Macroscopic & mesoscopic dynamics of financial markets

Teh, Boon Kin

2018

Teh, B. K. (2018). Macroscopic & mesoscopic dynamics of financial markets. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/73698>

<https://doi.org/10.32657/10356/73698>

NANYANG
TECHNOLOGICAL
UNIVERSITY

**Macroscopic & Mesoscopic Dynamics
of
Financial Markets**

Teh Boon Kin
School of Physical and Mathematical Sciences
Division of Physics and Applied Physics
2018

Macroscopic & Mesoscopic Dynamics of Financial Markets

Teh Boon Kin

School of Physical and Mathematical Sciences
Division of Physics and Applied Physics

A thesis submitted to the Nanyang Technological University
in fulfilment of the requirement for the degree of
Doctor of Philosophy

2018

*“Science is not only a disciple of reason,
but, also, one of romance and passion”*

Stephen Hawking, September 2010

Abstract

With hundreds trillion dollars of capital floating in the stock market, it is extremely important to understand market structures and dynamics of stock markets. In this thesis, we studied the macroscopic and mesoscopic dynamics of financial markets, from the econophysics (a marriage between physics and economics) point of view. When econophysicists study stock markets, they frequently borrow methods developed in other areas of physics. However, because of the nature of their problems, econophysicists sometimes also invent new methods. In this thesis, we have also contributed methodological innovations (MI), to contrast the phenomenological discoveries (PD) that we have also made. The first of these methodological innovations is (MI1) the method of partial hierarchical clustering (PHC), a supervised clustering method with the advantage of using multiple thresholds to determine clusters. Through the PHC results, we demonstrated (PD1) the existence of hierarchical structures in the Singapore Exchange and Hong Kong Stock Exchange: from market sectors, to country markets, and to global markets. Furthermore, we also investigated the dynamics of these hierarchical structures across market crashes. To do this, we (MI2) extend the complete-linkage hierarchical clustering algorithm, to obtain robust clusters of stocks with high intra-cluster homogeneity and high inter-cluster heterogeneity. By visualizing these robust clusters using (MI3) the fusion-fission diagram, we observed that (PD2) when approaching market crashes, the movements of stock prices become synchronized, causing most of stocks to merge into a giant cluster. Right after the crash, this giant cluster fragmented and thereafter mixed strongly. This discovery points us to the fusion-fission processes in the market, which we can exploit to forecast market crashes. We assume that the traders strategies form strategy clusters in the constantly changing strategy space, and the stock price movements are governed by the dynamics of these strategy clusters. Moreover, this dynamic can be described by a statistical physics fusion-fission model: the soup-of-groups model (SoG). We (MI4) derived a mean-field SoG forecasting equation and showed that some market crashes can be predicted. Specifically, by fitting the continuous returns of the component stocks of the Straits Times Index to the SoG forecasting equation, we (PD3) found episodes of heightened crash likelihoods close to the Chinese Correction (27 Feb 2007), beginning of the Subprime Crisis (17 Aug 2007), and the Asian Correction (9 Mar 2009), with early warning four to six months prior to the crashes.

Thesis Supervisor: Assoc. Prof. Cheong Siew Ann

Acknowledgments

*I*n this culmination of my Ph.D research carried out during the years 2013-2017 at School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore. I must express my greatest gratitude to my doctoral advisor Assoc. Prof. Cheong Siew Ann, for the guidance, encouragement, advice and wisdom he has provided over the years. Deeply appreciate his effort and time spent which influenced my character. I am deeply grateful to my Thesis Advisory Committee members: Assoc. Prof. Chew Lock Yue and Assoc. Prof. Charlie Charoenwong, for their guidance, knowledge and idea I gained from them.

I owe my deepest gratitude to my parents and family members, for allowing me to go after my dream. All the support they have provided me was the greatest gift I have ever received. Also, thank you to Bi Xiang, for all her love and support. Furthermore, I would like to address special thank to my friends Kenny Saw, Khi Pin, Darrell Tay, Teck Liang, Woon Pheng, Wenyuan, Chong Eu, Chee Yong, James, and Misha. Your companionship, support, and encouragement is what I value the most during the journey. I wish you all the best of luck in your future endeavors and manifesting of all your best intentions.

Lastly, I thank Nanyang Technological University for providing NTU Research Scholarship that funded this research, as well as the environment, facilities, and resources that ensure smooth research progress and amazing research experience.

Contents

1	Introduction	1
2	Macroscopic Dynamics of Financial Market	7
2.1	Introduction	8
2.2	The Chinese Correction of February 2007	10
2.2.1	Data	10
2.2.2	Clustering Algorithm: Partial Hierarchical Clustering	11
2.2.3	SGX and HKSE Clusters and Superclusters	13
2.2.4	Dynamics of SGX and HKSE Clusters	16
2.3	The Lehman Brothers Crisis in September 2008	21
2.3.1	Data	21
2.3.2	Clustering Algorithm: Interaction Hierarchical Clustering	21
2.3.3	Dynamics of SGX Clusters	26
2.3.4	Mixing Degree and Persistence	28
	Appendices	32
2.A	Component stocks of the 10 SGX clusters	32
2.B	Component stocks of the 15 HKSE clusters	33
3	Mesoscopic Dynamics of Financial Market	35
3.1	Introduction	36
3.2	Data and Terminologies	38
3.2.1	Data	38
3.2.2	Continuous Returns	39
3.2.3	Relation Between SoG Event and Continuous Return	40
3.3	SoG Forecasting	43
3.3.1	SoG Forecasting Model	43
3.3.2	SoG Forecasting Results	45
	Appendices	50
3.A	Continuous Return Distribution Fitting Results	50
3.B	Fits of SoG Forecasting Results	52
3.C	Forecasting Results	54
4	Conclusion and Future Works	58

5	Methodologies, Algorithms, and Models	62
5.1	Similarity Measures	63
5.1.1	Time Series Similarity: Pearson, Spearman, Kendall tau, and <i>Digital Correlation</i>	63
5.1.2	Spectral Similarity: Cosine Similarity and <i>Adjusted Cosine Similarity</i>	66
5.2	Clustering Analysis I (unsupervised)	68
5.2.1	Hierarchical Clustering	68
5.2.2	<i>Interaction Hierarchical Clustering</i>	71
5.2.3	<i>Clusters Identification</i>	72
5.3	Clustering Analysis II (Supervised)	75
5.3.1	<i>Partial Hierarchical Clustering</i>	75
5.3.2	PHC Sensitivity Analysis	77
5.4	Clusters Mixing Degree Measures	78
5.4.1	Rand Index, Adjusted Rand Index	78
5.4.2	<i>Degree of Mixing</i>	79
5.4.3	Cluster Mixing Degree Significance Testing	80
5.5	Soup-of-Groups Model	83
5.5.1	Soup-of-Groups (SoG) Model	83
5.5.2	<i>Stability of SoG Model</i>	85
5.5.3	<i>SoG Model Dynamics</i>	86
5.6	Distribution Statistical Significance Testing	89
5.6.1	Review of Probability Distributions and Parameters Estimation	89
5.6.2	Clauset, Shalizi, and Newman (CSN) algorithm	91
5.6.3	<i>CSN with Finite Sample Corrections</i>	93
5.6.4	Quantifying the Effect of Finite Largest of Elements	96
5.7	Goodness-of-fit Measures	101
5.7.1	Kolmogorov-Smirnov (KS) Distance	101
5.7.2	<i>Distribution Noise</i>	102
5.7.3	Relationship between Distribution Noise and KS Distance	105
	References	108

List of Figures

2.1.1 Location of Global Stock Exchanges	9
2.2.1 Asian Indices Correlation Matrix	11
2.2.2 Partial Hierarchical Clustering Results of SGX and HKSE	14
2.2.3 Super Clusters (MSTs) of SGX and HKSE	16
2.2.4 Number of Clusters in SGX and HKSE	17
2.2.5 Intra and Inter-Clusters Correlations Level for SGX and HKSE	20
2.3.1 Comparison of Correlation Matrix, Before and After Seriated	22
2.3.2 Fusion-Fission Diagram Illustration	24
2.3.3 SGX Financial Clusters Dynamics	25
2.3.4 Fusion-Fission Diagram of SGX	27
2.3.5 Mixing Degree of SGX Financial Clusters	29
2.3.6 Persistence of SGX Financial Clusters	30
3.2.1 Straits Times Index (daily) from Jan 2006 to Dec 2011	38
3.2.2 Illustration of Continuous Return	40
3.2.3 SoG and Continuous Returns Relation	41
3.2.4 Fits of Continuous Return Distribution	42
3.3.1 Fitting Results of Integrated Continuous Returns	44
3.3.2 Sensitivity Test of SoG Forecasting	46
3.3.3 Market Risk Heat Map of the Singapore Exchange	48
3.A.1Fits of Continuous Return Distribution All (I)	50
3.A.2Fits of Continuous Return Distribution All (II)	51
3.C.1Sensitivity Test of SoG Forecasting All (I)	54
3.C.2Sensitivity Test of SoG Forecasting All (II)	55
3.C.3Sensitivity Test of SoG Forecasting All (III)	56
5.1.1 Correlation Measures for Time Series with Extreme Event	64
5.1.2 Comparison Between Correlation measures	65
5.1.3 Illustration of Adjusted Cosine Similarity	66
5.2.1 Algorithm of Hierarchical Clustering	69
5.2.2 Illustration Clustering Results	70
5.2.3 Robust Clusters identification	72
5.2.4 Dendrogram base Clusters Identification	73
5.3.1 Illustration of Partial Hierarchical Clustering	76
5.3.2 Partial Hierarchical Clustering Kink Angle	77
5.4.1 Degree of Mixing Illustration	79
5.4.2 Comparison of Mixing Degree Measures, <i>Adj</i> and <i>DoM</i>	81
5.5.1 Illustration of the soup-of-Groups (SoG) model	83
5.5.2 SoG Equilibrium Distribution	84

5.5.3 Linear Stability Test of SoG Model	86
5.5.4 Growth Dynamics of Cumulative Events	87
5.5.5 Growth Dynamics of Giant Cluster	88
5.6.1 Statistical Significance Testing on Real Data (unadjusted)	93
5.6.2 KS Distance and Sample Size Relation	95
5.6.3 Statistical Significance Testing on Real Data (adjusted)	96
5.6.4 Approximated Relative Estimation Errors, ε_{RE}^{PL} and ε_{RE}^{EXP}	98
5.6.5 Comparison of Relative Estimation Errors, $\varepsilon_{RE}^{EXP}(\hat{\beta}_{unadj})$ and $\varepsilon_{RE}^{EXP}(\hat{\beta}_{adj})$	99
5.6.6 Comparison of KS Distance, KS_{unadj} and KS_{adj}	99
5.7.1 Illustration of Kolmogorov-Smirnov Distance	102
5.7.2 Distribution Noise Illustration	103
5.7.3 Distribution Noise and Sample Size Relation	104
5.7.4 Correlation between KS Distance and Distribution Noise	105

List of Tables

2.A.1Component stocks of the 10 SGX clusters	32
2.B.1Component stocks of the 15 HKSE clusters	33
3.1 SGX Stocks Codes and Transaction Details	39
3.B.1SoG Forecasting Results	52

List of Equations (Important Only)

2.2	Digital Correlation	12
2.4	Intra-Clusters Correlation	18
2.5	Inter-Clusters Correlation	18
2.8	Standard Deviation of Intra-Clusters Correlation	19
2.9	Standard Deviation of Inter-Clusters Correlation	19
2.10	Pearson Correlation	21
2.11	Robust Clusters Interaction Distance	23
3.1	Continuous Return	39
3.3	SoG Forecasting Equation	43
3.5	Empirical Integrated Continuous Returns	43
3.6	Weighted Mean and Weighted σ of Forecasted Crash Time, t_c	46
5.1	Pearson Correlation	63
5.2	Spearman Correlation	63
5.3	Kendall tau Correlation	64
5.4	Digital Correlation	64
5.5	Cosine Similarity	66
5.6	Adjusted Cosine Similarity	67
5.7	Distance Measures based on Similarity Measures	68
5.8	Robustness Length	71
5.9	Robust Clusters Criteria	71
5.10	Robust Clusters Interaction Distance	72
5.12	Adjusted Rand Index	78
5.13	Mixing Degree of a Cluster	80
5.18	Master Equations for the SoG Model	84
5.19	SoG Equilibrium Distribution	85
5.20	SoG Linearization Jacobian Matrix	85
5.21	Growth Dynamics Coefficient	87
5.25	Likelihood Function	90
5.33	Adjusted Estimated $\hat{\alpha}$ for Power Law Distribution	94
5.34	Adjusted Estimated $\hat{\beta}$ for Exponential Distribution	94
5.36	Relative Error of Estimated Parameters	96
5.43	Approximated Relative estimation errors, ε_{RE}^{PL} and ε_{RE}^{EXP}	98
5.44	Kolmogorov-Smirnov (KS) Distance	101
5.46	Distribution Noise	103

Chapter 1.

Introduction

Econophysics, a marriage between physics and economics, can be traced back the works of Jan Tinbergen, a former physicist who won the Nobel Memorial Prize in Economic Sciences in 1969. Evidently, his Ph.D research with the thesis titled “Minimumproblemen in de natuurkunde en de economie” (Minimization problems in Physics and Economics) [1] is trying to consider the analogies that would likely hold between physical and economic problems, and tackle economic problems as minimization problems that physicists usually solve [2]. More importantly, he put forth the gravity model of international trade [3]: The trade flow between country i and j , $E_{ij} = \alpha_o Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{-\alpha_3}$. Here Y represents the country gross national product and D is the distance measured by mobility, trade barriers, etc. Whereas α 's are estimated empirically. As compared to the Newton's Law of Universal Gravitation force $F_{ij} = Gm_i m_j D_{ij}^{-2}$, where G is the gravitational constant ($6.674 \times 10^{-11} \text{ N(m/kg)}^2$), m is the mass, and D is the physical distance between the bodies. As follow-up of the gravity-international trade model's solid empirical performance, there are attempts for theoretical economic foundation and extensions for the model [4–8].

As early as 1991, Mantegna showed that the distribution of returns of indices value has statistical properties compatible with a Levy random walk with super-diffusive price variations [9]. The following year, Takayasu *et al.* did a theoretical analysis of a microscopic market model and suggested that the financial market is at the edge of chaos [10]. Econophysics started attracting more attention in 1995, when H. Eugene Stanley coined the name during the *Statphys-Kolkata II* conference [11, 12]. This is in recognition of the strong econophysics movement within the 1994 Kolkata conference, where economists and physicists met to discuss the possibilities of applying physics methods to economic problems [12]. Since then, econophysics became widely acknowledged as a research field: methods and tools originating from statistical physics, non-linear chaotic dynamics, complexity and network science have been ported over to the study of economics with modifications and further developments [13, 14]. However, it is not widely known because the top rank physics journal, Physical Review Letters was shutting the door for all econophysics manuscripts without review simply as “the topic is not physics” at that time. The first international conference with econophysics in the title was held in 1997 in Budapest, and Physical Review Letters opened the door in 1997 to the paper by Takayasu *et al.* [15] that clarified why market price changes follow a power law by starting from a general physical point of view based on the Langevin equation. Since then, econophysics became accepted as a new field of physics or applied physics.

Econophysicists also dive into the study of business incomes and personal wealth. The power-law distribution of incomes of business firms became clear in 1999 [16] and later in 2000 personal income was also shown to follow the power law distribution [17]. Independent researches also show that distribution of the wealth, empirically have an initial rise (poverty region), followed by an exponential body (middle class region), then

power-law tail (rich region) [18, 19]. The exponential body of the wealth distribution can be explained by the wealth exchange dynamics in society, describable in terms of the thermodynamics of heat (wealth) exchange between ideal gas particles (agents) [20, 21]. Furthermore, extending this ideal-gas wealth model, Chakrabarti *et al.* showed that by introducing a propensity for agents to save, we will end up with a distribution that fits well with the empirical wealth distribution, from the poverty region to the rich region [22].

Besides the wealth distribution, econophysicists also study *interbank networks* and revealed the presence of a contagion risk, a systematic risk that one bank failure can cause cascading bank failures without intervention from the authorities [23–25]. Fortunately, Thurner and Poledna demonstrated that this systematic risk can be controlled, via risk-redistributing to reduce the *DebtRank* of the interbank liability network [26]. It is also worth mentioning the works done by Pietronero *et al.* in mapping out the income (gross domestic product)-fitness (country competitiveness) phase diagram, which traces the dynamics of economic complexity and evolution of countries [27, 28]. According to them, a country can fall within one of the four phases, (I) potentially risky economies, (II) poverty trap (underdeveloped economies), (III) emerging economies (frontier), and (IV) developed economies. This information is particularly important for policy makers, to know their country’s strengths and weaknesses, before tweaking plans for the country’s future. Finally, econophysicists are without a doubt making contributions in the study of financial markets as well. For instance, in understanding market microstructure and stock market dynamics, which we shall discuss in greater details as chapters introduction in Section 2.1 and 3.1.

With hundreds of trillions of dollars of capital floating in the financial market, it is extremely important to understand market structures and dynamics of financial markets. Because of this, extracting meaningful information from the sea of information is critical, especially when dealing with the vast dimensions of variables that are interrelated. One approach is via information filtering by removing less important information, such as keeping only the significant interactions (connections) between variables. The most common algorithms for doing this are the *minimum spanning tree* (MST) and the *planar maximum filter graph* (PMFG) from network science [29–38]. By applying MST and PMFG algorithms to analyze stock markets, studies have discovered that there are strong connections among stocks within same market sector, with weaker connections between market sectors. This result indicates that there are hierarchical organizations and structures in the stock markets. More importantly, these hierarchical structures are different before and after financial crises, as demonstrated by Onnela *et al.* on the stock market over the Black Monday crash (October 1987) [31], and Fenn *et al.* on the exchange market for several crises [39]. Although they successfully reduce the amount of information kept, both MST and PMFG algorithms keep only small number of links to make the structure of the connectivity easy to grasp, at the cost of removing a substantial amount of information. Another alternative is *clustering analysis*, which retains more connections, but reduce the dimensions by grouping similar variables into a single cluster.

Clustering algorithms have been in existence for decades and can be organized into two major branches: *hierarchical clustering* and *partitional* (segmentation) clustering. Two independent review papers focus on clustering methods used in management, business

studies, and business administration journals over the past decades. Both reviews show similar results, (I) for the first review, out of 45 publications 24 (53%) are hierarchical based, 8 (18%) are partitional based, and 6 (13%) are a combination of hierarchical and partitional clustering [40]; (II) for the second review, the number of publications used hierarchical clustering and partitional clustering are 83 (34%) and 90 (37%) respectively [41]. Both hierarchical clustering and partitional clustering have their own advantages and disadvantages. Hierarchical clustering provides us with a dendrogram, a tree-like diagram that reveals the structure of clusters at different levels, which is more informative than the unstructured set of clusters returned by flat clustering such as partitional clustering. On the other hand, the partitional clustering segregates variable into clusters (with a predefined number of clusters) via maximizing the homogeneity within clusters. Thus, while determining clusters, the degree of subjectivity for partitional clustering is less than in hierarchical clustering. In short, the former maximizes an utility function, but the latter relies on expert judgments in choosing the threshold [40, 42, 43].

In the study of financial markets, we chose to perform hierarchical clustering. This is because we expect hierarchical structures to be present in financial markets, as there are multi-levels of market structures, investors of various sizes, and different economy taxonomies. In order to reduce the degree of subjectivity in hierarchical clustering, we introduce the method of *partial hierarchical clustering* (PHC), that identify cluster by adding in next most similar variable one by one, and stop when the newly added variable is very different from those within the cluster. A cluster identified this way is called a natural cluster (Section 2.2.2). This jump in similarity measure as the stopping criterion for identifying cluster, is similar to the *stepsize* introduced by Johnson *et al.* [44] and Sneath *et al.* [45] to reduce the ambiguity in determining clusters. Furthermore, to identify more clusters, we repeat the PHC algorithm iteratively (after removing clusters that have been identified). Because of this, the PHC has the advantage of using multiple local thresholds to determine clusters. Unfortunately, often the distance increment is smooth, especially after a few natural clusters have been discovered, and subsequent natural clusters become hard to identify. This leads us to further invent the method of *interaction hierarchical clustering algorithm* (IHC) (Section 2.3.2). In IHC, we first identify robust clusters, which are similar to natural clusters identified by PHC. We then re-cluster these robust clusters based on the interactions among the robust clusters. Lastly, we identify clusters by taking into considerations distinction between the intra-cluster connections and the inter-cluster connections. We believe the IHC is a powerful tool for unsupervised clustering analysis, because it is a two-stage hierarchical clustering that has the same advantage as the PHC, as well as an objective way to account for both homogeneity and heterogeneity among clusters.

Armed with these methods, we first study the macroscopic dynamics of the financial market, particularly the dynamics of a large collection of individual securities within a market. This is done by applying the PHC and IHC algorithms (in Chapter 2) to the Hong Kong Stock Exchange (HKSE) and Singapore Exchange (SGX), where Chinese stocks are listed, for two different periods: (I) 2006 to 2007, and (II) 2008 to 2010. We chose these periods because we are interested in the dynamics of market structures across global financial crises, namely the Chinese Correction in February 2007, and the collapse

of Lehman Brothers in September 2009. Unsurprisingly, we demonstrate in Section 2.2.3, the stocks from similar sectors form clusters. At one level higher, these stock clusters further organized into a local (SGX/HKSE) dominated supercluster and a Chinese stock dominated supercluster. In addition, these two superclusters strongly connected as well due to market activities. This discovery highlights the hierarchical structures in the financial markets, from market sectors, to country markets, and to global markets. To go one step further, we analyzed the evolution of financial clusters toward market crashes. We learned that there are raising intra-cluster correlations, decreasing number of clusters when closer to crashes (see Section 2.2.4). These results were also discovered by Musmeci *et al.* in 2015, using different clustering algorithms. We visualized these changes in market hierarchies in the form of a *cluster fusion-fission diagram*, and showed the emergence of a giant cluster in SGX as we approached the Lehman Brothers crisis. This giant cluster fragmented into small clusters right after the crash (discussed in Section 2.3.3). In addition, we also discover strong mixing measured by the adjusted Rand index (*ARI*) [46, 47] (which measures the mixing degree at the global level) in the aftermath of crisis. We also introduce the *degree of mixing* (*DoM*) to measures mixing degree at the level of a single cluster. With this tool, we discovered that the high mixing degree right after the crisis is due to the major clusters (three largest clusters), indicating a wide spectrum of expectations and responses from the market participants (Section 2.3.4).

Up till this point, we have detected some precursors of market crashes: the rise of synchronization in market movements, the emergence of the giant cluster prior to market crashes, and disintegration of the giant cluster right after market crashes, with strong mixing. We want to exploit these discoveries to forecast market crashes and obtain early warning signals. Noting that market is made up from a large number of adaptive traders, who employ similar or competing trading strategies, and assuming that similar strategies form strategy clusters in the abstract strategy space, we reasoned that the stock price movement must be governed by the dynamics of these strategy clusters. For instance, when a strategy cluster grows larger, trader actions become increasingly correlated and this must be reflected as momentum in the stock price. We use the *soup-of-groups* (SoG) model introduced by Johnson *et al.* to describe the dynamics of strategy clusters. The SoG model has been used by Johnson *et al.* to describe many complex systems, including social networks, quantum many body systems, superconducting Cooper pairs, insurgents and terrorists, etc. [48–52]. Using the SoG model, they were able to explain the underlying dynamics of terrorist attacks, and reveal the escalation rate and timing of the fatal attacks from terrorist as following naturally SoG dynamics and is thus predictable [51, 53]. The SoG model was further used to describe the dynamics of fault plane elements, enabling the forecasting of the timings, magnitudes, and locations of large earthquakes [54]. These results thus motivated us to try to replicate the success of SoG forecasting in the financial market.

In Chapter 3 of this thesis, we study the mesoscopic dynamics of the financial market, focusing on 20 of the 30 component stocks of Straits Times Index (STI) traded in the SGX from 2006 to 2010. We called this mesoscopic dynamics as the scale is between the microscopic level (individual stocks and trading activities) and macroscopic level (market level). We first introduce the *continuous return* in Section 3.2.2, which is defined as a

continuous price movement in the same direction, to avoid the ambiguous choice of time intervals. We demonstrated that the continuous return follows the exponential truncated power law distribution, which is the equilibrium distribution of the SoG cluster size [52], that relate continuous returns with SoG clusters (Section 3.2.3). In order to forecast the October 2008 market crash due to the collapse of Lehman Brothers, we derived the mean-field *SoG forecasting model* in Section 3.3.1. In Section 3.3.2, we show that by fitting the continuous returns of the component stocks to the SoG forecasting equation, we can obtain crash predictions ranging from end October 2008 to mid-February 2009, with early warning four to six months prior to the crashes. Notably, we also find episodes of heightened market risks close to the STI low due to the Chinese Correction (27 Feb 2007), beginning of the Subprime Crisis (17 Aug 2007), and the Asian Correction (9 Mar 2009) through the *market risk heat map* created using the SoG forecasting results.

All in all, through the macroscopic and mesoscopic studies on financial markets, we gain a better understanding about the dynamics of financial markets. The implications are that we are now able to detect precursors of major market crashes, and can attempt to forecast these crashes. We conclude our thesis in Chapter 4, together with some suggestions for future works that can be explored to improve the SoG forecasting model. Lastly, in addition to the insights gained about the dynamics of financial market, other contributions from this thesis research include the revisions, extensions, and inventions of models. These are in the areas of similarity measures, clustering algorithms, cluster mixing degree measures, and statistical significance testing. We present these methodologies, algorithms, and models in Chapter 5. We also provide the codes that can be downloaded from <https://github.com/BoonKinTeh>.

Chapter 2.

Macroscopic Dynamics of Financial Market

Highlights

- Existence of clusters of local stocks and clusters of Chinese stocks in Singapore and Hong Kong stock markets, which interact with each other to form *superclusters* [Section 2.2.3].
- Endogenous and exogenous crashes have different intra-cluster and inter-cluster correlation signatures [Section 2.2.4].
- *Fusion-fission diagram* shows how financial cluster sizes evolve, the exchange of stocks between clusters, as well as how strongly the clusters mix [Section 2.3.3].
- A giant cluster emerged prior to the financial crisis, which remained persistently large until the Lehman Brothers Crisis. The giant cluster fragmented right after the Lehman Brothers Crisis, thereafter mixing strongly [Section 2.3.4].
- Two groups of strongly-correlated stocks, one where correlations decay with a lifetime of three weeks, and another where correlations decay with a lifetime of six weeks [Section 2.3.4].

Significance statements

By applying the method of partial hierarchical clustering, a supervised clustering algorithm that can use multiple thresholds to identify clusters, we demonstrated that stocks in the Singapore Exchange (SGX) and the Hong Kong Stock Exchange (HKSE) self-organized to form hierarchical structures: from market sectors, to country markets, and to global markets. Furthermore, we also investigate the changes of these hierarchical structures, and discover the precursors for financial crises. For instance, when we approach market crashes, there are increasing intra-cluster and inter-cluster correlation, decreasing number of clusters, and decreasing cluster mixing degree. More importantly, we present the cluster dynamics using the cluster fusion-fission diagram, and show the emergence of the giant cluster prior to the financial crisis, which remained persistently large until the Lehman Brothers Crisis. Right after the crisis, the giant cluster fragmented into small clusters associated with strong mixing among clusters in the aftermath of the crisis, indicating the wide spectrum of expectations and responses from the market participants.

2.1 Introduction

Ever question about the total amount of capital floating in the global stock market? According to the World Bank, “the market capitalization (also known as market value) is the share price times the number of shares outstanding (including their several classes) for listed domestic companies. Investment funds, unit trusts, and companies whose only business goal are to hold shares of other listed companies are excluded. Data are end of year values converted to U.S. dollars using corresponding year-end foreign exchange rates.” Based on this definition, the World Bank reported a 2016 global market capitalization of US\$ 64,819,639 millions, which is comparable to the 2016 global gross domestic product of US\$ 75,641,577 millions (85.7%). This gigantic amount underlines the importance to understand the financial market, such as the market microstructure and stock market dynamics.

In the econophysics literature a lot of work has been done to understand the financial market, that is highly heterogeneous, highly nonlinear, and strongly interacting, and notoriously non-stationary at human time scales [55–58]. Studies had discovered *stylized facts* in the financial markets (mostly qualitative), which are empirical statistical regularities or ‘truths’ discovered in the financial and economic systems across a wide range of financial instruments, markets, and time periods [59–65]. These discoveries hint that price dynamics is influenced by some underlying mechanisms, instead of being a purely random Brownian motion. Indeed, in 1999, Laloux *et al.* found that the residuals of the detrended financial returns can be described by the theory of random Wishart matrices [66–68], and for the very first time allows us to discriminate between signal and noise via *random matrix theory* (RMT) [69, 70]. Market signals are correlations that are much stronger than correlations we expect from the RMT from random (noisy) time series with the same variance [71, 72]. Furthermore, the signal eigenvectors effectively map out the industry sectors on the New York Stock Exchange. The nature of this signal became clearer after the second breakthrough, i.e. the discovery by Mantegna that there exists a hierarchical organization and structures of stocks in the financial market [29], which can be discovered by constructing the *minimal spanning tree* (MST) and *hierarchical clustering* [32, 73–79]. Indeed, other groups confirmed the existence of these hierarchical structures, which they discovered across different financial markets using more sophisticated methods [71, 80–82].

More importantly, researchers found that there are structural changes in the financial markets, such as their distributions of returns and their hierarchical structures after the financial crisis [30, 31, 83–88]. In addition, researchers also discovered precursor patterns prior to financial crises, such as an increase in the volatility of financial returns and the correlations between stocks [89–93], as well as an increase in the information dissipation level [94]. These observations not only can be found in stock markets, but also for interest rate swap and currency exchange markets [95]. We believe this is because of the wide spectrum of market expectations and traders reactions during and after market crashes (turbulent market periods), resulting in the active restructuring of trading strategies and portfolios. Therefore, it is very unlikely that the hierarchical structure of the stock market remains unchanged across market crashes. In these reorganizations, stocks that used to

be strongly correlated may become weakly correlated, while stocks that used to be weakly correlated may become strongly correlated.

In this chapter, we would like to examine the hierarchical structure of a financial markets, and map out its complex dynamics as the market crash unfolds, to see if there are universal lessons that can be learned. We first study (I) the global market crash at the end of February 2007, known as the Chinese Correction, and which economists attribute to structural problems in the Chinese economy [96–98]. This is a good case study to understand how a correction on the Shanghai Stock Exchange triggers a contagion effect on other markets around the world, when they are unlikely to have exposure to Chinese stocks. This brings us to look at the two stock markets outside of China, on which Chinese stocks are listed: the Singapore Exchange (SGX) and the Hong Kong Stock Exchange. Our second case study focuses on (II) the October 2008 crash, arguably due to the collapse of the Lehman Brothers in September 2008. The goal is to analyze the dynamics of hierarchy changes systematically using the SGX as a case study. The SGX was chosen because Singapore has a very open economy and reacts readily to global market movements, yet it is a small market with fewer than 1000 stocks, making it possible to do a full cross-section analysis.

Fig. 2.1.1. Location of some important stock exchanges, together with their daily index from 2006 to 2007. These are the Dow Jones Index, DJI traded in New York Stock Exchange [red], the FTSE 100 Index, FTSE traded in London Stock Exchange [blue], together with five major Asian indices: (I) the Shanghai Composite Index, SHCI and (II) the Shanghai B-Shares Index (SHBS) both traded in Shanghai Stock Exchange (and Shengzhen Stock Exchange) [purple]; (III) the Hang Seng Index, HSI for Hong Kong Stock Exchange [green]; (IV) the Straits Times Index, STI for Singapore Exchange [magenta]; and (V) the Nikkei 225 Index, NKX traded in Tokyo Stock Exchange [orange].

2.2 The Chinese Correction of February 2007

2.2.1 Data

According to the prevailing narrative, the Chinese correction was triggered by a large-scale correction in the Shanghai Stock Exchange (SSE) in China, widely regarded as the epicenter of the Chinese Correction. The SSE crash on 27 February 2007 then led over the next few days to cascades of market crashes all over the world, even though these markets do not actively trade Chinese stocks. In order to understand the Chinese Correction in greater details, we looked at the correlations between the daily index values of the Shanghai Composite Index (SHCI), FTSE 100 Index (FTSE), and Dow Jones Index (DJI) as the London Stock Exchange and the New York Stock Exchange are among the most important stock markets in the world. In addition, since the SSE is an Asian market we also looked at the indices of three other major Asian markets: the Hang Seng Index (HSI) for the Hong Kong Stock Exchange (HKSE), the Straits Times Index (STI) for the Singapore Exchange (SGX), the Nikkei 225 Index (NKX) for the Tokyo Stock Exchange (TSE). Finally, we also included in the comparison the Shanghai B-Share Index (SHBS) for Chinese stocks that can be traded by foreigners. In Fig. 2.1.1, we show the locations of these stock exchanges on a world map (obtained from Google Maps, <https://www.google.com/maps/>) together with the daily indices for the period from 2006 to 2007 (downloaded from Stooq, <http://stooq.com/>).

We computed the correlation between stock price index $\mathbf{P}_i = \{p_i(t_1), p_i(t_2), \dots, p_i(t_T)\}$ and $\mathbf{P}_j = \{p_j(t_1), p_j(t_2), \dots, p_j(t_T)\}$ using Pearson correlation define as:

$$C_{Pearson}(\mathbf{P}_i, \mathbf{P}_j) = \frac{\frac{1}{T} \sum_{k=1}^T (p_i(t_k) - \bar{\mathbf{P}}_i) (p_j(t_k) - \bar{\mathbf{P}}_j)}{\sqrt{\frac{1}{T} \sum_{k=1}^T (p_i(t_k) - \bar{\mathbf{P}}_i)^2} \sqrt{\frac{1}{T} \sum_{k=1}^T (p_j(t_k) - \bar{\mathbf{P}}_j)^2}}, \quad (2.1)$$

where $\bar{\mathbf{P}} = \frac{1}{T} \sum_{k=1}^T p(t_k)$. In Fig. 2.2.1, we show the correlations between daily index of SHCI, SHBS, HSI, STI, NKX, FTSE and DJI four months before and the four months after the Chinese Correction. Before the market crash we can see that the SHCI is strongly correlated with SHBS, STI, and HSI. The correlations between SHCI and NKX, DJI, and FTSE are weak. We are surprised by the fact that NKX is weakly correlated with the SHCI, in spite of the former being an Asian index. After the crash, the correlations between SHCI with DJI and FTSE increased but those with all Asian indices decreased, even though the SHCI-STI and SHCI-HSI correlations remain strong. These results suggest that the Chinese Correction had a broader origin than just the SSE, and is likely to have included HKSE and SGX. This is perhaps not surprising because Chinese stocks are listed and traded on these two markets. While it would certainly be interesting to study the life course of the Chinese Correction on the SSE, we are more interested in how the correction on SSE produced a contagion effect on other stock markets.

Hence we analyzed 546 stocks from the SGX and 1173 stocks from the HKSE over a two-year period from January 2006 to December 2007, using tick-by-tick data downloaded from Thomson Reuters Tick History (<http://thomsonreuters.com/tick-history/>).

With this choice of time period, we have roughly one year's worth of data before the Chinese Correction at the end of February 2007, and also roughly one year's worth of data after the Chinese correction. On the SGX and HKSE, the average numbers of trades per day are 150 and 210 in 2006, 210 and 490 in 2007 [99] respectively. This is considerably lower than the New York Stock Exchange (NYSE) or London Stock Exchange (LSE), where we find on average more than one trade per minute. Because of this lower trading frequency, we process the tick-by-tick data to obtain prices at 30-min time intervals. With this choice of time horizon, we have SGX price time series that are 8096-element long, and HKSE price time series that are 6409-element long, after excluding non-trading days such as public holidays and weekends.

Fig. 2.2.1. Pearson correlations between the daily index of five major Asian indices: the Shanghai Composite Index (SHCI), Shanghai B-Shares Index (SHBS), Hang Seng Index (HSI), Straits Times Index (STI), and Nikkei 225 Index (NKX), along with FTSE 100 Index (FTSE) and Dow Jones Index (DJI), for (a) four months before and (b) four months after the 27 Feb 2007 Chinese Correction.

2.2.2 Clustering Algorithm: Partial Hierarchical Clustering

2.2.2.1 Digital Correlations

To detect hierarchical structures in stock markets, we typically start from the similarity (correlation) measures of the price time series. The most commonly used correlations are the Pearson, Spearman, and Kendall tau correlations [100–103]. Please refer to Section 5.1 for detail discussions on these correlation measures. Notably, the *Pearson correlation* (Eq.(2.1)) is sensitive to extreme events, due to the mean and standard deviation being easily distorted by extreme events. On the other hand, the *Spearman correlation*, which is the Pearson correlation of the ranked variable, is less sensitive to extreme events because it works with the rank statistic, but still suffers from potential artifacts of non-stationary time series because of all elements from the time series are given a global ranking. Finally,

the *Kendall tau correlation*, like the Spearman correlation also works with rank statistics and is hence less sensitive to extreme events. But unlike the Spearman correlation, the Kendall tau correlation works with relative rankings, that is comparing the number of concordant pairs and discordant pairs over all possible sub-intervals. This process is very computationally expensive, but allows the Kendall tau correlation to be less sensitive to non-stationarities in the time series, as ranking is compared for all possible pairs and sub-intervals.

In this study, we use an alternative measure of correlations between two time series, which we call the *digital correlation* ($C_{Digital}$). Given two stocks i and j , with price time series $\mathbf{P}_1 = \{p_1(t_1), p_1(t_2), \dots, p_1(t_T)\}$ and $\mathbf{P}_2 = \{p_2(t_1), p_2(t_2), \dots, p_2(t_T)\}$, the digital correlation measures how frequently $\mathbf{P}_1$ and $\mathbf{P}_2$ move together, but ignoring the magnitude of the movements. We define it as

$$C_{Digital}(\mathbf{P}_1, \mathbf{P}_2) = \sum_{i=1}^{T-1} \mathbb{I}_{\{\Delta p_1(t_i) \Delta p_2(t_i)\}}, \mathbb{I}_{\{x\}} \begin{cases} = 0 & \text{if } x \leq 0; \\ = 1 & \text{if } x > 0, \end{cases} \quad (2.2)$$

where $\Delta p(t_i) = p(t_{i+1}) - p(t_i)$. By only accounting for the movement direction, the digital correlation limits the influences of extreme events, which can be subjected to idiosyncratic shocks for individual stock. In addition, unlike the Kendall tau correlation which incorporates comparisons at all time scales, the digital correlation focus on the local directional comparisons, hence also less sensitive to non-stationarities in the time series but the latter is less computationally expensive (see Section 5.1 of the comparisons among similarity measures). We also understand how uncorrelated noise will give rise to a background digital correlation level of $C_{Digital} = (T - 1)/2$, and one can choose to normalize the correlation level. However, this will not affect our ability to pick out significant correlations or anti-correlations from pairs of time series with consistent co-movements and the hierarchical clustering results, because they will rise about or fall below the background level.

2.2.2.2 Partial Hierarchical Clustering

Even for small markets like the SGX and HKSE, there are too many pairs of correlations to examine in detail. Instead of all correlations, we would like to understand the market hierarchical structure only in terms of the strongest correlations. Many filtering schemes have been proposed to help us do this, including the use of a simple threshold [104, 105], the *minimal spanning tree* (MST) [29–33], and the *planar maximally filtered graph* (PMFG) [34, 37]. Unlike the MST and the PMFG, which produce fully connected graphs of the stocks, threshold filtering typically gives us a collection of disconnected subgraphs. Stocks have strong correlations with other stocks within the same subgraph, but weaker correlations with stocks in other subgraphs. We understand that this is a form of clustering analysis, where stocks are organized into distinct clusters based on their correlations. Clustering analysis can also be done starting from the MST or the PMFG, to obtain a *minimum spanning forest* (MSF) [34] or the a *directed bubble hierarchical tree* (DBHT) [36, 38] respectively.

In the complex networks literature, we also find community detection methods that can be used for clustering. For example, Rosvall and Bergstrom identify clusters in a weighted network by examining how information flows on the network [106], while Radichi *et al.* developed a *global statistic significance* (GloSS) filter that keeps links that are statistically significant [107]. These clustering methods all attempt to tackle the problem of multiple correlation scales in stock markets and other complex systems, where it is not reasonable to assume a single universal threshold like what is done for threshold filtering. In thinking about this problem of identifying clusters with different densities, we realized that if we were to just identify a single cluster, and not parallelize the process, the single cluster can be identified naturally.

Thus, we introduce the method of *partial hierarchical clustering* (PHC), which aims to identify one cluster at a time instead of all clusters in parallel. The graphical illustration and description of PHC can be found in Section 5.3. Here, let us briefly explain how the method works, by imagining a system that contains three distinct clusters $\mathbf{cl}^{(1)} = \{E_1^{(1)}, E_2^{(1)}, E_3^{(1)}, \dots\}$, $\mathbf{cl}^{(2)}$, and $\mathbf{cl}^{(3)}$. We start by identifying the pair of elements that are closest together, and use them as the *seed* to grow the cluster. For instance, suppose we identified the seed to be $\mathbf{cl}_{seed} = \{E_1^{(1)}, E_2^{(1)}\}$. To grow it, we add the next closest element, which obviously comes from cluster $\mathbf{cl}^{(1)}$. The cluster has now grown to become $\mathbf{cl}_{seed/grown} = \{E_1^{(1)}, E_2^{(1)}, E_3^{(1)}\}$. When all elements within $\mathbf{cl}^{(1)}$ are exhausted, there will be a sharp jump in the linkage distance when an element of another cluster is added (e.g. $\mathbf{cl}_{seed/grown} = \{E_1^{(1)}, E_2^{(1)}, E_3^{(1)}, \dots, E_k^{(j \in [2,3])}\}$). This sharp jump allows us to accurately identify $\mathbf{cl}^{(1)}$ as a natural cluster. After that, we remove elements in $\mathbf{cl}^{(1)}$ from the system and repeat the PHC algorithm again. We do this until we find no further sharp jumps in the graph of linkage distance as a function of cluster size, and treat the remaining elements as a natural cluster or discard them from the analysis. Lastly, let us note that PHC requires human experts to justify the sharp jumps. Because of this it is considered a supervised clustering algorithm and is difficult to be fully-automated.

2.2.3 SGX and HKSE Clusters and Superclusters

2.2.3.1 SGX and HKSE Clusters

We show the complete-linkage correlation level as a function of the cluster size for the SGX and HKSE in Fig. 2.2.2. In both markets, we find a sharp initial drop in the correlation level at or below 10 stocks. These sharp drops mark the natural boundaries of strongly-correlated clusters in the two markets. The correlation level then decreases smoothly as the cluster size increases, until we get to more than 20 stocks, where we again find sharp drops. In general, these drops are sharper in the HKSE than in the SGX. These sharp drops mark the natural boundaries of moderately-correlated superclusters, and show the existence of hierarchical self-organization in the two markets.

Previous studies have shown that financial instruments tend to cluster accordingly to their countries of origin. For example, using the daily values of financial indices and exchange rates researchers found clusters forming primarily by the geographical regions Europe, America, Asia and Oceania [108, 109]. In addition, studies also showed that

financial stocks tend to cluster according to the industry sectors they belong to [81,110–115]. In our study, we observe both phenomena: the financial stocks cluster according to the country of origin as well as the industry sector. For the period from Jan 2006 to Dec 2007, we find 10 such self-organized clusters in the SGX and 15 self-organized clusters in the HKSE. The component stocks in these clusters are shown in Appendix Table 2.A.1 and Table 2.B.1.

Fig. 2.2.2. Complete-linkage correlation level as a function of the cluster size for (a) the SGX starting from Cluster 8 (see Table 2.A.1), and (b) the HKSE starting from Cluster 10 (see Table 2.B.1). In this figure, the cluster boundary is indicated by downward-pointing triangle. The supercluster boundary is very distinct in the HKSE, but much less so in the SGX.

For the Singapore market, stocks in some clusters are either issued by the same company, or they are in the same industry sector. For example, Cluster 1 comprises Singtel, Singtel 10, and Singtel 100, all issued by Singapore Telecommunications Limited, whereas Cluster 2 comprises Singapore Airline and Singapore Airline 200, both issued by Singapore Airlines Limited. Other examples include Cluster 5, comprising CapitaLand and City Development, both of which are in the Properties sector, and Cluster 6 comprising DBS Group Holdings, United Overseas Bank, Overseas Chinese Banking Corporation, and Wing Tai Holdings. Except for Wing Tai Holdings, the other three component stocks of Cluster 6 are local banks. More interestingly, we find a purely Chinese Cluster 7, comprising China Hongxing Sports and Pine Agritech, as well as mixed local-Chinese Clusters 3, 4, and 8. In Cluster 3, Celestial Nutri-Foods is local, but China Sun Bio-Chem Technologies Group is Chinese. In Cluster 4, Mirach Energy is a Southeast Asian regional company, but Sky China Petroleum Services, Ferrochina, and China Sky Chemical Fibre are all Chinese stocks. Cluster 8 is the most interesting, comprising local stocks like Neptune Orient Lines, ASL Marine Holdings, Singapore Exchange, Rotary Engineering, UOL Group, and Tat Hong Holdings, but also an Australian investment holding company AusGroup, and two Chinese stocks, Cosco Corporation Singapore and Fibrechem Technologies. There is therefore nontrivial interactions between local stocks and Chinese stocks in the year before and the year after the Chinese Correction.

A similar result is being obtained in HKSE as well, stocks in most clusters are either in the same industry sectors, or closely related. For example, Cluster 1 consists of Cheung Kong, Hang Seng Bank, Hutchison, Sun Hung Kai Properties, MTR Corporation, and China Mobile. These are companies controlled by Hong Kong business tycoon Li Ka-Shing (Li Ka-Shing owns close to 40% shares in Cheung Kong, which in turn owns close to 50% of Hutchison Whampoa's shares), or companies (the Hong Kong and Shanghai Banking Corporation (HSBC) owns more than 60% of the shares of Hang Seng Bank, and holds about 40% net stake in Sun Hung Kai Properties) who have made major deals with his companies within the 2006–2007 period (Cheung Kong and Hutchison Whampoa buying over HSBC and Hang Seng's stakes in the e-commerce setup iBusiness on July 27, 2007; MTR awarded a HK\$7 billion residential property project to Cheung Kong on Nov 28, 2007). Unlike the SGX, however, local clusters formed the minority on the HKSE, whose trading activities seem to be dominated by the clusters of Chinese stocks. There are nine clusters whose components are from the same industry sectors. These are the three finance clusters, Clusters 2 and 8 (Chinese), and Cluster 11 (foreign), one properties Cluster 3 (Chinese), one mining Cluster (Chinese), two industries Clusters 7 and 9 (Chinese), one consolidated enterprises Cluster 6 (Hong Kong), and one airline Cluster 15 (Chinese). Apart from Cluster 1, there are four mixed-industries clusters. Of these mixed-industries clusters, Cluster 10 consists purely of Chinese stocks, whereas Clusters 12, 13, and 14 all consist of a mix of local and Chinese stocks.

2.2.3.2 SGX and HKSE Superclusters

We have now identified clusters using the PHC algorithm, and observed that those clusters are either dominated by local stocks or by Chinese stocks. In order to investigate how interactions between local and Chinese stocks precipitated crashes on the SGX and HKSE, we also need to examine correlations between the clusters. To do this, we first determine the correlation matrix for the identified natural clusters, using the complete-linkage correlations as the pairwise correlations between clusters:

$$C_{Complete}^{(Cluster)}(\mathbf{cl}_i, \mathbf{cl}_j) = \max(C_{Digital}(\mathbf{cl}_i(k_i), \mathbf{cl}_j(k_j))). \quad (2.3)$$

We then construct the MST and MSF of the clusters. These reveal higher-level self-organizations in the stock markets, and can be thought of as the superclusters of stocks.

Across the 2006–2007 two-year period, the SGX MST/MSF is shown in Fig. 2.2.3(a) while the HKSE MST/MSF is shown in Fig. 2.2.3(b). From Fig. 2.2.3(a), we see that the Chinese Clusters 3, 4, and 7 formed a supercluster along with the local Cluster 10 in the SGX. This then interacted with a largely local supercluster consisting Clusters 1, 2, 5, 6, and 8. In fact, we see that this interaction was primarily between the Chinese Cluster 3 and the hybrid Cluster 8, both of which contain manufacturing stocks. Over in the HKSE, the picture is different, because of the large number of Chinese clusters. From Fig. 2.2.3(b), we see a large and predominantly Chinese supercluster made up of the Chinese Clusters 3, 4, 5, 9, 10, 13, 15, the hybrid Clusters 12, 14, the foreign Cluster 11, and the local Cluster 6, as well as a hybrid supercluster made up of the Chinese Clusters 2,

7, 8, and the local Cluster 1. The interaction between the two superclusters was primarily between Chinese Clusters 7 and 10. Again, looking at the component stocks, we realized that there is significant overlap between the business areas of the two clusters.

Fig. 2.2.3. The MSTs of (a) SGX and (b) HKSE over the period January 2006 to December 2007. In this figure, solid links are common to both the MST and the MSF, while dashed links are MST links that are omitted to give the MSF, while the area of each circle is proportional to the number of stocks contained in the cluster. A red cluster indicates that the majority of its components are local stocks, a green cluster indicates that the majority of its components are Chinese stocks, and a dark green cluster contains nearly balanced mix of local and Chinese stocks. Finally, a blue cluster contains foreign stocks from other countries.

It is important to note that we observe stocks from similar sectors form clusters. Moving up one level higher, these stock clusters further organized into a local (SGX and HKSE) dominated supercluster and a Chinese stock dominated supercluster. Moving up again, these two superclusters are connected act as a single market. This result highlighting *the hierarchical structures in the financial markets, from market sectors, to country markets, on to global markets.*

2.2.4 Dynamics of SGX and HKSE Clusters

From the results in Section 2.2.3, we find indeed a clean self-organization into local and Chinese clusters and superclusters in SGX, and a more complicated self-organization in the HKSE. Interactions between local and Chinese clusters/superclusters might have caused the two markets to crash on that fateful week at the end of February 2007. Since we have tick-by-tick data for both markets, we would like to understand with higher time-resolution how these hierarchies of clusters and superclusters, and their interactions evolved across the February 2007 Chinese Correction. To do this, we slide a two-month time window across the two-year period, one month at a time. Within each two-month time window, we compute the digital correlations of 30-minute stock prices, then perform PHC and superclustering analyses.

2.2.4.1 Number of SGX and HKSE Clusters

Because of the large number of time windows and clusters, we do not show the compositions of the clusters and the structures of the superclusters. Instead, we show how the numbers of clusters and MSF superclusters evolve over time for the two markets in Fig. 2.2.4. We expect the number of clusters and superclusters to decrease before the market crash, as they merge into one giant cluster. After the market crash, we expect this giant cluster to fragment into smaller clusters, and thus the number of clusters and superclusters should increase after the market crash. However, this is not we see from Fig. 2.2.4. In general, we see peaks instead of dips in the number of clusters around the dates of the three significant market events. Fortunately, these peaks are all preceded by dips in the number of clusters. Therefore, we believe this because of the giant cluster reaches its maximum size, and start fragmenting, before the significant market event is reported. This is reasonable, as early movers on a market frequently reverse their positions on stocks in the market before the market crash. Herding by other traders copying these decisions may then have precipitated the market crash.

Fig. 2.2.4. The number of financial clusters over overlapping two-month windows across Jan 2006 to Dec 2007 for the SGX (red) and the HKSE (blue). In this figure, the error bars are the numbers of ambiguous clusters identified through partial hierarchical clustering. The green vertical dashed-line marks a market correction that occur around May/Jun 2006, the red vertical dashed-line marks the February 2007 Chinese Correction, while the black vertical dashed-line marks the Subprime Crisis in July 2007.

We compare the clusters dynamics of the SGX and the HKSE. For the May/Jun 2006 correction, the number of clusters in the SGX and HKSE peaked in the same two-month window, although the ‘giant cluster’ for this event appears to have started fragmenting

earlier in the HKSE. For the February 2007 Chinese Correction, the ‘giant cluster’ reached its maximum size (dip in number of clusters) very quickly in the HKSE. The fragmentation (peak in number of clusters) also proceeded very quickly. These two processes were slower in the SGX, and we suspect this is due to there being more Chinese stocks on the HKSE. In fact, the number of clusters peaked only after the Chinese Correction in the SGX. Finally, for the July 2007 Subprime Crisis, the number of clusters in both markets dipped instead, with a larger dip in the SGX. In choosing the two markets, we were guided by the suspicion that there will be early warning signs of the Chinese Correction in the Shanghai Stock Exchange (SSE), and also in the HKSE and the SGX, two other markets where Chinese stocks are traded. We also expected knee-jerk reactions in other markets affected by the contagion, and that early warning signals would not be seen in these markets. To put in simply, we expect early warning signs only in markets that were origins of the contagion. In a sense, our observations that the number of SGX and HKSE clusters (as a proxy indicator of a giant cluster) dipped before the February 2007 Chinese Correction, but dipped during the July 2007 Subprime Crisis is consistent with such a source-contagion picture.

2.2.4.2 Intra-cluster and Inter-cluster correlations

Besides that, in order to obtain more insights from these clusters, we need to also consider correlations between clusters, and correlations between components of a cluster. We define the average intra-cluster correlation within cluster $\mathbf{cl}_i$,

$$C_{intra}^{(Cluster)}(\mathbf{cl}_i) = \frac{2}{N(\mathbf{cl}_i)(N(\mathbf{cl}_i) - 1)} \sum_{k_i=1}^{N(\mathbf{cl}_i)} C_{Digital}(\mathbf{cl}_i(k_i), \mathbf{cl}_i(k_i)), \quad (2.4)$$

and the inter-cluster correlation between clusters $\mathbf{cl}_i$ and $\mathbf{cl}_j$

$$C_{inter}^{(Cluster)}(\mathbf{cl}_i, \mathbf{cl}_j) = \frac{1}{N(\mathbf{cl}_i)N(\mathbf{cl}_j)} \sum_{k_i=1}^{N(\mathbf{cl}_i)} \sum_{k_j=1}^{N(\mathbf{cl}_j)} C_{Digital}(\mathbf{cl}_i(k_i), \mathbf{cl}_j(k_j)) \quad (2.5)$$

obtained for a particular time window. Here $C_{Digital}(\mathbf{cl}_i(k_i), \mathbf{cl}_j(k_j))$ is the digital correlation between the k_i^{th} element in cluster $\mathbf{cl}_i$ and the k_j^{th} element in cluster $\mathbf{cl}_j$. While, $N(\mathbf{cl}_i)$ and $N(\mathbf{cl}_j)$ the number of components in cluster $\mathbf{cl}_i$ and $\mathbf{cl}_j$. The standard errors of these average correlations are

$$\delta C_{intra}^{(Cluster)}(\mathbf{cl}_i) = \sqrt{\langle [C_{Digital}(\mathbf{cl}_i, \mathbf{cl}_i)]^2 \rangle - \langle C_{Digital}(\mathbf{cl}_i, \mathbf{cl}_i) \rangle^2} \quad (2.6)$$

and

$$\delta C_{inter}^{(Cluster)}(\mathbf{cl}_i, \mathbf{cl}_j) = \sqrt{\langle [C_{Digital}(\mathbf{cl}_i, \mathbf{cl}_j)]^2 \rangle - \langle C_{Digital}(\mathbf{cl}_i, \mathbf{cl}_j) \rangle^2} \quad (2.7)$$

where $\langle \mathbf{x} \rangle$ is the average of $\mathbf{x}$.

When there are NC number of clusters obtained, to track the global correlation

changes at the cluster level, we can average over all intra-cluster correlations ($\langle C_{intra}^{(Cluster)}(\mathbf{cl}_i) \rangle_{i=1}^{NC}$) and average over all inter-cluster correlations ($\langle C_{inter}^{(Cluster)}(\mathbf{cl}_i, \mathbf{cl}_j) \rangle_{i,j=1; i \neq j}^{NC}$). However, we must not use the standard deviations computed over all intra-cluster and inter-cluster correlations, because these are simply not uniformly spread over the range of correlations as suggested by the standard deviations. Instead, we define the standard deviation for intra-cluster correlation as

$$\delta \langle C_{intra}^{(Cluster)} \rangle = \langle C_{intra}^{(Cluster)} \rangle \left\langle \frac{\delta C_{intra}^{(Cluster)}(\mathbf{cl}_i)}{C_{intra}^{(Cluster)}(\mathbf{cl}_i)} \right\rangle_{i=1}^{NC}, \quad (2.8)$$

where the average of the coefficient of variation $\delta C_{intra}^{(Cluster)}(\mathbf{cl})/C_{intra}^{(Cluster)}(\mathbf{cl})$ is over all clusters $\mathbf{cl}_i$ for $i = 1$ to $i = NC$. On the other hand, the standard deviation for inter-cluster correlation is

$$\delta \langle C_{inter}^{(Cluster)} \rangle = \langle C_{inter}^{(Cluster)} \rangle \left\langle \frac{\delta C_{inter}^{(Cluster)}(\mathbf{cl}_i, \mathbf{cl}_j)}{C_{inter}^{(Cluster)}(\mathbf{cl}_i, \mathbf{cl}_j)} \right\rangle_{i,j \in MST; i \neq j}. \quad (2.9)$$

There are many pairs of inter-cluster correlations, we can of course average the coefficient of variation $\delta C_{inter}^{(Cluster)}(\mathbf{cl}_i, \mathbf{cl}_j)/C_{inter}^{(Cluster)}(\mathbf{cl}_i, \mathbf{cl}_j)$ over all cluster pairs. However, we should be more selective, as not all cluster pairs are equally important. In general, the most important inter-cluster correlations are those forming the minimal spanning tree (MST) of the clusters. We therefore average $\delta C_{inter}^{(Cluster)}(\mathbf{cl}_i, \mathbf{cl}_j)/C_{inter}^{(Cluster)}(\mathbf{cl}_i, \mathbf{cl}_j)$ only over the MST links.

In Fig. 2.2.5 we show the average intra-cluster correlation level and average inter-cluster correlation level in the SGX and HKSE. We expect the correlation level of the financial clusters to increase when closer to the financial crisis, reached a maximum during the market crash, and then falls off after the market crash. In addition, the gap between the average intra-cluster correlation level and the average inter-cluster correlation level decrease as we approach the market crash, reach a minimum during the market crash, and then increase again after the market crash. Indeed, increased correlations are found in both the SGX and HKSE for the three significant market events. This phenomena is most clear on the SGX, as we see two broad humps for the May/Jun 2006 market correction and the February 2007 Chinese Correction, and a sharp peak for the July 2007 Subprime Crisis. However, the HKSE only shows weak correlational reactions to the three significant market events. Finally, on the SGX the May/Jun 2006 market correction and the Feb 2007 Chinese Correction showed up as broad humps in the clustered correlations, but the Jul 2007 Subprime Crisis showed up as a sharp peak. This may suggests that the May/Jun 2006 and February 2007 events were endogenous to the SGX, while the July 2007 event was an exogenous shock.

As of now, we have investigated how these hierarchical structures in the SGX and the HKSE changed across the February 2007 Chinese Correction, by performing partial hierarchical clustering (PHC) based on the digital correlations between stocks. Particularly, when closer to the market crashes, we observed that the number of clusters dipped with raising intra-cluster and inter-cluster correlations. This suggests there are richer infor-

mation that can be revealed by using higher time resolution, including more stocks, and identifying more (natural) clusters. Unfortunately, this is difficult to do using the PHC for the reasons: (I) PHC is a supervised clustering algorithm, requiring judgments and inputs from human experts, which is not practical to use when dealing with a large number of time windows and clusters; (II) To identify natural clusters, we are relying on sharp changes in the linkages. Yet, often the distance increment is smooth, especially after few natural clusters have been discovered. Consequently, we need to discard a considerable number of stocks and information from the analysis. With these, it is difficult to construct a broader picture of how financial clusters interact, especially toward the market crashes. New visualization and analytical tools are necessary to meaningfully extract information from the clusters at different times. In the next Section, we will discuss our attempts to cope with these limitations, using the Lehman Brothers Crisis in September 2008 as a second case study.

Fig. 2.2.5. The average intra-cluster correlation level and average inter-cluster correlation level over overlapping two-month windows across Jan 2006 to Dec 2007 for (a) SGX and (b) HKSE. In this figure, the error bars of the intra-cluster and the inter-cluster correlations are estimated using Eq.(2.8) and Eq.(2.9). The green vertical dashed-line marks a market correction that occur around May/June 2006, the red vertical dashed-line marks the Feb 2007 Chinese Correction, while the black vertical dashed-line marks the Subprime Crisis in Jul 2007.

2.3 The Lehman Brothers Crisis in September 2008

2.3.1 Data

In this second case study, we work with data from the Singapore Exchange (SGX). There are three reasons for this choice. First, even though the model market econophysicists study is the New York Stock Exchange (NYSE), we are much more familiar with the stocks listed on the SGX, and the profiles and outlooks of their companies. Second, the SGX is a small market with fewer than 1000 stocks, making it possible to do a full cross-section analysis. Third, although the Singapore is an emerging market, it has a very open economy and strongly coupled to financial markets around the world, thus the SGX reacts readily to global market movements. For instance, the Straits Times Index (STI) is highly correlated with Dow Jones Industrial Average Index ($\rho : 0.91$ from 2006 to 2009). Based on the STI that tracks the 30 top performing stocks listed on the SGX, we know that the bourse reacted to the Lehman Brothers Crisis, losing 45% of its value from early August to the end of October 2008. Therefore, we chose the period of study to be from January 2008 to December 2009, roughly centered around the Lehman Brother Crisis.

Tick-by-tick data for more than 800 SGX stocks were downloaded from the Thomson-Reuters Tick History Database (<http://thomsonreuters.com/tick-history/>). We first filtered out stocks with very low trading frequencies, before extracting from the tick-by-tick data price time series at five-minute intervals. After cleaning the data, 561 stocks remains, with about 40,240 time intervals corresponding to 503 trading days. To understand how the SGX dynamics evolve across the market crashes, we break the 40,240-point time series into one-month time windows that overlap with their predecessors and successors by two weeks. Each time window contains more than 2,000 data points, which is sufficient for computing reliable correlation coefficients, as $Q = T/N = 2000/561 \approx 4 > 1$ [116]. The overlapping time windows also help us track hierarchy changes in the market in a smooth fashion.

2.3.2 Clustering Algorithm: Interaction Hierarchical Clustering

2.3.2.1 Correlation Measure

In this study we compute the Pearson correlation coefficient [100, 101] between stock prices. Given two stocks i and j , with price time series $\mathbf{P}_1 = \{p_1(t_1), p_1(t_2), \dots, p_1(t_T)\}$ and $\mathbf{P}_2 = \{p_2(t_1), p_2(t_2), \dots, p_2(t_T)\}$, their Pearson correlation is given by

$$C_{\text{Pearson}}(\mathbf{P}_1, \mathbf{P}_2) = \frac{\frac{1}{T} \sum_{i=1}^T (p_1(t_i) - \bar{P}_1) (p_2(t_i) - \bar{P}_2)}{\sqrt{\frac{1}{T} \sum_{i=1}^T (p_1(t_i) - \bar{P}_1)^2} \sqrt{\frac{1}{T} \sum_{i=1}^T (p_2(t_i) - \bar{P}_2)^2}}, \quad (2.10)$$

where $\bar{P} = \frac{1}{T} \sum_{i=1}^T p(t_i)$. These correlations can be organized into a correlation matrix $\mathbf{C}_{\text{Pearson}}$ and visualized as a color map. In such color maps, red represents strong correlations, blue indicates strong anti-correlations, and green means nearly uncorrelated. For other correlation measures and their comparison readers can refer to Section 5.1.

Fig. 2.3.1(a) show the color map of an actual correlation matrix, which contains hidden structures that we want to discover. In the literature we find many correlation filtering methods to discover these hidden structures: threshold filtering [105, 117], the minimum spanning tree (MST) [29], and the planar maximally filtered graph (PMFG) [34]. In addition, there are also cluster analysis methods that can discover hidden hierarchy and organization, for example centroid-based clustering [118], connectivity based clustering (hierarchical clustering) [119–122] and distribution-based clustering [123].

Fig. 2.3.1. The correlation matrix (a) for 561 stocks in the SGX from Jan 2008 to Dec 2009, where red indicates strong correlation, blue indicates strong anti-correlation and green means nearly uncorrelated. As the stocks are arranged in alphabetical order no structure can be discerned. After we perform CLHC, and rearrange the financial stocks in the order of the dendrogram, the structure in the CLHC seriated matrix (b) becomes obvious. The diagonal blocks are red, indicating strong intra-cluster correlations. However, we also see red off-diagonal blocks. This suggests strong interactions between correlations between clusters, which ideally should be grouped together or at least ordered close to each other. In the IHC seriated matrix (c) where we reordered the CLHC clusters to account for their interactions, we see less red color outside of the diagonal blocks, which are now more consistently red (i.e. intra-cluster correlations are more homogeneous).

2.3.2.2 Interaction Hierarchical Clustering

In this thesis, we chose to perform complete-linkage hierarchical clustering mainly because it provides more information about the hierarchical structures in the financial markets as described in Section 2.1. The agglomerative hierarchical clustering [29, 124] of N elements, start with N clusters containing one element each. Then the two most similar clusters (shortest distance between clusters) are merged into a larger cluster, and this process is repeated until we end up with a single cluster of N elements. When we compare two elements, the distance between them is unambiguous. However, for two clusters each containing more than one element, there are many ways to define measures of distances between them (or linkages). The common measures are furthest distance (*complete linkage*), shortest distance (*single linkage*), and average distance (*average linkage*) between elements within the clusters (refer Section 5.2.1 for the graphical illustrations). Among these linkages, Musmeci *et al.* shows that complete-linkage hierarchical clustering (CLHC)

perform the best in segregating stocks into their corresponding market sectors, as single-linkage hierarchical clustering (SLHC) often give heterogeneous clusters (e.g. majority of stocks are grouped into a single cluster), and similar for average-linkage which is a hybrid of CLHC and SLHC [38].

The correlation matrix in Fig. 2.3.1(a) contains much more information than its hierarchical clustering dendrogram suggests. In order not to lose this information, we reorder the rows and columns of the correlation matrix using the linear ordering suggested by the dendrogram. The reordering process is called *seriation* [125–127], and we see that the seriated correlation matrix in Fig. 2.3.1(b) is more structured and we can identify a few clusters just through visual inspection. This is evidence of self-organization within the financial market. Although CLHC does separate stocks into groups based on their correlations, the organization is not very clean. As shown in Fig. 2.3.1(b), there remains strong correlations in the off-diagonal blocks of the CLHC-seriated matrix. These strong off-diagonal correlations contain information on how clusters interact with each other, and are thus important.

In order to retain such information, we perform clustering on the CLHC clusters, focusing on the off-diagonal correlations, and call this higher-level clustering procedure *interaction hierarchical clustering* (IHC). The objective of IHC is to reorder the sequence of clusters, to obtain stronger diagonal blocks and weaker off-diagonal blocks. In order to do this, we (I) first determine the robust CLHC clusters, that meet three criteria: (i) a cluster must form only in the positive similarity (positive correlation) zone, (ii) no cluster will be broken into smaller clusters in the strong positive similarity (high positive correlation) zone, and (iii) most importantly the change in similarity for the newly added variables (clusters) is greater than the robustness length. The first two criteria are intuitive, as we should not define a cluster that contains dissimilar variables, nor break strongly-connected variables into multiple clusters. The third criterion is included to identify robust clusters, when the next added element (cluster) causes a sudden increase in the heterogeneity within the robust cluster. Next, (II) we regroup those robust clusters according to their interaction distance

$$ID(\mathbf{cl}_i, \mathbf{cl}_j) = \frac{1}{N_i N_j} \sum_{k_i=1}^{N_i} \sum_{k_j=1}^{N_j} 1 - C_{Pearson}(\mathbf{cl}_i(k_i), \mathbf{cl}_j(k_j)), \quad (2.11)$$

where $C_{Pearson}(\mathbf{cl}_i(k_i), \mathbf{cl}_j(k_j))$ is the correlation between the k_i th element in robust cluster $\mathbf{cl}_i$ and the k_j th element in robust cluster $\mathbf{cl}_j$. Like hierarchical clustering the pair of robust clusters with the shortest distance according to Eq.(2.11) will be merged first to form a bigger cluster, and thereafter we iterate this process until only one cluster remains. Please refer to Section 5.2.2 for the full description of interaction hierarchical clustering. As we can see comparing Fig. 2.3.1(b) and Fig. 2.3.1(c), diagonal blocks are better defined in the IHC seriated matrix than in the CLHC seriated matrix. This is because clusters with strong interactions, but are far apart in the dendrogram because of the history dependence of CLHC, are now brought closer together.

2.3.2.3 Cluster Identification and Fusion-Fission Diagram

So far we only discussed the constructions of the complete-linkage hierarchical dendrogram, finding robust clusters, and constructing the interaction-hierarchical dendrogram. Although robust clusters themselves can act as the identified clusters, if we want to further group those robust clusters we need an alternative approach. One of the most common ways is define a cutoff point in the dendrogram, so that the groups that remain below this cutoff are considered the financial clusters. In order to determine this cutoff point unambiguously, we compare the statistical similarity of intra-cluster correlations and inter-cluster correlations, by progressively lowering the cutoff threshold. We select the optimum cutoff to be the one at which the new inter-cluster correlations switch from being more similar to existing inter-cluster correlations to being more similar to the remaining intra-cluster correlations (refer to Section 5.2.3 for a full description of the procedure).

Fig. 2.3.2. In this figure, we show the 35 CLHC clusters and their member stocks in a given time window, that is clustered as seven identified financial (robust) clusters. Each financial cluster is then represented by a circle whose radius is proportional to the number of member stocks. In addition, we also color each circle based on the percentages of member stocks that have been randomly assigned the colors red, green, and blue.

As we are interested in the dynamics of the hierarchical organization of the financial market, we need to examine these over many sliding time windows. The large number of sliding windows makes keeping track of the clusters and their time evolutions tedious and difficult. Thus, to visualize the dynamics of these clusters, we introduce the *fusion-fission diagram*. Similar diagrams have been introduced by Mantegna *et al.* [128] in the econophysics and sociophysics literature. In our fusion-fission diagram, we show some of the larger robust clusters, represented by circles whose radii are directly proportional to the sizes of the robust clusters. We also show how many elements from each robust cluster in time window t go to robust clusters in time window $t + 1$. In addition, we assign

each stock randomly to one of the colors red, green, and blue, and thereafter color the circles based on the percentages of red, green, and blue elements. To illustrate how we construct the fusion-fission diagram, we show in Fig. 2.3.2 the IHC dendrogram for 35 CLHC clusters, that is grouped as seven identified financial (robust) clusters. Finally, the flow of the elements from one financial cluster into another financial cluster is indicated by a line, whose thickness is proportional to the number of elements that flow between them.

Fig. 2.3.3. (a) Straits Times Index downloaded from Yahoo Finance, <https://sg.finance.yahoo.com>. The vertical dashed lines are correspond to the lows index due to the Lehman Brothers Crisis on 27 Oct 2008 (red) and the Asian Correction on 9 Mar 2009 (blue), and also the peak on 5 Jan 2009 after the Lehman Brothers Crisis (green). (b) IHC-seriated correlation matrices before (Jun/2008 and Jul/2008), during (Sept/2008 and Oct/2008), and after (Nov/2008, Dec/2008) the Lehman Brothers Crisis where the Straits Times Index reached a low on 27 Oct 2008. (c) IHC seriated correlation matrices before (Jan/2009 and Feb/2009), during (Mar/2009 and May/2009) and after (Jun/2009 and Jul/2009) the Asian Correction.

With a fusion-fission diagram, we are able to condense a tremendous amount of information on the market dynamics into a simple diagram that is visually meaningful. Specifically, the number of circles and their sizes tell us how stocks are organized at any time, while the thicknesses of the lines tell us how strong the mixing dynamics is in the financial market. This mixing information is also encoded in the colors of the clusters. Such a diagram thus helps us visually detect key events in the stock market before and after the financial crisis.

2.3.3 Dynamics of SGX Clusters

In this study, we are interested in the description of financial market dynamics in terms of clusters. In their 2012 paper, Munnix *et al.* found that the US stock market switches abruptly between different ‘states’, which can be distinguished from each other by their correlation structures [129]. Furthermore, in 2015 Musmeci *et al.* found using the DHBT method strong mixing in the compositions of clusters in the US stock market, and proceeded to measure the rates at which the market reorganizes itself, and also how persistent strong correlations between stocks are [88]. However, they stop short of presenting this information in the form of a fusion-fission diagram, which we feel gives the most compact visual representation of the complexity and non-stationarity in financial markets. This dynamics is reminiscent of John Holland’s ideas that complex adaptive systems achieve perpetual novelty by taking different combinations of a small set of building blocks [130].

To tell a compelling story of how the Lehman Brother Crisis unfolded in the SGX, we split our two-year data set into one-month time windows that overlap with their predecessors and successors by two weeks. In total, we have 46 such time windows from Jan 2008 to Dec 2009. We then perform IHC within each time window to determine the robust clusters, and construct the fusion-fission diagram of the SGX within this time period. Within this period, two crises impacted the Singapore stock market, the Subprime Mortgage Crisis and the Asian Correction. From Fig. 2.3.3(a), we that after Jan 2008 the Straits Times Index reached its peak on 6 May 2008 with 3248.75. Thereafter, the index slid, and reached its lowest for 2008 on 27 Oct 2008 at 1600.29, as a result of the Lehman Brothers Crisis. After that, there are weak signs of recovery, but due to the Asian Correction, the index again drops from the after-crisis peak 1924.87 (5 Jan 2009) to the lowest of 1456.95 (9 Mar 2009). Fortunately, the Singapore market eventually recovered with a bull run after the Asian Correction. Over the period 9 Mar to 3 Aug 2009, the index saw an 85% rise to 2681.64. In Fig. 2.3.3(b~c), we show the IHC-seriated correlation matrices before, at the lowest, and after the two financial crises.

From the fusion-fission diagram shown in Fig. 2.3.4, we find repeated episodes of smaller clusters agglomerating into a giant cluster, which thereafter fragment into smaller clusters. In particular, after the fission episode in May 2008, the giant cluster persisted till Oct 2008 (the Lehman Brothers Crisis), where we find a sustained series of fission that lasted till Mar 2009 (the Asian Correction). These observations agreed with the findings of Viegas *et al.*, who found that a decrease of diversity in the market can cause market crisis [131]. After Mar 2009, a giant cluster re-emerges, interrupted by fission episodes that were closer and closer together. This suggests that the Lehman Brothers Crisis and Asian Correction are not separate events, but merely the beginning and end of an extended crisis phase. Consistent with this critical transition interpretation is the slowing down of the dynamical time scale before the Lehman Brothers Crisis, and the quickening of the dynamical time scale after the Asian Correction. During the recovery phase, we also observed the formation and persistence of a giant cluster, but its size is smaller than the giant cluster formed in Oct 2008.

Fig. 2.3.4. The fusion-fission diagram of the SGX from Jan 2008 to Dec 2009. This diagram is composed of three parts: (I) the innermost polar plot of the STI, where the index value is given by the radial distance, and additionally color coded such that blue is the highest (3249), green is intermediate (2353), and red is the lowest (1457). In this plot we also show as dashed lines the date (27 Oct 2008, cyan) when the STI is at its lowest due to the Lehman Brothers Crisis, and on 9 Mar 2009 (magenta) due to the Asian Correction; (II) the start and end dates of the time window, color coded by the average value of the STI within the time window; and (III) the robust clusters within each time window, and the exchanges of members with robust clusters of the previous and subsequent time windows.

2.3.4 Mixing Degree and Persistence

2.3.4.1 Mixing Degree

As we can see from Fig. 2.3.4, the fusion-fission diagram allows us to take what Gell-Mann called a *crude look at the whole* [132]. But it also helps us formulate the right detailed questions to ask. First, based on the visual representation of member stock exchanges between clusters, we realized that there are episodes of extensive mixing. Second, related to mixing and based on the colors of the clusters, we suspected that stocks do not remain in the same clusters for very long. Therefore we (I) quantify the degree of mixing in the SGX, based on the robust clusters identified; (II) we also quantify the persistence of strong correlations between stocks.

Given two cluster lists at time t_1 and t_2 as $\mathbf{CL}(t_1) = \{\mathbf{cl}_1(t_1), \mathbf{cl}_2(t_1), \dots, \mathbf{cl}_u(t_1)\}$ and $\mathbf{CL}(t_2) = \{\mathbf{cl}_1(t_2), \mathbf{cl}_2(t_2), \dots, \mathbf{cl}_v(t_2)\}$. Here $\mathbf{cl}$ represent a cluster contains at least one element, and $\cup_{i=1}^u \mathbf{cl}_i(t_1) = \cup_{j=1}^v \mathbf{cl}_j(t_2) = \mathbf{E}$, where $\mathbf{E} = \{E_1, E_2, \dots, E_N\}$ is the elements set. To examine the evolution of clusters, we can measure the cluster mixing degree between $\mathbf{CL}(t_1)$ and $\mathbf{CL}(t_2)$ using the *Adjusted Rand Index (ARI)* [47], which is defined as:

$$ARI(\mathbf{CL}(t_1), \mathbf{CL}(t_2)) = \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\binom{N}{2}}}{\frac{1}{2} \left(\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2} \right) - \frac{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\binom{N}{2}}}, \quad (2.12)$$

where N is the number of elements, n_{ij} is the number of shared elements between cluster $\mathbf{cl}_i(t_1)$ and $\mathbf{cl}_j(t_2)$, $n_{i.}$ is number of elements in cluster $\mathbf{cl}_i(t_1)$, and $n_{.j}$ is number of elements in cluster $\mathbf{cl}_j(t_2)$ (refer to Section 5.4.1 for the details). When $ARI = 0$ means the elements are randomly reclustered, and $ARI = 1$ when they remain in the same clusters for both t_1 and t_2 . Because of this, the mixing degree can be defined as $1 - ARI$.

Fig. 5.4.2 shows the mixing degree given by $1 - ARI$ (blue), and the results agreed with our expectations of high mixing degree after the financial crisis and low mixing degree during the bullish and bearish markets. In early 2008, the mixing degree was high and oscillating, and then started to decrease from Jul 2008 as SGX experienced the financial downturn. More importantly, right after the 2 Oct 2008-31 Oct 2008 window (lowest STI due to Lehman Brothers Crisis), we observe high mixing degree over a few months, followed by low mixing degree from 2 Mar 2009-31 Mar 2009 window (lowest STI due to Asian Correction then bullish market right after the correction) and back to oscillating high mixing degree in late 2009. In addition, we also learned that when the mixing degree is high in SGX, market participants were forced to update their portfolios more frequently. Moreover, as we can see most of the clusters are actually very small, and we are more interested in the mixing between the large clusters. Thus, we introduce another measurement for mixing degree named *Degree of Mixing (DoM)*, that is able to measure the mixing degree for a single cluster, please refer to Section 5.4.2 for the definition and details. The intuition for *DoM* is the mixing degree of a cluster $\mathbf{cl}(t_2)$ should be highest

($DoM = 1$) when its elements are equally contributed (size normalized) by all clusters from $\mathbf{CL}(t_1)$, and lowest when it is identical to or a subset of one of the clusters from $\mathbf{CL}(t_1)$ ($DoM = 0$).

Fig. 2.3.5. Mixing degree of SGX financial clusters, determined using different methods: (green) DoM using all robust cluster, (red) DoM using three largest financial clusters, and (blue) $1 - ARI$ based on all robust clusters. The vertical dashed lines indicate the windows where the STI is at the lowest due to the Lehman Brothers Crisis (2 Oct 2008 to 31 Oct 2008, cyan) and the Asian Correction (2 Mar 2009 to 31 Mar 2009, magenta).

Other than $1 - ARI$, Fig. 5.4.2 also shows the mixing degree measured by DoM using all financial clusters (green) and DoM using three largest financial clusters (red). As with for ARI , we run 5000 Monte Carlo simulations for each time window in the manner described in Section 5.4.3 to estimate the standard error of the DoM using all clusters and DoM using three largest clusters. Comparing these two measurements, we noticed that right after the Lehman Brothers Crisis, the mixing degree measured by DoM with all clusters is lower than that measured by DoM that only consider three largest clusters, which means that the strong mixing is mainly due to the large clusters. Meanwhile, during the bullish market the both DoM mixing degrees are close to each other, suggesting that the mixing degree due to large clusters is comparable to that of small clusters. All in all, the general trends are similar to $1 - ARI$, as the DoM measurements also show low mixing degrees during the bearish and bullish market, and high mixing degree right after the Lehman Brothers Crisis. We suspect these observations are due to behavioral synchronization of market participants during bullish and bearish markets, but after the financial crisis, there is a larger diversity in the market participant responses towards market recovery.

2.3.4.2 Persistence

We see high mixing degrees among SGX clusters from Jan 2008 to Dec 2009, as most of the time the mixing degree is above 0.7. This leads naturally to a question about the persistence of the clusters, such as how long a pair of stocks will remain together in the same cluster. This question is particularly important for portfolio selection, due to the important role played by the correlation in risk diversification. Whenever the correlation coefficient between two financial stocks is less than 1, there is a potential for diversifying the risks. Yet, correlations are frequently unstable, thus portfolios need to be updated from time to time, using the lifetimes of these correlations as a gauge on how frequently this is to be done.

For simplicity, we determine the *persistence of clusters* by measuring the lifetime of a pair of financial stocks that continuously remain within the same cluster. This is done by counting the number of pairs that remain in the same clusters across several time windows. For example, stocks FS_A and FS_B are within the same cluster at time t_n , t_{n+1} , and t_{n+2} ; but broken up into two different financial clusters at t_{n+3} and back to same financial cluster again at t_{n+4} . Based on this information, FS_A and FS_B contribute one pair of stocks from t_n to t_{n+2} , zero pair of stocks in t_{n+3} , and one pair of stocks in t_{n+4} , hence contribute one pair of stocks that last for 3 time window and another one last for 1 time window. The longer the stocks stay together the more persistent their correlation, which implies the market participants prefer to trade these stocks simultaneously.

Fig. 2.3.6. Distribution of the number of stock pairs that continuously remain within a same cluster across different duration. In this semilog plot, we show the double exponential fit $f(t) = A \exp(-t/\tau_1) + B \exp(-t/\tau_2)$ (red dashed curve) to the data (black solid curve), and compare it against a single exponential fit $f(t) = C \exp(-t/\tau_A)$ (blue dashed curve, offset). The double exponential decay function fits the data better and suggests that there are two scales governing the decay of correlations between stocks.

Since there are 46 time windows and there are 561 financial stocks, thus we can track in total $561 \times 561 \times 46$ different pairs of stocks from Jan 2008 to Dec 2009. Figure 2.3.6

shows the distribution of the number of pairs remaining within a same cluster continuously across different duration. On a semilog scale, we see that the distribution can be explained by a double exponential decay with two different time scales. Since each time window is one month long with two-week overlaps with the time windows before and after it, the data points are two weeks apart. The double exponential decay function suggests that there are two groups of correlations, the first with a short lifetime (2.94 ± 0.20 weeks) and long lifetime (5.85 ± 0.84 weeks). This implies that if a portfolio consists of stocks with short-lived correlations, the portfolio needs to be updated at least once every three weeks. Similarly, for a portfolio consisting stocks with long-lived correlations, the portfolio needs to be updated once every six weeks. In general, the high mixing degree among SGX clusters and short-lived correlations between stocks suggests that trading strategies need to be updated frequently.

Appendix

2.A Component stocks of the 10 SGX clusters

Table 2.A.1. Component stocks of the 10 SGX clusters between Jan 2006 and Dec 2007.

Cluster	Stock Code	Stock Name	Market Sector	Remarks
1	Z74	Singtel Singtel10 Singtel100	Communication Communication Communication	Singapore Singapore Singapore
2	C6L	Singapore Airline Singapore Airline 200	Airline Airline	Singapore Singapore
3	C56 C86	Celestial Nutri-Foods China Sun Bio Chem Tech	Manufacturing Manufacturing	Chinese Chinese
4	C68 W81 F33 E90	Mirach Energy Sky China Petroleum Svcs Ferrochina China Sky Chemical Fibre Co	Services Services Manufacturing Manufacturing	Hong Kong Chinese Chinese Chinese
5	C31 C09	CapitaLand City Development	Properties Properties	Singapore Singapore
6	D05 U11 O39 W05	DBS Gp Hldg United Oversea Bank Oversea Chinese Banking Corp Wing Tai Hldg	Finance Finance Finance Properties	Singapore Singapore Singapore Singapore
7	BR9 P39	China Hongxing Sports Pine Agritech	Manufacturing Manufacturing	Chinese Chinese
8	F83 N03 A04 S68 5GJ 5EE F12 R07 U14 T03	Cosco Corp (Singapore) Neptune Orient Lines ASL Marine Hldg Singapore Exchange Aus Group Aqua-Terra Supply Co Fibrechem Technologies Rotary Engineering UOL Gp Tat Hong Hldg	Shipping Shipping Manufacturing Finance Manufacturing Services Manufacturing Construction Properties Commerce	Singapore Singapore Singapore Singapore Foreign Singapore Chinese Singapore Singapore Singapore
9	L28 L46	LongCheer Holding Luzhou Bio-Chem Technology	Services Manufacturing	Chinese Chinese
10	F20 T06	Federal International (2000) Tiong Woon Corp Holdings	Commerce Shipping	Singapore Singapore

2.B Component stocks of the 15 HKSE clusters

Table 2.B.1. Component stocks of the 15 HKSE clusters between Jan 2006 and Dec 2007.

Cluster	Stock Code	Stock Name	Market Sector	Remarks
1	0001	Cheung Kong	Properties	Hong Kong
	0011	Hang Seng Bank	Finance	Hong Kong
	0013	Hutchison	Consolidated Enterprises	Hong Kong
	0016	Sun Hung Kai Ppt	Properties	Hong Kong
	0066	MTR Corporation	Utilities	Hong Kong
	0941	China Mobile	Consolidated Enterprises	Chinese
2	0762	China Unicom	Consolidated Enterprises	Hong Kong
	2318	Ping An	Finance	Chinese
	2628	China Life	Finance	Chinese
3	0688	China Overseas	Properties	Chinese
	1109	China Res Land	Properties	Chinese
	3383	Agile Property	Properties	Chinese
4	2899	Zijin Mining	Miscellaneous	Chinese
	3330	Lingbao Gold	Miscellaneous	Chinese
5	0323	Maanshan Iron	Industries	Chinese
	0347	Angang Steel	Industries	Chinese
6	0019	Swire Pacific A	Consolidated Enterprises	Hong Kong
	0087	Swire Pacific B	Consolidated Enterprises	Hong Kong
7	0857	PetroChina Co	Industries	Chinese
	0883	CNOOC	Industries	Chinese
8	0939	China Construction Bank	Finance	Chinese
	3328	Bank of Communication	Finance	Chinese
9	0297	Sinofert Hold Ltd	Industries	Chinese
	0338	Shanghai Pechem	Industries	Chinese
	1033	Yizheng Chem	Industries	Chinese
	2883	China Oilfield	Industries	Chinese
10	0165	China Eb Ltd	Construction	Chinese
	1171	Yanzhou Coal Mining	Oil and Gas	Chinese
	2328	Picc P&C	Properties	Chinese
11	0005	Hsbc Holdings	Finance	Foreign
	0945	Manulife Financial Corp	Finance	Foreign
	2888	StanChart Plc	Finance	Foreign
12	0017	New World Dev	Construction	Hong Kong
	0135	Cnpc(Hong Kong)	Oil and Gas	Hong Kong
	0489	Dongfeng Group	Construction	Chinese
	0906	China Netcom	Construction	Chinese
	1088	China Shenhua	Oil and Gas	Chinese
	2319	Mengniu Dairy	Construction	Hong Kong
13	0144	China Mer Hold	Finance	Chinese
	0363	Shanghai Ind H	Industries	Chinese
	0882	Tianjin Dev	Construction	Chinese
	1199	Cosco Pacific	Shipping	Hong Kong
	1919	China Cosco	Shipping	Chinese
14	0218	Shenyin Wanguo	Finance	Hong Kong
	0227	First Shanghai	Finance	Hong Kong
	0902	Huaneng Power	Energy	Chinese
	0991	Datang Power	Energy	Chinese
15	0670	China East Air	Airline	Chinese
	0753	Air China	Airline	Chinese
	1055	China South Air	Airline	Chinese

Chapter 3.

Mesoscopic Dynamics of Financial Market

Highlights

- We obtained the *SoG forecasting model* to forecast market crashes, using a mean field formula derived from a model of fusions and fissions in the strategy space [Section 3.3.1].
- Based on 20 (out of the 30) component stocks for STI, we obtain crash predictions ranging from end October 2008 to mid-February 2009 (the Lehman Brothers bankruptcy: 15 Sep 2008, and the Asian Correction: 9 Mar 2009), with early warning four to six months prior to the crashes [Section 3.3.2].
- We find episodes of heightened market risks close to the STI low due to the Chinese Correction (27 Feb 2007), beginning of the Subprime Crisis (17 Aug 2007), and the Asian Correction (9 Mar 2009) [Section 3.3.2].

Significance statements

Some of these market crashes (like the Asian Correction) can be predicted, by assuming the market is made up of a large number of adaptive traders, who employing similar or competing trading strategies, which in turn form strategy clusters in the abstract strategy space. Whereas, the stock price movement is governed by the dynamics of the strategy clusters. When a strategy group grows into a giant component, trader actions become increasingly correlated and this is reflected in the stock price. The fragmentation of this giant component will lead to a market crash. As the number of adherents for some strategies grow, others decline in the constantly changing strategy space, this leave trace for us to do forecasting. We developed the *SoG forecasting* model using a mean field formula derived from a model of fusions and fissions in strategy space. By fitting the *continuous returns* of 20 stocks traded in Singapore Exchange to the *SoG forecasting* equation, we obtain crash predictions ranging from end October 2008 to mid-February 2009, that is close to the Lehman Brothers bankruptcy (15 Sep 2008) and the Asian Correction (9 Mar 2009). Added to this, there are early warnings four to six months prior to the crashes, which can act as a buffer period and get prepared for the crash. It is also important to note that, by mapping out the *market risk heat map*, and we observe episodes of heightened market risks that corresponded to the Chinese Correction (27 Feb 2007), beginning of the Subprime Crisis (17 Aug 2007), and the Asian Correction (9 Mar 2009).

3.1 Introduction

In Chapter 2, we have demonstrated that financial market self-organized into hierarchical structures. Particularly, the stocks from the same sectors grouped into clusters. These clusters further organized into two superclusters, dominated by local stocks and by Chinese stocks each. Finally, these superclusters interact with each other as a single market. This series of interactions at various levels is the evidence of hierarchical structure in the financial market, from market sector, to the country market, and to the global market. Moreover, we also demonstrated the dynamics of financial markets toward the market crashes, such as increasing of inter-cluster and intra-cluster correlations, formation of the giant cluster that persisted for months, and persistent decreasing of cluster mixing degree. In addition, we also discussed the aftermath dynamics of the financial cluster, for example the fragmentation of the giant cluster into small clusters associated with high cluster mixing degree.

This information can act as the precursor for market crashes, and it is important to be able forecast financial crisis. The Global Financial Crisis of 2007-2008 wiped out US\$ 34 trillions across global financial markets [133]. This is equivalent to 45% of 2016 world gross domestic product (GDP) of total US\$ 76 trillions [134] (East Asia & Pacific US\$ 22.5 trillions [135], United States US\$ 18.5 trillions [136], Euro Union US\$ 16.4 trillions [137]) according to the World Bank databank. Arguably, the defining moment of this crisis was the failure of Lehman Brothers in September 2008, which sent shock waves through financial markets around the world in the form of global market crashes namely the October 2008 crash and the Asian Correction (March 2009). Had the Federal Reserve seen these crashes coming, they might have worked harder to bail out Lehman Brothers, and prevented the crashes altogether.

Financial economists largely believe that financial crises are due to exogenous factors, and are hence not predictable. However, Sornette and others classified all US market crashes, and found that most do not have exogenous causes [138–142]. We feel that the limitations come from the models chosen to do forecasting. For instance, a statistically stationary model [143–145] cannot produce booms and busts, whereas time series model with stochastic jumps [146–148] produces unpredictable booms and busts. Beyond statistical modeling, Palmer *et al.* introduced in 1994 the Santa Fe Stock Market Model [149], and showed that the artificial market can crash in the absence of exogenous shocks. We expect precursory signatures to exist for the endogenous booms and busts in the Santa Fe Stock Market Model, but no one has looked into this systematically. Alternatively, we can also try model-free forecasting of extreme market events using only empirical data [94]. The most successful of these is the Log Periodic Power Law (LPPL) method [150, 151], which was originally developed by Sornette in 1996 [152] to forecast earthquakes. As many have likened financial crashes to earthquakes [153–155], Johansen and Sornette fitted the LPPL to stock market indices, and found that the real market crashes times are very close to the finite-time singularities of the fitted LPPL. More importantly, they concluded that 25 out of the 49 crashes identified exhibit LPPL signatures and thus be classified as endogenous crashes [156].

In this Chapter, we focus on the mesoscopic dynamics of financial market. This is because we believe that market movements are influenced by the dynamics of trading strategies. Market participants read and react to market signals, during normal period they have diverse opinions, but they behave coherently when there are large price swings. These findings demonstrate that a financial market naturally self-organizes into clusters of strongly-correlated stocks (and trading strategy with similar opinions) that undergo fusions and fissions. This fusion-fission process can be modeled by the *soup-of-groups* (SoG) model introduced by Johnson *et al.* which has been employed to describe many dynamical complex systems, such as social network, quantum many-body systems, superconducting Cooper pairs, insurgence of terrorists [48–52], etc. The works that particularly inspired us, is the application of SoG model to explain the underlying dynamics in human insurgency, which suggested that the escalation rate and timing of the fatal attacks from terrorists follow naturally from SoG dynamics and hence is predicable [51, 53]. We are further encouraged by success in forecasting a large Taiwanese earthquake timing, magnitude, and location, by using the SoG model to describe the dynamics of fault plane elements [54]. Because of this, we are optimistic about the application of the SoG model in financial market and want to exploit the SoG forecasting power to forecast financial crashes.

3.2 Data and Terminologies

3.2.1 Data

In this study, we focus on data from the Singapore Exchange (SGX) because we are familiar with stocks listed in the SGX, in contrast to model markets (e.g. the London Stock Exchange (LSE) and the New York Stock Exchange (NYSE)) that are often studied by econophysicists. Furthermore, as discussed earlier, the SGX is an emerging market, it is strongly coupled to financial markets around the world, and readily reflects the global market movement. For instance, the Straits Times Index (STI) is highly correlated with the Dow Jones Index ($\rho \approx 0.91$), and slid more than 45% between August 2008 and October 2008 as the result of the October 2008 crash (see Fig. 3.2.1). The STI is made up of the top 30 stocks in SGX, but these components do change from time to time. Thus we only consider 20 component stocks that remain highly traded across January 2006 to December 2011. See Table.3.1 for the full list of these 20 stocks.

Fig. 3.2.1. Straits Times Index (daily) from January 2006 to December 2011 downloaded from Yahoo, Singapore Finance (<https://sg.finance.yahoo.com/q/hp?s=%5ESTI>) and Dow Jones Industrial Average Index (daily) downloaded from the Federal Reserve Bank of ST. Louis (<https://research.stlouisfed.org/fred2/series/DJIA/downloadaddata>). The correlation between STI and DJI is high ($\rho \approx 0.91$) indicating that the Singapore market is strongly affected by market activities in the United States. The dash vertical lines represent the lowest of STI due to: (red) 27 Feb 2007 Chinese Correction, (green) begin of landslide of Subprime Crisis-17 Aug 2007, (blue) Lehman Brothers crisis-27 Oct 2008, (magenta) Asian Correction-09 Mar 2009, and the black solid vertical line shows the collapse of Lehman Brothers, 15 Sep 2008.

The tick-by-tick data was downloaded from the Thomson Reuters Tick History database (<http://thomsonreuters.com/tick-history>). On average the number of transactions per stock is on the order of 10^6 over 1506 trading days. Unlike LSE and NYSE, the SGX is relatively illiquid, as the median time interval between transactions ranges from 6 s for most liquid stock to 47 s for least liquid stock (see Table.3.1). Across the period studied, the STI experience two crashes and tumbled to its lowest at 1600.29 (27 Oct 2008, October 2008 crash), with a slight recovery after that and slid further to 1456.29 (9 Mar 2009, Asian Correction). The first crash was larger thus we select 27 Oct 2008 as

our reference date for the market crash.

Table 3.1. Stocks codes and transaction details for the 20 stocks studied in this paper. Notes that Q1, Q2, and Q3 represent 25%, 50%, and 75% percentile across all transactions and continuous returns intervals from January 2008 to December 2008.

No	Stock Code	Stock Name	Transactions				Continuous Returns			
			Number	Interval (s)			Number	Interval (s)		
				Q1	Q2	Q3		Q1	Q2	Q3
01	CATL	CapitaLand	1436783	2	7	24	14017	103	679	3756
02	CMDG	ComfortDELGRO	435726	3	21	88	6795	359	2895	9453
03	COSC	Cosco Corp.	1137049	2	8	29	7592	262	2016	7905
04	CTDM	City Development	669882	3	13	52	11115	187	1210	4948
05	DBSM	DBS Bank	1304215	2	7	25	16635	77	492	3393
06	FRNM	Fraser & Neave	529932	3	17	70	8021	338	2249	7237
07	GAGR	Golden Agri	1137261	2	9	29	7799	129	1485	7026
08	HKLD	HongKong Land	547430	2	13	67	11450	233	1289	4775
09	JARD	Jardine Matheson	232975	5	30	149	9249	179	1260	5777
10	JCYC	Jardine C&C	210416	8	47	185	9164	276	1659	5924
11	KPLM	Keppel Corp.	1256012	1	7	26	13097	112	776	3909
12	NOBG	Noble Group	1333328	1	6	23	9376	128	1183	5778
13	OCBC	OCBC Bank	1013364	3	11	37	11852	152	1053	4489
14	SCIL	Sembcorp Industry	719008	2	13	53	8752	286	1928	6604
15	SGXL	Singapore Exchange	1175935	2	9	29	11504	122	918	4637
16	SIAL	Singapore Airline	647313	3	15	58	11526	179	1112	4546
17	STAR	StarHub	367062	4	24	101	7196	417	2940	8735
18	STEL	Singtel	1293144	2	9	29	13287	88	722	3991
19	UOBH	UOB Bank	949453	2	9	36	14016	120	747	3803
20	WLIL	Wilmar	1129634	1	8	29	8996	214	1412	5497

3.2.2 Continuous Returns

In finance, one of the most widely used definition for quantifying financial returns is the simple return, $r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$ and log return, $Lr_t = \log(1 + r_t)$. Stylized facts have been discovered, for instance the distribution of the return (log return) tend to have heavy tail for short intervals (less than a day) but approach the normal distribution for long intervals [62]. The choice of the time interval is therefore ambiguous. Because of this ambiguity, we introduce the *continuous returns* defined as the continuous prices movement in the same direction. Mathematically the continuous return ($CR_{t(m)}$) is given by

$$CR_{t(m)} = \frac{|P_{t(m+n)} - P_{t(m)}|}{P_{t(m)}}, \quad (3.1)$$

satisfying simultaneously

$$\begin{aligned} (P_{t(m-1)} - P_{t(m)}) \times (P_{t(m+n)} - P_{t(m+n+1)}) &> 0; \\ (P_{t(m-1)} - P_{t(m)}) \times (P_{t(m)} - P_{t(m+k)}) &< 0; \\ (P_{t(m+n)} - P_{t(m+n+1)}) \times (P_{t(m+n-k)} - P_{t(m+n)}) &< 0, \end{aligned}$$

where $P_{t(i)}$ is the i^{th} transaction price at time $t(i)$, $k \in 1, 2, 3 \dots n$, and at time t_{m+n} the direction of the price movement changes.

Fig. 3.2.2. A segment of stock price movement that illustrates the definition of a *continuous return*. A continuous return is a series of price movements in the same direction. In this figure, each point represents a transaction price and the time interval is different as transactions can occur at any time. Each vertical line indicates the transaction time when the price movement changes direction. Thus, each vertical line is considered as a continuous return $CR_{t(m_i)}$ with magnitude defined as $CR_{t(m_i)} = |P_{t(m_{i+1})} - P_{t(m_i)}|/P_{t(m_i)}$. Refer to Equation (3.1) for the details.

This definition of an event is similar with micro-trend used by da Cruz and Lind [157], or the ‘run’ used by Hashimoto *et al.* [158] In financial markets, market participants are heterogeneous, so though they read the same market signals they may arrive at different trading strategies. However, if the market signal is strong, then the traders will arrive at the same trading strategies thereby reinforcing the market signals, leading to a formation of strategy blocks. When a strategy block is executed, market participants within the strategy block will react in the same way, hence generating a continuous return with a magnitude that depends on the size of the strategy block. Fig. 3.2.2 illustrates a segment of stock price time series, where continuous return is the fractional price change between two red vertical lines, which denote the start and end a microtrend [159]. Over the study period, Jan 2006 to Dec 2011 the number of continuous returns on average across stocks is on the order of 10^4 , and the median of microtrend for different stocks ranged from 10 mins to 50 mins (see Table.3.1).

3.2.3 Relation Between SoG Event and Continuous Return

The soup-of-groups (SoG) model, as the name suggests, consists of a “soup” (system) of “groups” (clusters) with various sizes. The essence of SoG dynamics is the competition of fusion-fission processes among clusters. For instance, a cluster of size s_i can merge with a cluster of size s_j to give a cluster with size $s_i + s_j$ at a rate of ν_+ , or a cluster of size s_k can fragment completely into s_k clusters with size 1 at a rate of ν_- (refer to Section 5.5 for graphical illustration and the SoG master equation). In an infinite element equilibrium

system, Neil Johnson et. al. discovered that the equilibrium distribution for the SoG model is an *Exponential Truncated Power Law* (ETPL) distribution [52] as

$$N_s = f(S) \sim s^{-\frac{5}{2}} \exp(-S_o s); \quad S_o = -\ln \left(\frac{4(\nu_p + \nu_b)\nu_p}{(2\nu_p + \nu_b)^2} \right).$$

Fig. 3.2.3. (a) A schematic showing how SoG fusion and fission processes gives rise to Continuous Return (*CR*) events. Here, the selected cluster is colored green and the resultant clusters after fusion or fission process are colored blue. Each *CR* is due to a cluster fragmentation with a magnitude proportional to the cluster size. In contrast, a fusion process generates no *CR* but makes the cluster bigger, and possibly into a giant cluster. In (b) we show the curve generated by Eq (3.3) that illustrates how the Integrated Continuous Return (*ICR*) change when a giant cluster is present and growing. Initially, in the absence of the giant cluster, the *ICR* grows linearly due to the fragmentation of clusters with an equilibrium distribution of sizes. However, when a giant cluster forms at the expense of other clusters, we start to get only fragmentations of smaller clusters and hence a slow down in *ICR* growth. Finally at t_c where the giant cluster has exhausted all the resources, and reached the maximum allowed size, it must fragment completely in the next time step.

The market is made up from a large number of adaptive traders, who can either employ similar or competing trading strategies. We assume that similar strategies will form a strategy cluster in the constantly changing strategy space. In addition, the stock price movement is governed by the dynamics of these strategy clusters. For instance, when a strategy cluster grows larger, trader actions become increasingly correlated and this is reflected as the stock price momentum (continuous return). Thus, under the SoG context, each continuous return is caused by a fragmentation of the strategy block, therefore a market crash is associated with the fragmentation of a giant cluster. Fig. 3.2.3(a) illustrates the relation between the continuous return and the SoG fusion and fission process. During the fusion process two clusters merge and form a larger cluster, but generate no continuous return. On the other hand, each fission process will generate a continuous return. In these descriptions, a market crash is due to the fragmentation of a giant cluster (giant strategy block) that rapidly drives the price downwards. Thus if we can track the size and growth of the giant cluster we should be able to deduce the time when the market will crash, that is by determining the time when the giant cluster

will reach its maximum size. Assuming finite resources and hence a finite number of elements, as the giant cluster grows bigger it exhausts the larger clusters around it. This manifests itself as a slowing down in the growth of the integrated continuous returns (accumulated events) as shown in Fig. 3.2.3(b) (see Section 3.3 for greater details).

3.2.3.1 Distribution of Continuous Returns.

The SoG dynamics is universal, so we expect to be able to do forecasting with it even when the stock price dynamics is only approximately SoG. To demonstrate this approximate SoG behavior we check the distribution of the continuous return against ETPL distribution. For comparison we also fit the empirical data with power law (PL) and asymptotic exponential (EXP) distributions. The fitting results indicate that the continuous returns for the 20 stocks are best fitted by the ETPL. In Fig. 3.2.4 we show (a) a typical, (b) good, and (c) poor set of fits to the ETPL, PL, and EXP. When the fit is poor none of the distributions fit the continuous returns well. Fortunately there are few poor fits. The full fitting results are shown in Appendix 3.A. Overall, the average α_{ETPL} and β_{ETPL} across all stocks is 1.82 ± 0.23 and 0.37 ± 0.08 respectively. The reported $\hat{\alpha}_{ETPL}$ is less than the SoG exponent of $\alpha = 2.5$. If we take this result at face value, we must accept that the continuous return is not proportional to the size of the cluster, as required by the SoG model. But, if we assume that a cluster of size s produces a continuous return of s^ζ , where $\zeta = 2.5/\alpha$, then the SGX will indeed be a SoG system. This is equivalent to saying that a group of size s produces an effect that is superlinear in s , an assumption that is reasonable in the stock market. We will use this modified scaling to formulate the forecasting equation (Eq.(3.3)) in Section 3.3.1.

Fig. 3.2.4. Fits of continuous return distribution (black) with the exponential truncated power law, *ETPL* (red) benchmarked against the power law, *PL* (green) and asymptotic exponential, *EXP* (blue) distributions. In (a) typical and (b) good fits, the ETPL fits continuous returns the best, while the best PL decays too slowly and the best EXP decays too quickly. However, in (c) poor fits none of the distributions fit the data well.

3.3 SoG Forecasting

3.3.1 SoG Forecasting Model

In order to build the forecasting model, the inputs are: the size of the giant cluster $s_G(t)$, the size distribution of the rest of clusters $f(s)$, as well as the relations between the magnitude of the continuous return and the cluster size $CR_t = CR_t(s_t)$. We make several assumptions regarding these inputs. First, we assume that the giant cluster grows linearly with time as $s_G(t) = \gamma t$. Second, instead of using the ETPL distribution, which is difficult to work with, we use the simplified PL distribution $f(s) = As^{-2.5}$. Lastly, we assume that the continuous return is given as $CR_t \propto s_t^\zeta$ in order for the dynamics of continuous return to be approximately SoG. With these assumptions, we can write down the expected integrated continuous return $\hat{ICR}(0, t)$ from $time = 0$ to $time = t$ as

$$\begin{aligned} \hat{ICR}(0, t) &= \int_0^t dt \int_{s_o}^{s_{max}-s_G(t)} ds As^\zeta f(s) = \int_0^t dt \int_{s_o}^{s_{max}-s_G(t)} ds As^{\zeta-2.5} \\ &= \frac{A}{\zeta-1.5} \left\{ \frac{1}{\gamma(0.5-\zeta)} [(s_{max}-\gamma t)^{\zeta-0.5} - s_{max}^{\zeta-0.5}] - s_o^{\zeta-1.5} t \right\}. \end{aligned} \quad (3.2)$$

Assuming the giant cluster started growing at t_o , Eq (3.2) becomes

$$\begin{aligned} \hat{ICR}(t_o, t) &= \\ &= \frac{A}{\zeta-1.5} \left\{ \frac{1}{\gamma(\zeta-0.5)} [(S_{max}-\gamma\{t-t_o\})^{\zeta-0.5} - S_{max}^{\zeta-0.5}] - S_o^{\zeta-1.5} \{t-t_o\} \right\} + ICR_o. \end{aligned} \quad (3.3)$$

Eq.(3.3) is the mean-field *SoG forecasting equation* that we show in Fig 3.2.3(b). From Fig 3.2.3(b), we observe that before the giant cluster started growing, $\hat{ICR}(t)$ grows linearly with time. As the giant cluster grows, the growth rate of $\hat{ICR}(t)$ decreases. More importantly, a singularity occurs when the giant cluster reaches the maximum size of $S_G(t_c) = S_{max} = \gamma(t_c - t_o)$. We call this time the *forecasted crash time* (t_c) which is

$$t_c = \frac{S_{max}}{\gamma} + t_o. \quad (3.4)$$

From the data, we define the empirical *integrated continuous returns* from t_{Start} to t_{End} as

$$ICR(t_{Start}, t_{End}) = \sum_{k=i; t(i) \geq t_{Start}}^{N; t(N) \leq t_{End}} CR_{t(k)}, \quad (3.5)$$

which is the cumulative sum of the continuous returns. We fit the empirical integrated continuous returns $ICR(t_{Start}, t_{End})$ to our forecasting model $\hat{ICR}(t_{Start}, t_{End})$ (Eq (3.3)) to obtain the forecasted crash time t_c . For the fitting, we use the non-linear least square

method in MATLAB to estimate the parameter set that minimizes the overall residuals. In practice, we must only use data available up to the present moment. Therefore to mimic a real-time forecasting using historical data, we work with dynamic fitting windows starting at t_{Start} and ending at t_{End} , where t_{End} must be before the market crash we want to “forecast”. To do this, we fix a t_{Start} then we start with t_{End} one month forward and increase t_{End} two weeks at a time. We then obtain t_c as a function of t_{End} for a given t_{Start} , $t_c = t_c(t_{End}|t_{Start})$.

Fig. 3.3.1. Empirical Integrated Continuous Return ICR fitted to the forecasting model $\hat{ICR}$ (Eq (3.3)). For illustration purposes, we use DBSM.SI and a t_{Start} of 15 Dec 2006. t_{End} is then slowly increased with an interval of two weeks. The upper half of each sub-figure shows the empirical ICR (blue) and the fitted $\hat{ICR}$ (red dashed-curve), and the rescaled stock price is shown as the background (black). In addition, the black vertical dashed line represents the t_{End} and red vertical line indicates the corresponding t_c . The lower half shows the plot of $t_c(t_{End})$ as a function of t_{End} , where the Lehman Brothers Bankruptcy (15 Sept 2008) is represented by a blue horizontal dashed line and 27 October 2008 is represented by a black horizontal dashed line.

3.3.2 SoG Forecasting Results

Our objective is to use the SoG model to detect early warnings of market crashes by forecasting the crash times of the STI component stocks. This is a different approach compared to traditional finance and econometric models, wherein sudden changes are assumed to be brought about by exogenous shocks. In the SoG model, a market crash is endogenous, driven by the giant cluster formation and its fragmentation. Within the period that we study, from Jan 2006 to Dec 2011, two global financial crashes left their marks on the SGX. As shown in Fig. 3.2.1, the October 2008 crash and the Asian Correction that caused the STI to tumble to its lowest at 1600.29 (27 Oct 2008, October 2008 crash), with a slight recovery after that and further sliding to 1456.29 (9 Mar 2009, Asian Correction). The first crash was larger, thus we select 27 Oct 2008 as our reference date for the market crash.

We fix the start t_{Start} of the fitting window as 15 Dec 2006, which is about two years before the crash. We then slowly expand the fitting window by increasing the t_{End} of the fitting window. Fig. 3.3.1(a) to 3.3.1(c) show a series of fits for t_{End} as 15 Aug 2007, 15 Mar 2008, and 15 Dec 2008. In these figures, the black vertical line represents t_{End} and the red vertical line represents the corresponding forecasted crash time t_c . In the lower half of each sub-figure, we plot the corresponding t_c against t_{End} up to t_{End} as 15 Aug 2007, 15 Mar 2008, and 15 Dec 2008 with 15 Dec 2006 as the t_{Start} . From the graph, we can see that when far before the October 2008 crash, t_c is increasing linearly with t_{End} . This seems to suggest that the SGX is always critical, and ready to crash. This agrees with the suggestions that stock markets are self-organized critical (SOC) systems [160,161], as well as findings from calibrating agent-based models in stock markets [162].

As we approach the October 2008 crash, we see t_c growing at a slower rate and finally stagnating around the actual crash time (t_{Act} = 27 Oct 2008). This stagnation occurs a few months before t_{End} reaches t_{Act} . This is an interesting early warning signature that we must look out for in a system that is already bordering on criticality. This result gives a positive outlook for applying the SoG model to forecast market crashes, but rigorous tests are needed to establish the forecasting power of the SoG model.

3.3.2.1 Sensitivity Analysis

We perform a sensitivity test on the forecasting results by comparing the forecasted crash time t_c with the October 2008 crash, actual crash time t_{Act} = 27 Oct 2008. Given a t_{Start} , t_c is a function of t_{End} as $t_c = t_c(t_{End}|t_{Start})$. We can test the sensitivity of $t_c(t_{End}|t_{Start})$ to t_{End} , by calculating the weighted mean of t_c at particular t_{End} using various t_{Start} 's. The weighted mean and weighted standard deviation of $t_c(t_{End})$ is then calculated for each respective t_{End} as

$$\begin{aligned}\bar{t}_c(t_{End}) &= \sum_{i=1}^N w(t_{Start}^i|t_{End}) t_c(t_{End}|t_{Start}^i); \\ \sigma^2(t_c(t_{End})) &= \frac{\sum_{i=1}^N w(t_{Start}^i|t_{End}) [t_c(t_{End}|t_{Start}^i) - \bar{t}_c(t_{End})]^2}{N};\end{aligned}\tag{3.6}$$

where the weight is defined as

$$w(t_{Start}^i|t_{End}) = \frac{NumDays(t_{Start}^i, t_{End})}{\sum_{i=1}^N NumData(t_{Start}^i, t_{End})}.$$

In Eq.(3.6) the $NumDays(t_1, t_2)$ is the number of trading days between t_1 and t_2 , and weight $w(t_{Start}^i|t_{End})$ is introduced to take into account the different amount of data being used to do the forecast.

Fig. 3.3.2. In this figure we show the sensitivity test for stock DBSM.SI, where the blue dashed line represents the date of Lehman Brothers Bankruptcy (15 Sept 2008), black dashed line represents the our benchmark date-27 Oct 2008 when the STI at its lowest due to October 2008 crash, and red dashed line marks the date 9 Mar 2009 that the STI at its lowest due to Asian Correction. We plot the weighted average forecast crash time t_c against t_{End} showing the weighted standard deviation as the spread. We also show the p -value for our null hypothesis $H_o : t_c(t_{End}) = 27 \text{ Oct } 2008$ cutoff at 95% confidence level (orange curve).

In this study, we test the sensitivity of $t_c(t_{End}|t_{Start})$ to t_{End} using 14 t_{Start} 's at the beginning and middle of the months from 1 Nov 2006 to 16 May 2007. We compare the weighted mean of forecasted crash time, $\bar{t}_c(t_{End})$ s with the actual crash time $t_{Act} = 27 \text{ Oct } 2008$ at 95% confidence level, using the null hypothesis $H_o : \bar{t}_c = t_{Act}$, where the t -value and p -value for n observations are calculated as

$$t_{value}(t_{End}) = \frac{|\bar{t}_c(t_{End}) - t_{Act}|}{\sigma(t_c(t_{End}))};$$

$$pvalue(t_{End}) = 1 - \int_{-\infty}^{t_{value}(t_{End})} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \frac{1}{\sqrt{n\pi}} \frac{1}{\left(1 + \frac{x^2}{n-1}\right)^{\frac{n}{2}}} dx, \quad (3.7)$$

where $\Gamma(x)$ is the gamma function. When $p < 0.05$, t_c is significantly different from t_{Act} at 95% confidence level, and we reject H_o . Otherwise, we fail to reject H_o when p_{value} is greater than 0.05. Robust signatures for the market crash forecast can be observed in Fig. 3.3.2: $\bar{t}_c$ increases linearly with t_{End} when we are far from the market crash, and thereafter stagnates around t_{Act} when t_{End} is closer to the October 2008 crash. For earlier t_{End} 's the predicted $\bar{t}_c$ is statistically different from t_{Act} , whereas for t_{End} close to the actual crash there is no longer statistical difference between $\bar{t}_c$ and t_{Act} .

3.3.2.2 Discussions

The t_{End} when t_c first becomes statistically indistinguishable from t_{Act} is the *warning time*, t_w which in turn acts as the early warning signal for market crashes. In addition, in order to determine the accuracy of the SoG forecasting result, we calculate the average of the weighted mean of forecasted crash time $\langle \bar{t}_c(t_{End}) \rangle$: average of $\bar{t}_c(t_{End})$ for the t_{End} 's between the warning time t_w and the actual crash time t_{Act} (27 Oct 2008). If the $\langle \bar{t}_c(t_{End}) \rangle$ calculated is close to t_{Act} , it means the forecasting result is accurate. The results for all 20 component stocks are listed in the Table.3.B.1 in Appendix 3.B, where we find a range of four to six months of early warning prior the actual crash. Apart from the component stock DBSM.SI with $\langle \bar{t}_c(t_{End}) \rangle$ falling on 29 Oct 2008, the others predicted crash dates are after t_{Act} , as late as 12 Feb 2009. This is to be expected, as the SoG forecasting model predicts the latest possible crash time. However, in reality it is possible for a giant cluster to disintegrate before it reaches its maximum size.

In addition to the latest possible crash date for individual stocks that we are tracking, we can also combine the forecasting results from all stocks into an index to estimate the market-level risk. For example, when the majority of the STI component stocks predict a crash simultaneously on a particular date, the market risk heightens on that particular date. In order to do this cross sectional study, we first count the number of STI component stocks forecasted to crash on a particular day, for various t_{Start} 's and t_{End} 's. Going back to Fig. 3.3.1, we observe that when we are far from the actual crash, the forecasted crash date increases linearly with t_{End} and thus we get a uniform background count. In contrast, when t_{End} is close to t_{Act} , t_c stagnates leading to a high concentration of forecasted crashes in a narrow range of dates and hence a large number of counts on those dates. We show these counts as a heat map to visualize the market risk (*market risk heat map*), where a date with large counts hence high market risk is shown as a red pixel. In contrast, a blue pixel means background level of counts with low market risk.

On the left of Fig. 3.3.3, we vary t_{End} over the period Jan 2007 to Jun 2009. For each stock, we accumulate the forecasted crash time t_c into a histogram and these histograms are then represented as a heat map. In principle if we are at time t , we will not know the forecasted crash time for $t_{End} > t$. Thus on the right of Fig. 3.3.3, we show the average counts for all 20 STI component stocks, but only vary t_{End} from Jan 2007 up to t . We see that there are four red bands for t_c peaking at 25 Jan 2007, 14 Aug 2007, 14 Oct 2007, and 5 Mar 2009. These episodes of heightened market risks are close to the lowest of STI due to the Chinese Correction (27 Feb 2007), begin of the landslide of the Subprime Crisis (17 Aug 2007), and the Asian Correction (9 Mar 2009). However, there is no

heightened market risk near the collapse of Lehman Brothers bankruptcy (15 Sep 2008). We believe this is due to the October 2008 crash and the Asian Correction appearing as a single extended market crash to the forecasting model. More interestingly, early warning signals did appear a few months before the October 2008 crash, thus providing us the opportunity to prepare for the financial market crashes.

Fig. 3.3.3. Market risk heat map of the SGX. When a majority of the stocks simultaneously forecast a crash on a particular date, the market risk is high on that date. In this figure, a red pixel indicates a high market risk while the background risk level (low stress) is shown as blue pixels. For both figures, the dash horizontal lines represent the lowest of STI due to: (red) 27 Feb 2007 Chinese Correction, (green) begin of landslide of Subprime Crisis-17 Aug 2007, (blue) Lehman Brothers crisis-9 Oct 2008, (magenta) Asian Correction-9 Mar 2009, and the black solid horizontal line shows the collapse of Lehman Brothers, 15 Sep 2008. [Left] Heat map for all 20 stock with t_{End} running from Jan 2007 to Jun 2009. The white curve represents the sum of risk level for all stocks, which peaks at 25 Jan 2007, 14 Aug 2007, 14 Oct 2007, and 5 Mar 2009. [Right] Sum of all 20 stocks risk level with t_{End} running from Jan 2007 to time t . The mean and standard deviation of daily price for all stocks is shown as the black curve with white band. The increase in risk level can be seen a few months prior to the actual crash time, and thus can act as an early warning for market crashes.

All in all, these results show that the SoG forecasting model provides early warning in the form of slowing down of the integrated continuous returns. Furthermore the SoG model is also able to indicate the tension level for potential financial crashes via the market risk heat map. Together with other standard critical-slowness indicators like increasing volatility, increasing autocorrelation, red shift in power spectrum, and

increasing correlation [163–165], the combination of indicators can act as financial crisis early warning alarm for authorities, such that measures can be taken to prevent or at least soften the impact of financial crashes. Still, there is room for improvement, for example the false negative error for the October 2008 crash compared to Asian Correction. We do not expect that the SoG model to be the best model for describing market fusion-fission dynamics, particularly the complete fragmentation required by the SoG model. We believe partial fragmentation of clusters will be more realistic, and we are working to the test forecasting method based on SoG-like models. In summary, the SoG model forecasting is promising enough to merit further exploration.

Appendix

3.A Continuous Return Distribution Fitting Results

Fig. 3.A.1. Fits of continuous return distribution (black) with the exponential truncated power law, *ETPL* (red) benchmarked against the power law, *PL* (green) and asymptotic exponential, *EXP* (blue) distributions.

Fig. 3.A.2. Fits of continuous return distribution (black) with the exponential truncated power law, ETPL (red) benchmarked against the power law, PL (green) and asymptotic exponential, EXP (blue) distributions.

3.B Fits of SoG Forecasting Results

In Section 3.3, we perform a sensitivity test of the forecasting results by comparing the forecasted crash time t_c with the actual October 2008 crash $t_{Act} = 27 \text{ Oct } 2008$. In order to do this, we fit the empirical integrated continuous return with the forecasting model using various t_{Start} 's to obtain a distribution of t_c for a particular t_{End} . The weighted average $\bar{t}_c(t_{End})$ and weighted standard deviation $\sigma^2(t_c(t_{End}))$ are then being calculated. We compare $\bar{t}_c(t_{End})$ with t_{Act} at 95% confidence level with the null hypothesis as $H_o : \bar{t}_c = t_{Act}$. When the $p < 0.05$, t_c is significantly different from t_{Act} at 95% confidence level, and we reject H_o . Otherwise, we fail to reject H_o when p -value is greater than 0.05.

Table 3.B.1. The warning time, t_w that act as the precursor for for market crash and average forecasted crash time $\langle \bar{t}_c \rangle$ for the 20 STI component stocks studied. Compared with the actual market crash, t_{Act} (27 Oct 2008), the early warning is ranged from four to six months prior the actual crash time, while the average forecasted crash time $\langle \bar{t}_c \rangle$ range from 29 Oct 2008 to 12 Feb 2009. In this Table TDs stand for trading days.

No.	Index	t_w	$\langle \bar{t}_c \rangle$	No.	Index	t_w	$\langle \bar{t}_c \rangle$
01	CATL	15/05/2008	07/01/2009±17 TDs	11	CMDG	02/06/2008	04/12/2008±16 TDs
02	COSC	21/04/2008	13/01/2009±21 TDs	12	CTDM	07/05/2008	08/01/2009±19 TDs
03	DBSM	03/06/2008	29/10/2008±09 TDs	13	FRNM	27/05/2008	15/12/2008±13 TDs
04	GAGR	15/04/2008	12/02/2009±25 TDs	14	HKLD	26/05/2008	05/12/2008±13 TDs
05	JARD	26/05/2008	06/01/2009±21 TDs	15	JCYC	02/05/2008	19/12/2008±14 TDs
06	KPLM	25/06/2008	10/12/2008±13 TDs	16	NOBG	30/04/2008	30/01/2009±22 TDs
07	OCBC	04/06/2008	28/11/2008±12 TDs	17	SCIL	12/05/2008	11/12/2008±16 TDs
08	SGXL	12/05/2008	02/01/2009±16 TDs	18	SIAL	20/05/2008	15/12/2008±14 TDs
09	STAR	20/05/2008	17/12/2008±14 TDs	19	STEL	14/05/2008	18/11/2008±08 TDs
10	UOBH	30/05/2008	24/11/2008±12 TDs	20	WLIL	12/05/2008	30/01/2009±28 TDs

The sensitivity test for all 20 component stocks are shown in Appendix 3.C. Robust signatures for the market crash forecast can be observed, as t_c increased linearly with t_{End} when we are far from the market crash, and thereafter stagnated around t_{Act} when t_{End} is closer to the October 2008 crash. For earlier t_{End} 's the predicted t_c is statistically different from t_{Act} , whereas for t_{End} close to the actual crash there is no longer statistical

difference between t_c and t_{Act} . We call t_w , the t_{End} when t_c first becomes statistically indistinguishable from t_{Act} the *warning time*. We also calculate the average forecasted crash time $\langle \bar{t}_c(t_{End}) \rangle$ using t_{End} 's between t_w and t_{Act} (27 Oct 2008). If the $\langle \bar{t}_c(t_{End}) \rangle$ calculated is close to t_{Act} , it means the forecasting result is accurate. The results for all 20 stocks are listed in S Table 3.B.1. There are four to six months of early warning prior the actual crash, and apart from DBSM.SI, whose $\langle \bar{t}_C(t_{End}) \rangle$ falls on 29 Oct 2008, the others predicted crash dates after t_{Act} , as late as 12 Feb 2009.

3.C Forecasting Results

Fig. 3.C.1. The blue dashed line represents the date of Lehman Brothers Bankruptcy (15 Sept 2008), black dashed line represents the our benchmark date of 27 Oct 2008 when the STI at its lowest due to October 2008 crash, and red dashed line marks the date 9 Mar 2009 that the STI at its lowest due to Asian Correction. We plot the weighted average forecast crash time t_c against t_{End} showing the weighted standard deviation as the spread. We also show the p -value for our null hypothesis $H_0 : t_c(t_{End}) = 27 \text{ Oct } 2008$ cutoff at 95% confidence level (orange curve).

Fig. 3.C.2. The blue dashed line represents the date of Lehman Brothers Bankruptcy (15 Sept 2008), black dashed line represents the our benchmark date of 27 Oct 2008 when the STI at its lowest due to October 2008 crash, and red dashed line marks the date 9 Mar 2009 that the STI at its lowest due to Asian Correction. We plot the weighted average forecast crash time t_c against t_{End} showing the weighted standard deviation as the spread. We also show the p -value for our null hypothesis $H_0 : t_c(t_{End}) = 27 \text{ Oct } 2008$ cutoff at 95% confidence level (orange curve).

Fig. 3.C.3. The blue dashed line represents the date of Lehman Brothers Bankruptcy (15 Sept 2008), black dashed line represents the our benchmark date of 27 Oct 2008 when the STI at its lowest due to October 2008 crash, and red dashed line marks the date 9 Mar 2009 that the STI at its lowest due to Asian Correction. We plot the weighted average forecast crash time t_c against t_{End} showing the weighted standard deviation as the spread. We also show the p -value for our null hypothesis $H_o : t_c(t_{End}) = 27 \text{ Oct } 2008$ cutoff at 95% confidence level (orange curve).

Chapter 4.

Conclusion and Future Works

To summarize, in this thesis we first studied the macroscopic dynamics and then the mesoscopic dynamics of financial markets. In particular, we demonstrated the existence of hierarchical structures in the financial markets, from market sectors, to the country markets, and to the global markets. Furthermore, we also observed structural changes prior and post market corrections and crashes. These are persistently high synchronization in the market, persistently decreasing of cluster mixing degree, and the emergence of a giant financial cluster closer to the market crisis. Right after the crisis, the giant cluster fragmented into small clusters associated with strong clusters mixing degree. These observations prompted us to use the soup-of-groups (SoG) model to describe the financial market as a system with an enormous number of adaptive traders, who employ similar or competing trading strategies, forming strategy clusters in the strategy space. Stock price movements are thus governed by the dynamics of these strategy clusters. Based on this postulate, we developed the SoG forecasting model using a mean-field formula derived from a model of fusions and fissions in strategy space, and showed that some of the market crashes can be predicted, with warning signals ranges from four to six months prior to the market crashes.

In Chapter 2, we examine the macroscopic dynamics of financial markets toward market crashes, focusing on two case studies: (I) the Chinese Correction of February 2007 and (II) the Lehman Brothers Crisis in September 2008. To examine how the collapse of Chinese stock market was able to affect global markets, as most of them do not have exposure to Chinese stocks. To answer this, we studied two markets, the Singapore Exchange (SGX) and the Hong Kong Stock Exchange (HKSE) where Chinese stocks are listed. By applying the method of *partial hierarchical clustering*, a supervised clustering algorithm that can take multiple threshold to identify clusters, we segregated stocks into clusters based on their market sectors as well as their origin (local stock or Chinese stock). Using the *minimum spanning tree* (MST) and *minimum spanning forest* (MSF) algorithm, we show these stock clusters further organized into two superclusters. Interestingly, in the SGX, it is very clear that one supercluster is dominated by Chinese stocks, while another supercluster is dominated by local stocks. Furthermore, there are strong interacting (positive connection) between these superclusters, as the result of market activities. This showed that the market is self-organized and forms hierarchical structures, which allow the propagation of information, capital, risk, collapse, etc.

We went one step further, to study how these clusters evolve over shorter time windows from 2006 to 2007, and learned that the number of clusters decreases and thereafter increases, while stronger correlations within clusters weaken and weaker correlations between clusters strengthen before the Chinese Correction. To further support these findings, in Section 2.3, we studied how the hierarchy changed for 561 stocks in the SGX from 2008 to 2009, based on *robust clusters* identified by the method of *interaction hi-*

erarchical clustering. This period was chosen because within these two years, the SGX experienced the effects of the Lehman Brothers Crisis (Oct 2008) and the Asian Correction (Mar 2009). In the clusters *fusion-fission diagram*, we observed the emergence of a giant cluster, which remained persistently large until the Lehman Brothers Crisis. Then, right after the Lehman Brothers Crisis the giant cluster fragmented into small clusters. In the aftermath of crisis, we also find high mixing degree between a large number of small to mid-sized clusters, measured by the *adjusted Rand index* (a global measure) and the *degree of mixing*, *DoM* (focusing only on the largest clusters). We believe this high cluster mixing degree is due to the wide spectrum of expectations and responses from the market participants.

In Chapter 3, we turn our attention to the mesoscopic dynamics of the financial market, focusing on 20 component stocks of Straits Time Index (STI) traded from 2006 to 2010 in SGX. We assume the market is made up of a large number of adaptive traders, self-organizing into different strategy clusters in an ever changing abstract strategy space. When a strategy cluster grows larger, trader actions become increasingly correlated and this is reflected as the stock price momentum. Thus, the fragmentation of the giant strategy cluster will lead to a market crash. In other words, the stock price movements are governed by the dynamics of these strategy clusters, and we use the *soup-of-groups* (SoG) model to describe the dynamics of strategy clusters. In the study, to avoid the ambiguity of using different time intervals which will lead to different return distributions, we introduced the *continuous return*, which is defined as a continuous price movement in the same direction. Furthermore, we demonstrated that the continuous return fits well to the exponentially-truncated power-law distribution, which is the equilibrium distribution of SoG cluster size, hinting a relation between continuous return and SoG clusters.

Using a mean-field treatment of the SoG model, we derived the *SoG forecasting equation*, with the goal of forecasting global market crashes. By fitting the continuous return time series of the 20 component stocks to the SoG forecasting equation, we showed that some market crashes can be predicted. As we obtain crash predictions ranging from end October 2008 to mid-February 2009, which are close to the Lehman Brothers bankruptcy (15 Sep 2008) and the Asian Correction (9 Mar 2009), we realized that the SoG model is indeed able to forecast financial crises. Additionally, there are early warnings four to six months prior to the crashes, which can act as a buffer period to get ready for the crash. It is important to note that, we also find episodes of heightened market risks close to the STI low due to the Chinese Correction (27 Feb 2007), beginning of the Subprime Crisis (17 Aug 2007), and the Asian Correction (9 Mar 2009) by mapping out the *market risk heat map*. However, there is a false negative for the Lehman Brothers collapses, which we believe is due to signal convolution between the Lehman Brothers Crisis and the Asian Correction.

In spite of its success, there are signatures observed from the empirical data that cannot be explained by the SoG forecasting model. For example, large events seemingly not predicted by our SoG forecasting model can be found empirically prior to the giant event in the financial market. Therefore, as future work to understand the limitations of our SoG forecasting model, we will in Section 5.5 reexamine our SoG forecasting model, particularly the evolution of the event and the growth of giant cluster. We learned that

it is indeed possible for a large event to occur prior to the giant event, because of the existence of multiple giant clusters that we have not taken into consideration. Furthermore, in such a SoG system the giant clusters mostly experience accelerated growth and not linear growth as assumed in the SoG forecasting model. These observations improve our understanding about the SoG model, and indicates the needs for modifying the SoG forecasting model accordingly for higher predictive power. Besides this, we do not particularly believe that the SoG model is the best description for financial markets. The main feature of the SoG model that is probably not true in real world complex system is the complete fragmentation assumption. Perhaps a *SoG-like model* with partial fragmentation might be more realistic and better represents real-life trader dynamics. One can perform detailed analysis to gain insights of such SoG-like models, especially its dynamics, and thereafter modify the SoG forecasting model for more accurate forecasting of market crashes. Another dimension that can be explored is based on network theory for node-node aggregations and interactions, the Smoluchowski equation [166] which cover wider solutions than the SoG model. Besides that, it is also worth investigating alternative correlation measures between financial time series especially when their distributions are different [167]. Lastly, look at broader picture of financial markets by considering market-market interactions due to the high connectivity and integration among financial markets [168].

Chapter 5.

Methodologies, Algorithms, and Models

Highlights

- When the global distribution is known, the *adjusted cosine* similarity measure between two observations accounts for the similarity due to the global distribution [Section 5.1].
- We introduce the *Interaction-hierarchical clustering* as an unsupervised clustering algorithm, which involves two stages of clustering to account for interactions between clusters to yield more robust clustering result [Section 5.2].
- *Partial hierarchical clustering* in a supervised clustering algorithm, that provides multiple thresholds for identifying clusters [Section 5.3].
- The *Degree of Mixing, DoM* is useful in determining mixing degree with a single cluster (local mixing degree), unlike Adjusted Rand Index that only measure global mixing degree [Section 5.4].
- The exponential truncated power law distribution indeed is stable equilibrium distribution for the soup-of-groups model, but becomes marginally stable in the limit of large system [Section 5.5].
- The integrated cumulative event can undergo acceleration, linear, and deceleration growth depending on the fusion and fission rates. Therefore, the giant cluster can also follow a similar growth law but is more likely to experience acceleration growth [Section 5.5].
- We outline a procedure for statistical significance testing that accounts for finite sample effects, in order to reduce the risk of mistakenly rejecting plausible fits [Section 5.6].
- The Kolmogorov-Smirnov distance, d_{KS} measures statistical deviation at the cumulative density level, whereas the *distribution noise*, d_{DN} measure it at the probability density level [Section 5.7].

5.1 Similarity Measures

Often, we compute the similarity (correlations) between two observations, as an indicator of interaction strength. In this thesis, we discuss both the time series similarity (time sequence of the observations is important) and spectral similarity (focus on the distributions of observations). There are many existing correlation measures, but we only focus on the common similarity measures: (I) Pearson, Spearman, and Kendall tau for time series similarity in Section 5.1.1; (II) cosine similarity for spectral similarity in Section 5.1.2. In addition, we also introduced *Digital Correlation* (time series similarity) and *Adjusted cosine similarity* (spectral similarity).

5.1.1 Time Series Similarity:

Pearson, Spearman, Kendall tau, and *Digital Correlation*

Given two time series $\mathbf{O}_1 = \{o_1(t_1), o_1(t_2), \dots, o_1(t_T)\}$ and $\mathbf{O}_2 = \{o_2(t_1), o_2(t_2), \dots, o_2(t_T)\}$, one the most common measure of the similarity between their fluctuations is the *Pearson correlation* $C_{Pearson}$ [100, 101], defined as

$$C_{Pearson}(\mathbf{O}_1, \mathbf{O}_2) = \frac{\frac{1}{T} \sum_{i=1}^T (o_1(t_i) - \bar{\mathbf{O}}_1) (o_2(t_i) - \bar{\mathbf{O}}_2)}{\sqrt{\frac{1}{T} \sum_{i=1}^T (o_1(t_i) - \bar{\mathbf{O}}_1)^2} \sqrt{\frac{1}{T} \sum_{i=1}^T (o_2(t_i) - \bar{\mathbf{O}}_2)^2}}, \quad (5.1)$$

where $\bar{\mathbf{O}} = \frac{1}{T} \sum_{i=1}^T o(t_i)$.

One of the shortcomings of the Pearson correlation is that it is very sensitive to extreme events. As shown in Fig. 5.1.1, we can end up with the situation where two time series that are correlated nearly all the time give a negative Pearson correlation because their large responses to an exogenous event are in opposite directions, or where two time series that are anti-correlated nearly all the time give a positive Pearson correlation because their large responses to the exogenous event are in the same direction. This sensitivity is further confounded by the non-stationary time series, where a single standard deviation is not a good measure of fluctuations. To reduce the sensitivity to extreme events, we can use the *Spearman rank* correlation [102]

$$C_{Spearman}(\mathbf{O}_1, \mathbf{O}_2) = 1 - \frac{6 \sum_{i=1}^T \left(o_1^{(Ranked)}(t_i) - o_2^{(Ranked)}(t_i) \right)^2}{T(T^2 - 1)}, \quad (5.2)$$

which is the Pearson correlation between the ranked variables $\mathbf{O}_1^{(Ranked)}$ and $\mathbf{O}_2^{(Ranked)}$ respectively. The Spearman correlation is less sensitive to extreme events because it works with the rank statistic instead of the linear statistic. However, it suffers from potential artifacts to non-stationary time series because of all elements from the time series are given a global ranking.

Like the Spearman correlation, the *Kendall tau* correlation [103] also work with rank statistics and is hence less sensitive to extreme events. Unlike the Spearman correlation,

the Kendall tau correlation

$$C_{Kendall}(\mathbf{O}_1, \mathbf{O}_2) = \frac{N_{concordant} - N_{Disconcordant}}{\frac{1}{2}T(T-1)} \quad (5.3)$$

works with relative rankings. Here $N_{concordant}$ is number of concordant pairs and $N_{Disconcordant}$ is disconcordant pairs over all possible t and t' , where $t, t' = 1, \dots, T$ and $t \neq t'$. A concordant pair is when $o_1(t) > o_1(t')$ and $o_2(t) > o_2(t')$ or $o_1(t) < o_1(t')$ and $o_2(t) < o_2(t')$. On the other hand, if $o_1(t) > o_1(t')$ but $o_2(t) < o_2(t')$ or $o_1(t) < o_1(t')$ but $o_2(t) > o_2(t')$, the pair is considered disconcordant. Because ranking is performed for all possible pairs, the Kendall tau correlation is less sensitive to non-stationarities in the time series.

Fig. 5.1.1. Three time series simultaneously affected by an extreme event. The normal fluctuations of time series $\mathbf{O}_1$ and $\mathbf{O}_2$ are anti-correlated, but responses to the extreme event in the same direction. While the normal fluctuations of time series $\mathbf{O}_1$ and $\mathbf{O}_3$ are correlated, but responses to the extreme event in the opposite direction. The Pearson correlations reflect the direction of the extreme events (positive correlation between $\mathbf{O}_1$ and $\mathbf{O}_2$ and negative correlation between $\mathbf{O}_1$ and $\mathbf{O}_3$). In contrast, Spearman and Kendall tau correlations both provide correlations reflecting the normal fluctuations.

In this study, we use an alternative measure of correlations between two time series, which we call the *digital correlation* ($C_{Digital}$). The digital correlation measures how frequently $\mathbf{O}_1$ and $\mathbf{O}_2$ move together, but ignoring the magnitude of the movements. We define it as

$$C_{Digital}(\mathbf{O}_1, \mathbf{O}_2) = \frac{\sum_{i=1}^{T-1} \mathbb{I}_{\{\Delta o_1(t_i) \Delta o_2(t_i)\}} - \mathbb{I}_{\{-\Delta o_1(t_i) \Delta o_2(t_i)\}}}{T-1}, \quad \mathbb{I}_{\{x\}} \begin{cases} = 0 & \text{if } x \leq 0; \\ = 1 & \text{if } x > 0, \end{cases} \quad (5.4)$$

where $\Delta o(t_i) = o(t_{i+1}) - o(t_i)$. This definition of correlation has been quite popular from the early stages of econophysics, focusing only on the signs of price movement [169, 170]. We can check for the example in Fig. 5.1.1 that the digital correlation between time series $\mathbf{O}_1$ and $\mathbf{O}_2$ is negative (-0.81), whereas that between time series $\mathbf{O}_1$ and $\mathbf{O}_3$ is

positive (0.81), i.e. the digital correlation, by construction, is much less sensitive to isolated extreme events.

Fig. 5.1.2. The Pearson, Spearman, Kendall tau, and Digital correlations between two correlated Gaussian random variables $y_1 = x_1 \cos \theta + x_2 \sin \theta$ and $y_2 = x_1 \sin \theta + x_2 \cos \theta$ constructed out of two uncorrelated standard Gaussian random variables x_1 and x_2 . The theoretical linear correlation between y_1 and y_2 is $\rho = \langle y_1 y_2 \rangle = \sin 2\theta$, and θ can be tuned from $\theta = -\pi/4$ (perfect anti-correlation) to $\theta = +\pi/4$ (perfect correlation). For each value of linear correlation between y_1 and y_2 , we then computed the average nonlinear correlations (solid curves) over 10,000 samples of 100 data points for x_1 and x_2 . We also show the standard deviations of the nonlinear correlations (shaded areas) over the 10,000 samples.

A more systematic comparison between the three nonlinear correlations with the linear correlation is shown in Fig. 5.1.2 for a pair of time series sampled from a correlated two-dimensional Gaussian distribution. We see for this artificial data that the Spearman correlation is nearly identical to the Pearson correlation, whereas the digital correlation and the Kendall tau correlation are very nearly the same nonlinear function. Both are linear for weak correlations and nonlinear for strong correlations and anti-correlations, thus offer greater sensitivity to such correlations compared to the Pearson and the Spearman correlations. From the definitions of these correlations, we see that the Kendall tau correlation incorporates comparisons of all concordant and discordant pairs for all time scales, but the digital correlation only performs local comparisons. As result, digital correlation calculation is much lesser computationally expensive than computing the Kendall tau correlation.

GitHub: Similarity Measures

The codes for calculating correlation between variables are available at <https://github.com/BoonKinTeh/SimilarityMeasures>, and please refer to Teh *et al.* “The Chinese Correction of February 2007: How financial hierarchies change in a market crash” for more details [82], and cite the paper if you are using the codes.

5.1.2 Spectral Similarity:

Cosine Similarity and *Adjusted Cosine Similarity*

Another alternative of similarity measurement is the cosine similarity, commonly being used for spectral analysis (applicable for time series similarity). For example, given a set of possible outcomes $\mathbf{E}$, and two observations that measures the occurrence frequencies of $\mathbf{E}$, $\mathbf{f}_{O_i}$ and $\mathbf{f}_{O_j}$. The cosine similarity between the observations is defined as the inner product of their frequencies:

$$C_{Cosine}(O_i, O_j) = \frac{\sum_{e=1}^N f_{O_i}(E_e) f_{O_j}(E_e)}{\sqrt{\sum_{e=1}^N [f_{O_i}(E_e)]^2} \sqrt{\sum_{e=1}^N [f_{O_j}(E_e)]^2}}. \quad (5.5)$$

Here, $E_e \in \mathbf{E}$ is the e^{th} outcome, and N is total number of possible outcome. The $C(O_i, O_j)$ basically represent the angle between $\mathbf{f}_{O_i}$ and $\mathbf{f}_{O_j}$ in N -dimensional space.

Fig. 5.1.3. An illustration of cosine similarity and adjusted cosine similarity using a giant container (global distribution) with 50% – 50% chances to obtain black balls and red balls. The observations O_i and O_j are represented by small blue and red container, and from left to right indicate increasing similarity. (a) The cosine similarity Eq.(5.5) gives the value 0.0 for balls obtained that are completely different, and 1.0 for balls obtained that are completely the same. This result is reasonable, but it fails to account for the similarity caused by the high chances of withdrawing the same sample distribution as the global distribution. Also, it fails to distinguish the fact that it is harder to obtain sample distributions that are very different from the global distribution. (b) The adjusted cosine similarity, Eq.(5.6) discount the chances of obtaining the same distribution as the global distribution, and also increases the similarity magnitude if the observations deviated more from the global distribution.

In Fig. 5.1.3(a), we illustrate the cosine similarity between observations O_i and O_j drawn from a giant container that contains infinitely many red and black balls, where the probabilities to draw a red ball and a black ball are 0.5 and 0.5 respectively. Notice that the cosine similarity fails to incorporate (I) the high chances of obtaining observations (distributions) that are the same as the global distribution, and (II) to distinguish the fact that it is harder to obtain observations (distributions) that deviates more from the global distribution when the number of balls is increased, hence obtaining two such identical

observations should have a higher similarity measure. Because of this, we introduce the *adjusted cosine similarity* measure defined as:

$$C_{AdjCosine}(O_i, O_j) = \sum_{e=1}^N (f_{O_i}(E_e)f_{O_j}(E_e) - L_{O_i}L_{O_j} [P_G(E_e)]^2), \quad (5.6)$$

$$L_{O_i} = \sum_{e=1}^N f_{O_i}(E_e), \quad L_{O_j} = \sum_{e=1}^N f_{O_j}(E_e),$$

where $P_G(E_e)$ is the probability to obtain E_e from the global distribution. Using Eq.(5.6), we discount the similarity for two observations due to the expected drawn from the global distribution, and increase the magnitude of similarity when the observations deviate more from the global distribution.

GitHub: Similarity Measures

The code to calculate adjusted cosine similarity is available at <https://github.com/BoonKinTeh/SimilarityMeasures>. This code is intended to be published together with Teh *et al.* “Individual career progression can be more strategically planned”, please cite the paper (if published) if you are using the code.

5.2 Clustering Analysis I (unsupervised)

When dealing with situations with vast dimension of variables, we run a high risk of encountering problems such as multicollinearity and over-fitting. In order to reduce the dimensionality, clustering analysis become handy by grouping a set of elements into clusters, such that the similarity between elements within same cluster (group) are higher than the similarity between elements in different clusters (groups). There are different clustering algorithms, the first two most common classes are (I) hierarchical clustering (connectivity-based) [119,120,122] (II) partitional clustering (centroid-based) [118], follow by distribution-based, model-based clustering, density-based clustering, neural network classification [40–42, 123, 171–173], etc.

Algorithmically, for *partitional clustering*, the number of clusters k is predefined to give rise to k centrals, and the elements are clustered into different clusters by minimizing the utility function $U = \sum_{\mathbf{cl}_1}^{\mathbf{cl}_k} \sum_{j=1}^{N(\mathbf{cl}_i)} (P(E_j) - C(\mathbf{cl}_i))^2$. Here, the $\mathbf{cl}_i$ represent i^{th} cluster, E_j is the element belongs to $\mathbf{cl}_i$, $N(\mathbf{cl}_i)$ is the number of element in $\mathbf{cl}_i$, $P(*)$ is the ‘position’ of $*$, and $C(\mathbf{cl}_i)$ is the centroid (usually the average position) of all elements within $\mathbf{cl}_i$. On the other hand, in *distribution-based* or *model-based clustering*, elements are clustered together if they are most likely follow some distributions or models. Because of this, we have to choose appropriate distributions and models or we will run into the problem of overfitting. Lastly, *hierarchical clustering* gives an extensive hierarchy of clusters that formed at different distances and representing it as a dendrogram. This is the method we focus on in the thesis (see Section 5.2.1).

5.2.1 Hierarchical Clustering

The distance matrix, $\mathbf{D}$ is needed to perform hierarchical clustering, and we defined the distance between each pair of elements as

$$D(E_i, E_j) = 1 - C(E_i, E_j), \quad (5.7)$$

where $C(E_i, E_j)$ is the similarity between the elements (refer to Section 5.1 for details). Some authors prefer to define the distance as $D(E_i, E_j) = \sqrt{1 - C(E_i, E_j)}$ instead of Eq.(5.7) to satisfy the ultra-metric triangular inequality. Fortunately, the choice of definition of distance based on similarity does not affect the hierarchical clustering result, as long as it is one-to-one mapping and distances remain finite.

In agglomerative hierarchical clustering [29, 124] of N elements, we start with N clusters containing one element each. We then merge the two clusters with shortest distance to form a larger cluster. This process is repeated until we end up with a single cluster of N elements. There are three common measures of distances between clusters, called linkages that are defined based on the distances between elements. The *single linkage* or *minimum linkage* (SLHC), is defined as the minimum distance between an element in the first cluster and an element in the second cluster, is most commonly used for phylogenetic analysis, where we suspect that all elements are descended from the same ancestral element through series of small changes. Also commonly used for grouping variables in-

fluenced by common factors is the *complete linkage* or *maximum linkage* (CLHC), defined as the maximum distance between an element in the first cluster and an element in the second cluster. Finally, there is the *average linkage* (ALHC), defined as the average between distances between elements in the first cluster and elements in the second cluster. As its name suggests, the average linkage produces clusters that are intermediate between those produced by single and complete linkage. We illustrate in Fig. 5.2.1 hierarchical clustering using CLHC, SLHC, and ALHC.

Fig. 5.2.1. Five elements at positions 1, 3, 6, 10, 14.5 along a one-dimensional chain, and how they are hierarchically clustered using different algorithms: (a)-(d) complete-linkage (CLHC), (e)-(h) single-linkage (SLHC), and (i)-(l) average linkage (ALHC). The Euclidean distance between elements 1 and 2 is 2, which is the shortest amongst all pairs of elements. Thus in all three algorithms, they are merged to form the first cluster in (a), (e), and (i). In the CLHC, the linkage between element 3 and the cluster $\{1, 2\}$ is $\max\{6 - 1 = 5, 6 - 3 = 3\} = 5$ whereas the linkage between element 3 and element 4 is $10 - 6 = 4$. We therefore merge elements 3 and 4 to (b) form the next cluster. In contrast, for the SLHC, the linkage between element 3 and the cluster $\{1, 2\}$ is $\min\{6 - 1 = 5, 6 - 3 = 3\} = 3$, which is smaller than the distance of 4 between elements 3 and 4. Hence we merge element 3 with cluster $\{1, 2\}$ to (f) form the larger cluster $\{1, 2, 3\}$. For the ALCH the linkage between cluster $\{1, 2\}$ and element 3 is the average distance between element 3 with elements 1 and 2, hence 4 units, and in this case the element can either merge with cluster $\{1, 2\}$ or with element 4. In (j), we show the ALCH merging cluster $\{1, 2\}$ with element 3. For all three algorithms, the agglomeration process is iterated until we end up with a single cluster of all five elements.

In Fig. 5.2.2(a) we show an artificial correlation matrix that seem random but actually contains much more information that can be uncovered by hierarchical clustering. In order

to discover these hidden structures, we perform hierarchical clustering (single-linkage and complete-linkage) and reorder the correlation matrix using the ordering suggested by the dendrogram. The reordering process is called *seriation* [125–127], and we see that the seriated correlation matrix in Fig. 5.2.2(c)~(d) is more structured and we can identify a few clusters just through visual inspection.

Fig. 5.2.2. An artificial correlation matrix with elements drawn from three different Gaussian distributions. The objects the matrix represents are organized into four distinct clusters (red diagonal blocks). In addition, these clusters are also ‘interacting’ with one another (orange off-diagonal blocks). Our goal is to reorder the objects so that the stronger correlations are as close to the diagonal as possible, and the weaker correlations are as far away from the diagonal as possible. In this Figure, (a) shows the unordered correlation matrix, while the ordered correlation matrix are presented by: (b) using interaction hierarchical clustering, (c) using single-linkage hierarchical clustering, and (d) using complete-linkage hierarchical clustering.

Ideally we should have strongly-correlated (red) diagonal blocks and weakly-correlated (blue) off-diagonal blocks. However, although SLHC and CLHC do separate elements into groups based on their correlations, the organization is not very clean (see Fig. 5.2.2 (c)~(d)). There remain strong correlations in the off diagonal blocks of the seriated matrix. These are strong correlations overlooked by the CLHC and SLHC algorithms which considers a single distance between clusters to decide whether they are to be merged. These strong off-diagonal correlations contain information on how clusters interact with

each other, and are thus important. In order to retain such information, we perform clustering on the CLHC clusters, focusing on the off-diagonal correlations, and call this higher-level clustering procedure *interaction hierarchical clustering*.

5.2.2 *Interaction Hierarchical Clustering*

As we illustrate in Fig. 5.2.2, the objective of interaction hierarchical clustering (IHC) is to reorder the sequence of clusters, to obtain stronger diagonal blocks and weaker off-diagonal blocks. In order to do this, we first need to find the robust clusters from the seriated correlation matrix and then regroup those robust clusters according to their inter-cluster correlations.

The first step in IHC is to obtain the preliminary structure and also extract the robust clusters. A cluster is considered robust if the components in the cluster remain unchanged over a wide range of depth in the dendrogram. We quantify this notion of robustness length $l_{Robustness}(\mathbf{cl}_{i \cup j})$ of a cluster, $\mathbf{cl}_{i \cup j}$ as the linkage difference between the hierarchy level where it is first formed and the hierarchy level where it is merged with another cluster. Mathematically

$$l_{Robustness}(\mathbf{cl}_{i \cup j}) = d(\mathbf{cl}_{i \cup j}, \mathbf{cl}_k) - d(\mathbf{cl}_i, \mathbf{cl}_j), \quad (5.8)$$

where $d(\mathbf{cl}_i, \mathbf{cl}_j)$ is the linkage distance $\mathbf{cl}_i$ merged with $\mathbf{cl}_j$ to form $\mathbf{cl}_{i \cup j}$, and $d(\mathbf{cl}_{i \cup j}, \mathbf{cl}_k)$ is the linkage distance $\mathbf{cl}_{i \cup j}$ merged with other cluster.

Furthermore, we only consider $\mathbf{cl}_{i \cup j}$ as a robust cluster if the following criteria are met.

$$\begin{aligned} \text{I} \quad & l_{Robustness}(\mathbf{cl}_{i \cup j}) > \delta_{Robustness} \\ \text{II} \quad & d(\mathbf{cl}_i, \mathbf{cl}_j) \leq d_{MustBreak} \\ \text{III} \quad & d(\mathbf{cl}_i, \mathbf{cl}_j) \geq d_{MustNotBreak} \end{aligned} \quad (5.9)$$

Criterion I implies the robust cluster has to be robust in term of the robustness length, while criteria II and III are introduced as we only want cluster that formed between $d_{MustNotBreak}$ and $d_{MustBreak}$. This is because, clusters formed at distance above $d_{MustBreak}$ are weakly-correlated or anti-correlated, thus they should not grouped into one cluster. Meanwhile, clusters with distance below $d_{MustNotBreak}$ are strongly correlated and should not be broken up into more smaller clusters. In Fig. 5.2.3(a) we illustrate the *robustness length* ($l_{Robustness}$), *cluster cutoff point* ($d_{MustBreak}$), and *cluster unbreak point* ($d_{MustNotBreak}$). In Fig. 5.2.3(a), we also see how the 30 elements are organized into nine robust clusters based on the criteria we defined. These will be clustered a second time using the IHC to give the dendrogram in Fig. 5.2.3(b).

Once we have the robust clusters, the next step is to perform hierarchical clustering treating these robust clusters as the objects. To do this, we need to construct the distance matrix **ID** for robust clusters. We define the interaction distance $ID_{RobustCluster}(\mathbf{cl}_i^{(R)}, \mathbf{cl}_j^{(R)})$

between two robust clusters $\mathbf{cl}_i^{(R)}$ and $\mathbf{cl}_j^{(R)}$ as the average inter-cluster pair distance

$$ID^{(RobustCluster)}(\mathbf{cl}_i^{(R)}, \mathbf{cl}_j^{(R)}) = \frac{1}{N_i N_j} \sum_{k_i=1}^{N_i} \sum_{k_j=1}^{N_j} D(\mathbf{cl}_i^{(R)}(k_i), \mathbf{cl}_j^{(R)}(k_j)), \quad (5.10)$$

where $D(\mathbf{cl}_i^{(R)}(k_i), \mathbf{cl}_j^{(R)}(k_j))$ is the distance between the k_i^{th} element in robust cluster $\mathbf{cl}_i^{(R)}$ and the k_j^{th} element in robust cluster $\mathbf{cl}_j^{(R)}$. Like the hierarchical clustering of stocks described in Section 5.2.1 the pair of clusters with shortest distance according to Eq.(5.10) will be merged first to form a bigger cluster. This process is iterated until only one cluster remains, and we end up with the IHC dendrogram (see Fig. 5.2.3(b)). As we can see comparing Fig. 5.2.2(b) with Fig. 5.2.2(c)~(d), the diagonal blocks are better defined in the IHC-seriated matrix than in the SLHC-seriated and CLHC-seriated matrix. This is because clusters with strong interactions, but are far apart in the dendrogram because of the history dependence of SLHC and CLHC are now brought closer together.

Fig. 5.2.3. A schematic description (a) for *robustness length* ($l_{Robustness}$), *cluster cutoff point* ($d_{MustBreak}$), and *cluster unbreak point* ($d_{MustNotBreak}$). A cluster is considered robust if its robustness length is greater than a threshold, as it retains its identity over a large range of cutoffs. However, we do not consider clusters that are formed above the upper cutoff as robust, as their components are weakly or anti-correlated. Also, we do not admit cutoffs below the lower cutoff, because clusters that are so strongly-correlated should not be broken into smaller clusters. In (b), we show the IHC dendrogram of the 9 robust CLHC clusters, which form the leaves of the dendrogram.

5.2.3 Clusters Identification

So far we have only discussed the constructions of complete-linkage hierarchical dendrogram, finding robust clusters, and the construction of the interaction-hierarchical dendrogram. Although robust clusters themselves can be the clusters we identify, if we want to further group these robust clusters we need an alternative approach. One of the most common ways to do so is to define a cutoff point in the dendrogram, so that the groups

that remain below this cutoff will be the clusters we eventually identify. However, a natural question we must ask is whether there is a way to unambiguously specify this threshold.

Before answering the question, let us consider the artificial correlation matrix shown in Fig. 5.2.4(a). In this artificial correlation matrix, we find four robust clusters (RC_1, RC_4, RC_3, RC_2) with interaction blocks (IB_1, IB_2, IB_3, IB_4). Its IHC dendrogram is shown in Fig. 5.2.4(b). In general, the lower the cutoff the more robust clusters we obtain. In this process, the most important compromise we have to make is: if we break a cluster up into two clusters, then some intra-cluster correlations will become inter-cluster correlations. For example, when we lower the cutoff in Fig. 5.2.4(b) from Region 1 to Region 2, we go from two clusters, the first represented by the correlations in $\{RC_1, RC_4, RC_3, IB_1, IB_2\}$, and the second represented by the correlations in $\{RC_2\}$, to three clusters $\{RC_1, RC_4, IB_1\}$, $\{RC_3\}$, and $\{RC_2\}$. Here, note that the IB_2 correlations which were considered intra-cluster correlations in Region 1 has become inter-cluster correlations in Region 2. Similarly, when we lower the cutoff from Region 2 to Region 3, the IB_1 correlations will also go from being intra-cluster correlations to inter-cluster correlations.

Fig. 5.2.4. In (a) we show an artificial correlation matrix with four robust clusters: RC_1, RC_4, RC_3 , and RC_2 . Interactions between robust clusters are weak, except for that between RC_1 and RC_4 . In (b), we show the IHC dendrogram for this artificial correlation matrix, where we distinguish between three regions: Region 1 with two clusters, Region 2 and Region 3 with three and four clusters respectively. As the cutoff is lowered, more clusters will be formed, and previously intra-cluster correlations become inter-cluster correlations. In (c), we show the normalized distributions of existing inter-cluster correlations (green), newly inter-cluster correlations (red), and remaining intra-cluster correlations (blue). The top panel of (c) corresponds to the situation encountered when we lower the cutoff from Region 1 to Region 2. The bottom panel of (c) corresponds to the cutoff being lowered from Region 2 to Region 3.

More importantly, as Fig. 5.2.4(b) illustrates, when the cutoff is progressively lowered, we find initially the intra-cluster correlations that becomes inter-cluster correlations are those that are statistically more similar to the inter-cluster correlations than they are to the intra-cluster correlations. Eventually, the cutoff becomes so low that the intra-cluster correlations that become inter-cluster correlations are statistically more similar to the intra-cluster correlations than they are to the inter-cluster correlations. With this un-

derstanding, we select the optimum cutoff to be the one at which the new inter-cluster correlations switch from being more similar to existing inter-cluster correlations to being more similar to the remaining intra-cluster correlations. There are many ways to measure this statistical similarity. Here we use the overlaps between the normalized distribution of new inter-cluster correlations with the normalized distribution of existing inter-cluster correlations and the normalized distribution of remaining intra-cluster correlations as the measure of statistical similarity, see Fig. 5.2.4(c). With this we are now able to reliably identify clusters from those robust clusters obtain from complete-linkage hierarchical clustering.

GitHub: Interaction-Hierarchical Clustering

The codes for determine robust clusters, interaction-hierarchical clustering, and identifying cluster are available at <https://github.com/BoonKinTeh/InteractionHierarchicalClustering>, and please refer to Teh *et al.* “Cluster fusion-fission dynamics in the Singapore stock exchange” for more details [174], and cite the paper if you are using the codes.

5.3 Clustering Analysis II (Supervised)

As we have illustrated in Section 5.2, hierarchical clustering suffers from the problem of history dependence. Suppose two elements are close to each other, but are each closer to elements that are further away. Once they are clustered separately, they will sometimes end up in two different high-level clusters. This happens for various linkage hierarchical clustering (complete, single and average-linkage), in which two high-level clusters suggests that the two elements are not close to each other, when in actual fact this is not the case. Fortunately, this problem can be alleviated using multi-stage hierarchical clustering such as interaction-hierarchical clustering. Besides this history dependence problem, a single threshold is also commonly assumed in hierarchical clustering methods [175–177] for breaking up a dendrogram into the desired number of clusters. In a multi-scale complex system, different clusters may have different densities, and hence should be merged at different thresholds. The natural way to select clusters is therefore to use variable thresholds, but there are no good prescriptions for doing so.

5.3.1 *Partial Hierarchical Clustering*

In thinking about this problem of identifying clusters with different densities, we realized that if we were to just identify a single cluster, and not parallelize the process, the single cluster can be identified naturally. This is shown in Fig. 5.3.1, for three natural clusters of points. Since we are still performing hierarchical clustering, aiming to identify one cluster instead of all clusters in parallel, we call this method *partial hierarchical clustering*. In this method, we start by identifying the pair of elements that are closest together, and use as the seed to grow the cluster by continuously adding elements closest to the cluster. To identify the natural clusters, we then look out for sharp drops in the linkage correlation as the cluster grows.

As illustration, we use a system with 21 elements shown in Fig. 5.3.1. The seed cluster is $\{1, 17\}$ from cluster A . To grow this seed cluster, we first calculate the distance between elements with cluster A using maximum distance, then add the closest element into cluster A . This is then repeated to get a larger and larger cluster. Besides the complete-linkage, we can also use the single-linkage or the average-linkage algorithm for partial hierarchical clustering, but the features in the linkage distance-cluster size graph that demarcates a natural cluster of the rest of the points will be different from that shown in Fig. 5.3.1(b) and 5.3.1(c). In Fig. 5.3.1(b), we see that points are added to the growing cluster in the sequence $\{1, 17, 18, 12, 13, 5, 16, 10, 8, 2, 4, 7, 21, 6, 19, 9, 3, 14, 20, 11, 15\}$. The last member of cluster A added is 16. After this, there is a sharp jump in the complete-linkage distance as 10, which belongs to cluster B , is added. When the complete-linkage algorithm is used, this sharp jump is a natural signature that one cluster has been exhausted, and we are beginning to add members from a different cluster. We therefore accurately identified cluster A ($\text{cl}_A = \{1, 17, 18, 12, 13, 5, 16\}$) as a natural cluster. We also expect a sharp jump when cluster B is exhausted, and we move on to add a member of cluster C to the growing cluster. However, in Fig. 5.3.1(b), we find no such jump going from 19 (cluster B) to 9 (cluster C). This behavior is not generic,

but when it happens, full hierarchical clustering will potentially encounter the problem of history dependence.

Fig. 5.3.1. (a) 21 elements, organized into three clusters A $\{1, 5, 12, 13, 16, 17, 18\}$, B $\{2, 4, 6, 7, 8, 10, 19, 21\}$, and C $\{3, 9, 11, 14, 15, 20\}$. (b) Starting from the seed cluster $\{1, 17\}$, we grow the cluster one member at a time by incorporating the point whose maximum distance from members of the cluster is the minimum among all such maximum distances. The graph of maximum distance at which each member is admitted into the growing cluster shows a sharp jump after all members of cluster A have been included, and the first member of cluster B is added. (c) After cluster A is removed, we restart the partial hierarchical clustering with a seed cluster $\{4, 8\}$ from cluster B . The graph of maximum distance at which each member is admitted into the growing cluster shows a sharp jump after all members of cluster B have been included, and the first member of cluster C is added.

Looking more carefully at the distribution of points in Fig. 5.3.1(a), we see that when 19 is added to the growing cluster, it has seven large distances with members of cluster A , and seven small distances with members of cluster B . On the other hand, when 9 is added to the growing cluster, it has seven large distances with members of cluster A , and eight moderate distances with members of cluster B . This distinction between 19 and 9 is lost when we compute the complete-linkage distance, which is the maximum of all distances with members of the growing cluster. The result is that 19 and 9 have very similar complete-linkage distances from the growing cluster.

To circumvent this problem, which is caused by the presence of cluster A , we simply remove cluster A from the problem, and restart the partial hierarchical clustering using a new seed cluster. This is shown in Fig. 5.3.1(c), where we use $\{4, 8\}$ from cluster B as the new seed cluster. In this restarted partial hierarchical clustering, the remaining points are added to the growing cluster in the sequence $\{4, 8, 2, 21, 6, 7, 10, 19, 3, 9, 11, 20, 14, 15\}$. We also see from Fig. 5.3.1(c) that after cluster B is exhausted (after the addition of

19), there is a sharp jump in the linkage distance when a member of cluster C (3) is added. This sharp jump allows us to accurately identify $\mathbf{cl}_B = \{4, 8, 2, 21, 6, 7, 10, 19\}$ as a natural cluster. Removing cluster B from the problem, and restarting the partial hierarchical clustering using a seed cluster from C , we will find no further sharp jumps in the graph of linkage distance as a function of cluster size. This means that it is reasonable to treat the remaining points $\mathbf{cl}_C = \{3, 9, 11, 20, 14, 15\}$ as a natural cluster.

5.3.2 PHC Sensitivity Analysis

Since the partial hierarchical clustering is a supervised algorithm, that required human expert to determine the sharp jump (kink) in the linkage distance. Thus, after identifying the clusters we need to perform sensitivity analysis to confirm our results. In Fig. 5.3.2(a) we show the partial hierarchical clustering results in a typical case, and contrast it with the best (Fig. 5.3.2(b)) and worst cases (Fig. 5.3.2(c)) with kink angles. As a rule of thumb, the best and typical cases form dense bands away from 180° . We are certain therefore that the identification of these clusters are reliable. For the worst case scenario, the kink angles that fall within the ambiguous range of 160° to 200° , which we are less confident with the clusters identified.

Fig. 5.3.2. The comparison between a (a) typical, (b) good and (c) poor outcomes from partial hierarchical clustering, where the cluster boundary (kink) is marked with red star. The boundary is identified visually and is unlikely to be wrong for typical and good cases, whereas for poor case the cluster identified is ambiguous, whether this is done visually or automatically.

5.4 Clusters Mixing Degree Measures

In addition to clustering analysis, we also interested in the dynamics of the clusters, in particular how strongly they mix with each other as time progresses. The mixing degree act as an indicator of persistence level for how long two elements remain in the same cluster. Consequentially, the mixing degree reflects the diversity of opinions within the system, for instance high mixing degree often associated with high diversity and rapid changes of opinions.

5.4.1 Rand Index, Adjusted Rand Index

Suppose we have a set $\mathbf{E} = \{E_1, E_2, \dots, E_N\}$ with elements E_i . Furthermore, these elements form two different cluster lists $\mathbf{CL}(t_1) = \{\mathbf{cl}_1(t_1), \mathbf{cl}_2(t_1), \dots, \mathbf{cl}_u(t_1)\}$ at time t_1 and $\mathbf{CL}(t_2) = \{\mathbf{cl}_1(t_2), \mathbf{cl}_2(t_2), \dots, \mathbf{cl}_v(t_2)\}$ for time t_2 . Here $\mathbf{cl}$ represents a non-empty cluster, and $\cup_{i=1}^u \mathbf{cl}_i(t_1) = \cup_{j=1}^v \mathbf{cl}_j(t_2) = \mathbf{E}$. To measure the similarity between $\mathbf{CL}(t_1)$ and $\mathbf{CL}(t_2)$, the *Rand Index* (*RI*) (developed by Rand [46]) measures the percentages of pairs of elements that remain in same cluster or remain as different clusters in both t_1 and t_2 , mathematically

$$RI(\mathbf{CL}(t_1), \mathbf{CL}(t_2)) = \frac{V_1 + V_2}{V_1 + V_2 + V_3 + V_4} = \frac{V_1 + V_2}{\binom{N}{2}}, \quad (5.11)$$

where V_1 is the number of pairs remaining in the same cluster for both t_1 and t_2 ($\{E_m, E_n\} \in \mathbf{cl}_a(t_1)$ and $\{E_m, E_n\} \in \mathbf{cl}_b(t_2)$ for arbitrary a and b), V_2 is the number of pairs remaining in different clusters for both t_1 and t_2 ($\{E_m, E_n\}$ which $E_m \in \mathbf{cl}_a(t_1), E_n \in \mathbf{cl}_b(t_1)$ and $E_m \in \mathbf{cl}_c(t_2), E_n \in \mathbf{cl}_d(t_2)$ for arbitrary a, b, c , and d that $a \neq b$ and $c \neq d$). In addition, V_3 is the number of pairs in same cluster at t_1 but different clusters at t_2 ($\{E_m, E_n\} \in \mathbf{cl}_a(t_1)$ and $E_m \in \mathbf{cl}_c(t_2), E_n \in \mathbf{cl}_d(t_2)$ for arbitrary a, c , and d that $c \neq d$), and V_4 is the number of pairs in different clusters at t_1 but the same cluster at t_2 ($E_m \in \mathbf{cl}_a(t_1), E_n \in \mathbf{cl}_b(t_1)$ and $\{E_m, E_n\} \in \mathbf{cl}_c(t_2)$ for arbitrary a, b , and c that $a \neq b$).

One shortcoming of *RI* is that it is highly dependent on the number of clusters and the distribution of clusters, leading to some combinations having higher index values by chance. The well known *Adjusted Rand Index* (*ARI*), proposed by Hubert and Arabie [47], corrected the *RI* by removing the expected index and standardizing it for the sake of comparison. The *ARI* is defined as

$$\begin{aligned} ARI(\mathbf{CL}(t_1), \mathbf{CL}(t_2)) &= \frac{\text{Index} - \text{Expected Index}}{\text{Max Index} - \text{Expected Index}} \\ &= \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\binom{N}{2}}}{\frac{1}{2} \left(\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2} \right) - \frac{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\binom{N}{2}}}, \end{aligned} \quad (5.12)$$

where N is the number of elements, n_{ij} is the number of shared elements between cluster $\mathbf{cl}_i(t_1)$ and $\mathbf{cl}_j(t_2)$, n_i is number of elements in cluster $\mathbf{cl}_i(t_1)$, and n_j is number of elements in cluster $\mathbf{cl}_j(t_2)$. If the second list of clusters is obtained by randomly re-clustering the first list of clusters, the ARI is 0. In contrast, when the second list of clusters is identical to the first list of clusters, the ARI is 1. Since $ARI = 0$ when the elements are randomly reclustered, and $ARI = 1$ when they remain in the same clusters, the mixing degree can be defined as $1 - ARI$.

5.4.2 Degree of Mixing

Note that RI and ARI both assumed that the elements set $\mathbf{E}$ remains constant from time t_1 to t_2 . This is often not the case in real data, where old elements can be eliminated and new elements added. Furthermore, ARI is a global measurement for mixing, and does not tell us what the mixing is locally for each cluster in $\mathbf{CL}$. Yet, often we are only interested on the mixing degree of first k largest clusters, instead of all (small) clusters which can be treated as noise. Because of this, we introduced the *degree of mixing* (DoM) measure that can be used for individual clusters within a cluster list.

Fig. 5.4.1. A schematic illustration for mixing degree based on DoM for two adjacent windows: (a) before and (b) after normalization against the cluster sizes. Each cluster $\mathbf{cl}$ can be broken down to two sub-clusters $\mathbf{cl}^{In}$ and $\mathbf{cl}^{Ext}$, where $\mathbf{cl}^{In}$ consists of elements that appear in both windows. In contrast, elements within $\mathbf{cl}^{Ext}$ only appear in one of the windows and will be excluded when we compute the DoM . When the clusters are maximally mixing, it is equally likely for an element in a $\mathbf{CL}(t_1)$ cluster to move to any cluster in $\mathbf{CL}_{t+1}$, thus the number of elements from cluster $\mathbf{C}_i^{InI}$ transferred to cluster $\mathbf{C}_j^{InII}$ is proportional to size of cluster $\mathbf{C}_i^{InI}$. Else, after normalization against the cluster sizes, $\mathbf{NC}_j^{InII}$ receives equal contributions from clusters $\mathbf{NC}_i^{InI}$, $i = \{1, 2, 3\}$.

Suppose the cluster lists at time t_1 and t_2 are $\mathbf{CL}(t_1) = \{\mathbf{cl}_1(t_1), \mathbf{cl}_2(t_1), \dots, \mathbf{cl}_u(t_1)\}$ and $\mathbf{CL}(t_2) = \{\mathbf{cl}_1(t_2), \mathbf{cl}_2(t_2), \dots, \mathbf{cl}_v(t_2)\}$ respectively. Suppose further that we are only interested in the mixing between the k largest clusters in each list, $\mathbf{CL}(t_1) = \{\mathbf{cl}_{(1)}(t_1), \mathbf{cl}_{(2)}(t_1), \dots, \mathbf{cl}_{(k)}(t_1)\}$ and $\mathbf{CL}(t_2) = \{\mathbf{cl}_{(1)}(t_2), \mathbf{cl}_{(2)}(t_2), \dots, \mathbf{cl}_{(k)}(t_2)\}$, which are significant clusters in term of size. The *mixing degree*, MD for $\mathbf{cl}_{(j)}(t_2) \in \mathbf{CL}(t_2)$ due to

$\mathbf{CL}(t_1)$ is defined as

$$MD(\mathbf{cl}_{(j)}(t_2), \mathbf{CL}(t_1)) = \sqrt[k(\mathbf{CL}(t_1))]{\prod_{i=1}^{k(\mathbf{CL}(t_1))} f\left(\frac{n(\mathbf{cl}_{(i)}^{In}(t_1) \cap \mathbf{cl}_{(j)}^{In}(t_2)) / n(\mathbf{cl}_{(i)}^{In}(t_1))}{\sum_{i=1}^{k(\mathbf{CL}(t_1))} n(\mathbf{cl}_{(i)}^{In}(t_1) \cap \mathbf{cl}_{(j)}^{In}(t_2)) / n(\mathbf{cl}_{(i)}^{In}(t_1))}\right)}} \quad (5.13)$$

In Eq.(5.13), $k(\mathbf{CL}(t))$ is the k largest clusters in $\mathbf{CL}(t)$, $n(\mathbf{cl})$ is the number of elements within $\mathbf{cl}$. We also distinguish between two subclusters $\mathbf{cl}^{In}$ and $\mathbf{cl}^{Ext}$, where elements of $\mathbf{cl}^{In}$ appear in both $\mathbf{CL}(t_1)$ and $\mathbf{CL}(t_2)$, in contrast elements of $\mathbf{cl}^{Ext}$ only appear in either $\mathbf{CL}(t_1)$ or $\mathbf{CL}(t_2)$. Finally, $f(x)$ is given as

$$f(x) = \frac{x(x-1)\exp(-(A-x)^2)}{\alpha(\alpha-1)\exp(-(A-\alpha)^2)}; \quad \alpha = \frac{1}{k(\mathbf{CL}(t_1))}; \quad A = \frac{\alpha^3 - \alpha^2 - \alpha + \frac{1}{2}}{\alpha^2 - \alpha} \quad (5.14)$$

so that $f(0) = 0$, $f(1) = 0$, and $f(\alpha) = 1$ and also its derivative with respect to x is $f_x(x < \alpha) > 0$, $f_x(x = \alpha) = 0$, and $f_x(x > \alpha) < 0$.

Intuitively, the mixing degree of $\mathbf{cl}_{(j)}(t_2)$ is highest when its elements are contributed equally (size normalized) by all clusters from $\mathbf{CL}(t_1)$, and lowest when it is identical to or a subset of one of the clusters from $\mathbf{CL}(t_1)$ (see Fig. 5.4.1). There are three reasons why we define MD for $\mathbf{cl}_{(j)}(t_2)$ relative to $\mathbf{CL}(t_1)$ in this way: (I) it normalized against the cluster sizes of all clusters within cluster list $\mathbf{CL}(t_1)$, so that for maximum mixing degree, a larger cluster will contribute a larger portion in cluster $\mathbf{cl}_{(j)}(t_2)$; (II) it also normalized against the sizes of clusters in $\mathbf{CL}(t_2)$, such that is bounded between 0 and 1; (III) the function $f(x)$ peaks at $1/N(\mathbf{CL}(t_1))$, which corresponds to equal contributions from all clusters in $\mathbf{CL}(t_1)$, and decreases monotonically when x deviates from $1/N(\mathbf{CL}(t_1))$. Due to this, the geometric mean of $f(x)$ will peak at 1 if it is well mixed, and decreases from 1 monotonically if there are deviations. The mixing degree measured with Eq.(5.13) is focus on individual cluster in $\mathbf{CL}(t_2)$. In order to get the overall mixing degree between $\mathbf{CL}(t_1)$ and $\mathbf{CL}(t_2)$, we define the *degree of mixing* (DoM) as

$$DoM(\mathbf{CL}(t_1), \mathbf{CL}(t_2)) = \frac{1}{k(\mathbf{CL}(t_2))} \sum_{i=1}^{k(\mathbf{CL}(t_2))} MD(\mathbf{cl}_{(i)}(t_2), \mathbf{CL}(t_1)), \quad (5.15)$$

which is the average of mixing degree for the k largest clusters in $\mathbf{CL}(t_2)$.

5.4.3 Cluster Mixing Degree Significance Testing

When the cluster mixing degree changes from time window to time window, we want to know how significant this change is. Therefore, we estimate the uncertainty in the mixing degree through Monte Carlo simulations. First, we estimate the empirical transition probabilities between clusters in adjacent windows. Suppose at time t_1 the cluster list is $\mathbf{CL}(t_1) = \{\mathbf{cl}_1(t_1), \mathbf{cl}_2(t_1), \dots, \mathbf{cl}_u(t_1)\}$, while at time t_2 the cluster list is $\mathbf{CL}(t_2) = \{\mathbf{cl}_1(t_2), \mathbf{cl}_2(t_2), \dots, \mathbf{cl}_v(t_2)\}$, and $\Gamma(\mathbf{CL}_i^t, \mathbf{CL}_j^{t+1})$ is the transition probability from $\mathbf{cl}_i(t_1)$

to $\mathbf{cl}_j(t_2)$, defined as

$$\Gamma(\mathbf{cl}_i(t_1), \mathbf{cl}_j(t_2)) = \frac{n(\mathbf{cl}_i(t_1) \cap \mathbf{cl}_j(t_2))}{n(\mathbf{cl}_i(t_1))}, \quad (5.16)$$

where $n(\mathbf{cl})$ is the number of elements within $\mathbf{cl}$. In some sense, this is the maximum-likelihood Markov chain model of the observed reclustering from $\mathbf{CL}(t_1)$ to $\mathbf{CL}(t_2)$. In each Monte Carlo simulation, we then go through each $\mathbf{cl}_i(t_1)$, and choose which $\mathbf{cl}_j(t_2)$ its elements go to. After all elements have been reclustered, we compute the $1 - ARI$ mixing degree. After many Monte Carlo simulations, we have a distribution of $1 - ARI$, whose mean is our observed mixing degree, and whose standard deviation tells us how much spread we expect from the Markov chain model of reclustering between time windows t_1 and t_2 .

Fig. 5.4.2. Mixing degree measures using adjusted Rand index ($1 - ARI$) and degree of mixing (DoM) between $\mathbf{CL}(t_1)$ and $\mathbf{CL}(t_2)$. Clusters in $\mathbf{CL}(t_1)$ remain the same for all cases and are color coded with red, green, and blue. In contrast, clusters in $\mathbf{CL}(t_2)$ is changing with different transition rates (see Eq.(5.17)), and are be color coded according to the percentage of members coming from $\mathbf{cl}(t_1)$. Fig. 5.4.2 (a)~(c) shows the results for two clusters in $\mathbf{CL}(t_2)$, and Fig. 5.4.2 (e)~(f) shows the results for three clusters in $\mathbf{CL}(t_2)$.

In order to compare mixing degree measures using ARI and DoM , we simulated 1000 sets of $\mathbf{CL}(t_1)$ and $\mathbf{CL}(t_2)$ with different transition rates. Each simulation start

with $\mathbf{CL}(t_1)$ consists of three clusters $\mathbf{cl}_{Red}(t_1)$ (size 300), $\mathbf{cl}_{Green}(t_1)$ (size 300), and $\mathbf{cl}_{Blue}(t_1)$ (size 400). When $\mathbf{CL}(t_2)$ is constructed using the transition rates proportional to cluster size $\mathbf{cl}(t_1)$, the system is well mixed. In addition, we also introduced the deviation $\delta \in [0, 0.2] + \epsilon$ with normal distributed $\epsilon \sim N(\mu = 0, \sigma = 10^{-3})$, such that the transition matrix becomes

Fig. 5.4.2 (a)~(c)	$\mathbf{cl}_1(t_2)$	$\mathbf{cl}_2(t_2)$	Fig. 5.4.2 (e)~(f)	$\mathbf{cl}_1(t_2)$	$\mathbf{cl}_2(t_2)$	$\mathbf{cl}_3(t_2)$
$\mathbf{cl}_{Red}(t_1)$	$0.3 + \delta$	$0.3 - \delta$	$\mathbf{cl}_{Red}(t_1)$	$0.3 + \delta$	$0.3 + \delta$	$0.3 - 2\delta$
$\mathbf{cl}_{Green}(t_1)$	$0.3 - \delta$	$0.3 + \delta$	$\mathbf{cl}_{Green}(t_1)$	$0.3 - \delta$	$0.3 - \delta$	$0.3 + 2\delta$
$\mathbf{cl}_{Blue}(t_1)$	0.4	0.4	$\mathbf{cl}_{Blue}(t_1)$	0.4	0.4	0.4

(5.17)

for two clusters at t_1 Fig. 5.4.2 (a)~(c) and three clusters at t_2 Fig. 5.4.2 (e)~(f). $1 - ARI$ and DoM agree well for small deviations, but DoM decreased more rapidly than $1 - ARI$ as deviation increased for both two and three clusters at t_2 . This shows DoM is more sensitive on changes in clusters elements.

GitHub: Clusters Mixing Degree

The codes for computing clusters mixing degree using Adjusted Rand Index, ARI and Degree of Mixing, DoM are available at

<https://github.com/BoonKinTeh/ClustersMixingDegree>, and please refer to Teh *et al.* “Cluster fusion-fission dynamics in the Singapore stock exchange” for more details [174], and cite the paper if you are using the codes.

5.5 Soup-of-Groups Model

5.5.1 Soup-of-Groups (SoG) Model

The *soup-of-groups* (SoG) model was introduced by Johnson *et al.* to study emergent behaviors in vastly different complex systems. It is a macroscopic model, where it approaches the problem at the aggregate level without explicitly list down all the possible interactions between the agents. It was first applied to model the underlying dynamics of human insurgency [48], where terrorists are assumed to form groups that merge with each other in preparation for attacks, and disintegrate after a successful attack or to avoid detection by security forces. Johnson *et al.* found by comparing the model against a database of past terrorist attacks that the terrorist groups are equally effective, and the number of casualties depends only on the group size, which is distributed as a power law with exponent $\alpha = 2.5$. Following this, Johnson *et al.* also tried to explain contagion dynamics on social network in term of the SoG model, particularly in the intermediate regime where individual behaviors sensitively influence the structures of social groups [50]. In the follow up paper [51] they suggested that the escalation rate and timing of the fatal attacks from terrorist follow naturally from the SoG dynamics.

Fig. 5.5.1. Illustration of the soup-of-groups (SoG) model, as a system of clusters with various sizes, where clusters can experience either fusion or fission with probability ν_+ or ν_- respectively. The upper half represents the fusion process as two clusters with size s_i and s_j are selected and fuse into a larger cluster with size $s_i + s_j$. In contrast, during the fission process (lower half) the selected cluster with size s_i completely fragments into s_i size-1 clusters.

The SoG model, as the name suggests, consists of a “soup” (system) of “groups” (clusters) with sizes ranging from 1 to N , where N is the total number of particles within

the system. The essence of SoG dynamics is illustrated in Fig. 5.5.1, where a cluster of size s_i can merged with a cluster of size s_j to give a cluster with size $s_i + s_j$ at a rate of ν_+ , or a cluster of size s_k can fragment completely into s_k clusters with size-1 at a rate of ν_- . These fusion-fission processes can be described by the master equation

$$\begin{aligned}\frac{\partial N_1}{\partial t} &= \dot{N}_1 = \nu_- \sum_{k=2}^{\infty} k \frac{k}{N} N_k - 2\nu_+ \frac{1}{N} N_1 \sum_{k=1}^{\infty} \frac{k}{N} N_k, \\ \frac{\partial N_s}{\partial t} &= \dot{N}_s = \nu_+ \sum_{k=2}^{s-1} \frac{k}{N} N_k \frac{s-k}{N} N_{s-k} - 2\nu_+ \frac{s}{N} N_s \sum_{k=1}^{\infty} \frac{k}{N} N_k - \nu_- \frac{s}{N} N_s; \quad s \geq 2.\end{aligned}\tag{5.18}$$

Fig. 5.5.2. SoG equilibrium distribution for various fusion-fission rate, $\nu_+ \in (0, 1)$, $\nu_- = 1 - \nu_+$ and $N = 50,000$. (a) The exponential beta, S_o dependency on fusion-fission rate. (b) Also the dependency SoG equilibrium distribution, $f_{ETPL}(s|\nu_+)$ on fusion-fission rate from $\nu_+ = 0.1$ (blue) to $\nu_+ = 0.9$ (red).

In the master equation, we assume that all elements are equally likely to participate in fusion and fission processes, thus a larger cluster will have higher chances for being selected. A cluster of size k will disintegrate with a probability proportional to k/N to yield k elements (first term in $\dot{N}_1$ and last term in $\dot{N}_s$), whereas two clusters with sizes k and $s - k$ will merge with probability proportional to $k/N \cdot (s - k)/N$ to yield a cluster of size s (first term in $\dot{N}_1$ and last term in $\dot{N}_s$). In the equation of rate of change of cluster with size 1, $\dot{N}_1$, the first term indicates fragmentation of clusters with size larger than $s > 1$ that yields s clusters with size 1, and the second term indicates fusion process of cluster with size 1 with other clusters to form a larger cluster. The three terms in the rate of change equation of cluster with size $s \geq 2$, $\dot{N}_s$ correspond to merging of clusters with size $s - k$ and k that give a cluster with size s , fusion of size s cluster with other clusters, and complete fragmentation of size s cluster respectively. From this equation, Johnson *et al.* solved for the equilibrium distribution of an infinite system, which is an

exponentially-truncated power law distribution (*ETPL*) with expression as:

$$N_s = f_{ETPL}(s) \sim s^{-\frac{5}{2}} \exp(-S_o s), \quad S_o = -\ln \left\{ \frac{4\nu_+ (\nu_+ + \nu_-)}{(2\nu_+ + \nu_-)^2} \right\}, \quad (5.19)$$

and the dependency of S_o and hence $f_{ETPL}(s)$ on fusion-fission rate is shown in Fig. 5.5.2.

5.5.2 *Stability of SoG Model*

In order for *ETPL* distribution to be observed in the SoG system over a long period, we expect it to be stable under SoG dynamics. This means the equilibrium *ETPL* distribution act as an attractor in its dynamical phase space, such that over a long period the system will converge to this attractor. There are many ways to test the stability of a system. Here we use the common linear stability test which involved two steps: given the dynamical equation $\dot{\mathbf{x}} = f_{\mathbf{x}}$, we first (I) linearize the dynamical equations about the point of interest or fixed point $f_{\mathbf{x}} \approx f_{\mathbf{x}_{fixed}} + \mathbf{J}_{\mathbf{x}_{fixed}}(\mathbf{x}(t) - \mathbf{x}_{fixed}) = \mathbf{J}_{\mathbf{x}_{fixed}}(\mathbf{x}(t) - \mathbf{x}_{fixed})$ as $f_{\mathbf{x}_{fixed}} = 0$. After that we (II) diagonalize the Jacobian matrix $\mathbf{J}_{\mathbf{x}_{fixed}}$ to obtain its eigenvalues.

When the dynamical equations are expressed as $\dot{\mathbf{x}} = \mathbf{J}(\mathbf{x} - \mathbf{x}_{fixed})$, the solutions are $\mathbf{v}_i \cdot \mathbf{x}(t) = \exp(\lambda_i t) \mathbf{v}_i$, where λ is the eigenvalue with corresponding eigenvector $\mathbf{v}$ center at $\mathbf{x}_{fixed}$. The term $\exp(\lambda t)$ can be further decomposed to $\exp(\text{Re}(\lambda)t) \exp(\text{Im}(\lambda)t)$, and it is easy to see that for $t \rightarrow \infty$: $\exp(\lambda t) \rightarrow 0$ for $\text{Re}(\lambda) < 0$, and $\mathbf{x} \rightarrow \mathbf{x}_{fixed}$ represents a stable system (manifold). In contrast, for $\text{Re}(\lambda) > 0$ then $\exp(\lambda t) \rightarrow \infty$, $\mathbf{x} \rightarrow \infty$ indicates an unstable system (manifold). Thus, to test the SoG stability, we can determine the eigenvalues of its Jacobian matrix we derive from Eq.(5.18) as

$$\mathbf{J} = \begin{pmatrix} \frac{-2\nu_+}{N} & \frac{2^2\nu_-}{N} & \frac{3^2\nu_-}{N} & \cdots & \frac{c^2\nu_-}{N} & \cdots & \frac{N^2\nu_-}{N} \\ \frac{1(2-1)2\nu_+}{N^2} N_{2-1}^{(*)} & \frac{-(2\nu_+ + \nu_-)2}{N} & 0 & 0 & 0 & 0 & 0 \\ \frac{1(3-1)2\nu_+}{N^2} N_{3-1}^{(*)} & \frac{2(3-2)2\nu_+}{N^2} N_{3-2}^{(*)} & \frac{-(2\nu_+ + \nu_-)3}{N} & 0 & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \frac{1(c-1)2\nu_+}{N^2} N_{c-1}^{(*)} & \frac{2(c-2)2\nu_+}{N^2} N_{c-2}^{(*)} & \frac{3(c-3)2\nu_+}{N^2} N_{c-3}^{(*)} & \cdots & \frac{-(2\nu_+ + \nu_-)c}{N} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \frac{1(N-1)2\nu_+}{N^2} N_{N-1}^{(*)} & \frac{2(N-2)2\nu_+}{N^2} N_{N-2}^{(*)} & \frac{3(N-3)2\nu_+}{N^2} N_{N-3}^{(*)} & \cdots & \frac{c(N-c)2\nu_+}{N^2} N_{N-c}^{(*)} & \cdots & \frac{-(2\nu_+ + \nu_-)N}{N} \end{pmatrix} \quad (5.20)$$

where $N_s^{(*)}$ is the fixed points (equilibrium distribution) for cluster with size s .

As *ETPL* distribution is the equilibrium distribution for an infinitely large system, ideally we need to determine the stability of the SoG model for an infinite system. However, in practice, we cannot simulate an infinite large system. Fortunately we can overcome this limitation by using finite size scaling on the eigenvalue to extrapolate it to infinitely large systems. In Fig. 5.5.3, we show the largest eigenvalue $\lambda_{(1)}$ is real, negative, non-

degenerate, and scales with system size N as a power law ($-\lambda_{(1)} \propto N^{-\alpha}$ for all ν_+). This result implies that the *ETPL* equilibrium distribution for the SoG model is indeed stable as the largest eigenvalue is negative, but it becomes marginally stable in the limit of infinite system size.

Fig. 5.5.3. Linear stability test of SoG Model about its equilibrium distribution $N_s^{(*)} \sim f_{ETPL}(s)$. (a) The color-map of $-Re(\lambda)$ from largest to 25th largest in log scale, at different fusion-fission rates with $\nu_+ + \nu_- = 1$. Fig. 5.5.3(b) and (c) show how the largest ($-Re(\lambda_{(1)})$) and second largest eigenvalues ($-Re(\lambda_{(2)})$) scale with system size (N).

5.5.3 *SoG Model Dynamics* (Preliminary)

There are signatures observed from the empirical data that cannot be explained by the SoG forecasting model. For example, large events seemingly not predicted by our SoG forecasting model can be found empirically prior to the giant event in the financial market. Therefore, to understand the limitations of our SoG forecasting model, we need to understand the dynamics of the events as well as the giant clusters in greater details. To do this, we simulated the SoG dynamics using 50,000 elements for various fusion rates $\nu_+ \in [0.1, 0.9]$ with $\nu_- = 1 - \nu_+$. For each fusion rate, we run 30 sets of simulations, and for each simulation we run one million simulation steps. We assume an event is due to fragmentation of a cluster, with event magnitude equal to the size of fragmented cluster. Moreover, we define a giant event as the event with magnitude at $100 - 10^{-5}$ percentile or higher, and the corresponding cluster that fragmented is called as the giant cluster. To study the dynamics of giant clusters and events, we use the sequence of giant events to segment the times series. As illustration, for a time series for $t \in [1, T]$

with k giant events at $1 < t_1 < t_2 < \dots < t_k < T$, there are k segmented time series as $t^{(S_1)} \in [1, t_1], t^{(S_2)} \in [t_1 + 1, t_2], \dots, t^{(S_k)} \in [t_{k-1} + 1, t_k]$. Overall we collected about 200 time series segments for each fusion rate, and test the dynamic growth for the giant clusters as well as for the cumulative events within the segment.

Suppose we are given a (monotonically increasing) segmented time series of giant clusters and cumulative event for $t^{(S_i)} \in [t_{i-1}, t_i]$. To examine the growth we first partition the time series into two sub-segments ($t_{Sub1}^{(S_i)} \in [t_{i-1}, 0.5 * (t_{i-1} + t_i)]$ and $t_{Sub2}^{(S_i)} \in [0.5 * (t_{i-1} + t_i), t_i]$), then perform linear fits to both sub-segments to obtain their gradients g_1 and g_2 . By comparing these gradients we can gauge whether the growth is accelerating, decelerating, or strictly linear, if we define the growth dynamics coefficient as

$$GDC(g_1, g_2) = \frac{g_2}{g_1} - 1. \quad (5.21)$$

If $GDC > 0$ it indicates acceleration growth as the magnitude increase more rapidly in second half segment, conversely $GDC < 0$ means deceleration growth, and $GDC = 0$ represents linear growth.

Fig. 5.5.4. The growth dynamics coefficient, GDC for cumulative event as a (a) frequency heat map and (b) 3D-plot.

In Fig. 5.5.4, we plot the GDC for cumulative event as a (a) frequency heat map and (b) 3D-frequency plot. For low-fusion rate and high-fission rate ($\nu_+ \lesssim 0.5$, $\nu_- \gtrsim 0.5$), the GDC s for cumulative event are centered at zero indicate linear growth. This is because under this low-fusion high-fission environment it is difficult for clusters to grow bigger, hence the majority of clusters are small with small spread of the cluster size (see Fig. 5.5.2). However, as the fusion rate increases, the cumulative event can indeed experience acceleration, linear, and deceleration growths. This is due to the largest cluster growing bigger to form a giant cluster, leading to a larger spread of cluster sizes. Therefore, for a cumulative event to undergo acceleration or deceleration growth, there are two competitive factors: (I) the giant cluster exhausts the elements and leaving other clusters small, as a result we have deceleration growth; (II) on the other hand, when there can be more than one giant clusters, the growth rate for the cumulative event depends on the initial

distribution of clusters. If there are many large clusters the chances for acceleration growth is higher, whereas if there are only a few large clusters, deceleration growth becomes likely.

Meanwhile for the giant cluster growth dynamics, under low-fusion rate and high-fission rate environment ($\nu_+ \lesssim 0.5$, $\nu_- \gtrsim 0.5$), the *GDCs* for the gaint cluster are also centered about zero, indicating linear growth. Similar reason as cumulative event growth dynamics, as it is difficult for cluster to grow larger in low-fusion high-fission rate environment. However, unlike the growth dynamics of cumulative events, it is far more likely for a giant cluster to experience acceleration growth. This observation is not surprising, because the probability for selecting the giant cluster to merge with other cluster is higher when it grows bigger. Furthermore, because of there are more large clusters around, the giant cluster merge with larger and larger clusters as it grows.

Fig. 5.5.5. The growth dynamics coefficient, *GDC* for giant clusters as a (a) frequency heat map and (b) 3D-plot.

5.6 Distribution Statistical Significance Testing

The current era of Big Data has ushered in a new way to look at Science — and that is letting the data speak for itself. Because of this, we are now much more concerned about empirical distributions than we have in the past, and to check what the empirical distributions could be in statistically rigorous ways. In the past, statisticians commonly tested empirical data against the univariate normal distribution [178]. Some of these tests focus on the goodness-of-fit of higher order moments [179–181], while others compare the test statistics against an Empirical Distribution Function (EDF) [182–185]. In 2011, Nornadiah and Yap performed a systematic comparison of Anderson-Darling (AD), Lilliefors, Kolmogorov-Smirnov (KS), and Shapiro-Wilk (SW), using numerical simulation and concluded that the SW test is the best, followed closely by the AD test for a given significance [186].

Among these tests, the KS and Lilliefors tests can also be applied to non-normal distributions. In fact, many real-world data do not follow normal distributions. For instance, many social systems are known to have power-law distributions [187]. These include the financial returns [188–191], word count [192, 193], city size [194, 195], home price [196, 197], wealth and income [198, 199] distributions. One simple but naive way to detect a power law is to plot the data in log-log scale, fit it to a straight line and determine the goodness of fit. However, this simple method has three major flaws: (i) many distributions (eg. exponential, gamma, log-normal) can also look straight in log-log plot, especially if the range of data is small; (ii) the goodness of fit only quantifies how well the fit is visually but does not tell us how plausible the fit is; and (iii) if our data looks straight in both log-log and semi-log plots, the goodness of fit values obtained from the two cannot be directly compared since they were obtained from plots of different scales. Clauset, Shalizi and Newman (CSN) address precisely these three points in their 2009 paper [200], and the test they proposed is now considered by many the gold standard in curve fitting. We shall describe the main idea of the CSN technique in greater mathematical detail in Section 5.6.2.

5.6.1 Review of Probability Distributions and Parameters Estimation

Before discuss about the statistical significance testing, let first review the basic about the probability distribution. A probability distribution is a statistical function that describe all possible outcomes with its probability of occurrence (strictly more or equal to zero, and normalized to 1), it can discrete (probability mass function) or within a range of continuous regime (probability density function). Mathematically, Ω represent the set of all possible outcomes, for $\mathbf{E} \subseteq \Omega$ the probability distribution function $f(\mathbf{E})$ match $\mathbf{E}$ to $\Re \in [0, 1]$, that represent the probability obtaining $\mathbf{E}$ from Ω . If we define $\mathbf{E}$ as the random variable in Ω smaller than x , then $f(\mathbf{E}) = f(\Omega \leq x)$ is the cumulative distribution function $F(x)$.

In this thesis, we use power law (*PL*) and exponential (*EXP*) distributions as illus-

tration. The probability density function for them are

$$\begin{aligned} f_{PL}(x|x_{min}, \alpha) &= \frac{\alpha - 1}{x_{min}^{1-\alpha}} x^{-\alpha}, \\ f_{EXP}(x|x_{min}, \beta) &= \beta \exp(-\beta(x - x_{min})) \end{aligned} \quad (5.22)$$

for $x \in [x_{min}, \infty)$. Whereas their cumulative distribution functions ($\int_{x_{min}}^x f(x) dx$) are

$$\begin{aligned} F_{PL}(x|x_{min}, \alpha) &= 1 - \left(\frac{x}{x_{min}} \right)^{1-\alpha}, \\ F_{EXP}(x|x_{min}, \beta) &= 1 - \exp(-\beta(x - x_{min})). \end{aligned} \quad (5.23)$$

The probability integral transform state that, if $\mathbf{X} = \{x_1, x_2, \dots, x_N\}$ follows probability distribution function f_X with cumulative distribution function F_X , then $F_X(x) = \{F_X(x_1), F_X(x_2), \dots, F_X(x_N)\} \sim U(0, 1)$ is standard uniform distributed. This fact is particularly useful for statistical significance testing, as we can first integral transform $\mathbf{X}$ and compare it with standard uniform distribution $U(0, 1)$. Besides that, we can also generate $\mathbf{X}$ that is f_X distributed from standard uniform distribution using its inverse cumulative distribution function. For instance, the inverse cumulative distribution functions of PL and EXP are

$$\begin{aligned} F_{PL}^{-1}(u|x_{min}, \alpha) &= x_{min}(1 - u)^{\frac{1}{1-\alpha}}, \\ F_{EXP}^{-1}(u|x_{min}, \beta) &= x_{min} - \frac{\ln(1 - u)}{\beta}, \end{aligned} \quad (5.24)$$

which maps $u \sim U(0, 1)$ to $x \in [x_{min}, \infty)$ for the PL and EXP distribution respectively.

When give a sample set $\mathbf{X} = \{x_1, x_2, \dots, x_N\}$, that we expect is $f(x|\theta)$ distributed. In order to test the plausibility $\mathbf{X} \sim f(x|\theta)$, we need to estimate the parameters θ of $f(x|\theta)$ for performing probability integral transform hence statistical significance analysis. There are various ways for estimating the parameters, such as (I) method of moments [201] and (II) likelihood method [202–205]. In method of moments, the parameters are expressed in term of moments (calculated from the sample), and these equations are solved simultaneously for estimated parameters. Hence, the question remains is expressing the parameters in term of moments, which can quite challenging for some distributions. Another alternative is likelihood method, that is estimating the parameters by optimizing the likelihood function. The most common use likelihood function is defined as

$$\mathbb{L} = \prod_{i=1}^N f(x_i|\hat{\theta}), \quad (5.25)$$

and find θ such that the likelihood function is maximized, which is the highest chances to obtain such sample from $f(x)$ distribution.

The likelihood (log-likelihood) function for PL distribution is

$$\begin{aligned}\ln(\mathbb{L}_{PL}) &= \sum_{i=1}^N \ln \left\{ \frac{\hat{\alpha} - 1}{x_{min}} \left(\frac{x_i}{x_{min}} \right)^{-\hat{\alpha}} \right\} \\ &= N \{ \ln(\hat{\alpha} - 1) + \ln(x_{min}) \} - \hat{\alpha} \sum_{i=1}^N \ln \left(\frac{x_i}{x_{min}} \right)\end{aligned}\tag{5.26}$$

and for EXP distribution is

$$\begin{aligned}\ln(\mathbb{L}_{EXP}) &= \sum_{i=1}^N \ln \left\{ \hat{\beta} \exp \left(-\hat{\beta} (x_i - x_{min}) \right) \right\} \\ &= N \ln(\hat{\beta}) - \hat{\beta} \sum_{i=1}^N (x_i - x_{min}).\end{aligned}\tag{5.27}$$

Where the hat indicate estimated parameters. Applying the maximization condition $\frac{\partial \ln \mathbb{L}}{\partial \alpha} = 0$ for the PL and $\frac{\partial \ln \mathbb{L}}{\partial \beta} = 0$ for the EXP distribution, yield the estimated parameters,

$$\begin{aligned}\hat{\alpha} &= 1 + \frac{1}{\langle \ln \left(\frac{x}{x_{min}} \right) \rangle} \\ \hat{\beta} &= \frac{1}{\langle x \rangle - x_{min}}.\end{aligned}\tag{5.28}$$

where the angled brackets denote the mean. In addition, we also can see that the sufficient condition $\frac{\partial^2 [\ln \mathbb{L}]}{\partial \alpha^2} = -\frac{1}{(\hat{\alpha}-1)^2} < 0$ and $\frac{\partial^2 [\ln \mathbb{L}]}{\partial \beta^2} = -\frac{1}{\hat{\beta}^2} < 0$ are also satisfied by all $\hat{\alpha}$ and $\hat{\beta}$.

5.6.2 Clauset, Shalizi, and Newman (CSN) algorithm

In a recent statement, the American Statistical Association warned the scientific community that the statistical significance value or the p -value “was never intended to be a substitute for scientific reasoning” [206, para. 2], and outline six principles that can prevent its misuse [207]. A *Nature* commentary on this statement also added that “[r]esearchers should describe not only the data analyses that produced statistically significant results, . . . , but all statistical tests and choices made in calculations” [208, para. 3].

We heed the warning in this paper, but argue that when properly computed and interpreted, the p -value is useful in that it provides a quantitative and objective alternative to visual inspection of the fits. The latter is frequently subjective and biased. This utility becomes important when we are comparing fits of two or more data sets to two or more distributions, and have the ambiguity of being able to choose from two or more definitions of goodness of fit. This is why we need to go beyond the goodness of fits, to establish how plausible different distributions are for different data sets.

In 2009, Clauset, Shalizi and Newman (CSN) did precisely this by coming up with a

p -test model that use PL distribution as an illustration. However, this method can be extend to any distribution, and here we note down algorithm for the CSN p -test that involves four major steps:

CSN(i) Parameter Estimation

Given an empirical data with S observations, $\mathbf{Y} = \{y_1, y_2, \dots, y_S\}$ sorted such that $y_i \leq y_{i+1}$, the CSN algorithm (**CSN(ia)**) first constructs the S subsets $\mathbf{X}^{(j)} = \{x_1^{(j)} = y_j, x_2^{(j)} = y_{j+1}, \dots, x_N^{(j)} = y_S\}$. (**CSN(ib)**) For each $\mathbf{X}^{(j)}$, we estimate the parameters using the maximum likelihood estimation method that maximizes the log-likelihood function as described in Eq.(5.28).

CSN(ii) KS Distance

If $\mathbf{X}$ follows probability distribution function f_X with cumulative distribution function F_X , then its probability integral transform $u = F_X(x)$ is a standard uniform distribution function ($U(0, 1)$). For any f_X distributed sample $\mathbf{X} = \{x_1 = x_{min}, x_2, \dots, x_N\}$ with estimated $\hat{\theta}$, we (**CSN(ia)**) first transform the sample to $U^{(s)} = \{u_i^{(s)} = F_X(x_i|\hat{\theta})\}_{i=1}^N$. (**CSN(iib)**) Then we calculate the KS distance [209] (see Section 5.7 for more details)

$$d_{KS} = \forall_{i=1}^N \sup \left(\left| u_i - \frac{i}{N} \right| \right) \quad (5.29)$$

between $U^{(s)}$ and $U(0, 1)$. Here we make use of the fact that the CDF of $U(0, 1)$ is a linear function, $F_U(u) = u$.

CSN(iii) Determining x_{min}

Notice that the maximum likelihood function only able to estimate parameter $\hat{\theta}$ but not x_{min} . To determine x_{min} , (**CSN(iiia)**) we calculate the KS distance for each $\mathbf{X}^{(j)}$ with its corresponding $\hat{\theta}^{(j)}$. (**CSN(iiib)**) The set $\mathbf{X}^{(j)}$ that yields the lowest KS distance ($d_{KS}^{(em)}$) gives us $\hat{x}_{min}^{(em)} = y_j$ and $\hat{\theta}^{(em)} = \hat{\theta}^{(j)}$. The superscript ‘(em)’ indicates a parameter obtained from empirical data.

CSN(iv) Significance Testing

After $\hat{\theta}^{(em)}$ and $\hat{x}_{min}^{(em)}$ have been estimated from $\mathbf{Y} = \{y_1, y_2, \dots, y_S\}$, we test how plausible it is for $\mathbf{X} = \{x_1 = \hat{x}_{min}^{(em)}, x_2, \dots, x_N\} \subset \mathbf{Y}$ to be a sample taken from a f_X distribution. This is done by (**CSN(iva)**) sampling the f_X distribution M times using $\hat{\theta}^{(em)}$ and $\hat{x}_{min}^{(em)}$. (**CSN(ivb)**) For the m^{th} simulated sample we go through **CSN(i)** to **CSN(iii)** to obtain $d_{KS}^{(m)}$. (**CSN(ivc)**) The significance measure

$$p = \frac{1}{M} \sum_{m=1}^M \mathbb{I}_{\{d_{KS}^{(em)} < d_{KS}^{(m)}\}}. \quad (5.30)$$

is the fraction of simulated samples whose fits are poorer than that of the data.

Fig. 5.6.1. p -testing on [left] 2012 to 2014 Taiwan home price per square foot in order of $\times 10^5$ (fitted to EXP), [middle] 2012 Taiwan lower-tail income in million (fitted to EXP), and [right] 2009 to 2016 Strait Time Index normalized return (fitted to PL). The black dots represent empirical data while the blue dashed line represents the fit. All fits are visually good, yet only the p -value (P_{KS} in percentage) for Taiwan home price is appreciable.

We performed p -testing on the Taiwan home price per square foot (fitted to EXP), Taiwan income (fitted to EXP), and the Straits Times Index normalized return (fitted to PL). The fits and p -values are shown in Fig. 5.6.1. All fits are visually good yet only the p -value for the Taiwan housing is appreciable. We realized the reason for this is simple: while the EXP and PL distributions are defined over $(0, \infty)$, when we collect data from the real world we can only obtain a finite number of elements. Moreover, the largest element in the data is finite. However, existing tests for statistical significance generally do not account for the effects produced by having a *finite number of elements* (FNE) and a *finite largest element* (FLE). In the next section we will explain how the parameters and test statistics can be adjusted for FNE and FLE.

5.6.3 CSN with *Finite Sample Corrections*

We will illustrate the effects of FLE using an PL and EXP distribution, the same discussion can be generalized to other distributions. Refer to Eq.(5.28), the maximum likelihood estimator for $\hat{\alpha}$ and $\hat{\beta}$ are inversely proportional to the $\langle \ln(x) \rangle$ and $\langle x \rangle$ respectively. It is important to mention that Eq.(5.28) is derived for the domain of $x \in [0, \infty)$, however in real data the largest element is finite (x_{max}) hence $x \in [0, x_{max}]$. If we use the mean obtained from data ($\langle \ln(x) \rangle_{data}$ and $\langle x \rangle_{data}$), we will obtain the unadjusted estimator (α_{unadj} and β_{unadj}). However due to the FLE, we can only average up till x_{max} . As such, we underestimate the value of $\langle \ln(x) \rangle$ (for PL) and $\langle x \rangle$ (for EXP). This results in an overestimated $\hat{\alpha}$ (for PL) and $\hat{\beta}$ (for EXP). We correct for this by adding back the truncated

parts (from x_{max} to ∞) into $\langle \ln(x) \rangle$ and $\langle x \rangle$,

$$\begin{aligned} \left\langle \ln \left(\frac{x}{x_{min}} \right) \right\rangle_{adj} &= \left\langle \ln \left(\frac{x}{x_{min}} \right) \right\rangle_{data} \int_{x_{min}}^{x_{max}} f_{PL}(x) dx + \int_{x_{max}}^{\infty} \ln \left(\frac{x}{x_{min}} \right) f_{PL}(x) dx; \\ \langle x \rangle_{adj} &= \langle x \rangle_{data} \int_{x_{min}}^{x_{max}} f_{EXP}(x) dx + \int_{x_{max}}^{\infty} x f_{EXP}(x) dx. \end{aligned} \quad (5.31)$$

This yields the adjusted values,

$$\begin{aligned} \left\langle \ln \left(\frac{x}{x_{min}} \right) \right\rangle_{adj} &= \left\langle \ln \left(\frac{x}{x_{min}} \right) \right\rangle_{data} \left[1 - \left(\frac{x_{max}}{x_{min}} \right)^{1-\alpha} \right] + \frac{1}{\alpha-1} \left(\frac{x_{max}}{x_{min}} \right)^{1-\alpha} \left[(\alpha-1) \ln \left(\frac{x_{max}}{x_{min}} \right) + 1 \right] \\ \langle x \rangle_{adj} &= \langle x \rangle_{data} [1 - \exp(-\beta(x_{max} - x_{min}))] + \frac{1}{\beta} (\beta x_{max} + 1) \exp(-\beta(x_{max} - x_{min})), \end{aligned} \quad (5.32)$$

for the PL and EXP distribution respectively. By inserting (5.32) into (5.28), we obtain the nonlinear equations

$$\left\langle \ln \left(\frac{x}{x_{min}} \right) \right\rangle_{data} + \frac{1}{1 - \hat{\alpha}_{adj}} + \left(\frac{x_{max}}{x_{min}} \right)^{1-\hat{\alpha}_{adj}} \left[\ln \left(\frac{x_{max}}{x_{min}} \right) - \left\langle \ln \left(\frac{x}{x_{min}} \right) \right\rangle_{data} + \frac{1}{\hat{\alpha}_{adj} - 1} \right] = 0 \quad (5.33)$$

that can be solved to obtain the adjusted estimated $\hat{\alpha}$ for PL distribution. Similarly for EXP distribution, the adjusted estimated $\hat{\beta}$ can be obtain by solving

$$\left[\hat{\beta}_{adj} (x_{max} - \langle x \rangle_{data}) + 1 \right] \exp \left[-\hat{\beta}_{adj} (x_{max} - x_{min}) \right] + \hat{\beta}_{adj} (\langle x \rangle_{data} - x_{min}) - 1 = 0. \quad (5.34)$$

This is an extension of the series of work [210–212] on improving parameter estimation for PLs given both x_{min} and x_{max} .

In addition to the FLE effect, we also need to consider the bias caused by having a finite number of elements in the sample. This is because the goodness-of-fits measures (e.g KS distance) decreases as the sample size increases. Therefore, in order to have a fair comparison of the goodness of fit for various sample sizes, we need to determine how d_{KS} changes as a function of N . To do this, we simulated 10^6 samples of various sizes N from $U(0, 1)$. For each sample we determined d_{KS} using Eq.(5.29), so that for each N we end up with 10^6 KS distances. In Fig. 5.6.2 we show the KS distances at different deciles,

which exhibits the asymptotic behavior

$$d_{KS}(\wp_{KS}, N) = \frac{\left(\frac{100}{\wp_{KS}} - 1\right)^{-0.176} \exp(-0.274)}{N^{0.492}}, \quad N > 50 \quad (5.35)$$

that we settled for, after experimenting with several functional forms. This result agrees with our expectation that $d_{KS} \rightarrow 0$ as $N \rightarrow \infty$. It also suggests that if we have two samples with sizes N_1 and N_2 from the same distribution, we should compare $N_1^{0.492} d_{KS}^{(1)}$ against $N_2^{0.492} d_{KS}^{(2)}$. Otherwise, if $N_2 > N_1$ then naturally $d_{KS}^{(2)} < d_{KS}^{(1)}$ and we will be lead to the wrong conclusion that the N_2 sample fits the distribution better.

Fig. 5.6.2. Log-log plot of d_{KS} against N for different deciles going from the 10th percentile (blue) to the 90th (red), obtained from 10⁶ simulations.

We follow the steps outlined in the CSN algorithm (Section 5.6.2) to fit the empirical data, but with two important modifications: (I) the parameters (**CSN(ib)**) and goodness of fit (**CSN(iib)**) are adjusted for the finite largest element; and (II) the p -value (**CSN(ivc)**) is adjusted for the finite number of elements effect. Fig. 5.6.3 shows the fits and p -testing results for Taiwan housing price, Taiwan wealth, and Straits Times Index normalized returns. It is reassuring that after modifications the p -values of all distributions increased. In particular, the two distributions (Fig. 5.6.3(b),(c)) that did not meet the $p > 0.1$ criterion (as suggested by [200]) before modification, now have $p > 0.5$. This is in agreement with our visual assessment of the three fits. We also understand now that a large δ (small x_{max}) is the main reason for Taiwan wealth to fail p -testing before adjustment (although the fit is visually good). In general, our correction formulas perform the best when δ is large due to small sample sizes or truncations.

Fig. 5.6.3. p -testing results for (a) 2012 to 2014 Taiwan home price per square foot in order of $\times 10^5$ (fitted to EXP), (b) 2012 Taiwan lower-tail income in millions (fitted to EXP), and (c) 2009 to 2016 Straits Times Index normalized return (fitted to PL) before and after finite-sample adjustments. In this figure, the empirical CDF that is adjusted for FLE is shown as black dots, while the unadjusted and adjusted fits are shown as blue and red dashed line respectively. P_{KS} -values (in percentage) are for unadjusted (blue) and adjusted (red) fits.

5.6.4 Quantifying the Effect of Finite Largest of Elements

In previous section, we have discussion about the effect of FLE that cause us overestimating the parameters. In this section, we want to quantify this effect. Given a distribution simulated from a true parameter θ_T with estimated parameter $\hat{\theta}$, the relative estimation error is defined as

$$\varepsilon_{RE}(\hat{\theta}) = \frac{\sqrt{\langle (\hat{\theta} - \theta_T)^2 \rangle}}{\theta_T} \quad (5.36)$$

5.6.4.1 Relative Estimation Error-Analytical Approximation

In order to standardized the FLE, we define the largest element $X_{max}(\delta)$ as $F^{-1}(1 - \delta)$ for $\delta \in [0, 1]$ (refer to Section 5.6.1 for more details). In particular, for PL and EXP distribution

$$\begin{aligned} X_{max}^{PL} &= x_{min} \delta^{\frac{1}{1-\alpha_T}}; \\ X_{max}^{EXP} &= x_{min} - \frac{\ln(\delta)}{\beta_T}. \end{aligned} \quad (5.37)$$

Due to the presence of X_{max} , the terms $\langle \ln(x) - \ln(x_{min}) \rangle_{unadj}$ (for PL) and $\langle x \rangle_{unadj}$ (for EXP) only consider the domain of $x \in [x_{min}, x_{max}]$, so

$$\begin{aligned} \left\langle \ln \left(\frac{x}{x_{min}} \right) \right\rangle_{unadj} &= \int_{x_{min}}^{X_{max}^{PL}} \ln \left(\frac{x}{x_{min}} \right) f_{PL}(x|\alpha_T) dx; \\ \langle x \rangle_{unadj} &= \int_{x_{min}}^{X_{max}^{EXP}} x f_{EXP}(x|\beta_T) dx. \end{aligned} \quad (5.38)$$

Performing the integral for both equations gives,

$$\begin{aligned} \left\langle \ln \left(\frac{x}{x_{min}} \right) \right\rangle_{unadj} &= \frac{1}{\alpha_T - 1} \left\{ 1 - \left(\frac{X_{max}^{PL}}{x_{min}} \right)^{1-\alpha_T} \left[(\alpha_T - 1) \ln \left(\frac{X_{max}^{PL}}{x_{min}} \right) + 1 \right] \right\} \\ \langle x \rangle_{unadj} &= \frac{1}{\beta_T} \left\{ (\beta_T x_{min} + 1) - (\beta_T X_{max}^{EXP} + 1) \exp(-\beta_T (X_{max}^{EXP} - x_{min})) \right\}. \end{aligned} \quad (5.39)$$

To check the impact this has on the estimated parameters, we substitute Eq.(5.39) into Eq.(5.28) to obtain the unadjusted estimated parameter for PL distribution,

$$\begin{aligned} \hat{\alpha}_{unadj} &= 1 + \frac{\alpha_T - 1}{1 - \left(\frac{X_{max}^{PL}}{x_{min}} \right)^{1-\alpha_T} \left[(\alpha_T - 1) \ln \left(\frac{X_{max}^{PL}}{x_{min}} \right) + 1 \right]} \\ &\approx \alpha_T + (\alpha_T - 1) \left(\frac{X_{max}^{PL}}{x_{min}} \right)^{1-\alpha_T} \left[(\alpha_T - 1) \ln \left(\frac{X_{max}^{PL}}{x_{min}} \right) + 1 \right] \\ &\quad + \mathcal{O} \left\{ \left(\frac{X_{max}^{PL}}{x_{min}} \right)^{2(1-\alpha_T)} \left[(\alpha_T - 1) \ln \left(\frac{X_{max}^{PL}}{x_{min}} \right) + 1 \right]^2 \right\} \end{aligned} \quad (5.40)$$

and for EXP distribution,

$$\begin{aligned} \hat{\beta}_{unadj} &= \frac{\beta_T}{1 - (\beta_T X_{max}^{EXP} + 1) \exp(-\beta_T (X_{max}^{EXP} - x_{min}))} \\ &\approx \beta_T + \beta_T [\beta_T X_{max}^{EXP} + 1] \exp(-\beta_T (X_{max}^{EXP} - x_{min})) \\ &\quad + \mathcal{O} \left\{ [\beta_T X_{max}^{EXP} + 1]^2 \exp(-2\beta_T (X_{max}^{EXP} - x_{min})) \right\}. \end{aligned} \quad (5.41)$$

The approximations in Eq.(5.40) and Eq.(5.41) are estimated using first order Taylor expansion. Rewriting $\hat{\alpha}_{unadj}$ and $\hat{\beta}_{unadj}$ in terms of δ from Eq.(5.37) will lead to

$$\begin{aligned} \hat{\alpha}_{unadj} &\approx \alpha_T + (\alpha_T - 1) \delta [1 - \ln(\delta)] \\ \hat{\beta}_{unadj} &\approx \beta_T + \beta_T \delta [1 - \ln(\delta) + \beta_T x_{min}]. \end{aligned} \quad (5.42)$$

Hence, the relative estimation error for PL and EXP distribution in term of δ is

$$\begin{aligned}\varepsilon_{RE}^{PL}(\hat{\alpha}_{unadj}) &= \left(1 - \frac{1}{\alpha_T}\right) \delta [1 - \ln(\delta)], \quad \delta \in [0, 1] \\ \varepsilon_{RE}^{EXP}(\hat{\beta}_{unadj}) &= \delta [1 - \ln(\delta) + \beta_T x_{min}], \quad \delta \in [0, 1].\end{aligned}\tag{5.43}$$

While it is useful to note that both relative estimation errors are monotonic increasing function with $\varepsilon_{RE}^{PL}(\hat{\alpha}_{unadj}) \in [0, 1]$ and $\varepsilon_{RE}^{EXP}(\hat{\beta}_{unadj}) \in [0, \beta_T x_{min}]$, it is more important to observe that for a given δ , the $\varepsilon_{RE}^{PL}(\hat{\alpha}_{unadj})$ is always smaller or equal to $\varepsilon_{RE}^{EXP}(\hat{\beta}_{unadj})$ (refer to Fig. 5.6.4). This sensitivity of the estimated parameter (relative estimation error) on finite largest element ($x_{max}(\delta)$) provides us an insight on answering why power law distribution pass the significance testing more often compared to exponential distribution on unadjusted parameters.

Fig. 5.6.4. Approximated Relative estimation errors of (a) $\varepsilon_{RE}^{PL}(\hat{\alpha}_{unadj})$ and (b) $\varepsilon_{RE}^{EXP}(\hat{\beta}_{unadj})$ base on Eq.(5.43) for various parameters.

5.6.4.2 Relative Estimation Error-Simulation

To examine the relative estimation errors dependent on $x_{max}(\delta)$, and the performance of the adjustment formulas Eq.(5.33) and (5.34). (I) We first simulate M sample distributions using the inverse cumulative transform with a given θ_T , $F^{-1}(u|\theta_T)$ where $u \sim U(0, 1)$. As the $F^{-1}(u|\theta_T)$ map $u = 0$ to $x = x_{min}$ and $u = 1$ to $x = \infty$, hence in order to simulate $X_{max}(\delta)$ we use $u \sim U(0, 1 - \delta)$ instead $U(0, 1)$. (II) Then for each simulated sample distribution, we estimate the parameters unadjusted and adjusted for FLE. (III) Finally, we can determine the relative estimation error using these M sets for unadjusted and adjusted parameters.

For illustration, we choose to use EXP distribution as it is more sensitive towards $X_{max}(\delta)$. We simulated 1000 sets of EXP distributed data of $X_{max}(0.1)$, for $10^{-4} \leq \beta_T \leq 10^2$, with sample size $N \in [10^2, 10^4]$. Thereafter, we determine the relative estimation error $\varepsilon_{RE}^{EXP}(\hat{\beta}_{unadj})$ and $\varepsilon_{RE}^{EXP}(\hat{\beta}_{adj})$, which is shown in Fig. 5.6.5. The $\varepsilon_{RE}^{EXP}(\hat{\beta}_{unadj})$ is about 38% for small samples $N \sim 10^2$ and decreases to 34% for large samples $N \sim 10^4$. On

Fig. 5.6.5. Relative estimation errors of (a) $\hat{\beta}_{unadj}$ and (b) $\hat{\beta}_{adj}$ measured from 1000 simulated samples using different β_T and N with $x_{min} = 0$ and $x_{max} = F_{EXP}^{-1}(0.9)$. Due to the FLE, $\Delta\beta_{unadj}$ remains high (close to the theoretical relative estimation error of $\epsilon(\delta = 0.1, x_{min} = 0) = 0.1[1 - \ln(0.1)] \approx 33\%$) even for large N . In contrast, $\Delta\beta_{adj}$ decreases rapidly with increasing N .

the other hand, $\epsilon_{RE}^{EXP}(\hat{\beta}_{adj})$ starts at 20%, but decreases to 2% as the number of data points is increased. Although it can be shown that the $\hat{\beta}_{unadj}$ is an unbiased estimator [200, 213], we find it converging very slowly with increasing sample size in the unfortunate situation of a small x_{max} . In contrast, $\hat{\beta}_{adj}$ converges very quickly even for small x_{max} as we have accounted for the FLE.

Fig. 5.6.6. The median KS distances for (a) KS_{unadj} and (b) KS_{adj} measured from 1000 simulated samples using different β_T and N . The x_{min} is set to 0 and $x_{max} = F_{EXP}^{-1}(0.9)$. Because of the FLE, KS_{unadj} remains above $\delta = 0.10$ while KS_{adj} converges to zero for large sample size.

For a finite sample, $F_{EXP}(x) < 1$ for all $x < \infty$. Mathematically, this means that $F_{EXP}(x) \sim U(0, 1 - \delta)$, where $F_{EXP}^{-1}(x_{max}) = 1 - \delta$. This observation is important,

because d_{KS} is obtained by comparing $U^{(s)} = \{u_i^{(s)} = F_{EXP}(x_i|\hat{\beta})\}_{i=1}^N$ against $U(0, 1)$ (see Eq.(5.29)). This tell us that for a fair comparison, we need to rescale all elements in $U^{(s)}$ by a factor of $1/(1 - \delta)$. Fig. 5.6.6 shows the d_{KS} measured for the 1000 sets of EXP distributed data with finite largest element $x_{max} = F^{-1}(0.9)$ for various β_T and sample sizes N . For each sample, we use first to estimate the $\hat{\beta}_{adj}$ and then inverse cumulative transformed this data to $U^{(s)}$. After that, we measure d_{KS} with Eq.(5.29) to obtain non-rescaled KS distance, KS_{unadj} and rescaled KS distance, KS_{adj} using the non-rescaled and rescaled $U^{(s)}$ respectively. KS_{unadj} goes from 0.14 for small samples $N \sim 10^2$, to 0.10 for large samples $N \sim 10^5$. In contrast, KS_{adj} decrease from 0.06 for small samples to 0.006 for large samples.

GitHub: Statistical Significance Testing with Finite Sample Corrections

The codes are available at

<https://github.com/BoonKinTeh/StatisticalSignificanceTesting>, and please refer to Teh *et al.* “Finite Sample Corrections for Parameters Estimation and Significance Testing” for more details, and cite the paper if you are using the codes.

5.7 Goodness-of-fit Measures

In statistical significance testing, to test how plausible a sample distribution can be explained by the model (hypothetical) distribution, we cannot just stop at computing the likelihood or coefficient of variation, but also the goodness-of-fits. This is because in most cases, the likelihood and the coefficient of variation are distribution dependent, hence should not be used for comparing across different distributions. As a result, we need to have a goodness-of-fit measure that is independent of the distribution and is non-parametric for statistical significance testing.

In Section 5.6.1, we mentioned that if $\mathbf{X}$ is f_x distributed with cumulative distribution function F_x , then $F_x(\mathbf{X})$ follows a standard uniform distribution, $U(0, 1)$. This property is very useful as a goodness-of-fit measure, as it converts $\mathbf{X}$ (if F_x and F_x^{-1} exist) to $U(0, 1)$, if $\mathbf{X}$ is truly f_x distributed with cumulative distribution function F_x . With this information, given a sample $\mathbf{X}^{(e)} = \{x_1^{(e)}, x_2^{(e)}, \dots, x_N^{(e)}\}$ with hypothetical distribution function $f^{(e)}$ and cumulative distribution function $F^{(e)}$, to test the plausibility that $\mathbf{X}^{(e)}$ is withdrawn from $f^{(e)}$ can be done by testing the plausibility of $F^{(e)}(\mathbf{X}^{(e)}) = \{F^{(e)}(x_1^{(e)}), F^{(e)}(x_2^{(e)}), \dots, F^{(e)}(x_N^{(e)})\}$ is withdrawn from $U(0, 1)$. By doing this, now the plausibility test is independent on distribution, and we just need to compute the goodness-of-fits between $F^{(e)}(\mathbf{X}^{(e)})$ and $U(0, 1)$.

5.7.1 Kolmogorov-Smirnov (KS) Distance

One of the most common goodness-of-fits measures is the *Kolmogorov-Smirnov* (KS) distance [209], which is a non-parametric test of the equality between two continuous distributions. Given two distributions f_1 and f_2 , to compute the KS distance, we first determine their cumulative distributions F_1 and F_2 , and then compute the KS distance as the maximum distance between F_1 and F_2 . Fig. 5.7.1 shows the KS distance computed for (a) different distributions, (b) the same distribution function but with different parameters, (c) the same distribution function with the same parameters but small sample sizes, and (d) the same distribution function with the same parameters with large sample size. Not surprisingly, we see the KS distances are decreasing from (a) to (d) (goodness-of-fits increase from (a) to (d)), as the two distributions are closer to each other from (a) to (d). In addition, notice that the KS distance is size dependent, due to the law of large numbers, the empirical distribution approaches its true distribution as the sample size is increased.

For statistical significance testing purposes, we are comparing $F^{(e)}(\mathbf{X}^{(e)})$ and $U(0, 1)$. Making use of the properties of $F(x) = x$ for $x \sim U(0, 1)$, and assuming all data points carry equal weight then the KS distance is

$$d_{KS} = \forall_{i=1}^N \sup \left(\left| u_i - \frac{i}{N} \right| \right), \quad (5.44)$$

where N is the sample size, u_i is the ordered data point for $F^{(e)}(x_i^{(e)})$ which u_1 and u_N being the smallest and largest respectively.

Fig. 5.7.1. Illustration of the Kolmogorov-Smirnov distance between two distributions: (a) power-law and exponential distributions for $N = 1000$; (b) exponential distributions with $\beta = 2$ and $\beta = 3$ for $N = 1000$; (c) two $N = 100$ samples from an exponential distribution with $\beta = 2$; (d) two $N = 10,000$ samples from an exponential distribution with $\beta = 2$.

5.7.2 *Distribution Noise*

Besides having to work with finite samples and finite largest elements, we will also in practice encounter imperfections while collecting samples for various reasons, such as undetected samples, contamination by background noise, and recording errors. We call such noises that occur at the element level *elementary noise*. When we convert these samples to a distribution, noise will also be present at the distribution level that we refer to as *distribution noise*. In principle the information at the distribution level is more robust compared to the elementary level, as we expect random and thus uncorrelated noise to cancel each other. This means that the distribution is less sensitive to elementary noise, but we still worry whether the distribution noise may play an important role in our test of statistical significance. In order to account for the effects of distribution noise, we need to first be able to quantify distribution noise, and thereafter understand how it affects significance testing.

Back to the statistical significance testing, after integral probability transform and reordered $F^{(e)}(\mathbf{X}^{(e)}) = \{u_1 = F^{(e)}(x_{min}), u_2 = F^{(e)}(x_{(2)}), \dots, u_N = F^{(e)}(x_{max})\}$, a natural way to measure the distribution noise is to plot the histogram, count the frequency for each

bin, and compare it to the expected frequency from $U(0, 1)$ see illustration in Fig. 5.7.2(a)-(b). Since this can be more accurately done for smaller bin sizes, we use the intervals between sorted elements as a collection of non-uniform bins, as shown in Fig. 5.7.2(c)-(d). For a data set consisting of N elements, each bin carry a weight of $1/N$, evenly distributed within the interval $(u_{i-1}, u_i]$, such that the probability density is

$$f(u_{i-1}, u_i) = \frac{\frac{1}{N}}{u_i - u_{i-1}}. \quad (5.45)$$

As the theoretical probability density for $U(0, 1)$ is 1, we define the distribution noise d_{DN} mathematically to be

$$\begin{aligned} d_{DN} &= \sqrt{\frac{\sum_{i=1}^N (u_i - u_{i-1})^2 [f(u_{i-1}, u_i) - 1]^2}{\sum_{i=1}^N (u_i - u_{i-1})^2}} \\ &= \sqrt{\frac{\sum_{i=1}^N (u_i - u_{i-1})^2 \left(\frac{1}{N(u_i - u_{i-1})} - 1 \right)^2}{\sum_{i=1}^N (u_i - u_{i-1})^2}}, \end{aligned} \quad (5.46)$$

where $u_0 = 0$ and $u_N = 1$. We need to weigh the deviation of each bin by $(u_i - u_{i-1})^2$ because the bins are non-uniform. and to keep d_{DN} finite.

Fig. 5.7.2. Illustration of the distribution noise we would measure if we sample 10 elements from $U(0, 1)$, rescaled such that the largest element becomes 1. In (a)-(b) we use 5 uniform bins whereas in (c)-(d) we use the intervals between sorted elements as the bins. The counts are shown on the left, whereas the probability densities calculated using Eq.(5.45) are shown on the right.

For an ordered standard uniform distributed sample $(\{u_1, u_2, \dots, u_N\}$ with $u_i \leq u_j$ for $i < j$), we define $\Delta = \{\Delta_1 = u_1, \Delta_2 = u_2 - u_1, \Delta_3 = u_3 - u_2, \dots, \Delta_N = u_N - u_{N-1}\}$. The Δ is the spacing of ordered uniform distribution, which is $f_{\Delta}(\Delta) = N(1 - \Delta)^{N-1}$

distributed [214]. Rewriting Eq.(5.46) in term of Δ yield

$$d_{DN} = \sqrt{\frac{\sum_{i=1}^N \left(\Delta_i \left(\frac{1}{N\Delta_i} - 1 \right) \right)^2}{\sum_{i=1}^N \Delta_i^2}}, \quad (5.47)$$

in the limit where N is large ($N > 50$), we can apply the continuous approximation,

$$\begin{aligned} d_{DN}^{(*)} &= \sqrt{\frac{\int_0^1 \left(\Delta \left(\frac{1}{N\Delta} - 1 \right) \right)^2 N(1-\Delta)^{N-1} d\Delta}{\int_0^1 \Delta^2 N(1-\Delta)^{N-1} d\Delta}} \\ &= \sqrt{\frac{1}{2} + \frac{2-N}{2N^2}}. \end{aligned} \quad (5.48)$$

where the $(*)$ indicates the analytically derived distribution noise, we have performed the integration in the surd, using the properties of the Beta distribution.

Fig. 5.7.3. Relationship between distribution noise d_{DN} and sample size N at deciles going from the 10th percentile (blue) to the 90th (red), obtained from 10⁶ simulations. The d_{DN} value converges to $1/\sqrt{2}$ as N increases.

We simulated 10⁶ samples from $U(0, 1)$ with different sample size N . For each sample, we calculate the distribution noise d_{DN} using Eq.(5.46) and plot its deciles against N as shown in Fig. 5.7.3. After experimenting with several functional forms, we write down

the relationship between d_{DN} and N at percentile $\wp_{DN}$ as

$$d_{DN}(\wp_{DN}, N) = \langle d_{DN} \rangle + \Phi(\wp_{DN} - 50) \frac{\exp\left(-\frac{[50 - |\wp_{DN} - 50|]^{0.430}}{|\wp_{DN} - 50|^{0.302}}\right)}{N^{0.495}}, \quad (5.49)$$

where $\Phi(x)$ represents the sign of x , and

$$\langle d_{DN} \rangle = \sqrt{\frac{1}{2} + \frac{2 - N}{2N^2}} \left(\frac{N}{N + 0.5} \right) \quad (5.50)$$

is the mean of d_{DN} that converges to $1/\sqrt{2}$ as $N \rightarrow \infty$, and $N/(N + 0.5)$ is added in as a correction factor for small N . This result suggests that if we have two samples with sizes N_1 and N_2 with $N_2 > N_1$ from the same distribution, we should compare $N_1^{0.495}(d_{DN}^{(1)} - 1/\sqrt{2})$ against $N_2^{0.495}(d_{DN}^{(2)} - 1/\sqrt{2})$. Otherwise, we risk making wrong conclusion that the N_2 sample fits the distribution better if $d_{DN}^{(1)} > d_{DN}^{(2)}$.

5.7.3 Relationship between Distribution Noise and KS Distance

As measures for statistical deviations, d_{DN} and d_{KS} are different in that d_{DN} measures deviation at the probability density level, whereas the d_{KS} measure it at the cumulative density level. As a result, d_{KS} assigns more weight to the tail of the distribution, while d_{DN} is more sensitive to deviations in the body of the distribution. Therefore, if we wish to combine these two measures to estimate the significance level, we need to first investigate the relationship between d_{KS} and d_{DN} .

Fig. 5.7.4. Correlation between d_{KS} and d_{DN} at different sample sizes, obtained from the 10^6 simulated samples.

We do this by simulating 10^6 samples from $U(0, 1)$ for various sample sizes, and for each sample, we calculate d_{KS} and d_{DN} using Eq.(5.29) and Eq.(5.46) respectively, to obtain 10^6 pairs of d_{KS} and d_{DN} . We then compute the Pearson correlation between d_{KS} and d_{DN} (refer to Fig. 5.7.4 for the fits fits) and learned that

$$\rho_{d_{KS}, d_{DN}}(N) = \frac{e}{N^{0.481}}. \quad (5.51)$$

As expected, d_{KS} is positively correlated with d_{DN} . Since d_{KS} is a measure at the cumulative level, the random distribution noises cancel each other, thus the correlation between d_{KS} and d_{DN} vanishes as $N \rightarrow \infty$.

References

- [1] Jan Tinbergen. Minimumproblemen in de natuurkunde en de ekonomie. 1929.
- [2] Manuel Buitenhuis. Jan Tinbergen’s Move from Physics to Economics, 1922-1930. Master’s thesis, 2016.
- [3] Jan Tinbergen. Shaping the world economy; suggestions for an international economic policy. *Books (Jan Tinbergen)*, 1962.
- [4] James E Anderson. A theoretical foundation for the gravity equation. *The American Economic Review*, 69(1):106–116, 1979.
- [5] Jonathan Eaton and Samuel Kortum. Technology, geography, and trade. *Econometrica*, 70(5):1741–1779, 2002.
- [6] James E Anderson and Eric Van Wincoop. Gravity with gravitas: a solution to the border puzzle. *the American Economic Review*, 93(1):170–192, 2003.
- [7] Lorenzo Caliendo and Fernando Parro. Estimates of the Trade and Welfare Effects of NAFTA. *The Review of Economic Studies*, 82(1):1–44, 2015.
- [8] Mario Larch and Yoto Yotov. General equilibrium trade policy analysis with structural gravity. *WTO Working Paper ERSD-2016-08*, 2016.
- [9] Rosario Nunzio Mantegna. Lévy walks and enhanced diffusion in Milan stock exchange. *Physica A: Statistical Mechanics and its Applications*, 179(2):232–242, 1991.
- [10] Hideki Takayasu, Hitoshi Miura, Tadashi Hirabayashi, and Koichi Hamada. Statistical properties of deterministic threshold elements: the case of market price. *Physica A: Statistical Mechanics and its Applications*, 184(1):127–134, 1992.
- [11] Kausik Gangopadhyay. Interview with Eugene H. Stanley. *IIM Kozhikode Society & Management Review*, 2(2):73–78, 2013.
- [12] Bikas K Chakrabarti. Econophys-Kolkata: a short story. *Econophysics of Wealth Distributions*, pages 225–228, 2005.
- [13] Michael HR Stanley, Sergey V Buldyrev, Shlomo Havlin, Rosario N Mantegna, Michael A Salinger, and H Eugene Stanley. Zipf plots and the size distribution of firms. *Economics letters*, 49(4):453–457, 1995.
- [14] Michael HR Stanley, LuYs AN Amaral, Sergey V Buldyrev, Shlomo Havlin, Heiko Leschhorn, Philipp Maass, Michael A Salinger, and H Eugene Stanley. Scaling behaviour in the growth of companies. *Nature*, 379(6568):804–806, 1996.
- [15] Hideki Takayasu, Aki-Hiro Sato, and Misako Takayasu. Stable infinite variance fluctuations in randomly amplified Langevin systems. *Physical Review Letters*, 79(6):966, 1997.
- [16] Kazumi Okuyama, Misako Takayasu, and Hideki Takayasu. Zipf’s law in income distribution of companies. *Physica A: Statistical Mechanics and its Applications*, 269(1):125–131, 1999.

- [17] Hideaki Aoyama, Wataru Souma, Yuichi Nagahara, Mitsuhiro P Okazaki, Hideki Takayasu, and Misako Takayasu. Pareto's law for income of individuals and debt of bankrupt companies. *Fractals*, 8(03):293–300, 2000.
- [18] Yoshi Fujiwara, Wataru Souma, Hideaki Aoyama, Taisei Kaizoji, and Masanao Aoki. Growth and fluctuations of personal income. *Physica A: Statistical Mechanics and its Applications*, 321(3):598–604, 2003.
- [19] Victor M Yakovenko and A Christian Silva. Two-class structure of income distribution in the USA: Exponential bulk and power-law tail. *Econophysics of Wealth Distributions*, pages 15–23, 2005.
- [20] Anirban Chakraborti and Bikas K Chakrabarti. Statistical mechanics of money: how saving propensity affects its distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 17(1):167–170, 2000.
- [21] A Drăgulescu and Victor M Yakovenko. Evidence for the exponential distribution of income in the USA. *The European Physical Journal B-Condensed Matter and Complex Systems*, 20(4):585–589, 2001.
- [22] Bikas K Chakrabarti and Arnab Chatterjee. Ideal gas-like distributions in economics: effects of saving propensity. In *Application of Econophysics, Proc. 2nd Nikkei Econophys. Symp., Ed. Takayasu H, Springer, Tokyo*, pages 280–285. Springer, 2004.
- [23] Hajime Inaoka, Hideki Takayasu, Tokiko Shimizu, Takuto Ninomiya, and Ken Taniguchi. Self-similarity of banking network. *Physica A: Statistical Mechanics and its Applications*, 339(3):621–634, 2004.
- [24] Hans Degryse, Gregory Nguyen, et al. Interbank exposures: An empirical examination of contagion risk in the Belgian banking system. *International Journal of Central Banking*, 3(2):123–171, 2007.
- [25] Andrew G Haldane and Robert M May. Systemic risk in banking ecosystems. *Nature*, 469(7330):351, 2011.
- [26] Stefan Thurner and Sebastian Poledna. DebtRank-transparency: Controlling systemic risk in financial networks. *Scientific reports*, 3, 2013.
- [27] Matthieu Cristelli, Andrea Gabrielli, Andrea Tacchella, Guido Caldarelli, and Luciano Pietronero. Measuring the intangibles: A metrics for the economic complexity of countries and products. *PloS one*, 8(8):e70726, 2013.
- [28] Matthieu Cristelli, Andrea Tacchella, and Luciano Pietronero. The heterogeneous dynamics of economic complexity. *PloS one*, 10(2):e0117174, 2015.
- [29] Rosario N Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11(1):193–197, 1999.
- [30] J-P Onnela, A Chakraborti, K Kaski, and J Kertész. Dynamic asset trees and portfolio analysis. *The European Physical Journal B-Condensed Matter and Complex Systems*, 30(3):285–288, 2002.
- [31] J-P Onnela, Anirban Chakraborti, Kimmo Kaski, and Janos Kertesz. Dynamic asset trees and Black Monday. *Physica A: Statistical Mechanics and its Applications*, 324(1):247–252, 2003.
- [32] Jukka-Pekka Onnela, Anirban Chakraborti, Kimmo Kaski, Janos Kertesz, and Antti Kanto. Asset trees and asset graphs in financial markets. *Physica Scripta*, 2003(T106):48, 2003.
- [33] Tiziana Di Matteo, Tomaso Aste, ST Hyde, and Stuart Ramsden. Interest rates hierarchical structure. *Physica A: Statistical Mechanics and its Applications*, 355(1):21–33, 2005.

- [34] Michel Tumminello, Tomaso Aste, T Di Matteo, and RN Mantegna. A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30):10421–10426, 2005.
- [35] Tiziana Di Matteo, Francesca Pozzi, and Tomaso Aste. The use of dynamical networks to detect the hierarchical organization of financial market sectors. *The European Physical Journal B-Condensed Matter and Complex Systems*, 73(1):3–11, 2010.
- [36] Won-Min Song, T Di Matteo, and Tomaso Aste. Hierarchical information clustering by means of topologically embedded graphs. *PloS one*, 7(3):e31929, 2012.
- [37] F Pozzi, T Di Matteo, and T Aste. Spread of risk across financial markets: better to invest in the peripheries. *Scientific reports*, 3, 2013.
- [38] Nicolo Musmeci, Tomaso Aste, and Tiziana Di Matteo. Relation between financial market structure and the real economy: Comparison between clustering methods. *PloS one*, 10(3):e0116201, 2015.
- [39] Daniel J Fenn, Mason A Porter, Peter J Mucha, Mark McDonald, Stacy Williams, Neil F Johnson, and Nick S Jones. Dynamical clustering of exchange rates. *QUANTITATIVE FINANCE*, 12(10):1493–1520, 2012.
- [40] David J Ketchen Jr and Christopher L Shook. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal*, 17(6):441–458, 1996.
- [41] Sara Dolnicar. Using cluster analysis for market segmentation-typical misconceptions, established methodological weaknesses and some recommendations for improvement. *Australasian Journal of Market Research*, 11(2):5–12, 2003.
- [42] Girish Punj and David W Stewart. Cluster analysis in marketing research: Review and suggestions for application. *Journal of marketing research*, 20(2):134–148, 1983.
- [43] Glenn W Milligan and Martha C Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.
- [44] SC Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.
- [45] Peter HA Sneath, Robert R Sokal, et al. *Numerical taxonomy. The principles and practice of numerical classification*. 1973.
- [46] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.
- [47] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.
- [48] Juan Camilo Bohorquez, Sean Gourley, Alexander R Dixon, Michael Spagat, and Neil F Johnson. Common ecology quantifies human insurgency. *Nature*, 462(7275):911–914, 2009.
- [49] M Fratini, N Poccia, A Ricci, G Campi, M Burghammer, G Aeppli, and A Bianconi. Scale-free structural organization of oxygen interstitials in $\text{La}_2\text{CuO}_{4+y}$. *Nature*, 466(7308):841, 2010.
- [50] Zhenyuan Zhao, JP Calderón, Chen Xu, Guannan Zhao, Dan Fenn, Didier Sornette, Riley Crane, Pak Ming Hui, and Neil F Johnson. Effect of social group dynamics on contagion. *Physical Review E*, 81(5):056107, 2010.
- [51] Neil Johnson, Spencer Carran, Joel Botner, Kyle Fontaine, Nathan Laxague, Philip Nuetzel, Jessica Turnley, and Brian Tivnan. Pattern in escalations in insurgent and terrorist activity. *Science*, 333(6038):81–84, 2011.
- [52] Neil F Johnson, Josef Ashkenazi, Zhenyuan Zhao, and Luis Quiroga. Equivalent dynamical complexity in a many-body quantum and collective human system. *AIP Advances*, 1(1):012114, 2011.

- [53] Natasha Gilbert. Modellers claim wars are predictable: insurgent attacks follow a universal pattern of timing and casualties. *Nature*, 462(7275):836–837, 2009.
- [54] Siew Ann Cheong, Teck Liang Tan, Chien-Chih Chen, Wu-Lung Chang, Zheng Liu, Lock Yue Chew, Peter MA Sloot, and Neil F Johnson. Short-term forecasting of Taiwanese earthquakes using a universal model of fusion-fission processes. *Scientific Reports*, 4, 2014.
- [55] Zhuanxin Ding, Clive WJ Granger, and Robert F Engle. A long memory property of stock market returns and a new model. *Journal of empirical finance*, 1(1):83–106, 1993.
- [56] Mico Loretan and Peter CB Phillips. Testing the covariance stationarity of heavy-tailed time series: An overview of the theory with applications to several financial datasets. *Journal of empirical finance*, 1(2):211–248, 1994.
- [57] Arnoud WA Boot and Anjan V Thakor. Financial system architecture. *Review of Financial Studies*, 10(3):693–733, 1997.
- [58] Robert J Shiller. Human behavior and the efficiency of the financial system. *Handbook of macroeconomics*, 1:1305–1340, 1999.
- [59] Asli Demirgüç-Kunt and Ross Levine. Stock market development and financial intermediaries: stylized facts. *The World Bank Economic Review*, 10(2):291–321, 1996.
- [60] Timothy J Brailsford. The empirical relationship between trading volume, returns and volatility. *Accounting & Finance*, 36(1):89–111, 1996.
- [61] Elena Andreou, Nikitas Pittis, and Aris Spanos. On modelling speculative prices: the empirical literature. *Journal of economic surveys*, 15(2):187–220, 2001.
- [62] Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.
- [63] Damien Challet, Matteo Marsili, and Yi-Cheng Zhang. Stylized facts of financial markets and market crashes in minority games. *Physica A: Statistical Mechanics and its Applications*, 294(3):514–524, 2001.
- [64] Gong-meng Chen, Michael Firth, and Oliver M Rui. The dynamic relation between stock returns, trading volume, and volatility. *Financial Review*, 36(3):153–174, 2001.
- [65] Martin Sewell. Characterization of financial time series. *PResearch Note*, 11(1):1–35, 2011.
- [66] Laurent Laloux, Pierre Cizeau, Jean-Philippe Bouchaud, and Marc Potters. Noise dressing of financial correlation matrices. *Physical review letters*, 83(7):1467, 1999.
- [67] John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A:32–52, 1928.
- [68] Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.
- [69] Madan Lal Mehta and Michel Gaudin. On the density of eigenvalues of a random matrix. *Nuclear Physics*, 18:420–427, 1960.
- [70] Eugene P Wigner. Random matrices in physics. *SIAM Review*, 9(1):1–23, 1967.
- [71] Laurent Laloux, Pierre Cizeau, Marc Potters, and Jean-Philippe Bouchaud. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 3(03):391–397, 2000.
- [72] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luis A Nunes Amaral, Thomas Guhr, and H Eugene Stanley. Random matrix approach to cross correlations in financial data. *Physical Review E*, 65(6):066126, 2002.
- [73] Didier Sornette and Anders Johansen. A hierarchical model of financial crashes. *Physica A: Statistical Mechanics and its Applications*, 261(3):581–598, 1998.

- [74] RN Mantegna. Information and hierarchical structure in financial markets. *Computer Physics Communications*, 121(6535):153–156, 1999.
- [75] Didier Sornette, Prospero Simonetti, and Joergen Vitting Andersen. ϕ q field theory for portfolio optimization: fat tails and nonlinear correlations. *Physics Reports*, 335(2):19–92, 2000.
- [76] Giovanni Bonanno, Fabrizio Lillo, and Rosario N Mantegna. High-frequency cross-correlation in a set of stocks. *Quantitative Finance*, 1:96–104, 2001.
- [77] Edwin J Elton, Martin J Gruber, Stephen J Brown, and William N Goetzmann. *Modern Portfolio Theory and Investment Analysis*. John Wiley & Sons, 2009.
- [78] Michele Tumminello, Fabrizio Lillo, and Rosario N Mantegna. Correlation, hierarchies, and networks in financial markets. *Journal of Economic Behavior & Organization*, 75(1):40–58, 2010.
- [79] Khaldoun Khashanah and Linyan Miao. Dynamic structure of the US financial systems. *Studies in Economics and Finance*, 28(4):321–339, 2011.
- [80] Giovanni Bonanno, Guido Caldarelli, Fabrizio Lillo, and Rosario N Mantegna. Topology of correlation-based minimal spanning trees in real and model markets. *Physical Review E*, 68(4):046130, 2003.
- [81] Giovanni Bonanno, Guido Caldarelli, Fabrizio Lillo, Salvatore Micciche, Nicolas Vandewalle, and Rosario Nunzio Mantegna. Networks of equities in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):363–371, 2004.
- [82] Boon Kin Teh, Yik Wen Goo, Tong Wei Lian, Wei Guang Ong, Wen Ting Choi, Mridula Damodaran, and Siew Ann Cheong. The Chinese Correction of February 2007: How financial hierarchies change in a market crash. *Physica A: Statistical Mechanics and its Applications*, 424:225–241, 2015.
- [83] Michael J Naylor, Lawrence C Rose, and Brendan J Moyle. Topology of foreign exchange markets using hierarchical structure methods. *Physica A: Statistical Mechanics and its Applications*, 382(1):199–208, 2007.
- [84] Yiting Zhang, Gladys Hui Ting Lee, Jian Cheng Wong, Jun Liang Kok, Manamohan Prusty, and Siew Ann Cheong. Will the US economy recover in 2010? A minimal spanning tree study. *Physica A: Statistical Mechanics and its Applications*, 390(11):2020–2050, 2011.
- [85] Mustafa Keskin, Bayram Deviren, and Yusuf Kocakaplan. Topology of the correlation networks among major currencies using hierarchical structure methods. *Physica A: Statistical Mechanics and its Applications*, 390(4):719–730, 2011.
- [86] Ersin Kantar, Mustafa Keskin, and Bayram Deviren. Analysis of the effects of the global financial crisis on the Turkish economy, using hierarchical methods. *Physica A: Statistical Mechanics and its Applications*, 391(7):2342–2352, 2012.
- [87] Siew Ann Cheong, Robert Paulo Fornia, Gladys Hui Ting Lee, Jun Liang Kok, Woei Shyr Yim, Danny Yuan Xu, Yiting Zhang, et al. The Japanese economy in crises: A time series segmentation study. *Economics: The Open-Access, Open-Assessment E-Journal*, 6(2012-5):1–81, 2012.
- [88] Nicolás Musmeci, Tomaso Aste, and T Di Matteo. Risk diversification: a study of persistence with a filtered correlation-network approach. *Network Theory in Finance*, 1(1):77–98, 2015.
- [89] Didier Sornette. *Why stock markets crash: critical events in complex financial systems*. Princeton University Press, 2009.

- [90] Ip-Wing Yu, Kang-Por Fung, and Chi-Sang Tam. Assessing financial market integration in Asia–equity markets. *Journal of Banking & Finance*, 34(12):2874–2885, 2010.
- [91] Feng Guo, Carl R Chen, and Ying Sophie Huang. Markets contagion during financial crisis: A regime-switching approach. *International Review of Economics & Finance*, 20(1):95–109, 2011.
- [92] Leonidas Sandoval Junior and Italo De Paula Franca. Correlation of financial markets in times of crisis. *Physica A: Statistical Mechanics and its Applications*, 391(1):187–208, 2012.
- [93] Zeyu Zheng, Boris Podobnik, Ling Feng, and Baowen Li. Changes in Cross-Correlations as an Indicator for Systemic Risk. *Scientific reports*, 2, 2012.
- [94] Rick Quax, Drona Kandhai, and Peter MA Sloot. Information dissipation as an early-warning signal for the Lehman Brothers collapse in financial time series. *Scientific reports*, 3, 2013.
- [95] Dimitris Kenourgios, Aristeidis Samitas, and Nikos Paltalidis. Financial crises and stock market contagion in a multivariate time-varying asymmetric framework. *Journal of International Financial Markets, Institutions and Money*, 21(1):92–106, 2011.
- [96] Bill Powell. Behind China’s Stock Meltdown. *TIME Business and Money*, 2007.
- [97] Charles Hugh Smith. China Is Not Another Ascendant Superpower, It’s Just Another Nation with Structural Problems. *Business Insider, Finance*, 2009.
- [98] Mike Shedlock. China Facing Economic, Financial and Stock Market Crash Scenario. *The Market Oracle*, 2009.
- [99] HKSx. HKEx Fact Book. 2008.
- [100] Karl Pearson. Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, 60:489–498, 1896.
- [101] Stephen M Stigler. Francis Galton’s account of the invention of correlation. *Statistical Science*, 4(2):73–79, 1989.
- [102] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.
- [103] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.
- [104] James E Gentle. *Numerical linear algebra for applications in statistics*. Springer Science & Business Media, 2012.
- [105] Dong-Hee Kim and Hawoong Jeong. Systematic analysis of group identification in stock markets. *Physical Review E*, 72(4):046133, 2005.
- [106] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.
- [107] Filippo Radicchi, José J Ramasco, and Santo Fortunato. Information filtering in complex weighted networks. *Physical Review E*, 83(4):046101, 2011.
- [108] Takayuki Mizuno, Hideki Takayasu, and Misako Takayasu. Correlation networks among currencies. *Physica A: Statistical Mechanics and its Applications*, 364:336–342, 2006.
- [109] Dong-Ming Song, Michele Tumminello, Wei-Xing Zhou, and Rosario N Mantegna. Evolution of worldwide stock markets, correlation structure, and correlation-based graphs. *Physical Review E*, 84(2):026108, 2011.
- [110] Matteo Marsili. Dissecting financial markets: sectors and states. *Quantitative Finance*, 2(4):297–302, 2002.

- [111] Michele Tumminello, Claudia Coronello, Fabrizio Lillo, Salvatore Micciche, and Rosario N Mantegna. Spanning trees and bootstrap reliability estimation in correlation-based networks. *International Journal of Bifurcation and Chaos*, 17(07):2319–2329, 2007.
- [112] J-P Onnela, Kimmo Kaski, and Janos Kertész. Clustering and information in correlation based financial networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):353–362, 2004.
- [113] Vladimir Boginski, Sergiy Butenko, and Panos M Pardalos. Statistical analysis of financial networks. *Computational statistics & data analysis*, 48(2):431–443, 2005.
- [114] Michele Tumminello, Tiziana Di Matteo, Tomaso Aste, and RN Mantegna. Correlation based networks of equity returns sampled at different time horizons. *The European Physical Journal B-Condensed Matter and Complex Systems*, 55(2):209–217, 2007.
- [115] R Coelho, S Hutzler, P Repetowicz, and P Richmond. Sector analysis for a FTSE portfolio of stocks. *Physica A: Statistical Mechanics and its Applications*, 373:615–626, 2007.
- [116] Yakov Amihud. Illiquidity and stock returns: cross-section and time-series effects. *Journal of financial markets*, 5(1):31–56, 2002.
- [117] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.
- [118] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304, 1998.
- [119] Robin Sibson. SLINK: an optimally efficient algorithm for the single-link cluster method. *The computer journal*, 16(1):30–34, 1973.
- [120] Daniel Defays. An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366, 1977.
- [121] Tiziana Di Matteo and Tomaso Aste. How does the Eurodollar Interest Rate behave? *International Journal of Theoretical and Applied Finance*, 5(1):107–122, 2002.
- [122] Kriegel, Hans-Peter and Kröger, Peer and Sander, Jörg and Zimek, Arthur. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):231–240, 2011.
- [123] Stuart Lloyd. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2):129–137, 1982.
- [124] Lawrence Hubert. Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. *Journal of the American Statistical Association*, 69(347):698–704, 1974.
- [125] Robert C Dunnell. Seriation method and its evaluation. *American Antiquity*, 35(3):305–319, 1970.
- [126] Innar Liiv. Seriation and matrix reordering methods: An historical overview. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 3(2):70–91, 2010.
- [127] Innar Liiv, Rain Opik, Jaan Ubi, and John Stasko. Visual matrix explorer for collaborative seriation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(1):85–97, 2012.
- [128] Luca Marotta, Salvatore Micciche, Yoshi Fujiwara, Hiroshi Iyetomi, Hideaki Aoyama, Mauro Gallegati, and Rosario N Mantegna. Bank-firm credit network in Japan: an analysis of a bipartite network. *PloS one*, 10(5):e0123079, 2015.
- [129] Michael C Münnix, Takashi Shimada, Rudi Schäfer, Francois Leyvraz, Thomas H Seligman, Thomas Guhr, and H Eugene Stanley. Identifying states of a financial market. *Scientific reports*, 2(644), 2012.
- [130] John H Holland. *Signals and boundaries: Building blocks for complex adaptive systems*. MIT Press, 2012.

- [131] Eduardo Viegas, Misako Takayasu, Wataru Miura, Koutarou Tamura, Takaaki Ohnishi, Hideki Takayasu, and Henrik Jeldtoft Jensen. Ecosystems perspective on financial networks: Diagnostic tools. *Complexity*, 19(1):22–36, 2013.
- [132] Murray Gell-Mann. Let’s call it plectics’. *Complexity*, 1(5):96, 1995.
- [133] Henry C K Liu. The Crisis of Wealth Destruction. *Economy and Growth, Finance and Wealth, Roosevelt Institute*, 2010.
- [134] The World Bank. World, 2016. data retrieved from World Development Indicators, <http://data.worldbank.org/region/world>.
- [135] The World Bank. East Asia & Pacific, 2016. data retrieved from World Development Indicators, <http://data.worldbank.org/region/east-asia-and-pacific>.
- [136] The World Bank. United States, 2016. data retrieved from World Development Indicators, <http://data.worldbank.org/country/united-states>.
- [137] The World Bank. European Union, 2016. data retrieved from World Development Indicators, <http://data.worldbank.org/region/EUU>.
- [138] Jean-Philippe Bouchaud, Yuval Gefen, Marc Potters, and Matthieu Wyart. Fluctuations and response in financial markets: the subtle nature of random price changes. *Quantitative finance*, 4(2):176–190, 2004.
- [139] Didier Sornette. Endogenous versus exogenous origins of crises. In *Extreme events in nature and society*, pages 95–119. Springer, 2006.
- [140] Robert Engle and Riccardo Colacito. Testing and valuing dynamic correlations for asset allocation. *Journal of Business & Economic Statistics*, 24(2):238–253, 2006.
- [141] Wayne E Ferson and Campbell R Harvey. The risk and predictability of international equity returns. *Review of financial Studies*, 6(3):527–566, 1993.
- [142] Geert Bekaert and Campbell R Harvey. Time-varying world market integration. *The Journal of Finance*, 50(2):403–444, 1995.
- [143] Vedat Akgiray. Conditional heteroscedasticity in time series of stock returns: Evidence and forecasts. *Journal of Business*, 62(1):55–80, 1989.
- [144] Tim Bollerslev, Ray Y Chou, and Kenneth F Kroner. ARCH modeling in finance: A review of the theory and empirical evidence. *Journal of Econometrics*, 52(1):5–59, 1992.
- [145] James D Hamilton and Raul Susmel. Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, 64(1):307–333, 1994.
- [146] Andrew A Christie. The stochastic behavior of common stock variances: Value, leverage and interest rate effects. *Journal of Financial Economics*, 10(4):407–432, 1982.
- [147] Peter Carr, Hélyette Geman, Dilip B Madan, and Marc Yor. Stochastic volatility for Lévy processes. *Mathematical Finance*, 13(3):345–382, 2003.
- [148] Suzanne S Lee and Per A Mykland. Jumps in financial markets: A new nonparametric test and jump dynamics. *Review of Financial Studies*, 21(6):2535–2563, 2008.
- [149] Richard G Palmer, W Brian Arthur, John H Holland, Blake LeBaron, and Paul Tayler. Artificial economic life: a simple model of a stockmarket. *Physica D: Nonlinear Phenomena*, 75(1):264–274, 1994.
- [150] Anders Johansen and Didier Sornette. Financial” anti-bubbles”: Log-periodicity in gold and Nikkei collapses. *International Journal of Modern Physics C*, 10(04):563–575, 1999.
- [151] Didier Sornette, Anders Johansen, et al. Significance of log-periodic precursors to financial crashes. *Quantitative Finance*, 1(4):452–471, 2001.
- [152] H Saleur, CG Sammis, and D Sornette. Discrete scale invariance, complex fractal dimensions, and log-periodic fluctuations in seismicity. *Journal of Geophysical Research-All series*, 101(BB):17661–17677, 1996.

- [153] Didier Sornette. Predictability of catastrophic events: Material rupture, earthquakes, turbulence, financial crashes, and human birth. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2522–2529, 2002.
- [154] Fabrizio Lillo and Rosario N Mantegna. Power-law relaxation in a complex system: Omori law after a financial market crash. *Physical Review E*, 68(1):016119, 2003.
- [155] Bikas K Chakrabarti, Arnab Chatterjee, and Pratip Bhattacharyya. Two-fractal overlap time series: Earthquakes and market crashes. *Pramana*, 71(2):203–210, 2008.
- [156] Anders Johansen and Didier Sornette. Endogenous versus Exogenous Crashes in Financial Markets. *Social Science Research Network Working Paper Series*, 2002.
- [157] Joao P Da Cruz and Pedro G Lind. Self-organized criticality in a network of economic agents with finite consumption. *Physica A: Statistical Mechanics and its Applications*, 391(4):1445–1452, 2012.
- [158] Yuko Hashimoto, Takatoshi Ito, Takaaki Ohnishi, Misako Takayasu, Hideki Takayasu, and Tsutomu Watanabe. Random walk or a run. Market microstructure analysis of foreign exchange rate movements based on conditional probability. *Quantitative Finance*, 12(6):893–905, 2012.
- [159] Tobias Preis, Johannes J Schneider, and H Eugene Stanley. Switching processes in financial markets. *Proceedings of the National Academy of Sciences*, 108(19):7674–7678, 2011.
- [160] Jose A Scheinkman and Michael Woodford. Self-organized criticality and economic fluctuations. *The American Economic Review*, 84(2):417–421, 1994.
- [161] M Bartolozzi, DB Leinweber, and AW Thomas. Self-organized criticality and stock market dynamics: an empirical study. *Physica A: Statistical Mechanics and its Applications*, 350(2):451–465, 2005.
- [162] Thomas Lux and Michele Marchesi. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature*, 397(6719):498–500, 1999.
- [163] G William Schwert. Business cycles, financial crises, and stock volatility. In *Carnegie-Rochester Conference Series on Public Policy*, volume 31, pages 83–125. Elsevier, 1989.
- [164] Thomas C Chiang, Bang Nam Jeon, and Huimin Li. Dynamic correlation analysis of financial contagion: Evidence from Asian markets. *Journal of International Money and Finance*, 26(7):1206–1228, 2007.
- [165] Marten Scheffer, Jordi Bascompte, William A Brock, Victor Brovkin, Stephen R Carpenter, Vasilis Dakos, Hermann Held, Egbert H Van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59, 2009.
- [166] Hayato Goto, Eduardo Viegas, Henrik Jeldtoft Jensen, Hideki Takayasu, and Misako Takayasu. Appearance of Unstable Monopoly State Caused by Selective and Concentrative Mergers in Business Networks. *Scientific reports*, 7(1):5064, 2017.
- [167] Philippe Donnat, Gautier Marti, and Philippe Very. Toward a generic representation of random variables for machine learning. *Pattern Recognition Letters*, 70:24–31, 2016.
- [168] Arthur MYR Sousa, Hideki Takayasu, and Misako Takayasu. Influence Networks in the Foreign Exchange Market. In *Proceedings of the International Conference on Social Modeling and Simulation, plus Econophysics Colloquium 2014*, pages 3–13. Springer, 2015.
- [169] Toru Ohira, Naoya Sazuka, Kouhei Marumo, Tokiko Shimizu, Misako Takayasu, and Hideki Takayasu. Predictability of currency market exchange. *Physica A: Statistical Mechanics and its Applications*, 308(1):368–374, 2002.
- [170] Yi Zuo and Eisuke Kita. Up/down analysis of stock index by using Bayesian network. *Engineering Management Research*, 1(2):46–52, 2012.

- [171] S Prakash Sethi. Comparative cluster analysis for world markets. *Journal of Marketing Research*, 8(3):348–354, 1971.
- [172] Sangeeta Rani and Geeta Sikka. Recent techniques of clustering of time series data: a survey. *International Journal of Computer Applications*, 52(15), 2012.
- [173] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering—A decade review. *Information Systems*, 53:16–38, 2015.
- [174] Boon Kin Teh and Siew Ann Cheong. Cluster fusion-fission dynamics in the Singapore stock exchange. *The European Physical Journal B*, 88(10):1–14, 2015.
- [175] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.
- [176] M Levin. B. mirkin, mathematical classification and clustering. *Journal of Global Optimization*, 12(1):105–108, 1998.
- [177] Guojun Gan, Chaoqun Ma, and Jianhong Wu. *Data Clustering: Theory, Algorithms, and Applications*, volume 20. SIAM, 2007.
- [178] Mark Kac, Jack Kiefer, and Joseph Wolfowitz. On tests of normality and other tests of goodness of fit based on distance methods. *The Annals of Mathematical Statistics*, 26(2):189–211, 1955.
- [179] Ralph B D’Agostino. Transformation to normality of the null distribution of g1. *Biometrika*, 57(3):679–681, 1970.
- [180] Carlos M Jarque and Anil K Bera. A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, 55(2):163–172, 1987.
- [181] SS Shaphiro and MB Wilk. An analysis of variance test for normality. *Biometrika*, 52(3):591–611, 1965.
- [182] Theodore W Anderson and Donald A Darling. Asymptotic theory of certain” goodness of fit” criteria based on stochastic processes. *The annals of mathematical statistics*, 23:193–212, 1952.
- [183] T. W. Anderson and D. A. Darling. A Test of Goodness of Fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954.
- [184] Frank J. Massey Jr. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.
- [185] Hubert W Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, 62(318):399–402, 1967.
- [186] Nornadiah Mohd Razali, Yap Bee Wah, et al. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33, 2011.
- [187] Mark EJ Newman. Power laws, Pareto distributions and Zipf’s law. *Contemporary physics*, 46(5):323–351, 2005.
- [188] Rosario N Mantegna and H Eugene Stanley. Scaling behaviour in the dynamics of an economic index. *Nature*, 376(6535):46–49, 1995.
- [189] Vasiliki Plerou, Parameswaran Gopikrishnan, Luis A Nunes Amaral, Martin Meyer, and H Eugene Stanley. Scaling of the distribution of price fluctuations of individual companies. *Physical Review E*, 60(6):6519, 1999.
- [190] Parameswaran Gopikrishnan, Vasiliki Plerou, Luis A Nunes Amaral, Martin Meyer, and H Eugene Stanley. Scaling of the distribution of fluctuations of financial market indices. *Physical Review E*, 60(5):5305, 1999.

- [191] Boon Kin Teh and Siew Ann Cheong. The Asian Correction Can Be Quantitatively Forecasted Using a Statistical Model of Fusion-Fission Processes. *PloS one*, 11(10):e0163842, 2016.
- [192] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.
- [193] Ramon Ferrer i Cancho and Richard V Solé. The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1482):2261–2265, 2001.
- [194] Felix Auerbach. Das Gesetz der Bevölkerungskonzentration. *Petermanns Geographische Mitteilungen*, 59:74–76, 1913.
- [195] Xavier Gabaix and Yannis M Ioannides. The evolution of city size distributions. *Handbook of regional and urban economics*, 4:2341–2378, 2004.
- [196] Takaaki Ohnishi, Takayuki Mizuno, Chihiro Shimizu, and Tsutomu Watanabe. Power laws in real estate prices during bubble periods. *International Journal of Modern Physics: Conference Series*, 16:61–81, 2012.
- [197] Darrell Jiajie Tay, Chung-I Chou, Sai-Ping Li, Shang You Tee, and Siew Ann Cheong. Bubbles Are Departures from Equilibrium Housing Markets: Evidence from Singapore and Taiwan. *PloS one*, 11(11):e0166004, 2016.
- [198] Benoit Mandelbrot. The Pareto-Levy law and the distribution of income. *International Economic Review*, 1(2):79–106, 1960.
- [199] Victor M Yakovenko and J Barkley Rosser Jr. Colloquium: Statistical mechanics of money, wealth, and income. *Reviews of Modern Physics*, 81(4):1703, 2009.
- [200] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.
- [201] Jeffrey M Wooldridge. Applications of generalized method of moments estimation. *The Journal of Economic Perspectives*, 15(4):87–100, 2001.
- [202] Ronald Aylmer Fisher. On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41:155–160, 1912.
- [203] Bungon Kumphon. Maximum Entropy and Maximum Likelihood Estimation for the Three-Parameter Kappa Distribution. *Open Journal of Statistics*, 2(4):415–419, 2012.
- [204] Z Hradil and J Reháček. Likelihood and entropy for statistical inversion. In *Journal of Physics: Conference Series*, volume 36, page 55. IOP Publishing, 2006.
- [205] Hirotugu Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.
- [206] American Statistical Association. ASA P-Value Statement. <https://www.amstat.org/ASA/News/ASA-P-Value-Statement-Viewed-150000-Times.aspx>, 2016. Accessed: 2017-03-07.
- [207] Ronald L Wasserstein and Nicole A Lazar. The ASA’s statement on p-values: context, process, and purpose. *Am Stat*, 70(2):129–133, 2016.
- [208] Monya Baker. Statisticians issue warning over misuse of P values. *Nature*, 531:151, 2016.
- [209] Frank J. Massey Jr. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.
- [210] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. powerlaw: a Python package for analysis of heavy-tailed distributions. *PloS one*, 9(1):e85777, 2014.
- [211] Shan Yu, Andreas Klaus, Hongdian Yang, and Dietmar Plenz. Scale-invariant neuronal avalanche dynamics and the cut-off in size distributions. *PloS one*, 9(6):e99761, 2014.
- [212] Najja Marshall, Nicholas M Timme, Nicholas Bennett, Monica Ripp, Edward Lautzenhiser, and John M Beggs. Analysis of power laws, shape collapses, and neural complexity:

- new techniques and Matlab support via the ncc toolbox. *Frontiers in Physiology*, 7(250), 2016.
- [213] Edwin JG Pitman and Edwin James George Pitman. *Some basic theory for statistical inference*, volume 7. Chapman and Hall London, 1979.
- [214] Ronald Pyke. Spacings. *Journal of the Royal Statistical Society. Series B (Methodological)*, 27(3):395–449, 1965.