Asset Management and Graph Theory

Master 1 Econométrie-statistiques

Université Paris 1 Panthéon-Sorbonne

Armand L'Huillier, Camil Zahi, Yanis Rehoune

Supervised by Philippe De Peretti

January, 30th, 2022

Dedication

We would like to thank first of all Mr. Philippe De Peretti for his availability, his advice and his kindness throughout the semester. It was a challenging work thanks to which we learned many theoretical concepts and progressed in our way to approach a problem with SAS.

Contents

Abstract

This paper proposes to construct a portfolio of assets by linking econometric results, graph theory and portfolio allocation. In theory, the assets in a portfolio must be uncorrelated. This view excludes feedback and spillover effects and thus dynamic interactions. Focusing on the CAC40 index, we dynamically constructed a network (on a rolling window), the nodes being the assets and the arcs the cross-correlation at more or less one period. Once the topology was found, we constructed the adjacency matrix and found a relevant clustering measure. Following the results of this matrix, we selected the least systemic assets that will then be selected and allocated (equaly weighted) over a given time period. By rebalancing our portfolio every 6 weeks, we finally compared our results with the CAC40 benchmark, the objective being to beat the market for an equivalent risk.

Introduction

Before 1950 and the introduction of the first portfolio management models, management methods were known as "traditional". This was essentially based on the financial analysis of securities and their individual selection (stock-picking). Indeed, the allocation of the portfolio as well as the diversification was done at the manager's discretion. Performance was therefore dependent on the manager's talent and market knowledge. This period is associated with the "bottom-up" active management technique, which consists of investing in a company that is undervalued by the market despite its strong potential.

In the 1950s, modern management was born with the work of Harry Markowitz who developed the modern portfolio theory. This theory seeks to define the process for creating the portfolio as efficiently as possible, that is, the process that maximizes return for the lowest possible risk. This is also known as quantitative management.

In the 1960s, Sharpe, Lintner and Mossin extended Markowitz's work and created the "Capital Asset Pricing Model" (CAPM), which is the most famous and widely used valuation model. In parallel, the APT model (Arbitrage pricing theory) was created following the anomalies observed in the CAPM. This competing model is based on a multivariate analysis incorporating macroeconomic and specific factors. These models are associated with so-called "quantitative" management, which is developing more and more through the rise of asset management companies (BlackRock, State Street Global Advisors, etc.) and hedge funds (alternative management).

Nowadays, more and more firms are trying to develop proprietary algorithms that automatically manage the allocation and diversification of portfolios in a dynamic way. These advances seem to take precedence over the theoretical models stated earlier that are partly responsible for the various contemporary crises. Graph theory is proving to be a good way of schematizing the dynamic interactions between financial actors.

Through this paper, we will start by explaining the theoretical concepts of finance using graph modeling tools which will allow us in a second part to apply them to data collected from the CAC 40 over the period 01/2005-12/2009.

Chapter 1

Theoretical approach

1.1 Performance attribution models

1.1.1 The traditional measure

Modern financial literature traditionally measures the performance of a manager by his ability to outperform his investment universe, which is characterized by a reference index called "benchmark". Indeed, the Capital Asset Pricing Model (CAPM) developed by William F. Sharpe in 1964 is the basic financial asset pricing model. The market model corresponds to the empirical version of the model and relates the total return to the return of the market portfolio via the following equation :

$$R_{i,t} = \alpha_t + \beta_t * R_{m,t} + \epsilon_t$$

With :

$R_{i,t} : total fund return$

$\alpha_t : portion of return not specific to market changes$

$\beta_t : changes in the return on assets related to market portfolio returns$

$R_{m,t} : market portfolio return$

$\epsilon_t : idiosyncratic risk$

Modern financial literature assesses the performance of a fund through its alpha. Thus, the manager is judged against the benchmark that he seeks to outperform through a specific weighting of securities. The difference between the fund's performance and that of the index is called "tracking error".

1.1.2 Multifactor models

In addition, other multifactorial approaches have emerged in the wake of the CAPM to better measure performance. Indeed, the three-factor model of Fama and French (1993) as well as the four-factor model of Carhart (1997) show that there is not only one factor explaining performance. These approaches are symbolized by the following equation:

$$R_{i,t} = \alpha_t + \sum_{i=1}^{k} \beta_i * F_{i,t} + \varepsilon_t$$

With : F : what each factor yields

This approach is relatively close to the (APT) model according to which the expected return on a financial asset is a linear function of the sensitivity of the various risk factors influencing the asset. These models have fairly good results and manage to explain a large part of the returns of the equity market. In addition, econometric work that is close to multifactor models has emerged. The concept is to try to isolate the factors that really influence asset prices.

1.1.3 The limits of modern financial theory

The performance attribution models of modern financial theory are nevertheless limited. The main problem is that they are static models, i.e. the time dimension is not taken into account in the analysis. Although simple in appearance, the modern portfolio theory developed by Markowitz is disconnected from the reality of financial markets and is considered responsible for several crises (failure of LTCM, Lehman

Brothers). The variance-covariance matrices are biased and cannot be interpreted. Moreover, the assumption of normality of returns is also unrealistic in finance. There are many factors that impact the distribution of returns. However, in practice, some banks still use this type of model, which is supposed to work only if all the agents also use it (mimicry pheonomen).

1.2 Graphs Theory

1.2.1 General concept

Graph theory allows to synthesize interactions in a static or dynamic way. In particular, it allows to identify the central agents that broadcast information and those that are more "follower". The fields of application are numerous (transport, national defense, marketing) and extend more generally to any type of complex system that can be represented in the form of a network. More recently, we observe the development of applications of graph theory in Asset Management.

First, an important concept in market finance is that one cannot analyze a financial asset without considering the global aspect of the market. For these reasons, we can use copula functions and vector autoregressive models to model the structure of dependencies between assets. These methods allow us to diversify a portfolio in an efficient way. However, we will focus on an approach based on graph theory in order to identify the least systemic assets of a given index and then build a portfolio. We will review the different ways of using graph theory in Asset Management.

1.2.2 The concept of centrality

First of all, it is important to understand the intuitions that follow from the adjacency matrix. Indeed, the latter incorporates all the links of the network by indicating at each position if the node "i" broadcasts an information to the node "j". If we consider a stock market index, we can represent the correlations of assets 2 to 2 in the adjacency matrix as 1 and 0 according to the level of correlation at a given threshold.

The notion of node centrality is essential to identify how information propagates between assets. Techniques to observe the exchange of information between neighboring assets have emerged and are based on message passing algorithms (Pacreau, Lezmi, Xu,2021). This approach can help in the choice of assets that compose the portfolio. Indeed, we prefer to have assets that do not receive too much information.

The underlying idea is therefore to limit the propagation of information to reduce the systemic risk (Crisis of 2007-2008). The centrality of an asset is quantified by the degree of connection of the asset to its index (network). Cohen and Frazzini (2008), Wu and Birge (2014) and Wu (2015) have studied how centrality measures show the exposure of an asset to market variability. The most widely used measure is degree centrality.

$$C_d(i) = d(i) = \sum_{j \in V} A_{i,j}$$

1.2.3 The different types of clustering

Using clustering is probably the best way to use graph theory in Asset Management. Initially reserved for image processing, the work of Zhu et al. (2016) introduced graph clustering for time series. Today, there are several methods that we will now explain. A first approach consists in solving the minimum cut problem. The underlying idea is to find the cut that would separate the assets into two groups of unrelated assets. We look for the cut that minimizes a certain measure. The Kagner algorithm or the Stoer-Wagner method is often used to solve the problem.
In addition, we can use unsupervised classification methods. Indeed, spectral clustering algorithms based on the eigenvalues of the Laplacian matrix seem to give correct results (Pacreau, Lezmi, Xu, 2021). Some work also attempts to use the k-means algorithm to find non-systemic assets. The problem with these approaches is that the analytical framework is static.

However, we wish to conduct a dynamic analysis in order to study the cross-correlations between assets over time. This is why in the rest of this paper, we will focus on clustering measures developed by Giorgio Fagiolo (2006) which will allow us to synthesize the topology of the CAC40 and create a portfolio more interesting than the index itself.

Chapter 2

Empirical approach

2.1 The Dataset

Now that we have exposed the theoretical framework of asset management analysis using graph theory, we are going to move on to our next part, which is the study of the dependence of CAC40 assets. This study will be used later on to create a model to build portfolios.

First, the data set needs to be introduced. Values of components of the CAC40 from the 01/01/2005 to the 12/31/2009 have been recorded. Thus the data set consists of a five-year period with daily data. Some problems occurred when collecting data. For example, the CAC 40 composition is reviewed frequently therefore some assets enter the CAC40 while others have to leave it. Over those 5 years, only 27 assets stayed in the CAC40. A decision had to be made. One decision could have been to keep the 40 assets present on the 01/01/2005 throughout the period even though 13 assets left the CAC40 over the period. However, the decision taken was to consider only the data from 27 assets over the entire period. Moving on to the next part, we will try to model a proper way to manage assets, by considering the CAC40 as a complex network, where the nodes (bnp, vivendi, ...) in the network are connected by directed links. In particular, we will take into consideration that an asset A in a period (t-1) could have an influence on an asset B in a period (t).

2.2 Application: Outperform a stock index

2.2.1 Data Structure

It was explained in the first part of the paper that the returns do not follow a normal distribution. It was therefore decided to start by checking that our data from the CAC40 did not follow a normal distribution. This was done by performing the Kolmogorov-Smirnov test. The null hypothesis being that the CAC40 returns follow a normal distribution, this hypothesis is rejected at the 5 pct threshold:

Goodness-of-Fit for Normal Distribution		
Test	Statistique	p-value
Kolmogorov-Smirnov	D = 0 .05804118	< 0.010

Skewness = -0.2382617	Kurtosis= -0.9198411

In addition to the previous test, the kurtosis and the skewness of the approximated distribution are given: First, the skewness is negative, therefore the distribution of the index is more important below the mean than above. This can be explained by the time period over which our data extend. Indeed, the period covers the crisis of 2007-2008 which is characterized by negative returns. Second, the Kurtosis is around 2, which means that the distribution is flatter than the normal distribution. This distribution is platykurtic. To conclude, the returns are not stable.

2.2.2 Correlations and clustering coefficients

The objective of our model is to outperform our adjusted CAC40 index and to get a more attractive risk/reward profile than our benchmark. The method used is seeking to minimize the risk of a given portfolio of assets by eliminating assets that are too dynamically correlated with each other. This is where G. Fagiolo's article becomes interesting to us. Indeed, his work gives an accurate method to find clusters in any network. The CAC 40 can be defined as a network that links assets

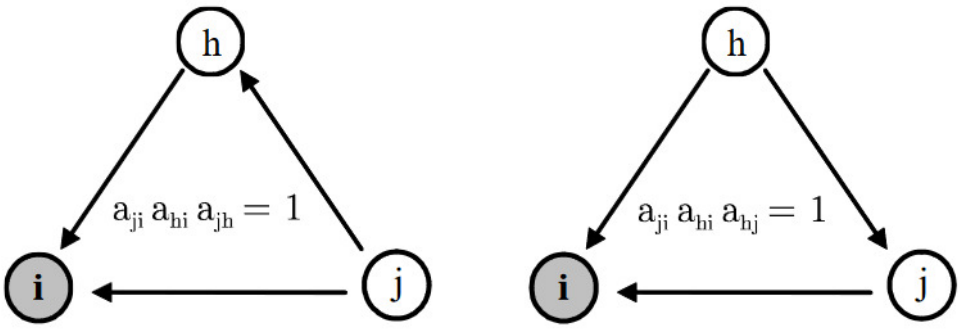
to each other. As such, Fagiolo's measures are useful in finding connections between multiple assets. Two main measures are defined by Fagiolo to detect whether or not a correlation exists between an asset i and an asset j. First, there is the clustering coefficient "out" that is useful in finding the level of correlation of a specific node to the others (one node has some influence on the others).The coefficient is calculated using EQUATION NUMBER J. In the given equation, A represent the adjacency matrix which will be defined later. Still in the question, $d_i^{out}$is the out-degree of asset i which is defined as the number of edges originating from i.

$$C_i^{out} = \frac{(A^2 A^T)_{ii}}{d_i^{out}(d_i^{out} - 1)}(Eq.K)$$

However, another measure seemed better. Indeed, Fagiolo's "in" clustering coefficient will be used is our model because in the context of portfolio studies and asset correlation, it is much more relevant to detect which assets receive an impact from another specific asset (a single node receives a certain amount of information from its peers). In addition to this intuition, both coefficients were tested in our final approach, and it is clear that the "in" clustering coefficient captures more information from our data than the other coefficient. Given that more links appear when the 'in' coefficient is used, it indicates that some information is lost when using the 'out' coefficient. To compute this clustering coefficient, equation k is used. A still represents the adjacency matrix. While $d_i^{in}$is the in-degree of node i, it measures the number of edges pointing towards the asset i.

$$C_i^{in} = \frac{(A^T A^2)_{ii}}{d_i^{in}(d_i^{in} - 1)}(Eq.K)$$

The underlying idea behind these clustering measures is to compute the number of triangles with a certain topology in the neighborhood of each asset. Getting all triangles of a certain topology is supposed to gather enough information on the correlation of asset i. The chosen clustering coefficient spots the following triangle formats:

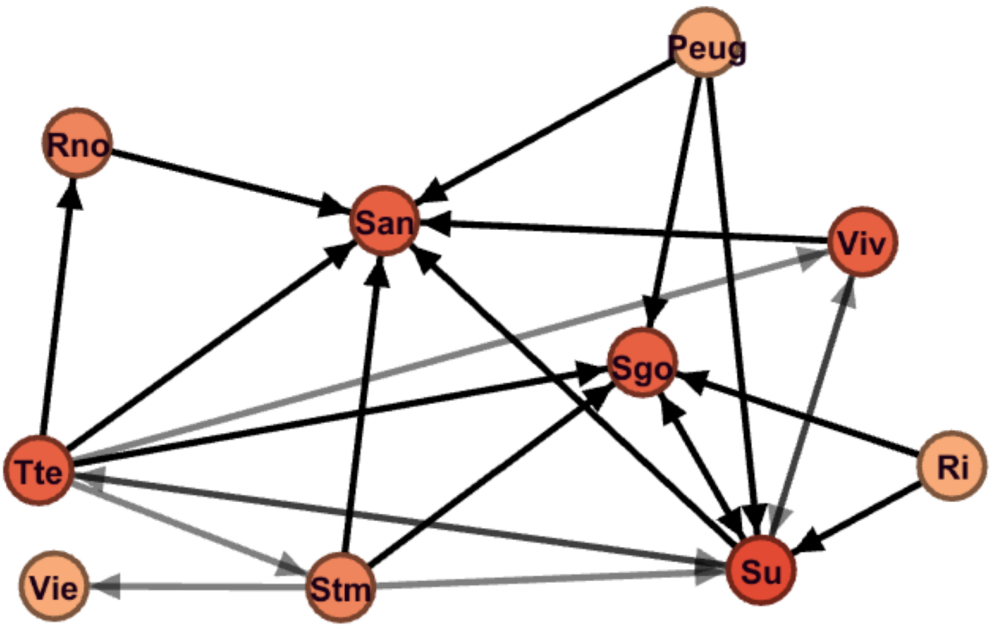


To calculate these coefficients, we must first define an adjacency matrix. Such a matrix is obtained by putting a 1 when a strong cross-correlation over the last 6 months is detected between two assets, and a 0 otherwise.

Our adjacency matrix thus indicates the cross-correlation between each pair of assets (X_{it-1}, X_{jt}) and (X_{jt-1}, X_{it}).

The cross-correlation is defined as strong when the Chi-square test suggests it. The threshold for this test was set at 10 %.

Indeed, after several simulations with different thresholds, it seems that the 10% threshold captures more information about the correlation than the 5% threshold. Concerning the window of calculation of the cross-correlations, we set it at 6 months because we noticed that it is a period which takes well into account the financial cycles.



The graph above is a partial representation of one of the adjacency matrices. We can easily identify the assets that receive more or less information from the other assets in the index. Thus, we can observe that 3 of the assets receive particularly a lot of directed linkage. These are Sanofi, Schneider Electric and Compagnie de Saint-Gobain and will be removed from the portfolio over this period.
Concerning the decision rule linked to the clustering coefficient, we realize a simulation of 100 observations in a normal distribution, then we calculate our clustering coefficients on it. Finally, the distribution of the CC shows that a relevant threshold is located at 0.166. Thus, all assets whose clustering coefficient is higher than 0.166 are removed from the portfolio. Because of the presence of autocorrelation, the main diagonal of the adjacency matrix was not filled with 0, this distorted our clustering coefficients because some correlations would be falsely taken into account. We thus had to set this diagonal to 0.

2.2.3 Management of a simulated portfolio

The data used in our research covers a 5-year period, starting in January 2005. Rebalancing the portfolio every two months proves to be effective as several periods were tested in our model. The results show that the best Sortino ratio and the best returns seem to appear when the rebalancing takes place every two months. Indeed, two weeks, one month and three months were tested in addition to the two-month period. As previously stated, the first time a portfolio is created is on 07/01/2005, using data from January to June. We start with an amount of 1000€ to invest in our "non sysmic" portfolio. Two months after investing this 1000€ in the right assets, we sell the assets purchased before the creation. The new amount of money defined by the quantity bought before and the new prices, is used again to invest on a new portfolio determined by the new adjacency matrix computed on the cross correlations of the last 6 months of this new period. In parallel we invest the same amount on our index in order to compare the performance over time.
Concerning the performance measures, we will use the Sortino ratio which is close

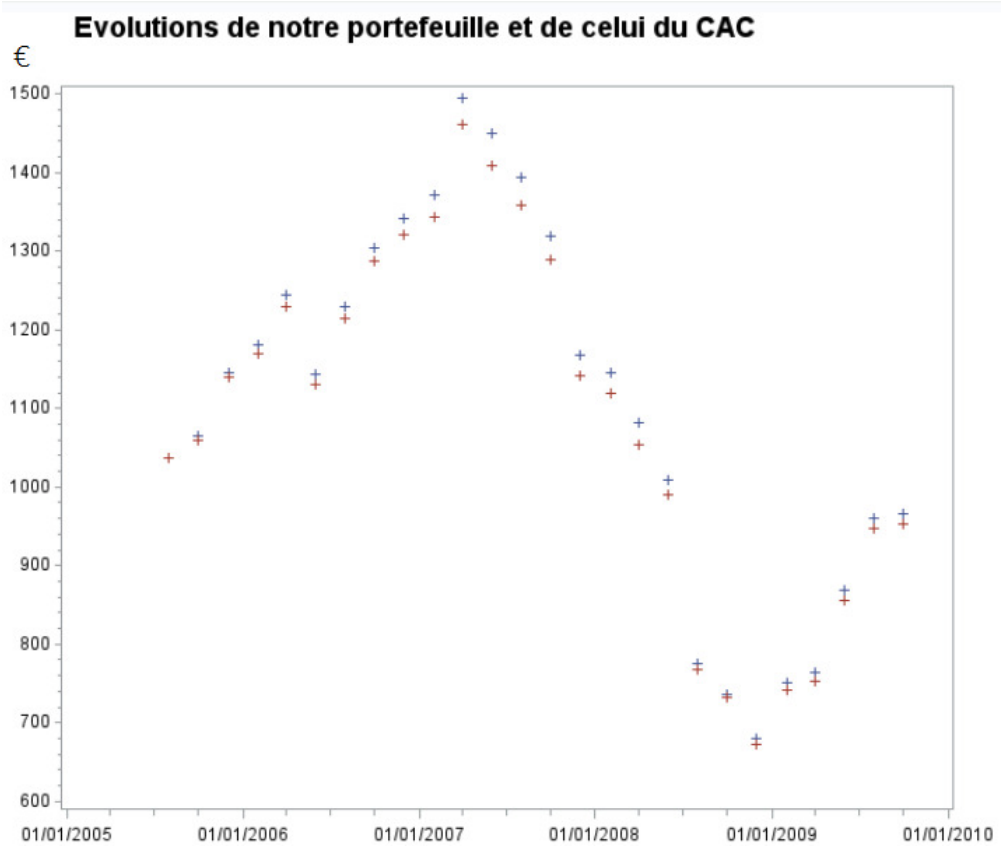
to the Sharpe ratio except that the risk measure used is not the volatility but the donwside deviation. This ratio is given by the following formula:

$$Sortino = \frac{E(R_t) - MAR}{\sigma_{MAR}}$$

It is a finer measure than the Sharpe ratio especially since MAR corresponds here to the benchmark (CAC40) rather than the risk-free rate. Moreover, the assumption of normality of returns is not necessary. Another interesting risk measure is the tracking error, which measures the difference in performance between a portfolio and its benchmark. It allows us to see to what extent the returns of a fund follow those of the index:

$$TrackingError = \sqrt{Var(r_p - r_b)} = \sqrt{E[(r_p - r_b)^2] - (E[r_p - r_b])^2}$$

Now we will analyze the performance of our portfolio against our benchmark. The following graph shows in blue the evolution of our balance and in red that of the index:



We can see that we outperform most of the time. As our database is focused on the crisis period, we cannot analyze the longer term. However, we can observe that in bullish periods, we outperform and in bearish periods, our portfolio tends little by little towards the performance of the index. This can be explained by the fact that in the equity market the correlations of returns increase when the markets are volatile (correlation smile) Cécile Kharoubi (2016).

For a management period of 2 months, at a threshold of 10%, we have:

dsd	Sortino
0.0940552	0.0076026

Tracking Error
0.0200533

We observe that we have a low but still positive Sortino ratio. This can be explained by the period we consider. Indeed, our returns over the period neutralize each other. We have good returns before 2007 and bad returns afterwards. The numerator is therefore quite low. Concerning the tracking error, it is at 2% which is credible considering our results which are interesting but not revolutionary. It would be interesting to apply our process to a larger database to see if we perform better than the CAC40.

Conclusion

Something very important to this essay is the data. It was hard to get, hard to adjust and hard to use properly. A problem faced is that the period of time was too short. But if we had used a longer period, it seems that the overall profitability would increase.The fact that during almost 2 years in the data used is a time of crisis shrinks a lot the profitabilities of portfolios. This is bad because it makes it harder to distinguish the portfolio management and the CAC40. Also, if the time period was longer, it could lead to a greater Sortino ratio as the numerator could increase significatively. Yet, the time-period covered in this paper gives very interesting results. For example, the portfolio management outperforms the CAC40 during bullish periods, but not so much in bearish periods. There are many improvements that can be suggested. Firstly, the model only takes into consideration the risk brought by cross-correlations with a lag of 1. But it would be relevant to take into account as well the performance and profitability. Moreover, methods to compute the best thresholds and time-periods should be implemented so that the risk is truly minimized using clustering coefficients and cross-correlations. Finally, the intuition suggests that our model should work better in times of great instability and in markets highly fickle. Thus, this intuition could be tested by using other data from other markets with the same time-period, proving whether or not our model works fine in times of crisis. Other data should be collected in order to prove whether our model works better when applied on some indexes rather than others. New approaches in deep learning applied to graph theory are emerging in asset management. It would be interesting to see the methods used in the next years.

Appendix

Portfolio performances depending on the management period

At 2 weeks	
dsd	**Sortino**
0.0520038	**0.0015784**

At 1 month	
dsd	**Sortino**
0.0784533	**0.0019585**

At 2 months	
dsd	**Sortino**
0.0940552	**0.0076026**

At 3 months	
dsd	**Sortino**
0.1248162	**0.0074253**

Portfolio performances depending on the threshold of the chi-2 test

At 1 percent	
dsd	**Sortino**
0.092406	**-0.001636**

At 5 percent	
dsd	**Sortino**
0.0939115	**-0.001399**

At 10 percent	
dsd	**Sortino**
0.0940552	**0.0076026**

Bibliography

Articles :

Cohen L., Frazzini A., (2008), Economic links and predictable returns, In Journal of Finance, volume 63(4), pages 1977–2011.

El Himdi K., Roy R., (1997), Tests for Noncorrelation of Two Multivariate ARMA Time Series, The Canadian Journal of Statistics / La Revue Canadienne de Statistique, Jun.,1997, Vol. 25, No. 2, pp. 233-256.

Fagiolo G., (2006), Clustering in Complex Directed Networks, Sant'Anna School of Advanced Studies,Laboratory of Economics and Management, Piazza Martiri della Libert'a 33, I-56127 Pisa, Italy.

Kharoubi C., (2016), Hedge Funds, une analyse critique, RB edition, .

Pacreau G., Lezmi E., Xu J., (2021), Graph Neural Networks for Asset Management, 119-2021, Amundi Asset Management, Working Paper.

Wu J., Birge J.R., (2014). Supply chain network structure and firm returns. SSRN, 238547.

Wu L., (2015), Centrality of the supply chain network. SSRN, 2651786.

Zhu Z., Peng Q., Guan X. (2016), A time series clustering method based on hypergraph partitioning. In 2016 International Conference on Progress in Informatics and Computing(PIC 2016), pages 27–31.