Centro de Investigação Operacional

A bi-objective hub-and-spoke approach for reconfiguring Web communities

Susana Colaço and Margarida Vaz Pato

CIO – Working Paper 4/2008

A bi-objective hub-and-spoke approach for reconfiguring Web communities

Susana Colaço [a, c,*], Margarida Vaz Pato [b, c, **]

[a] *address: Núcleo de Ciências Matemáticas e Naturais, Escola Superior de Educação, Instituto Politécnico de Santarém, Apartado 131, 2001 – 902 Santarém, Portugal*

**e-mail: susana.colaco@ese.ipsantarem.pt*

[b] *address: Departamento de Matemática, Instituto Superior de Economia e Gestão, Universidade Técnica de Lisboa, Rua do Quelhas, 6, 1200 – 781 Lisboa, Portugal*

[c] *Centro de Investigação Operacional – Faculdade de Ciências, Universidade de Lisboa, Portugal*

***e-mail: mpato@iseg.utl.pt*

Abstract

Web communities in general grow naturally, thus creating unbalanced network structures where a few domains centralise most of the linkups. When one of them breaks down, a significant part of the community might be unable to communicate with the remaining domains. Such a situation is highly inconvenient, as in the case of wishing to pursue distribution policies within the community, or for marketing purposes. In order to reduce the damages of such an occurrence, the Web community should be reconfigured, in such a way that a complete sub-network of main domains – the hubs - is identified and that each of the other domains of the community – the spokes – is doubly linked at least with a hub. This problem can be modellised through a bi-objective optimisation problem, the Web Community Reconfiguring Problem, which will be presented in this paper. A bi-objective mixed binary formulation will also be shown, along with a brief description of GRASP, tabu search and hybrid heuristics which were developed to find feasible solutions to the problem, possibly efficient solutions to the bi-objective problem. A computational experiment is reported, involving comparison of these metaheuristics when applied to several Web communities, obtained by crawling the Web and using epistemic boundaries and to other randomly generated ones. The heuristics revealed excellent quality for the small dimension cases whose efficient solutions were roughly all determined. As for the other

medium and higher dimension instances, the heuristics were successful in building a wide variety of feasible solutions that are candidate efficient solutions. The best behaviour was attained with the GRASP and the GRASP and tabu hybrid search. Comparison of some metrics before and after reconfiguration confirmed that the final structures are more balanced in terms of degree distribution reinforcing the connecting effect imposed by the reconfiguration process.

Keywords: multi-objective; heuristics; Web communities; hub-and-spoke models

This research was supported by PRODEP III, Medida 5 – Acção 5.3 and partially supported by POCTI–ISFL–1–152.

Introduction

Investigation into the structure of the World Wide Web reveals that, notwithstanding its arbitrary growth and its apparent disorganised structure, the Web has an unbalanced structure with a significant hierarchical nature (see for details (Kleinberg and Lawrence, 2001) and (Albert and Barabási, 2002)). Some relevant properties of the Web graph have been studied, such as power law degree distribution; short average path length and high values of clustering coefficient. Such research has underlined the autonomous organisation of this system, along with the constitution of Web communities at local level, with hubs or authorities playing a central role (Flake, Lawrence and Giles, 2000); (Flake, Lawrence, Giles and Coetzee, 2002); (Greco and Zumpano, 2004) and (Kumar, Raghavan, Rajagopalan, Tomkins, 1999).

As an alternative to the "natural" hierarchical structure of a specific Web community, a more balanced structure of the Web community network can be proposed through the Web Community Reconfiguring Problem (WRP, for short). In fact, a structure such as this can be achieved by performing a restructuring process, while respecting a minimum level of initial communication between each pair of community domains. This should be performed while minimising the costs of the action and as, much as possible, balancing the clusters formed by a hub domain and its respective spoke domains.

There are many applications of the problem addressed in this paper, namely:
- long-term preservation and availability of Web contents;
- implementation of distribution policies within the community, such as distribution among the members of a new software tool or a new education programme and planning, organisation and support of Web resources;
- performing marketing campaigns within the Web community;
- production of network indicators, allowing one to compare Web structure across different communities.

The paper proceeds with a presentation of the Web Community Reconfiguring Problem, followed by a section devoted to the formulation of the problem and proof of its NP-hardness. In the subsequent two sections the heuristics are briefly presented and the computational experiments are described. This includes results of the application of the heuristics to a set o instances of WRP real based and others obtained from random generation. The last section concludes the paper with comments.

The Web Community Reconfiguring Problem

A Web community is considered to be a set of Web pages providing resources on a specific topic. Additionally, it could be pages related to a specific topic. For the purpose of modelling, we regard the pages as being aggregated in domains. Domains can also be aggregated in upper level domains (see figure 1).

There is a hyperlink from one domain, say domain *i*, to another domain *j*, if at least one page from domain *i* points to one page of domain *j*. A parameter, called intensity, associated with any hyperlink is defined. This parameter is equal to the total number of hyperlinks connecting the pages of the two domains and sharing the same direction. In the case illustrated in figure 1 the intensity of the link from *i* to *j* is 2. The inverse of this parameter is called weakness. Here, the weakness W_{ij} of the linking up from *i* to *j* is one half.

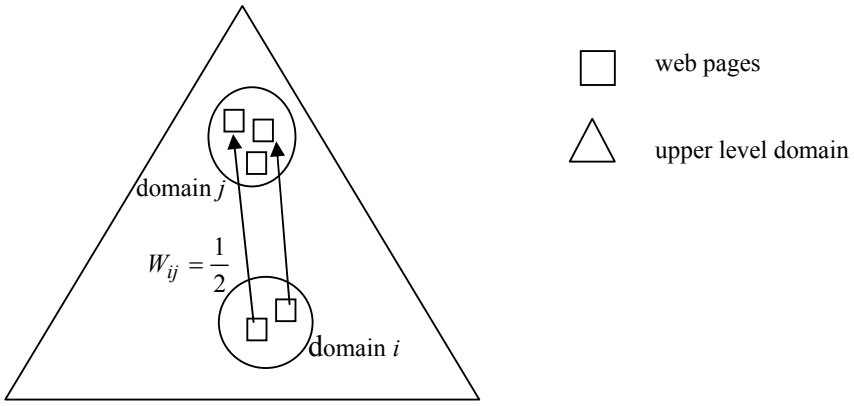

Figure 1 Two lower level domains of a Web community

The big question is: how can a specific Web community be reconfigured to counterbalance its hierarchical structure and, consequently, make it easy to preserve and share information among the elements of the community?

The authors thought that it would be beneficial to create a hub-and-spoke structure, where the domains playing the role of hubs aggregate a set of other domains – which we call spokes. In this reconfigured structure, each spoke is hyperlinked with its respective hub, working in both directions; the sub-network of the hubs is regarded as a complete directed network, that is, each pair of hubs is doubly linked. Moreover, a maximum bound on the weakness of the linkups between each pair of domains is imposed to enforce a minimum level of connection. The set formed by the hub and its spokes is known as a cluster.

In building the hub-and-spoke structure for the Web Community Reconfiguring Problem, two objectives must be born in mind: minimise costs and obtain more balanced clusters.

The reason for minimising costs is related with the minimisation of hubs and new arcs because the creation and maintenance of these structures involve costs. As for the balancing of clusters, in this case it is performed by equilibrating the number of links of the domains per cluster. Clearly the existence of a more balanced, equilibrated structure leads to a more homogeneous diffusion/sharing of information or of resources. Besides

this gain, a more balanced form can significantly improve the planning and organisation of structures, where the contribution of a domain with lower level of linkings is improved by aggregating it with others within an equilibrated cluster, thus creating several more homogeneous structures among the Web community.

The WRP is a hub covering type problem, similar to the one studied by Campbell in 1994 and is also an NP-hard problem as will be proved in the next section. Nevertheless, the WRP has some particular features which distinguish it from the standard hub covering problem: a covering criterion that does not verify the triangular inequality and the minimisation of two contradictory objectives.

There follows a formalisation for the Web Community Reconfiguring Problem.

A bi-objective mixed binary linear formulation

Consider the network $G = (N, A)$ with a node associated with each Web domain and an arc from domain i to domain j, if there is at least a hyperlink from a page of i to a page of j. A binary matrix $[A_{ij}]$ defines the arcs of the network, that is, A_{ij}=1 if arc (i,j) belongs to A, and A_{ij} =0, otherwise. The nodes of set N are characterised by two parameters: the indegree – the number of hyperlinks pointing to a node - and the outdegree – the number of arcs originating from a node; future work could use additional parameters such as flow betweenness, centrality, etc (for a compilation of network indicators see e.g. (Wasserman and Faust, 1994). As for the arcs of A, they are characterised by a single parameter, the weakness represented by parameters W_{ij}.

As mentioned earlier, the goal of the WRP is to redesign the network within a hub-and-spoke structure, by selecting hub nodes and, if necessary, by adding new arcs to the network. One must therefore determine the nodes by specifying/that specify the location of hubs, as well as the set of nodes allocated to each hub.

However, the resulting network must also comply with constraints that impose a maximum bound $\gamma > 0$ on the total weakness. In this way we can assume that the flow of resources within the entire community is facilitated. The weakness of hyperlinks

between hubs is decreased by a pre-determined factor $\alpha(0 \le \alpha \le 1)$, to allow for an improvement in the centrality and connectivity of these domains, presented in the following condition:

$$W_{ik} + \alpha W_{km} + W_{mj} \le \gamma \ . \qquad (1)$$

It is assumed that if $(i, j) \notin A$, then the weakness $W_{ij} = \varphi$, where $\varphi > 1$ is a pre-determined real parameter. Note that, to build a feasible solution one may choose any set of domains to play the role of hubs, provided the weakness constraint is satisfied.

The objectives of network redesigning within WRP are as follows:
- firstly, objective 1, minimisation of costs, assuming that costs are equal to the number of hubs plus the number of new arcs;
- and secondly, objective 2, building of balanced clusters, in keeping with the original node degree parameters, in other words, the cluster balancing goal is achieved by minimising the sum of the maximum degree values throughout the clusters.

Figure 2 illustrates one small Web community, already redesigned with its three hubs, domains *3, 4* and *8*. The pre-existing arcs are represented by black or grey arrows whereas the new ones, created to assure the hub-and-spoke structure, are depicted as doted lines.

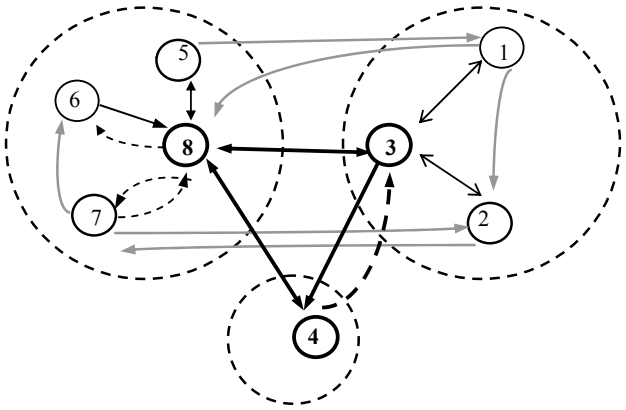

Figure 2 Network representing a restructured Web community

Here, we may see, for instance, what happens if a spoke domain is not initially hyperlinked both to and from its respective hub domain. Then, in the reconfigured Web network, at least one new hyperlink must be created. The cost of the solution illustrated is equal to three hubs plus four new arcs. As for the value of the second objective, 17=8+9, it is given by cluster 3, as its total indegree, before reconfiguration, is 8 (the sum of nodes' indegree in the cluster) and its total outdegree is 9.

The WRP can be formulated within a bi-objective mixed binary linear programming problem that is presented here:

$$Min \quad f_1 = \sum_{k=1}^{n} x_{kk} + \sum_{i=1}^{n} \sum_{(k=1 \wedge k \neq i)}^{n} z_{ik} \tag{2}$$

$$Min \quad f_2 = s + t \tag{3}$$

subject to

$$\sum_{k=1}^{n} x_{ik} = 1 \qquad \forall i \in N \tag{4}$$

$$x_{ik} \leq x_{kk} \qquad \forall i,k \in N \tag{5}$$

$$\sum_{i=1}^{n} x_{ik} Ind_i \le s \qquad \forall k \in N \tag{6}$$

$$\sum_{i=1}^{n} x_{ik} Out_i \le t \qquad \forall k \in N \tag{7}$$

$$W_{ik} x_{ik} \le ri_k \qquad \forall i,k(i \ne k) \in N \tag{8}$$

$$W_{ki} x_{ik} \le ro_k \qquad \forall i,k(i \ne k) \in N \tag{9}$$

$$ri_k + ro_k + \alpha W_{km}\left[(n+1)x_{kk} - n\right] \le \gamma\left[n - (n-1)x_{mm}\right] \qquad \forall k,m(k \ne m) \in N \tag{10}$$

$$2x_{ik} \le (A_{ik} + A_{ki}) + z_{ik}(1 - A_{ik}) + z_{ki}(1 - A_{ki}) \qquad \forall i,k \in N(i \ne k) \tag{11}$$

$$(2x_{kk} - 1) + (2x_{ii} - 1) \le (A_{ik} + A_{ki}) + z_{ik}(1 - A_{ik}) + z_{ki}(1 - A_{ki}) \qquad \forall k,i(k \ne i) \in N \tag{12}$$

$$ri_k, ro_k \ge 0 \qquad \forall k \in N \tag{13}$$

$$x_{ik}, z_{ik} \in \{0,1\}, \qquad \forall i,k \in N \tag{14}$$

$$s,t \ge 0. \tag{15}$$

The variable x_{ik} is equal to 1 if node *i* is assigned to node *k*, otherwise $x_{ik} = 0$; $x_{kk} = 1$ if node *k* is chosen to be a hub location, otherwise $x_{kk} = 0$; $z_{ik} = 1$ if it is necessary to create arc (*i*,*k*), otherwise $z_{ik} = 0$. The real non-negative variables $ri_k, ro_k, \forall k \in N$ represent the cover radius of hub *k* for the in and out-arcs, respectively. Cover radius of a specific cluster *k* represents the maximum weakness of arcs connecting spokes to the respective hub (arcs incident at the hub) or connecting the hub to the respective spokes (arcs starting at the hub). The variables *s* and *t* represent the maximum indegree and outdegree among all the clusters. These two variables are declared to be non-negative real numbers, although they are integer in any optimal solution, due to constraints (6), (7) and minimisation of f_2.

The functions f_1 and f_2 in (2) and (3) represent the two contradictory objectives to be minimised. The first one gives the cost involved in the Web community reconfiguration equal to the total number of new arcs and of clusters. The second one is related with the balancing of clusters and gives the sum of the maximum indegree and outdegree per cluster. Constraints (4) and (5) are usually employed in the single assignment hub-and-

spoke models to ensure the assignment of each spoke to precisely one hub k. In connection with the second objective function, f_2, constraints (6) and (7) guarantee more balanced clusters in relation to the original indegree and outdegree parameters of the respective spoke nodes, whilst constraints (11) and (12) with the first objective function (f_1), impose the new arcs required to link domains in both directions. Constraints (8) to (10) ensure that the total hyperlink weakness connecting node i to node j, via hubs k and m, is not superior to a given bound γ, using the variables ri_k and ro_k. Finally, constraints (13) to (15) specify the domains for the variables.

The proposed formulation (2)-(15) was based on the formulation of (Ernst, Jiang and Krishnamoorthy, 2005) for the hub covering problem using the concept of cover radius for each hub. As an alternative, two other models for the WRP have been studied. The two models were inspired by the work of (Kara and Tansel, 2003) which linearises the model for the hub covering problem initially presented by (Campbell, 1994). However, any of the abovementioned models provided worse results than (2)-(15) in terms of computational running time in the preliminary tests, mainly due to the use of four index variables that significantly enlarges the dimension of the already big instances. Such behavior is in keeping with the experiments found in (Ernst, Jiang and Krishnamoorthy, 2005).

The hub location problem with a non-fixed number of hubs, as is the case of WRP, has received less attention in the literature than the other problems of the class. However, some papers referring to solutions for one such problem were published: in (O'Kelly, 1992), (Abdinnour-Helm, 1998), (Abdinnour-Helm and Venkataramanan, 1998), (Topcuoglu, Corut, Ermis, and Yilmaz, 2005) different heuristics were explored; (Abdinnour-Helm and Venkataramanan, 1998) developed a branch-and-bound algorithm; (Klincewicz, 1996) used a dual algorithm; (Camargo, Miranda and Luna, 2008) used Benders decomposition algorithms for the uncapacitated multiple allocation hub location problem and (Rodríguez-Martín and Salazar-González, 2008) used a branch and cut algorithm and a heuristic approach for a capacitated hub location problem.

Despite the fact that the number of hubs is also a decision variable in WRP, there are other features involved which are not considered in the problems above: the need to obtain balanced clusters and minimise the number of new arcs.

As the WRP is a bi-objective optimisation problem, the two objectives are naturally contradictory: building a low cost structure will lead to highly non-balanced clusters. Hence, a single optimal solution for both objectives does not exist. As is known, the solutions for the bi-objective optimisation problem are the so-called efficient solutions. At the objectives space, they correspond to the non-dominated points; all the non-dominated points define the Pareto frontier of the problem (Ehrgott, 2005). As a result, to solve WRP one should determine all the efficient solutions, thus defining the Pareto frontier.

The WRP is an NP-hard problem. In fact, in (Colaço and Pato, 2006) it is proved that a version of the problem with a single objective function which is a weighted sum (with fixed weights, λ_1 and λ_2) of the two objective functions f_1 and f_2 is NP-hard. On the other hand, from multi-objective integer optimisation theory, optimization of the single objective weighted sum version of WRP, fixing the weights λ_1 and λ_2, produces a supported efficient solution for WRP. By changing the weights, one can generate all the supported efficient solutions, but other efficient solutions of WRP, unattainable by using that methodology, can exist. So, one may conclude that the WRP is at least as difficult as its single objective weighted sum version, and is therefore also an NP-hard problem.

Moreover, the real instances of WRP are, as a rule, of a very high dimension. Bearing this in mind, GRASP, tabu search and hybrid bi-objective metaheuristics, already developed for a single objective version of WRP, were explored to tackle the bi-objective nature of WRP. Of course, for solutions obtained by such non-exact approaches efficiency in the bi-objective context is not guaranteed. They generate simply candidate efficient solutions.

Heuristics for the WRP

Three metaheuristics were used to obtain an approximation to the Pareto frontier: a GRASP, a tabu search and a hybrid of the two.

All these search procedures start from a Greedy-Randomised Constructive heuristic, which is an important feature of the three metaheuristics. A brief description of this building heuristic is provided in figure 3. According to the pseudo-code, for a fixed number of p hubs, the constructive procedure follows two steps: step 1, performing a randomised choice of hubs and step 2, devoted to a randomised assignment of spokes. This constructive procedure stops when a feasible solution is achieved or a maximum number of iterations is attained (*maxiterconstr*). See, for further details, e.g. (Ebery, Krishnamoorthy, Ernst and Boland, 2000) and (Klincewicz, 1991) for the capacitated multiple allocation hub location problem and p-hub location problem respectively, and (Colaço and Pato, 2006) for the single objective weighted sum version of WRP.

```
procedure Greedy_Randomised Constructive ( p, α1, α2, solution_k)
        step 1. randomised choice of hubs
                repeat until p hubs have been chosen
                        build RCLH(α1) list based on free nodes' degree
                        select, at random, a hub from RCLH(α1)
                end
        step 2. randomised assignment of spokes
                niter=0
                while there is a node to be assigned and niter<maxiterconstr
                        for each node not yet assigned
                                compute incremental costs
                                choose the best feasible hub candidate
                        end
                        build RCL(α2) list
                        randomly select a spoke node from RCL(α2)
                        assign the spoke node to its hub candidate
                        niter=niter+1
                end
end Greedy_Randomised Constructive
```

Figure 3 Constructive procedure

Note that this is a single objective heuristic adapted to the bi-objective optimisation as, in step 2, the computing of the incremental costs is based on the weighted sum of the two objectives, plus the weakness values.

This heuristic can be used with an exclusive greedy component, when α_1=0 and α_2=0, or totally randomly when α_1=1 e α_2=1. A semi-random version was used as input for the GRASP and the hybrid heuristic, whereas a greedy version was adopted in the tabu search.

As for the metaheuristics, denoted by Grasp, Tabu and Hybrid, they were all local searches adapted from heuristics already developed for the single objective weighted sum version of WRP. Hence, each (bi-objective) metaheuristic procedure is based on a sequential application of the single objective metaheuristic, thus generating a set of feasible solutions for WRP, one for each pair of parameters λ_1 and λ_2. For the set of all these solutions, the respective points in the objectives space (f_1,f_2) are identified. From those points, the non-dominated ones are calculated. These are the candidates to be non-dominated for the bi-objective optimisation problem WRP and all define the so-called non-exact Pareto frontier generated by the heuristic, hence defining an approximation to the Pareto frontier of WRP.

Now, let us present a synthesis of the main characteristics of the searches performed within these metaheuristics, while taking into account a fixed choice of the parameters λ_1 and λ_2.

As mentioned above, a critical decision in the WRP involves the number of hubs, p. For this reason, the metaheuristics for a specific choice of the parameters λ_1 and λ_2 will be running for an appropriate range of $p \in [kmin, kmax]$ values and returning at the end the best solution found.

As is known, a GRASP is a multi-start metaheuristic with two phases per iteration: a construction phase and a local search phase (see, for details of GRASP heuristics in general, (Resende and Ribeiro, 2002)). Following the standard procedure, in the first

phase of our Grasp for the WRP, the Greedy_ Randomised Constructive procedure is used to obtain a feasible solution; in the second phase there follows a local search with shift movements chosen in the neighbourhood of that solution, until a local optimum for the weighted sum of the two objective functions is found (see for details (Colaço and Pato, 2006)).

Tabu search considers solutions and/or movements as a tabu, depending on the memory that keeps the solutions visited in the previous iterations, hence driving the search process to unexplored space regions (see, for details of general tabu search procedures, (Glover, 1989) and (Glover and Laguna, 1997)). The Tabu algorithm implemented for the WRP was inspired by the work of (Skorin-Kapov and Skorin-Kapov, 1994) with two phases: assignment and location phases searching, respectively, within the spoke shift-swap and hub location neighbourhoods. It is enhanced with strategic oscillation and some diversification strategies. A more detailed picture of the main components of this metaheuristic is found in (Colaço and Pato, 2006). This local search also uses a weighted sum of the two objectives to evaluate a solution or movement, as happens in the Grasp.

Finally, a combination of the two metaheuristics Grasp and Tabu, designated as Hybrid, was proposed.

Computational results

Computational tests were conducted for six real epistemic Web communities and another six randomly generated Web communities. The real Web communities were obtained using keyword search in several search-engines, as was the case of the Mathematics Education Web community in Portugal (Mat20, Mat 30 and Mat53) or obtained from an international Project at the Oxford Internet Institute, Climate Change Web community (Clim), HIV Web community (Hiv) and Poverty Web community (Pov) - see (Caldas, Schroeder, Mesch, and Dutton, 2006). The arcs' intensities were calculated using the «Galilei» software by (Caldas, 2005). To generate the other six Web communities the authors used the network analysis software «Pajek» due to (Batagelj and Mrvar, 1998)

All the heuristic algorithms were coded in C and the programmes, as well as the standard integer optimiser, ran on a PC Pentium IV, 512 Mb RAM, 2.6 GHz.

The exact Pareto frontier was determined for only one instance (Mat20 with 20 domains). Nevertheless, even with such a small instance, the time required to attain the exact Pareto frontier was significant, which makes it impossible to use it for larger instances.

Figures 4, 5 and 6 below show all the non-dominated points in the objectives space, corresponding to the supported and non-supported efficient solutions of the WRP instance. They were obtained by using the constraints method for bi-objective optimisation, see (Ross and Soland, 1980), applied to the formulation defined by (2)-(15) and running the CPLEX Optimizer version 8.0.

Figures 4, 5 and 6 also show approximations to the Pareto frontier for instance Mat20, obtained respectively by each of the metaheuristics: Grasp, Tabu and Hybrid.

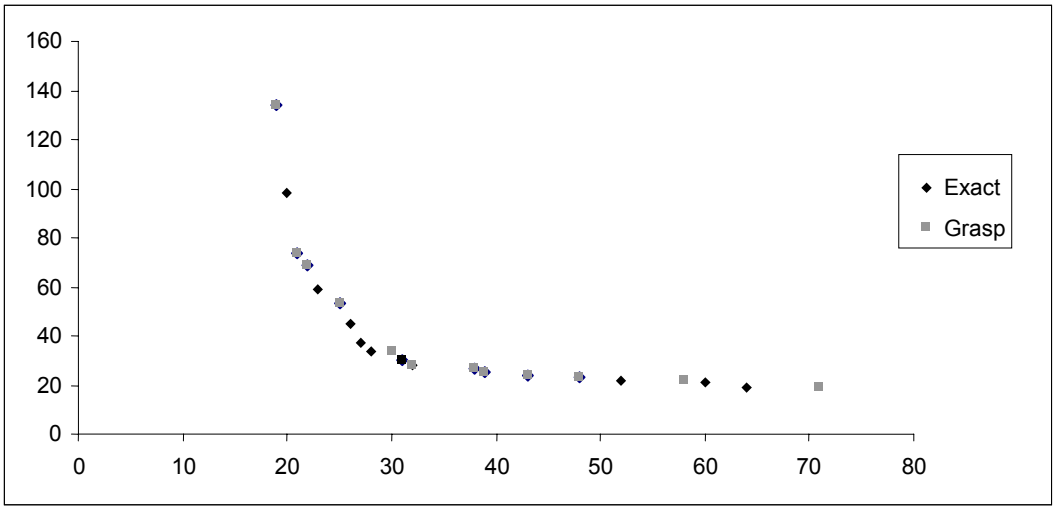

Figure 4 Pareto Frontier and Grasp approximation for Mat20

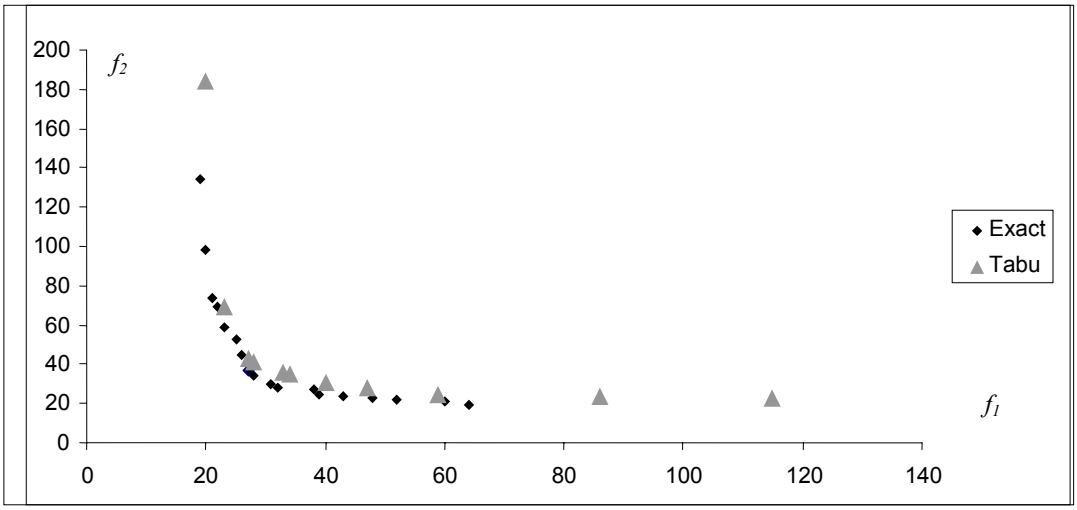

Figure 5 Pareto frontier and Tabu approximation for Mat20

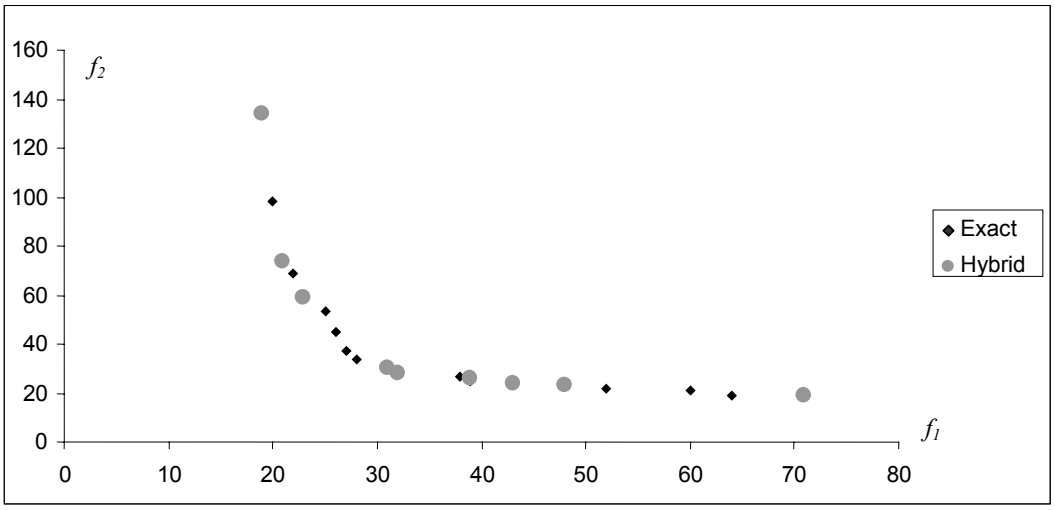

Figure 6 Pareto frontier and Hybrid approximation for Mat20

As may be seen in these graphs and later confirmed in Table 1, the three heuristics attained good results and, what is particularly noteworthy is the excellent approximations to the Pareto frontier given by the Grasp procedure.

Metrics taken from (Collette and Siarry, 2003 and 2005) were calculated for the purpose of analysing the behaviour of the heuristics, when the Pareto frontier is known or not:

- the global distance metric (*GDM*) represents the absolute distance between the Pareto frontier and the non-exact frontier

$$GDM = \frac{\left(\sum_{i=1}^{N} d_i^2\right)}{N}, \tag{16}$$

where d_i represents distance between the i^{th} point in the non-exact frontier and the closest point at the Pareto frontier and N is the total number of points in the non-exact frontier;

- the spacing metric, with two versions (SM_1 and SM_2) provides a way of measuring how points defining the non-exact frontier are distributed in the space of objectives - plane (f_1, f_2)

$$SM_1 = \left[\frac{1}{N-1} \times \sum_{i=1}^{N-1} \left(\overline{d}_1 - d_{1i}\right)^2\right]^{\frac{1}{2}}, \tag{17}$$

$$\text{with } d_{1i} = \min_j \left(\left|f_1(\vec{x}_i) - f_1(\vec{x}_j)\right| + \left|f_2(\vec{x}_i) - f_2(\vec{x}_j)\right|\right) \qquad \forall i, j \in \{1,\ldots,N\}\ (i \neq j)$$

$$SM_2 = \left[\frac{1}{N-1} \times \sum_{i=1}^{N} \left(1 - \frac{d_{2i}}{\overline{d}_2}\right)^2\right]^{\frac{1}{2}}, \tag{18}$$

$$\text{with } d_{2i} = \sqrt{\left(f_1(\vec{x}_i) - f_1(\overrightarrow{x_i + 1})\right)^2 + \left(f_2(\vec{x_j}) - f_2(\overrightarrow{x_j + 1})\right)^2} \qquad \forall i, j \in \{1,\ldots,N\}\ (i \neq j)$$

where d_{li}, for $l \in \{1, 2\}$, represents the distance from the i^{th} point of the non-exact frontier to its closest point, whereas $\overline{d}_l$, for $l \in \{1, 2\}$, is the average of d_{li} values for all N points of this non-exact frontier;

- the Pareto ratio metric (*PRM*), calculated only when the exact Pareto frontier is known, is given by the quotient between the total number of points in the non-exact frontier and the total number of points in the Pareto frontier.

Table 1 shows results for instance Mat20 relative to the three non-exact methods, Grasp, Tabu and Hybrid, as well as for the exact method for WRP (method indicated in column (1)) using the metrics *GDM*, SM_1, SM_2 and *PRM* (columns (2) to (5)). In column (6) the total number of non-dominated points in each frontier is given and the computational time for generating all the points of the exact or non-exact frontier is presented in column (7). Best results are formatted in bold in this table and also in Tables 2 and 3.

(1) **method**	**(2)** ***GDM***	**(3)** SM_1	**(4)** SM_2	**(5)** ***PRM***	**(6)** ***Card***	**(7)** **total CPU (sec.)**
Pareto frontier	0.0	8.8	1.2	1.0	18	306 044.3
non-exact frontier by Grasp	2.4	16.1	1.5	0.5	**13**	**15.1**
non-exact frontier by Tabu	6.9	33.5	1.3	**0.6**	11	524.0
non-exact frontier by Hybrid	**0.6**	**6.3**	**0.9**	0.5	9	236.4

Table 1 Comparing the Pareto and the non-exact frontiers for Mat20 Web community.

In this table, the Grasp heuristic provides the best results in terms of *Card,* as well as computational time. Nevertheless, the Hybrid heuristic gives better results for *GDM*, SM_1 and SM_2, and it is worse than Grasp in terms of computational time, but even so, better than Tabu. In fact, the latter heuristic presents the worst results with this instance related to CPU time, *GDM*, SM_1 and SM_2.

Despite the fact that these heuristics for each choice of parameters λ_1 and λ_2 minimise a weighted sum of two objectives - for this reason, they were designed to produce solutions that are candidate to be supported efficient solutions - they also attained in some cases non-supported efficient solutions. For example, in instance Mat20, 5 of the 13 non-dominated from Grasp in fact correspond to non-dominated and non-supported points for the bi-objective WRP.

As it was not possible to find the Pareto frontier for the instances with more than 20 domains studied in this paper, Tables 2, 3, 4 and 5 display only the results referring to the non-exact frontiers obtained by the heuristic methods.

In Tables 2 and 3, the columns (5) to (7) present some indicators calculated from the real communities and the randomly generated ones respectively. In addition, the nadir and utopic candidate points are also presented in columns (3) and (4). Again column (8) contains the computing time used to generate all points of the non-exact Pareto frontier.

(1) Web community/nº of domains	**(2) method**	**(3) nadir candidate**	**(4) utopic candidate**	**(5) *Card***	**(6) SM_1**	**(7) SM_2**	**(8) total CPU (sec.)**
Mat30/30	Grasp	**(76,122)**	(28,41)	8	**5.6**	**0.4**	**23.9**
	Tabu	(84,138)	(33,30)	**18**	6.4	0.7	3304.2
	Hybrid	(123,122)	**(28,26)**	14	5.9	0.7	1197.7
Mat53/53	Grasp	**(166,210)**	(55,113)	**12**	**7.1**	**0.6**	**63.6**
	Tabu	(164,295)	(59,52)	**12**	28.4	1.1	15834.0
	Hybrid	(203,231)	**(55,48)**	**12**	13.1	0.9	10880.2
Clim/68	Grasp	**(304,162)**	(124,81)	**20**	**6.5**	1.0	**70.6**
	Tabu	(313,162)	**(124,76)**	9	12.6	**0.5**	5066.4
	Hybrid	(316,162)	(124,77)	14	10.1	0.7	3903.3
Pov/59	Grasp	(280,152)	**(105,76)**	**18**	**9.2**	1.0	**60.2**
	Tabu	**(282,142)**	(105,82)	8	22.7	**0.9**	3368.6
	Hybrid	(280,152)	**(105,76)**	14	15.2	1.0	3199.9
Hiv/55	Grasp	**(285,107)**	**(100,50)**	**15**	**8.5**	0.9	**54.1**
	Tabu	(282,114)	(101,56)	7	35.6	1.1	7206.3
	Hybrid	(286,107)	**(100,50)**	12	9.7	**0.4**	2945.9

Table 2 Metrics calculated from the non-exact frontiers for the real Web communities

(1) Web community/nº of domains	(2) method	(3) nadir candidate	(4) utopic candidate	(5) *Card*	(6) SM_1	(7) SM_2	(8) total CPU (sec.)
Rnd20/20	Grasp	**(76,98)**	(20,23)	**11**	**6.4**	**0.6**	**13.4**
	Tabu	(76,130)	(21,22)	5	41.1	1.4	320.1
	Hybrid	(71,134)	**(19,19)**	9	19.1	1.3	236.6
Rnd30/30	Grasp	(178,273)	(37,31)	**18**	**8.5**	**0.7**	**29.5**
	Tabu	**(178,259)**	**(35,29)**	9	16.4	1.2	950.8
	Hybrid	(180,278)	(36,29)	17	33.2	1.8	1 212.7
Rnd40/40	Grasp	(240,207)	(67,103)	**18**	10.4	1.1	**41.7**
	Tabu	**(107,207)**	(67,157)	7	**7.1**	**0.6**	1 050.5
	Hybrid	(239,207)	**(67,101)**	12	20.8	1.4	1 832.2
Rnd50/50	Grasp	(250,350)	(83,179)	**22**	**5.6**	0.5	**70.8**
	Tabu	**(232,337)**	**(84,40)**	15	35.7	1.6	5 419.1
	Hybrid	(249,350)	(83,165)	17	7.6	**0.4**	5 741.4
Rnd60/60	Grasp	(276,384)	**(102,220)**	**21**	11.1	0.9	**107.0**
	Tabu	**(225,363)**	(103,240)	12	**8.9**	**0.6**	12 904.0
	Hybrid	(281,384)	(102,224)	16	10.7	**0.6**	11 884.3
Rnd150/150	Grasp	(403,829)	(282,389)	11	**4.1**	**0.3**	**1 283.4**
	Tabu	**(297,822)**	(281,402)	3	207.9	1.4	247 078.7
	Hybrid	(466,812)	**(281,361)**	**14**	98.2	2.5	63 2458.9

Table 3 Metrics calculated from the non-exact frontiers for the random Web communities

As can be seen from Tables 2 and 3 above, the Grasp wins in terms of the number of non-dominated solutions and in computing time, both for the real and for the random communities. As for the way in which points are spread along the non-exact frontier, measured by the spacing metrics SM_1 and SM_2, Grasp produced the better spacing for the smallest instances but not for all the others. Here, in some instances the Tabu and the Hybrid also attained good distributions along the non-exact frontier.

Tables 4 and 5 below include a comparison of each pair of heuristics in relation to the dominance of the respective non-exact Pareto frontiers. A metric proposed by Zitler, Deb and Thiele in 1999, called relative dominance metric (*RDM*), allows one to compare two non-exact Pareto frontiers in the following way: given two frontiers *F*' and *F*'', the function presented in (19) transforms the two frontiers into a real number

between 0 and 1, where 1 means that all points in *F''* are dominated or are equal to the points in *F'* and 0 otherwise.

$$RDM(F',F'') = \frac{\left|\left\{\overrightarrow{f''} \in F'' : \exists \overrightarrow{f'} \in F' : \overrightarrow{f'} \prec \overrightarrow{f''}\right\}\right|}{|F''|} \qquad (19)$$

where $|F|$ represents the cardinality of F and $\overrightarrow{f'} \prec \overrightarrow{f''}$ means that point $\overrightarrow{f'}$ dominates or is equal to $\overrightarrow{f''}$.

For example, in Mat30 the value for *RDM* (Hybrid, Grasp) is 0.60, which means that 60% of the points in the frontier generated by the Grasp are dominated or equal to the points in the frontier generated by the Hybrid.

(1) Web community/nº of domains	(2) method (*F'*)	(3) Grasp (*F''*)	(4) Tabu (*F''*)	(5) Hybrid (*F''*)
Mat30/30	Grasp		0.30	0.20
	Tabu	0.30		0.30
	Hybrid	0.60	0.40	
Mat53/53	Grasp		0.27	0.17
	Tabu	0.33		0.17
	Hybrid	0.50	0.27	
Clim/68	Grasp		0.56	0.71
	Tabu	0.35		0.43
	Hybrid	0.60	0.44	
Pov/59	Grasp		0.50	0.64
	Tabu	0.17		0.21
	Hybrid	0.50	0.50	
Hiv/55	Grasp		0.43	0.67
	Tabu	0.00		0.00
	Hybrid	0.27	0.57	

Table 4 Comparison between each pair of non-exact frontiers of the real Web communities

(1) Web community/nº of domains	(2) method (*F'*)	(3) Grasp (*F''*)	(4) Tabu (*F''*)	(5) Hybrid (*F''*)
Rnd20/20	Grasp		0.60	0.00
	Tabu	0.09		0.00
	Hybrid	0.60	1.00	
Rnd30/30	Grasp		0.11	0.20
	Tabu	0.28		0.12
	Hybrid	0.60	0.11	
Rnd40/40	Grasp		0.29	0.42
	Tabu	0.27		0.42
	Hybrid	0.30	0.43	
Rnd50/50	Grasp		0.40	0.24
	Tabu	0.41		0.41
	Hybrid	0.60	0.40	
Rnd60/60	Grasp		0.00	0.30
	Tabu	0.48		0.19
	Hybrid	0.80	0.33	
Rnd150/150	Grasp		0.00	0.06
	Tabu	0.55		0.21
	Hybrid	0.73	0.67	

Table 5 Comparison between each pair of non exact frontiers of the random Web communities

In particular, the figures given in Tables 4 and 5 reveal that none of the three methods dominate the others. It is nevertheless evident that, in these computational tests the Tabu heuristic, most of the time, is dominated by the Grasp and the Hybrid and only in the Rnd30 and Rnd50 instances does the Tabu dominate both Grasp and Hybrid heuristics.

Among the results obtained in these tests the dominance is equally distributed between the Grasp and Hybrid, though the latter heuristic does outperform the first, particularly in instances of larger dimensions (s.a. Rnd60 and Rnd150).

Note that, from all the experiments it is clear that, in terms of computational time, the best results are attained by Grasp, which displays some very low values when compared to the two other heuristics.

Conclusions and final remarks

In this paper the Web Reconfiguring Problem (WRP) is defined as a bi-objective hub-and-spoke model which is proved to be NP-hard. Three single objective metaheuristic approaches have been briefly presented, along with the way they are used to tackle the bi-objective optimisation characteristic of WRP. An application of the methodology, illustrated through the study of six epistemic and conceptual communities and another six random Web communities, was described.

The Grasp and Hybrid heuristics provided better results, both in quality and computational time, when compared with the Tabu heuristic. Due to computational limits we only generated all the efficient solutions (the exact optimal solutions) for one instance from the test set analysed here. Nevertheless, various indicators (*Card*, SM_1, SM_2 and *RDM* metrics) enabled us to compare the three heuristics and reveal their suitability in solving the bi-objective optimisation problem addressed here.

One may conclude from these experiments that all metaheuristics developed for WRP, and particularly the Grasp, generated a diversified set of candidate efficient solutions for the bi-objective optimisation. According to the specific policy goals (e.g. costs vs. balancing objectives), the decision-maker can, from this set of solutions, either select a more balanced reconfiguration or a low-cost solution, or a combination of both. In this respect, the heuristics represent a novel tool for decision-making.

Indicators of network distance, flow betweenness, centrality and cohesiveness of the Web communities (Wasserman and Faust, 1994) were later used in this investigation to analyse the impact of the reconfiguration process. The hub-and-spoke structure imposed a reconfiguration that pushes the indicators, specially those related to proximity and flow betweenness, towards better values when compared to the values of the initial Web community, in keeping with the goals proposed for the Web Balancing Problem.

Finally, the present study demonstrated that when reconfiguring Web communities, highly significant improvements were obtained in reducing the overall distances between any two domains in the network. Moreover, proximity indicators reinforce the

connecting effect imposed by the reconfiguration process. Any two domains become closer and better connected, which will promote democratisation of access to resources within the Web community or a more even distribution of information resources within the Web.

Hence, this innovative methodology promises interesting practical applications in the fields of information diffusion, network organization and network structure studies.

References

Albert R., Barabási A. (2002). Statistical mechanics of complex networks, *Reviews of Modern Physics*, 74, 47-97.

Abdinnour-Helm S. (1998). A hybrid heuristic for the uncapacitated hub location problem, *European Journal of Operational Research*, 106, 489-499.

Batagelj V. and Mrvar A. (1998). Pajek, a program for large network analysis. *Connections*; 21(2): 47–57.

Caldas A. (2005). Galilei – A Multi-agent System for the Discovery of Digital Knowledge Bases, working paper, Oxford Internet Institute.

Caldas A., Schroeder R., Mesch G. and Dutton W. (2005). The World Wide Web of Science and Democratisation of Access to Global Sources of Expertise, accepted for publication in *JCMC – Journal of Computer Mediated Communication*, Special Issue on Search Engines.

Camargoa R., Miranda G. and Lunac H. (2008). Benders decomposition for the uncapacitated multiple allocation hub location problem *Computers & Operations Research,* 35, 1047-1064.

Campbell J. (1994). Integer programming formulations of discrete hub location problems, *European Journal of Operational Research*; 72: 387–405.

Colaço S. and Pato M. (2006). GRASP and Tabu Search heuristics for redesigning Web communities, Working paper 10/2006 at Centro de Investigação Operacional, Faculdade de Ciências, Lisbon University.

Collette Y. and Siarry P. (2003). *Multiobjective optimization*, Springer, New York,.

Collette Y. and Siarry P. (2005). Three new metrics to measure the convergence of metaheuritsics towards the pareto frontier and the aesthetic of a set of solutions in biobjective optimization, *Computers & Operations Research*, 32, 773-792.

Ebery J., Krishnamoorthy M., Ernst A., and Boland N. (2000). The capacitated multiple allocation *hub* location problem: Formulations and algorithms, *European Journal of Operational Research*, 120, 614–631.

Ehrgott M. (2005). *Multicriteria Optimization*. Springer: Berlin.

Ernst A., Jiang H. and Krishnamoorthy M. (2005). Reformulations and computational results for uncapacitated single and multiple allocation hub covering problems, *unpublished*.

Flake G., Lawrence S. and Giles C. (2000). Efficient Identification of Web Communities in *6th International Conference on Knowledge Discovery and Data Mining* (ACM SIGKDD-2000), Boston, MA, USA: 150-160.

Flake G., Lawrence S., Giles C. and Coetzee F. (2002). Self-Organization and identification of Web communities, *Computer*, 35(3): 66–71.

Glover F. (1989). Tabu search-Part I, ORSA *Journal on Computing*, 1(3), 190–206.

Glover F. and Laguna M. (1997). Tabu search. Kluwer Academic Publishers: Boston.

Greco G. and Zumpano E. (2004). Web Communities:Models and Algorithms, *World Wide Web: Internet and Web Information Systems*, 7, 59-82.

ILOG (2002). CPLEX System, Version 8.0, User's Guide Standard (Command-line) Version Including CPLEX Directives.

Kara B. and Tansel B. (2000). On the single-assignment p-hub center problem, *European Journal of Operational Research*, 125, 648–655.

Kara B. and Tansel B. (2003). The single-assignment hub covering problem: Models and linearizations, *Journal of the Operational Research Society*, 54, 59-64.

Kleinberg J., Kumar R., Raghavan P., Rajagopalan S. and Tomkins A. (1999). The Web as a graph: measurements, models, and methods, in *Proceedings of 5th International Conference on Computing and Combinatorics*, Tokyo.

Kleinberg J. and Lawrence S. (2001). The structure of the Web, *Science*; 294, 1849-1850.

Klincewicz J. (1991). Heuristics for the *p*-hub location problem, *European Journal of Operational Research*, 53, 25–37.

Klincewicz J. (1996). A dual algorithm for the uncapacitated hub location problem, *Location Science*, 4(3), 173-184.

Kumar R., Raghavan P., Rajagopalan S. and Tomkins A. (1999). Trawling the Web for emerging cyber-communities, *Computer Networks*, 31, 1481-1493.

O'Kelly M. (1992). Hub facility location with fixed costs, *The Journal of the Regional Science Association International*, 71(3), 293-306.

Resende M. and Ribeiro C. (2002). Greedy randomized adaptive search procedures. In *State-of-the-Art Handbook of Metaheuristics*, (Eds), F. Glover, G. Kochenberger, Kluwer, 219-249.

Rodríguez-Martín I. and Salazar-González J. (2008). Solving a capacitated hub location problem, *European Journal of Operational Research*, 184, 468-479.

Ross G. and Soland R. (1980). A multicriteria approach to the location of public facilities, *European Journal of Operational Research*, 4-5, 307-321.

Skorin-Kapov D. and Shorin-Kapov J. (1994). On tabu search for the location of interacting hub facilities, *European Journal of Operational Research*, 73, 502-509.

Topcuoglu H., Corut F., Ermis M. and Yilmaz G. (2005). Solving the uncapacitated hub location problem using genetic algorithms, *Computers & Operations Research,* 32(4), 967-984.

Wasserman S. and Faust K. (1994). *Social Network Analysis,* Cambridge University Press, Cambridge.

Webopedia, online Dictionary for Computer and Internet Technology, 2005. Available at *URL:http://www.webopedia.com/TERM/d/domain.html* [Accessed 19 December 2005].

Zitler E., Deb K. and Thiele L. (1999). Comparison of multiobjective evolutionary algorithms: empirical results, technical report 70, *Computer Engineering and Networks Laboratory*, Swiss Federal Institute of Technology, Zurich.