Novel Forecasting Techniques Using Big Data, Network Science and Economics

Irena Vodenska[1,3,⋆], Andreas Joseph[2,3], Eugene Stanley[3], and Guanrong Chen[2]

[1] Administrative Sciences Department, Metropolitan College, Boston University, Boston, MA 02215 USA
[2] Center for Chaos and Complex Networks, Department of Electronic Engineering, City University of Hong Kong, Hong Kong S.A.R., China
[3] Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, USA
vodenska@bu.edu

Abstract. The combination of theoretical network approach with recently available abundant economic data leads to the development of novel analytic and computational tools for modeling and forecasting key economic indicators. The main idea of this study is to introduce a topological component into economic analysis, consistently taking into account higher-order interactions in the economic network. We present a multiple linear regression optimization algorithm to generate a relational network between individual components of national balance of payment accounts. Our model describes well annual country statistics using the explanatory power and best fits of related global financial and trade indicators. The proposed algorithm delivers good forecasts with high accuracy for the majority of the indicators.

Keywords: Complex Systems, Econometrics, Interdependent Networks, Finance and Economics.

1 Introduction

Since the latest global financial crisis of 2008 and the resulting European Sovereign Debt Crisis of 2011, policy makers, academics and the public have shown increased awareness of the strong and important interconnectedness and interdependence of the global financial and economic architecture [1,2,3,4]. Additionally, standard techniques in macroeconomics partly failed to describe or foresee these major downturns, as pointed out by Jean-Claude Trichet in his opening address at the ECB Central Banking Conference (Frankfurt, 18 November 2010): "Macro models failed to predict the crisis and seemed incapable of explaining what was happening to the economy in a convincing manner. As a policy-maker during the crisis, I found the available models of limited help. In fact, I would go further: in the face of the crisis, we felt abandoned by conventional tools." In recent years we

[⋆] Corresponding author.

V.M. Mladenov and P.C. Ivanov (Eds.): NDES 2014, CCIS 438, pp. 254–261, 2014.

have seen the rise of *Big Data*, i.e. the availability of large amounts of high-quality digitalized data, as well as the development of *network science*, which investigates the properties of systems composed of a large number of connected components. The combination of Big Data and network science offers potential applications for the design of data-driven analysis and regulatory tools, covering many aspects of our society [5]. More specifically, *econometrics* is a field where merging current analysis techniques with data- and network science is expected to provide large gains to modern economics.

The main idea behind network science is to generate a *network gain* by consistently considering multiple, as well as higher-order interactions between the individual components of a large system, such as users in a social network, banks in the financial system, or the interaction of entire economies on a global scale. Here we consider network dynamics and topological structure constructed by the flow of goods or cross-border investments between certain agents in a marketplace.

2 The MLR-Fit Network of Global Balance of Payment Accounts

Recently, the understanding of economic contagion has attracted growing interest, aiming to explain and measure the spreading of economic downturns between countries and across asset classes [6,7,3,8]. The modeling of global economic interactions between the different macro-components originating from multiple countries faces substantial difficulties due to the large number of possible interaction channels and due to the underlying computational complexity of the problem.

In this section, we present how a standard technique from economic analysis, namely multiple regression analysis (MLR), can be combined with Big Data and network science to tackle the above-stated problem by delivering an accurate phenomenological description of network relations and exhibiting good predictive power for macroeconomic indicators. The methodology that we apply here is not confined to any particular field, but has a large number of potential applications, whenever the criterion of sufficient data availability is fulfilled.

2.1 Balance of Payments Network Analysis

To study global trade and investment flows, we construct a relational network for 60 countries and eight financial and trade indicators for eleven consecutive years (2002-2012). These indicators constitute major parts of a country's balance of payments and represent nodes in our network. The eight indicators are the trade of goods (exports and imports) [9], inbound and outbound foreign direct investment (in- and out-FDI) [10] and inbound and outbound cross-border portfolio investment (in- and out-CPI) of equity and debt securities [11] on the nationally aggregated level, where we differentiate between in- and outbound relations. We show in Fig. 1 that the aggregated global trade flow within one year tracks the corresponding start-of-the-year investment positions, CPI and FDI. This is the reason why we will focus on a country's total exports and imports, as well as in- and outbound investment positions, in this analysis. Example indicators, which are included in this analysis, are the total exports of China (China: Exports) and the aggregated foreign holdings of US debt securities (USA: Debt (in)). All indicators have been adjusted for yearly changes in GDP, using the global GDP-deflator [12] (constant year-2012 values).

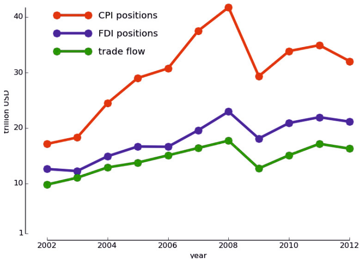

Fig. 1. Comparison of the **magnitudes of global trade flows and investment positions** (CPI and FDI) between 2002-2012. Because the total trade flow within one year tracks well the start-of-the-year positions of both types of investment (debt and equity), we study a combination of trade flows and investment positions.

We generate the links (relationships) between the trading and financial indicators (nodes) using a multiple linear regression (MLR) algorithm, connecting every indicator to one or several other indicators. We call the resulting relational network *Global Balance of Payments Network* (GBoPN). The network itself represents an evolution framework which describes and even forecasts most of the macroeconomic indicators with high accuracy. In algebraic terms, a network is represented by its *adjacency matrix* $\mathcal{A}$, where the element $a_{ij} \neq 0$ if there is an edge between node i and node j. One distinguishes between different classes of networks, depending on the possible values that the matrix elements a_{ij} are allowed to take. If $a_{ij} \in \{0, 1\}$, $a_{ij} = a_{ji}$ and $a_{ii} = 0$, the resulting network is called a simple graph, which captures the topological structure of the network in which connections between nodes are symmetric. If the only requirement is that $a_{ij} \in \{0, 1\}$, one obtains a directed graph, where connections between nodes are not required to be reciprocal. In the most general network, $a_{ij} \in \mathbb{R}$, meaning that one can have asymmetric connections between any two nodes accounting for possibly negative feedback. The underlying idea of constructing this network is that global financial and trade statistics are related to each other, such that it is possible to find a set of indicators (regressors) which best describes *another indicator* (regressand), using MLR.

2.2 MLR-fit Model for Forecasting Global Macroeconomic Indicators

Let $\boldsymbol{I}(t)$ denote the set of all indicators of all countries at time t, e.g. in 2007. We are interested in finding the best-fit network coefficient matrix $\boldsymbol{\beta} = \beta_{ij}$ which describes $\boldsymbol{I}(t+1)$. This can be written in terms of a linear matrix equation

$$I_i\,(t+1) \overset{\mathrm{MLR}}{=} \sum_{j=1, j\neq i}^{N} \beta_{ij}\, I_j\,(t) \,+\, c_i\,, \tag{1}$$

where N is the total number of indicators from all countries and c_i is the intercept of indicator I_i. A link between indicators I_i and I_j is established if $\beta_{ij} \neq 0$. Note that the model (1) is highly appealing from a mathematical point of view because of its simplicity, as well as its built-in predictive power, since the coefficient matrix $\boldsymbol{\beta}$ can be interpreted as an *evolution operator*, which takes the indicator vector $\boldsymbol{I}$ from $t \to t+1$.

However, finding the optimal $\boldsymbol{\beta}$ according to fixed statistical criteria on the MLR-fit is a computationally hard problem because the number of possible solutions is growing super-exponentially with the number of indicators. Finding a good solution to this problem is much more likely with the availability of a large amount of data, while an efficient search method will considerably shorten the time to find a solution. We use an iterative least-square algorithm, which is based on the assumption that a regressor which individually describes a regressand well (simple regression) is likely to be contained in a group of regressors (multiple regression). The optimization algorithm that we use is as follows: On each $I_i\,(t+1)$ time series we perform a simple linear regression (SLR) with each *other* time-lagged time series $I_j\,(t)$ (Step 1). We pick the regressor I_j which generates the smallest error (residuum) for the starting model $f^1\left(I_i^{t+1}\right)$ (Step 2). We then add additional regressors to the model,according to the ordering of residua from Step 1 (smallest to largest), and test the resulting model for error reduction, statistical significance of all regressors (t-test) and collinearity (variance inflation factor (VIF) and condition number of the normalized design matrix, Steps 3 and 4). Based on the test results, we update the model (or not) and go back to Step 3. We repeat this step for $N-2$ times (Step 5). Finally, we test the statistical significance (F-test) and error of the final model $f^n\left(I_i^{t+1}\right)$, where n is the final number of regressors, then accept or reject the model (Step 6).

This procedure yields one row of the coefficient matrix $\boldsymbol{\beta}$ for each indicator $I_i\,(t+1)$. The final result crucially depends on the chosen requirements regarding the maximally allowable error, the statistical significance of each coefficient, and on collinearity bounds where there is a general trade off between the maximal error on one side and the statistical significance and collinearity between regressors on the other. Depending on the achievable balance between these quantities, the MLR-fit model may be accepted or rejected.

We test our model on a set of 60 countries, selected according to a 95%-criterion on the cumulative amount of the total monetary value (in USD) of all eight indicators taken together. Theoretically, we obtain a total of $60 \times 8 = 480$

indicators. Unfortunately the data are partly incomplete, so that we are left with a total of 405 indicators, or about 84% of the expected number. This is still a considerable number and enough data to achieve good fits for the great majority of indicators. Rather strict statistical criteria had been set in order to accept or reject each single fit at each step. Namely, a significance level for the t- and F- tests of $\alpha = 2.5\%$, a maximal time-averaged final fit error of each I_i of 10% and a maximal condition number and VIF on the design matrix of ten and five, respectively.

In order to test the forecasting capability of the MLR-network model (1), we remove the year-2012 data before performing the fit to use it for an out-of-sample test later. Taking the one-year time shift between regressor and regressand variables into account, we are left with a series of nine data points to do the fitting, which will turn out to be sufficient to do proper forecasting.

The resulting GBoPN is generated from the coefficient matrix $\boldsymbol{\beta}$, where an edge from indicator i to indicator j is drawn if $\beta_{ij} \neq 0$. This methodology gives us a non-trivial insight into selected macroeconomic indicators from network perspective, since it tells us that any of the initial 405 indicators is significantly coupled to at least one other indicator and that the resulting network covers all indicators for each of the 60 countries. This analysis does not show a separation into geographically localized clusters, which could have been a valid expected outcome, but rather it depicts a picture of a globally interacting multi-layered economic system. A feature which strongly underlines the importance of cross-border economic relations is that basically all best-fit edges (more than 98.4%) are connecting indicators from different countries and about three quarter of the links connects indicators from different classes, such as trade and FDI.

The GBoPN turns out to be extremely sparse with an edge density[1] of less than 0.5%. This means that a very small number of relations between all possible pairs of indicators is enough to describe nearly all indicators with high accuracy. A great majority of 298 indicators are *tracked* by two indicators or even just one. On the other side of the spectrum, there is a smaller number of indicators which *track* a large number of other indicators. This, in the language of graph theory, means that these indicators have a large out-degree and might prove potentially useful for the purpose of macroeconomic monitoring and forecasting because their observation is expected to provide information about other indicators and their host-countries.

2.3 Tracking Centrality of Macroeconomic Indicators

An additional criterion for identifying nodes of interest is the monetary value of indicators to which they point, because the analyzed indicators vary by several orders of magnitude. To account for the number and size (in terms of their time-averaged values in USD) of a node's regressands, we define its *tracking centrality* as

[1] The number of maximally possible edges $N(N-1)$ dived by the number of actual edges, which, in our case, is the number of non-zero coefficient β_{ij}.

$$T_i \equiv \sum_{j=1, j\neq i}^{N} \sqrt{R_{ij}^2}\, S_j \,, \tag{2}$$

were R_{ij}^2 is the *coefficient of determination* between indicators i and j, and S_j is the time-averaged monetary value of indicator i. By definition, $R_{ij}^2 \neq 0$, whenever $\beta_{ij} \neq 0$. $\sqrt{R_{ij}^2}$ equals the absolute value of the Pearson product correlation coefficient between indicators i and j and measures the fraction of variation which is mutually described.

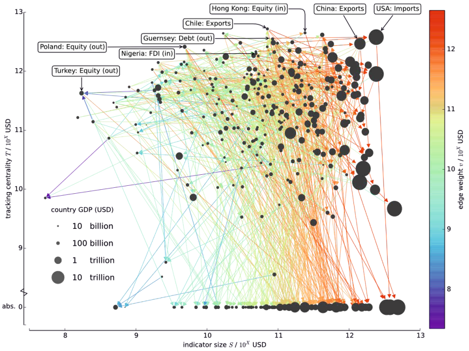

Fig. 2. Visualization of the **Global Balance of Payments Network** (GBoPN), constructed with the MLR-fit algorithm described in section 2. Indicators (nodes) are represented by dots with sizes corresponding to countries' GDP. The x- and y-positions of each node are set according to the indicator's time-averaged size S and tracking centrality T, respectively. Directed edges, pointing from a regressor i to its regressand(s) j, are color-coded according to the tracked value v_{ij}. A key-result is the high tracking capabilities of many indicators of small sizes or from small economies, which might be used as proxies or "thermometers" for larger economic changes. Highlighted examples include the outbound equity securities positions of Poland or the outbound debt securities positions of Guernsey. Indicators which have not been identified as statistically significant regressors, but are tracked by others, are positioned on the x-axis.

A comprehensive visualization of the entire GBoPN is shown in Fig. 2. Each indicator is represented by a dot of the size of its host-country's GDP. The x-position of each node is its size S and the y-position is its tracking centrality T, where a higher value means that a larger amount of economic value is quantitatively described by this indicator. The edges of the GBoPN are given through color-coded arrows pointing from a regressor to its regressand(s). The color is set according to the relation $v_{ij} = \sqrt{R_{ij}^2}\, S_j$, quantifying the value of this relationship. As such, the tracking centrality of an indicator is given by the values of its outbound connections. A key-result from Fig. 2 is the high tracking capabilities of many indicators of small sizes or from small economies, which might, in turn, be used as proxies or “thermometers” for larger economic changes. Such nodes can be easily spotted as having strong connection spanning a large horizontal distance (several orders of magnitude) from left to right. Highlighted examples include the outbound equity securities positions of Poland or the outbound debt securities positions of Guernsey. We note that the so-called off-shore financial centers [13] have high tracking centralities on average, due to their inherently strong coupling to the financial system.

Given the small average fit error, it is possible to make explicit use of the network structure, taking higher-order interactions into account, and track indicators over short paths, where the averaged errors are small. The maximally expected error is then the sum over all contributing errors along this path. A simple but instructive example is the tracking of Spain’s inbound FDI, using Greece’s imports with an aggregated error of 4.5% over 2002-2011, which is the only regressor in this case.

Besides having some outliers, the median forecasting error for the 372 macroeconomic indicators turns out to be 8.5% with an standard deviation of 15.5% excluding 3 indicators which have forecasting error of more than 100%. The forecasting capabilities described in this study are especially powerful for international trade, because trade flows are observed to lag behind cross-border investment positions, as seen in Fig. 1.

3 Conclusion

International trade is one of the main catalysts of globalization and social and economic development. The volume of global trade grew by more than a factor of five, in nominal terms, over the past 20 years[9], thereby closely linking up many of the involved countries and forming an international web of trade. In this sense, the investigation of the relation of global trade and economic growth is highly suited for the application of techniques from network science. In this study we use trade and financial (investment) flows between individual economies (nodes) to construct MLR-fit based edges of a *Global Balance of Payments Network* (GBoPN).

We introduce an MLR-fit algorithm to model multiple links between individual components of balance of payment accounts of a group of countries, which encompasses the majority of global macroeconomic activity. The derived network model delivers quite accurate description of most indicators, while the built-in time shift between regressors and regressands lead to good indicator forecasts. We have introduced the concept of an indicator’s tracking centrality, which allowed for the identification of “macroeconomic thermometers”, i.e. small indicators which track multiple large indicators with a relatively high precision. This novel methodology, in combination with other economic models, could be used as a monitoring tool of complex processes of economic cross-border interactions.

The methodology that we have developed here is not meant to stand alone, but to be merged with established economic analytics tools to generate a network-based approach in understanding global economic interactions. Similarly to [14], where the human organism is described as integrated network of complex physiological systems characterized by distinct network structure and topology, here we study the economic "organism" as dynamic complex network. By investigating the global balance of payment system as complex network, we introduce novel aspect of econometric analyses that may contribute to emergence of new research filed of *network economics.*

Acknowledgments. We would like to thank the European Commission FET Open Project FOC 255987 and FOC-INCO 297149 for financial support.

References

1. Catanzaro, M., Buchanan, M.: Network opportunity. Nature 9(3), 121–123 (2013)
2. Crotty, J.R.: Structural causes of the global financial crisis: A critical assessment of the 'new financial architecture'. Camb. J. Econ. 33(4), 563–580 (2009)
3. Preis, T., Kenett, D.Y., Stanley, H.E., Helbing, D., Ben-Jacob, E.: Quantifying the behavior of stock correlations under market stress. Sci. Rep. 2, 752 (2012), doi:10.1038/srep00752
4. Stulz, R.M.: Credit default swaps and the credit crisis. J. Econ. Perspect. 24(1), 73–92 (2010)
5. FuturICT: Global Computing for Our Complex World, `http://www.futurict.eu/`
6. Constancio, V.: Contagion and the european debt crisis. Financial Stability Review (16), 109–121 (2012)
7. Hartmann, P., Straetmans, S., de Vries, C.G.: Asset market linkages in crisis periods. The Review of Economics and Statistics 86(1), 313–326 (2004)
8. Kolanovic, M., et al.: Rise of cross-asset correlations. Global equity derivatives & delta one strategy report, J.P. Morgan Securities LLC (2011), `http://www.cboe.com/Institutional/JPMCrossAssetCorrelations.pdf` (accessed: November 15, 2013)
9. United Nations Statistics Division: `http://comtrade.un.org`: United Nations Commodity Trade Statistics Database (UN comtrade) (accessed: November 15, 2013)
10. United Nations Conference on Trade and Development Statistics (UNCTAD STAT): Inward and outward foreign direct investment stock, annual (1980-2012), `http://unctadstat.unctad.org/ReportFolders/reportFolders.aspx?sRF_ActivePath=P,5,27&sRF_Expanded=,P,5,27` (accessed: November 15, 2013)
11. International Monetary Fund: Coordinated Portfolio Investment Survey (Table 8): `http://cpis.imf.org` (acessed: April 1, 2013)
12. The World Bank, Data: `http://data.worldbank.org/indicator`: GDP (current USD), Stocks traded, total value (% of GDP), GDP deflator (annual %) (accessed: April-November 2013)
13. Zoromé, A.: Concept of offshore financial centers: In search of an operational definition. IMF Working Papers 07/87, International Monetary Fund (IMF) (2007): `http://www.imf.org/external/pubs/ft/wp/2007/wp0787.pdf` (accessed: November 15, 2013)
14. Bashan, A., Bartsch, R.P., Kantelhardt, J.W., Havlin, S., Ivanov, P.C.: Network physiology reveals relations between network topology and physiological function. Nature Communications 3, 702 (2012)