SCIENTIFIC REP✺RTS

Graphlet-based Characterization of Directed Networks

Anida Sarajlić[1,*], Noël Malod-Dognin[2,*], Ömer Nebil Yaveroğlu[3] & Nataša Pržulj[2]

We are flooded with large-scale, dynamic, directed, networked data. Analyses requiring exact comparisons between networks are computationally intractable, so new methodologies are sought. To analyse directed networks, we extend graphlets (small induced sub-graphs) and their degrees to directed data. Using these directed graphlets, we generalise state-of-the-art network distance measures (RGF, GDDA and GCD) to directed networks and show their superiority for comparing directed networks. Also, we extend the canonical correlation analysis framework that enables uncovering the relationships between the wiring patterns around nodes in a directed network and their expert annotations. On directed World Trade Networks (WTNs), our methodology allows uncovering the core-broker-periphery structure of the WTN, predicting the economic attributes of a country, such as its gross domestic product, from its wiring patterns in the WTN for up-to ten years in the future. It does so by enabling us to track the dynamics of a country's positioning in the WTN over years. On directed metabolic networks, our framework yields insights into preservation of enzyme function from the network wiring patterns rather than from sequence data. Overall, our methodology enables advanced analyses of directed networked data from any area of science, allowing domain-specific interpretation of a directed network's topology.

Deciphering the wiring patterns (also called topology) of large-scale, dynamic, directed, networked data coming from all domains of sciences is fundamental for understanding and predicting their functioning, emergent properties and controllability[1,2]. Topological analyses of networks are universal and provide insights in all areas of science that use network theory, including social science[3], politics[4], biology[5] and medicine[6].

In economics, yearly trade relationships between countries are captured in the world trade network (WTN), in which nodes represent countries and in which there is a directed edge from country u to country v if u exported some commodities (i.e., marketable items) to v in the considered year. In the literature, topological analyses of these directed networks focused on characterizing their organization. Garlaschelli and Loffredo[7] were the first ones to hypothesise the modular organization of WTN, due to its similarities with scale-free networks[8]. They showed that similar to scale free topology, WTN is characterized by a sharp power-law distribution of directed edges (e.g., many countries are involved in a few trades but only few countries are involved in large numbers of trades) and that trade relationships are disassortative (e.g., countries with many trade partners are connected to countries with few partners and vice-versa). This modular organization of WTN was refined by Kali and Reyes[9], who characterized the core-periphery organization[10] of trades using in-degree centrality: in the WTN, some countries are at the densely connected core (the central nodes in the WTN), forming rich-clubs of trading countries, while others (the non-central ones) are at the sparsely connected periphery. The core-periphery organization was later observed by using various statistics such as in-degree and betweenness centralities[11], random walkers and k-shell decomposition[12]. It was also shown to be affected by trade globalization and regionalization[11]. These studies also highlighted the link between the positioning of a country in the WTN and its economy. For example, Garlaschelli and Loffredo[7] proposed to predict the gross domestic product (GDP) of a country by a derivative of its number of trading partners (i.e., node degree).

In systems biology, metabolism is the set of all life-sustaining chemical reactions in a cell. These reactions, in which enzymes catalyse the transformation of substrates (input metabolites) into products (output metabolites), can be represented by metabolite-centred and by enzyme-centred metabolic networks. In metabolite-centred metabolic networks, two metabolites (nodes), u and v, are connected by a directed edge from u to v if there is an enzyme that catalyse a reaction having u as one of its substrates and v as one of its products. In enzyme-centred metabolic networks, two enzymes (nodes), u and v, are connected by a directed edge from u to v if some of the

[1]Department of Computing, Imperial College London, SW7 2AZ London, UK. [2]Department of Computer Science, University College London, WC1E 6BT London, UK. [3]Google UK, London, UK. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to N.P. (email: natasa@cs.ucl.ac.uk)

products of *u* are substrates of *v*. The topology of metabolic networks has been the focus of many studies (e.g., the review of Lacroix *et al.*[13]). In these studies, a popular network statistic is *network motifs*[14], which are defined as small partial sub-graphs that are over-represented in a network with respect to a given null model (partial sub-graph means that when selecting a sub-set of nodes from the large network, we can select any edges between them).

Similar to WTN, many studies focused on characterizing the organization of metabolic networks. Zhu and Qin[15] observed that metabolic networks have scale free topology with highly modular organization, coming from their power law distributions of in- and out-degrees. Recently, Shellman *et al.*[16] showed that this modular organization is related to the cellular organization: while the distributions of 3-node motifs in metabolic networks are similar across species, they are unique across organelles (i.e., different cellular components). Furthermore, the topological similarities between the metabolic networks of different species are shown to approximately reconstruct phylogenetic classification of species[17,18], which suggests a link between the divergence times between species and the dissimilarities between their metabolic networks. However, this observation must be toned down by the fact that metabolic networks are reconstructed from homology[19,20]. Recently, Pearcy *et al.*[21] related the similarities between the metabolic networks of 383 bacterial species, as measured by the similarity of their motif spectra, with their phenotypic variability (e.g., aquatic or terrestrial species, aerobic or anaerobic environment), which suggests that adaptation to environment during evolution may have shaped the topology of metabolic networks. Finally, motifs and their spectra were used to relate the positioning of enzymes and metabolites in metabolic networks and their biological functions. Shellman *et al.*[16] showed that some motifs are characteristic to specific metabolic functions of enzymes, and going further, Ganter *et al.*[22] proposed a hidden markov model[23] based framework for predicting metabolic functions from motif spectra.

Because exact comparison between complex networks has long been known to be computationally intractable[24], the topological analyses of complex networks use simple heuristics, commonly called network *statistics*, such as the in- and out-degree distributions, to approximately say whether the structures of networks are similar[25]. Recently, the concept of *node roles*, which associate nodes with well defined topological features, has been proposed for analysing networked data. For example, in control theory, 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]. In the same vein, Yan *et al.*[26] classify nodes in network as indispensable, neutral or dispensable, if their removal from the networks results in increasing, similar, or decreasing number of driver nodes in the resulting networks, respectively. In directed protein-protein interaction networks, indispensable proteins are shown to be the primary targets of disease-causing mutations, and this observation was used to predict novel cancer-driver genes. However, the limited number of controllability-based roles limits their interpretability. As seen above, nodes in metabolic networks can be characterized by their motif spectra[14]. These motif-based roles are used to relate the positioning of enzymes in metabolic networks with their biological functions[16,22]. However, motifs and their spectra have limited usability[27], as they are dependant on the choice of a null model, which is generally unknown for real-world data.

For analysing undirected networks, Yaveroğlu *et al.*[28] proposed a framework in which node roles are modelled by graphlets[29]. Graphlets are defined as small induced subgraphs of a large network that appear at any frequency; an induced sub-graph means that once you pick the nodes in the large network, you must pick all the edges between them to form the sub-graph. Within graphlets, *symmetry groups* of nodes called *automorphism orbits* are used to characterize different topological positions that a node can participate in. Orbits are used to generalize the notion of node degree: the *graphlet degrees* of a node are the numbers of times a node is found at each orbit position[30]. Graphlets with up-to five nodes and their degrees characterize 73 different node roles, and their interpretability is independent of a null model. In the node role framework, the dependencies between node roles (i.e., the correlations between the counts of graphlet orbits over all nodes in a network) are encoded in a symmetric matrix called the graphlet correlation matrix (GCM), which is shown to finely describe the topology of undirected networks[28]. GCMs have also been compared within the network statistics called graphlet correlation distance (GCD), which has been shown to be the most accurate network statistic for classifying undirected networks. Finally, canonical correlation analysis (CCA)[31] allows translating node roles in network into domain specific languages. In economics, CCA is used to relate the graphlet-based roles of countries in the undirected world trade network with their economic attributes[28]. In systems biology, this framework is used to uncover biological functions that are performed through similar patterns of protein-interactions across yeast and human (topologically orthologous biological functions)[32].

However, graphlets, and by extension, the whole node role framework, are only defined for undirected networks. Some real-world data are inherently directed and cannot be represented by undirected networks without information loss. For example, a trade between two countries is directed, from the exporting country towards the importing country, and these two situations are not equivalent; the trade represents an income for the exporting country, while it represents an expense for the importing country. When world trades are modelled as undirected networks, directionality is lost and the two situations are treated equally.

To better analyse directed data and to avoid information loss from modelling directed data with undirected networks, we extend the node role framework to directed networks. First, we define directed graphlets, and extend existing graphlet-based statistics to directed networks. Among them, we show that the directed graphlet correlation distance is superior for clustering directed networks. Second, we use our directed graphlets to extend the node role framework to directed networks, and demonstrate its superior descriptive and predictive power on real-world data in two domains, economics and biology.

New Methodology: Node Role Descriptors for Directed Networks

To adapt the node role framework to directed networks, we first generalize graphlets to directed ones as follows. We define directed graphlets as small induced sub-graphs of a larger directed network, without anti-parallel directed edges (i.e., if a graphlets contains a directed edge from node *u* to node *v*, it cannot contain the opposite
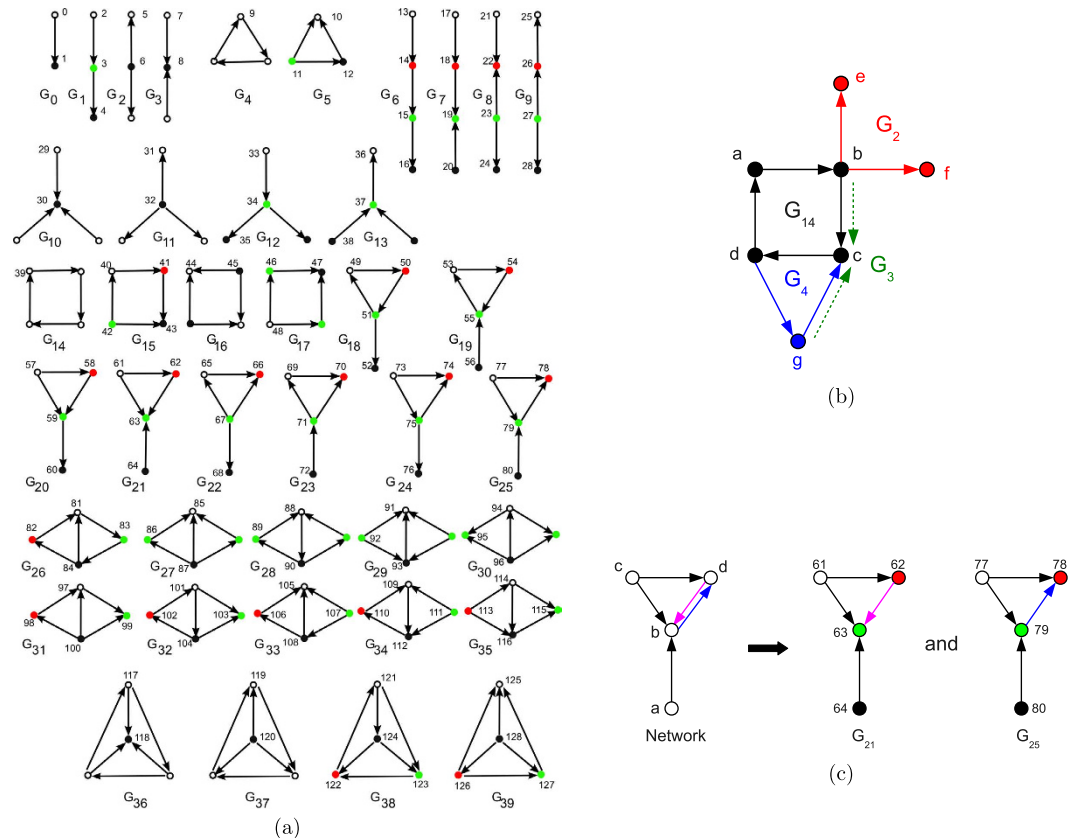
Figure 1. Illustration of directed graphlets. (**a**) The 40 two- to four-node directed graphlets $G_0, \ldots, G_{39}$. In each graphlet, node belonging to the same automorphism orbit are of the same colour. The 128 automorphism orbits are labelled from 0 to 127. (**b**) Illustration of how directed graphlets assemble together to form complex networks. The whole network can be created in three steps. First, we start with one graphlet G_{14} (nodes a,b,c and d, in black). Then, we add a graphlet G_2 (in red) by adding two new nodes, e and f, as heads of directed edges from node b. Finally, we add a graphlet G_4 (in blue) by adding a new node, g, as the head of a directed edge from node d and as the tail of a directed edge towards node c. Note that during this process, many new graphlets are created, e.g., graphlet G_3 (in green) between nodes b, c and g. (**c**) Illustration of our anti-parallel directed edge counting strategy. The two anti-parallel directed edges in the input network (in blue and in magenta) account for one graphlet G_{21} (when considering the magenta directed edge) and one graphlet G_{25} (when considering the blue directed edge), among all other induced graphlets.

directed edge from node v to node u). Within directed graphlets, because of symmetries (automorphisms), some nodes have identical wiring patterns and hence belong to the same automorphism orbit (*orbit* for brevity; formally defined in the Supplement). The 40 two-to-four node directed graphlets are presented in Fig. 1 (panel a) and are denoted from G_0 to G_{39}; their 129 orbits are denoted from 0 to 128. Real-world directed networks may contain anti-parallel directed edges, which we take into account in our counting strategy by inducing directed graphlets for each anti-parallel directed edge separately (Fig. 1, panel c). Analogous to the graphlet degree vector[33], a node in a directed network is described by its *Directed Graphlet Degree Vector* (DGDV), which is a 129-dimensional vector encoding the two- to four-node graphlet degrees of the node in the networks; e.g., the ith coordinate of the DGDV of node n, denoted by $DGDV_n[i]$ is the number of times a directed graphlet touches node n at orbit i. In our implementation, directed graphlets and their degrees are counted using a complete enumeration algorithm; for a given node, counting all directed graphlet degree for up-to k node directed graphlets requires visiting the $k-1$ neighbourhood of the node, which is done in $O(d^{k-1})$ time where d is the maximum degree of a node. Thus, for counting all directed graphlet degrees in a network with n nodes, the time complexity is $O(nd^{k-1})$, which is the same as the complexity of counting undirected graphlet degrees[28]. Memory wise, we need to store the network $(O(n^2))$ and the counts of the z graphlet degrees for each node ($z = 129$ when considering all 2- to 4-node direct graphlets), so the space complexity is $O(n^2 + nz)$.

Directed graphlet-based statistics. Directed graphlets are like Lego pieces that assemble with each other to build large networks; any directed network can be constructed by combining different directed graphlets at different directed-orbits (an example is given in Fig. 1, panel b). We exploit this observation to summarize the complex structures of networks and to compare them, by generalizing three graphlet-based network statistics to directed networks: the Relative Graphlet Frequency Distribution Distance[29], the Graphlet Degree Distribution Agreement[30] and the Graphlet Correlation Distance[28].

Directed Relative Graphlet Frequency Distribution Distance (DRGF) compares two networks according to their relative distributions of directed graphlet frequencies. Let $N_i(G)$ be the number of directed graphlets of type i in network G. We use the total number of 3- to 4-node directed graphlets in G, $T(G) = \sum_{i=1}^{39} N_i(G)$, to normalize $N_i(G)$: $\overline{N_i}(G) = \frac{\log N_i(G)}{\log T(G)}$ if $N_i(G) \neq 0$ and zero otherwise. Given two networks, G and H, the DRGF distance between these networks is defined as: $DRGF(G, H) = \sum_{i=1}^{39} |\overline{N_i}(G) - \overline{N_i}(H)|$.

Directed Graphlet Degree Distribution Agreement (DGDDA) compares two networks according to their distributions of directed graphlet degrees. Let d_G^j be the distribution of the jth automorphism orbit in network G, where $d_G^j(k)$ is the number of nodes in G that touch orbit j exactly k times. We scale d_G^j as $S_G^j(k) = \frac{d_G^j(k)}{k}$ to decrease the contribution of larger degrees, and then normalise it according to its total area $T_G^j = \sum_{k=1}^{\infty} S_G^j(k)$. This result in *normalised j^{th} distribution* for the network G, $N_G^j(k) = \frac{S_G^j(k)}{T_G^j}$. The agreement between normalised j^{th} distributions for the two networks G and H is: $D^j(G, H) = 1 - \frac{1}{\sqrt{2}} \cdot (\sum_{k=1}^{\infty} [N_G^j(k) - N_H^j(k)]^2)^{\frac{1}{2}}$, where the resulting value is between 0 and 1, 1 meaning that the jth directed graphlet degree distributions are identical, 0 that they are dissimilar. Finally, the directed graphlet degree distribution agreement between two networks is defined as the arithmetic mean of all agreements over the 129 automorphism orbits: $DGDDA(G, H) = \frac{\sum_{i=0}^{129} D^i(G, H)}{129}$.

Analogous to the case of undirected graphlets, the statistics of different orbits on directed graphlets are not independent of each other. The reason behind this is the fact that smaller graphlets are induced sub-graphs of larger graphlets. We report 23 such non-redundant dependencies in Supplementary Table 1. Within a network, we capture the dependencies between directed orbits within its *Directed Graphlet Correlation Matrix* (DGCM), which we construct as follow. We construct a matrix whose rows are the directed graphlet degree vectors over all nodes of the network. We calculate the Spearman's correlation[34] between each two pairs of columns in the resulting matrix, i.e., correlations between the orbits in the network. We present these correlations in a 129×129 dimensional Directed Graphlet Correlation Matrix (DGCM-129), which is symmetric and contains Spearman's correlation values in $[-1, 1]$ range. In addition to in-depth examination of network topology that can be qualitatively interpreted (see the Results section), we exploit the differences in DGCMs to compare directed networks. The *Directed Graphlet Correlation Distance* (DGCD) measures the distance between networks as the Euclidean distance between their DGCMs. Its superiority over other directed network distances is demonstrated in section "Evaluation of the methodology". We consider two versions of DGCD: DGCD-129, which uses the 129 two- to four-node directed graphlet orbits, and DGCD-13, which uses only the 13 two- to three-node directed graphlet orbits.

Directed graphlets can also be used to measure the similarity between the positioning of nodes in directed networks; the *Directed Graphlet Degree Vector Similarity* (DGDVS) measures the similarity between the positioning of two nodes by the similarity of their directed graphlet degree vectors. The similarity between the ith directed graphlet degrees of nodes u and v is computed as: $D_i(u, v) = w_i \times \frac{|log(DGDV_u[i] + 1) - log(DGDV_v[i] + 1)|}{log(max(DGDV_u[i], DGDV_v[i]) + 2)}$, where w_i is an orbit-specific weight that reduces the influence of dependent orbits (see Supplementary Material for the definition of w_i). DGDVS is calculated as: $DGDVS(u, v) = 1 - \frac{\sum_{i=0}^{128} D_i(u, v)}{\sum_{i=0}^{128} w_i}$.

Directed node role framework.

To bridge the gap between node roles (the wiring patterns around nodes that are captured by their directed graphlet degree) in a network and their (real-valued) domain specific annotations, we adapt the canonical correlation analysis (CCA) methodologies from Yaveroğlu *et al.*[28] and from Davis *et al.*[32] to uncover the linear relationships between them.

In our CCA framework, the node roles are encoded in a $n \times z$ node role matrix, R, and the domain specific annotations in a $n \times f$ attribute matrix, A. For both matrices, each row (called a *variable vector*) represents one of the n nodes of the network (for both matrices, the same row index represents the node). In the node role matrix, values in a variable vector (a row) are the $z = 129$ directed graphlet degrees of the node in the network, while in the attribute matrix, they correspond to the f annotations of the node.

Given n pairs of variable vectors from $R \times A$ for n nodes as an input, the CCA outputs two weight vectors, called a *canonical variate*, so that the weighted sum of R is maximally correlated with the weighted sum of A. The correlation between the two weighted sums is called their *canonical correlation*. After finding the first set of weights, CCA iterates $min \{z, f\}$ times to find more weight vectors, such that the resulting canonical variates are not correlated with any of the previous canonical variates. The weight matrices W_R and W_A, for R and A, respectively, are constructed by combining all of the identified weight vectors. We refer the interested reader to Weenink[35] for the mathematical aspects of CCA.

For a given canonical variate, the contribution of a variable (e.g., of a given directed graphlet degree) is measured by its *loading*, which is the Pearson's correlation[36] between the values of the variable and of the linear combination of variables that it participates in. For each canonical variate, the variables having the largest (positive or negative) loadings are the most related to each others, which allows uncovering the strongest relationships between node roles and annotations.

Finally, *the association matrix*, which is constructed as $W_R \times S \times W_A^+$, where S is the diagonal matrix of canonical correlations that weights the variates according to their correlation strength and where W_A^+ is the Moore-Penrose pseudoinverse[37] of W_A, is a multidimensional regression that allows predicting the domain specific annotations of a node from its directed graphlet degrees in the network.

Evaluation of the Methodology

We compare the performance of our novel directed graphlet-based network distances (DGCD-129, DGCD-13, DRGF and DGDDA) used for clustering networks coming from six directed network models (ER, SFBA-Sink, SFBA-Source, GEO, GEO-GD and SF-GD), which we define as follows.

- *Directed Erdös-Rènyi random model, **ER**,* represents uniformly distributed random interactions[38]. A directed ER network is generated by fixing the number of nodes in the network, and then by randomly adding directed edges between uniformly chosen pairs of nodes.
- *Scale free Barabàsi-Albert random model,* also called preferential attachment, generate networks based on the "rich-gets-richer" principle[8]. We use the directed preferential attachment model[39] to generate two distinct scale free topologies. In the first one, **SFBA-Sink**, we favour adding directed edges towards nodes having already large numbers of incoming edges, creating densely connected sink nodes. In the second one, **SFBA-Source**, we favour adding directed edges outgoing from nodes having already large numbers of outgoing edges, creating densely connected source nodes.
- *Directed scale-free random model with gene duplication and divergence, **SF-GD**,* is a scale-free model which mimics the gene duplication and the gene divergence processes from biology[40]. Starting from a small seed network, a node in the network is selected at random and a new node is created together with the connections to/from nodes that the "parent" node has (duplication step). A directed edge, whose directionality is randomly chosen, is added between the new node and his parent with probability *p*. Then, each directed edge that the new node "inherited" from its parent node is deleted with probability *q* (divergence step). This procedure is repeated until the desired number of nodes is achieved. Probabilities *p* and *q* are tuned to achieve the desired number of directed edges.
- *Directed geometric random model, **GEO**,* represents the proximity relationship between uniformly randomly distributed points in a *d*-dimensional metric space[41]. We generate GEO networks by putting points in 3-dimensional Euclidean unit cube uniformly at random that correspond to nodes of the networks and two nodes are connected by a directed edge, whose directionality is randomly chosen, if the Euclidean distance between the corresponding points in space is smaller than a specified distance threshold, *r*, which is chosen to achieve the desired number of directed edges.
- *Directed geometric random model with gene duplication, **GEO-GD**,* is a geometric model that mimics the gene duplication and divergence processes in biology[42]. Starting from a small number of nodes randomly distributed in a metric space, a new node is introduced as a copy of a randomly selected node in the network (duplication step); the new node is moved randomly from its *parent* node in the metric space (divergence step), at a random distance smaller or equal to $2 \times r$ (where *r* is the distance threshold used to generate GEO networks). This process is repeated until the desired number of nodes is achieved, and then directed edges between nodes are created following the rules of the directed geometric network model.

We extend the methodology of Yaveroğlu *et al.*[43]. We generate 10 networks for each model and for each of the following three numbers of nodes and two directed edge densities that mimic the sizes and densities of real-world networks: 500, 1000 and 2000 nodes, and 0.5% and 1% directed edge densities. Hence, the total number of synthetic networks that we consider is $10 \times 6 \times 3 \times 2 = 360$. We formally evaluate the performance of our directed graphlet-based network distances for grouping networks from these models together and compare them to the performances of other commonly used directed network distance measures: the in- and out-degree distribution distances (note that we only present results of in-degree distribution distance, as it produces almost identical results as its out-degree counterpart) and directed spectral distance[44]. We assess how well a distance measure groups networks of the same type by using the standard Precision-Recall and ROC curves[45]: for small increments of parameter $\varepsilon \geq 0$, if the distance between two networks is smaller than ε, then the pair of networks is declared to be similar (belong to the same cluster). For each ε, four values are computed: the *true positives* (TP), which are the numbers of correctly clustered pairs (i.e., that group together networks from the same model), the *true negatives* (TN), which are the numbers of correctly non-clustered pairs (i.e., that do not group together networks from different models), the *false positives* (FP), which are the numbers of incorrectly clustered pairs (i.e., that group together networks networks from different models), the *false negatives* (FN), which are the numbers of incorrectly non-clustered pairs (i.e., that do not group together networks from the same model). In *Precision-Recall curves*, for each ε, the precision ($PREC = \frac{TP}{TP+FP}$) is plotted against the recall, also-called true-positive rate, ($REC = \frac{TP}{TP+FN}$) and the quality of the grouping by a given distance measure is measured with the area under the Precision-Recall curve (AUPR), which is the average precision of the distance measure. In *ROC curves*, for each ε, the true-positive rate is plotted against the false-positive rate ($FPR = \frac{FP}{FP+TN}$) and the quality of the grouping by a given distance measure is measured with the area under the ROC curve (AUC), which is the probability that a randomly chosen pair of networks coming from the same model will have a distance smaller than a randomly chosen pair of networks coming from different models.

AUPRs and AUCs show that DGCD-13 is the most accurate among all tested measures (Fig. 2, panel a), which is also illustrated by its superior Precision-Recall curve (Fig. 2, panel b). The superiority of DGCD-13 over DGCD-129 may come from the large number of dependencies between three-nodes and four-nodes directed graphlet degrees (listed in the Supplementary Material), which may blur the (dis)similarities between the DGCMs used by DGCD-129. 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 distances and thus, that are retrieved first by the distance measure[46]. Both DGCD-13 and DGCD-129 show superiority in early retrieval over all measures (beginnings of the curves in Fig. 2, panel b). Furthermore, we confirm that such results could not have been achieved by undirected network measures by comparing the performances of

Classification quality

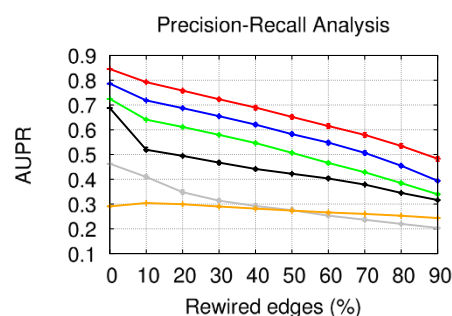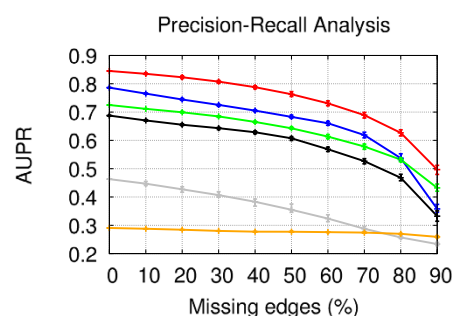| Directed Distances		AUPR	AUC	Undirected Distances		AUPR	AUC
	DGCD-13	**0.845**	**0.944**		GCD-15	**0.545**	**0.850**
	DGCD-129	0.786	0.917		GCD-73	0.518	0.820
	DRGF	0.725	0.899		RGF	0.503	0.832
	DGDDA	0.688	0.819		GDDA	0.503	0.795
	In-Degree Distribution	0.463	0.670		Degree Distribution	0.417	0.656
	Directed Spectral	0.291	0.533		Spectral	0.300	0.584

(a)
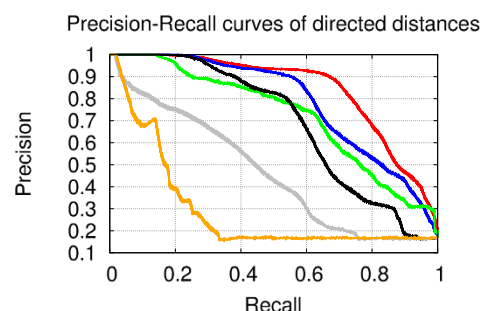
(b)
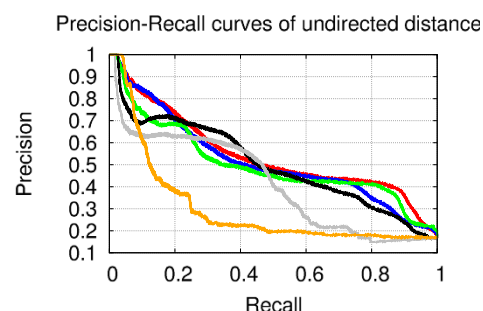
(c)

(d)

(e)

Figure 2. Quality of clustering of the 360 model networks using directed or undirected distance measures (colour coded in panel a). (a) For directed and undirected distance measures, the area under the ROC curves (AUC) and the area under the precision-recall curve (AUPR) achieved by a distance measure. (**b**) Precision-Recall curves achieved by the six directed distance measures. (**c**) Precision-Recall curves achieved by the six undirected distance measures. (**d**) For directed distance measures, AUPR for different percentages of noise (randomly rewired directed edges in 10% increments, horizontal axis) in model networks. (**e**) For directed distance measures, AUPR for different percentages of incompleteness (randomly removed directed edges in 10% increments, horizontal axis) in model networks.

directed distance measures to the ones of their undirected counterparts (when applied on the same networks but that are made undirected by removing the directionality of the edges). As expected, undirected network distances perform poorly, as evidenced by smaller AUCs and AUPRs (Fig. 2, panel a) and Precision-Recall curves (Fig. 2, panel c). Using directed graphlets allows for improving the average precision (AUPR) of classifying random networks by ≈55 percent, from AUPR = 0.545 when using the best undirected distance measure (GCD-15), to AUPR = 0.845 when using the best directed distance measure (DGCD-13).

Since real networks are incomplete and noisy[47,48], we evaluate the clustering quality of the above distance measures in the presence of noise. To simulate noise, we randomly rewire up to 90% of the directed edges of model networks, in increments of 10%. To account for the variability of randomizations, for each percentage of noise we repeat the randomisation process 30 times and report minimum, average and maximum AUPR for clustering networks from the same model. We follow the same procedure to simulate missing data by randomly removing up to 90% of the directed edges from model networks, in increments of 10%. These tests demonstrate robustness to noise and missing data and the superiority of DGCD over other measures on this sample of different network topologies, sizes and densities (Fig. 2, panel d,e).

We further validate our methodology by assessing its ability to correctly group real-world metabolic networks. We reconstruct the directed metabolic networks of all Eukaryotes (299 species) using the metabolic reaction data from KEGG database[49] (collected in December 2014). For each species, we model its metabolic reactions as an

enzyme-enzyme network in which two enzyme-coding genes (nodes) are connected by a directed edge if the first enzyme catalyses a reaction whose product is a substrate for a reaction catalysed by the second enzyme. We obtain the taxonomic classification of species from KEGG database, which classifies Eukaryotic species according to (from the most generic to the most specific) their Kingdom, Sub-phylum and Class. We use the six directed distance measures to compare our 299 directed metabolic networks and use Precision-Recall and ROC curve analyses to measure their agreement with their taxonomic classification. Since the metabolic reactions encoded in metabolic networks are mostly reconstructed from sequence homology relationships, we expect that the similarity between the topologies of our directed metabolic networks will relate to the evolutionary relationships between the corresponding species, validating our methodology. Indeed, as detailed in Supplementary Fig. 3, DGCD-129 and DGCD-13 achieve the highest agreements with the taxonomic classification of species. For example, DGCD-13 of directed metabolic networks agree with their Class classification with AUC of 0.945, validating the performance of the methodology.

Application to Real-World Networks

Economics: world trade network.   Using trade data from the United Nations Commodity Trade Statistics (UN Comtrade) database (collected in August 2014, from http://comtrade.un.org/), we generate 52 WTNs, one for each year between 1962 and 2013. To filter-out insignificant trades, we only consider the largest trades accounting for 90% of the total trade amount of a given year. While filterings based on the statistical significance of the trades[50,51] could be used within our framework, we do not use it because such methods rely on a null model, which is unknown for WTN. In the context of subgraph based analysis, such as ours, using a poorly-fitting null model has been argued to lead to invalid conclusions[27]. We obtain the economic indicators of country wealth from PENN World Table (PENN, version 7.1, downloaded in July 2014 from https://pwt.sas.upenn.edu/) and International Monetary Fund (IMF) World Economic Outlook Database (WEO, downloaded in July 2014 from http://www.imf.org/). For simplicity, we focus on the following nine economic indicators of countries: population size, employed population size, gross domestic product (GDP), GDP per capita (i.e., GDP of a country divided by its population size), debt, debt per capita (i.e., debt of a country divided by its population size), import expenses, export incomes and capital stock. Because of the limited availability of economic indicators, we do our analyses linking a country's wiring patterns in the WTN to its economic indicators on the subset of 32 WTNs from 1980 to 2011.

Uncovering the directed core-broker-periphery structure of WTN.   Undirected WTN is believed to possess a core-periphery organization[10], which is a binary classification of nodes according to their global positioning in the network: core nodes interact with each others and also with some peripheral nodes, while peripheral nodes tend to only interact with core nodes. This binary model was later refined into the core-broker-periphery model for the undirected WTN[28], which describes the positioning of nodes locally, as measured by their graphlet degrees. By considering different local neighbourhoods, nodes can simultaneously be involved in densely connected trade patterns (core nodes), in sparsely connected patterns (peripheral nodes), or in particular positions that enable them to mediate the trade between core and peripheral regions of the WTN (broker nodes). Rather than simply classifying nodes as either core, broker, or peripheral, using graphlet degrees allows quantifying the involvement of a node in all of these roles. We use the descriptive power of the directed graphlet correlation matrices (DGCMs) to assess whether the directed WTN also has a core-broker-periphery structure.

Each of graphlets G_{18} to G_{25} (Fig. 1, panel a) contains all three positions; the core is represented by the triangle's nodes (e.g., orbits 49, 50 and 51 on graphlet G_{18}), the periphery is represented by the node hanging from the triangle (e.g., orbit 52 on graphlet G_{18}), and the broker by the node from the core that is connected to the peripheral node (e.g., orbit 51 on graphlet G_{18}). We use these eight graphlets to assess if our directed WTN also possess the core-broker-periphery organization. To this aim, we focus on the DGCMs that are obtained when considering three sets of orbits: the eight peripheral orbits (orbits 52, 56, 60, 64, 68, 72, 76 and 80); the eight broker orbits (orbits 51, 55, 59, 63, 67, 71, 75 and 79); and the 16 core orbits that are not brokers (orbits 49, 50, 53, 54, 57, 58, 61, 62, 65, 66, 69, 70, 73, 74, 77 and 78). These orbits are illustrated in Fig. 3, panel a.

In our WTNs of each year, as presented in Fig. 3 (panels b, c and d), the broker orbits strongly correlate with the core orbits, showing that core trading countries are very likely to be involved in trade brokerage. On the opposite, the peripheral orbits are not correlated with core orbits and slightly anti-correlated with brokerage orbits, showing that in the WTN there is a clear distinction between core/broker and peripheral countries. Thus, similar to the undirected case, our directed WTNs are characterized by a core-broker-periphery structure. Moreover, we observe that peripheral countries form two subsets: the peripheral countries that import from the broker countries (orbits 52, 60, 68 and 76) and peripheral countries that export to the broker countries (orbits 56, 64, 72 and 80). The strong correlation within these two sets and the low correlation between them show that the peripheral countries tend to specialize as either peripheral importer or as peripheral exporter, but not both. This observation is only made possible by our new directed graphlet-based analysis and has eluded us thus far in undirected and directed studies. Over time, the correlations between the peripheral roles and the core/broker roles slightly increases, showing that peripheral countries are getting more integrated into the WTNs, which is likely to be an effect of globalization. However, while the correlation within the two sets of peripheral orbits get stronger over time, the correlations between them get weaker, which shows that peripheral countries became more specialised towards either import or export. Finally, the directed core-broker-periphery structure of WTNs does not appear in random networks (e.g., see the DGCM of a directed Erdös-Rènyi random network in Fig. 3 panel e).

Predictive power of directed wiring patterns.   We assess the ability of predicting the economic indicators of countries from their wiring patterns in the directed WTN for up-to ten years into the future using our canonical correlation analysis (CCA) framework. To predict the economic attributes y years into the future, we construct the node role matrix, R, and the economic attribute matrix, A, as follows. In R, a row corresponds to a node in the
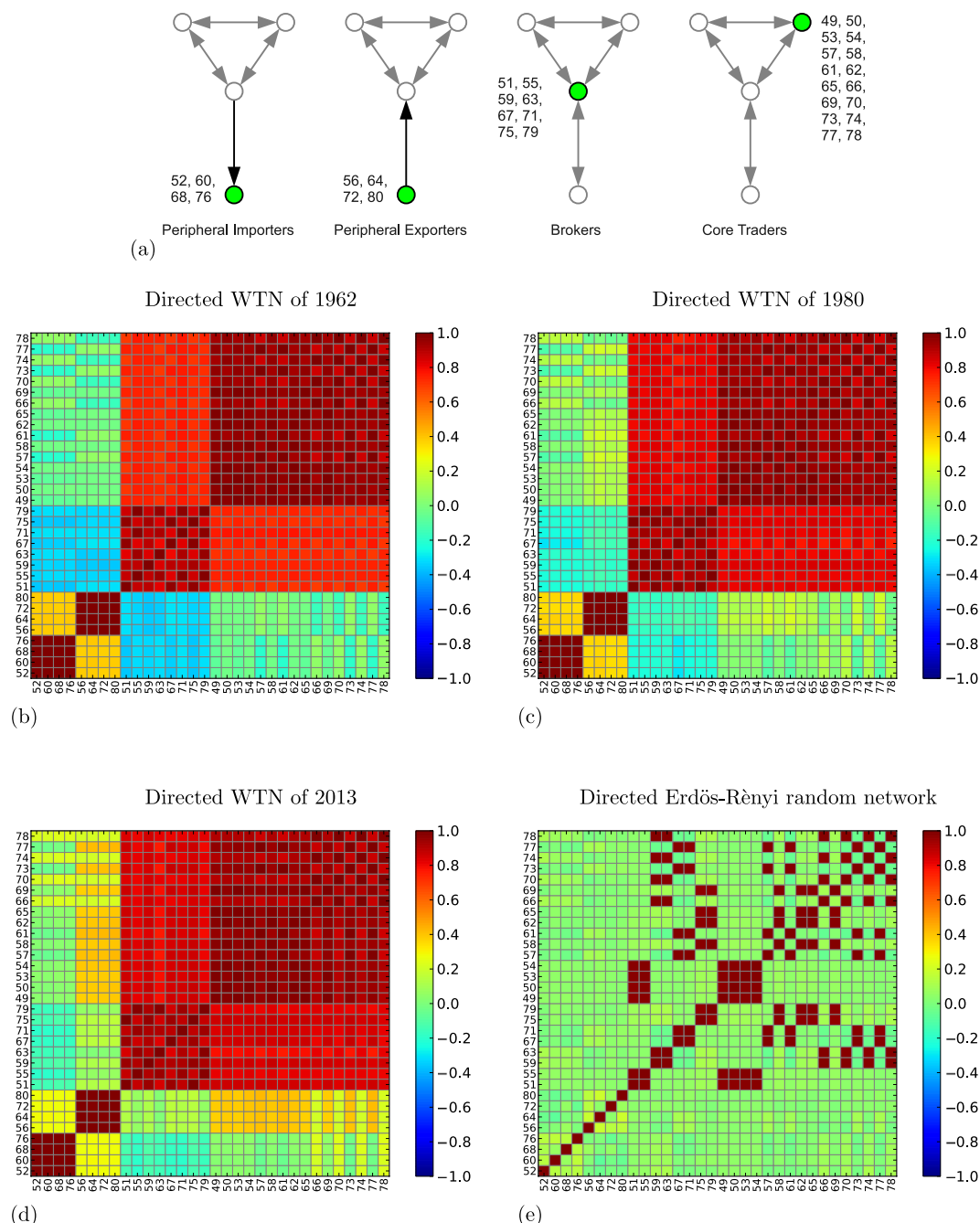
Figure 3. The directed Core-Broker-Periphery organization of WTN, with specialized peripheral importers and exporters. (**a**) The peripheral importer, peripheral exporter, broker and core orbits that we use to build the presented DGCMs. In the graphlets, bi-directional edges (in grey) mean that any directionality is allowed. The next panels present the corresponding DGCMs for WTN of 1962 (panel b), WTN of 1980 (panel c), WTN of 2013 (panel d) and for a directed Erdös-Rènyi random network (panel e).

WTN and it contains the 129 directed graphlet degrees of the country in the WTN of a given year, while in *A* the same row contains the nine economic indicators of that country *y* years into the future (e.g., when predicting 5 years in the future, a row in *R* contains the wiring patterns of, say, USA in the WTN of 2000, while the same row in *A* contains the economic indicators of USA in 2005).

We use a 10-fold cross-validation approach by randomly dividing the dataset into a learning set (90 percent of the data) and a validation set (the remaining 10 percent of the data); we obtain an association matrix transforming the directed graphlet degrees of a country into predicted economic indicators by applying our CCA framework on the learning set, and then we use the validation set to assess the predictive power of the association matrix, which we measure by the Pearson's correlations between predicted and observed economic indicator values. This random process is repeated ten times and we report the average of the obtained Pearson's correlations. As presented
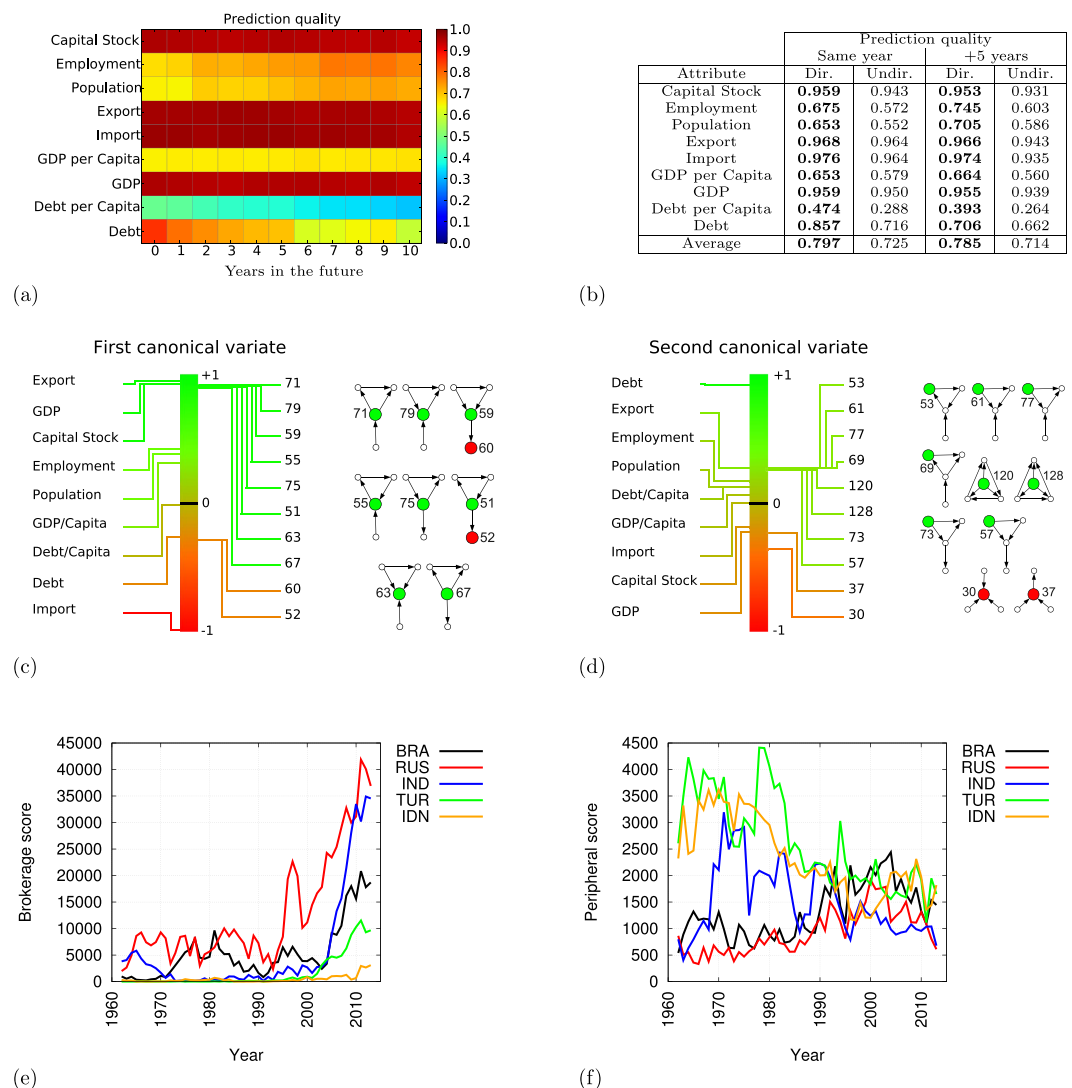
	Prediction quality			
	Same year		+5 years	
Attribute	Dir.	Undir.	Dir.	Undir.
Capital Stock	**0.959**	0.943	**0.953**	0.931
Employment	**0.675**	0.572	**0.745**	0.603
Population	**0.653**	0.552	**0.705**	0.586
Export	**0.968**	0.964	**0.966**	0.943
Import	**0.976**	0.964	**0.974**	0.935
GDP per Capita	**0.653**	0.579	**0.664**	0.560
GDP	**0.959**	0.950	**0.955**	0.939
Debt per Capita	**0.474**	0.288	**0.393**	0.264
Debt	**0.857**	0.716	**0.706**	0.662
Average	**0.797**	0.725	**0.785**	0.714

(a)   (b)

(c)   (d)

(e)   (f)

Figure 4. Analysis of directed world trade. (a) Predictive power of directed graphlets. Economic attributes are predicted for up-to ten years in the future from the wiring patterns of countries in the WTN, using the predictors that are learnt from canonical correlation analysis; we used 10-fold cross validation and the quality of the predictions is measured by the average of the Pearson's correlations between the predicted and the real economic attributes of the countries over ten random experiments. **(b)** Comparison of the predictive power of directed and undirected graphlets. The predictive powers are computed as in panel a. **(c,d)** The first and second canonical variates between economic attributes (on the left) and direct-graphlet orbits (on the right), computed using all countries and all years from 1980 to 2011 simultaneously; the middle bar is colour-coded value of correlation (loading), from −1 (in red) to +1 (in green). **(e,f)** Brokerage and peripheral scores of Brazil (BRA), Russia (RUS), India (IND), Turkey (TUR) and Indonesia (IDN).

in Fig. 4 (panel a), capital stock, import expenses, export incomes, and GDP are all consistently well-predicted from trade patterns up to ten years in the future. In contrast, debt and population-based economic indicators (population size, employment, GDP per capita, debt and debt per capita) are not well predicted, showing that these indicators do not depend only on trade patterns, but also on other geopolitical factors. Surprisingly, debt is best predicted from the wiring patterns of the same year, and then the prediction quality decreases as the prediction horizon increases, which suggests a short term effect of the trade pattern of a country on its debt level. In contrast, the prediction qualities of population and employment increase with the prediction horizon, which suggest a long term effect of the trade patterns on these indicators.

In the second step, we compare the quality of the obtained predictions to the ones that are obtained when WTNs are made undirected (by removing directionality of the edges) and when node roles are captured by undirected graphlet degrees. As presented in Fig. 4 (panel b), using directed node roles improves prediction quality of all economic indicators, justifying the introduction of directed node roles methodology. On average, when predicting an economic indicator of a country from its wiring pattern in the WTN, using directed graphlets instead of undirected ones improves prediction quality by 9.9 percent.

Among the many economic descriptors, GDP is of foremost importance, as it directly characterizes economic crises (e.g., the definition of global downturns[52]). Using all the available data (i.e., without the cross-validation approach), orbit 71 correlates the most strongly with GDP, having Pearson's correlation of 0.88. Using all directed orbits allows reaching Pearson's correlation of 0.97! These results, providing a strong evidence of the link between the GDP and a country's wiring in the world trade network, cannot be obtained by simpler, non-graphlet-based, measures of node wiring such as in- and out-degrees, or in- and out-edge closeness centralities that were previously used for analysing directed world trade networks[11]. In- and out-degrees correlate with GDP with Pearson's correlations of 0.71 and 0.72 respectively, and in- and out-edge closeness centralities with Pearson's correlations of 0.39 and 0.40 respectively.

These demonstrate that our directed graphlet-based framework finds more refined topological features than previously used methods[7,11,28], resulting in the best predictions of the economic attributes, such as GDP of a country, from its trade wiring patterns in the WTN. Thus, our framework enables prediction of how the changes in the trade policies of a country may affect its economy. However, despite the strong correlation between our predicted GDPs and the observed GDPs of the countries, our predictions are still not accurate enough for predicting the exact GDP values (e.g., see Supplementary Fig. 2). To further improve the quality of our predictions, additional non-trade related knowledge may be needed (e.g., external debt relations, bank exchanges or geopolitical situations). Such knowledge can be incorporated directly into our CCA framework by extending the node role matrix, R, with additional features, or by using data integration techniques, which are used in systems biology for fusing molecular data produced by various omics studies[53].

Economic interpretations of directed wiring patterns. We use the canonical variates (i.e., the linear combinations of graphlet degrees and their corresponding linear combinations of economic indicators) to bridge the gap between the trade patterns of the countries in the WTN and their economic attributes.

The first canonical variate, presented in Fig. 4 (panel c), is statistically significant, with canonical correlation of 0.987 and p-value ≈ 0. It highlights eight positions in the WTN enabling a country to be a middle-man in trade between non-trading countries (a broker position): orbits 55, 63, 71 and 79 correspond to countries mediating the trade between peripheral countries exporting to trade-linked countries, and orbits 51, 59, 67 and 75 correspond to countries mediating the trade between peripheral countries importing from trade-linked countries. In all these cases, the countries that are frequently seen in these broker positions in WTN have high GDPs and low debts by making profits from the transactions or added values by importing cheap raw materials and exporting expensive finished products, as highlighted by high export incomes and low import expenses. Also, there is a weak anti-correlation for the countries that are in peripheral positions (orbits 72, 80, 64, 56, 68, 76, 60 and 52; we only display orbits 52 and 60 in Fig. 4 because of space limitations), indicating that these peripheral countries tend to have low GDP, low export incomes, large debt and large import expenses.

The second canonical variate, presented in Fig. 4 (panel d), which is also statistically significant with canonical correlation of 0.95 and p-value ≈ 0, sheds light onto an unexpected aspect of the core-broker-periphery organisation of the WTN: countries that are in the core of the world trade (with large counts of orbits 53, 61, 77 and 69), but that do not intensively trade with peripheral countries (small counts orbits 30 and 37) tend to have debt. This observation has completely eluded us in previous directed and undirected studies of WTN, in which being a core trading country has always been considered as a sign of economic wealth. Our analysis suggests that in order to improve its economic wealth, a country should not only try to increase its trade relationships with the core-trading countries, but also should maximize its brokerage by trading with peripheral countries.

Tracking world trade dynamics. We demonstrated that the economy of a country is strongly related to its positioning in the WTN, highlighting favourable (broker) and dis-favourable (peripheral and core non-broker) trade positions. To quantify the strength of the brokerage position of a country in the WTN of each year, we extend the approach of Yaveroğlu *et al.*[28] and define the *brokerage score* of the country in a particular year as the weighted linear combination of broker graphlet degrees (i.e., orbits 51, 55, 59, 63, 67, 71, 75 and 79) using the coefficients obtained from the first canonical variate:

$$
\begin{aligned}
Brokerage(u) = {}& 0.903 \times DGDV_u[51] + 0.907 \times DGDV_u[55] + 0.907 \times DGDV_u[59] \\
& + 0.901 \times DGDV_u[63] + 0.897 \times DGDV_u[67] + 0.916 \times DGDV_u[71] \\
& + 0.904 \times DGDV_u[75] + 0.912 \times DGDV_u[79]
\end{aligned} \tag{1}
$$

Similarly, we quantify how *peripheral* a country is in the WTN of a particular year (by using orbits 52, 56, 60, 64, 68, 72, 76 and 80):

$$
\begin{aligned}
Peripherality(u) = {}& 0.286 \times DGDV_u[52] + 0.279 \times DGDV_u[56] + 0.284 \times DGDV_u[60] \\
& + 0.275 \times DGDV_u[64] + 0.279 \times DGDV_u[68] + 0.272 \times DGDV_u[72] \\
& + 0.281 \times DGDV_u[76] + 0.274 \times DGDV_u[80]
\end{aligned} \tag{2}
$$

These brokerage and peripheral scores enable us to track the 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 reconfirm previous observations obtained by undirected graphlet based analysis of WTN[28]. The first example is the loss of colonies of Great Britain (GBR), which was rivalling the USA as the world's leading broker of trades before decolonisation, that pushed it away from the world stage. The decline of trade brokerage position of Great Britain temporarily stabilized 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[54]. In contrast, the reunification of Germany (DEU) in 1991 transformed it to being the central economy of Europe[55]. Our scores also enable us to track the evolution of the emerging economies (Supplementary Fig. 1, panel b and c). The peripheral score of China drops after integrating Hong-Kong in 1984, which was the hub of trade between China and the rest of the world. China stopped being a peripheral trading country in 1995, after the adoption of the current foreign trade law in 1994, favouring international trade. Starting in 1995, China progressively became a broker in trade, and surpassed America in 2009, possibly benefiting from the fall of America's trade due to the global financial crisis. To refine the above presented brokerage score, we exploit the observation that in the WTN, peripheral countries are specialized towards export or import, and define import-specific and export-specific variants of the brokerage score according to the directionality of the trade between the broker role and the peripheral role (the formulas are presented in the Supplementary Material). Import- and export-specific variants of brokerage scores show that China's dominance as a broker in trade in the WTN is strongly driven by its export power, which is expected and hence validates our methodology (Supplementary Fig. 1, panels c and d).

In addition, the success of China as an emerging economy can be placed in parallel with other potential emerging economies, of Brazil (BRA), Russia (RUS), India (IND), Turkey (TUR) and Indonesia (IDN). All these countries have increased their brokerage scores in the last decade, albeit not as much as China. However, they remain largely peripheral in the WTN (Fig. 4, panel e and f). Moreover, import- and export-specific brokerage scores (Supplementary Fig. 1, panel e and f) show that while these potentially emerging economies have increased their export-brokerage scores (with Russia and India equalling or surpassing Great Britain), their import-brokerage scores remain far behind, which may prevent them from benefiting by buying (importing) commodities at low prices from peripheral exporting countries. Altogether, their peripherality in the WTN and their lack of import-brokerage power may explain why these countries do not meet the economists' expectations (see http://www.businessinsider.com/why-china-was-the-only-bric-to-succeed-2013-10?IR=T).

Biology: metabolic networks. *Linking wiring to function in metabolic networks.* In the cell, genes are transcribed into RNAs, which are translated into proteins, and these proteins interact with each other and with other molecules to perform their biological functions. Metabolic reactions are all chemical reactions that are needed to sustain the activity of the cell. These reactions are made possible (catalysed) by enzymes (specific proteins) and are represented as metabolic networks in which enzyme coding genes (nodes) are connected to each other with directed edges corresponding to the chains of chemical reactions that the enzymes participate in. As we already showed at the end of section "Evaluation of the methodology", similarity between the wiring patterns in metabolic networks of different species relates to their phylogenetic classification and could thus be used to refine our understanding of evolution.

Another key objective in biology is to understand the functions of genes. Because wet-lab experiments are costly, computational biology aims at predicting functions of the genes based on the similarity between their sequences, or on the similarity between the wiring patterns of the corresponding proteins in undirected protein-protein interaction networks[33]. While properly predicting and assessing enzyme-coding genes' functions from their wiring patterns in metabolic networks is out of the scope of this study, we investigate if directed graphlets could be used for such purpose.

To that aim, we cluster together human enzyme-coding genes according to the similarity of their wiring patterns in the human metabolic network and assess if the obtained clusters group together enzymes that have similar biological functions. We use the publicly available ascendant hierarchical classifier Chavl[56] to group together enzyme-coding genes from their directed graphlet degree vector similarity (see section "Directed graphlet-based statistics"). Unlike other common clustering methods, Chavl also proposes cuts of the classification tree based on likelihood linkage analysis[56]. On our data, Chavl produces two clusterings, a more generic one that groups together human genes into four clusters and a more specific one that groups human genes into 19 clusters. We use Gene Ontology (GO)[57] to annotate genes with GO terms representing the biological processes (BP) they participate in, their molecular functions (MF) and the cellular components (CC) in which they are localized. We downloaded experimentally confirmed GO annotations of the genes from NCBI (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz) in March 2015, which we standardized to the fifth level of the ontology[58–60]. In our experiment, we say that genes in a cluster share a biological function if the cluster is significantly enriched in a given GO term. We compute the probability of a GO term enrichment using the standard model of sampling without replacement[61]: $p = 1 - \sum_{i=0}^{X-1} \binom{K}{i}\binom{M-K}{N-i} / \binom{M}{N}$, where N is the size of the cluster (only annotated genes from the cluster are taken into account), X is the number of genes in the cluster that are annotated with the GO term in question, M is the number of all genes in the network that are annotated with any GO term, and K is the number of genes in the network that are annotated with the GO term in question. A cluster is significantly enriched in a given GO term if the corresponding p-value is smaller than or equal to 0.05. As presented in Fig. 5 (panel a), all clusters in our two clusterings are significantly enriched in GO terms. This confirms that enzymes involved in similar node roles in metabolic networks also perform similar biological functions, and this property could be used for transferring biological annotations from well-studied enzyme-coding genes to less well-studied ones.

Uncovering topologically orthologous functions. Recently, Davis *et al.*[32] showed that in protein-protein interaction networks, not only proteins having similar function are similarly wired in the network, but also that these function-related wiring patterns can be evolutionarily conserved across the networks of different species. They termed biological functions performed through evolutionarily conserved wiring patterns of their proteins as *topologically orthologous*. Uncovering topologically orthologous functions allows for transferring of biological
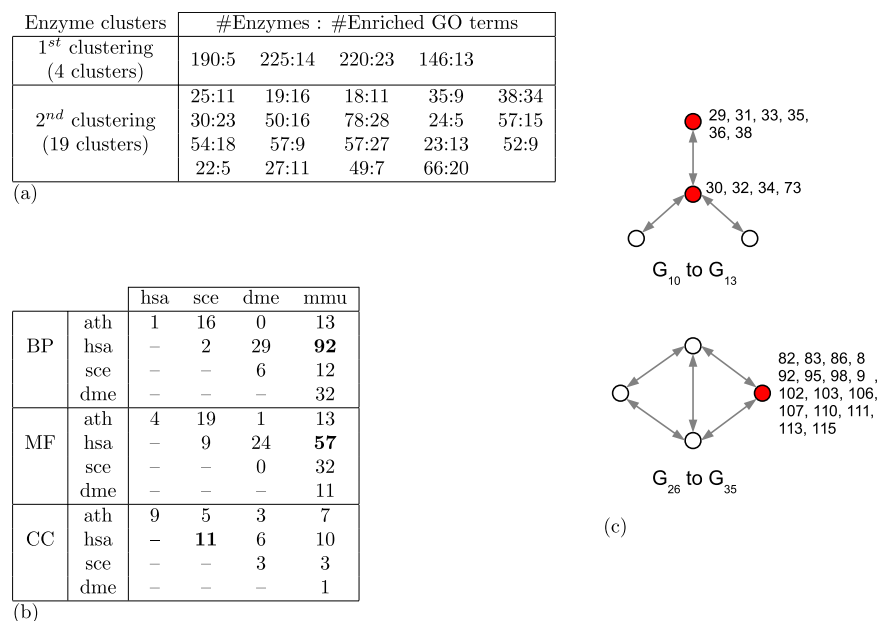
Enzyme clusters	#Enzymes : #Enriched GO terms			
1st clustering (4 clusters)	190:5	225:14	220:23	146:13
2nd clustering (19 clusters)	25:11	19:16	18:11	35:9	38:34
	30:23	50:16	78:28	24:5	57:15
	54:18	57:9	57:27	23:13	52:9
	22:5	27:11	49:7	66:20	

(a)

		hsa	sce	dme	mmu
BP	ath	1	16	0	13
	hsa	–	2	29	**92**
	sce	–	–	6	12
	dme	–	–	–	32
MF	ath	4	19	1	13
	hsa	–	9	24	**57**
	sce	–	–	0	32
	dme	–	–	–	11
CC	ath	9	5	3	7
	hsa	–	**11**	6	10
	sce	–	–	3	3
	dme	–	–	–	1

(b)

Figure 5. Analysis of metabolic networks. (**a**) GO term enrichment of the two node role based clusters of human enzymes. For each of the two clustering, the GO terms enrichment of each cluster is presented as a pair x:y, where x is the number of enzymes in the cluster and y is the number of enriched GO terms. (**b**) The number of topologically orthologous GO terms per species pairs. (**c**) The significant orbits (in red) for DNA metabolic process annotated enzymes; orbit numbers are presented next to the nodes in red (there are many orbits because our methodology treats different edge directions in separate graphlets).

annotations across networks of different species. We use our CCA framework to investigate if such topologically orthologous functions can be found in the metabolic networks of human and of four commonly used model organisms: baker's yeast (*Saccharomyces cerevisiae*), fruit fly (*Drosophila melanogaster*), thale cress (*Arabidopsis thaliana*) and mouse (*Mus musculus*).

For each species, we create its node role matrix, *R*, and its GO annotation matrix, *A*, in the following way. For both matrices, the same row index represents the same enzyme-coding gene. In the node role matrix, *R*, values in a variable vector are the 129 directed graphlet degrees of the enzyme in the corresponding metabolic network, while in the GO annotation matrix, *A*, they are the Gene Ontology annotations of the corresponding gene (each column correspond to a specific GO term, and the entry is set to 1 if the gene is annotated with the term and 0 otherwise). We considered separately the GO terms describing biological processes, BP, molecular functions, MF, and cellular components, CC, so each species has three different GO annotation matrices. We uncover topologically orthologous biological functions in two steps. In the first step, we apply our CCA framework between the node role matrix and biological annotation matrix of each species separately and use the corresponding association matrix (which predict the biological annotations of a gene from the directed graphlet degrees of the corresponding enzyme in the metabolic network) to uncover biological functions that are statistically significantly predicted from topology (statistical significance is computed according to 1,000 randomized experiments in which rows in the node role matrix are randomly shuffled; a biological function is significantly predicted if it's p-value, after Benjamini Hochberg correction, is lower than 0.05). In the second step, biological functions that are statistically significantly predicted in the two species and that have predictions based on similar combinations of directed graphlet degrees, are considered as topologically orthologous.

As presented in Fig. 5 (panel b), human and mouse have the largest number of biological functions that are topologically orthologous. This is expected since human and mouse are the closest species among the five considered ones. A further validation is that we show that DNA metabolic process, which is one of the essential process for all living organisms, is topologically orthologous across all the considered species (appearing as topologically orthologous in eight out of the ten pairs of species); its significantly associated node roles (orbits) are presented in Fig. 5 (panel c). Our topologically orthologous biological functions may be used to refine the inter-species based function predictions. However, a deeper investigation on the meaning of the uncovered relationships between node roles and biological functions is a subject of future research.

Other domains. Directed network data are abundant in many other domains, e.g. brain networks, e-mail communication networks, citation networks, world wide web networks, and internet peer-to-peer networks, to name a few. Research on uncovering their organizational principles and function will continue. Applying our new directed graphlet-based methods, or devising new methods for their analyses[62,63] are a subject of future research.

In this study, we use our directed graphlets to propose three new alignment-free network similarity measures (DGCD, DRGF and DGDDA), which directly compare directed networks without searching for their aligned regions. Finding the aligned regions between networks is done by alignment-based methods[64], which compute

the node-to-node correspondences (called alignments) between the networks. Since some of the aligners for undirected networks use graphlets and graphlet degrees to guide their node mapping processes, our directed graphlets could also be used to define novel aligners for directed networks.

Concluding Remarks

By generalising the graphlet-based node role framework to directed graphlets, we have enabled advanced descriptive and predictive analyses of directed networked data that could not have been achieved by using undirected graphlets or other directed network statistics. We have shown that correlations between directed graphlet based node roles, which we capture in the directed graphlet correlation matrices, can be used to sensitively and robustly group networks from the same model and separate those from different models. Analysing WTNs over years, our framework allows us to: (1) uncover the directed core-broker-periphery organisation of the world trade, in which peripheral countries are specialized in either import or export roles; (2) predict the economic attributes of countries from their positioning in the WTN better than by using traditional directed descriptors or undirected graphlets that may inform regulators about benefits of trade agreements and predict success of an emerging economy; (3) identify relationships between the roles of countries in the WTN and their economic attributes; (4) finely track the dynamics of the WTN, yielding insights into dominating in trade as brokers of import or export. Our methodology is general and works in other domains. We observe that enzymes involved in similar patterns of metabolic reactions in the metabolic network of human are also involved in similar biological functions, and we identify conserved biological functions performed by enzymes that are similarly wired in metabolic reactions across species. Hence, our directed graphlet based node role framework is universal and promises to deliver insights in the other domains of data science.

References

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. *Nature Physics* **9,** 126–128 (2013).
3. Scott, J. *Social Network Analysis* (Sage, 2012).
4. Ward, M. D., Stovel, K. & Sacks, A. Network analysis and political science. *Annual Review of Political Science* **14,** 245–264 (2011).
5. Junker, B. H. & Schreiber, F. *Analysis of Biological Networks*, vol. 2 (John Wiley & Sons, 2011).
6. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* **12,** 56–68 (2011).
7. Garlaschelli, D. & Loffredo, M. I. Structure and evolution of the world trade network. *Physica A: Statistical Mechanics and its Applications* **355,** 138–144 (2005).
8. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286,** 509–512 (1999).
9. Kali, R. & Reyes, J. The architecture of globalization: a network approach to international economic integration. *Journal of International Business Studies* **38,** 595–620 (2007).
10. Borgatti, S. P. & Everett, M. G. Models of core/periphery structures. *Social Networks* **21,** 375–395 (1999).
11. De Benedictis, L. & Tajoli, L. The world trade network. *The World Economy* **34,** 1417–1454 (2011).
12. Della Rossa, F., Dercole, F. & Piccardi, C. Profiling core-periphery network structure by random walkers. *Scientific Reports* **3,** 1467 (2013).
13. Lacroix, V., Cottret, L., Thébault, P. & Sagot, M.-F. An introduction to metabolic networks and their structural analysis. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* **5,** 594–617 (2008).
14. Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298,** 824–827 (2002).
15. Zhu, D. & Qin, Z. S. Structural comparison of metabolic networks in selected single cell organisms. *BMC bioinformatics* **6,** 8 (2005).
16. Shellman, E. R., Burant, C. F. & Schnell, S. Network motifs provide signatures that characterize metabolism. *Molecular BioSystems* **9,** 352–360 (2013).
17. Heymans, M. & Singh, A. K. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics* **19,** i138–i146 (2003).
18. Zhang, Y. *et al.* Phylophenetic properties of metabolic pathway topologies as revealed by global analysis. *BMC Bioinformatics* **7,** 252 (2006).
19. Francke, C., Siezen, R. J. & Teusink, B. Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology* **13,** 550–558 (2005).
20. Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L. & Palsson, B. Ø. Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology* **7,** 129–143 (2009).
21. Pearcy, N., Crofts, J. J. & Chuzhanova, N. Network motif frequency vectors reveal evolving metabolic network organisation. *Molecular BioSystems* **11,** 77–85 (2015).
22. Ganter, M., Kaltenbach, H.-M. & Stelling, J. Predicting network functions with nested patterns. *Nature Communications* **5** (2014).
23. Ghahramani, Z. An introduction to hidden markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence* **15,** 9–42 (2001).
24. Cook, S. A. The complexity of theorem-proving procedures. In *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, STOC 71, 151–158 (ACM, New York, NY, USA, 1971).
25. Newman, M. *Networks: An Introduction* (Oxford University Press, Oxford, 2009).
26. Yan, G. *et al.* Spectrum of controlling and observing complex networks. *Nature Physics* **11,** 779–786 (2015).
27. 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).
28. Yaveroğlu, O. N. *et al.* Revealing the hidden language of complex networks. *Scientific Reports* **4** (2014).
29. Pržulj, N., Corneil, D. & Jurisica, I. Modeling interactome: Scale-free or geometric? *Bioinformatics* **20,** 3508–3515 (2004).
30. Pržulj, N. Biological network comparison using graphlet degree distribution. *Bioinformatics* **23,** 177–183 (2007).
31. Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. & Tatham, R. L. *Multivariate Data Analysis*, vol. 6 (Pearson Prentice Hall Upper Saddle River, NJ, 2006).
32. Davis, D., Yaveroğlu, O. N., Malod-Dognin, N., Stojmirovic, A. & Pržulj, N. Topology-function conservation in protein-protein interaction networks. *Bioinformatics* **31,** 1632–1639 (2015).
33. Milenković, T. & Pržulj, N. Uncovering biological network function via graphlet degree signatures. *Cancer Informatics* **6,** 257 (2008).
34. Spearman, C. The proof and measurement of association between two things. *The American Journal of Psychology* **15,** 72–101 (1904).
35. Weenink, D. Canonical correlation analysis. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, vol. 25, 81–99 (2003).
36. Pearson, K. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* 240–242 (1895).

37. Albert, A. *Regression and the Moore-Penrose pseudoinverse* (Elsevier, 1972).
38. Erdös, P. & Rényi, A. On random graphs. *Publicationes Mathematicae* **6,** 290–297 (1959).
39. Bollobás, B., Borgs, C., Chayes, J. & Riordan, O. Directed scale-free graphs. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 132–139 (Society for Industrial and Applied Mathematics, 2003).
40. Vázquez, A., Flammini, A., Maritan, A. & Vespignani, A. Modeling of protein interaction networks. *Complexus* **1,** 38–44 (2002).
41. Penrose, M. Random geometric graphs. *Oxford Studies in Probability* **5** (2003).
42. Pržulj, N., Kuchaiev, O., Stevanovic, A. & Hayes, W. Geometric evolutionary dynamics of protein interaction networks. In *Pacific Symposium on Biocomputing*, vol. 2009, 178–189 (World Scientific, 2010).
43. Yaveroğlu, Ö. N., Milenković, T. & Pržulj, N. Proper evaluation of alignment-free network comparison methods. *Bioinformatics* btv170 (2015).
44. Wilson, R. C. & Zhu, P. A study of graph spectra for comparing graphs and trees. *Pattern Recognition* **41,** 2833–2841 (2008).
45. Fawcett, T. An introduction to roc analysis. *Pattern Recognition Letters* **27,** 861–874 (2006).
46. 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 Research* **34,** 5966–5973 (2006).
47. Stumpf, M. P. H., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M. & Wiuf, C. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences* **105,** 6959–6964 (2008).
48. Röttger, R., Rückert, U., Taubert, Jan. & Baumbach, J. How little do we actually know? On the size of gene regulatory networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **9,** 1293–1300 (2012).
49. Kanehisa, M. Toward pathway engineering: a new database of genetic and molecular pathways. *Science & Technology Japan* **59,** 34–38 (1996).
50. Serrano, M. A., Boguñá, M. & Vespignani, A. Patterns of dominant flows in the world trade web. *Journal of Economic Interaction and Coordination* **2,** 111–124 (2007).
51. Piccardi, C. & Tajoli, C. Existence and significance of communities in the World Trade Web. *Physical Review E* **85,** 066119 (2012).
52. Freund, C. L. The trade response to global downturns: historical evidence. *World Bank Policy Research Working Paper Series, Vol* (2009).
53. Gligorijević, V., Malod-Dognin, N. & Pržulj, N. Integrative methods for analysing big data in precision medicine. *Proteomics* (2015).
54. Kindleberger, C. P. Government policies and changing shares in world trade. *The American Economic Review* **70,** 293–298 (1980).
55. Mundell, R. A. A reconsideration of the twentieth century. *American Economic Review* **90,** 327–340 (2000).
56. Lerman, I.-C. Foundations of the likelihood linkage analysis (lla) classification method. *Applied Stochastic Models and Data Analysis* **7,** 63–76 (1991).
57. Ashburner, M., Ball, C. A., Blake, J. A. *et al.* Gene ontology: tool for the unification of biology. *Nature Genetics* **25,** 25–29 (2000).
58. Singh, R., Xu, J. & Berger, B. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences* **105,** 12763–12768 (2008).
59. Liao, C.-S., Lu, K., Baym, M., Singh, R. & Berger, B. IsorankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics* **25,** i253–i258 (2009).
60. Alkan, F. & Erten, C. Beams: backbone extraction and merge strategy for the global many-to-many alignment of multiple PPI networks. *Bioinformatics* **30,** 531–539 (2014).
61. Kuchaiev, O., Milenković, T., Memišević, V., Hayes, W. & Pržulj, N. Topological network alignment uncovers biological function and phylogeny. *Journal of The Royal Society Interface* **7,** 1341–1354 (2010).
62. Aparício, D., Ribeiro, P. & Silva, F. Extending the Applicability of Graphlets to Directed Networks. *IEEE ACM Transactions on Computational Biology and Bioinformatics* **PP,** 1–1 (2016).
63. Trpevski, I., Dimitrova, T., Boshkovski, T. & Kocarev, L. Graphlet characteristics in directed networks. *arXiv* 1603.05843 (2016).
64. Clark, C. & Kalita, J. A comparison of algorithms for the pairwise alignment of biological networks. *Bioinformatics* **30,** 2351–2359 (2014).

Acknowledgements

Author Contributions

A.S. generalized graphlets, network distances and network models to the directed case and performed the validation on model networks and the study of metabolic networks. N.M.-D. performed the validation against undirected distances and the study of world trade data, and wrote the manuscript. O.N.Y. helped conducting computations. N.P. conceived and directed the study. All authors analysed the results and reviewed the manuscript.

Additional Information