Many objective optimization and complex network analysis

Maulana, A.

PART II

5

Modularity Maximization in Multiplex Network Analysis Using Many-Objective Optimization

5.1 · Introduction

In many disciplines, complex systems can be studied through network modeling and analysis. This yields a better understanding of complex phenomena, including conflicting sociology phenomena, spreading of disease, conflicting economic situations, telecommunication systems, biological systems, and networks in engineering. The networks or a collection of nodes are joined in pairs by edges. Clustering such groups of nodes in the network has become an important area of research. Network data becomes increasingly available but is also complex due to the omnipresence of data measurement and inquiry as a recent trend. In this chapter, we will focus on a special class of networks so-called multiplex networks. Often for the same set of network nodes, several or many network layers can be defined. Networks defining trade of different types of commodities is an example and it will provide a case study for this chapter. Other examples include multiplex networks

- in communication via different channels (social media, telephone, peer-to-peer),
- in biology, the different types of signaling networks of trees or plants (via scents, via insects, via underground root networks),
- in sociology, defined by different types of relationships, such as personal friends, relatives, business relationships, which might partially overlap.

This chapter presents a first step in the analysis of such multiplex networks by means of modularity optimization, where modularity is a measure of the quality of how well a partition of a network is representing communities. We consider the optimization of modularity for the different layers as the objective functions. Optimizing several $(2, 3)$ objectives simultaneously can be addressed by multi-objective optimization and many (>3) objectives by many-objective optimization resulting in a high dimensional Pareto front. By computing the Pareto fronts of pairs of different layers we find relationships between the objectives. Layers can be in conflict with each other, meaning that they yield very different optimal modularity structures. They can be also complementary, meaning that maximizing the modularity of the one layer also maximizes the modularity of the other layer. In this case, it is possible to merge the layers without losing essential information. Finally, it is also possible that the maximization of modularity of one layer does not affect the optimization of the modularity of another layer, in which case the problem could be easily decomposed.

5.2 • Related Work

To optimize many objectives simultaneously various approaches have been developed. Some of them aim at reducing complexity, such as Objective Reduction in Many-objective Optimization: Linear and Nonlinear Algorithms [50], Reducing Complexity in Many-Objective Optimization Using Community Detection [40], and Objective Reduction Based on Nonlinear Correlation Information Entropy [57]. Other approaches are based on Evolutionary Multi-objective optimization (EMO) extended for dealing with many objectives, cf. [33]. In this chapter, the CoDEMO framework from Chapter 3 is applied. The objective functions are the modularities achieved for different layers.

5.3 • Many Objective Optimization Approach to Community Detection in Complex Networks

Our research approach is to perform many-objective optimization of network modularity by computing and visualizing a matrix of Pareto fronts for pairs of objectives. Then we use community detection algorithms to group objective functions in order to understand and visualize the conflict or correspondence of community structures w.r.t. different edge sets. For every edge set, one objective function is defined, which is to maximize the modularity of this edge set. The search space X is the space of all partitioning of the node sets. In this way for a multiplex network G with layers $G_1,...,G_M$

we define M objective functions $Q_1 : X \to \mathbb{R}_0^+$, $Q_2 : X \to \mathbb{R}_0^+$,...,$Q_M : X \to \mathbb{R}_0^+$. All objective functions are to be maximized. Our first goal is to compute Pareto optimal solutions. Then we analyze projections to pairs of objective functions (corresponding to pairs of layers), in order to understand the relationship between layers in terms of modularity structure. In this way, we aim to gain insight into essential aspects of the community structure of a given multiplex network.

5.4 · Network Analysis Method

Given as an input a multiplex network with M layers represented by a set of graphs G_1,..., G_M, the approach is called Pareto front Modularity for Multiplex Network (PaMoPlex). Similar to the CoDeMO approach, discussed in chapter 3, it is a workflow consisting of several subsequent analysis steps. It is summarized in a work flow which consists of two major phases: (1) Preparation of data by optimization, (2) Analysis of data. The preparation of data in step (1) of the analysis consists of solving optimization tasks to find non-dominated solutions. In order to get more precise results, we also compute single objective optima and marginal Pareto fronts for every pair of two objective functions (between modularities as objective functions associated with two layers, each). The first phase is summarized in the next three steps:

- Single Objective Optimization: Optimize the modularity of each network separately using evolutionary single objective optimization based on a genetic algorithm.
- Many-Objective Optimization: Optimize the modularity of network, all layers together, as one unit in a multiplex network. For this, we use M-objective optimization algorithms.
- Pairwise Pareto-Front Computation: Optimize modularity for pairs of objectives.

The optimization methods are evolutionary multi-objective optimization based on NSGA-II [16], MOEA/D [31] and SMS-EMOA [56], [8] (population size: 100, number of generations: 2000). For small examples, we use a complete enumeration of partitions. Since NSGA-II is not really appropriate for a Many-Objective Optimization problem we rely on the MOEA/D and SMS-EMOA algorithms for the experiment.

In the second phase, the obtained data are analyzed. This is conducted in the following three steps:

- Matrix of Pareto Fronts Analysis: Visualization of Pareto Fronts is done on a plot matrix, where each tile with $j \in 1, \ldots, M, i \in 1, \ldots, M, j > i$ consists of a plot of a Pareto front of tradeoffs between objectives Q_i and Q_j (see Figure 5).

- Correlation Heat Map Analysis: Computation of the correlation coefficients matrix from the projections of the output of many-objective optimization. The heat map has as many rows and columns as the number of network layers (or objectives). The Pearson correlation coefficients of the projected 2-objective function vectors have values in the range of [-1,1] for each pair of objective functions; see Table 2 for an example. In the heat map, see Figure 5.6 for an example, blue color represents positive correlations, whereas red color represents negative correlations. The intensity (darkness) and size of the colored square in each matrix cell grow with the absolute value.

- Community Analysis: This tool is based on the result of the correlation analysis. The correlation matrix is used for community detection by the graph-theoretic algorithm to detect communities using the information of correlation coefficients matrix and interpreting it as edge weights. Here the analysis proposed by Maulana et al. [40] is used, where the edge weight is determined by the absolute value of the correlation coefficient. This leads to a separation of independent communities of layers. Conflicting communities are placed opposite to each other (see Figure 5.5).

Further details on the analysis of examples and interpretation of results will be discussed in the subsequent sections.

5.5 · Case Study and Analysis

As an illustrative example on how to interpret results of multi-objective modularity optimization, we computed the exact Pareto fronts for three synthesized multiplex networks consisting of only two layers each. The networks and the corresponding Pareto fronts are displayed in Figure 5.1, Figure 5.2, and Figure 5.3. Red edges denote edge weights of 3, blue edges represent edge weights 1, and omitted edges have weight 0. A complete enumeration of all 203 possible partitioning was used to compute the exact Pareto fronts (cf. Bell 1934 [7]).

5.5.1 • Analysis on Synthesized Multiplex Networks

The first network in Figure 5.1 is a multiplex network where the maximization of modularity is conflicting, due to non-overlapping communities w.r.t. both layers. The linear Pareto front indicates a strong conflict between the maximization of two types and it is difficult to find a compromise solution that optimizes both objectives at the same time.

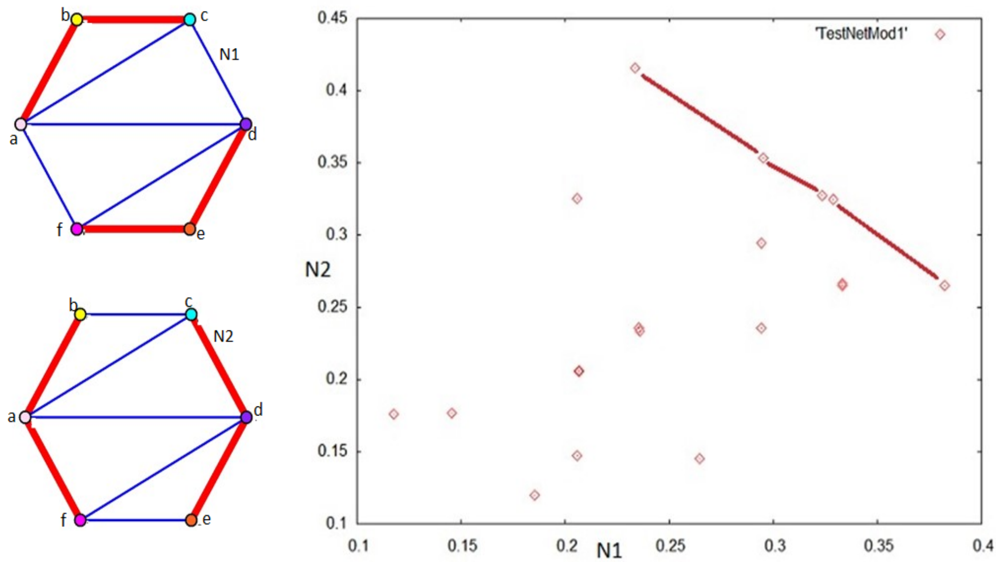

Figure 5.1 A visual depiction of the Pareto front for network modularity between two network layers N1 and N2 corresponding to highly conflicting objectives function

In the second example, in Figure 5.2, the optimal modularity for the first network is achieved by grouping the upper nodes in the graph, while for the second network it is important to group the lower nodes. Thereby the value of the modularity is widely indifferent to how the remaining nodes are grouped. This represents a case where the modularity optimization for the two layers is almost independent and the Pareto front has a knee point solution where both objective functions almost obtain their maximum. The correlation is close to zero. Finally, the third example in Figure 5.3 shows a multiplex network consisting of two equal edge sets. Here, solutions can be found that cluster for one layer optimally w.r.t. modularity necessarily also do so for the modularity of the second network. In other words, optimizing one network coincides

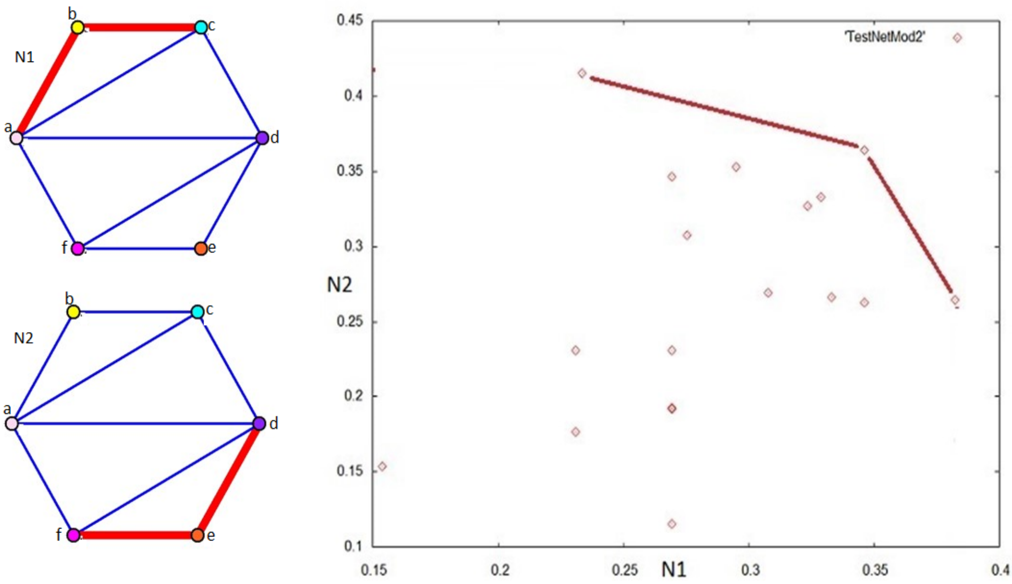

Figure 5.2 A visual depiction of the Pareto front for network modularity between two network layers N1 and N2 corresponding to highly correlated objectives function

with optimizing the other network. This is indicated by a perfect correlation between the modularities of sampled points even for random inputs. The Pareto front consists of only a single solution. In real-world applications, it is of course not so obvious how the structure of the Pareto front looks like. These three examples should be seen as boundary cases, which can help to interpret and understand the observed shape of Pareto fronts in such real-world networks.

5.5.2 • Economic Trade Multiplex Network Analysis

Next, a full PaMoPlex analysis on an economic dataset is provided. The data originates from network economy (trade data) using import-export Commodityodities network between countries in 2011 (see [39], Appendix). The data represents the import-export relationships between countries of the world, disaggregated for different traded commodities. This network can be defined as a multiplex network composed of many layers, where each layer is given by a different commodity. The nodes are given by 207 countries. A link between two countries in the $i-th$ layer defined as the weight will exist if there is trade between them in the $i-th$ commodity, for $i \in 1, \ldots, 11$. Data are presented in matrix form: rows and columns represent countries, and the entries of

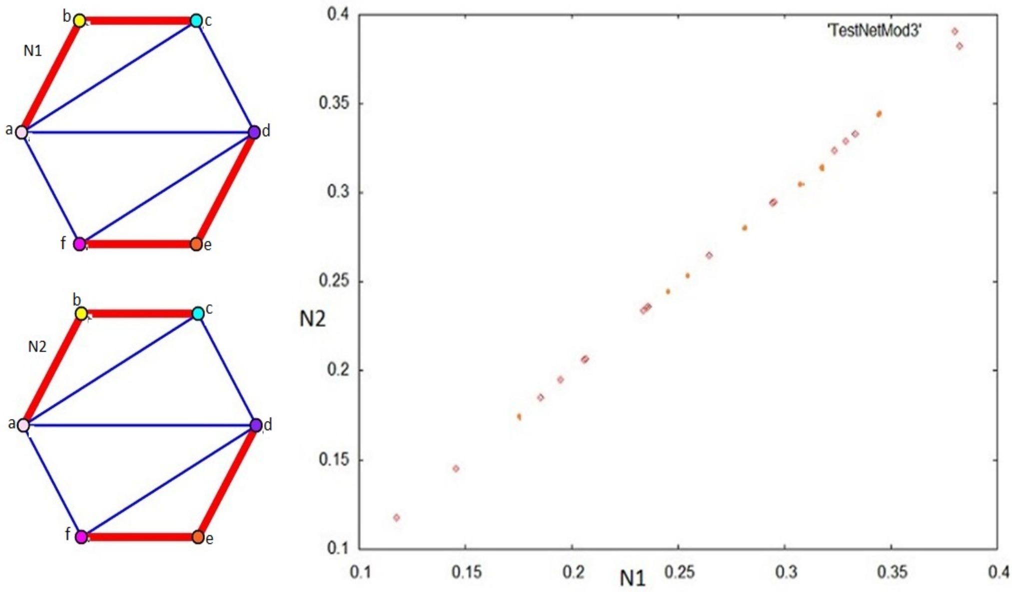

Figure 5.3 A visual depiction of the Pareto front for network modularity between two network layers N1 and N2 corresponding to very strong correlation of objectives function

the matrices are the volumes of trade. It is, therefore, a weighted multiplex network. The general classification is based on 96 different commodities. The classification is performed by grouping together similar commodities; this procedure leads to 11 aggregated 'super-commodities'.

The single objective optimization was conducted by a standard Louvain method and by a genetic algorithm. In all cases, the genetic algorithm found a better result. The results are summarized in Table 5.1 A typical number of communities when maximizing modularity are between 5 and 9.

The genetic algorithm is from the software package JMetal (gGA). It has population size 2000 and 100 generations were conducted. The default parameter settings for the genetic operators were used (http://jmetal.sourceforge.net/, February 2015). We suppose that by tuning of parameters better results can be achieved, but defer such studies to future research in order to focus more on the overall analysis method in this chapter.

The many-objective optimization yields a Pareto front that is embedded in an 11-dimensional space. The analysis of the correlation and community between objectives was conducted following the approach mentioned in [40]. From this, we compute the

Trading network	Louvain method		Genetic Algorithm	
	Modularity	NC	Modularity	NC
Trade 1	0.34392	9	0.35162	9
Trade 2	0.34794	9	0.35225	9
Trade 3	0.30513	9	0.30801	8
Trade 4	0.33691	7	0.33771	7
Trade 5	0.29084	6	0.29968	6
Trade 6	0.26811	5	0.27008	5
Trade 7	0.24781	7	0.24873	7
Trade 8	0.18622	6	0.18863	5
Trade 9	0.29881	5	0.29882	4
Trade 10	0.22961	5	0.22966	4
Trade 11	0.15493	4	0.15494	4

Table 5.1 A modularity for each single network based on single objective optimization using genetic algorithm. From the table, NC is a number of community

heat map of correlations between objectives (Figure 5.6) and the community structure (Figure5.5). The results are also reflected in the Pareto front plot matrix (Figure 5.4). Our interpretation of these results is as follows: Strong conflicts occur between Q_3 and Q_8, Q_3 and Q_9, Q_1 and Q_8, Q_4 and Q_5. Q_1 and Q_2, Q_1 and Q_3, Q_4 and Q_{11}, Q_4 and Q_{10}, Q_4 and Q_{11}. From the analysis we can, for instance, conclude that for trade-networks of Q_3 and Q_8 the countries cannot be clustered in a way that community structures for both groups of commodities are well represented. On the contrary, for Q_1 and Q_2 there exists a clustering that represents the community structures for both communities very well (See the description of the data). It seems logical that the main agricultural products of a group Q_1 and Q_2 appear to adhere to similar trade community structures, whereas for the very disjoint products in group Q_3 and Q_8, it might have been difficult to predict a priori how their trade networks will overlap.

5.6 • Summary

This chapter showed how to apply many-objective optimization for the analysis of multiplex networks. Different ways on how to analyze the community structure in multilayer networks were introduced, all relying upon data from many-objective optimization. First, we discussed the meaning of the Pareto fronts between modularities

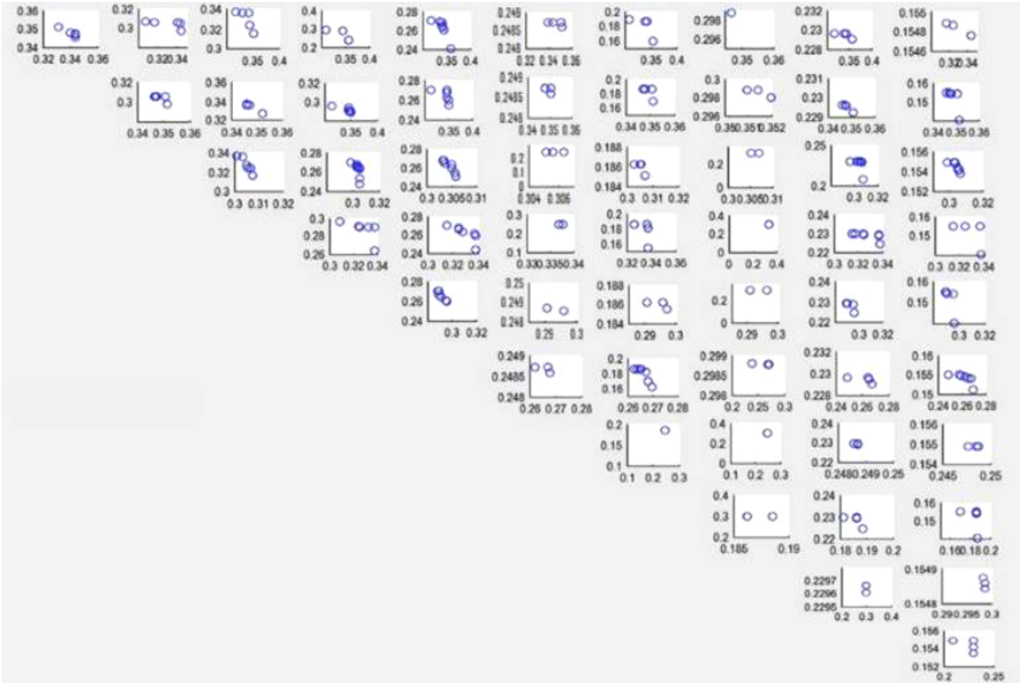

Figure 5.4 Pairwise Pareto Fronts Matrix for Economic Trade Network Analysis

by exact computations of Pareto fronts on three illustrative examples, which represent important boundary cases. Then, on the example of trade networks for commodities, we performed a full analysis. First, we generated data using many-objective optimization, bi-objective optimization (of any pair of layers), and single objective optimization (of any single layer). The results were analyzed using three tools suggested here: Correlation heatmap, the community of objectives analysis, and the Pareto-front plot matrix. These were computed for an economic trade network with 11 groups of commodities. Clearly, a grouping emerges in terms of complementarity and/or in terms of indifference. NSGA-II, SMS-EMOA, and single-objective genetic algorithms can be used as a search engine.

Description of the data

This section describes The description of the data from selected commodities in trade network: Due to space limitations, we will not go in detail about economic trade

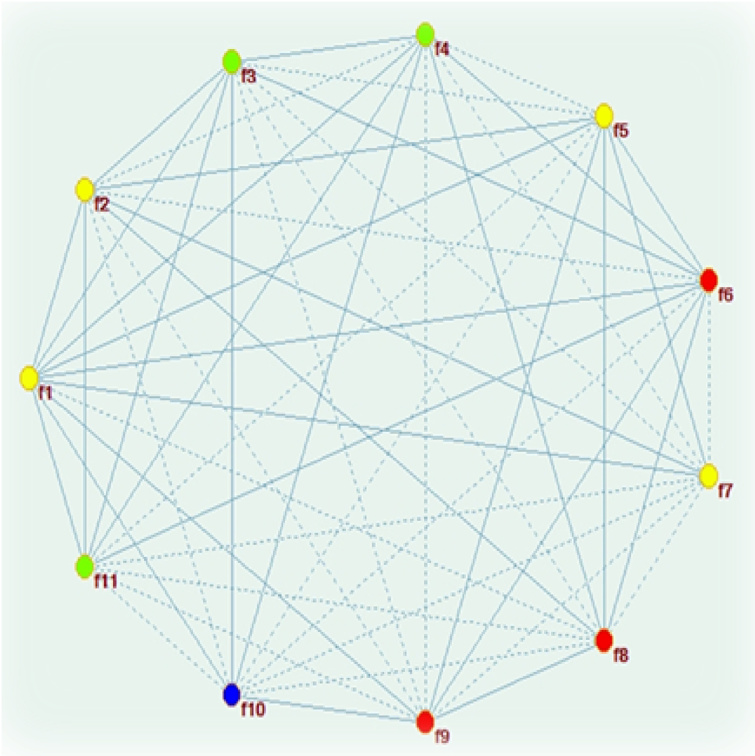

Figure 5.5 community structure for many-objective optimizations of the 11 node trade network

data, but briefly describe those mentioned above:. The data are about economics trading commodities between countries in 2011. Number of country are 207, the numbers of commodities are 96 commodities and grouping in 11 group of commodities described by Q_1 to Q_{11}. For brief explanation, we describe some group of commodities

- Q_1 = Live animals, Meat and edible meat offal, Fish, crustaceans and aquatic invertebrates, Dairy produce; birds eggs; honey and other edible animal products

- Q_2 = Live trees, plants; bulbs, roots; cut flowers and ornamental foliage tea and spices; Edible vegetables and certain roots and tubers; Edible fruit and nuts; Citrus fruit or melon peel; Coffee, tea, mate and spices; Cereals; Milling products; malt; starch; insulin; wheat gluten; Oil seeds and oleaginous fruits; miscellaneous grains, seeds and fruit; Industrial or medicinal plants; straw and fodder

- Q_3 = Lac; gums, resins and other vegetable sap and extracts Vegetable plaiting

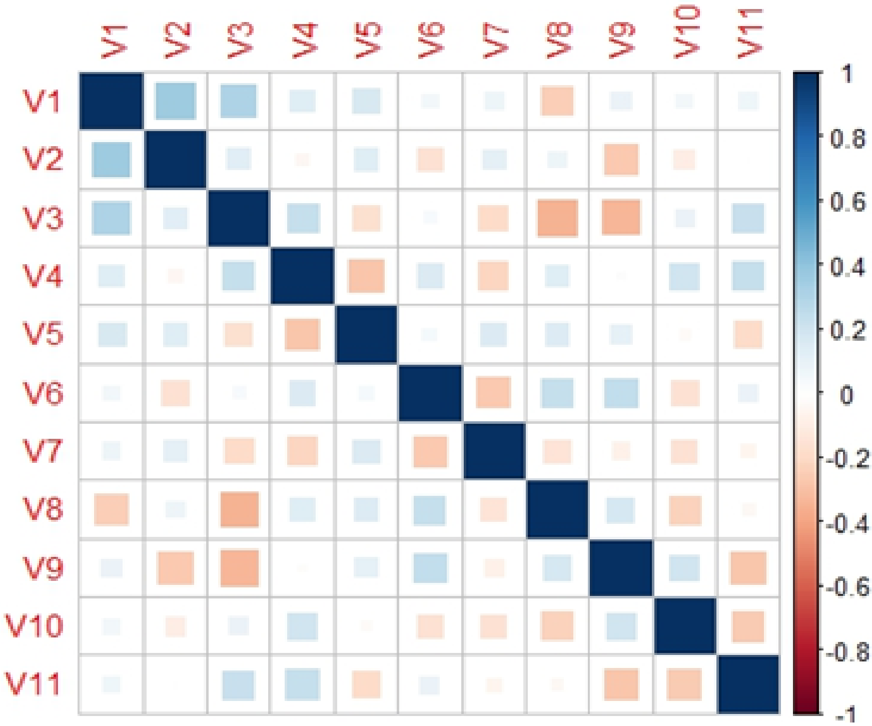

Figure 5.6 Correlation heat map for many-objective optimizations of 11 node trade network

materials and other vegetables products; Animal, vegetable's fats and oils, cleavage product, etc; Edible preparations of meat, fish, crustaceans, mollusc's's or other aquatic invertebrates; sugars and sugar confectionery; Cocoa and cocoa preparations; Preparation of cereals, flour, starch or milk; bakers wares; Preparations of vegetables, fruit, nuts or other plant parts; Miscellaneous edible preparations; Beverages, spirits and vinegar; Food industry residues and waste; prepared animal feed; Tobacco and manufactured tobacco substitutes

- Q_4 = Salt; sulfur; earth and stone; lime and cement plaster, Ores, slag and ash, Mineral fuels, mineral oils and products of their distillation; bitumen substances;mineral wax, Inorganic chemicals; organic or inorganic compounds of precious metals, rare-earth metals, of radioactive elements or of isotopes, Organic chemicals, Pharmaceutical products, Fertilizers, Tanning or dyeing extracts; tannins and derivatives; dyes, pigments and coloring matter; paint and varnish; putty and other mastics; inks, Essential oils and resinoids; perfumery, cosmetic or toilet preparations, Soap; waxes; polish; candles; modeling pastes; dental preparations with basic of plaster, Albuminoidal substances; modified starch; glues; enzymes

- Q_7 = Silk, including yarns, woven, fabric thereof Wool, animal hair, including yarn and woven fabric, Cotton, including yarn, woven fabric thereof, Other vegetable textile fibers; paper yarn and woven fabrics of paper yarn, Man-made filaments, including yarns and woven fabrics, Man-made staple fibers, including yarns and woven fabrics, Wadding, felt and non-wovens; special yarns; twine, cordage, ropes and cables and article thereof.

- Q_{11} = Optical, photographic, cinematographic, measuring, checking, precision, medical or surgical instruments/apparatus; parts and accessories, Clocks and watches and parts thereof, Musical instruments; parts and accessories thereof, Arms and ammunition, parts and accessories thereof, Furniture; bedding, mattresses, cushions, etc.; other lamps and light fitting, illuminated signs and nameplates, prefabricate buildings, Toys, games and sports equipment; parts and accessories, Miscellaneous manufactured articles, Works of art, collectors pieces and antiques.

(See COMTRADE 96 Classification of commodities for 2011 on http://comtrade.un.org/db/mr/rfCon
Moreover we use the following grouping of commodities

- from 1 to 5: Commodity01;
- from 6 to 12: Commodity02
- from 13 to 24: Commodity03
- from 25 to 35: Commodity04
- from 36 to 40: Commodity05
- from 41 to 49: Commodity06
- from 50 to 56: Commodity07
- from 57 to 67: Commodity08

- from 68 to 82: Commodity09
- from 83 to 88: Commodity10
- from 89 to 96: Commodity11

The commodity data we used was from 2011 for all 207 countries.