Parameterless Pruning Algorithms for Similarity-Weight Network and Its Application in Extracting the Backbone of Global Value Chain

Lizhi Xing^{1,2†}, Yu Han¹

¹College of Economics and Management, Beijing University of Technology, Beijing 100124, China

²International Business School, Beijing Foreign Studies University, Beijing 100089, China

Abstract

Purpose: With the availability and utilization of Inter-Country Input-Output (ICIO) tables, it is possible to construct quantitative indices to assess its impact on the Global Value Chain (GVC). For the sake of visualization, ICIO networks with tremendous low-weight edges are too dense to show the substantial structure. These redundant edges, inevitably make the network data full of noise and eventually exert negative effects on Social Network Analysis (SNA). In this case, we need a method to filter such edges and obtain a sparser network with only the meaningful connections.

Design/methodology/approach: In this paper, we propose two parameterless pruning algorithms from the global and local perspectives respectively, then the performance of them is examined using the ICIO table from different databases.

Findings: The Searching Paths (SP) method extracts the strongest association paths from the global perspective, while Filtering Edges (FE) method captures the key links according to the local weight ratio. The results show that the FE method can basically include the SP method and become the best solution for the ICIO networks.

Research limitations: There are still two limitations in this research. One is that the computational complexity may increase rapidly while processing the large-scale networks, so the proposed method should be further improved. The other is that much more empirical networks should be introduced to testify the scientificity and practicability of our methodology.

Practical implications: The network pruning methods we proposed will promote the analysis of the ICIO network, in terms of community detection, link prediction, and spatial econometrics, etc. Also, they can be applied to many other complex networks with similar characteristics.

Citation: Xing, L.Z., & Han, Y. (2022). Parameterless pruning algorithms for similarity-weight network and its application in extracting the Backbone of Global Value Chain. *Journal of Data and Information Science*, 7(1), 57–75. <https://doi.org/10.2478/jdis-2022-0002>

Received: Jul. 11, 2021

Revised: Aug. 25, 2021;

Oct. 2, 2021

Accepted: Oct. 11, 2021

[†] Correspondence author: Lizhi Xing (Email: itwasa@163.com; ORCID: 0000-0001-9554-5414).

Research Paper

Originality/value: This paper improves the existing research from two aspects, namely, considering the heterogeneity of weights and avoiding the interference of parameters. Therefore, it provides a new idea for the research of network backbone extraction.

Keywords Inter-Country Input-Output Table; Global Production System; Global Value Chain Network; Backbone Extraction

1 Introduction

With trade barriers being removed and driven by the scientific and technological revolution, the global economic integration based on the Global Value Chain (GVC) is already established. In order to analyze various economic phenomena happened within and across countries or regions on the GVC, it is very important to find and adopt appropriate database to illustrate the intricate inter-sector economic activities. In recent years, with the penetration of Complex Network theory in various disciplines, scholars found that the matrix structure of Input-Output (IO) data can be directly or slightly modified to be the adjacency matrix of the network models, which will be the basis of Social Network Analysis (SNA). Unfortunately, lots of related studies were negatively influenced by the features of IO network, which are the highly dense structure and the heterogeneous edge weight distribution. For instance, the degree of each node is the same (equal to the maximum), so as the between centrality, closeness centrality, clustering coefficient, etc. Therefore, we need to propose effective methods to extract the backbone of IO network, by reducing the scale of network and retaining only the meaningful connections.

Ideally, we hope that the crucial information of network could be retained as much as possible with the decline of number of edges. Sometimes, a weighted edge is numerically insignificant but functionally significant, and reckless deletion will result in a useless broken structure. Therefore, the effective network backbone extraction method has to be a trade-off between the number of retained edges and the gross of retained weights.

In this paper, considering that the edges of the IO networks represent the flow of bilateral economic activities, we obtain the IO network backbones by retaining the Strongest Relevance Path Length (SRPL) (Xing et al., 2019) between industry sectors, which is called the Searching Paths (SP) method. In addition, there is an unequal relationship between inputs and outputs. We propose a new heuristic filtering algorithm Filtering Edges (FE) based on the assumption that a minority of edges hold the most weights in the network. The SP method and FE method come from global and local perspectives respectively, and both of them can converge naturally through algorithms rather than rely on artificial parameters.

The remainder of this paper is organized as follows: Section 2 contains the literature review and the shortcomings of the present study. Section 3 introduces the two methods and their mixed strategies, as well as the specific implementation process of network pruning. Section 3 describes the data and modeling process. Section 4 compares the method proposed in this paper with the existing classical methods. Section 5 is the conclusions, in which the main conclusions and further research direction of this paper are summarized.

2 Literature review

The problems that higher network density causes to SNA are not unique to IO network. For example, regionally clustered edges pose challenges to explicitly depict the overall structure of geographical network (Hennemann, 2013). Higher network density is a serious obstacle for cluster and spectrum analyses, which are usually applied in the sparse networks. Therefore, scholars in different fields including sociology (Nick et al., 2013), economics (Marotta et al., 2015), and computer science (Foti, Hughes, & Rockmore, 2011) proposed a series of methods for extracting network backbones. In detail, these methods can be classified into three categories—network sampling, coarse graining, and filtering. First, Network sampling means obtaining the main network structure through sampling representative nodes or edges, which is extensively overviewed in the study of Ahmed, Neville, and Kompella (2014). Second, to turn the original network into a smaller skeleton, coarse graining means merging nodes or edges according to certain rules (Itzkovitz et al., 2004), such as the attribute of node or edge (Toivonen et al., 2012), community structure (Blagus et al., 2012). The most common one is the filter-based method. It typically defines a criterion for a node or edge to examine their importance or relevance to the network, after which the redundant ones are removed in a stepwise procedure (Dai, Derudder, & Liu, 2018). In practice, there are two filtering perspectives.

An important angle is to preserve the core by removing structures that are under an overall topological characteristic, which can be called the global method. For nodes, a range of indices such as degree (Carmi et al., 2007; Kitsak et al., 2010; Siganos, Tauro, & Faloutsos, 2006), local centrality (Chen et al., 2012) and Leader Rank (Li et al., 2014; Lv et al., 2011) have been applied to identify influential nodes. Further, backbone network can be obtained by reserving the smallest possible subset of highly influential nodes and their relevant links (Malang et al., 2020). As for edges, link weights (Zhang & Zhu, 2013; Zhang et al., 2018; Zhao et al., 2014) and network motifs (Cao, Ding, & Shi, 2019) were used to extract functional backbones while edge betweenness (Kim, Noh, & Jeong, 2004) and the shortest path (Grady,

Research Paper

Thiemann, & Brockmann, 2011; Zhang et al., 2014) were used to preserve more structural features. Among them, the simplest method is to remove all edges with weight below a certain threshold, i.e. the Global Threshold (GT) method. Zhang et al. defined h-backbone method that extract both the high-strength links and network bridges, which can preserve important information about both network function and structure (Zhang, Stanley, & Ye, 2018).

The other is to retain the components which are statistically significant at the local level. Disparity filter (DF) is the most representative method that focuses on the local distribution of edge weights, preserving the edges that represent statistically significant deviations with respect to a null model (Serrano, Boguna, & Vespignani, 2009). The improvement of DF comes from different aspects (Coscia & Neffke, 2017). For example, Radicchi, Ramasco, & Fortunato, proposed a weight-filtering technique based on a global null model, which accounts for the full topology while preserving the heterogeneity of the weight distribution (Radicchi, Ramasco, and Fortunato, 2011). Foti, Hughes, and Rockmore presented an improvement relying solely on the empirical distribution to judge statistical significance instead of any null model (Foti, Hughes, & Rockmore, 2011). Bu et al. designed a stricter filter rule and an iterative local search model to avoid outliers and improve efficiency (Bu et al., 2014). Zhang et al. and Cao, J., Ding, C. and Shi, B. identified which edges for each node should be preserved by evaluating whether or not their link involvement or motif weights are compatible with the null model respectively (Cao, Ding, & Shi, 2019; Zhang et al., 2011).

However, most of the previous studies on backbone extraction have the following aspects to be further considered: 1) Most of the proposed methods are general without considering the actual meaning of edge weight. The distribution of industrial relationship in international trade networks, such as IO network, is obviously unequal or heterogeneous. 2) Artificial parameters are required in many methods, which make the backbone network not unique. 3) Both global and local perspectives have their disadvantages. The global methods may overlook the local heterogeneity and multiscale structures of the network, and the local methods may ignore “the crucial weak ties” (Granovetter, 1973) that connect different communities. For these reasons, we need backbone extraction methods that are sensitive to the heterogeneity of weights and independent of parameters.

3 Methodology

3.1 Searching paths

According to the comparison between real IO relations and SRPL paths in Xing and Han (2021), we propose a global backbone extraction method is efficiently to

reserving the main industrial relevance with the fastest reaction speed and the most frequent traffic.

$$SRPL_{ij}^{(k)} = \max_{i,j,k \in \{1,2,\dots,N\}} \left\{ w_{ij}^{(k-1)}, \frac{w_{ik}^{(k-1)} w_{kj}^{(k-1)}}{w_{ik}^{(k-1)} + w_{kj}^{(k-1)}} \right\} \quad (1)$$

where $\frac{w_{ik}^{(k-1)} w_{kj}^{(k-1)}}{w_{ik}^{(k-1)} + w_{kj}^{(k-1)}}$, namely Relevance Path Length (RPL) uses the edge weight product to measure the impact from the industrial structure, and uses the edge weight sum to provide distance information and serve as a barrier function. If it is greater than $w_{ij}^{(k-1)}$, we keep the record as it is, otherwise just equalize it to $w_{ij}^{(k-1)}$. $SRPL_{ij}^{(k)}$ is the SRPL between nodes i and j , representing the shortest Industrial Value Chain (IVC) with each link between the two industrial sectors closely connected, which achieves convergence through iteration of Revised Floyd-Warshall Algorithm (RFWA) algorithm (Xing et al., 2019). According to the Eq. (1), the SRPL of any two nodes in the global network can be calculated, and finally a new matrix is formed:

$$SRPL^{(N)} = \begin{pmatrix} SRPL_{11}^{(N)} & \cdots & SRPL_{N1}^{(N)} \\ \vdots & \ddots & \vdots \\ SRPL_{N1}^{(N)} & \cdots & SRPL_{NN}^{(N)} \end{pmatrix} \quad (2)$$

When the corresponding elements of the new matrix $SRPL'$ and the edge weight matrix W are exactly equal, it means that the optimal path is the most direct propagation path between the two nodes, that is, there is no need to pass through another node. From the perspective of GVC, we believe that between specific upstream and downstream sectors, whether it is value stream or material stream, it tends to travel along the SRPL path rather than in a random manner.

$$\tilde{w}_{ij} = \begin{cases} w_{ij} & , \quad w_{ij} = SRPL_{ij}^{(N)} \\ 0 & , \quad otherwise \end{cases} \quad (3)$$

Taking $\tilde{W} = \{\tilde{w}_{ij}\}$ as an adjacency matrix, we will acquire a new network model. According to Burt's Structural Holes theory, brokers on the propagation paths have gained two types of advantages, namely, greater information advantages and more intermediate benefits (Burt, 1992). There is no doubt that too many brokers will damage the efficiency of the network. Therefore, the strong correlation in the industrial complex network can be extracted by retaining the direct IO relationship that carries the inter- industry intermediate products.

3.2 Filtering edges

Enlightened by H-Index (Hirsch, 2005) and the Pareto Principle, we present a novel heuristic algorithm to effectively prune the dense and weighted IO network,

Research Paper

which is named X-Index Filtering Algorithm (XIFA). XIFA algorithm is a sort of mixed quantitative index that takes into account the scope and intensity of industrial sectors and extracts the main topology of IO network according to the heterogeneity of IO relations. From the perspective of complex networks, for the nodes owning diversely weighted edges, we only need to retain a small number of them with extremely large weights, while for the nodes owning similarly weighted edges, many but no more than 50% of them will stay, as shown in Figure 1.

Figure 1. Three possible situations in the application of XIFA algorithm. Note: (a) The source node has only one weighted edge connected to it, and 100% of its strength is allocated on it; (b) Top 20% of weighted edges carry 80% of the strength of source node. (c) Any 50% of weighted edges carry 50% of the strength of source node.

In addition to the case in Figure 1 (a), the larger the X-index, the more even distribution of edge weights, and the more edges are retained; the smaller the X-index, the more concentrated distribution of edge weights, and the fewer edges are retained. The range of the X-index is theoretically (0,0.5]. The core idea of XIFA is as follows: An industrial sector has backward index x (XI^B) / forward index x (XI^F) if top x percent of its relations to all upstream/downstream sectors occupy at least $(1 - x\%)$ of its total input/output amount of intermediate goods (Guan, Ren, & Xing, 2021). Therefore, when performing network pruning, we retain only the edge of the top XI^B present or XI^F present. Formula deduction and practical calculation procedures are presented in Table 1.

This algorithm serves a dual purpose. One is to retain the key outward connections from the perspective of provider, and the other is to retain the key inward connections from the perspective of consumer. It should be noted that if all the key export-oriented links cannot touch the industrial sectors of some weak countries as consumers, then these sectors will be separated from the GVC, which undermines the integrity of GVC. Therefore, we carry out the merge after network pruning from the perspective of out-degree and in-degree respectively. However, according to the nature of this algorithm, this pruning method is still based on the relations of adjacent nodes, rather than global information, so it is still possible to ignore the “critical” and “weak” edges.

Table 1. The procedure of network pruning based on XIFA.

Procedure	Column Deletion of Input Relations	Row Deletion of Output Relations
Network	$W = (w_{ij})_{N \times N} \quad i, j \in [1, N]$	
Refactoring	$\bar{w}_1 = \text{descend}(w_{11}, w_{21}, \dots, w_{N1})^T$ $\bar{w}_2 = \text{descend}(w_{12}, w_{22}, \dots, w_{N2})^T$ $\dots$ $\bar{w}_N = \text{descend}(w_{1N}, w_{2N}, \dots, w_{NN})^T$ $\bar{W} = (\bar{w}_1, \bar{w}_2, \dots, \bar{w}_N) = (\bar{w}_{sj})_{N \times N}$ $s, j \in [1, N]$	$\bar{w}_1 = \text{descend}(w_{11}, w_{12}, \dots, w_{1N})$ $\bar{w}_2 = \text{descend}(w_{21}, w_{22}, \dots, w_{2N})$ $\dots$ $\bar{w}_N = \text{descend}(w_{N1}, w_{N2}, \dots, w_{NN})$ $\bar{W}^T = \begin{pmatrix} \bar{w}_1 \\ \bar{w}_2 \\ \dots \\ \bar{w}_N \end{pmatrix} = (\bar{w}_{it})_{N \times N}$ $i, t \in [1, N]$
Conditions	$\forall a_1, a_2, \dots, a_j \in [1, N]$ $\begin{cases} \frac{\sum_{s=1}^{a_j} \bar{w}_{sj}}{\sum_{s=1}^N \bar{w}_{sj}} \geq 1 - \frac{a_j}{N} \\ \frac{\sum_{s=1}^{a_j-1} \bar{w}_{sj}}{\sum_{s=1}^N \bar{w}_{sj}} < 1 - \frac{a_j-1}{N} \end{cases}$	$\forall b_1, b_2, \dots, b_i \in [1, N]$ $\begin{cases} \frac{\sum_{t=1}^{b_i} \bar{w}_{it}}{\sum_{t=1}^N \bar{w}_{it}} \geq 1 - \frac{b_i}{N} \\ \frac{\sum_{t=1}^{b_i-1} \bar{w}_{it}}{\sum_{t=1}^N \bar{w}_{it}} < 1 - \frac{b_i-1}{N} \end{cases}$
Definition	$XI_j^B = \frac{a_j}{N}$ $XI^B = (XI_j^B)_{N \times 1}$	$XI_i^F = \frac{b_i}{N}$ $XI^F = (XI_i^F)_{N \times 1}$
Pruning	$\bar{w}_{ij} = \begin{cases} w_{ij} & , \quad w_{ij} = \bar{w}_{sj} \text{ and } s \leq a_j \\ 0 & , \quad \text{otherwise} \end{cases}$	$\bar{w}_{ij} = \begin{cases} w_{ij} & , \quad w_{ij} = \bar{w}_{it} \text{ and } t \leq b_i \\ 0 & , \quad \text{otherwise} \end{cases}$
Merging	$\tilde{w}_{ij} = \begin{cases} w_{ij} & 0 \quad \bar{w}_{ij} \neq 0 \text{ or } \bar{w}_{ij} \neq \\ 0 & , \quad \text{otherwise} \end{cases}$	
Result	$\tilde{W} = (\tilde{w}_{ij})_{N \times N}$	

Note: “descend ()” means to arrange the vectors in descending order.

3.3 Mixed strategy

In consideration of balancing the overall and local, the preferable algorithm in theory for extracting the backbone of GVC with enough structural and functional information has three steps.

- Step 1. Find the edges acting as crucial bridges between nodes or communities based on the RFWA.
- Step 2. Find the edges retaining the great majority of weights based on the XIFA.
- Step 3. Merge the edges of Step 1 and 2 and add their adjacent nodes.

Research Paper

To illustrate this mixed strategy, Figure 2 (a) shows a five-node directed and weighted network with self-loops, in which the larger the edge weight, the closer the asymmetric interactions. When conducting the SP method, we search globally for key chains seeing both the efficiency and effectiveness. An edge will be preserved if it happens to be located on the most direct path between two nodes and the strongest relevance path as well (see Figure 2 (b)). The FE method, from the perspective of local information, retains the pivotal links that play an important role in the adjacent edges of each node, by examining them in two ways according to the XIFA (see Figure 2 (c)). In this case, the mixed strategy, denoted by SPFE, increases another two edges on the basis of the FE-based backbone (see Figure 2 (d)).

Figure 2. Getting the SPFE from an original network.

4 Data and modeling

With the availability and utilization of global IO database, especially Inter-Country Input-Output (ICIO) tables, relevant research is no longer limited to the economic system of a single country, but extends to inter-regional and even global economic systems. Typical IO or ICIO table includes three different areas, namely value-added (VA), intermediate use (Z) and final demand (Y). It is possible that the whole global economic system can be abstracted to a Multi-Layer Network as shown in Figure 3 (b), which includes three layers: the value-added layer, the intermediate use layer, and the final demand layer. Commonly, the GVC-related researches are concerned with the global production system, which is well profiled by the intermediate use part of ICIO table as shown in Figure 3 (a).

Two Countries with Two Sectors		Intermediate Use (Z)				Final Demand (Y)			
		Country A		Country B		Country A		Country B	
		Sector 1	Sector 2	Sector 1	Sector 2	Sector 1	Sector 2	Sector 1	Sector 2
Country A	Sector 1								
	Sector 2								
Country B	Sector 1								
	Sector 2								
Value-Added (VA)									

(a) ICIO Table Including Two Countries with Two Sectors

(b) Tripartite Valued Graph Based on ICIO Table

(c) Graph Form of Intermediate Use Part

Figure 3. Relations between ICIO table, multi-layer and single-layer ICIO network.

To embody the non-linear relations between worldwide industrial sectors, we adopt the intermediate use part of ICIO table as the basis of network modeling. In this ICIO network, a sector within a country or region is considered as a node and the inter- industry IO relation as an edge, and its edge weight represents the sale and purchase relations between producers and consumers, quantified by cash flux. Thus, a graph $G = (V, E, W)$ containing n nodes is drawn and denotes a node set V . Pairs of nodes are linked by edges to reflect their interdependencies, thereby forming an asymmetric edge set E . In this valued graph, the function of set E can be substituted by that of weight set W , which can be directly extracted from the region of inter-country and inter- industry use and supply in ICIO table. Finally, we name this directed and weighted network the Global Industrial Value Chain Network (GIVCN) model as shown in Figure 3 (c).

By the comparison of the main ICIO databases available for now, the year of 2014 as the latest and the greatest common divisor exists in World Input-Output

Research Paper

Database (WIOD), OECD-WTO Database on Trade in Value-Added (OECD-TiVA), Eora Multi-Region Input-Output Table Database (Eora-MRIO) and Asian Development Bank Multi-Regional Input-Output Tables (ADB-MRIO), so we take them as the basis of modeling and empirical analysis. Two of them, WIOD and TiVA, provide more granular sectoral information and are more rooted in official statistics, and are widely used to exploit the analytical richness of trade-focused empirical analysis. Although Eora covers 189 economies, the sectoral data provided is highly aggregated. ADB, as an extension to WIOD, facilitates the production and analysis of GVC-related statistics for more Asian economies and becomes a rich source of economic information for research and policymaking.

To facilitate the calculation and visualization in empirical analysis, sectors are further merged into four categories, including Agriculture (SC1), Mining (SC2), Manufacturing (SC2) and Services (SC4). The four GIVCN models based on the ICIO databases of the latest version are hence named “GIVCN-WIOD-SC4-2014”, “GIVCN-TiVA-SC4-2014”, “GIVCN-Eora-SC4-2014” and “GIVCN-ADB-SC4-2014” respectively, and statistics for the structural properties of full networks includes the number of edges $|E|$, average distance $\langle d \rangle$, mean node degree $\langle K \rangle$ and clustering coefficient $\langle C \rangle$, which are shown in Table 3.

Table 2. The basic information of ICIO databases.

Database	Version	Time Span	Country / Region	Industrial Sector	Abbr
WIOD	2016 Release	2000–2014	44	56	WIOD2016
	2013 Release	1995–2011	41	35	WIOD2013
OECD-TiVA	2020 Release	1995–2018	Unknow	Unknow	TiVA2020
	2018 Release	2005–2015	65	36	TiVA2018
	2016 Release	1995–2011	64	34	TiVA2016
	2015 Release	1995, 2000, 2005, 2008–2011	62	34	TiVA2015
Eora-MRIO	V199.82	1990–2015	189	26	Eora26
ADB-MRIO	Updated to 2019	2000, 2007–2019	63	35	ADB2019

Note: In addition to sovereign states, the Rest of World (RoW) is taken as an independent economic entity in WIOD, OECD and ADB, most of which belong to the developing countries; In TiVA2018, for “intermediates”, “value added” and “output”, data for Mexico and China are split into MX1, MX2 and CN1, CN2, respectively.

Table 3. Illustration of properties of the original networks.

Networks	$ E $	$\langle d \rangle$	$\langle K \rangle$	$\langle C \rangle$
GIVCN-WIOD-SC4-2014	30,795 (30,976)	1.006 (1.000)	174.972 (176.000)	1.000 (1.000)
GIVCN-TiVA-SC4-2014	66,600 (67,600)	1.007 (1.000)	257.143 (260.000)	0.996 (1.000)
GIVCN-Eora-SC4-2014	571,536 (571,536)	1.000 (1.000)	756.000 (756.000)	1.000 (1.000)
GIVCN-ADB-SC4-2014	57,761 (63,504)	1.069 (1.000)	231.972 (252.000)	0.971 (1.000)

If the information of edge weight is not considered, the GIVCN model has a high density of connections, which may give it access to the shorter average path and the

greater clustering coefficient (see Table 3). Therefore, both of them are strong evidence that this model is a small-world network. If the weight is considered, the heavy tails in Figure 4 reveal that the distributions of edge weight are extremely heterogeneous. Most IO relations carry on very little value flow, while some pass massive intermediate goods on the GVC. This phenomenon reflects the GIVCN model is a sort of scale-free network in nature. In pruning, the above two points require us to consider both the removal of redundancy and the retention of weights.

Figure 4. Edge weight distribution of ICIO networks under the double logarithmic coordinate.

5 Results and discussions

5.1 Preservation of structural information

Three backbone extraction methods are applied to the “GIVCN-” models together with the simplest GT method and the classical DF method. Despite the subnetworks still look a bit like a ball of yarn, it is possible to analyze the heterogeneous characteristics of global production system based on the simplified topological structure (see Appendix). Given that E_{SP} and E_{FE} are the edge sets of GIVCN-SP-2014 model and GIVCN-FE-2014 model, respectively, that of the whole one will be $E_{SPFE} = E_{FE} \cup E_{SP}$. We use three sorts of colored lines to differentiate

Research Paper

which model the links come from, and Figure 5 shows that GIVCN-SP-2014 is completely contained in GIVCN-FE-2014 in most cases. As an exception, GIVCN-Eora-SC4-2014 has six edges that only belong to the sub-network based on SP method, because the size of the model increases the heterogeneity of edge weight. After all, there are always some puny industry sectors or industrial linkages that exist independently, but it is really difficult to capture them from the perspective of weight.

Figure 5. MDS of GIVCN-SPFE-2014 models. Note: E_G denoted by the green line is the intersection of E_{SP} and E_{FE} , i.e., $E_G = E_{FE} \cap E_{SP}$; E_B denoted by the blue line is the edge set only belonging to E_{FE} but not E_{SP} , i.e., $E_B = E_{FE} - E_G$; E_R denoted by the red line, contrary to E_G , is the edge set only belonging to E_{SP} but not E_{FE} , i.e., $E_R = E_{SP} - E_G$.

Take the GIVCN-Eora-SC4-2014 model as an example, in the GT method, the minimum threshold is one ten-thousandth of the maximum edge weight, which is equal to $9.728\text{E}+05$, and the second smallest one is $1.946\text{E}+06$, and so on. Then, each threshold is selected in order from small to large to conduct network pruning. In other words, the parameters in the GT method can be regarded as specific removal rules. DF method computes the probability $P_{ij=(1-w_{ij}/s_i)^{k_i-1}}$ that such a value is compatible with the null model: if $p_{ij} < \alpha$, where α is a parameter representing the desired significance level, the edge is preserved, otherwise it is removed. Lower values of α lead to sparser networks, as fewer edges are preserved. Although an edge is connected to two nodes, its direction determines there exists only one value for p_{ij} , that is, based on the out-degree and out-strength of the source node.

Table 4 reports the structural similarity of different backbone networks in terms of the percentage of non-isolated nodes ($N_i\%$) and their Quadratic Assignment Procedure (QAP) correlation with the original network, where the QAP is an often-used approach to measure the extent to which two networks are correlated or have a similar pattern of connections (Dai, Derudder, & Liu, 2018). According to the results, the QAP correlations of the five algorithms (all P value < 0.001) are extremely high, that is, they can all maintain a high similarity with the original network in the overall structure level. Except for GT, other methods can better protect the bridges in the network and hardly make nodes isolated.

Table 4. Structural similarity of various backbones and original network.

Networks	$N_i\%$ no self loops					QAP				
	GT	DF	SP	FE	SPFE	GT	DF	SP	FE	SPFE
GIVCN-WIOD-SC4-2014	88.068	100.000	100.000	100.000	100.000	1.000	0.998	0.994	1.000	1.000
GIVCN-TiVA-SC4-2014	91.539	99.615	99.615	99.615	99.615	1.000	0.991	0.996	1.000	1.000
GIVCN-Eora-SC4-2014	65.079	100.000	100.000	100.000	100.000	1.000	1.000	0.999	1.000	1.000
GIVCN-ADB-SC4-2014	77.778	98.413	98.810	98.810	98.810	1.000	0.998	0.996	1.000	1.000

5.2 Preservation of functional information

The number of residual edges is recorded in the new network and the percentage of the residual edge weights for the original network is shown in Figure 6. Take the GIVCN-Eora-SC4-2014 model as an example, in the network extracted at the global threshold of $9.728\text{E}+05$ (one ten-thousandth of the maximum edge weight), the number of edges quickly decreases from 571,536 to 2,779 and the decline rate reaches 99.510%, while the total edge weight of the network decreases by only 3.370%. Obviously, setting thresholds is acceptable if following the principle of

Research Paper

“significantly reducing the number of edges while retaining most of the total edge weight”. The edge filtering procedure retains more edges than setting global thresholds, and the proportion of residual edge- weights of the former with $\alpha \geq 0.3$ or $\alpha \leq 0.1$ is higher or lower than the latter. The proportion of residual edge weights is only 87.758% while simplifying the GIVCN model by the SP. However, with the addition of FE method, GIVCN-SPFE can retain the weight to the maximum extent, almost reaching 99.154%.

Figure 6. Comparative result of network pruning of GIVCN models. Note: Since the pruning results of SPFE and FE method are almost the same in our models, the FE method is not shown in the figure.

Moreover, we propose the Average Edge Weight (AEW) to measure the weight information included by unit edge:

$$AEW = \frac{\sum_{(i,j) \in E} w_{(i,j)}}{|E|} \quad (4)$$

where $|E|$ is the number of edges in E and $w_{(i,j)}$ is the weight of edge (i,j) . Moreover, given that AEW_{np} represents the average edge weight of the backbones which extracted by the pruning methods and AEW_{rand} is the result obtained in a null model where the edges with the same proportion as skeleton networks are randomly

removed, then the ratio of them, $AEW_{ratio} = AEW_{np}/AEW_{rand}$, is of great significance to evaluate a specific method's efficiency. Deviations of AEW_{ratio} above or below 1 demonstrate positive or negative pruning results, respectively.

A thousand times comparisons between each method and null model are shown by boxplot graph in Figure 7. Although all of them precede their null models, the SP method as the best and the most consistent one manifests it catches a small number of critical paths but with extremely high edge weights. In conclusion, among the two parameterless methods, the global SP method can preserve the structural integrity of the network well, but the entire robustness and local function of network are inevitably influenced since it yields so few links. FE method can basically embrace SP and supplement more weight information, and thus has more significance in extracting the backbone of ICIO network.

Figure 7. The ratio of the average edge weight of the pruning methods and the null model.

6 Conclusions

Modeling and analyzing complex systems based on Complex Network theory has become a research hotspot. With the advent of the era of big data, the rapid expansion

Research Paper

of network scale is becoming a new challenge for network science research. Backbone extraction is an adequate tool for studying large network. It is used to detect significant edges and obtain a sparser network with only the crucial information. Different from the existing methods, both the heterogeneity of weights and denseness of ICIO network are considered in this paper. We strive to propose solutions that do not rely on artificial parameters and find a better one between the global and local methods.

The SP algorithm measures the strength of relations across or within sectors from the global perspective of IVC, retaining the links with closer IO relations and lower transaction costs. By combining H-index and Pareto rule, the FE algorithm performs the filtering process at the node-local level and converges naturally without depending on parameters. Through comparison between the two and existing methods, it is proved that the SP algorithm is already an efficient and effective tool, which can not only keep the structural similarity with the original network on the global level but also ensure that retained unit edge carries enough information. However, high-efficiency redundancy removal makes it tortuous to reach some marginal nodes inevitably. The FE algorithm covers the results of SP, while retaining higher weight on the premise of the number of remaining edges is close to other methods. Therefore, the FE algorithm can basically replace the mixed strategy and become the better of the two proposed methods.

In addition, the two special network pruning algorithms are useful to explore more possibilities for the study of highly dense and heterogeneously weighted networks, such as information networks, communication networks, citation networks, social media networks, etc. Also, except for visualization and network topology analysis, the obtained sparse network backbone will promote the relevant analyses in terms of community detection, link prediction, spatial econometrics, network embedding edge2eve, etc.

Acknowledgment

The author acknowledges support from National Natural Science Foundation of China (Grant No. 71971006), Humanities and Social Science Foundation of Ministry of Education of the People's Republic of China (Grant No. 19YJCGJW014).

Author contributions

Lizhi Xing (itwasa@163.com) initiated the idea, designed the research, and wrote the paper, Yu Han (17864192095@163.com) collected data, processed figures and tables and wrote the paper.

References

- Ahmed, N.K., Neville, J., & Kompella, R. (2014). Network sampling: From static to streaming graphs. *ACM Trans. Knowl. Discov. Data*, 8(2), 1–56.
- Blagus, N., Šubelj, L., & Bajec, M. (2012). Self-similar scaling of density in complex real-world networks. *Physica A*, 391(8), 2794–2802.
- Bu, Z., Wu, Z.A., Qian, L.Q., Cao, J., & Xu, G.D. (2014). A backbone extraction method with local search for complex weighted networks. In: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 85–88.
- Burt, R.S (1992). *Structural holes: The social structure of competition*. Harvard University Press.
- Cao, J., Ding, C.L., & Shi, B.Y. (2019). Motif-based functional backbone extraction of complex networks. *Physica A*, 526, 121123.
- Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y., & Shir, E. (2007). A model of internet topology using k-shell decomposition. In: *Proceedings of the National Academy of Sciences of the United States of America*, 104(27), 11150–11154.
- Chen, D.B., Lü, L.Y., Shang, M.S., Zhang, Y.C., & Zhou, T. (2012). Identifying influential nodes in complex networks. *Physica A*, 391(4), 1777–1787.
- Coscia, M., & Neffke, F. (2017). Network backboning with noisy data. In: 2017 IEEE 33rd International Conference on Data Engineering, 425–436.
- Dai, L., Derudder, B., & Liu, X. (2018). Transport network backbone extraction: A comparison of techniques. *Journal of Transport Geography*, 69, 271–281.
- Foti, N.J., Hughes, J.M., & Rockmore, D.N. (2011). Nonparametric sparsification of complex multiscale networks. *Plos One*, 6(2), e16431.
- Ghalmane, Z., Cherifi, C., Cherifi, H., & Hassouni, M. (2020). A backbone extraction method for complex weighted networks, MARAMI.
- Grady, D., Thiemann, C., & Brockmann, D. (2011). Robust classification of salient links in complex networks. *Nat Commun*, 3(1), 199–202.
- Granovetter, M. (1973). The Strength of Weak Ties. *American Journal of Sociology*, 78(6), 1360–1380.
- Guan, J., Ren, J.Q., & Xing, L.Z. (2021). Measuring the nestedness of global production system based on bipartite network. In: *The 9th International Conference on Complex Networks and Their Application*, 547–558.
- Hennemann, S. (2013). Information-rich visualisation of dense geographical networks. *Journal of Maps*, 9, 68–75.
- Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. In: *Proceedings of the National Academy of Sciences of the United States of America*, 102, 16569–16572.
- Itzkovitz, S., Levitt, R., Kashtan, N., Milo, R., Itzkovitz, M., & Alon, U. (2004). Coarse-graining and self-dissimilarity of complex networks. *Physical Review E*, 71, 016127.
- Kim, D.H., Noh, J.D., & Jeong, H. (2004). Scale-free trees: The skeletons of complex networks. *Physical Review E*, 70(4), 046126.
- Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., & Makse, H.A. (2010). Identification of influential spreaders in complex networks. *Nature Phys.*, 6(11), 888–893.
- Li, Q., Zhou, T., Lü, L.Y., & Chen, D.B. (2014). Identifying influential spreaders by weighted LeaderRank. *Physica A*, 404, 47–55.

Research Paper

- Lv, L.Y., Zhang, Y.C., Yeung, C.H., & Zhou, T. (2011). Leaders in social networks, the delicious case. *Plos One*, 6(6), e21202.
- Malang, K., Wang, S.H., Phaphuangwittayakul, A., Lü, Y.Y., Yuan, H.N., & Zhang, X.Z. (2020). Identifying influential nodes of global terrorism network: A comparison for skeleton network extraction. *Physica A*, 545, 123769.
- Marotta, L., Miccichè, S., Fujiwara, Y., Lyetomi, H., Aoyama, H., Gallegati, M., & Mantegna, R.N. (2015). Backbone of credit relationships in the Japanese credit market. *EPJ Data Sci.*, 5(1), 1–14.
- Nick, B., Lee, C., Cunningham, P., & Brandes, U. (2013). Simmelian backbones: Amplifying hidden homophily in facebook networks. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 525–532.
- Radicchi, F., Ramasco, J.J., & Fortunato, S. (2011). Information filtering in complex weighted networks. *Physical Review E*, 83(4), 046101–046101.
- Serrano, M.A., Boguna, M., & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. In: *Proceedings of the National Academy of Sciences of the United States of America*, 106, 6483–6488.
- Siganos, G., Tauro, S.L., & Faloutsos, M. (2006). Jellyfish: A conceptual model for the AS internet topology. *Journal of Communications and Networks*, 8(3), 339–350.
- Toivonen, H., Zhou, F., Hartikainen, A., & Hinkka, A. (2012). Network compression by node and edge mergers. In: M.R. Berthold (Eds.), *Bisociative Knowledge Discovery*, Springer, Berlin, Heidelberg, 199–217.
- Xing, L.Z., Dong, X.L., Guan, J., & Qiao, X.Y. (2019). Betweenness centrality for similarity-weight network and its application to measuring industrial sectors' pivotability on the global value chain. *Physica A*, 516, 19–36.
- Xing, L.Z., & Han, Y. (2021). Extracting the backbone of global value chain from high-dimensional inter-country input-output network. In: *The 9th International Conference on Complex Networks and Their Application*, 559–570.
- Zhang, R.J., Stanley, H.E., & Ye, F.Y. (2018). Extracting h-backbone as a core structure in weighted networks. *Sci Rep*, 8(1), 1–7.
- Zhang, X.H., Zhang, Z.C., Zhang, H., Wang, Q., & Zhu, J. (2014). Extracting the globally and locally adaptive backbone of complex networks. *Plos One*, 9(6), e100428.
- Zhang, X.H., & Zhu, J. (2013). Skeleton of weighted social network. *Physica A*, 392(6), 1547–1556.
- Zhao, S.X., Zhang, P.L., J, L., Tan, A.M., & Ye, F.Y. (2014). Abstracting the core subnet of weighted networks based on link strengths. *Journal of the Association for Information Science and Technology*, 65(5), 984–994.
- Zhou, Y., Cheng, H., & Yu, J.X. (2009). Graph clustering based on structural/attribute similarities. In: *Proceedings of the VLDB Endowment*, 718–729.

This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Appendix

Table 5. Illustration of properties of the GT backbone.

Database	$ E $	$\langle d \rangle$	$\langle K \rangle$	$\langle C \rangle$
WIOD	2,009	5.4774	6.9659	0.5291
TiVA	3,211	4.3517	8.5598	0.4798
Eora	2,779	4.1916	43.8587	0.4397
ADB	2,361	4.2087	8.0484	0.4858

Table 6. Illustration of properties of the DF backbone.

Database	$ E $	$\langle d \rangle$	$\langle K \rangle$	$\langle C \rangle$
WIOD	1,385	5.4774	6.9659	0.5291
TiVA	2,471	4.3517	8.5598	0.4798
Eora	33,289	4.1916	43.8587	0.4397
ADB	2,224	4.2087	8.0484	0.4858

Table 7. Illustration of properties of the SP backbone.

Database	$ E $	$\langle d \rangle$	$\langle K \rangle$	$\langle C \rangle$
WIOD	533	4.8292	2.3864	0.4569
TiVA	806	4.9457	2.4093	0.3788
Eora	2,496	5.1258	2.6481	0.5381
ADB	752	4.7738	2.3695	0.4463

Table 8. Illustration of properties of the FE backbone.

Database	$ E $	$\langle d \rangle$	$\langle K \rangle$	$\langle C \rangle$
WIOD	3,825	1.9275	20.8125	0.6231
TiVA	7,477	1.9451	27.8764	0.6258
Eora	54,747	1.9504	71.4762	0.5985
ADB	6,288	1.9675	24.3253	0.616

Table 9. Illustration of properties of the SPFE backbone.

Database	$ E $	$\langle d \rangle$	$\langle K \rangle$	$\langle C \rangle$
WIOD	3,825	1.9275	20.8125	0.6231
TiVA	7,477	1.9451	27.8764	0.6258
Eora	54,753	1.9485	71.4841	0.5993
ADB	6,288	1.9675	24.3253	0.616