Instituto de Engenharia de Sistemas e Computadores de Coimbra

Institute of Systems Engineering and Computers

INESC - Coimbra

Carlos Gomes da Silva and João Clímaco

Finding communities in the efficient solutions network

No. 1 2008

ISSN: 1645-2631

Instituto de Engenharia de Sistemas e Computadores de Coimbra

INESC - Coimbra

Rua Antero de Quental, 199; 3000-033 Coimbra; Portugal

www.inescc.pt

Finding communities in the efficient solutions network

Carlos Gomes da Silva[1,3,*] and João Clímaco[2,3]
(1) Escola Superior de Tecnologia e Gestão de Leiria
Morro do Lena, Alto Vieiro, 2401-951 Leiria, Portugal
Phone: +351 244 820 300, Fax: +351 244 820 310
E-mail: cgsilva@estg.ipleiria.pt
(2) Faculdade de Economia, Universidade de Coimbra
Av. Dias da Silva, 165, 3004-512 Coimbra, Portugal
Phone: +351 239 790 500, Fax: +351 239 790 514
(3) INESC-Coimbra
Rua Antero de Quental, 199
3000-033 Coimbra, Portugal
Phone: +351 239 851 040, Fax: +351 239 824 692
E-mail: jclimaco@inescc.pt

January 14, 2008

Abstract

Can we expect to find any transition rule between efficient solutions of {0,1}-multiobjective linear problems? Some highlights on this issue are drawn in the paper. The presence of meaningful communities in the efficient solutions network is investigated and a procedure for interpreting them is proposed. The bi-criteria {0,1}-knapsack problem is used and extensive computational experiences are performed.

Key-Words: Networks, {0,1}-multiobjective problems, community structure, modularity function

1 Introduction

In multiobjective linear programming problems the efficient solutions set is already perfectly characterized, benefiting from the connectedness property, which assures that an efficient solution can be obtained from another one by performing an efficient pivoting (Steuer, 1986). In general, for {0,1}-multiobjective linear problems the above results do not hold. As a result, the computation of the efficient solutions set of {0,1}-multiobjective linear problems has been

*Corresponding author

a difficult task. In general, only small size instances can be exactly solved by the available traditional methods such as dynamic programming or branch-and-bound based algorithms. A considerable effort has been given to the development of strengthened formulations of subproblems used to find efficient solutions, aiming at using efficiently the well-known properties of the linear programming theory. Despite these efforts a large amount of computational time is still required, and medium-large size instances remain unsolved in practice. This difficulty is also present in previously proposed heuristics, since these methods pay a significant price in terms of the accuracy of the approximate solutions.

In some bi-criteria problems (knapsack, transportation,...), empirical observations have revealed that near efficient solutions in the objective space also share many similarities, but a transition rule among efficient solutions remains to be found.

An interesting line of research consists of studying the composition of the set of efficient solutions in order to find some sort of rule or constancy which could be used to propose some alternative resolution procedures. In Gomes da Silva et al. (2007) this line of research is followed. Converting the efficient solutions into a network, where the nodes, n_i, represent efficient solutions, x^i, and the edges the adjacency of the solutions x^i and x^j, in the following sense: if an edge exists between n_i and n_j then, $d(n_i, n_j) = \sum_{k=1}^{n} \left| x_k^i - x_k^j \right| \leq \gamma$, where γ is a constant previously defined, named neighborhood radius, it was found, for example, that the networks of efficient solutions share some properties of the small-world phenomenon: a small shortest path between any pair of nodes, an high clustering coefficient, and a node degree approximated by a power law distribution.

In this paper we go a step ahead and try to explore the existence of communities in the efficient network and, more importantly, try to link them to the objective space of the problem. The communities uncovered may hopefully inspire researchers to build theoretical fundaments for the efficient generation of efficient solutions for {0,1}–multiobjective linear problems.

The search for communities is not a recent area and it is connected with the graph partioning problems and the techniques for discovering groups in social networks. The interest of identifying communities is present in several diverse areas such as computer science, sociology, biochemistry, taxonomy and the WWW (please refer to the works by Newman and Girvan, 2004, and Donetti and Muñoz, 2004, for references concerning these applications).

By community it is meant a subset of nodes highly linked among them, comparing to the links to the rest of the network (Newman and Girvan, 2004 and Donetti and Munõz, 2004), and is important noting that the distance among the nodes is not sufficient to detect communities, i.e., a community is not necessarily composed by the nearest nodes (Wu and Huberman, 2004).

Several methods have been proposed to the NP-complete problem of finding communities in networks (see Newman, 2004, for references and comparisons). Newman and Girvan (2004) is one of the most recent proposals and has the following steps: 1) calculate the betweenness (the number of times that an edge is in the shortest path between all pairs of nodes is one possibility for the measure) scores for all edges in the network; 2) find the edge with the highest score and remove it from network; 3) recalculate betweenness for all remaining edges; 4) repeat from step 2). To assess the strength of the community structure it was proposed the named modularity function, $Q(k)$:

$$Q(k) = \sum_{i=1}^{k} \left(e_{ii} - a_i^2 \right) \tag{1}$$

where e_{ij} is the fraction of all edges in the network that link vertices in community i to vertices in community j, $a_i = \sum_j e_{ij}$ and k is the number of communities.

The modularity function measures the quality of the clustering partition of the network and as Newman and Girvan (2004) showed using several applications, the modularity function enabled the correct detection of communities in the underlying networks and that in real networks, the optimal number of communities is three, $k^* = 3$ and $Q(k^*)$ is usually between 0.3 and 0.7.

Since $Q(k)$ belongs to the range $[0, 1]$ and the greater $Q(k)$ the better it is, the optimal number of communities, k^*, is the one which is associated with the highest value of $Q(k)$. Thus, when $Q(k) = 1$ no edge exists among the k communities and when $Q(k) = 0$ the number of edges among the k communities is not better than random. The modularity function gives a way to find the optimal number of clusters in the network.

Due to the greedy-heuristic nature of the procedure aiming optimizing the modularity function, it may exist some mistakes in the number of the clusters and the assignment of the individuals to the clusters.

White and Smyth (2005) showed that the maximization of the modularity function is related to spectral cluster methods and proposed two new clustering techniques. Donetti and Muñoz (2004) propose a method for finding communities that combine spectral analysis and clustering algorithms to optimize the modularity function. The procedure by Wu and Huberman (2004) is inspired in "voltage drops across networks", and has the ability of discovering the community of a given node.

One of the major differences among the available methods is their computational complexity. The empirical problems where they were applied lead to mixed results. Here we used the procedure by Newman and Girvan (2004), which is implemented in the software NetDraw (Borgatti, S.P., 2002. NetDraw: Graph Visualization Software. Harvard: Analytic Technologies).

The paper is organized as follows. Section 2 presents the used procedure for finding and interpreting communities in the {0,1}-bi-criteria knapack problem . Section 3 presents the computational experiments and Section 4 concludes the paper.

2 Communities in bi-criteria {0,1}-knapsack efficient networks

The experiments that follow are about the efficient solutions networks of the well-known bi-criteria {0,1}-knapsack problem instances, which can be mathematically formulated as:

$$
\begin{aligned}
&\max\ z_1(x_1,...,x_j,...,x_n) = \sum_{j=1}^{n} c_j^1 x_j \\
&\max\ z_2(x_1,...,x_j,...,x_n) = \sum_{j=1}^{n} c_j^2 x_j \\
&s.t.: \\
&\sum_{j=1}^{n} w_j x_j \leq W \\
&x_j \in \{0,1\}, j = 1,...,n
\end{aligned} \tag{2}
$$

where c_j^i represents the value of item j on criterion $i, i = 1, 2, x_j = 1$ if item j $(j = 1, ..., n)$ is included in the knapsack and $x_j = 0$ otherwise, w_j is the weight of item j and W is the overall knapsack capacity.

For the set of efficient solutions a connected network was built by means of adapting the neighborhood radius γ (see Introduction section).

Given a connected network, firstly we start by computing the optimal number of communities in the network; then, we investigate the composition of the communities and its relation with the objective space of the problem.

As the considered networks are artificial, the interpretation of the communities is not as straightforward as in real networks. We need to have some hypothesis to interpret the communities. And we consider the following one:

Hypothesis: *The communities are organized according to an order, ψ, of the efficient solutions in the objective space.*

The above hypothesis depends on the chosen order and several orders can be defined. The most straightforward is the Euclidean. Other possibilities are the cardinality of the efficient solutions and the sum of the criteria values.

To test the stated hypothesis, i.e., in order to evaluate and quantify the association between a given order, ψ, and community orders we proceed as described below:

1) The communities are built using the procedure by Newman and Girvan (2004) and are named, $C^1, ..., C^j, ..., C^{k^*}$, such that, $\min\left\{\psi\left(x^i\right) : x^i \in C^1\right\} \leq ... \leq \min\left\{\psi\left(x^i\right) : x^i \in C^j\right\}$ $\leq ... \leq \min\left\{\psi\left(x^i\right) : x^i \in C^{k^*}\right\}$.

   Within each community, the solutions are also sorted according to non-decreasing values of $\psi(x)$ - community order.

2) For each solution, x^i, are defined two parameters, u_i and v_i. Parameter u_i gives the order of x^i using ψ and v_i gives the order of x^i using the community order.

3) The coefficient of linear correlation between u and v, r_{uv}, is computed, as the measure of association between the ψ and community orders:

$$r_{uv} = \frac{\sum_{i=1}^{n} (u_i - \overline{u})(v_i - \overline{v})}{\sqrt{\sum_{i=1}^{n} (u_i - \overline{u}) \sum_{i=1}^{n} (v_i - \overline{v})}} \tag{3}$$

The coefficient r_{uv} is between [-1,1] and $r_{uv} = 1$ or $r_{uv} = -1$ gives an exact linear relation between u and v.

2.1 An illustration using a random instance

Let us consider a random instance of the bi-criteria {0,1}-knapsack problem with 50 variables/items. All the coefficients were randomly generated from the range [1,100] and the knapsack capacity was set at 50% of the total weight coefficients.

This instance has a total of 25 efficient solutions. The network induced by these solutions, which is presented in Figure 1, is connected when $\gamma = 3$. Thus, in the network each node represents an efficient solution and any two nodes, n_i, n_j are linked if $d(n_i, n_j) \leq 3$.

From the modularity function (1) presented in Figure 2, it can be seen that the optimal number of communities is 3 $(k^* = 3)$, which is associated to the value of k corresponding to the maximum value of $Q(k)$ $(Q(k^*) = 0.381)$.

In the network showed in Figure 1, the nodes of each community are represented by the same mark. It is interesting to observe that these communities are also very well organized in the objective space of the problem. As shown in Figure 3 the communities (C_1, C_2, C_3) are, in this case, composed of a given number of nearest solutions, according to the Euclidean distance, in the objective space. This result gives a direct interpretation of the found communities. This shows that adjacent solutions in the objective space are more densely linked than non-adjacent solutions, revealing a similar neighborhood structure in the decision space, since the network was built based on the Hamming distance between solutions in the decision space

It is easy to see that r_{uv} is equal to 1, revealing a perfect linear association between the two orders: euclidean and community.

3 Computational experiments

The example of section 2.1 revealed that when the communities were converted into the objective space, the solutions of the same community are the nearest neighbors, showing that the geometric placement of the efficient solutions in the objective space is highly connected with the structure of the underlying efficient network. And considering the way the networks were built, one can say that the nearest solutions in the objective space are also closer in the decision space (the number of connection is higher).

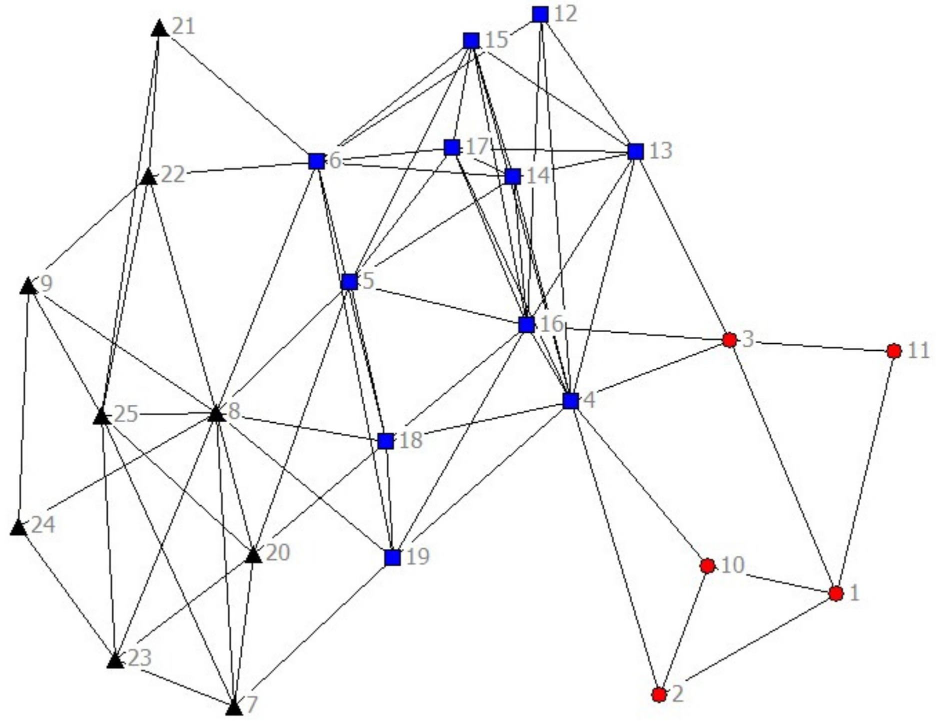

Figure 1: Efficient network and communities aggregation

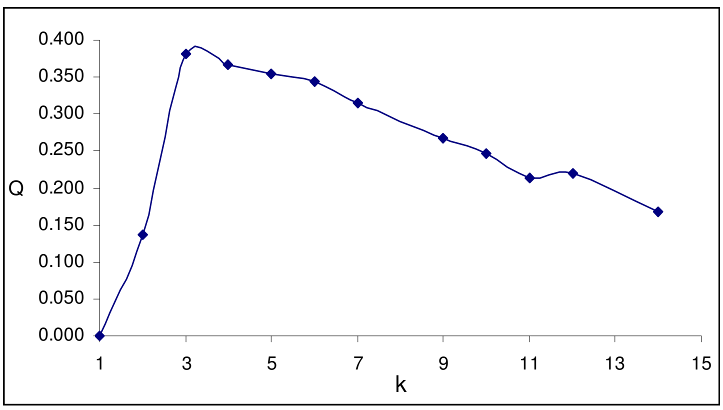

Figure 2: Modularity function

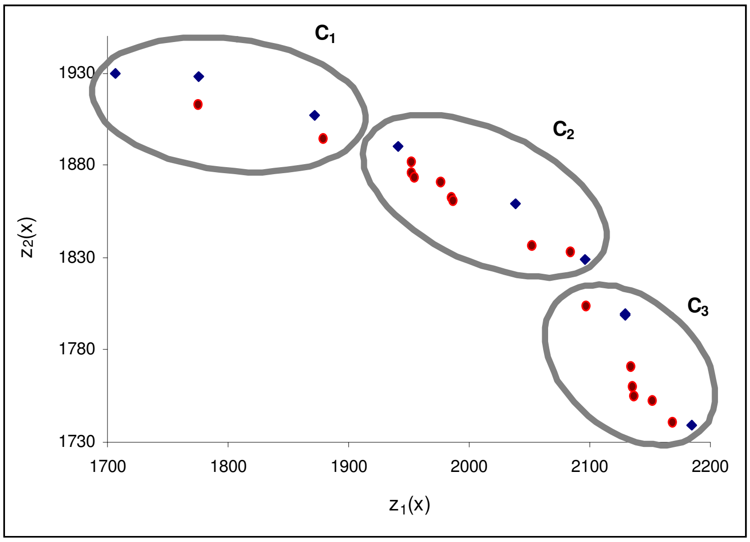

Figure 3: Pareto frontier and communities association

In order to test the hypothesis stated in Section 2, concerning the presence of possible consistent governing rules in general instances of the bi-criteria {0,1}-knapsack problem, we used 12 randomly generated instances from the uniform distribution within the range [1,100], $U(1,100)$, of the following three types:

Type 1: $c_j^1, c_j^2, w_j \sim U(1,100), j = 1, ..., n$ (uncorrelated instances)

Type 2: $c_j^1, c_j^2 \sim U(1,100), w_j = 100, j = 1, ..., n$ (uncorrelated criteria functions and constant weight)

Type 3: $c_j^1, w_j \sim U(1,100), c_j^2 = 101 - c_j^1, j = 1, ..., n$ (uncorrelated criteria and weight functions and strongly correlated criteria functions)

The number of variables was set to 100 ($n = 100$) for Type 1 and 2 instances, and 27 ($n = 27$) for Type 3 instances. The obtained networks are described in Table 1.

Efficient solutions of Type 2 instances, have the particularity of having the same cardinality, $|x^i|$, i.e., the same number of included items. In Type 3 instances, the sum of the criteria values of efficient solutions is proportional to its cardinality. Consequently, the decision space is organized by lines such that $z_1(x) + z_2(x) = k|x|$. This property will be used below to validate the community interpretation. Type 3 instances are characterized by having a considerably greater number of efficient solutions and are much more difficult to solve, comparing to Type 1 and 2 instances. The efficient solutions of Type 1 instances do not have any *a priori* known property, namely in terms of the cardinality and the criteria values.

Three different orders are considered: euclidean, cardinality and the sum of the criteria values. As efficient solutions of Type 2 instances have the same cardinality, the cardinality order

	Type 1		Type 2		Type 3	
Instance	Nodes	Edges	Nodes	Edges	Nodes	Edges
1	113	1043	277	1525	93	458
2	148	877	263	1544	135	994
3	137	727	538	3853	139	808
4	73	369	295	1674	206	1854
5	141	706	312	1833	211	1604
6	138	712	331	1894	235	1974
7	73	307	370	2363	264	1946
8	142	671	380	2395	48	196
9	133	1391	296	1634	416	4068
10	156	727	296	1634	135	1170
11	122	662	330	2068	368	3071
12	177	944	242	1371	182	1694

Table 1: Networks structure

is not meaningful. For Type 3 instances the cardinality order is equivalent to the sum of the criteria order.

3.1 Modularity function and the optimal number of communities

Figures 4, 5 and 6 show the modularity function of the 12 used instances of each type. As can be seen the Type 2 instances have the lowest dispersion of Q coefficient and Type 3 instances the highest one. Besides that, Type 3 instances have the lowest Q values.

The optimal number of communities, according to the procedure by Newman and Girvan (2004) are presented in Table 2.

The networks related to Type 1 instances can be divided into a smaller number of communities. On average, Type 1 instances has 6 communities, Type 2, 7 and Type 3 a considerably higher value: 11.

The distribution of the nodes, in percentage values, of each instance type, among the diverse communities, sorted according non-increasing relative values, is presented in Tables 3, 4 and 5. In general, there are three communities containing the significant part of the nodes. There are communities with only one node (communities 8 and the following 2 of instance 8, are examples), showing the presence of weakly connected nodes. The presence of relatively unimportant communities is particularly present in Type 3 instances.

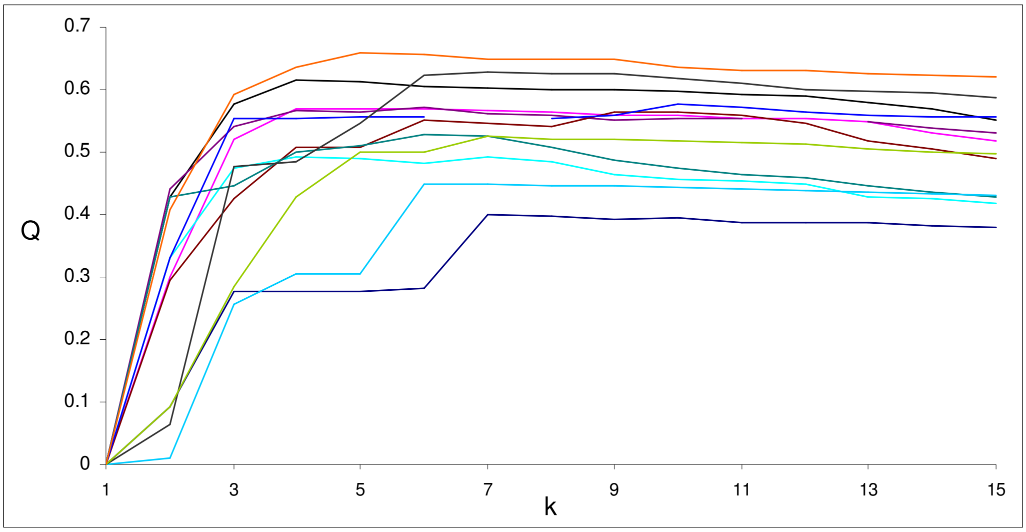

Figure 4: Modularity functions - Type 1 instances

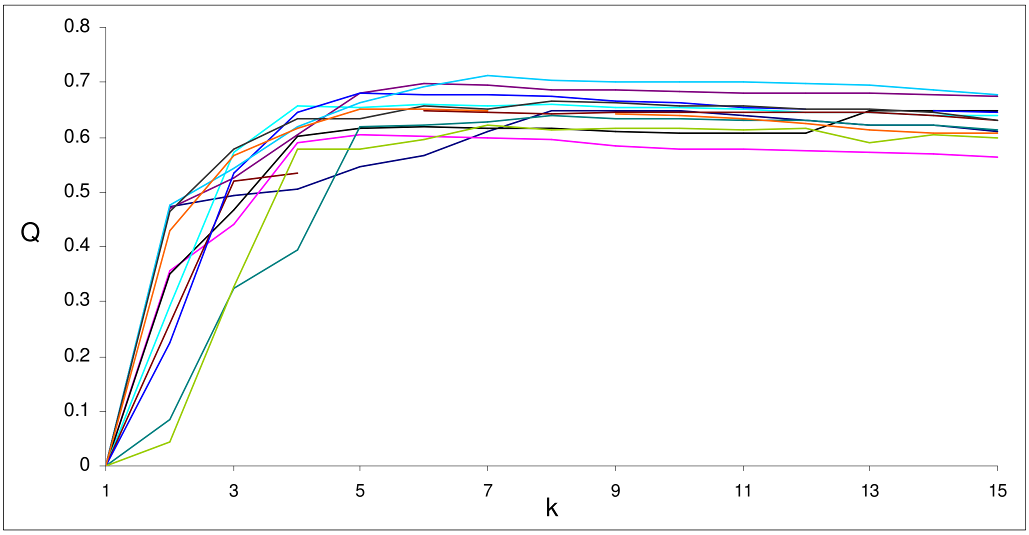

Figure 5: Modularity functions - Type 2 instances

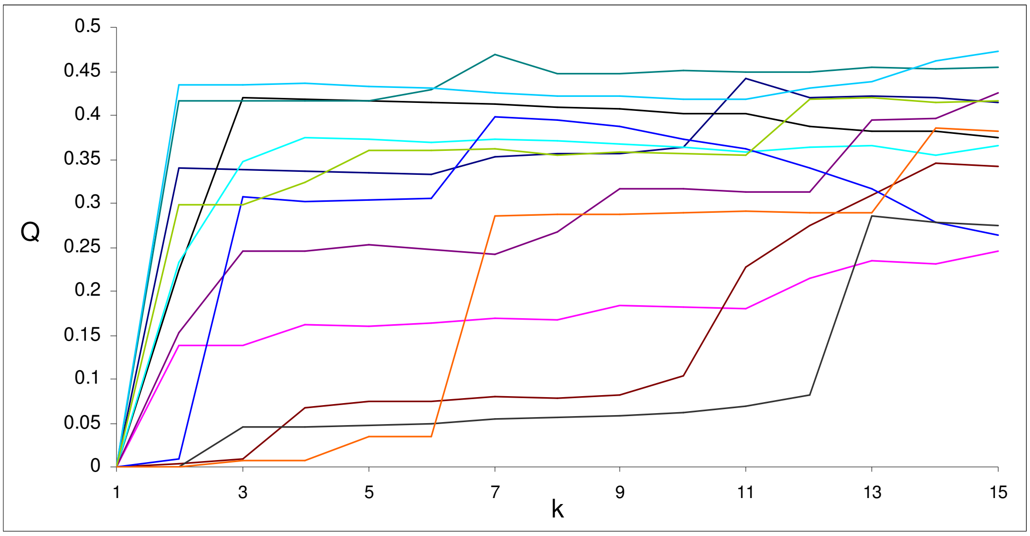

Figure 6: Modularity functions - Type 3 instances

	Type 1		Type 2		Type 3	
Instance	k^*	$Q(k^*)$	k^*	$Q(k^*)$	k^*	$Q(k^*)$
1	7	0.40	9	0.65	11	0.44
2	4	0.57	5	0.60	16	0.27
3	4	0.62	13	0.65	3	0.42
4	4	0.49	8	0.66	4	0.37
5	6	0.57	6	0.70	15	0.43
6	9	0.57	5	0.65	14	0.35
7	6	0.53	8	0.64	7	0.47
8	10	0.58	5	0.68	7	0.40
9	6	0.45	7	0.71	15	0.47
10	7	0.63	8	0.67	17	0.31
11	7	0.53	7	0.62	13	0.42
12	5	0.66	5	0.65	14	0.39
Average	6.3	0.55	7.2	0.66	11.3	0.39

Table 2: Optimal number of communities by instance type

	Community								
Instance	1	2	3	4	5	6	7	8	9
1	46.9	21.2	18.6	10.6	0.9	0.9	0.8		
2	29.1	25.7	23.6	21.6					
3	30.7	27.7	24.8	16.8					
4	34.2	30.1	26.0	9.6					
5	31.9	27.7	20.6	14.9	4.3	0.7			
6	26.8	16.7	14.5	13.0	11.6	7.2	7.2	2.2	0.7
7	28.8	21.9	15.1	15.1	12.3	6.8			
8	28.9	27.5	20.4	8.5	7.0	3.5	2.1	(0.7)	
9	38.3	24.1	23.3	9.8	3.8	0.8			
10	23.1	22.4	20.5	19.9	6.4	4.5	3.2		
11	32.0	16.4	15.6	12.3	11.5	11.5	0.8		
12	26.0	22.6	20.9	(15.3)					

Table 3: Percentual composition of communities - Type 1 instances

	Community										
Instance	1	2	3	4	5	6	7	8	9	10	11
1	25.3	22.0	14.8	9.4	9.4	5.8	5.4	4.3	3.6		
2	33.5	25.1	16.0	14.1	11.4						
3	21.7	16.9	11.9	11.2	9.3	8.6	8.4	6.3	4.6	0.6	(0.2)
4	23.4	21.7	16.9	16.6	11.5	4.4	3.7	1.7			
5	22.4	22.1	21.2	17.0	9.3	8.0					
6	28.7	22.7	20.2	16.6	11.8						
7	29.2	26.5	15.4	7.8	7.6	5.4	4.3	3.8			
8	29.5	21.3	16.6	16.6	16.1						
9	19.9	16.6	16.6	14.5	11.8	10.8	9.8				
10	22.6	22.0	19.9	11.8	11.1	9.8	2.0	0.7			
11	26.1	23.9	22.7	11.5	10.0	5.2	0.6				
12	29.3	21.9	19.4	17.8	11.6						

Table 4: Percentual composition of communities - Type 2 instances

	Community													
Instance	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	23.7	21.5	19.4	18.3	6.5	3.2	3.2	(1.1)						
2	36.3	30.4	14.8	4.4	2.2	2.2	1.5	1.5	(0.7)					
3	46.0	30.9	23.0											
4	38.8	30.6	28.6	1.9										
5	19.4	18.5	15.2	15.2	9.5	8.1	5.2	4.7	1.4	(0.5)				
6	21.7	21.7	18.7	10.2	8.1	7.7	4.7	2.1	1.7	1.3	0.9	(0.4)		
7	42.0	24.2	21.2	11.0	0.8	(0.4)								
8	29.2	29.2	22.9	8.3	4.2	4.2	2.1							
9	24.8	24.8	14.7	7.9	5.5	4.8	4.1	4.1	3.4	2.2	1.4	1.2	0.7	(0.2)
10	33.3	27.4	17.0	8.1	2.2	1.5	1.5	1.5	(0.7)					
11	35.3	35.1	10.6	8.2	5.2	2.4	1.1	0.5	0.5	(0.3)				
12	32.4	29.1	25.8	4.4	1.6	1.1	1.1	1.1	(0.5)					

Table 5: Percentual composition of communities - Type 3 instances

3.2 Communities and the objective space

The coefficients of linear correlation between orders were computed for each network and are presented in Tables 6, 7 and 8.

For the Euclidean order, the obtained r_{uv} coefficients are considerably high for instances Type 1 and 2, with its value being above 0.9 and in the majority of the cases, above 0.99.

The correlation is not casual. Indeed, if we try to interpret the communities of Type 1 instances, using the cardinality order (Table 7), one obtain smaller coefficients of linear correlation between the two orders.

Figures 7 and 8 show a typical association, for this type of instances, between cardinality and community orders, revealing the differences of adjustment.

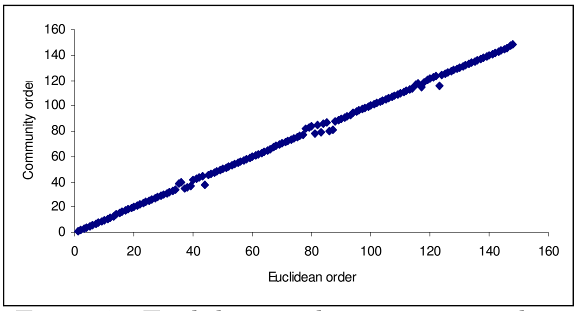

Figure 7: Euclidean and community orders

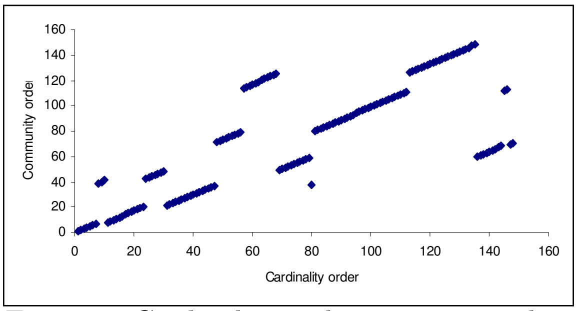

Figure 8: Cardinality and community orders

Instance	Type 1	Type 2	Type 3
1	0.9952	0.9999	0.3484
2	0.9994	0.9996	0.8007
3	0.9994	0.9995	0.3838
4	0.9940	0.9998	0.7996
5	0.9901	0.9993	0.8719
6	0.9931	0.9998	0.7953
7	0.9405	0.9994	0.8389
8	0.9902	0.9999	0.3735
9	0.9870	0.9998	0.8789
10	0.9825	0.9917	0.4230
11	0.9078	0.9972	0.8935
12	0.9993	0.9999	0.5119

Table 6: Coefficient of linear correlation between euclidean and community orders

Instance	Type 1	Type 2	Type 3
1	0.5899	–	0.9789
2	0.7647	–	1.0000
3	0.8268	–	1.0000
4	0.7367	–	1.0000
5	0.7374	–	1.0000
6	0.8551	–	0.9980
7	0.8146	–	0.9999
8	0.7847	–	0.9970
9	0.7220	–	0.9927
10	0.7978	–	1.0000
11	0.9028	–	0.9978
12	0.8394	–	1.0000

Table 7: Coefficient of linear correlation between cardinality and community orders

Instance	Type 1	Type 2	Type 3
1	0.9308	0.9757	0.9789
2	0.9259	0.9580	1.0000
3	0.9728	0.9531	1.0000
4	0.9169	0.9741	1.0000
5	0.8317	0.9100	1.0000
6	0.9059	0.9313	0.9980
7	0.9321	0.9587	0.9999
8	0.9183	0.9688	0.9970
9	0.9222	0.9647	0.9927
10	0.9025	0.6344	1.0000
11	0.9452	0.9415	0.9978
12	0.9577	0.9614	1.0000

Table 8: Coefficient of linear correlation between sum of criteria and community orders

For Type 3 instances the results, using the Euclidean order, were not consistent among the instances. Some have very high coefficients but others reveal no association between the community order and the Euclidean one (Tables 7 and 8). For these instances, the results improved significantly when the cardinality (sum of the criteria) orders are used. Now the association is almost perfect. All the correlation coefficients are above 0.97. The communities seems to capture the structure of the objective space, in concordance with the formulated hypothesis.

The sum of the criteria leads to high correlation coefficients also for Type 1 instances, where no a priori strucuture is known.

4 Conclusions

To the best of our knowledge, this is the first attempt to find communities within efficient solutions, at least with the described methodology. We believe that efficient solutions set of {0,1}-multiobjective problems have some governing rules which, when uncovered, will enable more efficient resolution procedures.

The paper showed that the organization of the connected network of efficient solutions is related with the organization of the objective space. By using different orders (interpretation hypothesis) it was seen that some are better than others, which brings the problem of proposing the most accurate interpretation of the communities. These results also showed that for the same problem, different structures of communities can be found.

The empirical observation that, in general instances, adjacent solutions, in the objective space, have many similar components is already known by the multicriteria combinatorial researchers. Nevertheless, it is also interesting to recognize these results using a different new approach.

The evaluation of the efficient networks of other {0,1}-multiobjective problems, the proposal of new communities interpretation since different problems may have different rules governing the arrangement of its efficient network and the construction of search algorithms inspired on the presented results are interesting future research lines.

References

[1] Amaral, L., Scala, A., Barthélémy, M., Stanley, H. (2000) Classes of small-world networks, *Proceedings of National Academy Science*, 97, 21, pp. 11149-11152.

[2] Donetti, L., Muñoz, M. (2004) Detecting network communities: a new systematic and efficient algorithm, *Journal of Statistical Mechanics: Theory and Experiment*, 1742-5468/04/P10012.

[3] Gomes da Silva, C. (2005) *Finding the set of non-dominated solutions in the bi-criteria {0,1}-knapsack: an approach with meta-heuristics and exact methods*, PhD Dissertation (in Portuguese), Faculty of Economics, University of Coimbra.

[4] Gomes da Silva, C., Clímaco, J., Almeida Filho, A. (2007) The small-world of efficient solutions network, *Research Report* 6, INESC-Coimbra (available from the site www.inescc.pt).

[5] Newman, M. (2004) Detecting community structure in networks, *European Physical Journal B*, 38, pp. 321-330.

[6] Newman, M., Girvan, M. (2004) Finding and evaluating community structure in networks, *Physical Review E*, 69, pp. 026113.1- 026113.15.

[7] Steuer, R. (1986) *Multiple Criteria Optimization, Theory, Computation and Application*, John Wiley & Sons, New York.

[8] White, S., Smyth, P. (2005) A spectral clustering approach to finding communities in graphs, SIAM International Conference on Data Mining, Newport Beach, CA.

[9] Wu, F., Huberman, B.A. (2004) Finding communities in linear time: a physic approach, *The European Physical Journal B*, 38, pp. 331-338.