

Degree in Mathematics

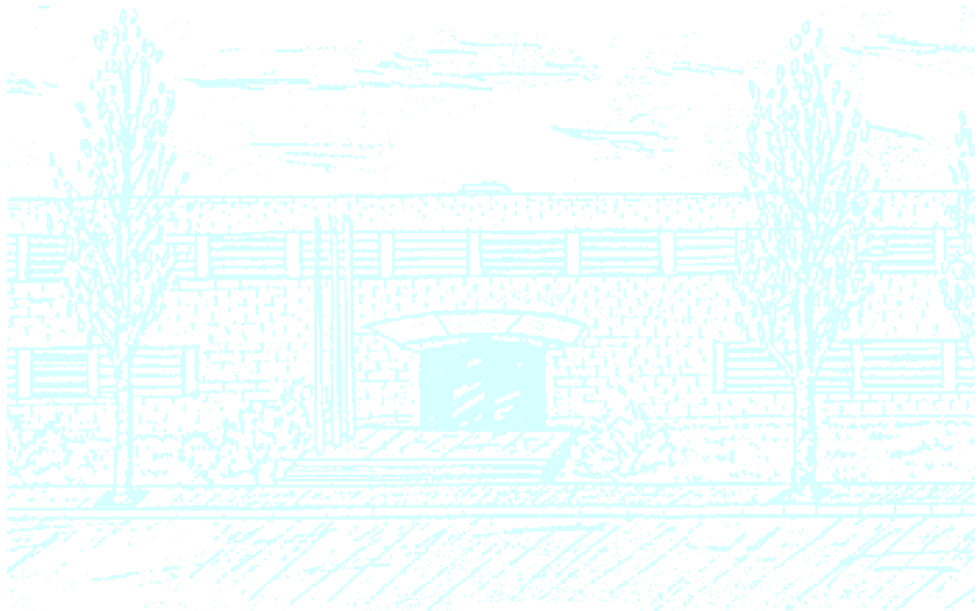
Title: Detecting Clusters and Their Dynamics in the Forex Market

Author: Martí Renedo Mirambell

Advisor: Argimiro Arratia Quesada

Department: Ciències de la computació

Academic year: 2015-2016



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Facultat de Matemàtiques i Estadística

Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Degree in Mathematics
Bachelor's Degree Thesis

Detecting Clusters and Their Dynamics in the Forex Market

Martí Renedo Mirambell

Supervised by Argimiro Arratia Quesada
Department of Computer Science & BGSMath

June, 2016

Thanks to Professor Argimiro Arratia for helping and supervising this project.

Abstract

This project studies and implements the clustering methods in [1] to detect correlations in the foreign exchange market. To deal with the potentially non linear nature of currency time series dependance, we propose two alternative similarity metrics to use instead of the Pearson linear correlation. We observe how each of them responds over several years of currency exchange data and find significant differences in the resulting clusters.

Keywords

Community detection, Distance correlation, Kendall correlation, FX market

1. Introduction

The foreign exchange market (Forex, or FX) is a decentralized worldwide market for the trading of currencies. With a global activity of \$5.3 trillion per day in 2013, it is the largest market in the world by a significant margin, which results in a very high liquidity. As opposed to other markets, the foreign exchange market operates 24 hours a day five days per week (from Sunday 22:00 GMT to Friday 22:00 GMT). This is due to the location of market centers in different time zones. It is worth noting, though, that the largest financial centres (United Kingdom, United States, Singapore and Japan) account for most of the trade (71% in April 2013 [2]).

The most traded currency in the foreign exchange market is the US dollar, which was in one side of 87% of trades in April 2013, followed by the euro (33% in April 2013, but having lost share compared to the previous years) and Japanese yen (23% share). Currencies from emerging economies, like the Mexican peso or the Chinese yuan have significantly increased their activity.

The aims of this project are to study and implement techniques of clustering time series of currency exchanges. In particular we study the clustering algorithm of [1] and propose some potential improvements. This method consists of applying generalized modularity minimization techniques to a weighted similarity graph obtained from the Pearson Correlation between the time series of returns of the exchange rates. We apply this algorithm to similarity graphs built from both the Kendall rank correlation and the distance correlation and analyse the differences in the results.

2. The Forex Network

Given our set of currency exchange rates (see section 2.1) and the time series of their prices over a time interval, we are going to work with their log returns to compare their behaviour.

Definition 2.1. The log return (or continuously compounded return) r_t of an asset of price P_t over a time step $[t - 1, t]$ is defined as:

$$r_t = \ln \frac{P_t}{P_{t-1}} = \ln P_t - \ln P_{t-1} \quad (1)$$

Then, once all time series of returns have been obtained, for each pair of currencies a measure of similarity between them (see section 2.2) is calculated. The network will be built with every exchange rate as a node, assigning to the edges between exchanges weights corresponding to their similarity. This will result in a weighted undirected complete graph.

2.1 Choice of Currencies

The currencies included in this study will be the 13 of the most traded in April 2013¹[2]: US dollar (USD), euro (EUR), yen (JPY), pound sterling (GBP), Australian dollar (AUD), Swiss franc (CHF), Canadian dollar (CAD), Mexican peso (MXN), Chinese yuan (CNY), New Zealand dollar (NZD), Swedish krona (SEK), Hong Kong dollar (HKD) and Singapore dollar (SGD).

¹Of the 15 most traded currencies, the 13 for which daily data was available in the U.S. Federal Reserve Economic Data were selected. The omitted currencies are Russian rouble (RUB) and Turkish lira (TRY).

The set of nodes in the network will be formed by the exchange rate of every pair of those currencies (resulting in $n = 78$ nodes). While it would be possible to just use one base currency and assign one node to every other currency corresponding to its exchange rate over the base, this could overlook interactions between some of the currencies. Moreover, the resulting network would strongly depend on the choice of base currency. This alternative approach [1] prevents those issues and introduces all the interactions between every pair of the selected currencies into the analysis.

The data used was published by the Federal Reserve Economic Data and it was downloaded into R using Quandl [3]. Since it can be difficult to find data for the exchange rate of every pair of currencies, only the exchange rates with the US dollar are used, and the rest are obtained by converting to the US dollar as an auxiliary step. That is, to obtain the exchange rate between currencies XXX and YYY where neither of them are USD, we define $XXX/YYY = \frac{XXX/USD}{YYY/USD}$.

While it is possible to obtain data for the exchange rates of most currency pairs, the fact that markets around the world have different opening hours and holidays would mean that the times at which data was taken would most likely present mismatches between all pairs. Taking only data from the US Federal Reserve limits the study to the New York opening hours but eliminates any potential inconsistencies and makes calculating correlations between pairs more accurate.

2.2 Similarity Metrics

To be able to form clusters of similar exchange rates, the edges of the network will have weights assigned according to a similarity measure. In this context, similarity is a metric built from a statistic measure of dependence, which is defined as follows[4]:

Definition 2.2. Given a set X , a function $s : X \times X \rightarrow \mathbb{R}$ is a similarity metric if, for all $x, y, z \in X$, it satisfies:

- $s(x, y) = s(y, x)$,
- $s(x, x) \geq 0$,
- $s(x, x) \geq s(x, y)$,
- $s(x, y) + s(y, z) \leq s(x, z) + s(y, y)$,
- $s(x, x) = s(y, y) = s(x, y)$ if and only if $x = y$.

We analyze three measures of dependence from which we can construct a similarity metric.

Pearson Correlation

The approach used in [1] is based on the Pearson correlation $\rho(r^i, r^j)$ of returns of exchange rates r^i, r^j over the given time interval:

$$\rho(r^i, r^j) = \frac{\text{Cov}(r^i, r^j)}{\sqrt{\text{Var}(r^i)\text{Var}(r^j)}} \quad (2)$$

Then, the weighted similarity matrix among the exchange rate returns A^ρ is given by

$$A_{ij}^\rho = \frac{1}{2}(\rho(r^i, r^j) + 1) - \delta_{ij} \quad (3)$$

which scales the Pearson correlation from $[-1, 1]$ to $[0, 1]$, while the Kronecker delta δ_{ij} removes self-edges. In the graph with adjacency matrix A^ρ exchange rates with positively linearly correlated returns will be connected by edges of weight close to 1, and weight near 0 if the correlation is negative. Edges connecting non correlated exchanges will have weights closer to the center of the interval $[0, 1]$.

This method, however, will fail to detect similarities between exchange rate returns that are not linearly correlated, which are not uncommon in financial time series.

Kendall Rank Correlation

A correlation measure alternative to Pearson's is the Kendall correlation (or rank correlation) τ [5], which measures concordance of the variables instead of linear relations between them:

Definition 2.3. Given two random variables X and Y , their Kendall correlation coefficient is

$$\tau(X, Y) = p_c - p_d \quad (4)$$

where for any two independent pairs of values $(X_i, Y_i), (X_j, Y_j)$,

$$p_c = P((X_j - X_i)(Y_j - Y_i) > 0) \quad \text{and} \quad p_d = P((X_j - X_i)(Y_j - Y_i) < 0)$$

are the probabilities of them being concordant and discordant respectively.

Definition 2.4. For our series of returns r^i, r^j over m time steps, the Kendall correlation is estimated by

$$\tau_m(r^i, r^j) = \sum_{1 \leq s < t \leq m} \frac{\text{sgn}(r_t^i - r_s^i) - \text{sgn}(r_t^j - r_s^j)}{n(n-1)} \quad (5)$$

The Kendall correlation coefficient $\tau(r^i, r^j)$ takes values in $[-1, 1]$ (equalling -1 and 1 when r^i and r^j are completely discordant or concordant, respectively) so, as before, we can define the adjacency matrix A^τ :

$$A_{ij}^\tau = \frac{1}{2}(\tau(r^i, r^j) + 1) - \delta_{ij} \quad (6)$$

For both the networks defined by A_{ij}^ρ and A_{ij}^τ and given two of their exchange rates XXX/YYY and ZZZ/TTT , the following equality holds

$$A\left(\frac{XXX}{YYY}, \frac{ZZZ}{TTT}\right) = 1 - A\left(\frac{XXX}{YYY}, \frac{TTT}{ZZZ}\right), \quad (7)$$

so if XXX/YYY is similar to ZZZ/TTT (their mutually adjacent edge has a high -close to 1- weight), it will be considered dissimilar to its inverse TTT/ZZZ (their mutually adjacent edge will have a low weight, close to 0). This means that, since we cannot *a priori* determine whether two exchange rates will be correlated directly or with one of them inverted, we have to include all inverses in the network.

Distance Correlation

Another possible choice is the distance correlation \mathcal{R} , a measure of correlation introduced in 2007 [6] and which presents several advantages. For all X, Y \mathcal{R} satisfies $0 \leq \mathcal{R}(X, Y) \leq 1$, being only equal to 0 when X and Y are independent (this is not satisfied by most correlation measures, like Pearson's, for which a correlation of 0 doesn't necessarily imply independence). The distance correlation can detect both linear and nonlinear correlations, which is relevant for this study because while many exchange rates tend to influence each other's prices, this relation doesn't necessarily have to be linear.

Definition 2.5. The empirical distance covariance between two samples X, Y is defined by

$$\mathcal{V}_n^2(X, Y) = \frac{1}{n^2} \sum_{k,l=1}^n B_{kl} C_{kl}. \quad (8)$$

where

$$\begin{aligned} b_{kl} &= |X_k - X_l|, & \bar{b}_{k\cdot} &= \frac{1}{2} \sum_{l=1}^n b_{kl}, \\ \bar{b}_{\cdot l} &= \frac{1}{n} \sum_{k=1}^n b_{kl}, & \bar{b}_{\cdot\cdot} &= \frac{1}{n^2} \sum_{k,l=1}^n b_{kl}, \\ B_{kl} &= b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}. \end{aligned} \quad (9)$$

and $c_{kl} = |Y_k - Y_l|$, $C_{kl} = c_{kl} - \bar{c}_{k\cdot} - \bar{c}_{\cdot l} + \bar{c}_{\cdot\cdot}$ are defined analogously for Y

$\mathcal{V}_n^2(X)$ is defined by:

$$\mathcal{V}^2(X) = \mathcal{V}^2(X, X) = \frac{1}{n^2} \sum_{k,l=1}^n B_{kl}^2. \quad (10)$$

Definition 2.6. The empirical distance correlation $\mathcal{R}_n(X, Y)$, between two samples X and Y , is the square root of

$$\mathcal{R}_n^2(X, Y) = \begin{cases} \frac{\mathcal{V}_n^2(X, Y)}{\sqrt{\mathcal{V}_n^2(X) \mathcal{V}_n^2(Y)}}, & \mathcal{V}_n^2(X) \mathcal{V}_n^2(Y) > 0 \\ 0, & \mathcal{V}_n^2(X) \mathcal{V}_n^2(Y) = 0 \end{cases} \quad (11)$$

Remark 2.7. It should be observed that the distance correlation is not a distance in the metric sense. It is by definition a correlation among all distances of the samples, and one can easily see that $\mathcal{R}(X, X) = 1$, violating the identity of indiscernibles of a distance. The distance correlation is in fact a similarity metric.

Using \mathcal{R} for studying currency exchange rates the following property is satisfied: given XXX/YYY , ZZZ/TTT exchange rates, $\mathcal{R}(\frac{XXX}{YYY}, \frac{ZZZ}{TTT}) = \mathcal{R}(\frac{XXX}{YYY}, \frac{TTT}{ZZZ})$. This implies that there is no need to include the inverses of the exchange rates (like in [1], where it is needed or else some correlations could be overlooked), since the correlations detected will be the same for XXX/YYY and YYY/XXX .

Since \mathcal{R} already has the properties of a similarity metric[6], the exchange rate network is simply built from the matrix of distance correlations, $A^{\mathcal{R}}$, removing self edges; that is, for each pair of exchange rate returns r^i, r^j ,

$$A_{ij}^{\mathcal{R}} = \mathcal{R}(r^i, r^j) - \delta_{ij}$$

3. Community detection

To partition the graph into communities we will use the Potts method. It consists on minimizing an objective function, the Potts Hamiltonian, which evaluates the strength² of a partition of the graph. This can be seen as a generalization of the modularity function[7].

²Considering a strong partition one that has strong links inside the communities and weak links between them.

Definition 3.1. The modularity of the partition \mathcal{P} of a weighted undirected graph with adjacency matrix A is given by

$$Q(\mathcal{P}) = \frac{1}{2m} \sum_{ij} [A_{ij} - P_{ij}] \delta(c_i, c_j)$$

where c_i is the community of the node i in the partition \mathcal{P} (so $\delta(c_i, c_j)$ is 1 when i and j are in the same community and 0 otherwise), P_{ij} is the expected weight of the edge ij in a null model and m is the sum of the weights of all edges in the graph.

Definition 3.2. The Hamiltonian of the Potts system of the partition \mathcal{P} of a weighted undirected graph with adjacency matrix A is given by

$$H(\mathcal{P}) = - \sum_{ij} [A_{ij} - \gamma P_{ij}] \delta(c_i, c_j)$$

where γ is a parameter.

Note that when $\gamma = 1$, $Q(\mathcal{P}) = -\frac{1}{2m} H(\mathcal{P})$, so in this case the problem of minimizing the Potts Hamiltonian is equivalent to the maximization of the modularity. This means the Potts method is a generalization of the modularity maximization method. The advantage of this generalization is that by varying the value of γ the partition that minimizes the Hamiltonian will contain bigger or smaller communities (corresponding with lower and higher values of γ respectively).

3.1 Selecting the null model P

The modularity and Potts Hamiltonian functions depend on null models to estimate the expected edge weight, and then form communities where the actual weight is significantly larger. In the context of community identification in networks, a popular choice of null model is the Newman-Girvan model [7]. For the networks A_{ij}^ρ and A_{ij}^τ , the Newman-Girvan null model results in

$$P_{ij} = \frac{(\sum_l A_{il})(\sum_l A_{jl})}{\sum_{i,j} A_{ij}} = \frac{n-2}{2n} \quad (12)$$

a constant value for all edges. This happens because, for a given node, the two edges connecting it to any other exchange rate and its inverse will have sum 1. Then, the weight of a node³ will depend only on the size of the network and not on the correlations between exchange rates at a given time step.

In the case of the Newman-Girvan null model of the network $A_{ij}^\mathcal{R}$, however, this property is generally not satisfied. For simplicity, we can use the average edge weight as a uniform null model:

$$P^\mathcal{R} = \frac{\sum_{i,j} A_{ij}}{n(n-1)} \quad (13)$$

3.2 Selecting the value of γ

To select an appropriate value of γ for the study, we use a sample network built from daily data of January 2010 at a single time step for each of the similarity metrics. The minimization algorithm is applied for values of γ ranging from 0.4 (all nodes are in a single community) to 2.2 (each node forms its own community) in 0.01 steps. The total number of communities for each value of γ are shown in figures 1, 2 and 3.

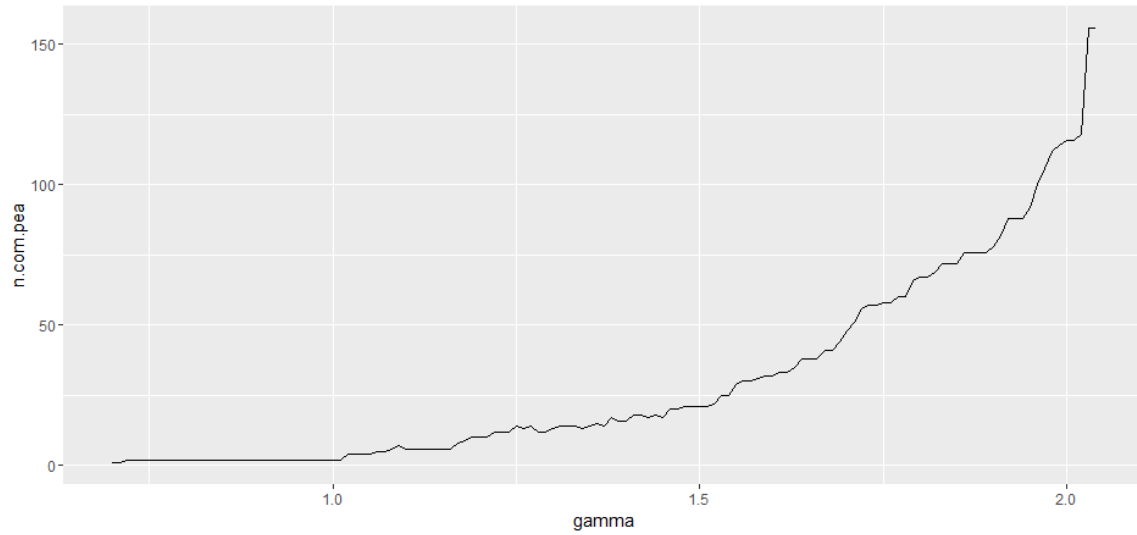


Figure 1: Number of communities of the Pearson correlation network for each γ

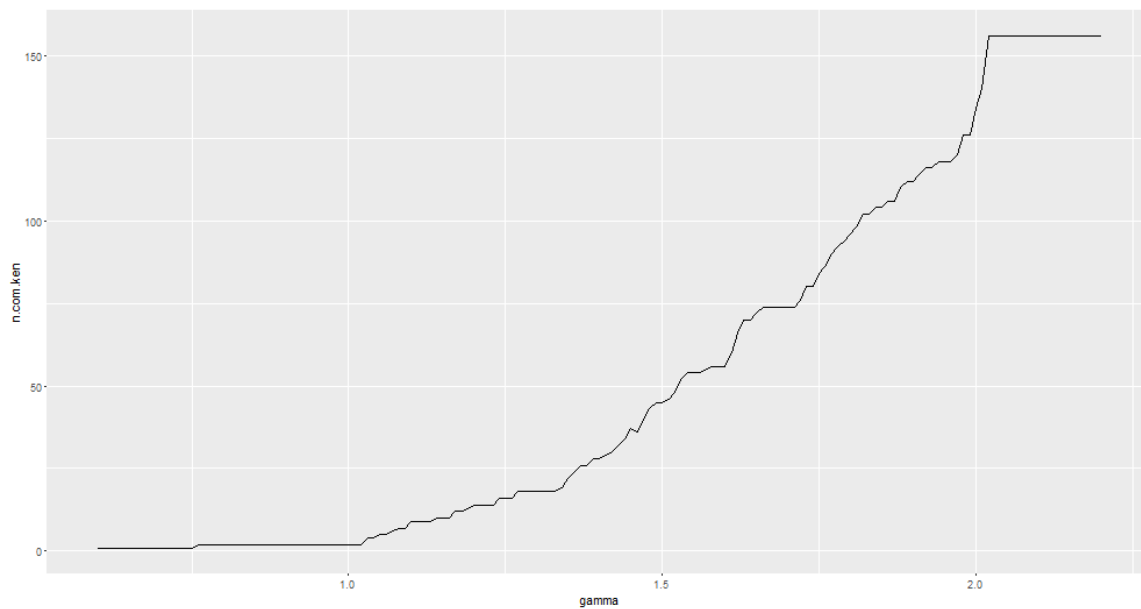


Figure 2: Number of communities of the Kendall correlation network for each γ

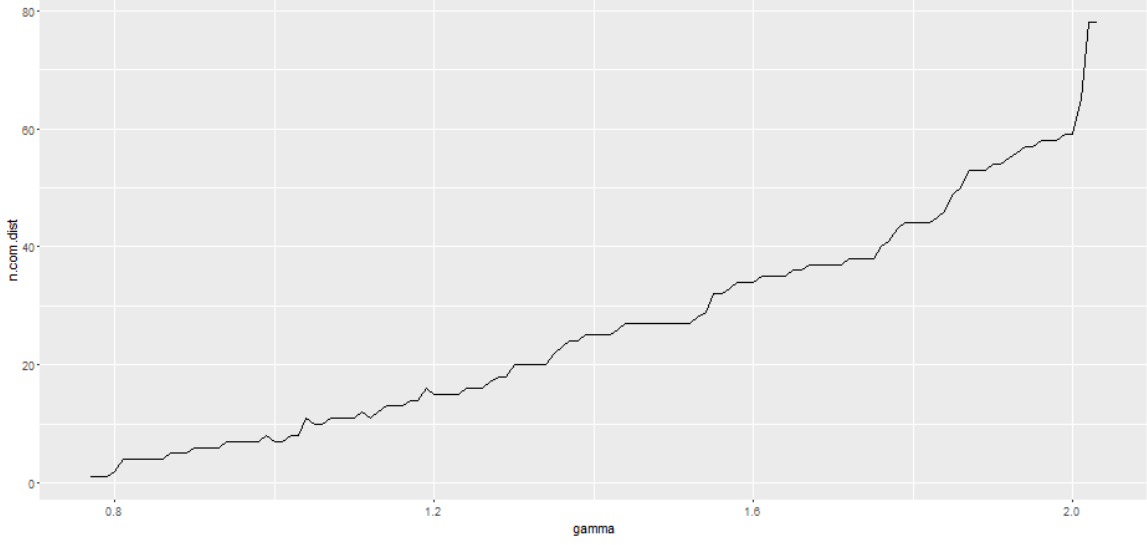


Figure 3: Number of communities of the distance correlation network for each γ

Then, we can expect to obtain more robust communities by selecting a value of gamma inside an interval where the number of communities stabilizes (referred to as plateaus in [1]). Excluding the trivial cases at the extremes of the plot with just one community containing all nodes and every node in a single community⁴, the widest non trivial plateaus are the intervals $\sim (1.48, 1.51)$, $\sim (1.27, 1.33)$ and $\sim (1.44, 1.52)$ for the Pearson, Kendall and distance correlations respectively. We then select values of gamma $\gamma_p = 1.495$, $\gamma_\tau = 1.3$ and $\gamma_R = 1.48$.

3.3 Potts Hamiltonian minimization algorithm

The algorithm used to minimize the Potts Hamiltonian is an adapted version of the modularity maximization algorithm in [8].

For a given node, we want to see if moving it to another community would give an overall decrease in the Hamiltonian (definition 3.2). The change in the Hamiltonian caused by moving the node i to a community C is given by the expression:

$$\Delta H = -\frac{1}{2m} \left[\sum_{j|c_j=C} (A_{ij} - P_{ij}) - \sum_{k|c_i=c_k} (A_{ik} - P_{ik}) \right] \quad (14)$$

which adds the contribution of the new edges ij and subtracts the contribution of the old edges ik .

We begin with each node of the network in a separate community and then we start the following iterative procedure: for each node i , we compare the variation of the Hamiltonian (14) caused by moving it to each community in the graph, and then move it to the one with the smallest negative value. If all computed values are positive, then the node remains in its community. When all nodes have been checked,

³The weight of a node is defined as the sum of the weights of its incident edges

⁴In the networks built from the Pearson and Kendall correlations, there is also a big interval where the number of communities stabilizes at two: one containing half of the nodes and the other its inverses.

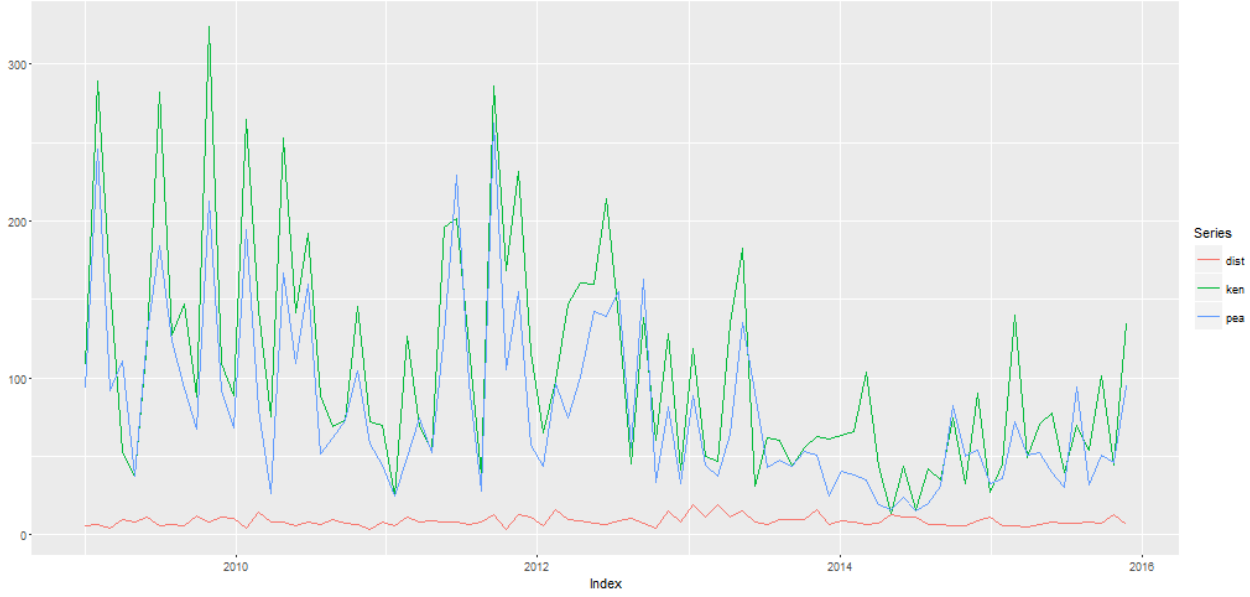


Figure 4: Variance of the size of communities from daily data at monthly time steps, over a 5 year period.

the algorithm starts again at the first node, and stops when one iteration is completed without any nodes switching communities.

4. Results

4.1 Comparison of similarity measures

In this section we will compare the results of clustering with each of the three proposed correlation measures. After choosing appropriate values of γ (see section 3), the algorithm has been applied to each of the networks obtained with daily data from the period 2009-2015 at monthly time steps (that is, for each month a network is built using daily returns of the currency exchanges).

One variable that presents significant changes between the networks is the number of nodes within communities. While in the distance correlation network the resulting communities have similar sizes among them, those of the Pearson and Kendall correlation present bigger differences. This can be easily seen by calculating the variances of the number of elements in communities for each of the three networks at every time step (figure 4).

4.2 Exploring short-term community dynamics

In this section we apply the proposed methods to one of the events studied in [1], the 2007 credit crisis to observe how the changes in the market are detected by each of the three versions of the algorithm. Figures 5, 6 and 7 show the evolution from the 2007/07/15 - 2007/08/15 to the 2007/08/15 - 2007/09/15 monthly periods.

Like in the results of [1], in the network built from the Pearson linear correlation, there is a big cluster

of relevant carry trade currencies (JPY, AUD, NZD) which at the second time step incorporates more exchanges with these currencies at one side. In the Kendall correlation networks we observe a similar behaviour; a big carry trade community that gains nodes at the second time step. In the case of the distance correlation networks, most of the carry trade currency exchanges are split into two communities, one dominated by Australian dollar exchanges and the other by the Japanese yen. At the second time step, the Australian dollar community loses nodes and ends up with only AUD exchange rates, while the Japanese yen community grows and attracts some of the nodes lost by the AUD community.

5. Centrality measures

In this section, we try to determine which nodes play important roles in the network using centrality measures.

The concept of betweenness is based on a distance between nodes and calculating shortest paths between nodes (where the length of a path is the sum of the node distances along it). In our network, edge weights A_{ij} represent similarity between nodes, so the distance d_{ij} can be taken as^[1]:

$$d_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1/A_{ij} & \text{otherwise} \end{cases} \quad (15)$$

We define G_{st} as the number of shortest paths from node s to node t , and g_{st}^i as the number of shortest paths from s to t passing through i .

Definition 5.1. The betweenness centrality b_i of a node i is:

$$b_i = \sum_{s \neq i} \sum_{t \neq s, i} \frac{g_{st}^i}{G_{st}}.$$

The betweenness centrality measures how often a given node is in the middle of shortest paths connecting other nodes, and as a consequence its role in communications between the network. This makes it relevant when studying how nodes influence each other and which ones are the most influential.

Consider $J_{ij} = A_{ij} - \gamma - P_{ij}$ and its spectral decomposition⁵ $J = UDU^T$, where D is the diagonal matrix of eigenvalues β_i and U the corresponding matrix of eigenvectors. Define q as the number of positive eigenvalues of D .

Definition 5.2. The community centrality^[9] of node i is given by the magnitude $|x_i|$, where x_i is a node vector of dimension q with j -th element given by

$$[x_i]_j = \sqrt{\beta_j} U_{ij}, \quad j \in \{1, 2, \dots, q\} \quad (16)$$

Note that the community centrality measures the connection of a node to all its neighbours regardless of their community membership. To compensate for the fact that q can vary between time steps, to calculate the averages in table 1 we normalize $|x_i|$ by its maximum at each time step.

⁵We consider the decomposition in which the eigenvalues appear in D in decreasing order, so if there are q positive eigenvalues, they will have indices $1, \dots, q$ in D

[illegible][illegible]

11

Clustering of exchange rates

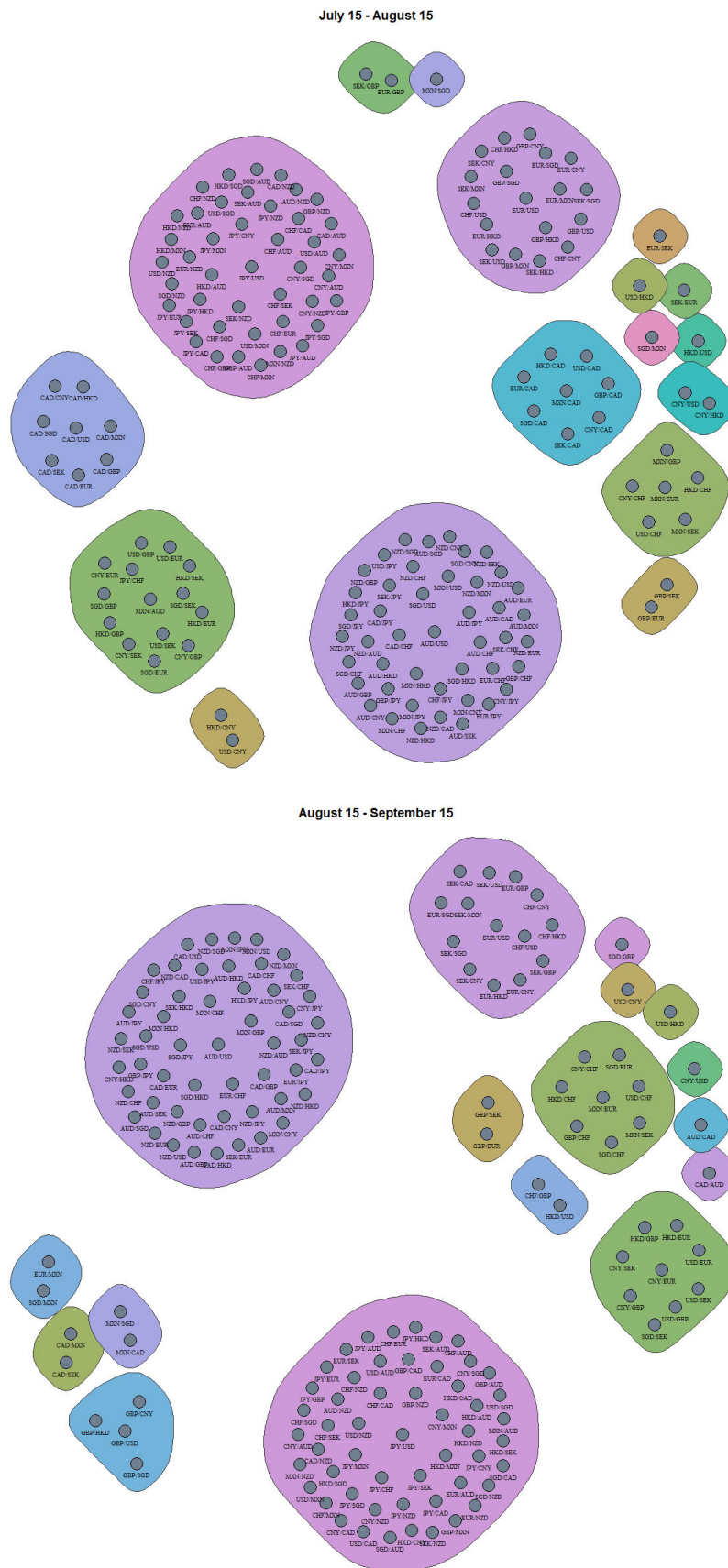


Figure 6: Kendall Correlation

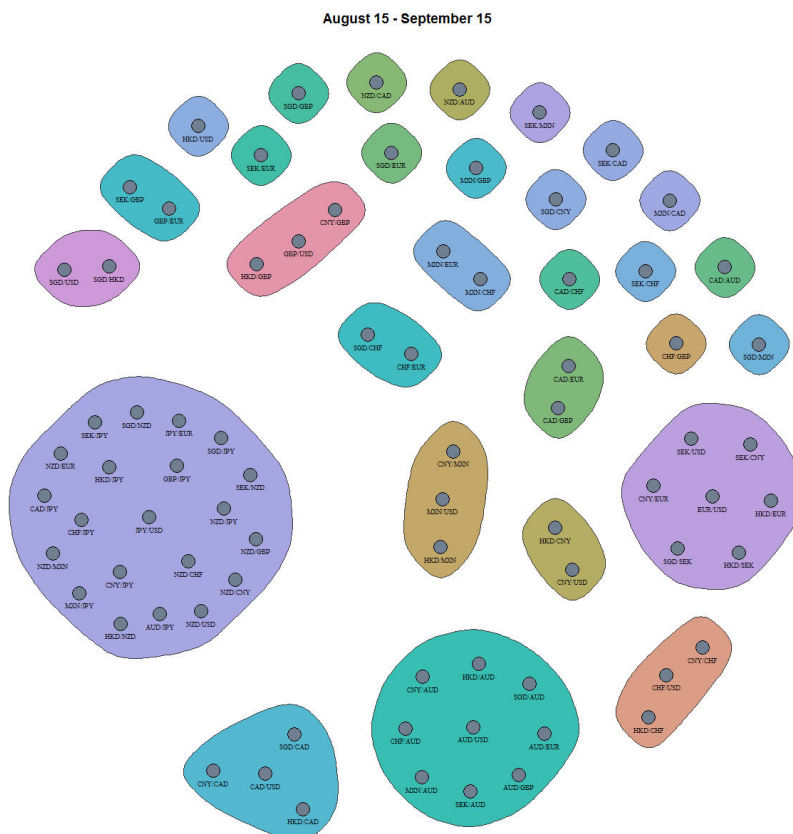
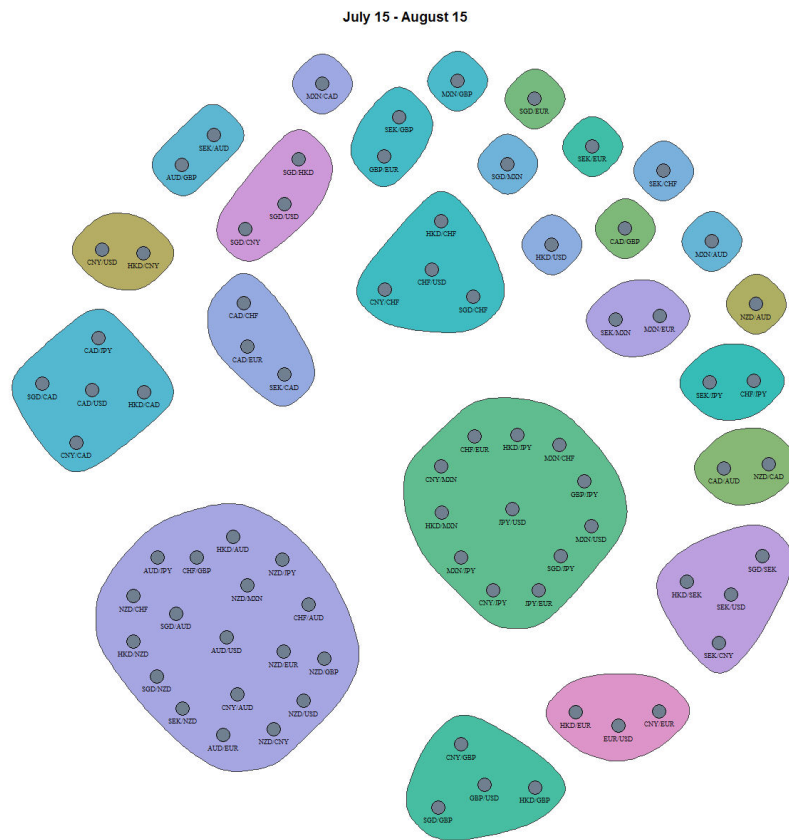


Figure 7: Distance Correlation

Table 1: Top currency exchanges sorted by average betweenness and average community centrality over the 2009-2015 period.

rank	betweenness			community centrality		
	distance	Kendall	Pearson	distance	Kendall	Pearson
1	MXN/JPY	NZD/AUD	NZD/AUD	HKD/JPY	NZD/USD	NZD/USD
2	SEK/JPY	SEK/AUD	SEK/AUD	JPY/USD	HKD/NZD	HKD/NZD
3	NZD/JPY	MXN/AUD	SEK/NZD	CNY/JPY	HKD/AUD	AUD/USD
4	SEK/MXN	SEK/NZD	MXN/AUD	HKD/CHF	HKD/EUR	HKD/AUD
5	SEK/NZD	GBP/EUR	NZD/MXN	CHF/USD	EUR/USD	SEK/USD
6	AUD/JPY	SGD/GBP	GBP/EUR	MXN/CHF	AUD/USD	NZD/CNY
7	NZD/CHF	SGD/CAD	SGD/GBP	CNY/CHF	SEK/USD	HKD/SEK
8	CAD/JPY	CAD/AUD	CAD/GBP	SGD/SEK	HKD/SEK	EUR/USD
9	MXN/CHF	NZD/MXN	SGD/CAD	SGD/CHF	CHF/USD	HKD/EUR
10	CHF/AUD	CAD/GBP	MXN/CAD	HKD/SEK	HKD/CHF	SEK/CNY

Results for the 2009-2015 period

We calculate the betweenness at every monthly time step over the 2009-2015 period for each of the three networks. The listed nodes are those with the highest mean betweenness over the period. For the Kendall and Pearson networks, betweenness of a currency exchange is equal to the betweenness of its inverse (because of the symmetry in the networks, see section 2.2), so for each pair only one representative has been included. Results are in table 1. Exchanges with AUD or NZD at one side occupy the highest positions in the Kendall and Pearson rankings for both centrality measures, while in the distance correlation network JPY exchanges dominate.

6. Conclusions

The proposed similarity metrics applied to the community detection algorithm give some differences in the results, which could be explained by their ability to detect non linear dependences between currency exchange rates. In the case of the distance correlation, being able to work with a network of half the size (due to not needing to add the inverses of exchange rates) greatly reduces the computational cost⁶ of running the algorithm, which can be significant when working with large sets of currencies.

6.1 Software used for the data analysis

All the algorithms in this project were implemented in *R*[10] with the additional packages *igraph*[11], *Quand*[3], *mcclust*[12] and *ggplot2*[13].

⁶Given that the graph is complete, halving the number of nodes divides the number of edges approximately by four. Since the algorithm used to minimize the Potts Hamiltonian is roughly linear on the number of edges, the total computational cost of clustering the distance correlation network is approximately one fourth compared to that of the Pearson and Kendall correlation networks.

References

- [1] P. J. Mucha M. McDonald S. Williams N. F. Johnson N. S. Jones D. J. Fenn, M. A. Porter. Dynamical clustering of exchange rates. 2009.
- [2] Monetary Bank for International Settlements and Economic Department. Triennial Central Bank Survey . Foreign exchange turnover in 2013: preliminary global results, 2013.
- [3] Quandl. Federal reserve economic data.
- [4] L. Ilie, G. Rozenberg, A. Salomaa, K. Salomaa, Shihyen Chen, Bin Ma, and Kaizhong Zhang. Formal languages and applications: A collection of papers in honor of sheng yu on the similarity metric and the distance metric. *Theoretical Computer Science*, 410(24):2365 – 2376, 2009.
- [5] Argimiro Arratia. *Computational Finance, An Introductory Course with R*. Atlantis Press, 2014.
- [6] N. K. Bakirov G. J. Szekely, M. L. Rizzo. Measuring and testing dependency by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.
- [7] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, Feb 2004.
- [8] R. Lambiotte E. Lefebvre V. D. Blondel, J. Guillaume. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008.
- [9] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74, May 2006.
- [10] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.
- [11] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *Inter-Journal, Complex Systems*:1695, 2006.
- [12] Arno Fritsch. *mcclust: Process an MCMC Sample of Clusterings*, 2012. R package version 1.0.
- [13] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.