Detecting a Currency's Dominance or Dependence using Foreign Exchange Network Trees

Mark McDonald[1], Omer Suleman[2], Stacy Williams[3], Sam Howison[1] and Neil F. Johnson[2]
[1]*Mathematics Department, Oxford University, Oxford, OX1 2EL, U.K.*
[2]*Physics Department, Oxford University, Oxford OX1 3PU, U.K.*
[3]*FX Research and Trading Group, HSBC Bank, 8 Canada Square, London E14 5HQ, U.K.*

(Dated: March 11, 2005)

In a system containing a large number of interacting stochastic processes, there will typically be many non-zero correlation coefficients. This makes it difficult to either visualize the system's inter-dependencies, or identify its dominant elements. Such a situation arises in Foreign Exchange (FX) which is the world's biggest market. Here we develop a network analysis of these correlations using Minimum Spanning Trees (MSTs). We show that not only do the MSTs provide a meaningful representation of the global FX dynamics, but they also enable one to determine momentarily dominant and dependent currencies. We find that information about a country's geographical ties emerges from the raw exchange-rate data. Most importantly from a trading perspective, we discuss how to infer which currencies are 'in play' during a particular period of time.

PACS numbers: 89.75.Fb, 89.75.Hc, 89.65.Gh

I. INTRODUCTION

There is enormous interest in the properties of complex networks [1–3]. There has been an explosion of papers within the physics literature analyzing the structural properties of biological, technological and social networks; the main results of which are summarized in [3]. Such networks or 'graphs', contain n nodes or 'vertices' $\{i\}$ connected by M connections or 'edges'. In the case of physical connections, such as wires or roads, it is relatively easy to assign a binary digit (i.e. 1 or 0) to the edge between any two nodes i and j according to whether the corresponding physical connection exists or not. However, for social networks such as friendship networks [3], and biological networks such as reaction pathways [3], the identification of network connections is less clear. In fact it is extremely difficult to assign any particular edge as being a definite zero or one – instead, all edges will typically carry a weighting value ρ_{ij} which is analog rather than binary, and which is in general neither equal to zero nor to one. The analysis of such weighted networks is in its infancy, in particular with respect to their functional properties and dynamical evolution [4]. The main difficulty is that the resulting network is fully-connected with $M = n(n-1)$ connections between all n nodes. In symmetric situations where $\rho_{ij} \equiv \rho_{ji}$, this reduces to $M = n(n-1)/2$ connections, but is still large for any reasonable n.

An interesting example of such a fully-connected weighted network is provided by the set of correlation coefficients between n stochastic variables. Each node i corresponds to the stochastic variable $x_i(t)$ where $i = 1, 2, \ldots, n$, and each of the $n(n-1)/2$ connections between pairs of nodes carries a weight given by the value of the correlation coefficient ρ_{ij} (see definition below). For any reasonable number of nodes the number of connections is very large (e.g. for $n = 110$, $n(n-1)/2 = 5995$) and hence it is extremely difficult to deduce which correlations are most important for controlling the overall dynamics of the system. Indeed, it would be highly desirable to have a simple method for deducing whether certain nodes, and hence a given subset of these stochastic processes, are actually 'controlling' the correlation structure [5]. In the context of financial trading, such nodal control would support the popular notion among traders that certain currencies can be 'in play' over a given time period. Clearly such information could have important practical consequences in terms of understanding the overall dynamics of the highly-connected FX market. It could also have practical applications in other areas where n inter-correlated stochastic process are operating in parallel.

With this motivation, we present here an analysis of the correlation network in an important real-world system, namely the financial currency (i.e. FX) markets. Although the empirical analysis presented is obtained specifically for this financial system, the analysis we provide has more general relevance to any system involving n stochastic variables and their $n(n-1)/2$ correlation coefficients. There is no doubt that currency markets are extremely important [6] – indeed, the recent fall in the value of the dollar against other major currencies is quite mysterious, and has attracted numerous economic 'explanations' to reason away its dramatic decline. The currency markets, which represent the largest market in the world, have daily transactions totalling trillions of dollars, exceeding the yearly GDP (Gross Domestic Product) of most countries.

The technical approach which we adopt, is motivated by recent research within the physics community by Mantegna and others [7–14] and concerns the construction and analysis of Minimum Spanning Trees (MST), which contain only $n-1$ connections. Mantegna and co-workers focused mainly on equities – by contrast, we consider the case of FX markets and focus on what the time-dependent properties of the MST can tell us about the FX market's evolution. In particular, we investigate the stability and time-dependence of the resulting MST, and introduce a methodology for inferring which currencies are 'in play' by analyzing the clustering and leadership structure within the MST network.

The application of MST analysis to financial stock (i.e. equities) was introduced by the physicist Rosario Mantegna [7]. The MST gives a 'snapshot' of such a system; however, it is the temporal evolution of such systems, and hence the evolution of the MSTs themselves, which motivates our research. In a series of papers [10–12], Onnela et al. extended Mantegna's work to investigate how such trees evolve over time in equity markets. Here we follow a similar approach for FX markets. One area of particular interest in FX trading – but which is of interest for correlated systems in general – is to identify which (if any) of the currencies are 'in play' during a given period of time. More precisely, we are interested in understanding whether particular currencies appear to be assuming a dominant or dependent role within the network, and how this changes over time. Since exchange rates are always quoted in terms of the price of one currency compared to another, this is a highly non-trivial task. For example, is an increase of the value of the euro versus the dollar primarily because of an increase in the intrinsic value of euro, or a decrease in the intrinsic value of the dollar, or both? We analyze FX correlation networks in an attempt to address such questions. We believe that our findings, while directly relevant to FX markets, could also be relevant to other complex systems containing n stochastic processes whose interactions evolve over time.

II. MINIMUM SPANNING TREE (MST)

Given a correlation matrix (e.g. of financial returns) a connected graph can be constructed by means of a transformation between correlations and suitably defined distances [8]. This transformation assigns smaller distances to larger correlations [8]. The MST, which only contains $n-1$ connections, can then be constructed from the resulting hierarchical graph [8, 15]. Consider n different time-series, x_i where $i \in \{1, 2, ...n\}$, with each time-series x_i containing N elements (i.e. N timesteps). The corresponding $n \times n$ correlation matrix C can easily be constructed, and has elements $C_{ij} \equiv \rho_{ij}$ where

$$\rho_{ij} \ = \ \frac{\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle}{\sigma_i \sigma_j} \qquad (1)$$

where $\langle ... \rangle$ indicates a time-average over the N datapoints for each x_i, and σ_i is the sample standard deviation of the time-series x_i. From the form of ρ_{ij} it is obvious that C is a symmetric matrix. In addition,

$$\rho_{ii} = \frac{\langle x_i^2 \rangle - \langle x_i \rangle^2}{\sigma_i^2} \ \equiv \ 1, \quad \forall \ i \qquad (2)$$

hence all the diagonal elements are identically 1. Therefore C has $n(n-1)/2$ independent elements. Since the number of relevant correlation coefficients increases like n^2, even a relatively small number of time-series can yield a correlation matrix which contains an enormous amount of information – arguably 'too much' information for practical purposes. By comparison, the MST provides a skeletal structure with only $n-1$ links, and hence attempts to strip the system's complexity down to its bare essentials. As shown by Mantegna, the practical justification for using the MST lies in its ability to provide economically meaningful information [7, 8]. Since the MST contains only a subset of the information from the correlation matrix, it cannot tell us anything which we could not (in principle) obtain by analyzing the matrix C itself. However, as with all statistical tools, the hope is that it can provide an insight into the system's overall behavior which would not be so readily obtained from the (large) correlation matrix itself.

To construct the MST, we first need to convert the correlation matrix C into a 'distance' matrix D. Following Refs. [7, 8], we use the non-linear mapping:

$$d_{ij}(\rho_{ij}) \ = \ \sqrt{2(1-\rho_{ij})} \qquad (3)$$

to get the elements d_{ij} of D [16]. Since $-1 \leq \rho_{ij} \leq 1$, we have $0 \leq d_{ij} \leq 2$. This distance matrix D can be thought of as representing a fully connected graph with edge weights d_{ij}. In the terminology of graph theory, a 'forest' is a graph where there are no cycles [17] while a 'tree' is a connected forest. Thus a tree containing n nodes must contain precisely $n-1$ edges [3, 17]. The minimum spanning tree $\mathbf{T}$ of a graph is the tree containing every node, such that the sum $\sum_{d_{ij} \in \mathbf{T}} d_{ij}$ is a minimum. There are two methods for constructing the MST — Kruskal's algorithm and Prim's algorithm [9]. We used Kruskal's algorithm, details of which are given in [18].

CLUSTER ANALYSIS

Whilst the impetus for this research came from the MST work of Mantegna and colleagues in the econophysics community, the task of finding a hierarchical clustering of a set of timeseries falls firmly within the established field of cluster analysis. There are two distinct steps necessary in a cluster analysis. First one must define a meaningful distance between the objects one wishes to cluster (the *distance measure*), then one can implement a clustering procedure to group the objects together. An introduction to the most common distance measures and clustering methods is given in [19], which also contains evidence that the choice of clustering procedure has more effect on the quality of the clustering than the distance measure chosen.

The clustering procedure used to form the MST is known in cluster analysis as the *single-linkage* clustering method (also known as the *nearest-neighbour* technique) [20, 21]. This is the simplest of an important group of clustering methods known collectively as Agglomerative Hierarchical Clustering methods. The main problem with the MST (single-linkage method) is that it has a tendency to link poorly clustered groups into 'chains' by successively joining them through their nearest neighbours. Hence, one would expect the hierarchy produced by the MST to represent larger distances (anti-correlated) less reliably than the smaller distances (highly correlated). Since we are attempting to identify highly-clustered groups this will not be a problem. However, in other situations — for example, if one were attempting to use an MST to identify poorly correlated or anti-correlated stocks for use in portfolio theory — it may be preferable to use a more sophisticated clustering method.

III. DATA

1. Raw Data

The empirical currency data that we investigated are hourly, historical price-postings from HSBC Bank's database for nine currency pairs together with the price of gold from 01/04/1993 to 12/30/1994 [22]. Gold is included in the study because there are similarities in the way that it is traded, and in some respects it resembles a very volatile currency. The currency pairs under investigation are AUD/USD, GBP/USD, USD/CAD, USD/CHF, USD/JPY, GOLD/USD, USD/DEM, USD/NOK, USD/NZD, USD/SEK [23]. In the terminology used in FX markets [23], USD/CAD is counter-intuitively the number of Canadian dollars (CAD) that can be purchased with one US dollar (USD). We must define precisely what we mean by hourly data, as prices are posted for different currency pairs at different times. We do not want to use average prices since we want the prices we are investigating to be prices at which we could have executed trades. Hence for hourly data, we use the last posted price within a given hour to represent the (hourly) price for the following hour.

We emphasize that the n stochastic variables which we will analyze correspond to currency *exchange* rates and hence measure the *relative* values of any two currencies. It is effectively meaningless to ask the *absolute* value of a given currency, since this can only ever be measured with respect to some other financial good. Thus each currency pair corresponds to a node in our network. We are concerned with the correlations between these currency exchange rates, each of which corresponds to an edge between two nodes. A given node does *not* correspond to a single currency.

2. Data Filtering

As with all real-world systems, the issue of what constitutes correct data is complicated. In particular, there are some subtle data-filtering (or so-called 'data-cleaning') issues which need to be addressed. Such data problems are, by contrast to the physical sciences, a reality in most disciplines which deal with human timescales and activity. In our specific case, we are interested in calculating both the instantaneous and lagged correlations between exchange-rate returns. Hence it is neccessary to ensure that (a) each time series has an equal number of posted prices; (b) the n'th posting for each currency pair corresponds, to as good an approximation as possible, to the price posted at the same timestep t_n for all $n \in \{1, ..., N\}$. For some of the hourly timesteps, some currency pairs have missing data. The best way to deal with this is open to interpretation. Is the data missing simply because there has been no price change during that hour, or was there a fault in the data-recording system? Looking at the data, many of the missing points do seem to occur at times when one might expect the market to be illiquid. However, sometimes there are many consecutive missing data points — even an entire day. This obviously reflects a fault in the data recording system. To deal with such missing data we adopted the following protocol. The FX market is at its most liquid between the hours of 08:00 and 16:00 GMT [24]. In an effort to eradicate the effect of 'zero returns' due to a lack of liquidity in the market – as opposed to the price genuinely not moving in consecutive trades – we only used data from between these hours [25]. Then, if the missing data were for fewer than three consecutive hours, the missing prices were taken to be the value of the last quoted price. If the missing data were for three or more consecutive hours, then the data for those hours were omitted from the analysis. Since we must also ensure completeness of the data at each point, it is then necessary that the data for those hours are omitted from *all* currency pairs under investigation [26]. We believe that this procedure provides a sensible compromise between the conflicting demands of incorporating all relevant data, and yet avoiding the inclusion of spurious zero-returns which could significantly skew the data.

3. Stationarity

Once we have cleaned up the exchange-rate data for each currency pair, whose associated price we will henceforth label as $P_i(t)$, we turn this value into a financial return

$$r_i(t_n) = \ln\left(\frac{P_i(t_{n+1})}{P_i(t_n)}\right). \tag{4}$$

Since we will be calculating correlations, we need to be confident that our return distributions are stationary. A useful probe of stationarity is to calculate the autocorrelation for each time-series. A stationary time-series will have an autocorrelation function which rapidly decays to zero [27], whereas a non-stationary time-series will have an autocorrelation function which decays to zero very slowly (if at all). The autocorrelation is defined as

$$\rho_i(\tau) = \frac{\langle x_i(t+\tau)x_i(t)\rangle - \langle x_i(t+\tau)\rangle\langle x_i(t)\rangle}{\sigma_{i,\tau}\sigma_i} \tag{5}$$

where $\langle ...\rangle$ indicates a time-average over the $N-\tau$ elements and $\sigma_{i,\tau}$, σ_i are the sample standard deviations of the time-series $x_i(t+\tau)$ and $x_i(t)$, respectively. To illustrate this point, Figure 1 shows the autocorrelation of both the price and return for the currency pair AUD/USD over the two year period of interest. It can be clearly seen that this confirms the assertion that the returns are stationary whilst the prices themselves are not. Thus we can, with confidence, calculate correlations between different currency-pair returns.

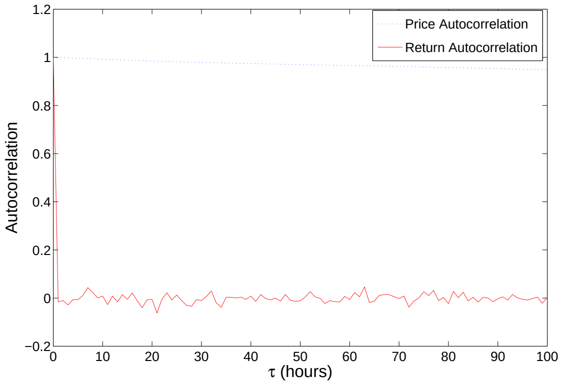

FIG. 1: (Color online) Autocorrelation of the price and return for the currency pair AUD/USD (i.e. the Australian dollar versus the US dollar).

4. Foreign Exchange Data

In addition to the problems outlined above which are common to the analyses of all such real-world data, there are further issues that are specific to FX data and which make the study of FX and equities fundamentally different. When producing the MST for the returns of the stock which make up the FTSE100 index, one calculates the returns from the values of the price of the stock *in the same currency* — specifically, UK pounds (GBP). With FX data, however, we are considering exchange rates between currency pairs. Thus should we consider GBP/USD or USD/GBP? And does it indeed make a difference which one we use? Since the correlation is constructed to be normalized and *dimensionless*, one might be tempted to think that it does not matter since the value of the correlation will be the same and only the sign will be different. However, it is important when constructing the MST since there is an asymmetry between how positive and negative correlations are represented as distances. In particular, the MST picks out the smallest distances — i.e. the highest correlation. A large negative correlation gives rise to a large distance between nodes. Thus a connection between two nodes will be missing from the tree even though it would be included if the other currency in the pair were used as the base currency.

Consider the following example. There is a large *negative* correlation between the returns of the two currency pairs GBP/USD and USD/CHF [28]. Conversely, if we put them both with USD as the base currency, we get a large *positive* correlation between USD/GBP and USD/CHF. Thus our choice will give rise to a fundamentally different tree structure. For this reason, we perform the analysis for all possible currency-pairs against each other. Since we are analysing ten currency pairs, this gives us eleven separate currencies and hence 110 possible currency pairs (and hence $n = 110$ nodes). However, there are constraints on these timeseries and hence an intrinsic structure is imposed on the tree by the relationships between the timeseries. This is commonly known as the 'triangle effect'. Consider the three exchange-rates USD/CHF, GBP/USD and GBP/CHF. The nth element of the timeseries for GBP/CHF is simply the product of the nth elements of USD/CHF and GBP/USD. This simple relationship between the timeseries gives rise to some relationships between the correlations. More generally, with three time-series $P_1(t)$,$P_2(t)$,$P_3(t)$ such that $P_3(t) = P_1(t)P_2(t)$, there exist relationships between the correlations and variances of the returns. If we define the returns r_i such that $r_i = \ln P_i$ for all i, then we have:

$$r_3 = r_1 + r_2\ . \tag{6}$$

Thus

$$\begin{aligned} \mathrm{Var}(r_3) &= \mathrm{Var}(r_1 + r_2) & (7)\\ &= \mathrm{E}((r_1+r_2)^2) - (\mathrm{E}(r_1+r_2))^2 & (8)\end{aligned}$$

For currency pairs, it is valid to assume that the expected value of the return is zero [29]. Hence this expression simplifies to

$$\begin{aligned} \sigma_3^2 &= \mathrm{E}(r_1^2 + r_2^2 + 2r_1 r_2) && (9)\\ &= \sigma_1^2 + \sigma_2^2 + 2\mathrm{Cov}(r_1, r_2) && (10)\\ &= \sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2\rho_{12} && (11)\end{aligned}$$

where σ_1, σ_2, σ_3 are the variances of the returns $r_1(t)$, $r_2(t)$, $r_3(t)$ while ρ_{12} is the correlation between the returns $r_1(t)$ and $r_2(t)$. Finally we obtain

$$\rho_{12} = \frac{\sigma_3^2 - (\sigma_1^2 + \sigma_2^2)}{2\sigma_1\sigma_2} \qquad (12)$$

Hence there is a structure forced upon the market by the triangle effect. This is not a problem since all the cross-rates we include in the tree *do* exist and the correlations calculated are the true correlations between the returns. Even though the values of these correlations have some relationships between them, they should be included in the tree since it is precisely this market structure that we are attempting to identify. We do, however, need to confirm that this structure which is being imposed on the market is not dominating our results.

5. Checking Data Validity

We performed a simple check on the data by calculating the minimum and maximum return for each currency pair. For example, if the rate for USD/JPY was entered as 1.738 instead of 173.8 then this would give rise to returns of approximately ±1. As a result of this check, we could confirm that there were no such errors in our dataset. We then drew scatter plots of the returns against time, plus histograms of the return distribution in order to check that there were no unusual points on the graph. The next check that we performed is slightly more subtle. The correlation between two variables is related to the gradient of the regression line between the variables [27]. However, this gradient is very susceptible to outliers so we need to ensure that our data does not contain points which are sufficiently outlying as to justify deletion. To check this, one can plot scatter-graphs of returns from different currency pairs against each other. Since we have 110 currency pairs, there are 5995 possible scatter-graphs (one for each of the $n(n-1)/2$ independent correlation coefficients). Obviously it is unfeasible to plot all of these. However, as we have seen, all the cross-rates are formed from the USD rates. Therefore we can just plot the 45 scatter-graphs of USD/A against USD/B where A $\neq$ B. Figure 2 shows this scatter plot for USD/DEM vs USD/GBP. It can be seen that there are no such outlying points. This figure is typical of all the 45 possible plots.

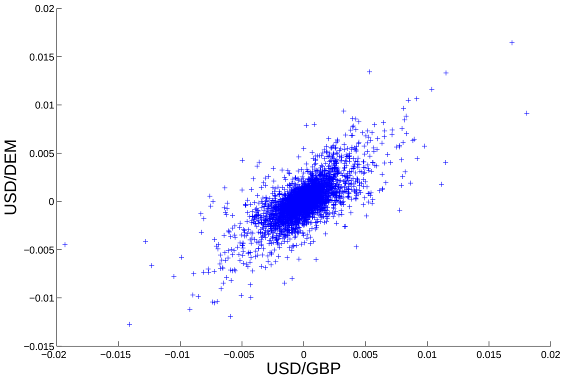

FIG. 2: (Color online) Scatter plot of returns for USD/DEM (i.e. the US dollar versus the German mark) against USD/GBP (i.e. the US dollar versus the UK pound)

IV. DIRECTED TREES

In [30], the Minimum Spanning Tree approach was generalized by considering a directed graph. Lagged correlations were investigated in an attempt to determine whether the movement of one stock price 'preceded' the movement in another stock price. We now investigate whether this approach yields useful results here. First we should define what we mean by lagged correlation. If we have two time-series, $x_i(t)$ and $x_j(t)$ where both time-series contain N elements, the τ-lagged correlation is given by

$$\rho_{ij}(\tau) = \frac{\langle x_i(t+\tau)x_j(t)\rangle - \langle x_i(t+\tau)\rangle\langle x_j(t)\rangle}{\sigma_{i,\tau}\sigma_j} \qquad (13)$$

where $\langle ...\rangle$ indicates a time-average over the $N-\tau$ elements and $\sigma_{i,\tau}$, σ_j are the sample standard deviations of the time-series $x_i(t+\tau)$ and $x_j(t)$, respectively. Note that autocorrelation is simply the special case of this where $i = j$. Armed with this definition, we can now look at our data to see whether there are any significant lagged correlations between returns of different currency pairs. Figure 3 shows the lagged correlation between the returns of each pair of currencies when the prices of those currencies are given with GBP as the base currency. In the figure, AUD vs USD (lagged) refers to the lagged correlation between GBP/AUD (at time $t+\tau$) and GBP/USD (at time t). The results in this figure are representative of the results from all currency pairs included in our study.

Figure 3 clearly shows that the approach considered in [30] will not yield anything useful here for FX. If such lagged correlations do exist between currency pairs, they occur over a timescale smaller than one hour. In other words, the FX market is very efficient. This should not come as a surprise — the FX market is approximately 200 times as liquid as the equities market [6].

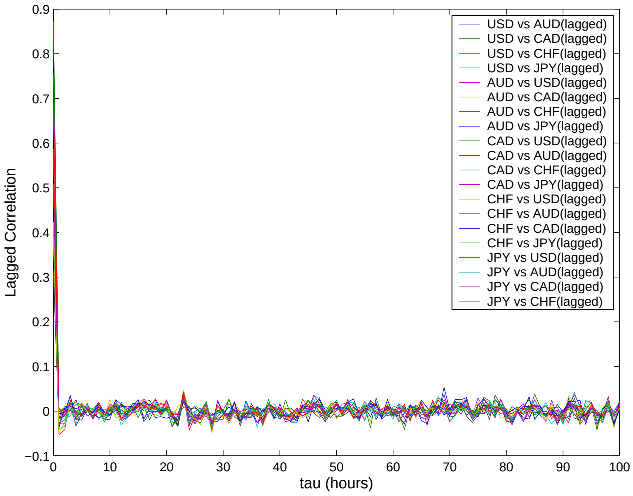

FIG. 3: (Color online) Lagged correlation between different currency pairs when GBP is the base currency. As explained in the text, AUD vs USD (lagged) refers to the lagged correlation between GBP/AUD (at time $t+\tau$) and GBP/USD (at time t)

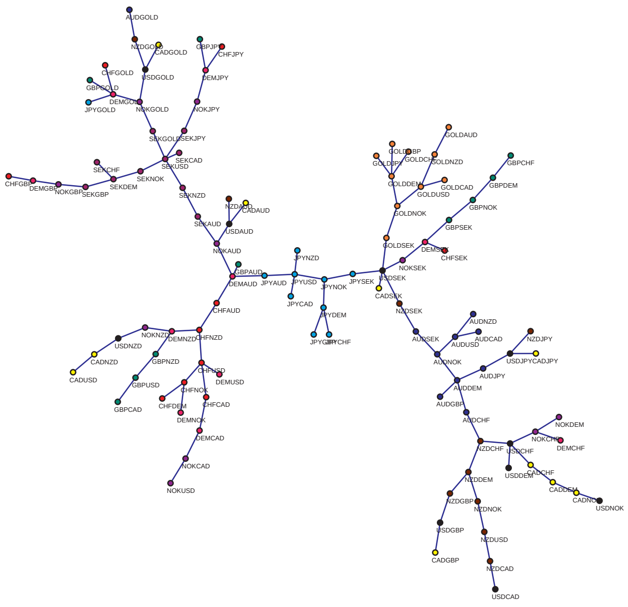

FIG. 4: (Color) The Minimum Spanning Tree representing the correlations between all hourly cross-rate returns from the years 1993 and 1994.

V. THE CURRENCY TREE

Creating all the possible cross-rates from the 11 currency pairs gives rise to a total of $n = 110$ different time-series. It is here that the approach of constructing the MST comes into its own, since 110 different currencies yields an enormous correlation matrix containing 5995 separate elements. This is far too much information to allow any practical analysis by eye. However, as can be seen from Figure 4, the hourly FX tree is quite easy to look at. Rather than a mass of numbers, we now have a graphical representation of the complex system in which the structure of the system is visible.

Before analysing the tree in detail, it is instructive to consider first what effect the constraints of Eq. (12) (the 'triangle effect') will have on the shape of the tree. Figure 5 illustrates this. The data used in this figure is the same data as in Fig. 4, however the price-series for the currencies in USD were randomized before the cross-rates were formed. This process gives prices for the various currencies in USD which are random, and will hence have negligible correlation between their returns. Thus Fig. 5 shows the structure forced on the tree by the triangle effect.

This tree resulting from randomizing data as described above, is actually very different in character from the true tree in Figure 4. At first glance it might appear that some aspects are similar — currencies show some clustering in both cases. However, in the tree of real cross-rates there are currency-clusters forming about any node, whereas in Figure 5 there are only clusters centred on the USD node. This is not surprising: after all, what do the 'CHF/everything' rates all have in common in the case of random prices other than the CHF/USD rate? The best way to interpret Fig. 5 is that we have a tree of USD nodes (which are spaced out since their returns are poorly correlated) and around these nodes we have clusters of other nodes which have the same base currency, and which are effectively the information from the USD node plus noise. This exercise shows us that the MST results are not dominated by the triangle effect. In an effort to show this in a more quantitative way, we investigate the proportion of links that are present in both trees. Less than one third of the edges in Fig. 4 are present in Fig. 5.

Another more quantitative comparison is to compare the degree distribution of the tree from the random price series with that of the tree from real price data. This is shown graphically in Fig. 6. Again, this further highlights the differences between the two trees.

Now that we have produced the tree, how does one interpret it? Despite the initial impression, the tree is actually very easy to interpret. It contains coloured nodes, each of which represents a particular currency pair. For the reasons explained earlier, currency pairs are quoted both ways round: USD/JPY appears with USD as the base currency, as is normal market convention, but so does JPY/USD. This gives all currencies the chance to stand out as a cluster, as will be seen shortly. The currency-pair nodes are each colour coded, according to the labelled base currency. Broadly speaking, each node is linked to the nodes representing the currency-pairs to which it is most closely correlated. The observation that

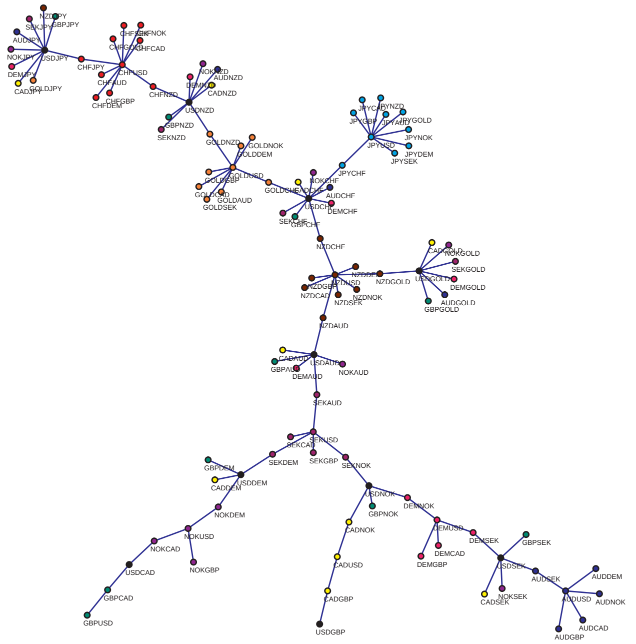

FIG. 5: (Color) The Minimum Spanning Tree formed from randomized data for the USD prices. This shows only the structure imposed on the tree by the triangle effect

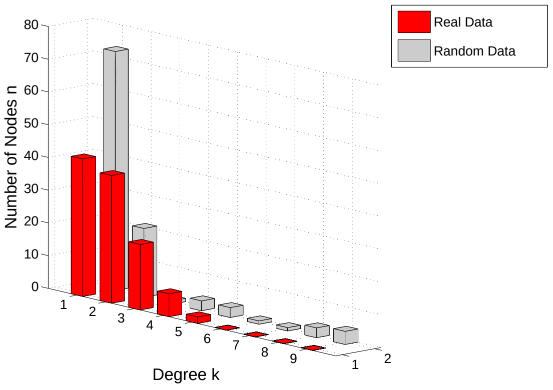

FIG. 6: (Color online) Comparison of the degree distributions for the trees shown in Figure 4 (Real Data) and Figure 5 (Randomized Data)

certain currency-pairs cluster together means that they have been moving together consistently over the monitored period.

The most interesting feature of Figure 4 is the clustering of nodes which have the same base currency. For example, one can see a cluster of 9 AUD nodes. This observation demonstrates that over this two year period, the Australian Dollar has been moving systematically against a range of other currencies during this time. To use the prevailing industry term, the AUD is 'in play'. The same is also true for the SEK, JPY and GOLD clusters.

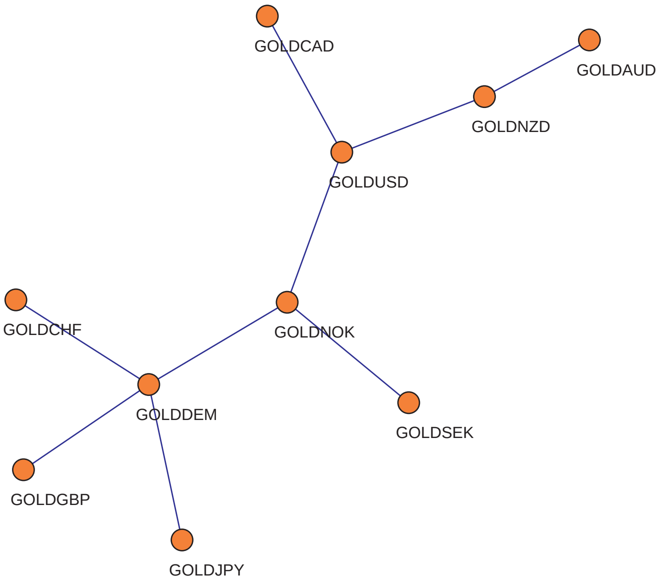

FIG. 7: (Color online) The cluster of GOLD exchange-rates from Fig. 4

It is encouraging that the cluster of Gold exchange rates links currencies in a sensible way. This cluster is re-drawn in Figure 7. It can be seen that the nodes in this cluster are grouped in an economically meaningful way: remarkably, there is a geographical linking of exchange-rates. The Australisian nodes, AUD and NZD, are linked, as are the American ones (USD and CAD). The Skandinavian currencies, SEK and NOK, are also linked. Finally, there is a European cluster of GBP, CHF and EUR. This provides a useful check that our results are sensible. Indeed if such geographical clustering had not arisen, it would be a good indication that something was wrong with our methodology.

Now that it is possible to identify clusters of currencies, we would like to quantify how clustered they are. This can be done by finding the level one has to partition the hierarchical tree associated with the MST [15] to get all the nodes with, for example, USD as the base currency into the same cluster. This results in a *self-clustering distance* for each currency. The smaller this distance is, the more tightly all the nodes for that currency are clustered. An alternative way to think of this is as the maximum ultrametric distance between any two nodes for that currency.

We are now in a position to compare the results produced by the MST with those from the original distance matrix. Let us compare the self-clustering distance for each currency with the maximum Euclidean Distance between any two nodes with that base currency and also with the average Euclidean Distance between all nodes with that base currency. This is shown in Figure 8. It can be seen that the agreement between the two results

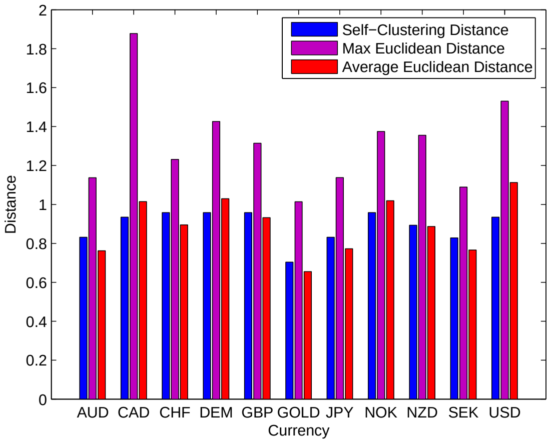

FIG. 8: (Color) Comparison of the results from the MST with those from the original correlation matrix

is very good. Not only does the MST rank the clusters in the same way as the original distance matrix does, it gives results which agree better with the average distance than with the maximum Euclidean distance. Hence, the results for the MST and the original distance matrix are not only in agreement, the MST results are also robust with respect to a single, large edge being contained between two nodes with the same base currency. As mentined above, the MST has the advantage over standard network representations since it only requires $n-1$ connections.

VI. STABILITY AND TEMPORAL EVOLUTION OF THE CURRENCY TREE

We now investigate the single-step survival ratio of the edges

$$\sigma_{\delta t} = \frac{|E_t \cap E_{t+\delta t}|}{|E|} \qquad (14)$$

where E_t and $E_{t+\delta t}$ represent the set of edges present in the trees formed from a dataset of length T=1000 hours [31] beginning at times t and $t+\delta t$ respectively, in order to see how this ratio depends on the value chosen for δt. This ratio must tend to one as δt approaches 0 for our results to be meaningful. The results are plotted in Fig. 9 and it can be seen that it is indeed the case that this ratio tends to one as δt approaches 0. Thus the topology of the MST is stable.

Next we investigate the time-dependence of the tree. Onnela [9] defined the k multi-step survival ratio to be:

$$\sigma_{\delta t,k} = \frac{|E_t \cap E_{t+\delta t} \cap \ldots \cap E_{t+k\delta t}|}{|E|} \qquad (15)$$

Thus if a link disappears for only one of the trees in the time t to $t+\delta t$ and then comes back, it is not

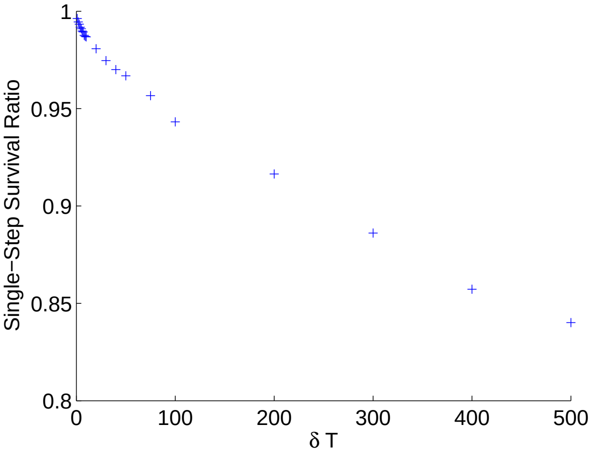

FIG. 9: (Color online) Single Step survival ratio as a function of δT

counted in this survival ratio. This seems a possibly overly-restrictive definition which might underestimate the survival. We will therefore also consider the more generous definition

$$\sigma_{\delta t,k} = \frac{|E_t \cap E_{t+k\delta t}|}{|E|} \quad . \qquad (16)$$

This quantity will, for large values of k, include cases where the links disappear and then come back several timesteps later. It therefore tends to overestimate the survival since a reappearance after such a long gap is more likely to be caused by a changing structure than by a brief, insignificant fluctuation.

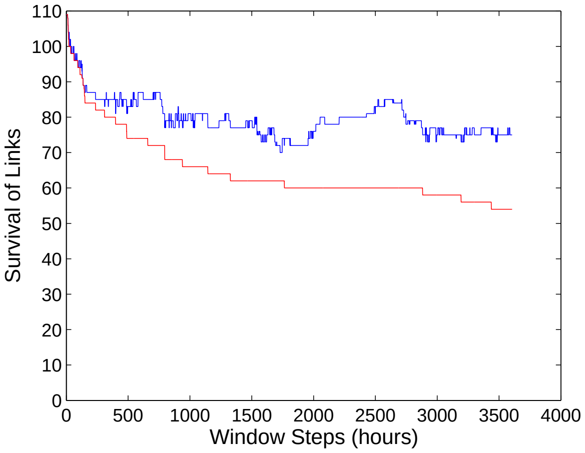

FIG. 10: (Color) Multi-step survival ratio of the FX tree's connections, as a function of time. The graph shows the two definitions described in the text, which tend to overestimate (blue) and underestimate (red) the survival effect.

Figure 10 shows both definitions, and uses a time-window of length $T = 1000$ hours and a time-step $\delta t = 1$

hour. It can be seen from the figure that the two lines form a ‘corridor’ for the multi-step survival ratio. This is because the over-restrictive definition of Eq. (15) underestimates the survival and the over-generous definition of Eq. (16) over-estimates the result. It is particularly noteworthy that even with the over-restrictive definition of Eq. (15), the survival of links after the end of two years is only just below fifty percent (i.e. 54/109). In other words, there are strong correlations existing between exchange-rate returns that are extremely long-lived.

VII. INTERPRETATION OF TREES FROM RECENT DATA

We know from our analysis that clusters occur in the MST, and that these clusters change over time. Next we illustrate the significance of this in practice. We will approach this by analyzing two trees which are one calendar month apart.

Figure 11 shows an example of a currency tree from a period in June 2004. Figure 11 shows a strong, brown, NZD cluster near the bottom of the tree. The NZD is ‘in play’, to use the prevailing industry term introduced above. The self-clustering distance for NZD is 0.845. The same is true for the yellow-coloured Canadian dollar (CAD), which has also formed a cluster, with a self-clustering distance of 0.932. Other clusters are also evident, including a red Swiss franc (CHF) cluster, which has formed near the top of the tree. In contrast, the Sterling currency-pairs are dispersed around the tree, indicating that there is little in common in their behaviour. In short, Sterling is not ‘in play’.

If the trees were static, this would be the end of the story. However, we have already shown that the trees do change over time. Figure 12 shows the equivalent currency-tree one month later. The CAD cluster is still evident and, in fact, has strengthened: all nine CAD nodes are linked together and the self-clustering distance is now at the smaller value of 0.808. The NZD cluster is still evident, this time near to the top of the Figure and is, in fact, slightly stronger with a self-clustering distance of 0.78. More interesting are the clusters which have changed. The CHF cluster has completely disintegrated; the CHF nodes are scattered over the tree. Hence the Swiss franc is no longer in play. Conversely, there is now a American dollar (USD) cluster which has formed, indicating that the dollar has become more important in determining currency moves.

In short, it has become possibly to identify currencies which are actively in play and are effectively dominating the FX market. Sometimes, when currencies are in play, it will be obvious to traders: for example, when there is a large and sustained USD move. However, this is not always the case, and our currency-trees are able to provide an indication of how important (i.e. how much in play) a particular currency is. In addition to using the tree as a graphical tool, it is possible to quantify how clustered a particular currency is by calculating the self-clustering distance.

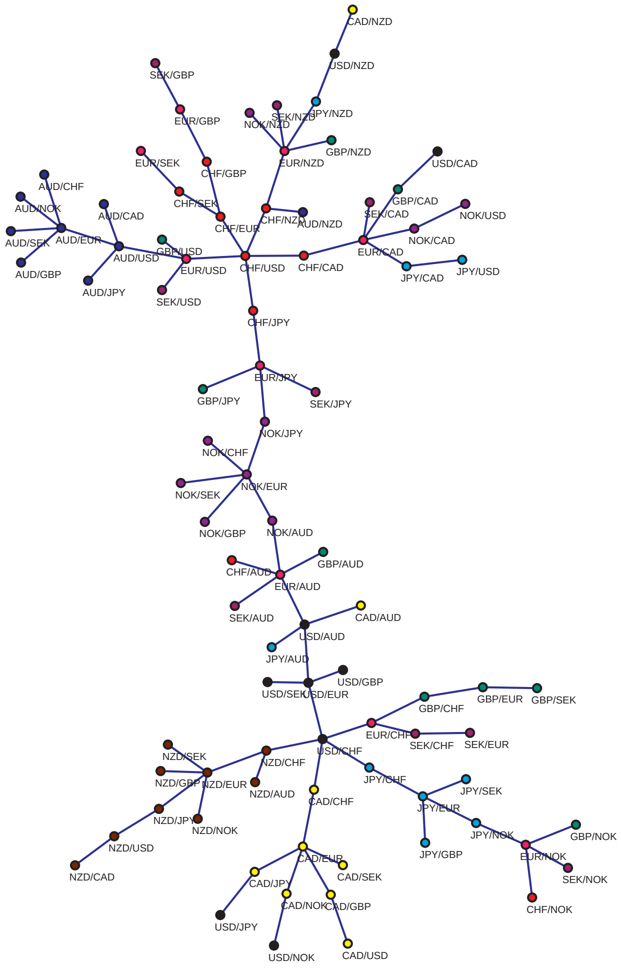

FIG. 11: (Color) Currency tree (MST) for a 2 week period in June 2004.

VIII. CONCLUSIONS

We have provided a detailed analysis of the Minimal Spanning Trees associated with empirical Foreign Exchange data. This analysis has highlighted various data-related features which make this study quite distinct from earlier work on equities.

We have shown that there is a clear difference between the currency trees formed from real markets and those formed from randomized data. For the trees from real markets, there is a clear regional clustering. We have also investigated the time-dependence of the trees. Even though the market structure does change rapidly enough

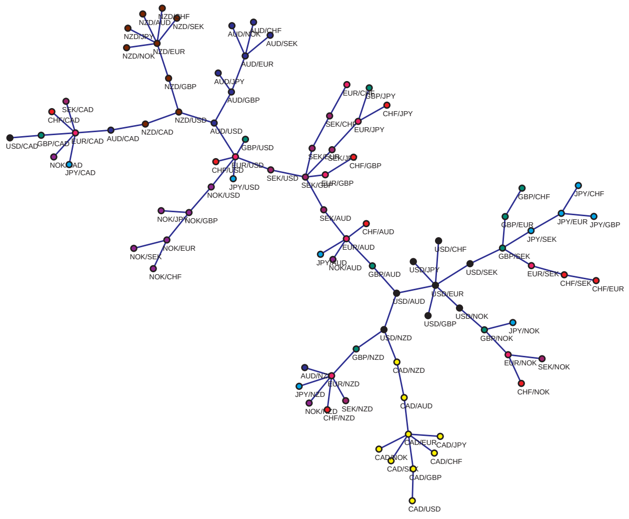

FIG. 12: (Color) Currency tree (MST) for a 2 week period in July 2004.

to identify changes in which currency-pairs are clustering together, there are links in the tree which last over the entire two year period. This shows that there is a certain robust structure to the FX markets. We have also developed a methodology for interpreting the trees which has practical applications: the trees can be used to identify currencies which are in play. Whilst this does not have predictive power, it helps one to identify more accurately the state the market is currently in. Armed with this information, one can be more confident of the predictions made from other models. In future work, we will look at trying to isolate the effect of news on the FX market – in other words, the extent to which external news 'shakes' the FX tree. Of particular interest is whether particular clusters have increased robustness over others, or not. In addition, we shall be investigating how tree structure depends on the frequency of the data used.

[1] D.J. Watts and S.H. Strogatz, Nature **393**, 440 (1998).

[2] D.S. Callaway, M.E.J. Newman, S.H. Strogatz, and D.J. Watts, Phys. Rev. Lett. **85**, 5468 (2000).

[3] S.N. Dorogovtsev and J.F.F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).

[4] See www.unifr.ch/econophysics for an up-to-date listing of all network papers and preprints.

[5] We assume that the correlation structure is specified by the correlation coefficients.

[6] P. Evans, *Introduction to Treasury and Capital Markets*, HSBC Bank Training Manual (2003).

[7] R.N. Mantegna, Eur. Phys. Jour. B **11**, 193 (1999).

[8] R.N. Mantegna and H.E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, Cambridge, 2000).

[9] J.-P. Onnela, *Taxonomy of Financial Assets*, M.Sc. Thesis, Department of Electrical and Communications Engineering, Helsinki University of Technology (2002).

[10] J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertesz, Phys. Scrip. T **106**, 48 (2003).

[11] J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertesz, Physica A **324**, 247 (2003).

[12] J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertesz, Phys. Rev. E **68**, 056110 (2003).

[13] G. Bonanno *et at.*, Phys. Rev. E **68**, 046130 (2003).

[14] G. Bonanno *et at.*, Eur. Phys. J. B **38**, 363 371 (2004).

[15] R. Rammel and G. Toulouse, and M.A. Virasoro, Rev. Mod. Phys. **58**, 765 (1986).

[16] For a discussion as to why this function is used as the definition of distance and not, say, $d_{ij} = 1 - \rho_{ij}$, see Chapter 13 of Ref. [8]. This formula simply approximates to the Standardized Euclidean Distance Metric between two time series x_i and x_j. The only reason this formula appears so counter-intuitive is because the result is given in terms of the correlation between the two time-series rather than as a sum over the elements of the series. In fact, for the MST, monotonic transformations of the distance function will give the same MST [21], with the edge weights all suitably transformed. Thus it makes no difference whether $d_{ij} = \sqrt{2(1 - \rho_{ij})}$ or $d_{ij} = 1 - \rho_{ij}$. We chose $d_{ij} = \sqrt{2(1 - \rho_{ij})}$ for compatability with previous econophysics MST work.

[17] B Bollobas, *Graph Theory, An Introductory Course*, (Springer-Verlag, 1979).

[18] T.H. Cormen, C.E. Leiserson, and R.L. Rivest, *Introduction to Algorithms*, (MIT Press, Cambridge MA, 1990).

[19] P. Hirst, *Cluster Analysis of Financial Securities*, M.Sc. Thesis, Oxford Centre for Industrial and Applied Mathematics, Oxford University (2003)

[20] L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data. An Introduction to Cluster Analysis*, (Wiley-Interscience, 1990)

[21] J.C. Gower and G.J.S. Ross, Applied Statistics, **18, 54–64 (1969)**

[22] This is over 12000 datapoints.

[23] The acronyms are standard ones in the FX market, where AUD means Australian dollar, GBP is UK pound, CAD is Canadian dollar etc. The exception is GOLD/USD which is quoted as XAU/USD in the markets – however for clarity, we have replaced the rather obscure symbol XAU with GOLD. Also, for the trees from 1993–1994 DEM is presetn in the tree; however, in the recent trees this currency has been replaced with EUR. For more information on terminology and functioning of the FX markets, see F. Taylor, *Mastering FX and Currency Options* (Financial Times Press, London, 1997).

[24] M.H. Jensen, A. Johansen, F. Petroni, and I. Simonsen, Inverse Statistics in the Foreign Exchange Market, Physica A 340 678 (2004).

[25] If there were no missing data points, this would give rise to a dataset of 4680 data points.

[26] This gave rise to the final dataset which contained 4608 datapoints. Therefore, the data cleaning procedure removed less than 2 percent of the 4680 possible datapoints.

[27] R.S. Pindyck, and D.L. Rubinfeld, *Econometric Models and Economic Forecasts*, (McGraw-Hill, New York, 1998).

[28] Even though quoting the two exchange-rates in this form seems slightly idiosyncratic, it is actually how the rates are quoted on the market. Most currencies are quoted with USD as the base currency. However GBP/USD is quoted in this unusual form because historically GBP was not based on a decimal system.

[29] We note that the interest-rate differentials are close to zero and the time-period of one hour is tiny. We also note that when price-changes are small, as discussed in Ref. [8], it makes little difference in terms of the overall behavior as to which precise definition of returns one uses.

[30] L. Kullmann, J. Kertesz, and K. Kaski, Phys. Rev. E 66 026125 (2002).

[31] T =1000 hourly datapoints corresponds to just under half a year of data.