	Snap-on Incorporated	SNA	BBG000BT7JW9
146	SNAP-ON INCORPORATED	Snap-on Incorporated	SNA	BBG000BT7JW9
147	SVB Financial Group	SVB Financial Group	SIVB	BBG000BT0CM2
148	Teledyne Technologies Incorporated	Teledyne Technologies Incorporated	TDY	BBG000BMT9T6
149	Texas Instruments Incorporated	Texas Instruments Incorporated	TXN	BBG000BVV7G1
150	Thermo Fisher Scientific Inc.	Thermo Fisher Scientific Inc.	TMO	BBG000BVDLH9
151	Stanley Black & Decker, Inc.	Stanley Black & Decker Inc.	SWK	BBG000BTQR96
152	THOR Industries Inc.	Thor Industries Inc.	THO	BBG000BV6R84
153	Tiffany & Co	Tiffany & Co.	TIF	BBG000BV75B7
154	Triumph Bancorp, Inc.	Triumph Bancorp Inc.	TBK	BBG000QS6MN9
155	Union Pacific Corporation	Union Pacific Corporation	UNP	BBG000BW3299
156	Universal Forest Products, Inc.	Universal Forest Products Inc.	UFPI	BBG000BL0T06
157	Bank of Montreal	Bank of Montreal	BMO	BBG000DLY9B9
158	Whirlpool Corporation	Whirlpool Corporation	WHR	BBG000BWSV34
159	YRC Worldwide Inc.	YRC Worldwide Inc.	YRCW	BBG000BX6PW7
160	Kellogg Company	Kellogg Company	K	BBG000BMKDM3
161	Verizon Communications Inc.	Verizon Communications Inc.	VZ	BBG000HS77T5
162	Bank of Nova Scotia	Bank of Nova Scotia	BNS	BBG000C2RV03
163	MCKESSON CORPORATION	McKesson Corporation	MCK	BBG000DYGNW7

	ICIJ Entry	Listed Company	Ticker	OpenFIGI
164	MICROSOFT CORPORATION	Microsoft Corporation	MSFT	BBG000BPH459
165	METHODE ELECTRONICS INC.	Methode Electronics Inc.	MEI	BBG000BNY197
166	QIAGEN N.V.	QIAGEN NV	QGEN	BBG000GTYWL7
167	TSAKOS ENERGY NAVIGATION LIMITED	Tsakos Energy Navigation Limited	TNP	BBG000BRM155
169	STEALTHGAS INC.	StealthGas Inc.	GASS	BBG000BN0Y47
170	HILL INTERNATIONAL INC	Hill International Inc.	HIL	BBG000Q3X4V5
171	EUROSEAS LTD	Euroseas Ltd.	ESEA	BBG000C9GH56
172	STAR BULK CARRIERS CORP	Star Bulk Carriers Corp.	SBLK	BBG000L5R950
173	PROOFPOINT, INC.	Proofpoint Inc.	PFPT	BBG000RQ2GY7
174	SIGNET JEWELERS LIMITED	Signet Jewelers Limited	SIG	BBG000C4ZZ10
175	GLOBUS MARITIME LIMITED	Globus Maritime Limited	GLBS	BBG000QP8KT1
176	GALMED PHARMACEUTICALS LTD	Galmed Pharmaceuticals Ltd.	GLMD	BBG005ZD02W1
177	NAVIOS MARITIME HOLDINGS INC.	Navios Maritime Holdings Inc.	NM	BBG000QQV8V7
178	ADECOAGRO S.A.	Adecoagro S.A.	AGRO	BBG001DCNPK3
179	JONES LANG LASALLE INCORPORATED	Jones Lang LaSalle Incorporated	JLL	BBG000C2L2L0
180	VIASAT INC.	ViaSat Inc.	VSAT	BBG000HHLBF9
181	TWIN DISC INCORPORATED	Twin Disc incorporated	TWIN	BBG000BV06P7
182	Fluor Corporation	Fluor Corporation	FLR	BBG000BB1TH9