Rifki *et al. Economic Structures (2017) 6:3*
DOI 10.1186/s40008-017-0062-2

Journal of Economic Structures

RESEARCH Open Access

The robustest clusters in the input–output networks: global CO_2 emission clusters

Omar Rifki*, Hirotaka Ono and Shigemi Kagawa

*Correspondence: omar.rifki@gmail.com
Faculty of Economics, Kyushu University, 6-19-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan

Abstract

Finding environmentally significant clusters in global supply-chain networks of goods and services has been investigated by Kagawa et al. (Soc Netw 35(3):423–438, 2013a; Econ Syst Res 25(3):265–286, 2013b; Glob Environ Chang, 2015), using the popular clustering method of nonnegative matrix factorization, which actually yields sensitive cluster assignments. Due to this sensitivity issue, there is a danger of overfitting of the results. In order to confirm the robustness of the obtained clusters, which in fact have strong implications for international climate change mitigation, especially for the US-induced Chinese clusters, we design a simulation-based experiment. Empirical findings of the proposed approach are compared with those of Kagawa et al. (Glob Environ Chang, 2015). The environmental implications are reported as well.

Keywords: Robustness, CO_2 emissions, Cluster analysis, Simulation

1 Background

Graph partitioning methods or clustering methods in general have been widely used for understanding and visualizing fundamental features of social and economic network complexity, e.g., Newman and Girvan (2004), Kagawa et al. (2013a, b), Liang et al. (2015). A striking environmental study has been provided by Kagawa et al. (2015); the authors identified CO_2 emission clusters within global supply-chain networks formed by the final demand impulse of a specific final product and argued how the identified emission clusters have contributed to increasing CO_2 emission transfers and have grown over time [see also Davis et al. (2011) and Peters et al. (2011) for the analysis of CO_2 emission transfers]. The authors applied the nonnegative matrix factorization (NMF) approach (Lee and Seung 1999) and obtained certain clusters whose normalized cut value (Ncut value) is minimized, which implies that the obtained clusters could best explain the environmentally important supply chains (Kagawa et al. 2015).

Although Kagawa et al. (2015) provided important emission clusters for climate change mitigation, there is a crucial problem that the obtained results highly depend on the employed algorithms and parameters. To see the problem, we show an example: Suppose that we apply a typical clustering algorithm such as the K-means method (MacQueen et al. 1967) for the analysis. By setting the parameter $K = 10$, we can obtain a set of 10 clusters, but if we instead set the parameter $K = 11$, the obtained 11 clusters could include very different sectors from the ones of $K = 10$. In such a situation, which

© The Author(s) 2017. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

"clusters" really reflect the actual economic structure? Or which "clusters" are plausible for the analysis? We have the same problem for not only the value of K but also the many other parameters used in the employed algorithms. It is worth noting that the K-means algorithm is indeed used in the NMF method.

The same problem is seen for the quality of the datasets. Economic network data such as input–output tables usually contain errors, or they always just constitute an approximation, which is a central issue in input–output analysis (e.g., Dietzenbacher 1995, 2006). Due to the errors, the same problem mentioned above appears in cluster analysis. That is, the clustering analysis could be very sensitive to the employed algorithms and datasets. In fact, the actual datasets used for constructing the supply-chain networks of Kagawa et al. (2015) are estimations derived from the multi-regional input–output framework (e.g., Lenzen et al. 2012; Dietzenbacher et al. 2013). If the employed clustering technique is quite sensitive to changes in the input–output data, which is our case, we need to be careful to claim that the resulted clusters are plausible.

This paper investigates this problem and proposes a method to obtain clusters that are "stable" with respect to errors or noise in the data and parameters of the algorithm. That is, even if we slightly perturb the values in datasets, clusters obtained by our method still have a good Ncut value; even though the original data may contain errors or noise, the obtained clusters are still reliable if the noise is small enough. The idea of our approach is rather simple and is based on simulations. It can be interpreted as applying a Monte Carlo-type simulation to obtain stable clusters in terms of perturbations by noise in data or choices of parameters. We also propose two criteria and a diagram based on the criteria to guarantee the robustness of the obtained clusters. The details will be described in Sect. 2. It should be noted that, due to its generality, this diagram could provide a new guideline to measure the reliability of analysis results, used in various fields where clustering analyses are applied, such as economic and social networks.

As a case study, we focused on an adjacency matrix obtained by using a multi-regional input–output analysis (Kagawa et al. 2015). The proposed two criteria were applied to obtain robust CO_2 emission clusters within global supply-chain networks. The robustness and performance of our clustering results are compared to those of Kagawa et al. (2015). We particularly evaluate the difference in terms of cluster compositions, which carry strong environmental implications. The remaining sections are as follows: Section 2 describes the methodology in this study, Sect. 3 provides a numerical example, Sect. 4 presents the obtained empirical results and discussions, and Sect. 5 concludes this paper.

2 Methods

2.1 Constructing an adjacency matrix

An economic transaction between geographically distributed industries is defined as $\mathbf{Z} = (Z_{ij}^{rs})$ $(i, j = 1, \ldots M; r, s = 1, \ldots, N)$, which represents a product sale from industry i in country r to industry j in country s. Here, M is the number of industries and N is the number of countries. If geographical input coefficients are defined by $\mathbf{A} = (a_{ij}^{rs})$ with $a_{ij}^{rs} = Z_{ij}^{rs}/x_j^s$, where x_j^s denotes domestic output of industry j in country s, the widely used interregional input–output (IRIO) model (e.g., Miller and Blair 2009) can be formulated as

$$x_i^r = \sum_{s=1}^{N}\sum_{j=1}^{M} a_{ij}^{rs} x_i^r + \sum_{s=1}^{N} f_i^{rs}, \tag{1}$$

or $\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{f}$ in matrix notation, where $\mathbf{x} = (x_i^r)$, $\mathbf{f} = (\sum_{s=1}^{N} f_i^{rs})$, and $f_i^{rs}(i = 1, \ldots, M; r, s = 1, \ldots, N)$ represents final demand from industry i of country r to country s.

Solving the IRIO model in Eq. (1) yields $\mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{f} = \mathbf{B}\mathbf{f}$. Here $\mathbf{I}$ is the identity matrix, and $\mathbf{B} = (\mathbf{I} - \mathbf{A})^{-1}$ is the direct and indirect requirement matrix, in which each element b_{ij}^{rs} represents how many units of the products of industry i in country r are needed to produce one unit of the products of industry j in country s.

Using the above IRIO framework, we can further formulate the unit structure model based on the IRIO model (e.g., Kagawa et al. 2015):

$$\mathbf{X}_j^s = \mathbf{A}\mathrm{diag}(\mathbf{b}_j^s f_j^s), \tag{2}$$

where $\mathbf{b}_j^s$ can be easily obtained as the $((s-1)M + j)$-th column vector in the direct and indirect requirement matrix $\mathbf{B}$, and f_j^s is the final demand of products produced by industry j in country s. The matrix $\mathbf{X}_j^s$ shows the economic transactions between geographically distributed industries that are triggered by the final demand on industry j located in country s.

If the direct emission coefficient vector showing CO_2 emissions generated per unit of output of industry i in country r is defined as $\boldsymbol{\alpha} = (\alpha_i^r)$, the CO_2 emissions embedded in the economic transactions are obtained as

$$\mathbf{G} = \left(g_{ij}^{rs}\right) = \mathrm{diag}(\boldsymbol{\alpha})\mathbf{X}_j^s, \tag{3}$$

where diag represents the diagonalization. Using this formulation, we define adjacency matrix $\mathbf{W} = \left(w_{ij}^{rs}\right)$, where

$$w_{ij}^{rs} = \begin{cases} 0 & i = j, r = s, \\ g_{ij}^{rs} + g_{ji}^{sr} & \text{otherwise.} \end{cases} \tag{4}$$

2.2 Clustering input–output analysis

A graph is a discrete structure that consists of vertices and edges that connect two vertices. In the context of economic analysis, a vertex and an edge correspond to a sector and a transaction between the corresponding two sectors, respectively. Graph clustering concerns finding for a graph similar vertices that can be arranged in dissimilar groups. This problem has multiple variants, algorithms, and applications; see Schaeffer (2007). Spectral and NMF-based clusterings have become popular in recent years, especially in the field of machine learning, e.g., Shi and Malik (2000), Ng et al. (2002), Kannan et al. (2004), Ding et al. (2005), Von Luxburg (2007), Zhang and Jordan (2008). However, spectral clustering can be traced back to the field of computer science for the graph partitioning problem, due to the work of Donath and Hoffman (1973) and Fiedler (1973).

Suppose an undirected weighted network $G = (V, E)$ of order $n = |V|$ with edge weights w_{uv}. In the context of clustering IO analysis (CIOA), a vertex u corresponds to a sector of an industry i in a country r. We denote this by $u = (i, r)$. Here, V and E, respectively, represent the set of all the sectors and the set of all the transactions between two sectors, and $|V| = n = M \times N$. The edge weights represent the amounts of CO_2 emissions associated with the corresponding transactions. It is also possible to consider unweighted graphs with zero-one edge weights. A central instrument of the spectral and NMF clustering framework is the use of *Laplacian matrices*, which are matrix representations of graphs. If $\mathbf{W} = (w_{uv})_{1 \leq u,v \leq n}$ is the *adjacency matrix* of the network G and $\mathbf{D} = \mathrm{diag}(\mathbf{d})$ is the diagonal *degree matrix* of G, with $\mathbf{d} = (d_u)_{1 \leq u \leq n} = (\sum_{v=1}^{n} w_{uv})_{1 \leq u \leq n}$ being the vector of vertices' degrees, then the Laplacian matrix of G can be defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$. The normalized version of $\mathbf{L}$ is given by $\mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$ (Shi and Malik 2000; Ng et al. 2002).

Such clustering is considered graph partitioning. A family of subsets $U_1, \ldots, U_k$ of set V is called a *(k-)partition* of V if $\bigcup_{p=1}^{k} U_p = V$ and $U_p \cap U_q = \emptyset$ for $1 \leq p, q \leq k, p \neq q$. A graph partition is a partition of vertices. The objective is to minimize for each cluster its total weights to the rest of the graph, which is called *cut* in graph theory and expressed as $\mathrm{cut}(U, \bar{U}) = \sum_{u \in U, v \in \bar{U}} w_{uv}$ for a subset $U \subset V$ of vertices and its complement $\bar{U}$. In this vein, Shi and Malik (2000) introduced the *normalized cut* criterion, abbreviated as Ncut, that produces when it is minimized clusters of reasonable sizes. For k partition $U_1, U_2, \ldots, U_k$ of V, the Ncut is defined as

$$\mathrm{Ncut}(U_1, U_2, \ldots, U_k) = \sum_{p=1}^{k} \frac{\mathrm{cut}(U_p, \bar{U_p})}{\sum_{u \in U_p} d_u},$$

where the denominators $\sum_{u \in U_p} d_u = \sum_{u \in U_p, v \in V} w_{uv}$ are implicitly implementing the objective of increasing the connectivity within clusters. Graph clustering into k clusters can be formulated as the following combinatorial problem:

$$\begin{array}{ll} \min\limits_{U_1, U_2, \ldots, U_k} & \mathrm{Ncut}(U_1, U_2, \ldots, U_k) \\ \text{subject to} & U_1 \cup U_2 \cup \cdots \cup U_k = V \text{ and } U_i \cap U_j = \emptyset \ (i \neq j). \end{array} \tag{5}$$

Unfortunately, this problem has been proven to be NP-hard (Shi and Malik 2000), unlike the abovementioned "min-cut" problem, which can be solved efficiently. However, problem (5) can be converted to matrix form as a minimization of the Rayleigh quotient (Shi and Malik 2000), which can be in its turn expressed as a trace matrix (Ding et al. 2005; Von Luxburg 2007; Zhang and Jordan 2008; Ding et al. 2008). For instance, using the notation of Ding et al. (2005), problem (5) becomes

$$\min_{\mathbf{H}} \mathrm{Tr}\left(\mathbf{H}^T \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}} \mathbf{H}\right) \text{ subject to } \mathbf{H}^T\mathbf{H} = \mathbf{I}_k, \tag{6}$$

where $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_k)$ is the $(n \times k)$ matrix defined by $\mathbf{h}_i = \mathbf{D}^{\frac{1}{2}}\mathbf{q}^{(i)}/||\mathbf{D}^{\frac{1}{2}}\mathbf{q}^{(i)}||$. Here, $\mathbf{H}^T$ is the transpose matrix of $\mathbf{H}$, and $\mathbf{q}^{(i)} = (q_1^{(i)} q_2^{(i)} \cdots q_n^{(i)})^T$ is the n-dimensional indicator vector of the cluster U_i; $q_u^{(i)} = 1$ if a vertex u belongs to cluster U_i, 0 otherwise. Note that $\mathbf{H}$ is a nonnegative matrix; that is, all the elements of $\mathbf{H}$ are nonnegative.

Problem (6) can be rewritten in an easily solvable form, specifically, accordingly to Ding et al. (2005),

$$\min_{\mathbf{H}\geq \mathbf{0}} \left\| \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}} - \mathbf{H}\mathbf{H}^T \right\|_F^2, \tag{7}$$

where $||.||_F$ is the Frobenius matrix norm. In conformity with the algorithm of Lee and Seung (2001), problem (7) can be solved using update rules, which are iterative improvements converging to local optima solutions. In this study, we make use the following rule given by Ding et al. (2005) and used by Kagawa et al. (2013a, 2015):

$$h_{ij} \leftarrow h_{ij}\left(1 - \beta + \beta \frac{(\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}})_{ij}}{(\mathbf{H}\mathbf{H}^T\mathbf{H})_{ij}}\right), \tag{8}$$

where β is a parameter such that $0 < \beta \leq 1$. We set $\beta = 0.5$ and initialize the matrix $\mathbf{H}$ according to the previous studies (Ding et al. 2008; Kagawa et al. 2013a, 2015). After iteratively applying (8), an approximated real-valued solution matrix $\hat{\mathbf{H}}$ is reached. Although we can expect to obtain $\hat{\mathbf{H}}$ in realistic running time, $\hat{\mathbf{H}}$ itself does not give any clustering due to the non-integer property. To obtain a concrete clustering, we apply the well-known and well-used K-means algorithm, which we call the *rounding step* in our algorithm and explain below.

The K-means algorithm introduced by MacQueen et al. (1967) is one of the most popular hierarchical clustering methods. This algorithm starts by selecting k initial clusters identified by their cluster centers and then iteratively refining them as follows. Given a target dataset $\{d_1, d_2, \ldots, d_n\}$ to be clustered, each iteration of the algorithm aims to minimize the within-cluster sum of squared distance, which is expressed as

$$\sum_{p=1}^{k} \sum_{i\in U_p} ||d_i - m_p||^2, \tag{9}$$

where m_p is the center of the cluster U_p, which is defined as the mean value of the elements belonging to U_p, i.e., $m_p = \sum_{i\in U_p} d_i/|U_p|$, and $||.||$ is a norm function on the dataset space. Minimizing this distance involves assigning each data instance d_i to its closest cluster, i.e., U_p such that the distance $||d_i - m_p||$ is minimum. The m_p centers are updated thereafter. The final clusters describe a partition of the dataset. This algorithm converges once no further assignment of instances is applied. The K-means can actually be proven to converge to a local minimum of expression (9) (Bottou and Bengio 1994). However, this algorithm suffers from several drawbacks, mainly its sensitivity to the initial conditions, which can lead to potentially misleading results (Bradley and Fayyad 1998). This issue is actually common among hill-climbing algorithms, where according to Duda et al. (1995): "different starting points can lead to different solutions and one never knows whether or not the best solution has been found." Thus, a bad choice of the initial cluster centers can easily converge to a poor cluster assignment. A second issue concerns the best value of the parameter k to be chosen. A bad choice here also can lead to poor results.

The basic steps of the NMF method presented in this section are summarized as the following algorithm.

NMF-based clustering algorithm

Input: A network $G = (V, E)$ of order $n = |V|$, an integer k number of clusters.
Output: A set of clusters $\{U_1, U_2, ..., U_k\}$.

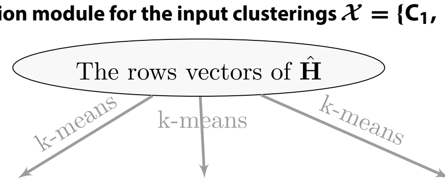
1: Initialize the matrix $\mathbf{H}$ in accordance with (Ding et al., 2008; Kagawa et al., 2015);
2: Use iteratively the update rule of (8);
3: **Rounding step:** Run the K-means algorithm to partition the n rows of $\hat{\mathbf{H}}$, which correspond to the n vertices of G, into k clusters $U_1, U_2, \ldots, U_k$;
4: Return $\{U_1, U_2, \ldots, U_k\}$;

2.3 Simulation module

This section analyzes the rounding step of the NMF algorithm described in the previous section. A sampling-based simulation procedure is performed in the rounding step, instead of running one instance of the K-means algorithm on the matrix $\hat{\mathbf{H}}$, as prescribed in the NMF algorithm. In order to simulate uncertain environments for the input clusterings, which will be introduced in Table 1, small perturbations in $\mathbf{W} = (w_{uv})_{1 \le u,v \le n}$, the adjacency matrix of the network G, are generated N times. Consequently, perturbed adjacency matrices can be obtained as $\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_N$, such that N denotes the number of the samples from now on. To assess the performance of a clustering $\mathrm{C} = \{U_1, U_2, \ldots, U_k\}$, obtained from the "initial" adjacency matrix $\mathbf{W}_0 = \mathbf{W}$, we propose the modified Ncut criterion as follows: $\mathrm{Ncut}(\mathrm{C}, \mathbf{W}_I) = \sum_{p=1}^{k} \left(\sum_{u \in U_p} \sum_{v \notin U_p} x_{uv} / \sum_{u \in U_p} \sum_{v \in V} x_{uv} \right)$. It should be noted that this modified Ncut value represents the goodness achieved in a network that includes a small amount of noise in the edge weights, i.e., the perturbed adjacency matrix $\mathbf{W}_I = (x_{uv})_{1 \le u,v \le n}$, under the given clustering C. Thus, under the same clustering C, this Ncut criterion is well distinguished for different perturbed matrices $\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_N$.

Table 1 illustrates the simulation module developed in this study. First, we endogenously determine the matrix $\hat{\mathbf{H}}$ by applying NMF to the initial adjacency matrix $\mathbf{W}_0$ and subsequently repeatedly apply the K-means algorithm to the n-rows of the matrix $\hat{\mathbf{H}}$, M times. The clusterings $\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_M$, termed input clusterings, are thus obtained. These clusterings are assigned differently due to the instability of repeatedly conducting K-means rounding. The perturbation scenario (I_I) implies that the matrix $\mathbf{W}_I$ is used for

Table 1 Description of the simulation module for the input clusterings $\mathcal{X} = \{\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_M\}$



	C_1	C_2	...	C_M
(I_0 scenario) $\mathbf{W}_0$	$\mathrm{Ncut}(\mathrm{C}_1, \mathbf{W}_0)$	$\mathrm{Ncut}(\mathrm{C}_2, \mathbf{W}_0)$	...	$\mathrm{Ncut}(\mathrm{C}_M, \mathbf{W}_0)$
(I_1 scenario) $\mathbf{W}_1$	$\mathrm{Ncut}(\mathrm{C}_1, \mathbf{W}_1)$	$\mathrm{Ncut}(\mathrm{C}_2, \mathbf{W}_1)$	...	$\mathrm{Ncut}(\mathrm{C}_M, \mathbf{W}_1)$
(I_2 scenario) $\mathbf{W}_2$	$\mathrm{Ncut}(\mathrm{C}_1, \mathbf{W}_2)$	$\mathrm{Ncut}(\mathrm{C}_2, \mathbf{W}_2)$	...	$\mathrm{Ncut}(\mathrm{C}_M, \mathbf{W}_2)$
...	...	...	...	...
(I_N scenario) $\mathbf{W}_N$	$\mathrm{Ncut}(\mathrm{C}_1, \mathbf{W}_N)$	$\mathrm{Ncut}(\mathrm{C}_2, \mathbf{W}_N)$	...	$\mathrm{Ncut}(\mathrm{C}_M, \mathbf{W}_N)$
$\widehat{R_1^{\mathcal{X}}}(\mathrm{C})$	-	-	...	-
$\widehat{R_2^{\mathcal{X}}}(\mathrm{C})$	-	-	...	-

the Ncut computation. We suppose that the perturbed matrices $\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_N$ take values from the following uncertainty set:

$$U = \{(x_{uv}) \in \mathbb{R}^{n \times n} : |x_{uv} - w_{uv}| \leq \xi \, |w_{uv}|, \ \ \forall u, v \in [|1, n|]\}, \tag{10}$$

where ξ is a noise magnitude parameter, which usually takes small values, e.g., 0.1. From the set U, deviations are symmetric around the values of the initial matrix $\mathbf{W}_0$, such that each element $(x)_{uv}$ of the randomly perturbed adjacency matrices is limited within the interval $\Big[(1-\xi)w_{uv},\ (1+\xi)w_{uv}\Big]$. In practice, to obtain the matrices $\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_N$, we draw N independent and identically distributed samples from the set U.

Two measures of robustness are introduced in the simulation module, $R_1^{\mathcal{X}}$ and $R_2^{\mathcal{X}}$, such that the set $\mathcal{X} = \{\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_M\}$ denotes the M input clusterings. The first measure, $R_1^{\mathcal{X}}$, reports for a given clustering $\mathrm{C} \in \mathcal{X}$ the fractional degree that it is the best clustering within $\mathcal{X}$ across the perturbation scenarios. For instance, C being the best clustering within $\mathcal{X}$ for the perturbation scenario (I_I) means that C yields the smallest Ncut value among the Ncut values computed using the M clusterings of $\mathcal{X}$ and the perturbed adjacency matrix $\mathbf{W}_I$, i.e., $\mathrm{C} \in \arg\min_{\mathrm{D} \in \mathcal{X}} \mathrm{Ncut}(\mathrm{D}, \mathbf{W}_I)$. We introduce the indicator function $I_{\mathcal{X}}(\mathrm{C}, \mathbf{W}_I)$ that takes value one if C is the best clustering within $\mathcal{X}$ for perturbed matrix $\mathbf{W}_I$, zero otherwise. The indicator function is expressed as follows:

$$I_{\mathcal{X}}(\mathrm{C}, \mathbf{W}_I) = \begin{cases} 1, & \text{if } \mathrm{C} \in \arg\min\limits_{\mathrm{D} \in \mathcal{X}} \mathrm{Ncut}(\mathrm{D}, \mathbf{W}_I) \\ 0, & \text{otherwise} \end{cases}. \tag{11}$$

We can formulate the probability of appearance or being best, i.e., robustness measure, for a given clustering C over the initial and the perturbed N matrices as $\widehat{R_1^{\mathcal{X}}}(\mathrm{C})$ as follows:

$$\widehat{R_1^{\mathcal{X}}}(\mathrm{C}) = \frac{1}{N+1} \sum_{I=0}^{N} I_{\mathcal{X}}(\mathrm{C}, \mathbf{W}_I). \tag{12}$$

This expression is obviously dependent on the number of samples N. Thus, the measure expressed by (12) can be viewed as an estimation of the accurate measure $R_1^{\mathcal{X}}$. We use the caret symbol ($\widehat{}$) to indicate this approximation. The accurate measure $R_1^{\mathcal{X}}(\mathrm{C})$ is assumed to be reached for an infinite number of samples: $\widehat{R_1^{\mathcal{X}}}(\mathrm{C}) \xrightarrow[N \to \infty]{} R_1^{\mathcal{X}}(\mathrm{C})$.

The second measure, $R_2^{\mathcal{X}}(\mathrm{C})$, reports the average value of the *performance ratio*. We define the performance ratio of a clustering $\mathrm{C} \in \mathcal{X}$ under the perturbed adjacency matrix $\mathbf{W}_I$ as the ratio of the Ncut value of C to the smallest Ncut value computed for the perturbed matrix $\mathbf{W}_I$. The ratio expression for a given clustering C and a given matrix $\mathbf{W}_I$ is given by

$$\frac{\mathrm{Ncut}(\mathrm{C}, \mathbf{W}_I)}{\min_{\mathrm{D} \in \mathcal{X}} \{\mathrm{Ncut}(\mathrm{D}, \mathbf{W}_I)\}}. \tag{13}$$

It should be noted that if the given clustering C yields the smallest Ncut value under the perturbed adjacency matrix $\mathbf{W}_I$, the performance ratio takes value one. Using the performance ratios of a given clustering C, we can formulate the average of performance ratios, i.e., robustness measure, for C over the initial and the perturbed N matrices as $\widehat{R_2^{\mathcal{X}}}(\mathrm{C})$ as follows:

$$\widehat{R_2^{\mathcal{X}}}(\mathrm{C}) = 1/\Big(\frac{1}{N+1}\sum_{i=0}^{N}\frac{\mathrm{Ncut}(\mathrm{C},\mathbf{W}_I)}{\min_{\mathrm{D}\in\mathcal{X}}\{\mathrm{Ncut}(\mathrm{D},\mathbf{W}_I)\}}\Big). \tag{14}$$

Here, $\widehat{R_2^{\mathcal{X}}}(\mathrm{C})$ ranges within the interval between 0 and 1. If $\widehat{R_2^{\mathcal{X}}}(\mathrm{C})$ takes value one, the given clustering C always gives the smallest Ncut value under the initial and the perturbed adjacency matrices. Any lower value of $\widehat{R_2^{\mathcal{X}}}(\mathrm{C})$ implies that the given clustering C does not yield a better graph partition. The measure $\widehat{R_1^{\mathcal{X}}}(\mathrm{C})$ is always between 0 and 1. Both measures can be then seen as percentages. Note that the caret symbol is used also for measure $R_2^{\mathcal{X}}$, due to the dependence on the number of samples N. The more a clustering C is robust within $\mathcal{X}$, the higher are the estimated values of $R_1^{\mathcal{X}}(\mathrm{C})$ and $R_2^{\mathcal{X}}(\mathrm{C})$. To see why both measures are useful, we provide a numerical example with a focus with simplified network data in the next section.

3 Numerical example

We use the following simplified adjacency matrix:

$$\mathbf{W}_0 = \begin{bmatrix}
0.000 & 10.466 & 0.249 & 0.328 & 0.264 & 4.198 & 0.249 & 0.318 & 0.127 & 4.091 & 0.176 & 0.228 & 0.417 & 0.189 & 0.267 & 0.461 \\
10.470 & 0.000 & 10.840 & 10.950 & 15.000 & 12.750 & 10.980 & 13.260 & 11.590 & 11.720 & 10.480 & 14.010 & 13.750 & 11.720 & 11.640 & 13.270 \\
0.249 & 10.836 & 0.000 & 0.355 & 0.157 & 0.239 & 0.304 & 0.121 & 0.166 & 0.175 & 3.039 & 0.301 & 0.256 & 4.094 & 4.218 & 0.197 \\
0.328 & 10.947 & 0.355 & 0.000 & 0.336 & 0.406 & 2.353 & 0.335 & 2.555 & 0.210 & 0.189 & 0.375 & 3.893 & 0.176 & 0.159 & 3.215 \\
0.264 & 15.002 & 0.157 & 0.336 & 0.000 & 0.082 & 0.407 & 2.983 & 0.104 & 0.441 & 0.348 & 4.132 & 0.332 & 0.258 & 0.330 & 0.182 \\
4.198 & 12.746 & 0.239 & 0.406 & 0.082 & 0.000 & 0.254 & 0.320 & 0.251 & 5.245 & 0.537 & 0.244 & 0.428 & 0.356 & 0.558 & 0.394 \\
0.249 & 10.978 & 0.304 & 2.353 & 0.407 & 0.254 & 0.000 & 0.325 & 2.069 & 0.094 & 0.130 & 0.155 & 3.185 & 0.103 & 0.296 & 3.811 \\
0.318 & 13.261 & 0.121 & 0.335 & 2.983 & 0.320 & 0.325 & 0.000 & 0.236 & 0.465 & 0.240 & 3.168 & 0.363 & 0.233 & 0.200 & 0.338 \\
0.127 & 11.585 & 0.166 & 2.555 & 0.104 & 0.251 & 2.069 & 0.236 & 0.000 & 0.283 & 0.167 & 0.098 & 3.185 & 0.131 & 0.184 & 3.429 \\
4.091 & 11.716 & 0.175 & 0.210 & 0.441 & 5.245 & 0.094 & 0.465 & 0.283 & 0.000 & 0.082 & 0.222 & 0.288 & 0.271 & 0.391 & 0.444 \\
0.176 & 10.476 & 3.039 & 0.189 & 0.348 & 0.537 & 0.130 & 0.240 & 0.167 & 0.082 & 0.000 & 0.127 & 0.285 & 3.630 & 3.658 & 0.070 \\
0.228 & 14.006 & 0.301 & 0.375 & 4.132 & 0.244 & 0.155 & 3.168 & 0.098 & 0.222 & 0.127 & 0.000 & 0.107 & 0.152 & 0.239 & 0.310 \\
0.417 & 13.751 & 0.256 & 3.893 & 0.332 & 0.428 & 3.185 & 0.363 & 3.185 & 0.288 & 0.285 & 0.107 & 0.000 & 0.251 & 0.328 & 3.823 \\
0.189 & 11.716 & 4.094 & 0.176 & 0.258 & 0.356 & 0.103 & 0.233 & 0.131 & 0.271 & 3.630 & 0.152 & 0.251 & 0.000 & 3.692 & 0.443 \\
0.267 & 11.640 & 4.218 & 0.159 & 0.330 & 0.558 & 0.296 & 0.200 & 0.184 & 0.391 & 3.658 & 0.239 & 0.328 & 3.692 & 0.000 & 0.219 \\
0.461 & 13.270 & 0.197 & 3.215 & 0.182 & 0.394 & 3.811 & 0.338 & 3.429 & 0.444 & 0.070 & 0.310 & 3.823 & 0.443 & 0.219 & 0.000
\end{bmatrix}$$

Using this initial adjacency matrix, $\mathbf{W}_0$, we can depict a network with 16 vertices. A problem is how to find the robust clusterings from the network data. Following Ding et al. (2005) and Kagawa et al. (2013a, b), the clustering method based on the NMF of the normalized adjacency matrix is useful for achieving this goal. We apply the NMF method to the normalized adjacency matrix $\mathbf{D}^{-1/2}\mathbf{W}_0\mathbf{D}^{-1/2}$ as follows:

$$
\begin{aligned}
\mathbf{D}^{-1/2}\,\mathbf{W}_0\,\mathbf{D}^{-1/2} &= \begin{bmatrix}
0.000 & 0.165 & 0.011 & 0.014 & 0.011 & 0.175 & 0.011 & 0.014 & 0.005 & 0.176 & 0.008 & 0.010 & 0.016 & 0.008 & 0.011 & 0.018 \\
0.165 & 0.000 & 0.161 & 0.159 & 0.221 & 0.184 & 0.164 & 0.205 & 0.173 & 0.176 & 0.161 & 0.212 & 0.183 & 0.171 & 0.168 & 0.178 \\
0.011 & 0.161 & 0.000 & 0.014 & 0.006 & 0.009 & 0.012 & 0.005 & 0.007 & 0.007 & 0.127 & 0.012 & 0.009 & 0.162 & 0.165 & 0.007 \\
0.014 & 0.159 & 0.014 & 0.000 & 0.013 & 0.016 & 0.093 & 0.014 & 0.101 & 0.008 & 0.008 & 0.015 & 0.138 & 0.007 & 0.006 & 0.114 \\
0.011 & 0.221 & 0.006 & 0.013 & 0.000 & 0.003 & 0.016 & 0.124 & 0.004 & 0.018 & 0.014 & 0.168 & 0.012 & 0.010 & 0.013 & 0.007 \\
0.175 & 0.184 & 0.009 & 0.016 & 0.010 & 0.000 & 0.010 & 0.013 & 0.010 & 0.207 & 0.022 & 0.010 & 0.015 & 0.014 & 0.021 & 0.014 \\
0.011 & 0.164 & 0.012 & 0.093 & 0.013 & 0.010 & 0.000 & 0.014 & 0.084 & 0.004 & 0.005 & 0.006 & 0.115 & 0.004 & 0.012 & 0.139 \\
0.014 & 0.205 & 0.005 & 0.014 & 0.010 & 0.013 & 0.014 & 0.000 & 0.010 & 0.020 & 0.010 & 0.135 & 0.014 & 0.004 & 0.008 & 0.013 \\
0.005 & 0.173 & 0.007 & 0.101 & 0.207 & 0.010 & 0.084 & 0.010 & 0.000 & 0.012 & 0.007 & 0.004 & 0.116 & 0.010 & 0.007 & 0.125 \\
0.176 & 0.176 & 0.007 & 0.008 & 0.022 & 0.207 & 0.004 & 0.020 & 0.012 & 0.000 & 0.003 & 0.009 & 0.010 & 0.005 & 0.015 & 0.016 \\
0.008 & 0.161 & 0.127 & 0.008 & 0.010 & 0.022 & 0.005 & 0.010 & 0.007 & 0.003 & 0.000 & 0.005 & 0.011 & 0.011 & 0.148 & 0.003 \\
0.010 & 0.212 & 0.012 & 0.015 & 0.168 & 0.010 & 0.006 & 0.135 & 0.004 & 0.009 & 0.005 & 0.000 & 0.004 & 0.149 & 0.010 & 0.011 \\
0.016 & 0.183 & 0.009 & 0.138 & 0.012 & 0.015 & 0.115 & 0.014 & 0.116 & 0.010 & 0.011 & 0.004 & 0.000 & 0.006 & 0.011 & 0.124 \\
0.008 & 1.171 & 0.162 & 0.007 & 0.010 & 0.014 & 0.004 & 0.010 & 0.005 & 0.011 & 0.149 & 0.006 & 0.009 & 0.000 & 0.142 & 0.016 \\
0.011 & 0.168 & 0.165 & 0.006 & 0.013 & 0.021 & 0.012 & 0.008 & 0.007 & 0.015 & 0.148 & 0.010 & 0.011 & 0.142 & 0.000 & 0.008 \\
0.018 & 0.178 & 0.007 & 0.114 & 0.007 & 0.014 & 0.139 & 1.013 & 0.125 & 0.016 & 0.003 & 0.11 & 0.124 & 0.016 & 0.008 & 0.000
\end{bmatrix} \\
&\approx \begin{bmatrix}
0.041 & 0.000 & 0.321 \\
0.350 & 0.308 & 0.327 \\
0.001 & 0.345 & 0.028 \\
0.301 & 0.038 & 0.004 \\
0.110 & 0.088 & 0.173 \\
0.041 & 0.015 & 0.340 \\
0.296 & 0.040 & 0.000 \\
0.109 & 0.076 & 0.168 \\
0.296 & 0.038 & 0.000 \\
0.037 & 0.000 & 0.345 \\
0.000 & 0.328 & 0.036 \\
0.106 & 0.083 & 0.173 \\
0.326 & 0.045 & 0.004 \\
0.001 & 0.346 & 0.035 \\
0.000 & 0.344 & 0.043 \\
0.329 & 0.039 & 0.006
\end{bmatrix} \\
&\times \begin{bmatrix}
0.041 & 0.350 & 0.001 & 0.301 & 0.110 & 0.041 & 0.296 & 0.109 & 0.296 & 0.037 & 0.000 & 0.106 & 0.326 & 0.001 & 0.000 & 0.329 \\
0.000 & 0.308 & 0.345 & 0.038 & 0.088 & 0.015 & 0.040 & 0.076 & 0.038 & 0.000 & 0.328 & 0.083 & 0.045 & 0.346 & 0.344 & 0.039 \\
0.321 & 0.327 & 0.028 & 0.004 & 0.173 & 0.340 & 0.000 & 0.168 & 0.000 & 0.345 & 0.036 & 0.173 & 0.004 & 0.035 & 0.043 & 0.006
\end{bmatrix} \\
&= \hat{\mathbf{H}}\hat{\mathbf{H}}'
\end{aligned}
$$

Here, it should be noted that the matrix size of $\hat{\mathbf{H}}$ is 16 by 3. From this matrix, we obtain the following 16 feature vectors corresponding to $\hat{\mathbf{H}}$ row vectors:

$$
\begin{aligned}
&\mathbf{y_1} = (0.041, 0.000, 0.321);\ \mathbf{y_2} = (0.350, 0.308, 0.327);\ \mathbf{y_3} = (0.001, 0.345, 0.028); \\
&\mathbf{y_4} = (0.301, 0.038, 0.004);\ \mathbf{y_5} = (0.110, 0.088, 0.173);\ \mathbf{y_6} = (0.041, 0.015, 0.340); \\
&\mathbf{y_7} = (0.296, 0.040, 0.000);\ \mathbf{y_8} = (0.109, 0.076, 0.168);\ \mathbf{y_9} = (0.296, 0.038, 0.000); \\
&\mathbf{y_{10}} = (0.037, 0.000, 0.345);\ \mathbf{y_{11}} = (0.000, 0.328, 0.036);\ \mathbf{y_{12}} = (0.106, 0.083, 0.173); \\
&\mathbf{y_{13}} = (0.326, 0.045, 0.004);\ \mathbf{y_{14}} = (0.001, 0.346, 0.035);\ \mathbf{y_{15}} = (0.000, 0.344, 0.043); \\
&\mathbf{y_{16}} = (0.329, 0.039, 0.006)
\end{aligned}
$$

We applied the K-means method using these 16 feature vectors ten times and obtained the ten clusterings listed in Table 2.

We notice in the case of Table 2 that the clusterings C_1 and C_5 describe the same cluster assignment, and that this is also true for C_6 and C_{10}. Our interest is which clustering among the eight different cluster assignments gives the best graph partition. If we examine the Ncut values for the initial adjacency matrix $\mathbf{W}_0$, the best performance within our

Table 2 Clusterings resulting from applying 10 *K*-means instances to the feature vectors, $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{16}$, associated with the simplified network example

Clustering	Cluster	Industry 1	Industry 2	Industry 3	Industry 4	Industry 5	Industry 6	Industry 7	Industry 8	Industry 9	Industry 10	Industry 11	Industry 12	Industry 13	Industry 14	Industry 15	Industry 16	Total
C_1	U_1		●	●								●			●	●		5
	U_2				●			●		●				●			●	5
	U_3	●				●	●		●		●		●					6
C_2	U_1	●	●		●	●	●	●	●	●	●		●	●			●	12
	U_2			●											●			2
	U_3											●				●		2
C_3	U_1	●	●		●	●	●	●	●	●	●		●	●			●	12
	U_2			●											●	●		3
	U_3											●						1
C_4	U_1		●	●		●			●			●	●		●	●		8
	U_2				●			●		●				●			●	5
	U_3	●					●				●							3
C_5	U_1		●	●								●			●	●		5
	U_2				●			●		●				●			●	5
	U_3	●				●	●		●		●		●					6
C_6	U_1		●		●			●		●				●			●	6
	U_2	●				●	●		●		●		●					6
	U_3			●								●			●	●		4
C_7	U_1		●		●	●		●	●	●			●	●			●	9
	U_2			●								●			●	●		4
	U_3	●					●				●							3
C_8	U_1	●		●		●	●		●		●	●	●		●	●		10
	U_2				●			●		●				●			●	5
	U_3		●															1
C_9	U_1	●	●			●	●		●		●		●					7
	U_2				●			●		●				●			●	5
	U_3			●								●			●	●		4
C_{10}	U_1		●		●			●		●				●			●	6
	U_2	●				●	●		●		●		●					6
	U_3			●								●			●	●		4

input clusterings $\mathcal{X} = \{C_1, \ldots, C_{10}\}$ is exhibited by clustering C_3 according to the following table:

Clustering	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
Ncut(C, $\mathbf{W}_0$)	1.611	0.556	0.551	1.09	1.611	1.412	0.889	1.791	1.225	1.412

In order to draw the perturbed adjacency matrices, we first set the noise magnitude parameter, e.g., $\xi = 0.5$. We have sampled the following perturbed adjacency matrix $\mathbf{W}_I$ from the set U given in expression (10):

$$
\mathbf{W}_I = \begin{bmatrix}
0.001 & 10.789 & 0.271 & 0.381 & 0.368 & 3.085 & 0.145 & 0.388 & 0.140 & 3.397 & 0.174 & 0.327 & 0.521 & 0.179 & 0.205 & 0.534 \\
6.386 & 0.001 & 16.251 & 14.012 & 9.982 & 11.930 & 14.839 & 16.849 & 16.659 & 16.750 & 12.510 & 10.320 & 19.125 & 9.537 & 6.226 & 10.406 \\
0.300 & 6.503 & 0.001 & 0.303 & 0.125 & 0.237 & 0.183 & 0.124 & 0.190 & 0.136 & 3.210 & 0.386 & 0.306 & 3.120 & 2.384 & 0.139 \\
0.459 & 13.928 & 0.483 & 0.001 & 0.389 & 0.383 & 1.414 & 0.238 & 2.283 & 0.114 & 0.127 & 0.350 & 5.541 & 0.173 & 0.096 & 4.308 \\
0.296 & 11.070 & 0.133 & 0.388 & 0.001 & 0.061 & 0.308 & 4.271 & 0.145 & 0.339 & 0.270 & 3.791 & 0.318 & 0.148 & 0.438 & 0.249 \\
5.599 & 11.842 & 0.194 & 0.372 & 0.051 & 0.001 & 0.319 & 0.423 & 0.285 & 5.369 & 0.442 & 0.227 & 0.401 & 0.330 & 0.335 & 0.513 \\
0.335 & 5.908 & 0.261 & 1.897 & 0.382 & 0.374 & 0.001 & 0.210 & 2.791 & 0.063 & 0.081 & 0.098 & 1.888 & 0.083 & 0.368 & 3.114 \\
0.430 & 10.070 & 0.081 & 0.290 & 3.450 & 0.268 & 0.170 & 0.001 & 0.286 & 0.447 & 0.297 & 2.776 & 0.414 & 0.190 & 0.109 & 0.316 \\
0.065 & 12.180 & 0.237 & 1.647 & 0.069 & 0.126 & 2.406 & 0.162 & 0.001 & 0.406 & 0.147 & 0.115 & 2.460 & 0.074 & 0.245 & 4.841 \\
4.261 & 11.207 & 0.155 & 0.176 & 0.374 & 3.433 & 0.110 & 0.468 & 0.214 & 0.001 & 0.072 & 0.260 & 0.263 & 0.321 & 0.228 & 0.250 \\
0.118 & 11.475 & 4.469 & 0.191 & 0.382 & 0.582 & 0.067 & 0.333 & 0.104 & 0.067 & 0.001 & 0.145 & 0.207 & 1.976 & 2.955 & 0.076 \\
0.188 & 8.264 & 0.251 & 0.215 & 2.442 & 0.329 & 0.205 & 3.806 & 0.092 & 0.253 & 0.167 & 0.001 & 0.157 & 0.191 & 0.333 & 0.273 \\
0.242 & 11.988 & 0.378 & 4.734 & 0.215 & 0.271 & 4.094 & 0.263 & 4.358 & 0.273 & 0.318 & 0.152 & 0.001 & 0.249 & 0.347 & 5.340 \\
0.146 & 10.752 & 5.464 & 0.186 & 0.339 & 0.322 & 0.143 & 0.226 & 0.173 & 0.235 & 3.868 & 0.105 & 0.131 & 0.001 & 3.320 & 0.319 \\
0.217 & 6.439 & 4.882 & 0.212 & 0.454 & 0.682 & 0.396 & 0.221 & 0.248 & 0.244 & 4.061 & 0.127 & 0.204 & 2.069 & 0.001 & 0.202 \\
0.303 & 7.390 & 0.211 & 4.056 & 0.164 & 0.432 & 4.447 & 0.406 & 3.398 & 0.532 & 0.066 & 0.352 & 2.882 & 0.617 & 0.293 & 0.000
\end{bmatrix}.
$$

For this specific perturbed adjacency matrix, clustering C_2 is deemed to be the best clustering according to the Ncut results, which are shown in the following table:

Clustering	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
$\mathrm{Ncut}(C, \mathbf{W}_I)$	1.615	0.550	0.564	1.170	1.615	1.334	0.909	1.738	1.315	1.334

We could thus confirm that random noise in the edges of this simplified network clearly affects the choice of the best clustering. Next, we consider which among

C_2 and C_3 is a more reliable clustering. To answer this question, we proceed by examining the values of the indicator function and the performance ratio, which are given, respectively, in expressions (11) and (13). Our proposed robustness measures are entirely based on these values. For the current perturbed matrix $\mathbf{W}_I$, the values of the indicator function and the performance ratio are shown in the following table:

Clustering	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
$I_{\mathcal{X}}(C, \mathbf{W}_I)$	0	1	0	0	0	0	0	0	0	0
Performance Ratio	2.933	1	1.024	2.124	2.933	2.422	1.651	3.156	2.387	2.422

The indicator function tacked on clusterings can be intuitively derived. It merely indicates through a 0–1 binary representation which clusterings yield the best performance for the current perturbation scenario. In the current matrix $\mathbf{W}_I$, clustering C_2 obviously takes value one, while the other cluster assignments take value zero. For the performance ratio, the emphasis is instead put on the relative span to the best clusterings; a smaller value of the ratio indicates a better clustering. The clusterings C_3 and C_7 seem to be better in this regard for the perturbed matrix $\mathbf{W}_I$.

Our proposed simulation module iteratively draws a perturbed matrix $\mathbf{W}_I$ from the set U and computes the values of the indicator function and the performance ratio for the input clusterings and the matrix $\mathbf{W}_I$. The robustness measures $\widehat{R_1^{\mathcal{X}}}$ and $\widehat{R_2^{\mathcal{X}}}$ as explained in the previous section are averages of these iteratively computed values. The following table shows the results we obtained for the current simplified network when the sample size is set to $N = 100$ perturbation scenarios:

Clustering	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
$\widehat{R_1^{\mathcal{X}}}(C)$	0	0.247	0.752	0	0	0	0	0	0	0
$\widehat{R_2^{\mathcal{X}}}(C)$	0.333	0.985	0.996	0.492	0.333	0.384	0.611	0.304	0.443	0.384

According to both robustness measures, clustering C_3 performs better throughout the perturbation scenarios. However, C_3 is closely followed by clustering C_2 according to measure $\widehat{R_2^{\mathcal{X}}}$, even if in terms of measure $\widehat{R_1^{\mathcal{X}}}$ clustering C_2 appears to be the best for only 24.7% of the randomly perturbed adjacency matrices.

4 Empirical results

In this section, we introduce the experiment scheme that we designed in order to obtain clusterings that not only are robust as shown in the simulation module of Sect. 2.3 but also have a good enough normalized cut performance. Subsequently, the empirical results of Kagawa et al. (2015) are discussed and compared to those of the experiment run on the same datasets.

4.1 Experiment scheme

Robustness and performance are conflicting objectives. In order to not sacrifice the performance while finding robust clusterings, we incorporate the simulation module in the bigger experiment scheme shown in Fig. 1. Prior to starting the experiment, we apply the NMF method to the initial adjacency matrix in order to obtain the approximate matrix $\hat{\mathbf{H}}$, which has been described in Sect. 2.2. Our experiment is iterative. It starts by initializing the two clusterings C_{R_1} and C_{R_2} that are the output of the experiment, which are the overall best clusterings according to respective measures $R_1^{\mathcal{X}}$ and $R_2^{\mathcal{X}}$. Clusterings $\mathrm{C}_{R_1}^{\text{simul}}$ and $\mathrm{C}_{R_2}^{\text{simul}}$ are similarly top clusterings, but only for one simulation run. Clusterings C_{R_1} and C_{R_2} are initially set by random initializations of the K-means algorithm applied to matrix $\hat{\mathbf{H}}$. The experiment continues by generating M clusterings, using same initialization process as for C_{R_1} and C_{R_2}, to construct the set of input clusterings $\mathcal{X} = \{\mathrm{C}_{R_1}, \mathrm{C}_{R_2}, \mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_M\}$. Afterward, the simulation module of Table 1 is performed on the set $\mathcal{X}$. The best resulting clusterings $\mathrm{C}_{R_1}^{\text{simul}}$ and $\mathrm{C}_{R_2}^{\text{simul}}$ thereafter become C_{R_1} and C_{R_2}, respectively. The set $\mathcal{X}$ and the simulation are iteratively computed. Convergence is reached when the same $\mathrm{C}_{R_1}^{\text{simul}}$ and $\mathrm{C}_{R_2}^{\text{simul}}$ are obtained for a number of consecutive iterations. In our study, we choose a minimum of five iterations. Iterative schemes such as ours are widely used for practical solving of optimization problems. The basic idea is to continuously refine the approximate solutions. For instance, the various metaheuristic solvers such as genetic algorithms, simulated annealing, and iterated local search could be mentioned (Gendreau and Potvin 2010).

4.2 Results and discussion

Kagawa et al. (2013a, b) have been among the first to apply clustering methods to environmental analysis of economic systems. Their approach connects input–output models to techniques of network partition. As mentioned in introduction, highly important clusters in terms of CO_2 emissions have been identified by Kagawa et al. (2015). While

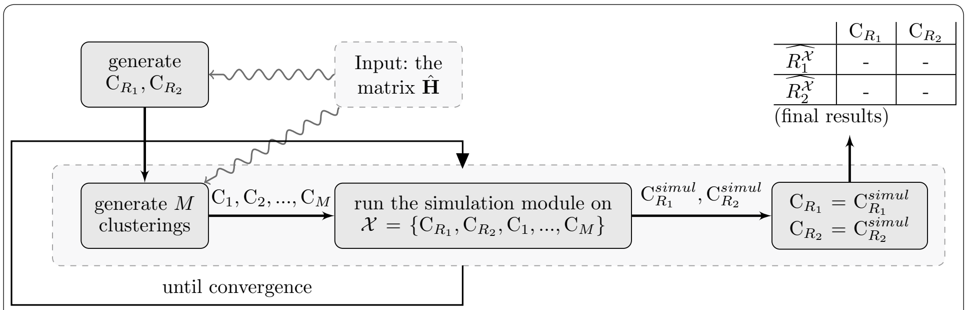


Fig. 1 Flowchart of the experiment scheme

China in 2009 was the largest emitter of CO_2 production-based emissions (Kagawa et al. 2015), the authors' goal was then to identify which country contributed most to these emissions and through which supply chains. To quantify CO_2 intensity for a cluster, the *within-cluster sum* is used. This is defined for a cluster U_p of a supply-chain network with the adjacency matrix $\mathbf{W} = (w_{uv})_{1 \leq u,v \leq n}$ as the summation $\sum_{u \in U_p} \sum_{v \in U_p} w_{uv}$.

In Kagawa et al. (2015), it was found among the 4756 industry clusters induced by the final demand of various good and services in the five developed countries of the USA, the UK, Germany, France, and Japan that both the US construction industry and the US transport equipment industry generate prominent Chinese clusters that are among the clusters with the 15th highest within-cluster sums. These two Chinese clusters have nevertheless the highest annual growth rates (also within the top 15), equal to 57.5 and 41.7% for US construction and transport equipment demands, respectively. In the top 15 clusters, there is only one other Chinese cluster, that one induced by the Japanese construction demand, but this cluster has a lower growth rate and a smaller within-cluster sum compared to both of the abovementioned US-induced Chinese clusters.

Therefore, for data, we consider the two adjacency matrices of CO_2 emissions induced by US construction and transport equipment demands for the year 2009, which were also used by Kagawa et al. (2015). We shall refer to them as the US construction and US transport datasets. Each adjacency matrix characterizes a network in which vertices are specified by a combination of a country plus an industry category (country–industry), with a total of 41 countries and 35 industries. The considered categories of countries and industries are listed in the supporting materials.

To detect the appropriate number of clusters K to be used in the experiment, we rely on the modularity index similarly as done by Kagawa et al. (2013a, 2015). This index, which has been developed by Newman and Girvan (2004), is optimal (maximized) for the correct number of clusters. The modularity index can be formulated for a network $G = (V, E)$ of an adjacency matrix $\mathbf{W} = (w_{uv})_{1 \leq u,v \leq n}$ as

$$Q(K) = \sum_{k=1}^{K} (p_{kk} - q_k^2),$$

where $p_{kk} = \left(\sum_{u \in U_k} \sum_{v \in U_k} w_{uv} / \sum_{u \in V} \sum_{v \in V} w_{uv} \right)$ represents the within-cluster ratio for the k-th cluster and $q_k = \left(\sum_{u \in U_k} \sum_{v \in V} w_{uv} / \sum_{u \in V} \sum_{v \in V} w_{uv} \right)$ represents the betweenness ratio for the k-th cluster. For each dataset case, we compute the modularity index for the instances $1 \leq K \leq 200$. Each K instance involves performing NMF clustering to obtain the approximate matrix $\hat{\mathbf{H}}$ and then averaging the modularity index over 10 runs of K-means rounding. For the US construction dataset, the best index value is reached at $K = 66$ for $Q = 0.197$. We opted, although, for $K = 64$ as in Kagawa et al. (2015) in order to ensure adequate comparison. Actually, $Q(K = 64) = 0.195$ is very close in value to the best case $Q(K = 66)$. For the US transport dataset, K is set to the maximum index value, $K = 68$, which coincides exactly with the Kagawa et al. (2015) choice. The plots of $Q(K)$ are available in the supporting materials.

Adjacency matrices considered here represent CO_2 emissions in interindustries induced by the final demand of products for a certain industry in a certain country. These matrices are based in an atomic level on three elements that are all obtained from the World Input–Output Database (Kagawa et al. 2015; Dietzenbacher et al. 2013; Tukker and Dietzenbacher 2013): the quantities of product sale between industries of different countries, quantities of product sale to final consumers, and the amounts of carbon dioxide (CO_2) emission of industries in different countries. All these data are actually estimates that could possibly suffer from statistical biases such as noise and missing values. Some weaknesses such as differences between countries in price concept or in import–export-processing activities are tackled by Dietzenbacher et al. (2013). However, the risk that estimates do not match reality always exists. This can be an issue if the employed analysis is quite sensitive to errors in the used estimates, which is exactly our case.

Actually, we rely on the update rule (8) for the solving process, which is nothing more than an iterative improvement based on a gradient-descent procedure (Lee and Seung 2001). This later method is famously known to be highly sensitive to the initial starting solution (Avriel 2003), which is set in our case according to Kagawa et al. (2015), Ding et al. (2008) to the indicator matrix solution obtained by spectral clustering and the application of K-means (plus a constant matrix). Furthermore, spectral clustering is sensitive to errors in the input adjacency matrix (Von Luxburg 2007, p. 18). Following this reasoning, the NMF method is equally sensitive to errors in the input adjacency matrix. On the other hand, we already mentioned about the sensitive nature of K-means to the initial starting points.

Given these different sources of uncertainty in the employed analysis, we could easily suspect a risk of overfitting dataset instances and model initial choices when reporting Kagawa et al. (2015) cluster assignments. Plus, the reported clusters convey significant information about entry points for mitigating global warming, e.g., which industry sectors could be starting points or priority targets when implementing policies of CO_2 emissions reduction involving the USA and China. Due to the relevance of the results' implications, special care needs to be taken regarding the impact of errors in input adjacency matrices on the output clustering results.

Instead of relying on heuristic procedures to reduce the model's uncertainties, which is in a sense done by Kagawa et al. (2013a, 2015) when generating M clusterings corresponding to M runs of the K-means algorithm and then choosing the one with the optimal Ncut, a further systematic mechanism could be more reliable. By simultaneously encapsulating the simulation module and iteratively improving the normalized cut performance for clusterings, our method provides a more rigorous way to approach the sensitivity issue. Robustness against noise in the adjacency matrix is evaluated through a very large number of scenarios. Our overall method can be used as a black-box procedure for clustering methods relying on K-means rounding, which also include spectral clusterings.

The setting of the experiment parameters is as follows. In order to cover a large range of noise magnitudes ξ, six instances are considered, $\xi = 0.0$ to 0.5 in steps of 0.1. The distribution of perturbations used to sample the perturbed adjacency matrices $\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_N$ is taken to be uniform. The parameters N the number of samples and M the number of clusterings are, respectively, set at 1000 and 100, which allow experiments to be run within an acceptable CPU time. All experiment runs were done in a Java environment on a 3-GHz CPU processor with 8 GB of memory. In the remainder of this section, we discuss the obtained results.

After the experiment is run for each instance of ξ, we construct a solution pool $\mathcal{X}$ composed by the found solutions plus the Kagawa et al. (2015) clustering. Notice that in all experiment runs, we obtain $C_{R_1} = C_{R_2}$. The results of performing the simulation module with three noise magnitudes, $\xi = 0.1$ (small), 0.5 (large), and 5 (huge), on the constructed set $\mathcal{X}$ are shown in Figs. 2, 3, and 4 and the corresponding Tables 3, 4, and 5 in the case of the US construction dataset. The results on the US transport dataset are available in the supporting materials. The first striking observation for both datasets is the low normalized cut performance of Kagawa et al. (2015) clustering for both the nominal adjacency matrix and across the perturbation scenarios. These Ncut values are almost twice the Ncut values of the experiment top clusterings, and those for the US transport case are much higher. Ncut values of Kagawa et al. (2015) clusterings are centered around approximately 21.56 and 45.7, compared with the robustest clusterings of the experiment, which gravitated around 9.8 and 10.9 for, respectively, US construction and transport demands. This means that in uncertain environments Kagawa et al. (2015) clusterings do not perform well, even if their Ncut values are close on average to the nominal case for low and large magnitude cases. It should be noted that the mean Ncut can be misleading due to the presence of outliers. These observations are elucidated clearly via the simulation assessments $R_1^{\mathcal{X}}$ and $R_2^{\mathcal{X}}$ of Figures 2, 3, and 4. While the clustering induced by the $\xi = 0.4$ experiment for the US construction case exhibits the most robust behavior across the various noise magnitudes, the Kagawa et al. (2015)

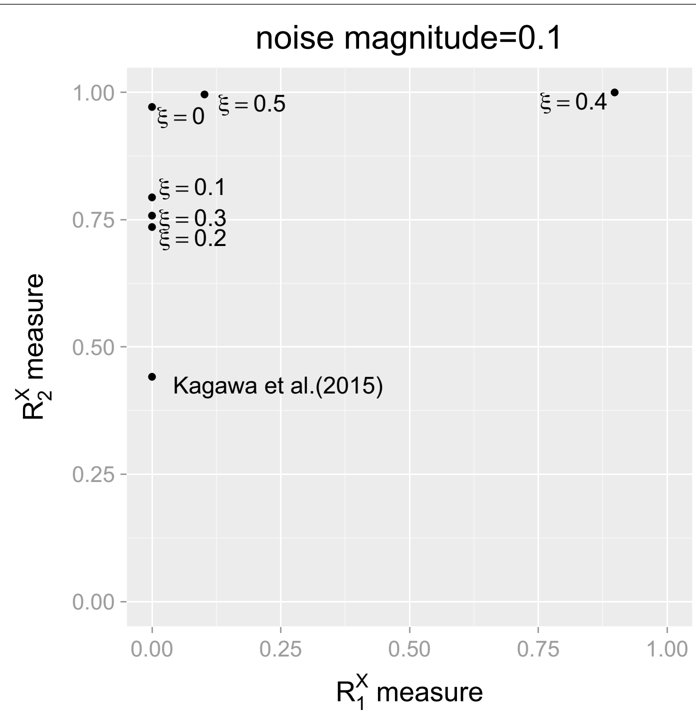


Fig. 2 Results of a simulation run for the noise magnitude $\xi = 0.1$ and the US construction dataset such that the set $\mathcal{X}$ is taken to be the top clusterings found by the experiment (ξ varying from 0.0 to 0.5) in addition to the Kagawa et al. (2015) clustering

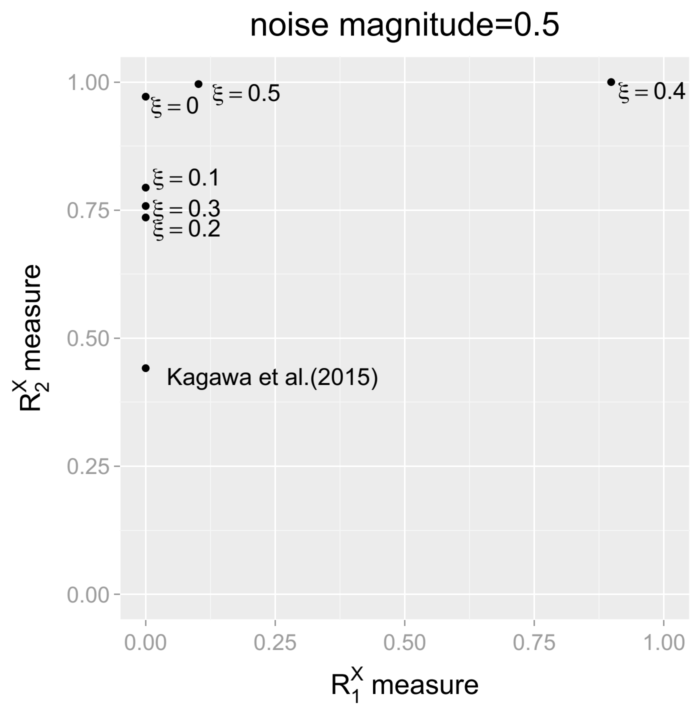


Fig. 3 Results of a simulation run for the noise magnitude $\xi = 0.5$ and the US construction dataset such that the set $\mathcal{X}$ is taken to be the top clusterings found by the experiment (ξ varying from 0.0 to 0.5) in addition to the Kagawa et al. (2015) clustering

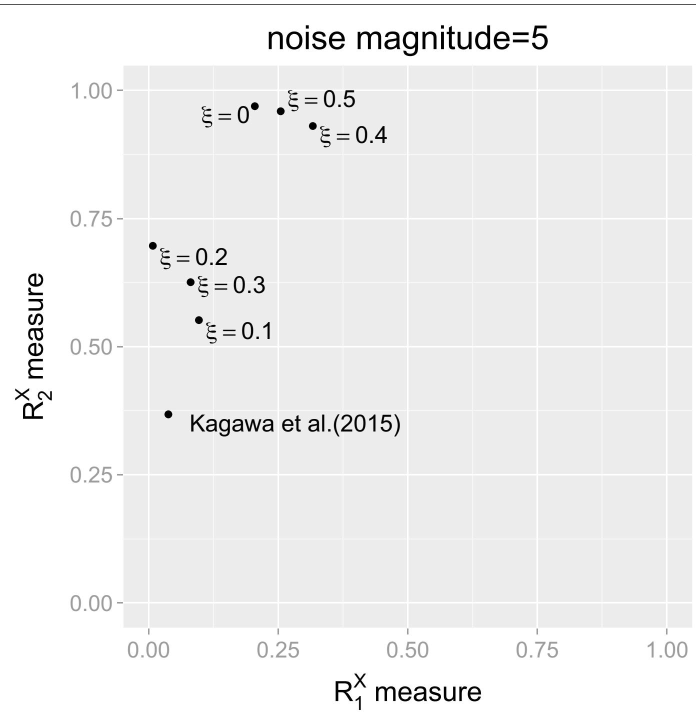


Fig. 4 Results of a simulation run for the noise magnitude $\xi = 5$ and the US construction dataset such that the set $\mathcal{X}$ is taken to be the top clusterings found by the experiment (ξ varying from 0.0 to 0.5) in addition to the Kagawa et al. (2015) clustering

Table 3 Ncut values for the nominal scenario (nominal Ncut) and averaged over all scenarios (mean Ncut) for the clusterings of the simulation run corresponding to Fig. 2

Clustering	$\xi = 0$	$\xi = 0.1$	$\xi = 0.2$	$\xi = 0.3$	$\xi = 0.4$	$\xi = 0.5$	Kagawa et al. (2015)
Nominal Ncut	9.800	11.991	12.931	12.557	9.521	9.557	21.564
Mean Ncut	9.802	11.994	12.943	12.561	9.522	9.560	21.568

Table 4 Ncut values for the nominal scenario (nominal Ncut) and averaged over all scenarios (mean Ncut) for the clusterings of the simulation run corresponding to Fig. 3

Clustering	$\xi = 0$	$\xi = 0.1$	$\xi = 0.2$	$\xi = 0.3$	$\xi = 0.4$	$\xi = 0.5$	Kagawa et al. (2015)
Nominal Ncut	9.800	11.991	12.931	12.557	9.521	9.557	21.564
Mean Ncut	9.810	12.006	12.953	12.566	9.529	9.567	21.578

Table 5 Ncut values for the nominal scenario (nominal Ncut) and averaged over all scenarios (mean Ncut) for the clusterings of the simulation run corresponding to Fig. 4

Clustering	$\xi = 0$	$\xi = 0.1$	$\xi = 0.2$	$\xi = 0.3$	$\xi = 0.4$	$\xi = 0.5$	Kagawa et al. (2015)
Nominal Ncut	9.800	11.991	12.931	12.557	9.521	9.557	21.564
Mean Ncut	12.498	14.523	14.811	11.658	5.751	10.753	31.696

clustering is found to be in an unfavorable position of below and to the left of all other plots of $(R_1^{\mathcal{X}}, R_2^{\mathcal{X}})$.

After examining the consistency of the clusters order throughout our experiment runs, which is available in the supporting materials, we compare the cluster components of the Kagawa et al. (2015) clusterings to those of our robustest clusterings for US transport and US construction in Tables 6 and 7, respectively. The US clusters, i.e., induced and generated by US industries, of (C1) and (C5) have the highest positions in Tables 6 and 7, respectively. Each cluster generated in 2009 more than 162 million tonnes of CO_2 emissions in US territories. Their compact structure and high within-cluster emissions are the main differences with the Kagawa et al. (2015) US clusters. In fact, there were two US clusters in the Kagawa et al. (2015) clustering for the US construction dataset.

On the other hand, the two Chinese clusters of (C2) and (C6) have exactly the same components, except for two additional Korean elements, c1 and c3, in the US construction case. The differences between the (C2) and (C6) components and the Kagawa et al. (2015) Chinese clusters are small. The quasi-totalities of the elements are similar. More importantly, the two strong supply chains reported in Kagawa et al. (2015)—(1) c17 (CHN) $\Rightarrow$ c12 (CHN) and (2) c17 (CHN) $\Rightarrow$ c9 (CHN)—are still present in our Chinese clusters. c17 refers to Electricity, Gas and Water Supply, c12 to Basic Metals and Fabricated Metal, and c9 to Chemicals and Chemical Products. Both supply chains are major contributors in the CO_2 emissions within the Chinese clusters. Figure 5 illustrates the Chinese emission cluster (C2) of Table 6, which corresponds to the US transport case.

By obtaining similar results for the most important supply chains, we confirm the environmental policy conclusions reached by Kagawa et al. (2015). Mitigating global

Table 6 CO_2 clusters with the highest within-cluster emissions (Kt CO_2 eq.) induced by the final demand of the US transport dataset for the following two cases: the Kagawa et al. (2015) clustering and the robustest clustering of the experiment ($\xi = 0.3$)

Rank	Kagawa et al. (2015)			Experiment ($\xi = 0.3$)		
	Cluster name	Industrial sectors	Within-cluster sum	Cluster name	Industrial sectors	Within-cluster sum
1	American and polonese cluster	USA: Mining and Quarrying	49,201	American cluster (C1)	USA: Mining and Quarrying	164,072
		USA: Pulp, Paper, Paper, Printing and Publishing			USA: Pulp, Paper, Paper, Printing and Publishing	
		USA: Coke, Refined Petroleum and Nuclear Fuel			USA: Coke, Refined Petroleum and Nuclear Fuel	
		USA: Chemicals and Chemical Products			USA: Chemicals and Chemical Products	
		USA: Rubber and Plastics			USA: Rubber and Plastics	
		USA: Other Non-Metallic Mineral			USA: Other Non-Metallic Mineral	
		USA: Basic Metals and Fabricated Metal			USA: Basic Metals and Fabricated Metal	
		USA: Transport Equipment			USA: Wood and Products of Wood and Cork	
		USA: Electricity, Gas and Water Supply			USA: Electricity, Gas and Water Supply	
		USA: Wholesale Trade and Commission Trade,			USA: Wholesale Trade and Commission Trade,	
		Except of Motor Vehicles and Motorcycles			Except of Motor Vehicles and Motorcycles	
		USA: Inland Transport			USA: Inland Transport	
		USA: Other Supporting and Auxiliary Transport			USA: Retail Trade, Except of Motor Vehicles and	
		Activities; Activities of Travel Agencies			Motorcycles; Repair of Household Goods	
		USA: Financial Intermediation			USA: Construction	
		USA: Renting of M&Eq and Other Business Activities			USA: Air Transport	
		POL: Mining and Quarrying			USA: Other Supporting and Auxiliary Transport	
		POL: Coke, Refined Petroleum and Nuclear Fuel			Activities; Activities of Travel Agencies	
		POL: Chemicals and Chemical Products			USA: Post and Telecommunications	
		POL: Rubber and Plastics			USA: Financial Intermediation	
		POL: Other Non-Metallic Mineral			USA: Renting of M&Eq and Other Business Activities	
		POL: Basic Metals and Fabricated Metal			USA: Other Community, Social and Personal Services	
		POL: Machinery, Nec				
		POL: Electrical and Optical Equipment				
		POL: Transport Equipment				
		POL: Electricity, Gas and Water Supply				
		POL: Wholesale Trade and Commission Trade,				
		Except of Motor Vehicles and Motorcycles				
		POL: Retail Trade, Except of Motor Vehicles and				

Table 6 continued

Rank	Kagawa et al. (2015)			Experiment ($\xi = 0.3$)		
	Cluster name	**Industrial sectors**	**Within-cluster sum**	**Cluster name**	**Industrial sectors**	**Within-cluster sum**
		Motorcycles; Repair of Household Goods				
		POL: Inland Transport				
		POL: Real Estate Activities				
		POL: Renting of M&Eq and Other Business Activities				
		POL: Other Community, Social and Personal Services				
2	Big cluster	(664 elements)	18,488	Chinese cluster (C2)	CHN: Agriculture, Hunting, Forestry and Fishing	13,035
					CHN: Mining and Quarrying	
					CHN: Food, Beverages and Tobacco	
					CHN: Textiles and Textile Products	
					CHN: Wood and Products of Wood and Cork	
					CHN: Pulp, Paper, Paper, Printing and Publishing	
					CHN: Coke, Refined Petroleum and Nuclear Fuel	
					CHN: Chemicals and Chemical Products	
					CHN: Rubber and Plastics	
					CHN: Other Non-Metallic Mineral	
					CHN: Basic Metals and Fabricated Metal	
					CHN: Machinery, Nec	
					CHN: Electrical and Optical Equipment	
					CHN: Electricity, Gas and Water Supply	
					CHN: Hotels and Restaurants	
					CHN: Inland Transport	
					CHN: Renting of M&Eq and Other Business Activities	
3	Chinese cluster	CHN: Agriculture, Hunting, Forestry and Fishing	12,805	Rest of the world cluster (C3)	DNK: Water Transport	5963
		CHN: Mining and Quarrying			FRA: Agriculture, Hunting, Forestry and Fishing	
		CHN: Food, Beverages and Tobacco			FRA: Food, Beverages and Tobacco	
		CHN: Textiles and Textile Products			FRA: Wood and Products of Wood and Cork	
		CHN: Leather, Leather and Footwear			FRA: Pulp, Paper, Paper, Printing and Publishing	
		CHN: Pulp, Paper, Paper, Printing and Publishing			FRA: Coke, Refined Petroleum and Nuclear Fuel	
		CHN: Coke, Refined Petroleum and Nuclear Fuel			FRA: Chemicals and Chemical Products	
		CHN: Chemicals and Chemical Products			FRA: Rubber and Plastics	
		CHN: Rubber and Plastics			FRA: Other Non-Metallic Mineral	

Table 6 continued

Rank	Kagawa et al. (2015)			Experiment ($\xi = 0.3$)		
	Cluster name	Industrial sectors	Within-cluster sum	Cluster name	Industrial sectors	Within-cluster sum
		CHN: Other Non-Metallic Mineral			FRA: Basic Metals and Fabricated Metal	
		CHN: Basic Metals and Fabricated Metal			FRA: Machinery, Nec	
		CHN: Machinery, Nec			FRA: Electrical and Optical Equipment	
		CHN: Electrical and Optical Equipment			FRA: Transport Equipment	
		CHN: Transport Equipment			FRA: Manufacturing, Nec; Recycling	
		CHN: Electricity, Gas and Water Supply			FRA: Electricity, Gas and Water Supply	
		CHN: Inland Transport			FRA: Sale, Maintenance and Repair of Motor Vehicles	
		CHN: Renting of M&Eq and Other Business Activities			and Motorcycles; Retail Sale of Fuel	
					FRA: Wholesale Trade and Commission Trade,	
					Except of Motor Vehicles and Motorcycles	
					FRA: Retail Trade, Except of Motor Vehicles and	
					Motorcycles; Repair of Household Goods	
					FRA: Hotels and Restaurants	
					FRA: Inland Transport	
					FRA: Financial Intermediation	
					FRA: Renting of M&Eq and Other Business Activities	
					FRA: Other Community, Social and Personal Services	
					KOR: Water Transport	
					ROW: Agriculture, Hunting, Forestry and Fishing	
					ROW: Mining and Quarrying	
					ROW: Food, Beverages and Tobacco	
					ROW: Wood and Products of Wood and Cork	
					ROW: Pulp, Paper, Paper, Printing and Publishing	
					ROW: Chemicals and Chemical Products	
					ROW: Rubber and Plastics	
					ROW: Basic Metals and Fabricated Metal	
					ROW: Machinery, Nec	
					ROW: Electrical and Optical Equipment	
					ROW: Electricity, Gas and Water Supply	
					ROW: Wholesale Trade and Commission Trade,	
					Except of Motor Vehicles and Motorcycles	

Table 6 continued

Rank	Kagawa et al. (2015)			Experiment ($\xi = 0.3$)		
	Cluster name	Industrial sectors	Within-cluster sum	Cluster name	Industrial sectors	Within-cluster sum
					ROW: Retail Trade, Except of Motor Vehicles and	
					Motorcycles; Repair of Household Goods	
					ROW: Hotels and Restaurants	
					ROW: Inland Transport	
					ROW: Water Transport	
					ROW: Air Transport	
					ROW: Other Supporting and Auxiliary Transport	
					Activities; Activities of Travel Agencies	
					ROW: Post and Telecommunications	
					ROW: Financial Intermediation	
					ROW: Renting of M&Eq and Other Business Activities	
					ROW: Other Community, Social and Personal Services	
4	Rest of the world cluster	ROW: Mining and Quarrying	4278	Indian cluster (C4)	IND: Textiles and Textile Products	1570
		ROW: Rubber and Plastics			IND: Wood and Products of Wood and Cork	
		ROW: Basic Metals and Fabricated Metal			IND: Pulp, Paper, Paper, Printing and Publishing	
		ROW: Electricity, Gas and Water Supply			IND: Coke, Refined Petroleum and Nuclear Fuel	
		ROW: Inland Transport			IND: Chemicals and Chemical Products	
		ROW: Water Transport			IND: Rubber and Plastics	
		ROW: Renting of M&Eq and Other Business Activities			IND: Other Non-Metallic Mineral	
					IND: Basic Metals and Fabricated Metal	
					IND: Machinery, Nec	
					IND: Electrical and Optical Equipment	
					IND: Transport Equipment	
					IND: Manufacturing, Nec; Recycling	
					IND: Electricity, Gas and Water Supply	
					IND: Construction	
					IND: Inland Transport	
					IND: Post and Telecommunications	
					IND: Financial Intermediation	
					IND: Renting of M&Eq and Other Business Activities	
					ROW: Manufacturing, Nec; Recycling	
5	Canadian cluster	CAN: Agriculture, Hunting, Forestry and Fishing	2110	Mexican cluster	MEX: Agriculture, Hunting, Forestry and Fishing	1302

Table 6 continued

Rank	Kagawa et al. (2015)			Experiment ($\xi = 0.3$)		
	Cluster name	**Industrial sectors**	**Within-cluster sum**	**Cluster name**	**Industrial sectors**	**Within-cluster sum**
		CAN: Mining and Quarrying			MEX: Mining and Quarrying	
		CAN: Food, Beverages and Tobacco			MEX: Food, Beverages and Tobacco	
		CAN: Wood and Products of Wood and Cork			MEX: Textiles and Textile Products	
		CAN: Pulp, Paper, Paper, Printing and Publishing			MEX: Wood and Products of Wood and Cork	
		CAN: Coke, Refined Petroleum and Nuclear Fuel			MEX: Pulp, Paper, Paper, Printing and Publishing	
		CAN: Chemicals and Chemical Products			MEX: Coke, Refined Petroleum and Nuclear Fuel	
		CAN: Rubber and Plastics			MEX: Chemicals and Chemical Products	
		CAN: Other Non-Metallic Mineral			MEX: Rubber and Plastics	
		CAN: Basic Metals and Fabricated Metal			MEX: Other Non-Metallic Mineral	
		CAN: Machinery, Nec			MEX: Basic Metals and Fabricated Metal	
		CAN: Electrical and Optical Equipment			MEX: Machinery, Nec	
		CAN: Transport Equipment			MEX: Electrical and Optical Equipment	
		CAN: Electricity, Gas and Water Supply			MEX: Transport Equipment	
		CAN: Wholesale Trade and Commission Trade,			MEX: Wholesale Trade and Commission Trade,	
		Except of Motor Vehicles and Motorcycles			Except of Motor Vehicles and Motorcycles	
		CAN: Retail Trade, Except of Motor Vehicles and			MEX: Retail Trade, Except of Motor Vehicles and	
		Motorcycles; Repair of Household Goods			Motorcycles; Repair of Household Goods	
		CAN: Inland Transport			MEX: Inland Transport	
		CAN: Water Transport			MEX: Manufacturing, Nec; Recycling	
		CAN: Air Transport			MEX: Electricity, Gas and Water Supply	
		CAN: Other Supporting and Auxiliary Transport			MEX: Hotels and Restaurants	
		Activities; Activities of Travel Agencies			MEX: Sale, Maintenance and Repair of Motor	
		CAN: Post and Telecommunications			Vehicles and Motorcycles; Retail Sale of Fuel	
		CAN: Financial Intermediation			MEX: Financial Intermediation	
		CAN: Real Estate Activities			MEX: Renting of M&Eq and Other Business Activities	
		CAN: Renting of M&Eq and Other Business Activities				
		CAN: Other Community, Social and Personal Services				

Table 7 CO_2 clusters with the highest within-cluster emissions (Kt CO_2 eq.) induced by the final demand of the US construction dataset for the following two cases: the Kagawa et al. (2015) clustering and the robustest clustering of the experiment ($\xi = 0.4$)

Rank	Kagawa et al. (2015)			Experiment ($\xi = 0.4$)		
	Cluster name	Industrial sectors	Within-cluster sum	Cluster name	Industrial sectors	Within-cluster sum
1	American cluster (1)	USA: Mining and Quarrying	120,033	American cluster (C5)	USA: Mining and Quarrying	162,275
		USA: Coke, Refined Petroleum and Nuclear Fuel			USA: Coke, Refined Petroleum and Nuclear Fuel	
		USA: Other Non-Metallic Mineral			USA: Other Non-Metallic Mineral	
		USA: Basic Metals and Fabricated Metal			USA: Basic Metals and Fabricated Metal	
		USA: Electricity, Gas and Water Supply			USA: Electricity, Gas and Water Supply	
		USA: Construction			USA: Construction	
		USA: Inland Transport			USA: Inland Transport	
					USA: Wood and Products of Wood and Cork	
					USA: Pulp, Paper, Paper, Printing and Publishing	
					USA: Chemicals and Chemical Products	
					USA: Rubber and Plastics	
					USA: Wholesale Trade and Commission Trade,	
					Except of Motor Vehicles and Motorcycles	
					USA: Retail Trade, Except of Motor Vehicles and	
					Motorcycles; Repair of Household Goods	
					USA: Air Transport	
					USA: Other Supporting and Auxiliary Transport	
					Activities; Activities of Travel Agencies	
					USA: Financial Intermediation	
					USA: Renting of M&Eq and Other Business Activities	

Table 7 continued

Rank	Kagawa et al. (2015)			Experiment ($\xi = 0.4$)		
	Cluster name	**Industrial sectors**	**Within-cluster sum**	**Cluster name**	**Industrial sectors**	**Within-cluster sum**
					USA: Other Community, Social and Personal Services	
2	Chinese cluster	CHN: Agriculture, Hunting, Forestry and Fishing	12,900	Chinese cluster (C6)	CHN: Agriculture, Hunting, Forestry and Fishing	13,037
		CHN: Mining and Quarrying			CHN: Mining and Quarrying	
		CHN: Food, Beverages and Tobacco			CHN: Food, Beverages and Tobacco	
		CHN: Textiles and Textile Products			CHN: Textiles and Textile Products	
		CHN: Wood and Products of Wood and Cork			CHN: Wood and Products of Wood and Cork	
		CHN: Pulp, Paper, Paper, Printing and Publishing			CHN: Pulp, Paper, Paper, Printing and Publishing	
		CHN: Coke, Refined Petroleum and Nuclear Fuel			CHN: Coke, Refined Petroleum and Nuclear Fuel	
		CHN: Chemicals and Chemical Products			CHN: Chemicals and Chemical Products	
		CHN: Rubber and Plastics			CHN: Rubber and Plastics	
		CHN: Other Non-Metallic Mineral			CHN: Other Non-Metallic Mineral	
		CHN: Basic Metals and Fabricated Metal			CHN: Basic Metals and Fabricated Metal	
		CHN: Machinery, Nec			CHN: Machinery, Nec	
		CHN: Electrical and Optical Equipment			CHN: Electrical and Optical Equipment	
		CHN: Electricity, Gas and Water Supply			CHN: Electricity, Gas and Water Supply	
		CHN: Inland Transport			CHN: Inland Transport	
		CHN: Renting of M&Eq and Other Business Activities			CHN: Renting of M&Eq and Other Business Activities	
					CHN: Hotels and Restaurants	
					KOR: Agriculture, Hunting, Forestry and Fishing	

Table 7 continued

Rank	Kagawa et al. (2015)			Experiment ($\xi = 0.4$)		
	Cluster name	**Industrial sectors**	**Within-cluster sum**	**Cluster name**	**Industrial sectors**	**Within-cluster sum**
					KOR: Food, Beverages and Tobacco	
3	American cluster (2)	USA: Agriculture, Hunting, Forestry and Fishing	7458	Big cluster (C8)	(570 elements)	3816
		USA: Food, Beverages and Tobacco				
		USA: Textiles and Textile Products				
		USA: Wood and Products of Wood and Cork				
		USA: Pulp, Paper, Paper, Printing and Publishing				
		USA: Chemicals and Chemical Products				
		USA: Rubber and Plastics				
		USA: Wholesale Trade and Commission Trade,				
		Except of Motor Vehicles and Motorcycles				
		USA: Retail Trade, Except of Motor Vehicles and				
		Motorcycles; Repair of Household Goods				
		USA: Hotels and Restaurants				
		USA: Air Transport				
		USA: Other Supporting and Auxiliary Transport				
		Activities; Activities of Travel Agencies				
		USA: Post and Telecommunications				
		USA: Financial Intermediation				
		USA: Renting of M&Eq and Other Business Activities				
		USA: Other Community, Social and Personal Services				
		ROW: Chemicals and Chemical Products				

Table 7 continued

Rank	Kagawa et al. (2015)			Experiment ($\xi = 0.4$)		
	Cluster name	**Industrial sectors**	**Within-cluster sum**	**Cluster name**	**Industrial sectors**	**Within-cluster sum**
4	Rest of the world cluster	ROW: Mining and Quarrying	3869	Rest of the world cluster (C7)	ROW: Mining and Quarrying	3264
		ROW: Electricity, Gas and Water Supply			ROW: Electricity, Gas and Water Supply	
		ROW: Inland Transport			ROW: Inland Transport	
		ROW: Water Transport			ROW: Water Transport	
		ROW: Rubber and Plastics				
		ROW: Renting of M&Eq and Other Business Activities				
5	Canadian cluster	CAN: Mining and Quarrying	1533	Russian cluster	RUS: Mining and Quarrying	1297
		CAN: Pulp, Paper, Paper, Printing and Publishing			RUS: Basic Metals and Fabricated Metal	
		CAN: Coke, Refined Petroleum and Nuclear Fuel			RUS: Coke, Refined Petroleum and Nuclear Fuel	
		CAN: Electricity, Gas and Water Supply			RUS: Electricity, Gas and Water Supply	
		CAN: Wholesale Trade and Commission Trade,			RUS: Wholesale Trade and Commission Trade,	
		Except of Motor Vehicles and Motorcycles			Except of Motor Vehicles and Motorcycles	
		CAN: Chemicals and Chemical Products			RUS: Chemicals and Chemical Products	
		CAN: Inland Transport			RUS: Inland Transport	
		CAN: Agriculture, Hunting, Forestry and Fishing				
		CAN: Wood and Products of Wood and Cork				
		CAN: Basic Metals and Fabricated Metal				
		CAN: Retail Trade, Except of Motor Vehicles and				
		Motorcycles; Repair of Household Goods				

Table 7 continued

Rank	Kagawa et al. (2015)			Experiment ($\xi = 0.4$)		
	Cluster name	**Industrial sectors**	**Within-cluster sum**	**Cluster name**	**Industrial sectors**	**Within-cluster sum**
		CAN: Renting of M&Eq and Other Business Activities				
		CAN: Other Community, Social and Personal Services				

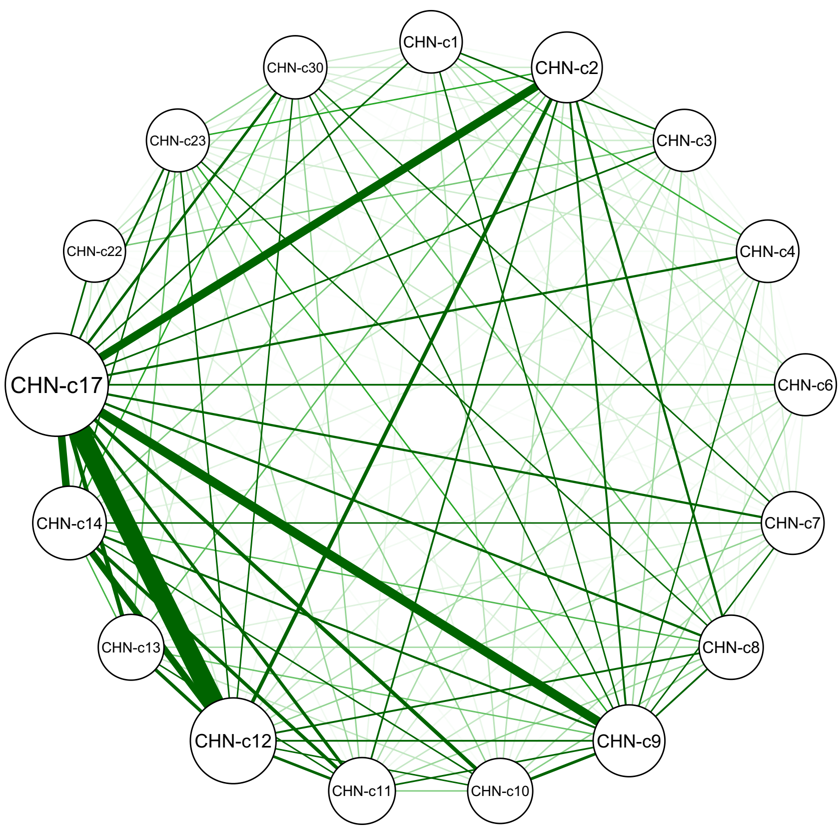


Fig. 5 Chinese emission cluster (C2) induced by US transport equipment demand obtained in Table 6. This figure is drawn using the "qgraph" R package (Epskamp et al. 2012)

warming through supply-chain transfers that could be based on reducing the amount of CO_2 emissions for supply chains (1) and (2) has, from our point of view, a good immunity against random deviations of the input–output data. Additionally, our superior Ncut performance reached for the robustest clusterings confers more trust in the environmental conclusions about the policies suggested in Kagawa et al. (2015).

5 Conclusion

In this study, we establish a sampling-based procedure in order to examine the robustness of clusterings that could be found using nonnegative matrix factorization or spectral clustering methods. An application of the procedure is provided here by re-examining/comparing the analysis of Kagawa et al. (2015). In their paper, significant clusters in terms of CO_2 emissions that are rapidly growing over time were found. Here, our procedure is applied to the datasets of Kagawa et al. (2015) that have strong environmental implications, namely the two CO_2 emissions networks induced by the US construction and US transport equipment sectors. In our empirical results, we find clusterings that have much better normalized cut performance and robustness assessments than those of Kagawa et al. (2015). Some differences in the components between the compared clusters are observed. However, the main supply-chain paths on which Kagawa et al. (2015) based their recommendations for mitigating global warning still persist. These recommendations concern the significant Chinese clusters linked to our target US demands. In summary, from a robustness perspective, we concur with the Kagawa et al. (2015) environmental conclusions regarding policies.

Authors' contributions

OR, HO, and SK proposed the methodology and provided discussions. Omar Rifki was in charge of data collection and conducted data analysis. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by a Grant-in-Aid for research [Nos. 26241031 and 16H01797] from the Ministry of Education, Culture, Sports, Science and Technology in Japan.

Received: 21 September 2016 Accepted: 26 January 2017
Published online: 14 February 2017

References

Avriel M (2003) Nonlinear programming: analysis and methods. Courier Corporation
Bottou L, Bengio Y (1994) Convergence properties of the k-means algorithms. In: Advances in neural information processing systems 7,[NIPS conference, Denver, Colorado, USA, 1994], pp 585–592
Bradley PS, Fayyad UM (1998) Refining initial points for k-means clustering. In: ICML, vol 98. Citeseer, pp 91–99
Davis SJ, Peters GP, Caldeira K (2011) The supply chain of CO_2 emissions. Proceedings of the National Academy of Sciences, 201107409
Dietzenbacher E (1995) On the bias of multiplier estimates. J Reg Sci 35(3):377–390
Dietzenbacher E (2006) Multiplier estimates: to bias or not to bias? J Reg Sci 46(4):773–786
Dietzenbacher E, Los B, Stehrer R, Timmer M, De Vries G (2013) The construction of world input–output tables in the wiod project. Econ Syst Res 25(1):71–98
Ding C, Li T, Jordan MI (2008) Nonnegative matrix factorization for combinatorial optimization: spectral clustering, graph matching, and clique finding. In: 8th IEEE International conference on data mining, 2008. ICDM'08, pp 183–192. IEEE
Ding CH, He X, Simon HD (2005) On the equivalence of nonnegative matrix factorization and spectral clustering. In: SDM, vol 5. SIAM, pp 606–610
Donath WE, Hoffman AJ (1973) Lower bounds for the partitioning of graphs. IBM J Res Dev 17(5):420–425
Duda RO, Hart PE, Stork DG (1995) Pattern classification and scene analysis, 2nd edn. Wiley Interscience, New York
Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D (2012) qgraph: network visualizations of relationships in psychometric data. J Stat Softw 48(4):1–18
Fiedler M (1973) Algebraic connectivity of graphs. Czechoslov Math J 23(2):298–305
Gendreau M, Potvin J-Y (2010) Handbook of metaheuristics, vol 2. Springer, New York
Kagawa S, Okamoto S, Suh S, Kondo Y, Nansai K (2013a) Finding environmentally important industry clusters: multiway cut approach using nonnegative matrix factorization. Soc Netw 35(3):423–438

Kagawa S, Suh S, Hubacek K, Wiedmann T, Nansai K, Minx J (2015) CO_2 emission clusters within global supply chain networks: implications for climate change mitigation. Glob Environ Chang
Kagawa S, Suh S, Kondo Y, Nansai K (2013b) Identifying environmentally important supply chain clusters in the automobile industry. Econ Syst Res 25(3):265–286
Kannan R, Vempala S, Vetta A (2004) On clusterings: good, bad and spectral. J ACM 51(3):497–515
Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788–791
Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems, pp 556–562
Lenzen M, Moran D, Kanemoto K, Foran B, Lobefaro L, Geschke A (2012) International trade drives biodiversity threats in developing nations. Nature 486(7401):109–112
Liang S, Feng Y, Xu M (2015) Structure of the global virtual carbon network: revealing important sectors and communities for emission reduction. J Ind Ecol 19(2):307–320
MacQueen J et al (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley symposium on mathematical statistics and probability, vol 1. Oakland, CA, USA, pp 281–297
Miller RE, Blair PD (2009) Input–output analysis: foundations and extensions. Cambridge University Press, Cambridge
Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113
Ng AY, Jordan MI, Weiss Y et al (2002) On spectral clustering: analysis and an algorithm. Adv Neural Inf Process Syst 2:849–856
Peters GP, Minx JC, Weber CL, Edenhofer O (2011) Growth in emission transfers via international trade from 1990 to 2008. In: Proceedings of the national academy of sciences, 201006388
Schaeffer SE (2007) Graph clustering. Comput Sci Rev 1(1):27–64
Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell 22(8):888–905
Tukker A, Dietzenbacher E (2013) Global multiregional input-output frameworks: an introduction and outlook. Econ Syst Res 25(1):1–19
Von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416
Zhang Z, Jordan MI et al (2008) Multiway spectral clustering: a margin-based perspective. Stat Sci 23(3):383–403

Submit your manuscript to a SpringerOpen journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com