The Chinese correction of February 2007 : how financial hierarchies change in a market crash

Choi, Wen Ting; Damodaran, Mridula; Cheong, Siew Ann; Teh, Boon Kin; Goo, Yik Wen; Lian, Tong Wei; Ong, Wei Guang

2015

Teh, B. K., Goo, Y. W., Lian, T. W., Ong, W. G., Choi, W. T., Damodaran, M., et al. (2015). The Chinese correction of February 2007 : how financial hierarchies change in a market crash. *Physica A : Statistical Mechanics and its Applications*, 424, 225-241.

<https://hdl.handle.net/10356/96390>

<https://doi.org/10.1016/j.physa.2015.01.024>

© 2015 Elsevier B.V. This is the author created version of a work that has been peer reviewed and accepted for publication by *Physica A: Statistical Mechanics and its Applications*, Elsevier. It incorporates referee's comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at:
[<http://dx.doi.org/10.1016/j.physa.2015.01.024>].

Downloaded on 26 Apr 2023 22:19:50 SGT

The Chinese Correction of Feb 2007: How Financial Hierarchies Change in a Market Crash

Boon Kin Teh^{a,e,*}, Yik Wen Goo^{b,d}, Tong Wei Lian^{c,d}, Wei Guang Ong^{a,d}, Wen Ting Choi^a,
Mridula Damodaran^a, Siew Ann Cheong^{a,e}

^a*Division of Physics and Applied Physics, School of Physical and Mathematical Sciences, Nanyang Technological University, 21 Nanyang Link, Singapore 637371, Republic of Singapore*

^b*School of Mechanical and Aerospace Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Republic of Singapore*

^c*School of Electronic and Electrical Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Republic of Singapore*

^d*CN Yang Scholars Program, Nanyang Technological University*

^e*Complexity Institute, Nanyang Technological University, 60 Nanyang View, Singapore 639673, Republic of Singapore*

Abstract

We analyzed 546 stocks in the Singapore Stock Exchange (SGX) and 1173 stocks in the Hong Kong Stock Exchange (HKSE) in 2006 and 2007, to understand how financial hierarchies on these two markets change over market corrections and crashes. To do so, we introduced the *digital cross correlation* as a measure of the comovement tendencies between stock pairs, and also the method of *partial hierarchical clustering* to iteratively identify strongly-correlated clusters of stocks. From daily prices over the 2006–2007 period, we found the existence of clusters of local stocks as well as clusters of Chinese stocks traded on the two markets. We further discovered the Chinese clusters organizing into a Chinese supercluster, interacting less strongly with a supercluster dominated by local clusters. Going down to 30-minute prices within two-month overlapping time windows over 2006 and 2007, we found dips in the number of clusters before market corrections and crashes, followed by peaks in the number of clusters afterwards. On the SGX, we also found the stronger intracluster correlation weakening, and the weaker intercluster correlation strengthening before the Feb 2007 Chinese Correction. These features are in qualitative agreement with a chemical reactions picture in which clusters of stocks ‘react’ to form large superclusters of stocks that ‘dissociate’ during market crashes. Finally, on the SGX we found broad humps in the intracluster and intercluster correlations for the May/Jun 2006 market correction and the Feb 2007 Chinese Correction, but a sharp peak for the Jul 2007 Subprime Crisis. This suggests that the earlier events were endogeneous to the SGX, while the latter event was an exogeneous shock.

Keywords: Singapore Stock Exchange, Hong Kong Stock Exchange, digital cross correlations, partial hierarchical clustering, minimal spanning tree, Chinese Correction, Subprime Crisis

PACS: 05.45.Tp, 87.23.Ge, 89.65.Gh, 89.75.Fb

---

*Corresponding author

Email address: S130005@e.ntu.edu.sg (Boon Kin Teh)

1. Introduction

From statistical mechanics, we know that a system that explores its N -dimensional configuration space uniformly has the highest possible entropy. In such a system, we can make no better long-time prediction than a random guess. Therefore, we have the least possible information on such a system. If the system visits a specific region in the configuration space more frequently than others, the entropy of the system will be lower. Moreover, the concentration of probability allows us to make better predictions. We thus have more information on such a system. Essentially, the stock market is all about making predictions, whether we are chartists betting on a short-lived bull or bear runs, or fundamentalists investing in long-term trends. The more information we can distill from the seemingly stochastic stock price and volume fluctuations, the more accurate our predictions will be, and the more money we expect to make.

The first major breakthrough on this prediction problem came when econophysicists applied random matrix theory (RMT) [1, 2] to financial cross correlations to tease apart signals from noise. In their paper, Plerou et al. found market signals that produce cross correlations much stronger than what RMT predicts from an uncorrelated cross section of stocks with the same amount of return fluctuations [3, 4]. They also discovered that 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 major breakthrough, i.e. the discovery by Mantegna that there exists a hierarchical organization of stocks in the financial market [5]. By visualizing the cross correlations in terms of a minimal spanning tree (MST), and also by performing ultrametric hierarchical clustering, Mantegna showed that beyond the strong cross correlations between stocks within the same sectors (already apparent from RMT studies), there exist weaker cross correlations between different sectors, and cross correlations between groups of sectors that are weaker still.

After the two seminal papers, initial efforts were focused on showing that the cross correlations and the hierarchies they represent are stable in time [6–10]. This is reasonable when the market is steadily growing, but during market crashes and other turbulent market periods, a large cross section of traders actively restructure their portfolios and strategies. Therefore, it is very unlikely that the strongly-correlated clusters of stocks remain 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. More importantly, this ‘chemical reaction’ picture of stock market dynamics suggests by analogy the existence of ‘reaction pathways’ and ‘elementary reactions’, whereby the new correlations between all stocks involved must rise to the levels of the old correlations before the latter decay, signaling completion of the ‘chemical reaction’. In fact, this increase in cross correlations before the market crashes has been observed by many econophysicists for many market crashes [11–14]. Recently, Quax et al. also took a big step forward to map out the ‘reaction pathway’ near the tipping point, by developing an information-theoretic approach to the market prediction problem [15].

Why do markets crash? Sornette and others classified all US market crashes, and found that most do not have exogenous causes [16–20]. The endogenous dynamics of a stock market are clearly complex. But even so, we would like to map out this complex dynamics as a market crash unfolds, to see if there are universal lessons that can be learned. For this reason, we identified a global market crash at the end of Feb 2007 for our case study. This has since come to be known as the Chinese Correction [22], and economists attribute this to structural problems in the Chinese economy [23–25]. However, why then are most markets around the world affected, when they

are unlikely to have exposure to Chinese stocks? To better understand how a correction on the Shanghai Stock Exchange triggers a contagion effect on other markets, we look at the only two stock markets outside of China, on which Chinese stocks are listed: the Singapore Stock Exchange (SGX) and the Hong Kong Stock Exchange (HKSE). To understand how these small cross sections of Chinese stocks trigger crashes on both markets, we examine the interactions between Chinese stocks and local stocks. To do this, we perform time series clustering in different time windows over 2006 and 2007, to see how the hierarchical organization of stocks changes with time, and in particular across the crash. To avoid common problems associated with the linear cross correlations, we introduce a new correlation measure called the *digital cross correlation*, which weighs all synchronous changes equally. We also introduce a new clustering method to map out the hierarchical organization of stocks.

To tell the story of how the Chinese Correction unfolded in the SGX and HKSE, we organize our paper as follows. In Section 2, we describe the data as well as methods used in this research. Specifically, we define the digital cross correlation and explain how it can be computed, and also how *partial hierarchical clustering* can be done by identifying natural thresholds in the resulting digital cross correlations. In Section 3, we report the clusters of stocks identified in SGX and HKSE over the entire 2006–2007 time window. On both markets, we found the existence of clusters of local stocks, as well as clusters of Chinese stocks. These are further organized into a supercluster dominated by Chinese clusters, and a supercluster dominated by local clusters. In Section 4, we go one step further, to study how these clusters evolve over shorter time windows with the 2006–2007 time period. We find dynamical features consistent with a chemical reactions picture, i.e. the number of clusters decreases and thereafter increases, while stronger correlations within clusters weaken and weaker correlations between clusters strengthen before a market crash. Through this dynamical analysis of cross correlations, we show that it might be possible to tell apart endogeneous crashes from exogeneous shocks. Finally, we conclude in Section 5.

2. Data and Methods

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 Feb 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 cross 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 Stock 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. These daily index data were downloaded from Stooq (<http://stooq.com/>) for the four months before and the four months after the Chinese Correction.

Figure 1 show the cross correlations between daily index values of SHCI, SHBS, HSI, STI, NKX, FTSE and DJI. Before the market crash we can see that the SHCI is strongly correlated with

Figure 1: Pearson cross correlations between the daily index values 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.

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 [21] 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. In Section 3, we compute the digital cross correlations over the entire two-year period, whereas in Section 4, we use sliding windows that are two-months long, such that successive time windows overlap each other by one month. This is to ensure we can follow the changes in the hierarchical structure as we move from window to window.

Figure 2: Three time series simultaneously affected by an extreme event. The normal fluctuations of time series (a) are anticorrelated with those in time series (b), but their responses to the extreme event are in the same direction. Because of this, the Pearson (linear) correlation between (a) and (b) is positive. Conversely, the normal fluctuations of time series (c) are correlated with those in time series (b), but their responses to the extreme event are in opposite directions. As a result, the Pearson correlation between (b) and (c) is negative.

2.2. Linear and Digital Cross Correlations

To detect hierarchical structures in stock markets, we typically start from the cross correlations of the price time series. Given two price time series

$$\mathbf{P}_i = (P_i(1), P_i(2), \dots, P_i(t), \dots, P_i(T)), \quad \mathbf{P}_j = (P_j(1), P_j(2), \dots, P_j(t), \dots, P_j(T)) \quad (1)$$

for stocks i and j , the most common measure of the similarity between their fluctuations is the *Pearson cross correlation* (PC) or *linear cross correlation*

$$PC(i, j) = \frac{\frac{1}{T} \sum_{t=1}^T [P_i(t) - \bar{P}_i] [P_j(t) - \bar{P}_j]}{\sqrt{\frac{1}{T} \sum_{t=1}^T [P_i(t) - \bar{P}_i]^2} \sqrt{\frac{1}{T} \sum_{t=1}^T [P_j(t) - \bar{P}_j]^2}}, \quad (2)$$

where $\bar{P}_i = \frac{1}{T} \sum_{t=1}^T P_i(t)$ and $\bar{P}_j = \frac{1}{T} \sum_{t=1}^T P_j(t)$ are the means of $\mathbf{P}_i$ and $\mathbf{P}_j$ respectively.

However, the Pearson correlation is very sensitive to extreme events. As shown in Figure 2, 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 fact that financial time series are typically nonstationary, and consists of a mixture of calm and crash phases. A single standard deviation is thus not a good measure of fluctuations within the nonstationary financial time series. To reduce the sensitivity to extreme events, we can use the Spearman rank correlation [26]

$$SC(i, j) = 1 - \frac{6 \sum_{t=1}^T [RP_i(t) - RP_j(t)]^2}{T(T^2 - 1)}, \quad (3)$$

which is the Pearson correlation between the ranked variables $RP_i(t)$ and $RP_j(t)$ of $\mathbf{P}_i$ and $\mathbf{P}_j$ 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 when dealing with non-stationary time series, because all elements from the time series are given a global ranking.

Like the Spearman correlation, the Kendall tau correlation [27] also work with rank statistics and is hence less sensitive to extreme events. Unlike the Spearman correlation, the Kendal tau correlation

$$KC(i, j) = \frac{N_{\text{concordant}} - N_{\text{Disconcordant}}}{\frac{1}{2}T(T-1)} \quad (4)$$

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

In this study, we introduce a new measure of cross correlations between financial stocks, which we call the *digital cross correlation* (DC) [28]. The digital correlation measures how frequently $\mathbf{P}_i$

and $\mathbf{P}_j$ move together, ignoring the magnitude of the price movements. We define it as

$$DC(i, j) = \sum_{t=1}^T \Phi(\Delta P_i(t) \Delta P_j(t)), \quad \Phi(x) \begin{cases} = 0 & \text{if } x \leq 0; \\ = 1 & \text{if } x > 0, \end{cases} \quad (5)$$

where $\Delta P_i(t) = P_i(t+1) - P_i(t)$. We can easily check for the example in Figure 2 that the digital correlation between times series (a) and (b) is negative, whereas that between time series (b) and (c) is positive, i.e. the digital correlation, by construction, is much less sensitive to isolated extreme events.

Figure 3: The nonlinear (Spearman, Kendall tau and digital) and linear (Pearson) 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 Gaussian random variables x_1 and x_2 . In this figure, we have set $\langle x_1 \rangle = \langle x_2 \rangle = 0$, $\langle x_1^2 \rangle = \langle x_2^2 \rangle = 1$, so that $\langle y_1 \rangle = \langle y_2 \rangle = 0$. 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 Figure 3 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. Comparing the Kendall Tau correlation, which is bounded between -1 to 1 , the digital correlation scales as the size of the data, and can potentially offers higher resolution. In addition, from the definitions of these correlations, we see that the Kendall tau correlation incorporates comparisons at all time scales, but the digital correlation only performs local comparisons. In particular, extreme events are troublesome when dealing with the price time series of stocks, because they are subjected to idiosyncratic shocks. Since we are aiming to identify clusters of stocks with consistent comovements, the use of digital correlations limit the influences of such extreme events on our clustering analysis. We also understand how uncorrelated noise will give rise to a background digital correlation level of $DC = T/2$. This will not affect our ability to pick out significant correlations or anti-correlations from pairs of time series with consistent comovements, because they will rise about or fall below the background level.

2.3. *Partial Hierarchical Clustering*

Even for small markets like the SGX and HKSE, there are too many pairs of cross correlations to examine in detail. Instead of all cross correlations, we would like to understand the market hierarchical structure only in terms of the strongest cross correlations. Many filtering schemes have been proposed to help us do this, including the use of a simple threshold [29], the minimal spanning tree (MST)[5, 30–33], and the planar maximally filtered graph (PMFG) [34, 35]. 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 cross correlations with other stocks within the same subgraph, but weaker cross 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 cross correlations.

Clustering analysis can also be done starting from the MST or the PMFG, to obtain a minimum spanning forest (MSF) [35] or the a directed bubble hierarchical tree (DBHT) [36] 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 [41], while Radichi et al. developed a *global statistic significance* (GloSS) filter that keeps links that are statistically significant [42]. 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.

A single threshold is also commonly assumed in hierarchical clustering methods [38–40] for breaking up a dendrogram into the desired number of clusters. In a multiscale 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. Besides this threshold problem, full hierarchical clustering also suffers from the problem of *history dependence*. Suppose two points are close to each other, but are each closer to points further away. Once they are clustered separately, they will sometimes end up in two different high-level clusters. This happens whether we use complete-linkage hierarchical clustering (CLHC, more suitable for grouping objects reacting to similar environmental influences [43]), average-linkage hierarchical clustering (ALHC), or single-linkage hierarchical clustering (SLHC, more suitable for grouping objects derived with variations from the same parents [44, 45]). The two high-level clusters suggests that the two points are not close to each other, which is not the case.

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 Figure 4, 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 data points that are closest together, and use as the *seed* to grow the cluster. In the example shown in Figure 4, the seed cluster is $c = \{1, 17\}$ from cluster A. To add the next data point to c , we first determine the maximum distances

$$d_{\max}^c(j) = \max_{i \in c} d_{ij} \quad (6)$$

for all points $j \notin c$ to points in the cluster c . We then pick the point j^* , such that

$$d_{\max}^c(j^*) = \min_{j \notin c} d_{\max}^c(j), \quad (7)$$

to add to c . This is the *complete-linkage algorithm* for hierarchical clustering. We can also use the *single-linkage algorithm* 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 Figures 4(b) and 4(c). To simplify discussions, we will use the complete-linkage algorithm exclusively in this paper.

Figure 4: (a) 21 data points, 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.

In Figure 4(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 $A = \{1, 17, 18, 12, 13, 5, 16\}$ 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 Figure 4(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.

Looking more carefully at the distribution of points in Figure 4(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 Figure 4(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 Figure 4(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 $B = \{4, 8, 2, 21, 6, 7, 10, 19\}$ 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 $C = \{3, 9, 11, 20, 14, 15\}$ as a natural cluster.

To perform partial hierarchical clustering of the SGX and HKSE, we can start from the normalized cross correlations $PC_{ij} = PC(i, j)$ between stocks i and j , compute the ultrametric distances [5]

$$d_{ij} = \sqrt{2(1 - PC_{ij})}, \quad (8)$$

to discover the natural clusters of stocks. However, since d_{ij} is a monotonic decreasing function of PC_{ij} , we can modify the complete-linkage algorithm to work directly with correlations:

1. determine the minimum correlations

$$C_{\min}^c(j) = \min_{i \in c} C_{ij} \quad (9)$$

for all stocks $j \notin c$ to stocks in the cluster c ; and

2. pick the stock j^* such that

$$C_{\min}^c(j^*) = \max_{j \notin c} C_{\min}^c(j), \quad (10)$$

to add to c .

Since this algorithm does not require the cross correlations to be normalized, we can also apply it to the digital cross correlations $DC_{ij} = DC(i, j)$ between the stock i and stock j , as is done in this paper. To identify the natural clusters, we then look out for sharp drops in the complete-linkage correlation as the cluster grows. A MATLAB graphical user interface was developed to help us identify visually such sharp drops and pick out the natural clusters.

2.4. Analysis of Intracluster and Intercluster Correlations

While partial hierarchical clustering allows us to determine a natural and reliable set of clusters in time window, the clusters alone tell a very limited story. In order to obtain more insights from these clusters, we need to also consider correlations between clusters, and correlations between components of a cluster. First, let us consider the average intracluster correlation within cluster I ,

$$C_{intra}(I) = \frac{1}{N_I(N_I - 1)/2} \sum_{i, i' \in I} DC(i, i'), \quad (11)$$

and the intercluster correlation between clusters I and J ,

$$C_{inter}(I, J) = \frac{1}{N_I N_J} \sum_{i \in I} \sum_{j \in J} DC(i, j), \quad (12)$$

obtained for a particular time window. Here i and i' are component stocks from cluster I , j is a component stock from cluster J , N_I is the number of components in cluster I , N_J is the number of components in cluster J , and $DC(k, l)$ is the digital correlation between the stock k and stock l . The standard errors of these average correlations are

$$\delta C_{intra}(I) = \sqrt{\langle [DC(i, i')]^2 \rangle - \langle DC(i, i') \rangle^2} \quad (13)$$

and

$$\delta C_{inter}(I, J) = \sqrt{\langle [DC(i, j)]^2 \rangle - \langle DC(i, j) \rangle^2}, \quad (14)$$

where $\langle . \rangle$ is the average.

Because the clusters are discovered using the complete-linkage criterion, correlations within a cluster are fairly homogeneous. Correlations between clusters on the other hand can be heterogeneous, and $\delta C_{inter}(I, J)$ is an appropriate measure of this heterogeneity. In addition, there is also heterogeneity at the level of clusters as well. Let us explain the effect of this cluster-level heterogeneity by using the example shown in Figure 5. As we can see, the three clusters (C1, C2, and C3) are fairly homogeneous, because of the small spreads in their intracluster correlations. The intercluster correlations are moderately homogeneous, with moderate spreads in values. However, the more important feature we need to note is the large spread in intracluster correlations going from C1 to C3, and the large spread in intercluster correlations going from C1-C2 to C1-C3 to C2-C3. In such a scenario, we would obtain a large standard deviation if we compute the spread over *all* intracluster correlations, and an even larger standard deviation if we compute the spread over *all* intercluster correlations. Given that our interest is in understanding the correlation dynamics at the cluster level, which measure of spread should we use?

Assuming that clusters are robust within their individual lifetimes, we expect two kind of changes to the intracluster correlations of each cluster: (1) the average intracluster correlation

Figure 5: In the inset to this figure, we show the digital correlation matrix of three strongly-correlated clusters, C1, C2, and C3. Their average intracluster correlations are shown in red, orange, and yellow respectively. The average intracluster correlations are shown in green (C1-C2), cyan (C1-C3), and blue (C2-C3). In the figure proper, we show the histogram of the digital correlations, where intracluster correlations are stronger with smaller spreads while intercluster correlations are weaker with larger spreads.

increase or decrease; and (2) the spread of intraclass correlations increase or decrease. As the clusters interact, we also expect (3) the average intercluster correlation to increase or decrease; and (4) the spread of intercluster correlations to increase or decrease, with intercluster correlations weaker than intraclass correlations. As the market approaches a crash or correction, we expect interactions to become stronger and thus intraclass correlations weaken while intercluster correlations strengthen. To track such global changes at the cluster level, we can average over *all* intraclass correlations (red, orange, and yellow) and average over *all* intercluster correlations (green, cyan, and blue). However, we must *not* use the standard deviations computed over *all* intraclass and intercluster correlations, because these are simply not uniformly spread over the range of correlations as suggested by the standard deviations. To avoid potential confusions, we therefore monitor the median intraclass correlation

$$AC_{intra} = Md(C_{intra}(I)) \quad (15)$$

over all intraclass correlations and over all clusters, and the median intercluster correlation

$$AC_{inter} = Md(C_{inter}(I, J)). \quad (16)$$

Figure 6: The minimal spanning tree (MST) of six clusters, constructed from the intercluster correlations $C_{inter}(I, J)$. In this figure, the MST links (which are the most important intraclass correlations) are blue, while all other links are gray.

We also estimate the error for AC_{intra} as

$$\delta AC_{intra} = AC_{intra} \left\langle \frac{\delta C_{intra}(I)}{C_{intra}(I)} \right\rangle, \quad (17)$$

where the average of the coefficient of variation $\delta C_{intra}(I)/C_{intra}(I)$ is over all clusters. To estimate the error for AC_{inter} ,

$$\delta AC_{inter} = AC_{inter} \left\langle \frac{\delta C_{inter}(I, J)}{C_{inter}(I, J)} \right\rangle, \quad (18)$$

we observe from Figure 6 that there are typically many more pairs of intracluster correlations than there are clusters. We can of course average the coefficient of variation $\delta C_{inter}(I, J)/C_{inter}(I, J)$ over all pairs (I, J) , but we also observe in Figure 6 that not all $C_{inter}(I, J)$ are equally important. In general, the most important intercluster correlations are those forming the minimal spanning tree (MST) of the clusters. We therefore average $\delta C_{inter}(I, J)/C_{inter}(I, J)$ only over the MST links.

3. Clusters and Superclusters of Stocks

3.1. SGX and HKSE Clusters

In Figure 7, we show the complete-linkage correlation level as a function of the cluster size for the SGX and HKSE. 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.

Figure 7: Complete-linkage correlation level as a function of the cluster size for (a) the SGX starting from Cluster 8 (see Table 1), and (b) the HKSE starting from Cluster 10 (see Table 2). 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.

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 [47, 48]. In addition, studies also showed that financial stocks tend to cluster according to the industry sectors they belong to [49–56]. 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 Table 1 and Table 2. As we can see in Table 1 for SGX, 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. As we can see from Table 2, 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 Jul 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.

3.2. SGX and HKSE Superclusters

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 clusters. To do this, we construct the MST and MSF of the clusters, using the complete-linkage correlations as the pairwise correlations between clusters. Both the MST and MSF reveals higher-level self-organizations in the stock markets, and can be thought of as *superclusters* of stocks.

Across the 2006–2007 two-year period, the SGX MST/MSF is shown in Figure 9a while the HKSE MST/MSF is shown in Figure 9c. From Figure 9a, 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

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

Table 2: 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

Figure 8: The MSTs of (a) SGX and (b) HKSE over the period Jan 2006 to Dec 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.

manufacturing stocks. Over in the HKSE, the picture is different, because of the large number of Chinese clusters. From Figure 9c, 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.

4. Dynamics of Clusters and Superclusters

The results in Section 3 provide interesting insights into the Feb 2007 Chinese Correction in SGX and HKSE. In particular, 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 Feb 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 Feb 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 cross correlations of 30-minute stock prices, before doing the clustering and superclustering analyses.

In Figure 9(a) we show the partial hierarchical clustering results in a typical case, and contrast it with the best (Figure 9(b)) and worst cases (Figure 9(c)). We then plot the kink angles associated with all clusters identified from SGX and HKSE over all time windows in Figure 9(d), where we see that 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, we are less confident with the clusters identified, because many of them have kink angles that fall within the ambiguous range of 160° to 200° . Fortunately less than 10% of the clusters identified belong to this category. For each time window, we also take the number of ambiguous clusters to be the standard error for the number of 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 Figure 10. From the naive chemical reactions picture described in Section 1, 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 Figure 10. 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 can still fit our observations within the chemical reactions picture, if we concede that 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.

Within this modified chemical reaction, we can now compare the reactions of the SGX and HKSE. For the May/Jun 2006 correction, the number of clusters in the SGX and HKSE peaked

Figure 9: 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. As a naive measurement of how well partial hierarchical clustering performs, we measure the kink angle associated with the cluster, and show these in (d) as a scatter plot for clusters identified from SGX and HKSE in different time windows. The closer the angle to 180° the more ambiguous the cluster identified. Based on the density of points on the scatter plot, we chose the band of ambiguous clusters to be between 160° and 200°, and use these to estimate the errors in the numbers of clusters for different time windows.

in the same two-month window, although the ‘giant cluster’ for this event appears to have started fragmenting earlier in the HKSE. For the Feb 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 Jul 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 Feb 2007 Chinese Correction, but dipped during the Jul 2007 Subprime Crisis is consistent with such a source-contagion picture.

To look for further supporting evidence for such a picture of source-contagion chemical reactions, we also calculate the average intracluster and intercluster correlation level, as shown in Figure 11a and 11b the average intracluster correlation level and average intercluster correlation level in the SGX and HKSE. Again, from the chemical reactions picture, we expect the correlation level of the financial clusters increased when closer to the financial crisis, reached a maximum during the market crash, and then falls off after the market crash. In addition, the chemical reactions picture also suggests that the gap between the average intracluster correlation level and the average intercluster 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. The chemical reactions picture is most clear on the SGX, as we see two broad humps for the May/Jun 2006 market correction and the Feb 2007 Chinese Correction, and a sharp peak for the Jul 2007 Subprime Crisis. We do not understand why the HKSE has such weak correlational reactions to the three significant market events. More interestingly, from Figure 11a we see the gap between the average intracluster and intercluster correlation levels close at two times, Sep/Oct 2006, and Nov/Dec 2006. Going back to Figure 10, we find the number of SGX clusters dipping at these two times. This gap-closing feature is not seen for the May/Jun 2006 market correction and the Jul 2007 Subprime Crisis.

(a)

(b)

Figure 10: The number of financial clusters over overlapping two-month windows across Jan 2006 to Dec 2007 for (a) SGX and (b) HKSE. In this figure, the error bars are the numbers of ambiguous clusters identified through partial hierarchical clustering. The green vertical line marks a market correction that occur around May/June 2006, the red vertical line marks the Feb 2007 Chinese Correction, while the black vertical line marks the Subprime Crisis in Jul 2007.

(a)

(b)

Figure 11: The average intracluster correlation level and average intercluster 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 intracluster and the intercluster correlations are estimated using Equations (17) and (18). The green vertical line marks a market correction that occur around May/June 2006, the red vertical line marks the Feb 2007 Chinese Correction, while the black vertical line marks the Subprime Crisis in Jul 2007.

5. Conclusions

To conclude, we investigated how the hierarchical organizations in the SGX and HKSE changed across the Feb 2007 Chinese Correction by performing complete-linkage partial hierarchical clustering of the digital cross correlations between stocks. Over the entire 2006–2007 period, we find the existence of clusters of Chinese stocks listed on the two markets, in addition to the expected clusters of local stocks. These clusters are then organized into a Chinese supercluster and a local supercluster, whose interactions might have been causally responsible for the Feb 2007 crash in SGX and HKSE. To examine these interactions in detail, we repeated the clustering and superclustering analyses over overlapping two-month windows. We find the number of clusters dipped before significant market corrections and crashes, and thereafter increased. On the SGX, we also find the gap between the intracluster and intercluster correlations closing before the Feb 2007 Chinese Correction. These features are qualitatively consistent with a chemical reactions picture, where we expect clusters to merge into a giant cluster before market corrections and crashes. From this chemical reactions picture, we also expect the stronger intracluster correlations to weaken, and the weaker intercluster correlations to strengthen, leading to the gap between the two correlations closing during the market event. 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 suggests that the May/Jun 2006 and Feb 2007 events were endogenous to the SGX, while the Jul 2007 event was an exogenous shock.

Even though we do not believe there can be an exact mapping of stock market dynamics to chemical reaction dynamics, we have demonstrated in this paper how the chemical reaction picture can be used to guide our thinking about processes in the stock markets, what quantities to measure and present, and help us focus on insightful questions to ask. In addition, in our time-resolved analysis of how the market hierarchy changes across market crashes, we ended up with a large number of time windows and clusters. We were thus limited to plotting the time evolutions of the number of clusters and the intracluster and intercluster correlations. With these alone, it is difficult to construct a broader picture of how interactions between Chinese and local stocks led to the market crashes on SGX and HKSE. New visualization and analytical tools are necessary to meaningfully extract information from the clusters at different times. We have done this, and will report our results in another paper.

Acknowledgements

This research is supported by the Nanyang Technological University startup grant SUG 19/07, and also by the Nanyang Technological University’s CN Yang Scholars Program. We have had helpful discussions with Low Buen Sin, Charlie Charoenwong, Gerald Cheang Hock Lye, and Chris Kok Jun Liang.

References

- [1] M. L. Mehta and M. Gaudin, “On the density of eigenvalues of a random matrix,” *Nuclear Physics*, vol. 18, pp. 420–427, 1960.
- [2] E. P. Wigner, “Random matrices in physics,” *SIAM Review*, vol. 9, no. 1, pp. 1–23, 1967.

- [3] L. Laloux, P. Cizeau, M. Potters, and J.-P. Bouchaud, “Random matrix theory and financial correlations,” *International Journal of Theoretical and Applied Finance*, vol. 3, no. 03, pp. 391–397, 2000.
- [4] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley, “Random matrix approach to cross correlations in financial data,” *Physical Review E*, vol. 65, no. 6, pp. 066–126, 2002.
- [5] R. N. Mantegna, “Hierarchical structure in financial markets,” *The European Physical Journal B*, vol. 11, pp. 193–197, 1999.
- [6] M. Tumminello, F. Lillo, and R. N. Mantegna, “Correlation, hierarchies, and networks in financial markets,” *Journal of Economic Behavior & Organization*, vol. 75, no. 1, pp. 40–58, 2010.
- [7] D. Sornette and A. Johansen, “A hierarchical model of financial crashes,” *Physica A: Statistical Mechanics and its Applications*, vol. 261, no. 3, pp. 581–598, 1998.
- [8] D. Sornette, P. Simonetti, and J. V. Andersen, “ ϕ q field theory for portfolio optimization: fat tails and nonlinear correlations,” *Physics Reports*, vol. 335, no. 2, pp. 19–92, 2000.
- [9] E. J. Elton, M. J. Gruber, S. J. Brown, and W. N. Goetzmann, *Modern portfolio theory and investment analysis*. John Wiley & Sons, 2009.
- [10] G. Bonanno, F. Lillo, and R. N. Mantegna, *High-frequency cross-correlation in a set of stocks*. Taylor & Francis, 2001.
- [11] D. Sornette, *Why stock markets crash: critical events in complex financial systems*. Princeton University Press, 2009.
- [12] L. Sandoval and I. D. P. Franca, “Correlation of financial markets in times of crisis,” *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 1, pp. 187–208, 2012.
- [13] J. Maskawa and W. Souma, “Large correlations as a signal of instability in stock market,” *Preprint*, 2010.
- [14] Z. Zheng, B. Podobnik, L. Feng, and B. Li, “Changes in cross-correlations as an indicator for systemic risk,” *Scientific Reports*, vol. 2, Article No. 888, 2012.
- [15] R. Quax, D. Kandhai, and P. M. Sloot, “Information dissipation as an early-warning signal for the lehman brothers collapse in financial time series,” *Scientific reports*, vol. 3, Article No. 1898, 2013.
- [16] J.-P. Bouchaud, Y. Gefen, M. Potters, and M. Wyart, “Fluctuations and response in financial markets: the subtle nature of random price changes,” *Quantitative Finance*, vol. 4, no. 2, pp. 176–190, 2004.
- [17] D. Sornette, “Endogenous versus exogenous origins of crises,” in *Extreme events in nature and society, The Frontiers Collection 2006*. Springer, pp. 95–119, 2006.

- [18] R. Engle and R. Colacito, “Testing and valuing dynamic correlations for asset allocation,” *Journal of Business & Economic Statistics*, vol. 24, no. 2, pp. 238–253, 2006.
- [19] W. E. Ferson and C. R. Harvey, “The risk and predictability of international equity returns,” *Review of financial Studies*, vol. 6, no. 3, pp. 527–566, 1993.
- [20] G. Bekaert and C. R. Harvey, “Time-varying world market integration,” *The Journal of Finance*, vol. 50, no. 2, pp. 403–444, 1995.
- [21] “HKEx Fact Book,” *HKEx*, 2008.
- [22] “Wall street slammed after china stock woes,” *Fox News, Associated Press*, 2007.
- [23] B. Powell, “Behind china’s stock meltdown,” *TIME Business and Money*, 2007.
- [24] C. H. Smith, “China is not another ascendant superpower, it’s just another nation with structural problems,” *Business Insider, Finance*, 2009.
- [25] M. Shedlock, “China facing economic, financial and stock market crash scenario,” *The Market Oracle*, 2009.
- [26] Spearman C , “The proof and measurement of association between two things”, *The American journal of psychology, JSTOR*, vol. 15, no. 1, pp. 72–101, 1904.
- [27] Kendall, Maurice G , “A new measure of rank correlation”, *Biometrika, JSTOR*, Vol. 30, No. 1/2 (Jun), pp. 81–93, 1938.
- [28] S. K. Lai, Y.-T. Lin, P. J. Hsu, and S.A. Cheong, “Dynamical study of metallic clusters using the statistical method of time series clustering”, *Computer Physics Communications*, vol. 182, pp. 1013–1026, 2011.
- [29] D.-H. Kim and H. Jeong, “Systematic analysis of group identification in stock markets”, *Physical Review E*, vol. 72, 046133, 2005.
- [30] J. M. Lattin, J. D. Carroll, and P. E. Green, *Analyzing multivariate data*. Thomson Brooks/Cole Pacific Grove, CA, 2003.
- [31] J. E. Gentle, *Numerical Linear Algebra with Applications in Statistics*. Springer, 1998.
- [32] J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertesz, “Dynamic asset trees and black monday,” *Physica A: Statistical Mechanics and its Applications*, vol. 324, no. 1, pp. 247–252, 2003.
- [33] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, and A. Kanto, “Asset trees and asset graphs in financial markets,” *Physica Scripta*, vol. 2003, no. T106, pp. 48–54, 2003.
- [34] F. Pozzi, T. Di Matteo, and T. Aste, “Spread of risk across financial markets: better to invest in the peripheries,” *Scientific reports*, vol. 3, Article No. 1665, 2013.

- [35] M. Tumminello, T. Aste, T. Di Matteo, and R. Mantegna, “A tool for filtering information in complex systems,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 30, pp. 10421–10426, 2005.
- [36] W.-M. Song, T. Di Matteo, T. Aste, “Hierarchical information clustering by means of topologically embedded graphs”, *PLoS ONE*, vol. 7, Issue. 3, e31929, 2012.
- [37] T. Di Matteo, F. Pozzi, T. Aste, “The use of dynamical networks to detect the hierarchical organization of financial market sectors”, *The European Physical Journal B*, vol. 73, no. 1, pp. 3–11, 2010.
- [38] L. Kaufmann and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, 2009.
- [39] B. Mirkin, *Mathematical Classification and Clustering: From How to What and Why*, Springer, 1998.
- [40] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, SIAM, 2007.
- [41] M. Rosvall, C.T Bergstrom, “Maps of random walks on complex networks reveal community structure”, *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.
- [42] F. Radicchi, José J. Ramasco, S. Fortunato, “Information filtering in complex weighted networks”, *Physical Review E*, vol. 83, no. 4, 046101, 2011.
- [43] F.B. Baker, “Numerical taxonomy for educational researchers”, *Review of Educational Research*, vol. 42, no. 3, pp. 345–358, 1972.
- [44] P.H.A. Sneath, “The application of computers to taxonomy”, *Journal of General Microbiology*, vol. 17, pp. 201–226, 1957.
- [45] S.C. Johnson, “Hierarchical clustering schemes”, *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.
- [46] R. N. Mantegna and H. E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance*, Cambridge University Press, 2000.
- [47] T. Mizuno, H. Takayasu, M. Takayasu, “Correlation networks among currencies”, *Physica A: Statistical Mechanics and its Applications*, vol. 364, pp. 336–342, 2006.
- [48] Song D.M, Tumminello M, Wei-Xing Z, Mantegna R.N., “Evolution of worldwide stock markets, correlation structure, and correlation-based graphs”, *Physical Review E*, vol. 84, no. 2, pp. 26–108, 2011.
- [49] Marsili M, “Dissecting financial markets: sectors and states”, *Quantitative Finance*, vol. 2, no. 4, pp. 297–302, 2002.

- [50] Bonanno G, Caldarelli G, Lillo F, Micciché S, Vandewalle N, Mantegna R.N., “Networks of equities in financial markets”, *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 38, no. 2, pp. 363–371, 2004.
- [51] Tumminello M, Coronello C, Lillo F, Micciche S, Mantegna R.N., “Spanning trees and bootstrap reliability estimation in correlation-based networks”, *International Journal of Bifurcation and Chaos*, vol. 17, no. 7, pp. 2319–2329, 2007.
- [52] Onnela, J.P, Kaski K, Kertész J, “Clustering and information in correlation based financial networks”, *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 38, no. 2, pp. 353–362, 2004.
- [53] Boginski V, Butenko S, Pardalos P.M., “Statistical analysis of financial networks”, *Computational statistics & data analysis*, vol. 48, no. 2, pp. 431–443, 2005.
- [54] Tumminello M, Di Matteo T, Aste T, Mantegna R.N., “Correlation based networks of equity returns sampled at different time horizons”, *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 55, no. 2, pp. 209–217, 2007.
- [55] Coelho R, Hutzler S, Repetowicz P, Richmond P, “Sector analysis for a FTSE portfolio of stocks”, *Physica A: Statistical Mechanics and its Applications*, vol. 373, pp. 615–626, 2007.
- [56] Jung W.S., Kwon O, Wang F, Kaizoji T, Moon H.T., Stanley H.E., “Group dynamics of the Japanese market”, *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 2, pp. 537–542, 2008.