SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
CELL BIOLOGY
COMPUTER SCIENCE
COMPUTATIONAL SCIENCE

Received
17 February 2014

Accepted
13 March 2014

Published
1 April 2014

Correspondence and requests for materials should be addressed to N.P. (n.przulj@imperial.ac.uk)

Revealing the Hidden Language of Complex Networks

Ömer Nebil Yaveroğlu[1], Noël Malod-Dognin[1], Darren Davis[2], Zoran Levnajic[1,6], Vuk Janjic[1], Rasa Karapandza[3], Aleksandar Stojmirovic[4,5] & Nataša Pržulj[1]

[1]Department of Computing, Imperial College London, UK, [2]Computer Science Department, University of California, Irvine, USA, [3]Department of Finance, Accounting & Real Estate EBS Business School, Germany, [4]National Center for Biotechnology Information (NCBI), USA, [5]Janssen Research and Development, LLC, Spring House, PA, USA, [6]Faculty of Information Studies in Novo mesto, Novo Mesto, Slovenia.

Sophisticated methods for analysing complex networks promise to be of great benefit to almost all scientific disciplines, yet they elude us. In this work, we make fundamental methodological advances to rectify this. We discover that the interaction between a small number of roles, played by nodes in a network, can characterize a network's structure and also provide a clear real-world interpretation. Given this insight, we develop a framework for analysing and comparing networks, which outperforms all existing ones. We demonstrate its strength by uncovering novel relationships between seemingly unrelated networks, such as Facebook, metabolic, and protein structure networks. We also use it to track the dynamics of the world trade network, showing that a country's role of a broker between non-trading countries indicates economic prosperity, whereas peripheral roles are associated with poverty. This result, though intuitive, has escaped all existing frameworks. Finally, our approach translates network topology into everyday language, bringing network analysis closer to domain scientists.

Detecting and interpreting the patterns of change in complex networks may yield insight into their underlying function, emergent properties, and controllability[1,2]. However, this is a challenging task, since a complete comparison between complex networks has long been known to be computationally intractable[3]. Hence, simple heuristics, commonly called network *properties* or network *statistics*, such as the degree distribution, have been used to approximately say whether the structure of networks is similar[4]. The most sophisticated statistics are based on graph spectra[5,6] and small subnetworks including network motifs[7] and graphlets[8]. However, none of the current methods are sufficient for characterizing the structure and extracting information hidden in the topology of complex networks.

Real-world networks often have few types of nodes with well defined topological characteristics, also called *roles*. For example, the set of driver nodes that can control and move the networks into specific states has been identified and shown to be of low degree[1]. Also, world trade networks are proposed to have a core-periphery structure, with some countries (nodes in the network) being at the dense core, forming rich-clubs of trading countries, while others are at the sparsely connected periphery[9]. Such node roles are differently correlated in different types of networks[9]. Hence, we seek to design a method that will reveal and exploit these phenomena.

We cannot utilize graph spectra to design such a method, since spectra do not provide a direct real-world interpretation of network structure[5]. While *network motifs* and their spectra[7,10] can be used to define node roles, their interpretation is highly dependent on the choice of a network null model, which limits their usability[11]. This is because network motifs are defined as small partial subgraphs that are overrepresented in the real network compared to a chosen network null model; a partial subgraph means that once you pick a set of nodes in the large network, you can pick any subset of edges between them. *Graphlets* do not suffer from these drawbacks, can be used to define node roles and to design methods for linking the network structure with real-world function[12]. They are defined as small induced subgraphs of a large network that appear at any frequency and hence are independent of a null model (denoted by G_0 to G_{29} in Fig. 1 d); an induced subgraph means that once you pick the nodes in the large network, you must pick all the edges between them to form the subgraph. We define and utilize the correlations between graphlets (detailed below) to create a superior network measure that, unlike other simple or complex measures, makes network structure directly interpretable and provides its clear translation into everyday language. As such, our new measure can uncover novel relationships between seemingly unrelated networks from different domains. Furthermore, it can be used to track the dynamics and explain the evolution of any network, which we demonstrate on the world trade network example (Fig. 1 a-c). Our methodology is

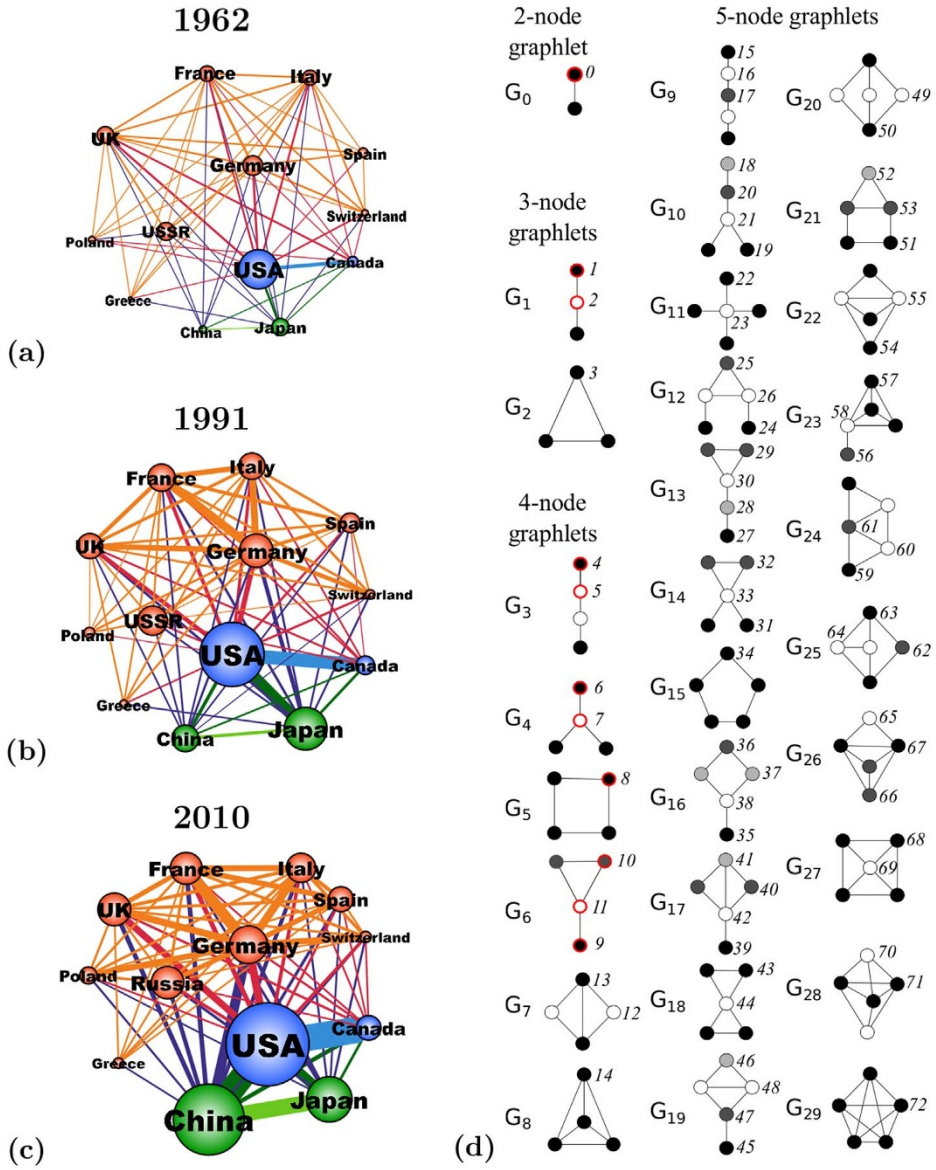


Figure 1 | Illustrations of subnetworks of the world trade networks. (a) 1962, (b) 1991, and (c) 2010. Node colors correspond to the continents: orange for Europe, green for Asia, blue for America. The node size corresponds to the GDP of the country. The edge thickness corresponds to the volume of the trade between the countries. (d) The thirty 2- to 5-node graphlets G_0, G_1, G_2, …, G_{29}. In each graphlet, nodes belonging to the same automorphism orbit are of the same shade. The 73 automorphism orbits of the 30 graphlets are labelled from 0 to 72[8]. Some orbits are redundant (their counts in a network can be derived from the counts of other orbits); the 11 red orbits illustrate the non-redundant ones for up to 4-node graphlets – there are several ways to choose non-redundant orbits, but that choice does not impact further analysis.

universal and can provide insight in all areas of science that use network theory, including biology, medicine, social sciences, and security.

New network statistic: Graphlet Correlation Distance. The distributions of graphlet frequencies in networks have been compared in the network statistic called Relative Graphlet Frequency Distribution[13]. To increase sensitivity at the same computational cost, *symmetry groups* of nodes within graphlets, called *automorphism orbits* [For a node x of network G, the automorphism orbit of x is the set of nodes of G that can be mapped to x by an automorphism, an isomorphism of a network with itself; i.e., a bijection of nodes that preserves node adjacency. Automorphism orbits of graphlets are illustrated as 0 to 72 in Fig. 1 d.], have been used to generalize the degree distribution into the spectrum of 73 Graphlet Degree Distributions that correspond to the 73 orbits for up to 5-node graphlets: the first of these distributions is the familiar degree distribution, the second gives the number of nodes in the network that touch k orbits 1 of graphlet G_1 for all values of k, etc. for all 73 orbits. Then, Graphlet Degree Distributions of two networks are compared over all orbits in the network statistic called Graphlet Degree Distribution Agreement[8]. A related concept is that of the Graphlet Degree Vector of a node that has been used to link wiring around a node with its real-world function[12]: it has 73 coordinates, each of which measures the number of times the node is touched by a particular orbit of a graphlet (so the first coordinate is the degree of the node, the second is the number of 3-node paths that it touches at an end node etc.).

We design a superior graphlet-based measure by identifying and eliminating redundancies and exploiting dependencies between orbit counts in a network. For example, if we denote by C_i the i^{th} *graphlet degree* of a node (where $i \in \{0, 1, \ldots, 72\}$, Fig. 1 d), which is the number of times the node is touched by orbit i[8], then if we consider

the degree of the node, C_0, we can argue as follows. The neighbours of the node are either connected, or they are not: if they are connected, then they contribute to counts of triangles, C_3, that the node touches; if they are not connected, then they contribute to C_2 for the node, the number of times the node is touched by the middle of a 3-node path (orbit 2 in graphlet G_1, Fig. 1 d). Since these are the only options for connectedness of neighbours of a node, the number of ways in which C_0 neighbours of the node can be connected, $\binom{C_0}{2}$, is equal to the sum of C_2 and C_3 for the node: $\binom{C_0}{2} = C_2 + C_3$. Hence, if we know two of C_0, C_2 and C_3, we can derive the third, so one of them is redundant and does not need to be included in graphlet-based statistics. Similarly, we obtain a system of 17 linear equations describing all orbit redundancies (see Supplementary Information). When we solve it for the 73 orbits, 56 orbits remain non-redundant. There are 15 orbits for up to 4-node graphlets and 11 are non-redundant (red ones in Fig. 1 d). Similar redundancies, but in orbits of partial 4-node subgraphs have been reported[14].

We identify and exploit the dependencies (correlations) between graphlets as follows. First, we note that there are fewer dependencies between the 11 non-redundant orbits for up to 4-node graphlets than between the 56 non-redundant orbits for up to 5-node graphlets (also see below). Hence, they introduce less noise in the corresponding new network statistic, so we construct a network statistic using the 11 non-redundant orbits for up to 4-node graphlets. However, we contrast it with analogous statistics that include redundant and up to 5-node graphlet orbits as well.

Then, we devise a network statistic based on correlations between various node properties contained in non-redundant orbit counts, over all nodes, as follows. For each node in a network, first we construct its Graphlet Degree Vector consisting of 11 coordinates corresponding to the 11 non-redundant orbits. Then we construct a matrix whose rows are the above described Graphlet Degree Vectors, so the number of rows in the matrix is equal to the number of nodes in the network and it has 11 columns. The existence of correlations between non-redundant orbits over all nodes is exploited for constructing a new network statistic: for a given network N_1, we compute Spearman's Correlation coefficients between all pairs of columns of the above described matrix and present them in an 11×11 symmetric matrix that we term the *Graphlet Correlation Matrix (GCM)* of network N_1, GCM_{N_1}. In this way, we can summarize the topology of a network of any size into an 11×11 symmetric matrix with values in the interval $[-1, 1]$ (illustrated in Supplementary Fig. S2).

Different real and model networks generally have very different orbit dependencies and hence very different GCMs (Fig. 2 a–d). For example, in agreement with known properties of scale-free Barabási-Albert (SF-BA) networks[15], orbits 0, 2, 5, and 7, which are characteristic to existence of hubs, form a cluster of dependent orbits (as illustrated by their correlation coefficients being close to 1 in Fig. 2 a); also, orbits 10 and 11, which are characteristic to existence of clustering near hubs, form a cluster of dependent (i.e., correlated) orbits; and finally, orbits 1, 4, 6, and 9, characteristic to existence of a large number of degree 1 nodes, are dependent as well. The picture is quite different for geometric random graphs (GEO)[16] of the same size, which have Poisson degree distributions and hence the structure not dominated by a large fraction of degree 1 nodes and a small number of hubs (Fig. 2 b) (see Supplementary Information).

Uncovering orbit dependencies in real-world networks is much more interesting, since they can reveal currently unknown organisational principles of these networks. Indeed, the world trade network of 2010[17] (explained in Supplementary Information) contains two large clusters of dependent orbits, {0, 2, 5, 7, 8, 10, 11} and {4, 6, 9}, while there is no correlation between orbits {4, 6, 9} and orbits {0, 2, 5, 7, 8, 10, 11} (Fig. 2 c). We ask what this means and notice that orbits 4, 6 and 9 correspond to *peripheral*, degree 1 nodes that are "hanging" from graphlets G_3, G_4 and G_6 (Fig. 1 d), while members of the large cluster of correlated orbits, {0, 2, 5, 7, 8, 10, 11}, correspond to higher degree, either clustered (in a densely linked neighbourhood), or *broker*-type (*mediators* between nodes that are not directly linked) orbits. Since these two clusters are not correlated, we can conclude that countries are either clustered/brokers, or on the periphery of world trade, but not both. Hence, GCM unveils a hidden structure of this network that can be further interpreted qualitatively: through further analysis presented below, we interpret this observation on 49 world trade networks corresponding to trade data from 1962 to 2010. In contrast, the topology of the human metabolic network (see Supplementary Information) is very different from the topology of world trade networks: the correlations between all orbits are high, indicating that constituent bio-molecules can be at the same time both peripheral and clustered/broker (Fig. 2 d).

In addition to in-depth examination of network topology that can be qualitatively interpreted, the demonstrated differences in GCMs enable us to define a new measure of distance between topologies of two networks. For networks N_1 and N_2, we define their network *distance* by taking the Euclidean distance of the upper triangle values of GCM_{N_1} and GCM_{N_2} and we term it *Graphlet Correlation Distance (GCD)* between two networks. GCD is clean of redundancies and elegantly encodes much information about local network topology. We demonstrate that it outperforms other measures both on synthetic and real networks and we illustrate its utility on real-world problems (detailed below).

Evaluation on synthetic and real networks. To evaluate the performance of GCD for clustering networks of the same type, first we compare its results to those produced by other network statistics on synthetic networks belonging to seven different, commonly used, network models: Erdös-Rènyi random graphs (ER)[18], generalized random graphs with the same degree distribution as the data (ER-DD)[4], Barabási-Albert scale-free networks (SF-BA)[15], scale-free networks that model gene duplications and mutations (SF-GD)[19], geometric random graphs (GEO)[16], geometric graphs that model gene duplications and mutations (GEO-GD)[20], and stickiness-index based networks (STICKY)[21]. For each model, we generate 30 networks for each of the following four numbers of nodes and three edge densities that mimic the sizes and densities of real-world networks: 1000, 2000, 4000, and 6000 nodes, and 0.5%, 0.75%, and 1% edge density. Hence, the total number of synthetic networks that we compare using GCD (and other network statistics) is $7 \times 4 \times 3 \times 30 = 2,520$. Once we find GCD distances between all pairs of the 2, 520 networks, to illustrate the grouping (clustering) of these networks produced by GCD (a formal evaluation is presented below), we use the standard method of multi-dimensional scaling (MDS)[22] and embed the 2, 520 networks as points into 3-dimensional space so that their GCD distances are preserved as best as possible (Fig. 2 e). As illustrated in Fig. 2 e, networks belonging to the same model are grouped together in space regardless of size and edge density; model networks of the same size and density are grouped even better (Supplementary Fig. S5).

To illustrate its performance for grouping real networks from the same domain, we compute GCDs between all pairs of 11, 407 real-world networks from five different domains: 733 autonomous networks of routers that form the Internet, Facebook networks of 98 universities, metabolic networks of enzymes of 2, 301 organisms, 8, 226 protein structure networks, and 49 world trade networks corresponding to years 1962 to 2010 (detailed in Supplementary Information). As before, GCD-based MDS embedding of the 11, 407 networks shows clear groupings of networks from the same domain (Fig. 2 f). We interpret the grouping and the evolution of the world trade networks later in the text.

We formally evaluate the performance of GCD for clustering networks from the same model or real-world domain and systematically compare it to the performance of six other commonly used, or

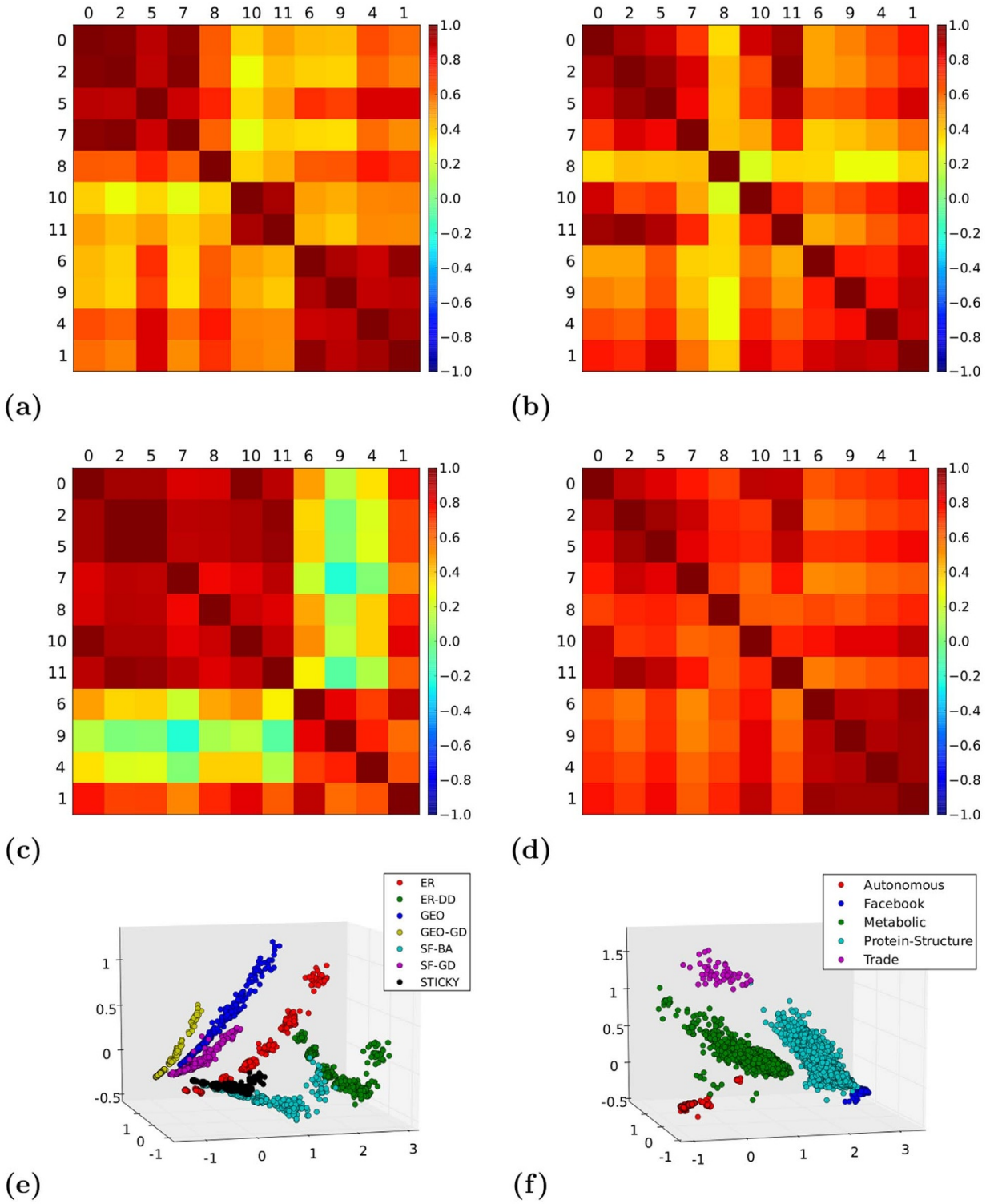


Figure 2 | **Illustrations of GCMs.** (a) a scale-free Barabàsi-Albert (SF-BA) network with 500 nodes and 1% edge-density; (b) a geometric random network (GEO) of the same size and density as network in (a); (c) the world trade network of 2010; and (d) the human metabolic network. Note that for SF-BA, GEO and metabolic networks, all the orbit correlations are statistically significant (p-values ≤ 0.05). This is not the case in the world trade network, where some correlations involving orbits 4 and 9 (the green cells in the GCM) have larger p-values. Illustrations of Graphlet Correlation Distance-based clustering of: (e) the 2,520 networks of various sizes and densities from 7 different random graph models: ER (red), ER-DD (green), GEO (blue), GEO-GD (yellow), SF-BA (light blue), SF-GD (purple), STICKY (black); (f) 11,407 real-world networks from 5 different domains: autonomous systems (red), Facebook (blue), metabolic networks (green), protein structure networks (light blue), and world trade networks (purple).

sensitive and robust network comparison measures (see Supplementary Information): degree distribution[4], clustering coefficient[4], network diameter[4], spectral distance[5], Relative Graphlet Frequency Distribution[13], and Graphlet Degree Distribution Agreement[8]. To make the comparison complete and evaluate what is gained by exclusion of redundant orbits or 5-node graphlets, we present comparison of the performance of GCD that includes the 11 non-redundant orbits for up to 4-node graphlets (that we term GCD-11) with the

performance of GCD constructed by using the full set of 73 orbits for all up to 5-node graphlets (termed GCD-73). Also, we make comparisons with GCDs constructed from all 15 orbits of up to 4-node graphlets (GCD-15) and from the 56 non-redundant orbits of up to 5-node graphlets (GCD-56): GCD-11 outperforms all measures for comparing networks of similar size and density, which is the most relevant for modelling network data, as models need to mimic sizes and densities of the data (see Supplementary Information).

In particular, one can test how well a distance measure groups networks of the same type by using the standard Precision-Recall curve: for small increments of parameter $\epsilon > 0$, if the distance between two networks is smaller than ϵ, then the pair of networks is retrieved. For each ϵ, precision is the fraction of correctly retrieved pairs (i.e., grouping together two networks from the same model), while recall is the fraction of the correctly retrieved pairs over all correct ones. The Area Under the Precision-Recall curve (AUPR), also called *average precision*, standardly measures the quality of the grouping by a given distance measure. We chose Precision-Recall curve analysis as it is known to be more robust to large numbers of negatives (in our case, negatives would be pairs of networks from different models that are grouped together) than Receiver Operator Characteristic (ROC) curve analysis[23].

Precision-Recall curves show that GCD-11 is the most precise among all tested measures (Fig. 3 a–b). Since the closest objects are the first to start forming clusters, we are interested in distance measures that optimize the number of correctly clustered pairs of networks that are at the shortest distance and hence are *retrieved first* by the distance measure[24]. Both GCD-11 and GCD-73 exhibit superiority in early retrieval over all other measures (beginning of the curves in Fig. 3 a). GCD-11 outperforms GCD-73 in this regard, because it contains fewer orbit dependencies and also has no redundancies, which introduce noise in GCD-73. Hence, GCD-11 is clearly the most sensitive measure for clustering networks. In addition, it is computationally efficient, since it involves counting only up to 4-node graphlets (see Supplementary Information).

Robustness to noise and missing data. Since real networks are noisy and incomplete, we evaluate the clustering quality of the above distance measures in the presence of noise. To simulate noise, we would like to randomize each of the above described 2, 520 synthetic networks 30 times (detailed below). However, if we were to randomize each of these networks 30 times, evaluating the results on the set of $2,\ 520 \times 30 = 75,\ 600$ networks would be computationally prohibitive. Hence, we use a subset of 280 out of the 2, 520 synthetic networks: for each of the 7 network models, we generate 10 networks for each of the following node sizes and edge densities: 1000 and 2000 nodes, and 0.5% and 1% edge density. We use these node sizes and edge densities because they correspond to networks that are more difficult to cluster than larger networks, so that if we show the methodology to be robust under these stringent conditions, we can be confident that it will be robust on real-world networks as well.

To simulate noise, we randomly rewire up to 90% of edges in the model networks of various sizes and densities described above and show that on these rewired networks, GCD-11 outperforms all other measures with respect to AUPR (Fig. 3 c; numbers on the vertical axis are not the same as those in column 2 of Fig. 3 b, since they correspond to the 280 networks described above, while those in Fig. 3 b correspond to the full set of 2, 520 networks). Similarly, it outperforms other measures on networks with missing data, which we simulate by randomly removing up to 90% of edges from model networks (Supplementary Fig. S7 a). Since many real networks are both noisy and incomplete[25,26], we ask how robust the measures are to missing edges in the presence of noise in the data. To answer that, we first randomly rewire 40% of edges in model networks to simulate noise and then randomly remove a percentage of edges to simulate missing data in the noisy networks. Again, GCD-11 outperforms all other measures even for networks with 40% of random noise that are missing up to 80% of edges at random (Fig. 3 d).

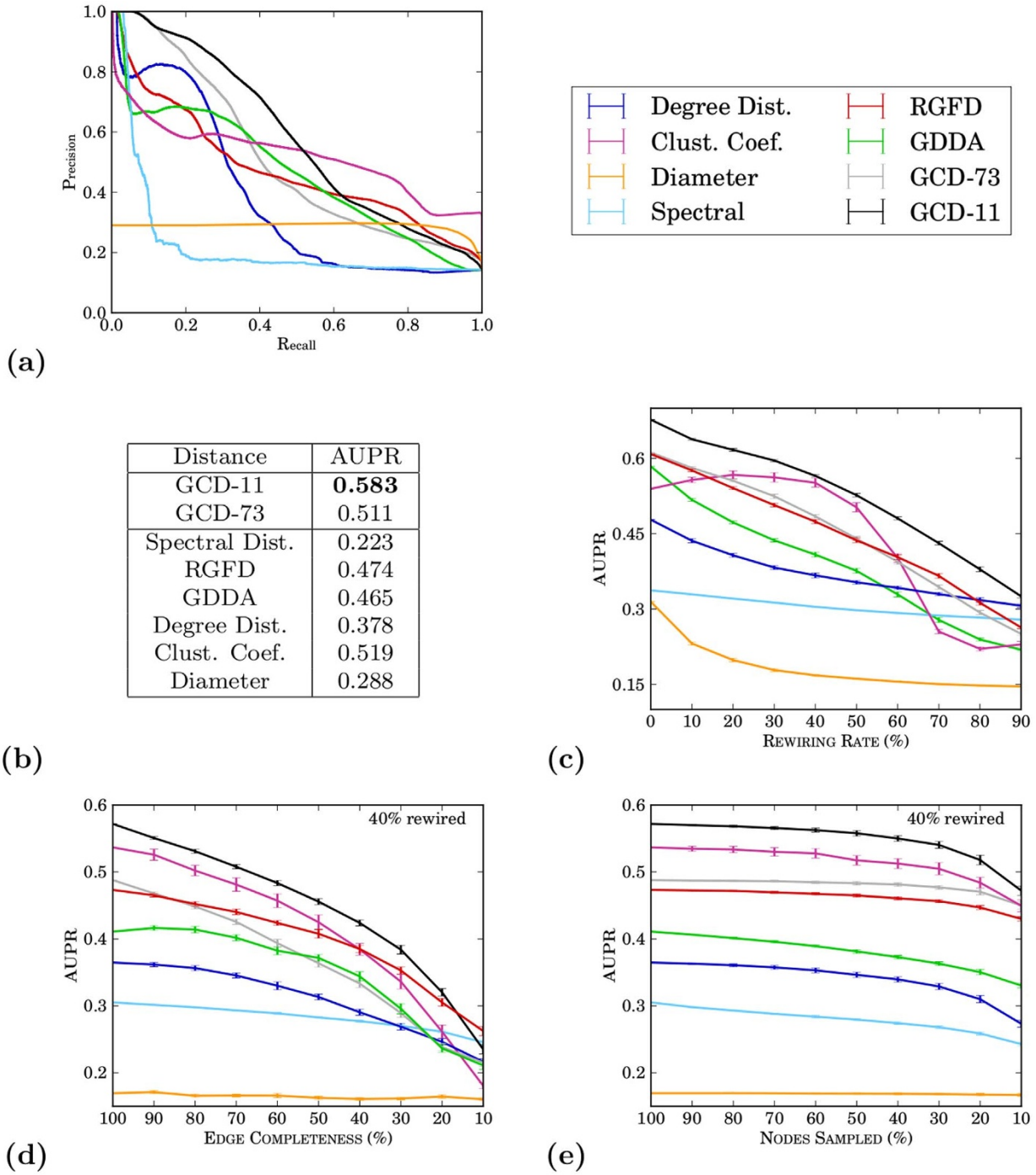
Furthermore, a surprising speed up in computational time can be obtained without loss in the clustering quality: by taking Graphlet Degree Vectors of as few as 30% of randomly chosen nodes in a model network to form GCM-11 (instead of taking Graphlet Degree Vectors of all nodes in the network), AUPR of GCD-11 only slightly decreases compared to when all nodes are used, and also it outperforms all other measures (Supplementary Fig. S7). In addition, for noisy and incomplete networks described above, the clustering obtained by GCD-11 not only outperforms those obtained by all other measures, but also it does not deteriorate even if we randomly sample as few as 30% of Graphlet Degree Vectors to form GCD-11 (Fig. 3 e) (see Supplementary Information).

These tests demonstrate robustness to noise and missing data and superiority of GCD-11 over other measures on a wide array of different network topologies, sizes and edge densities. The results improve further if we consider only networks of the same size and density (Supplementary Fig. S6).

World trade and other real network examples. Since GCM-11 is fast to compute and superior to other measures for clustering diverse networks even in the presence of large amounts of noise, we apply it to real networks in several domains.

Modelling networks from five domains. We use GCD-11 as a distance measure to evaluate the fit of network models to the above described 11, 407 real-world networks from five different domains. We use the state-of-the art non-parametric test to evaluate the fit[27,28] (Supplementary Fig. S8). Surprisingly, we find that networks from very different domains, Facebook, metabolic, and protein structure, are all best modelled by three network models: geometric random graphs (GEO), geometric graphs that mimic gene duplications and mutations (GEO-GD), and scale-free networks that also mimic gene duplications and mutations (SF-GD). While it is not difficult to explain why biological networks are the best fit by networks that model evolutionary processes, it may be surprising that Facebook networks seem to be organized by the same principles. A possible explanation is that Facebook grows as follows: when a person joins Facebook, he/she links to a group of his/her friends, which mimics a gene duplication, but he/she hardly ever has exactly the same friends as another person, which mimics the evolutionary process of divergence, or mutation. The fit of GEO to both Facebook and biological networks is perhaps more straightforward to explain, since all biological entities are subject to spatial constraints[20].

Crises and the topology of the world trade network. To gain insight into the relationship between economic crises and the world trade network (WTN), we apply GCD-11 to examine the dynamic changes of the WTN from year 1962 to 2010. We ask if rewiring of the WTN happens during crises and seek potential causes for the rewiring, or impacts of the rewiring. In particular, we test for correlations between time series of crude oil price changes and the topological changes in the WTN obtained by GCD-11. We shift these time series by up to 3 years forward and backward in time to see whether the change of WTN follows the change of oil price or vice versa and in what time interval. We test for all year shifts in $\{-3, -2, -1, 0, 1, 2, 3\}$ and report only statistically significant correlations (p-value ≤ 0.05) by using Spearman's correlation coefficients, which take into account the size of the change, and Phi correlation coefficients, which detect upward or downward trends only. Also, to cope with yearly data variability, we test the above correlations by grouping years in blocks of size 1, 2, and 3 years and report only statistically significant correlations (see Supplementary Information): for example, for changes in oil price in block sizes of 2 years for year 1990, we group year 1990 with the previous year, 1989, and find the average of the absolute



Distance	AUPR
GCD-11	**0.583**
GCD-73	0.511
Spectral Dist.	0.223
RGFD	0.474
GDDA	0.465
Degree Dist.	0.378
Clust. Coef.	0.519
Diameter	0.288

Figure 3 | Quality of clustering the 2,520 model networks using eight network distance measures (color coded and listed in the top panel). RGFD denotes Relative Graphlet Frequency Distribution and GDDA denotes Graphlet Degree Distribution Agreement. Error bars in panels (c) to (e) are one standard deviation above and below the mean. (a) Precision-Recall curves. (b) For each distance measure (the first column), the Area Under the Precision-Recall curve (AUPR, second column) achieved by a distance measure. (c) AUPR for different percentages of noise (randomly rewired edges, horizontal axis) in model networks (in 10% increments). (d) For "noisy" model networks, with 40% of edges randomly rewired, AUPR when x% of edges (horizontal axis) are kept in the network and 100 – x% are randomly removed (in 10% increments). (e) For "noisy" model networks, with 40% of edges randomly rewired, AUPR when a percentage of randomly sampled nodes (horizontal axis) is used to construct a distance measure; e.g., we obtain Graphlet Degree Distributions for x% of randomly chosen nodes to make up GCM-11 of the network and this is done for all networks before GCD-11 is computed between all pairs of networks.

values of the differences in oil prices between these two years and the two years that follow 1990, i.e., 1991 and 1992 [that is, $\frac{1}{4}(|price('91) - price('89)| + |price('91) - price('90)| + |price('92) - price('89)| + |price('92) - price('90)|)$].

We find that changes in crude oil price are correlated with changes in WTN topology and that they affect the WTN one and two years later (Fig. 4 a). Since WTN consists of trades in many commodities, different commodities are affected differently by the oil price (Supplementary Fig. S9 and Fig. S10), with the strongest and imme-

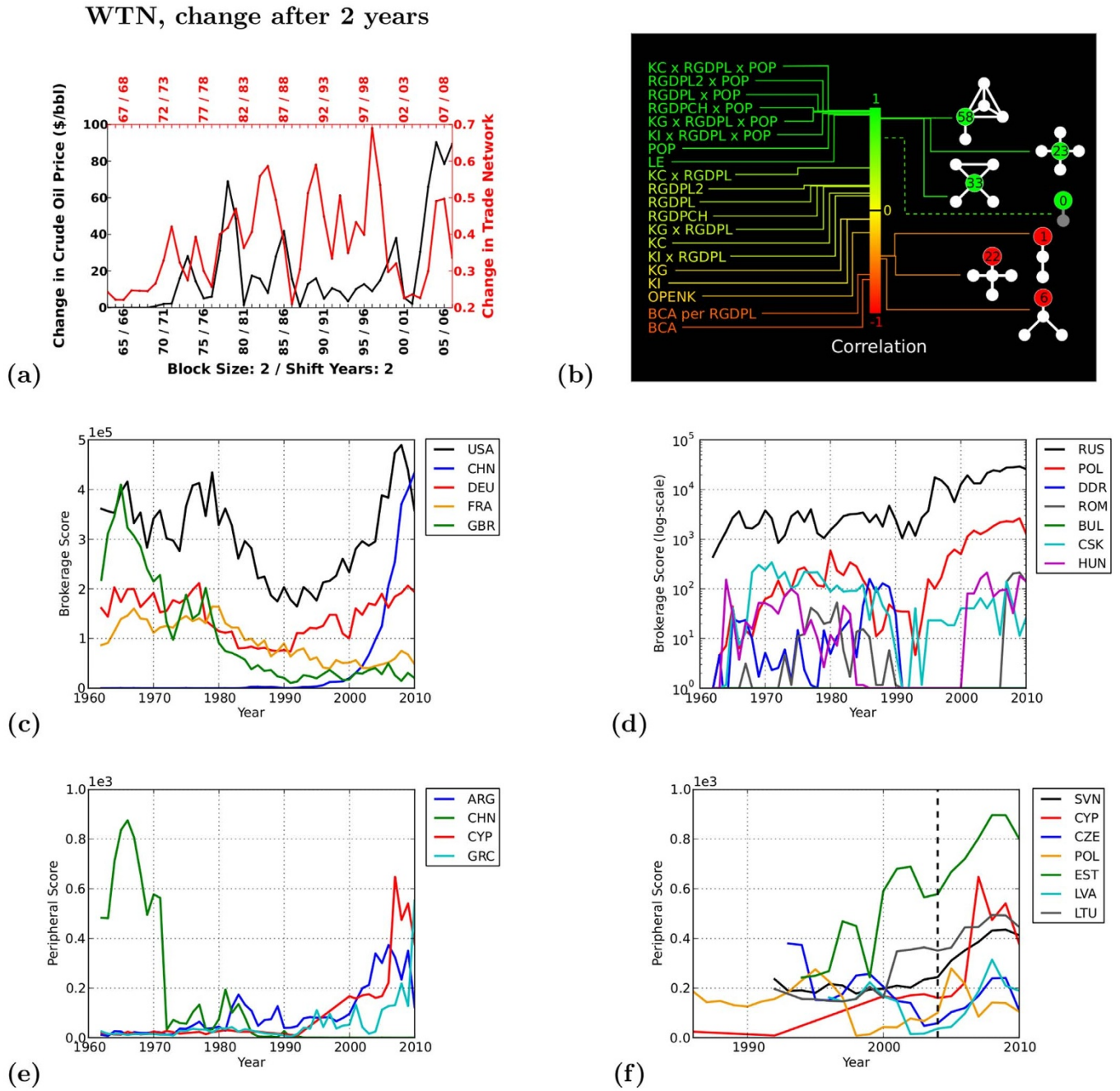


Figure 4 | Results of world trade network analysis. (a) Correlation of changes in crude oil price and changes in the structure of WTN, with the block size of two years and a two year shift; the Spearman correlation coefficient is 0.414 with the p-value of 0.005. (b) CCA correlations between economic attributes on the left (described in Supplementary Information) and graphlet degrees on the right; the middle bar is color-coded value of correlation. (c) Brokerage scores of the United States (USA), China (CHN), Germany (DEU), France (FRA), and the United Kingdom (GBR) from 1962 to 2010. (d) Brokerage scores of the Eastern Bloc from 1962 to 2010: the Soviet Union until 1991 replaced by Russia afterwards (RUS), Poland (POL), Eastern Germany (DDR), Romania (ROM), Bulgaria (BUL), Czechoslovakia until 1991 replaced by the sum of Czech Republic and Slovakia afterwards (CSK), and Hungary (HUN). (e) Peripheral scores of Argentina (ARG), China (CHN), Cyprus (CYP), and Greece (GRC) from 1962 to 2010. (f) Peripheral scores of countries that joined EU in 2004 and show an increase in their peripheral scores right before and after joining the EU: Slovenia (SVN), Cyprus (CYP), Czech Republic (CZE), Poland (POL), Estonia (EST), Latvia (LVA), and Lithuania (LTU).

diate effect (in the same year in which oil price changes) being on the trade of "Food and Live Animals" (Supplementary Fig. S10 a). This may be explained by agriculture needing oil, as well as by increase in demand for bio-fuels as oil price increases[29]. We further confirm this by observing that the correlation between oil price and the structure of the network of trade in "Food and Live Animals" increases over time, as agriculture becomes more oil dependent: Phi correlation coefficient rises from 0.31 in years 1962 to 1986, to 0.51 in years 1986 to 2007.

We ask if we can get similar results by using network similarity measures other than GCD-11. To that end, we seek for correlations between changes in crude oil price and changes in WTN structure measured by each of the above described network similarity measures: degree distribution, clustering coefficient, network diameter,

spectral distance, RGFD, and GDDA. The only relevant result is that GDDA uncovers a potentially interesting, but hard to explain correlation: a change in WTN structure (as reported by GDDA) is followed by a change in crude oil price 3 years later. Explaining this observation is a subject of future research. All other network similarity measures produce irrelevant correlations, such as "Beverage and Tobacco" trade network changes (observed by RGFD) correlating with changes in oil price two years later. The list of all correlations found by each of the similarity measures is available in the Supplementary Data.

We recall our previous observation about a country in the WTN being either peripheral or clustered/broker, but not both, and offer a qualitative explanation. In particular, we use the standard method of Canonical Correlation Analysis (CCA)[30] to correlate economic indicators of the development of a country[31,32] with its graphlet-based position in the WTN (see below). Interestingly, the indicators of economic wealth (e.g., gross domestic product, level of employment, consumption share of purchasing power parity; described in Supplementary Information) strongly correlate with a country being in a brokerage relationship (i.e., a mediator between unconnected countries), or within a cluster of densely connected countries, while the indicators of economic poverty (e.g., current account balance) correlate with a country being peripheral in the network, i.e., linked only to one other country by a trade relationship (Fig. 4 b). Since a country is either peripheral or clustered/broker, this may indicate that one of the factors that contribute to the wealth of a country could be its brokerage/clustered position in the WTN.

To evaluate if the above result linking GDP to a country's wiring in the WTN can be obtained by simpler, non-graphlet-based, previously used measures of node wiring, such as node degree, clustering coefficient, and betweenness centrality[33–35], we compute the Pearson's correlation coefficients (PCCs) between the GDPs of countries and each of these node statistics. To assess the quality of the CCA analysis described above, we measure the PCC between GDP and the graphlet degree of orbit 58, since it has the largest coefficient reported by CCA that links it to GDP. We demonstrate superiority of our method over others, since we find that orbit 58 outperforms all other statistics, achieving with GDP the PCC of 0.869, followed by betweenness centrality achieving PCC of 0.816, node degree (i.e., orbit 0) achieving PCC of 0.690, and clustering coefficient achieving PCC of -0.136. This demonstrates that our graphlet-based method finds more refined topological features than previously used methods[33–35] and that even betweenness centrality gives a more coarse-grained insight in to the function of WTN.

To quantify the strength of the brokerage position of a country in the WTN of each year, we define the *brokerage score* of the country in a particular year as the weighted linear combination of broker graphlet degrees (i.e., C_{23}, C_{33}, C_{44}, and C_{58}) using the coefficients obtained from CCA. Similarly, we quantify how *peripheral* a country is in the WTN of a particular year by using C_{15}, C_{18}, and C_{27}. Since we have demonstrated above that a country is either a broker or peripheral in each year, these brokerage and peripheral scores enable us to track changes in the position of a country in the WTN over years. We analyse if the changes in brokerage and peripheral scores of a country over years coincide with economic crises and other events impacting the economy of the country.

Indeed, we find that during 1980s, brokerage scores of the world's highest brokers fall (Fig. 4 c), for which we find support in the economics literature. For example, in the USA during the first Reagan administration, a mix of monetary policy and loose budgets sky-rocketed the dollar and sent international balances in the wrong direction. The merchandise trade deficit rose above $100 billion in 1984 and remained there throughout the decade. The ratio of the USA imports to exports during the eighties peaked at 1.64, a disproportion not seen since the War between the States. Such a drop in the export power of the USA, and thus the change of its position in the trade network (drop of its brokerage score in the WTN, black line in Fig. 4 c), had no precedent in modern USA history[36]. Another example is that of Great Britain. There is a huge drop in its brokerage score as it loses the Empire in the 1960s, seeing a small improvement in 1973 when the Conservative Prime Minister, Edward Heath, led it into the European Union (EU). However, the downward trend induced by the dissolution of the colonial superpower has continued[37]. In contrast, the reunification of Germany transformed it from being in the shadow of the Second World War a peripheral economy of Western Europe, with most of the decisions in Europe having been made by France and the UK, to being the central economy of Europe[38]. Among the countries of the former Eastern Bloc, USSR has been the most dominant broker, with both Russia and Poland sharply gaining in brokerage scores after the fall of communism (Fig. 4 d; y-axis is in logarithmic scale).

Similarly, peripheral scores (Fig. 4 e) are consistent with economic reality. China's peripheral score dropped sharply in the early 1970s, which coincides with President Nixon's international legitimization of China[39]. This was a turning point that changed China's closed economy to one deeply integrated with global financial markets[40], as evident not only by its fallen peripheral score (Fig. 4 e), but also by its increased brokerage score that has surpassed that of the USA in 2009 (Fig. 4 c). Conversely, raising peripheral scores of Argentina, Cyprus and Greece coincide with their recent economic crises. By year 2001, poor management in great part led to Argentina's real GDP shrinkage, unemployment sky-rocketed, and the international trade plunged, so Argentina turned into a peripheral economy[41]. Less than a decade later, Cyprus and Greece went the "South American way:" the similarities, starting with the fixed exchange regime followed by the bank runs, were striking[42].

Interestingly, accession of countries into the EU makes them more peripheral in the WTN, as evident by increases in their peripheral scores before and after accession (Fig. 4 f). Even though all trade within the EU is exempt from import taxes, at the time of accession new members are required to leave other advantageous free trade associations (e.g., BAFTA, CEFTA, CISFTA, EFTA). The fact that a country has to leave free trade agreements with other non-EU member countries leads to the destruction of trade connections while the positive effects of EU accession on trade need time to materialize. In other words, since trade connections are easy to break, but much more difficult to build, EU accession increases the peripheral score of a country and whether and when the country will recover remains an open question.

We assess if similar can be observed by other node measures previously applied to WTNs, such as betweenness and closeness centralities[35], and find that it cannot. In particular, we plot betweenness centrality of a country over years (and also its betweenness peripherality, that we define by subtracting from 1 the value of its betweenness centrality) and find that we cannot detect the events that can be detected by our brokerage and peripheral scores described above, such as the drop of export power of the USA, or the fall of Argentina, Cyprus and Greece (see Supplementary Fig. S11 a and b). Similarly, closeness centrality and peripherality (defined analogous to betweenness peripherality described above) can also not detect these events (Supplementary Fig. S11 c and d).

Final remarks on GCD. We have shown that by exploiting correlations between node characteristics in a network (e.g., broker/clustered and peripheral nodes in the world trade network), we can sensitively and robustly uncover the network type and track network dynamics. This is possible because real-world networks generally have few types of nodes with well defined characteristics and because the node characteristics are differently correlated in different types of networks. We have uncovered some of the node characteristics, in particular broker/clustered and peripheral ones in the world trade network, that are amenable to economic

interpretation. In particular, during a crisis, a country becomes more peripheral and less of a broker in the WTN than in economically stable periods. Accession of a country to the EU has a similar effect. The methodology promises to deliver insight in many other areas; for example, it can help detect online or telephone-based terrorist activities, because it can robustly and sensitively typify a newly formed network by identifying the most similar known network group.

1. Liu, Y.-Y., Slotine, J.-J. & Barabási, A.-L. Controllability of complex networks. *Nature* **473**, 167–173 (2011).
2. Galbiati, M., Delpini, D. & Battiston, S. The power to control. *Nat. Phys.* **9**, 126–128 (2013).
3. Cook, S. A. The complexity of theorem-proving procedures. In *Proceedings of the Third annual ACM symposium on Theory of Computing,* 151–158 (ACM, 1971).
4. Newman, M. *Networks: An Introduction* (Oxford University Press, Oxford, 2009).
5. Wilson, R. C. & Zhu, P. A study of graph spectra for comparing graphs and trees. *Pattern Recogn.* **41**, 2833–2841 (2008).
6. Thorne, T. & Stumpf, M. P. Graph spectral analysis of protein interaction network evolution. *J. R. Soc. Interface* **9**, 2653–2666 (2012).
7. Milo, R. *et al.* Network motifs: Simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
8. Pržulj, N. Biological network comparison using graphlet degree distribution. *Bioinformatics* **23**, 177–183 (2007).
9. Della Rossa, F., Dercole, F. & Piccardi, C. Profiling core-periphery network structure by random walkers. *Sci. Rep.* **3**, 1467 (2013).
10. Milo, R. *et al.* Superfamilies of evolved and designed networks. *Science* **303**, 1538–1542 (2004).
11. Artzy-Randrup, Y., Fleishman, S. J., Ben-Tal, N. & Stone, L. Comment on "Network motifs: simple building blocks of complex networks" and "Superfamilies of evolved and designed networks" *Science* **305**, 1107–1107 (2004).
12. Guerrero, C., Milenković, T., Pržulj, N., Kaiser, P. & Huang, L. Characterization of the proteasome interaction network using a qtax-based tag-team strategy and protein interaction network analysis. *Proc. Nat. Acad. Sci. U.S.A.* **105**, 13333–13338 (2008).
13. Pržulj, N., Corneil, D. G. & Jurisica, I. Modeling interactome: scale-free or geometric? *Bioinformatics* **20**, 3508–3515 (2004).
14. Marcus, D. & Shavitt, Y. RAGE – a rapid graphlet enumerator for large networks. *Comput. Netw.* **56**, 810–819 (2012).
15. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
16. Penrose, M. *Random geometric graphs* (Oxford University Press, Oxford, 2003).
17. United Nations, United nations commodity trade statistics (COMTRADE) database., (2010) (Date of access: 15/11/2011) URL: http://comtrade.un.org.
18. Erdös, P. & Rényi, A. On random graphs. *Publ. Math. Debrecen* **6**, 290–297 (1959).
19. Vázquez, A., Flammini, A., Maritan, A. & Vespignani, A. Modeling of protein interaction networks. *Complexus* **1**, 38–44 (2002).
20. Pržulj, N., Kuchaiev, O., Stevanovic, A. & Hayes, W. Geometric evolutionary dynamics of protein interaction networks. *Pac. Symp. on Biocomput.* **2009**, 178–189 (2010).
21. Pržulj, N. & Higham, D. J. Modelling protein–protein interaction networks via a stickiness index. *J. R. Soc. Interface* **3**, 711–716 (2006).
22. Cox, T. F. & Cox, M. A. *Multidimensional Scaling* (CRC Press, Florida, 2010).
23. Davis, J. & Goadrich, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning (ICML '06)*, 233–240 (ACM, New York, NY, USA, 2006).
24. Yu, Y.-K., Gertz, E. M., Agarwala, R., Schäffer, A. A. & Altschul, S. F. Retrieval accuracy, statistical significance and compositional similarity in protein sequence database searches. *Nucleic Acids Res.* **34**, 5966–5973 (2006).
25. Han, J.-D. J., Dupuy, D., Bertin, N., Cusick, M. E. & Vidal, M. Effect of sampling on topology predictions of protein-protein interaction networks. *Nat. Biotechnol.* **23**, 839–844 (2005).
26. Stumpf, M. P., Wiuf, C. & May, R. M. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Nat. Acad. Sci. U.S.A.* **102**, 4221–4224 (2005).
27. Rito, T., Wang, Z., Deane, C. M. & Reinert, G. How threshold behaviour affects the use of subgraphs for network comparison. *Bioinformatics* **26**, i611–i617 (2010).
28. Hayes, W., Sun, K. & Pržulj, N. Graphlet-based measures are suitable for biological network comparison. *Bioinformatics* **29**, 483–491 (2013).
29. Headey, D. & Fan, S. Anatomy of a crisis: the causes and consequences of surging food prices. *Agr. Econ.* **39**, 375–391 (2008).
30. Hair, J. F., Anderson, R. E., Tatham, R. L. & William, C. *Multivariate Data Analysis* (Prentice-Hall International, WC, 1998).
31. Heston, A., Summers, R. & Aten, B. PENN world table, (2002) (Date of access: 15/11/2011) URL: https://pwt.sas.upenn.edu/.
32. Fund, I. M. World economic outlook (WEO) database, (2006). (Date of access: 15/10/2012) URL: http://www.imf.org/external/pubs/ft/weo/2012/02/weodata/index.aspx.
33. Serrano, M. A. & Boguñá, M. Topology of the world trade web. *Phys. Rev. E* **68**, 015101 (2003).
34. Fagiolo, G., Reyes, J. & Schiavo, S. World-trade web: Topological properties, dynamics, and evolution. *Phys. Rev. E* **79**, 036115 (2009).
35. De Benedictis, L. & Tajoli, L. The world trade network. *World Econ.* **34**, 1417–1454 (2011).
36. Destler, I. US trade policy-making in the eighties. In *Politics and Economics in the Eighties*, 251–284 (University of Chicago Press, 1991).
37. Kindleberger, C. P. Government policies and changing shares in world trade. *Am. Econ. Rev.* **70**, 293–298 (1980).
38. Mundell, R. A. A reconsideration of the twentieth century. *Am. Econ. Rev.* **90**, 327–340 (2000).
39. Cukierman, A. & Tommasi, M. When does it take a Nixon to go to China? *Am. Econ. Rev.* **88**, 180–97 (1998).
40. Prasad, E. S. & Rajan, R. G. Modernizing China's growth paradigm. *Am. Econ. Rev.* **96**, 331–336 (2006).
41. Arellano, C. Default risk and income fluctuations in emerging economies. *Am. Econ. Rev.* **98**, 690–712 (2008).
42. Berka, M., Devereux, M. B. & Engel, C. Real exchange rate adjustment in and out of the eurozone. *Am. Econ. Rev.* **102**, 179–85 (2012).

Acknowledgments

We thank Michael Stumpf, Dimitris Achlioptas, and Des Higham for their comments and assistance with this work. Supported by the European Research Council (ERC) Starting Independent Researcher Grant 278212 and the USA National Science Foundation (NSF) Cyber-Enabled Discovery and Innovation (CDI) grant OIA-1028394; EU Creative Core FISNM-3330-13-500033, ARRS Program P1-0383 and Project J1-5454 (Z.L.); and the intramural program of the USA National Library of Medicine (A.S.).

Author contributions

Ö.N.Y. performed all the analyses except the canonical correlation analysis on world trade networks; N.M.D. was involved in experimental design; D.D. performed the canonical correlation analysis on world trade networks; V.J. collected the world trade network datasets; R.K. interpreted the results of the world trade network analysis; A.S. and N.P. designed and supervised the study, and analysed the results; Z.L. and N.P. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Yaveroğlu, Ö.N. *et al.* Revealing the Hidden Language of Complex Networks. *Sci. Rep.* **4**, 4547; DOI:10.1038/srep04547 (2014).