International Journal of Operations Research

International Journal of Operations Research Vol. 5, No. 3, 138-153 (2008)

Grasp and Tabu Search for Redesigning Web Communities

Susana Colaço[1, 3, *] **and Margarida Vaz Pato**[2, 3]

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

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

[3]Centro de Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Bloco C6, Piso 4, Campo Grande, 1749 – 016 Lisboa, Portugal

Received May 2007; *Revised October 2007*; *Accepted February 2008*

Abstract—Web topologies are commonly characterised by hierarchical structures and highly unbalanced compositions, as illustrated by the difference of centrality and connectivity as to their elements. The major interest of the problem addressed in this paper lies in restructuring web communities to reduce these initial disequilibria so as to democratise information access or even for the purpose of preserving contents distributed on the Internet. Discussion of this issue thus leads to a hub location problem, formalised by network and integer programming models. Due to its highly complex nature, a GRASP and a tabu search heuristics were developed to find good quality feasible solutions to the problem. The set of test instances includes web communities obtained by crawling the web and using epistemic boundaries, as well as other randomly generated communities, built with specific network analysis software. The experiment demonstrated that the metaheuristics produced low costs and balanced structures, at least for the lower dimension web communities considered. All the redesigned web communities are more closely connected than before and the average distance among their elements reduced

Keywords—Heuristics, Grasp, Tabu search, Web communities, Hub-and-spoke networks

1. INTRODUCTION

Several studies on the structure of the Internet, and particularly the World Wide Web, have revealed that, irrespective of its arbitrary growth, the web comprises a significant hierarchical structure, dominated by hubs and authorities, with highly skewed degree distributions (see, for instance, Kumar et al. (1999), Kleinberg and Lawrence (2001)). Some of these authors have particularly focused on identification and characterisation of web communities (Kleinberg et al. (1999), Kumar et al. (1999) and Flake et al. (2002)). However, far less research has been conducted into studying how the structure of these communities may be redesigned so as to counterbalance the network disequilibrium in terms of degree distributions. This may be accounted for by the fact that few pages have high indegree and outdegree values, whereas many pages register low indegree and outdegree values. Our research on the Web Community Balancing Problem, abbreviated as WBP, is precisely dedicated to the goal of redesigning web communities. The methodology proposed is supposed to be used as an analytical tool that can be employed to support the planning of a web community reconfiguration and not a tool for the engineering or effective construction of information systems forcing balanced structures on the web.

Applications of such a problem can be found in the diffusion of information or resources among the elements of a community, the implementation of distribution policies within the community or even used as a support when planning, organising and preserving web resources, as the authors will briefly explain next.

The WBP can be applied to long term preservation and availability of contents distributed on the Internet. Information available on the web changes rapidly and dynamically, with the systematic addition and deletion of information, nodes and hyperlinks. In the medium/long term, this situation generates acute information instability on a web community. In this context, identification of a set of hubs aggregating parts of the web community can clearly enhance long term web resource preservation because the hubs and all information they aggregate can be controlled and preserved in a more systematic manner, as in making periodic backups of this information and links. This is one among various practical situations that have recently been studied from the socioeconomic point of view by Caldas (2007) and Day (2003) and will be the basis of a future joint project with the first author.

* Corresponding author's email: susana.colaco@ese.ipsantarem.pt

The WBP can also be employed as a means of defining the organising procedures required to implement some recent Education policies in Portugal, where most teachers pertaining to the elementary system are scattered around the country, working in small schools attached to isolated villages, where the web may be regarded as the privileged form of communication. As for the process of updating scientific and pedagogical information for education, it is crucial to reach all elements of this community. Here, the planning of information diffusion and the receiving of teacher feedback may be organised through the web, by resorting to a hub-and-spoke type model to ensure that the information both flows and reaches all elements through a more balanced structure. In this context, an example of a Mathematics Education web community is studied in Section 6.2.

There is an additional benefit in redesigning web communities. This concerns the availability of indicators, which allows one to compare web structure heterogeneity across different communities. These indicators could be of special value when comparing different types of communities (e.g. academic or business communities) or the same community over different times, as pointed out, for instance, by Toyoda and Kitsuregawa (2001).

In the following paragraphs, the Web Community Balancing Problem is presented within a hub-and-spoke approach. Here, a web community is regarded as a collection of web domains, along with their respective hyperlinks. A web page is a document (text, image, video, etc.) that belongs to a specific domain. Web pages are connected through hyperlinks and a hyperlink is a reference (word, figure, etc.) on a web page pointing to the same page, to other pages of the same domain or even to pages of other domains. Hence, one considers that a link from domain *i* to domain *j* exists when there is at least one hyperlink from a web page in *i* to a web page in *j*. Associated with any oriented pair of domains there is a parameter called intensity, which is expressed by the total number of hyperlinks connecting pages of the two domains and sharing the same direction. The inverse of this parameter is the weakness. Figure 1 shows the weakness values for the links (i, j) and (j, i), respectively $1/2$ and 1. Without loss of generality, it is assumed that the weakness from one domain to another is a real value greater than 1, if no such hyperlink exists. The purpose of this paper is to present one way of redesigning a given web community by selecting some controller domains so as to create balanced clusters in relation to the original links, thus improving the communication equilibrium within the web community (see, for illustrative purposes, a small web community in Figure 2 below). The new design, which will be explained next, not only sets out to achieve the linking structure balance but also to minimise costs, determined by the number of new links that must be created and the amount of controller domains to be selected. Redesigning is attained with a hub-and-spoke structure built from the given web community, where the hubs are domains aggregating a set of other domains – the spokes – which are hyperlinked with them. The hubs act as controllers, that is, they are consolidation and dissemination centers, designed to receive, process and distribute information. Each spoke domain is associated with one hub only, which is supposed to control all information leaving or entering the spoke.

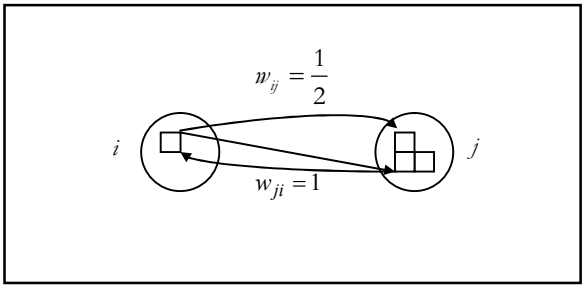

Figure 1. The weakness of the linkings between domains *i* and *j*.

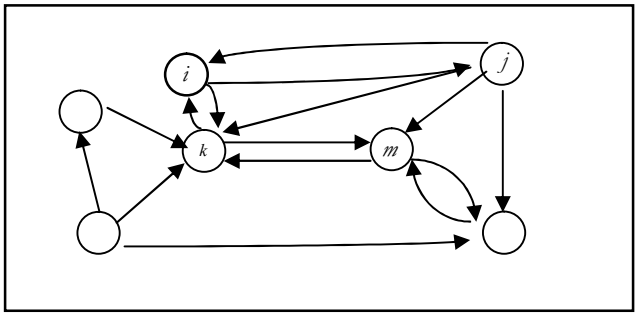

Figure 2. An illustrative web community.

The connection from one spoke domain to another spoke domain that does not belong to the same hub, is always routed via a pair of hubs, which are assigned to serve at least these two domains. If a spoke is not initially hyperlinked, both from and to the respective hub, then one or two new hyperlinks must be created in the redesigned web network. As for the hubs within the redesigned network, all will be linked in both directions. This hub-and-spoke reconfiguration is adequate for the WBP because it assigns each non-hub domain to only one hub, which works as a center of dissemination and aggregation of a cluster of domains. When a domain has to redirect its respective flow via one hub, then it is forced to share information and resources with all the web community. The diagram presented in Figure 3 shows the small web community of Figure 2 already redesigned within a hub-and-spoke structure of two clusters, where the hubs are represented by bold black circles, the pre-existing linkings by arrows and the new hyperlinks by dotted arrows. The weakness of links between hubs is reduced by a pre-determined factor α with $0 \le \alpha \le 1$, to account for improved domain centrality and connectivity. By imposing a maximum bound, represented by γ, on the total weakness of the connections between each origin/destination pair of spoke domains (calculated from the original hyperlinks), routed via hubs, a minimum level of connection is guaranteed. Consequently, the flow of resources within the entire community is facilitated. For example, let us suppose that one establishes the connection from *i* to *j*, two spoke domains not allocated to the same hub, through the path *i–k–m–j*, where *k* and *m* are hubs. Then this path possesses a total weakness equal to $w_{ik} + \alpha w_{km} + w_{mj}$, as illustrated in Figure 4. When *i* and *j* are allocated to the same hub, the formula for the total weakness is the same, assuming that *k* = *m* and $w_{kk} = 0$. Note that, even if *i* is originally linked to *j*, meaning that locally it is faster to reach *j* directly from *i*, globally, i.e. in terms of the overall web community, such a situation does not guarantee that other domains have access to information originating from *i*. In this case, the hub covering model provides the answer to this problem.

In the above context, as already mentioned, from all the domains in the web community one endeavours to select those that will act as hubs. The remaining domains, spokes, will be allocated to these controllers, thus building non-overlapping clusters. This may lead to the creation of new hyperlinks between pages of the domains, besides which, the resulting structure designed for the web community must possess balanced clusters. For this purpose, two parameters associated with a domain were considered, the indegree – the number of incoming links from all the other domains in the original web community – and the outdegree – the number of outgoing links from this domain to others. Balancing is related to the indegree and outdegree parameters of the initial web community and not to these parameters of the restructured community. In fact, it was considered that equilibrium based on the indegree and outdegree of the pre-existing links is more realistic and has more impact than equilibrium from the existing links plus the newly created links. In fact, the initial in and outdegree of each domain are more representative of the role that the specific domain actually plays within the web community. To sum up, the Web Community Balancing Problem redesigns the web community within a hub-and-spoke structure by minimising the number of hubs and the number of new hyperlinks required to reconfigure the web community, besides minimising the disequilibrium between clusters, while ensuring that the total weakness of the connection between each origin/destination pair of domains respects a pre-defined bound.

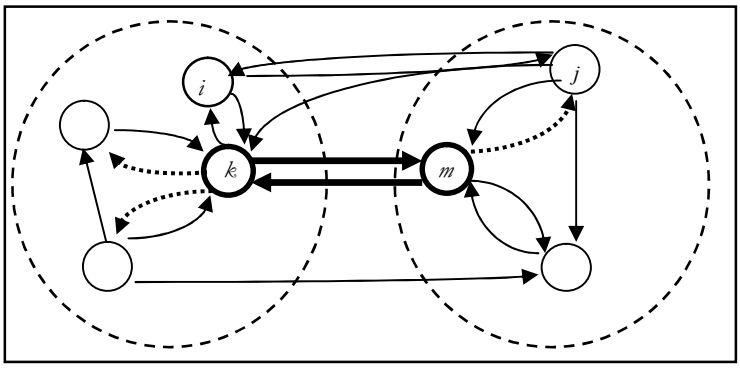

Figure 3. A redesigned web community.

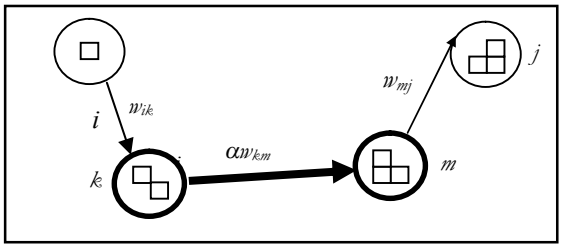

Figure 4. Total weakness associated with the path *i– k–m–j*.

It should be observed that minimisation of costs leads to less created links and to a low number of hubs, which requires that the web be controlled efficiently. Such a feature of the model was motivated by the aim of interfering as little as possible in the structure of the web community. On the other hand, the presence of more hubs is important insofar as it ensures a better balance among the web community, which is why these goals reveal a contradictory nature.

In Section 2 models are presented for the WBP. In view of the high complexity of this problem and on the strength of the experiments that used exact methods but failed to solve even the medium dimension instances, the solving option favoured the non-exact methods. Subsequently, Sections 3, 4, and 5 are devoted to metaheuristics, along with the computational experiments in Section 6, followed by comments in Section 7.

2. NETWORK AND INTEGER PROGRAMMING MODELS

Let $G = (N, A)$ be a directed network associated with the existing web community, where $N = \{1, 2, ..., n\}$ is the set of nodes associated with the domains and A the set of arcs relative to the links represented by the binary matrix $[a_{ij}]_{i,j\in N}$ in which $a_{ij} = 1$ if there is at least one hyperlink from a page of i to a page of j in the original community and $a_{ij} = 0$, otherwise. Assigned to each node $i \in N$ are two parameters, already defined in the previous section, related with the linkings of the node within the network G: the outdegree out_i and the indegree ind_i. The parameter $1 \geq w_{ij} > 0$ is associated with each arc $(i, j) \in A$ and represents the weakness of the link from i to j in the original network. As referred to above, it is assumed that when $(i, j) \notin A$, then $w_{ij} = \varphi$, where $\varphi > 1$ is a pre-determined real parameter.

Within the WBP resolution one must therefore determine the set $H \subset N$, which represents the nodes specifying the location of hubs, besides, for each $h \in H$ the set of spoke nodes allocated to that hub, that is, its cluster. These assignments can be performed by using the existing arcs of A or by creating new arcs. In doing so, the redesigned network must satisfy a maximum value $\gamma > 0$ for the total weakness between each pair of spoke nodes and, at the same time, minimise costs given by the number of new arcs and of hubs, as well as the maximum of total indegree and outdegree, among all clusters. As mentioned in the previous section, the total weakness is calculated by using the parameter $\alpha (0 \leq \alpha \leq 1)$, the discount factor for the hub to hub weakness. In other words, the aim of the WBP is to construct a partially new network $G' = (N, A')$, where $A' \subset (N \times N)$ is the set composed by the old arcs, along with the new necessary arcs. Some of them represent the links between the hubs and their spokes, as well as between hubs, and act in both directions at all times. Hence, G' is a directed network containing a hub-and-spoke structure, with balanced disjoint clusters of nodes redesigned at a low cost, while respecting the weakness constraint.

The WBP is similar to the hub covering problem which belongs to the class of discrete hub location problems, first introduced by Campbell (1994). Kara and Tansel (2003) recently studied these issues and pointed out that the main difference among the hub location problems depends on the criterion used to optimise the system's performance. The standard hub covering problem is a hub location issue in which the number of hubs must be minimised, while respecting a maximum bound on the travelling time between any two nodes – the so-called covering criterion. Campbell (1994) produced an earlier formulation for the hub covering problem, while Kara and Tansel (2000) proposed a combinatorial formulation, as well as a new integer programming model. Recently, Wagner (2004) improved the formulations for the hub covering problem and Kara and Tansel (2003) proved that it is *NP*-hard.

Besides the fact that the WBP is, in some way, like the hub covering problem, the weakness values used in the covering criterion do not verify the triangular inequality, as the weakness is not a metric. Moreover, the WBP is based on a pre-existing network that will be rebuilt according to specific features regarding the optimisation objective, which are not that often found in hub location problems. Here, the objective amounts to minimising hubs, plus new arcs and at the same time, minimising the maximum of the clusters' indegree and outdegree values.

In order to clarify certain aspects of the above WBP, and bearing in mind the formulation due to Kara and Tansel (2003), an integer programming formulation was developed as follows, in (1) to (10). Here, $X_{ik} = 1$ if node i is assigned to hub k and $X_{ik} = 0$, otherwise; $X_{kk} = 1$ if node k is chosen to be a hub location, otherwise, $X_{kk} = 0$; $Z_{ik} = 1$ if arc (i, k) is added to the original network, otherwise, $Z_{ik} = 0$; Y and T represent, respectively, the maximum indegree and outdegree among all clusters.

$$Min\ Z = b_0\left(\sum_{k=1}^{n} X_{kk} + \sum_{i=1}^{n}\sum_{k=1}^{n} Z_{ik}\right) + b_1(Y + T) \tag{1}$$

subject to

$$\sum_{k=1}^{n} X_{ik} = 1 \qquad \forall i \in N \tag{2}$$

$$X_{ik} \leq X_{kk} \qquad \forall i, k \in N \tag{3}$$

$$(w_{ik} + \alpha w_{km} + w_{mj}) X_{ik} X_{jm} \leq \gamma \quad \forall i, k, m, j \in N\,(i \neq j) \tag{4}$$

$$\sum_{i=1}^{n} X_{ik} ind_i \leq Y \qquad \forall k \in N \tag{5}$$

$$\sum_{i=1}^{n} X_{ik} out_i \leq T \qquad \forall k \in N \tag{6}$$

$$2X_{ik} \leq (a_{ik} + a_{ki}) + Z_{ik}(1 - a_{ik}) + Z_{ki}(1 - a_{ki})$$
$$\forall i, k \in N\ (i \neq k) \tag{7}$$

$$2X_{ii} X_{kk} \leq (a_{ik} + a_{ki}) + Z_{ik}(1 - a_{ik}) + Z_{ki}(1 - a_{ki})$$
$$\forall i, k \in N\ (i < k) \tag{8}$$

$$X_{ik}, Z_{ik} \in \{0,1\} \qquad \forall i,k \in N \tag{9}$$

$$T, Y \quad \text{non-negative integers.} \tag{10}$$

The first component of the objective function, whose weight is b_0, corresponds to the cost given by the number of hubs plus new arcs. The second component, weighted by b_1, is related to the minimisation of the maximum clusters' indegree and outdegree, where $b_0 + b_1 = 1$. Constraints (2) and (3) assign any node to precisely one hub and a node i (spoke node) is assigned to a node k only if k is chosen to be a hub location. Constraint (4) ensures that the total hyperlink weakness connecting node i to node j, via hubs k and m, is not greater than a given bound γ. As regards the objective function, constraints (5) and (6) are responsible for balancing the clusters in relation to the original indegree and outdegree parameters of the respective spoke nodes. Because the WBP is defined from a pre-existing network, constraints (7) and (8) with the objective function impose the obligation of creating new arcs, albeit as few as possible. Though they may not exist in the initial network, they are required to link spoke domains in both directions with their hubs and to link the hubs, thus creating a complete sub-network of hubs. Finally, constraints (9) and (10) specify values for the variables.

From the computational theory perspective this problem shares the high complexity of other hub-and-spoke issues. In fact, it is an *NP*-hard problem as is proved below through identification of a particular case of WBP, which is *NP*-hard. First, one can take WBP with weight parameters $b_0 = 1$ and $b_1 = 0$. In this case, the components of the model relative to the balancing clusters' goal become inactive. Now, consider that the network $G = (N, A)$ is a complete graph, hence new arcs will not be added. The aforementioned simplification includes elimination of the variables Y, T and Z_{ij}, constraints (5) to (8) and the second sum in (1). Thus the particular problem assumes the following simplified form:

$$Min\ Z = \sum_{k=1}^{n} X_{kk} \tag{11}$$

subject to (2), (3), and (4)

$$X_{ik} \in \{0,1\} \quad \forall i,k \in N \tag{12}$$

Moreover, if one assumes that within the problem (11), (2), (3), (4) and (12) the parameters w_{ij} associated with the arcs of the network satisfy triangular inequality and symmetry, then it becomes precisely the problem of hub covering. Since it has already been proven that the hub covering is *NP*-hard (see Kara and Tansel (2003)), the WBP is also *NP*-hard.

As one can see, the model for WBP, (1) to (10) is not linear due to (4) and (8). Though several possible linearisations exist for a standard formulation of the hub covering problem (see Kara and Tansel (2003)), unfortunately the one that has produced the best results cannot be used for WBP, as the weakness used in the covering criterion is not a metric. Others, however, could be employed here. One, proposed by Campbell (1994) and adapted by Kara and Tansel for the hub covering problem involves introducing coefficients v_{ijkm} such that $v_{ijkm} = 1$ if $w_{ik} + \alpha w_{km} + w_{mj} \leq \gamma$ and 0, otherwise, and four-index variables Y_{ijkm} replacing the products $X_{ik}X_{jm}$, as is the case when linearising quadratic functions. Another is an adaptation for the hub covering problem of a formulation from Shorin-Kapov et al. (1996) for p-hub median problems, and is due to Kara and Tansel (2003). This also involves the above four-index variables.

However, the one adopted here for the WBP was proposed in Ernst et al. (2005) and is based on the concept of maximum in and out-radius covering of hub k, $Rind_k$ and $Rout_k$, respectively. The values of these new variables, for each hub, are calculated on the basis of the maximum weakness value of the arcs incoming to and outgoing from that specific hub. This enjoys the advantage of using $2(n^2 + n + 1)$ variables and $8n^2 + 2$ constraints, in other words far less additional variables than the formulations referred to in the previous paragraph, besides a smaller number of constraints and proved to be more efficient than those in preliminary tests.

Here, non-linear constraints (4) and (8) in the earlier formulation for WBP are replaced by:

$$(2X_{kk} - 1) + (2X_{ii} - 1) \leq (a_{ik} + a_{ki}) + Z_{ik}(1 - a_{ik}) + Z_{ki}(1 - a_{ki})$$
$$\forall i,k \in N (i < k) \tag{13}$$

$$w_{ik}X_{ik} \leq Rind_k \quad \forall i,k\ (i \neq k) \tag{14}$$

$$w_{ki}X_{ik} \leq Rout_k \quad \forall i,k\ (i \neq k) \tag{15}$$

$$Rind_k + Rout_m + \alpha w_{km}X_{kk} \leq \gamma\ X_{kk}$$
$$\forall k,m \in N\ (k \neq m) \tag{16}$$

$$Rind_k, Rout_k \geq 0 \quad \forall k \in N \tag{17}$$

As with the hub covering issue, the hub location problems with a non-fixed number of hubs have been less studied as compared, for example, to the p-hub location problem. Nevertheless, some heuristics have been developed for these problems by O'Kelly (1992), Abdinnour-Helm (1998), Abdinnour-Helm and Venkataramanan (1998), Topcuoglu et al. (2005) and Rodríguez-Martin and Salazar-González (2008). Also, Klincewicz (1996) used a dual algorithm, Abdinnour-Helm and Venkataramanan (1998) developed a branch-and-bound method, Ernst et al. (2005) presented a new formulation with branch and bound techniques, and Camargo et al. (2008) consider an efficient algorithm based on benders decomposition for the uncapacitated multiple allocation hub location problem.

As mentioned earlier, although the number of hubs is also a decision variable in WBP, there are other features involved which are not addressed in the standard hub location problems: the weakness does not verify the triangular inequality, one optimisation goal is to obtain balanced clusters, and minimisation of costs is partly given by the number of new arcs. This explains why the authors did not use the very same methods already developed for

other hub problems. Due to the complexity of the WBP and to the high dimension of the real instances addressed, heuristic approaches are the appropriate choice for this case. Taking this into account, GRASP and tabu search metaheuristics were developed for the WBP and are presented in Sections 3, 4, and 5.

3. NEIGHBOURHOODS

Both GRASP and tabu search are local search heuristics. They are therefore based on movements performed within specific neighbourhoods, as is usual. The neighbourhood of a solution is defined as the set of solutions obtained from it by:

- reassigning a spoke node to another hub (shift movement) – in the case of the shift neighbourhood;
- swapping two spoke nodes (swap movement) – in the swap neighbourhood;
- replacing one hub by a spoke node (location movement) – in the location neighbourhood.

Selection of movements during the local search is performed by calculating the respective savings. As for the savings resulting from the shift or swap movements, they are given in terms of the sum of the number of new arcs and the balancing of the clusters' component of the objective function, along with the respective penalising coefficients b_0 and b_1.

The location neighbourhood is defined by changes in the location of hubs, though hubs are only replaced by spoke nodes assigned to them, apart from the case of a hub without spokes. In such a case, the evaluation is extended to all the other spoke nodes in the network. The savings value related to the replacement of a hub by one of its spoke nodes is computed from the sum of the new arcs' savings. In the second case, where the hub could be replaced by any other spoke node, due to the change in cardinality of the two clusters, the savings are now computed on the basis of the savings resulting from the new arcs, as well as on the savings related to the clusters' degree balancing, weighted by b_0 and b_1. Note that, in the three neighbourhoods the number of hubs never changes so as to reduce the computational expenses.

4. GRASP

A standard GRASP was implemented. Through such a metaheuristic, a feasible solution is obtained with a greedy-randomised construction, followed by a local search in a neighbourhood of that solution, until a local optimum is found. This process repeats over several iterations, for more details see Resende and Ribeiro (2002).

The GRASP main procedure, Global GRASP, has to run for an appropriate range of $p \in [kmin, kmax]$ because, as explained earlier, the number of hubs, p, is not fixed. The pseudo-code presented in Figure 5 refers to the Global GRASP for WBP, which iteratively constructs a maximum of ($kmax - kmin + 1$) solutions, each one for p hubs. Within the procedure denoted as GRASP (p; *bestsolution*), in each iteration a feasible solution with p hubs is built on the basis of a constructive procedure with randomness – Greedy-Randomised Constructive (p, *maxitconstr*, α_1, α_2; *solution_k*) –, followed by a local search procedure – Exchange (*solution_k*, *maxitexch*; *bestsolution_k*). When this GRASP process is concluded, following a fixed number of iterations, *maxit*, the best solution with p hubs is returned or no feasible solution is found for that specific p, and the Global GRASP procedure continues with another value of p. For a fixed number of p hubs, the Greedy-Randomised Constructive procedure was implemented to choose a specific set of p hubs (step 1). Within each iteration of this process, a hub is selected from a restricted candidate list, $RCLH(\alpha_1)$, similar to the prioritised hub candidates' list in Ebery et al. (2000).

The list of «good» hub candidates $RCLH(\alpha_1)$ is built for the current iteration, while taking into account the node's degree parameters (indegree+outdegree), as in Klincewicz (1991), and a threshold parameter $\alpha_1 \in [0,1]$. This means that hub h belongs to $RCLH(\alpha_1)$ if

$$d_h = (ind_h + out_h) \in \left[d^{\max} - \alpha_1 \left(d^{\max} - d^{\min} \right), d^{\max} \right] \qquad (18)$$

where $d^{\min} = \min_{h \in N} d_h$ and $d^{\max} = \max_{h \in N} d_h$. Then a hub location is randomly selected from the elements of the restricted list. Note that, when $a_1 = 1$ this process is completely random and when $a_1 = 0$ it is totally greedy.

The next step (step 2) consists in assigning each spoke node to a hub. First, for each spoke i, the incremental costs regarding each hub are calculated. Then, the best feasible hub candidate for i is chosen, which is precisely the feasible hub with the lowest incremental cost. Here, the incremental costs are based on an optimisation criterion and satisfaction of the weakness constraint. More specifically, the incremental cost c_i for a spoke node i, is calculated as a weighted sum of the number of new arcs needed to assign i to its candidate hub (weighted by b_0), plus the difference between the maxima of the sum of the indegree and the outdegree of the clusters, before and after assignment of i to the candidate hub (weighted by b_1), plus the total weakness of the path linking the spoke to the hub. Following these computations, a restricted candidate list of spokes is created for the best possible assignments on the basis of the lowest incremental costs, $RCL(\alpha_2)$. This list is again associated with a threshold parameter $\alpha_2 \in [0,1]$, hence node i belongs to $RCL(\alpha_2)$ if

$$c_i \in \left[c^{\min}, \; c^{\min} + \alpha_2 \left(c^{\max} - c^{\min} \right) \right] \qquad (19)$$

As above, when $\alpha_2 = 1$ this process is completely random, whereas, when $\alpha_2 = 0$, it is greedy.

This Greedy-Randomised Constructive procedure runs until a feasible solution is achieved or a maximum number of iterations is attained (*maxitconstr*). When this last situation occurs, another set of p hubs must be taken from the list $RCLH(\alpha_1)$, that is, step 1 runs again.

```
Global GRASP
    for p = kmin to kmax do
        procedure GRASP (p; bestsolution)
            for k=1 to maxit do
                procedure Greedy-Randomised Constructive (p, maxitconstr, α1, α2; solution_k)
                    step 1. randomised selection of hubs
                    step 2. randomised assignment of spokes
                procedure Exchange (solution_k, maxitexch; bestsolution_k)
            end
            if bestsolution_k better than bestsolution then bestsolution← bestsolution_k
        end GRASP
    end
    return bestsolution
end Global GRASP
```

Figure 5. Global GRASP algorithm.

```
Global Tabu procedure
    K←∅
    for p = kmin to kmax do
        procedure Greedy Constructive (p; solution_p)
            step 1. selection of hubs
            step 2. assignment of spokes
              UpdateK (t, solution_p; K)
    end
    for each solution_k ∈ K do
        bestsolution←solution_k
        procedure Assignment Phase (maxitassi, maxd, solution_k, maxrepsol, bestsolution; bestsolution_k)
            solution_k ←bestsolution_k
            procedure Location Phase (maxitloca, solution_k, maxrepsol, bestsolution; bestsolution_k)
            if bestsolution_k better than bestsolution then bestsolution← bestsolution_k
    end
    return bestsolution
end Global Tabu
```

Figure 6. Global Tabu algorithm.

Exchange heuristics are local search procedures widely used in facility location problems (Klincewicz (1991)). Here, in the Exchange procedure, starting from the current solution resulting from the random constructive, one evaluates the process of performing a shift movement. If this movement improves the value of the WBP objective function while maintaining feasibility, then the shift is performed – first-improving criterion – and the process continues until no positive savings exist in the shift neighbourhood of the current solution to guarantee a local optimum. To avoid consuming excessive computation time, a limit on the number of movements is imposed (*maxitexch*). Note that only the feasible region is searched by this heuristic.

At the end of Exchange the best solution is kept, i.e., a local optimum for the shift neighbourhood (with p hubs) may be achieved. A particular characteristic of this GRASP method is the permanent restart, combined with a large degree of randomness. This is why the neighbourhood of a solution explored by one such method is based only on shift movements. On the other hand, within the tabu search method detailed in the following section, the focus is placed not on randomness but on the previously conducted search, combined with a more extensive neighbourhood exploration, not only based on shift movements but also on swap and hub location movements.

5. TABU SEARCH

5.1 General aspects

Tabu search is a local search metaheuristic incorporating the concept of memory by considering solutions and/or movements as a tabu, depending on the solutions visited in the previous iterations. The use of memory usually drives the searching process in different directions within the hitherto unexplored space region, in such a way that the closest local optimum can be overreached (see, for details, Glover (1989) and Glover and Laguna (1997)).

There follows in Figure 6 a very brief description and a condensed pseudo-code of the Global Tabu algorithm designed for WBP. This tabu search method is based on previous work, namely a two-step method adopted in Skorin-Kapov and Skorin-Kapov (1994), with assignment and location phases even if, during the assignment phase, both authors have only implemented the swap movements. To construct the initial solution one profits from the work of Klincewicz (1991) for a hub location problem, where the total number of hubs is previously pre-defined at the outset. Cortinhal and Captivo's article (2003) is a

fundamental reference for strategic oscillation in location problems, in particular when setting criteria to enter unfeasible regions. Nevertheless, as the WBP comprises more complex constraints compared to the problem addressed in that work, here recovery of the feasible solutions becomes more difficult, requiring heavier computational resources.

Each main iteration of the Global Tabu algorithm is used with the purpose of improving each solution from a set of initial feasible solutions and, as mentioned above, comprises two phases: assignment and location. The t initial solutions that enter as input in this procedure, defining set K, are designed by a two-step Greedy Constructive (*p*; *solution_p*) procedure. The first step concerns the choice of the hubs according to the degree parameter, inspired by the work of Klincewiz (1991), while the second step deals with the assignment of each spoke node to one hub only, in order to verify the weakness constraint between each pair of nodes. This procedure, called ($kmax - kmin + 1$) times, returns feasible solutions for a range of *kmin* to *kmax* hubs. On the basis of these solutions, set K is defined by the t best solutions, through UpdateK procedure. In each main iteration of the Global Tabu algorithm, for each *solution_k* in K, the process searches the three types of neighbourhoods, never changing the number of hubs. This search is based on shift and swap movements (in the Assignment and Location Phases) and location movements (in the Location Phase), up to a maximum number of movements per phase. Moreover, as opposed to an ordinary local search procedure, Global Tabu allows movements to neighbour solutions that do not improve the objective value of the current solution, but can orient the search into unexplored regions of the feasible set.

After performing a tabu search from each solution in K, the best solution found, in terms of the objective function, is returned. The next subsections detail the components of this metaheuristic.

5.2 Local search

In each Assignment Phase iteration, the shift or swap type movement to be used is randomly selected. This random choice produced better results in computational tests than either alternated shift-swap movements or the use of a single type of movement. The process continues until the number of movements reaches a maximum, *maxitassi*.

The best movement corresponds to the feasible and non-tabu movement with the highest savings. If the movement is classified as tabu it may be selected, but only if it verifies the aspiration criterion. A filter was implemented to limit the local search to the movements with a high probability of solution improvement. This device evaluates the nodes whose sum of indegree and outdegree is higher than a predefined value *maxd*.

In the Location Phase the search is performed within the location neighbourhood. Taking into consideration the heuristics developed in Skorin-Kapov and Skorin-Kapov (1994) and Cortinhal and Captivo (2003), a movement in the Location Phase is always followed by a shift or swap movement. However, in this case, only improving movements are allowed. This process continues until the number of iterations reaches a maximum, *maxitloca*.

5.3 Tabu lists, aspiration criterion and diversification

In order to reach unexplored regions, some movements and solutions are regarded as tabu, on the basis of previous searches. This leads to a dynamic change in the neighbourhood, as mentioned in Glover and Laguna (1997). Here the same tabu list T_1 is used for both shift and swap movements. For location movements another tabu list is used, this time with a different structure, represented by T_2. The tabu list T_2 stores the set of hubs from the previous solutions. Both lists, T_1 and T_2, were implemented with a fixed length.

The aspiration criterion allows one to select a movement in tabu status if the move affords better solution than the best one found so far.

In a tabu search, diversification offsets the effects of short-term memory represented by tabu lists, and its goal is to reach a distant solution. In this case, diversification is achieved by reinitialising the search, using new solutions. As such, in some randomly chosen circumstances, the current solution can be replaced during the Location Phase by another solution derived from running the Greedy Constructive procedure for the current set of hubs. The following probabilities are used: 0.75 for continuing to use the current solution and 0.25 for reinitializing.

5.4 Strategic oscillation

Strategic oscillation was incorporated in this algorithm with a view to crossing the boundary of feasibility, by relaxing the weakness constraint, and afterwards returning to the feasible region at another point.

If the best solution is not improved during a fixed number of consecutive iterations in Assignment Phase or Location Phase, *maxrepsol*, then the Strategic Oscillation procedure will take place. Two distinct movements are used to cross the boundary of feasibility: shift movements and location movements. The first type is used when the Strategic Oscillation is activated during the Assignment Phase and the second one when activated during the Location Phase. Inspired by the work of Cortinhal and Captivo (2003), the best movement to cross the feasibility boundary, chosen from all the movements that improve current solution quality, is the one that minimises the degree of unfeasibility. After entering the unfeasible region, the iterative search process continues and stops once a critical value for the unfeasibility degree is achieved or a maximum number of iterations is reached.

Once the process of exploring the unfeasible region is completed, the search comes back to the feasible region. In this case, the best movement to guide the return to feasibility is the one that minimises a weighted sum of degree of unfeasibility and reduction in the objective function value. The movements used to recover feasibility

are also shift or location movements. To ensure that the process returns to the feasible region at another point, the reverse movements used to enter the unfeasible region are declared to be tabu, hence defining tabu list T_3. In this situation, there is no aspiration criterion. If recovery of feasibility terminates unsuccessfully, then the movements used to cross the feasibility boundary are inserted in T_3, the process recommences, and other movements are chosen to cross the boundary.

The solution used when one decides to enter Strategic Oscillation may be the best solution found so far or the current solution. This depends on whether it was used or not in a previous Strategic Oscillation process. Hence, tabu list T_3 is stored in vectors of three components. Each first component stores the objective function value corresponding to a solution already explored through this procedure, whereas the second and third components are 0 or 1 if the solution with that objective function value has been used or not during this process in an Assignment or a Location Phase.

6. COMPUTATIONAL EXPERIMENTS

6.1 Case studies and implementation

Computational tests in this work were based on real epistemic web communities: three Mathematics and Mathematics Education communities in Portugal and, on an international level, Climate Change, Poverty and the HIV/AIDS communities. Six other randomly generated hypothetical web communities were also tested.

The Mathematics and Mathematics Education web communities in Portugal (Mat20, Mat30 and Mat53, with 20, 30 and 53 domains) were generated using the following procedure:

- keyword search in two search-engines (Google and Altavista) of a collection of web pages within the domain "*.pt*";
- additional selection of a subset of web domains based on expert knowledge;
- web crawling of each web domain, using the software "Galilei", presented in Caldas (2005), followed by generation of a database of hyperlinks for each of these domains;
- shrinking network analysis, thus obtaining a subnetwork and, ultimately, computation of arc intensities.

The web communities of Climate Change (Clim, for short), Poverty (Pov) and HIV/AIDS (Hiv) were obtained from an international project at the Oxford Internet Institute – World Wide Web of Science (Caldas et al. (2007a)). The methodology used was identical to that of the mathematics communities' case. As intensities between web domains were not initially available in these datasets, the "Galilei" software was used once more to calculate intensities.

A network analysis software "Pajek", due to Batagelj and Mrvar (1998), was used to generate six random communities. In this case, the links between the nodes were randomly generated according to density parameters. They represent six hypothetical web communities, some with a low density and others with higher density, corresponding to different real situations studied to date. In the first two cases, Rnd20 and Rnd30, with 20 and 30 domains respectively, each domain has a number of outgoing arcs, randomly generated between 0 and 10. The following, Rnd40, Rnd50 and Rnd60, with 40, 50 and 60 domains respectively, have a number of outgoing arcs, randomly generated between 0 and 8. Finally, Rnd150 was generated with 150 domains and the number of outgoing arcs was randomly generated between 0 and 6. For all these randomly generated networks, each arc's intensity value was randomly generated between 1 and 5.

Table 1 displays information on the general characteristics of the web communities used in the computational experiments. In order to tune some parameters of the methodology, a preliminary experiment was undertaken with all these instances. From these runs the maximum number of hubs, *kmax*, for the Global Grasp as well as for the Global Tabu algorithms, was set equal to 10 when $n \leq 20$, otherwise it was set equal to 15. As for the minimum, $kmin = 1$. The bound parameter γ for the weakness constraint WBP was fixed at 2.5 for all instances of web communities with $n \leq 20$ and at 2.8 otherwise because, when the nodes increase, the weakness constraint must be less restrictive. In fact, the study of the effect of the weakness constraint revealed that this is an important feature for this problem and cannot be discarded.

The value used for maximum weakness bound γ in computational tests was the one which ensures that, after reconfiguration, a path with three arcs (worst scenario) between any two nodes built with new arcs only does not exist. The algorithms relative to the metaheuristics were implemented in C programming language and the programs ran on a single PC Pentium IV processor with 512 Mb RAM and 598 MHZ.

6.2 Experiments with a mathematics education web community

The Mathematics Education community in Portugal constitutes a privileged field of application of the reconfiguration model with the main aim of improving the process of updating scientific and pedagogical information among their elements, according to Ministry of Education policies.

This section follows with a brief description of the methodology used to build and afterwards redesign an illustrative web community MAT20. Figure 7 shows the entire ground conceptual and epistemic web community focused on Mathematics and Mathematics Education in Portugal, prior to selection of seed domains.

From this network, a smaller web community with 20 web domains was obtained by using the steps described in Section 6.1, Mat20, and may be seen in Figure 8 below (after redesigning). In Mat20 the number of arcs is 74 and the network density is equal to 19.5%, the average in and outdegree per node being 3.7.

Table 1. Topological properties of web communities tested

Web community	Number of domains (n)	Number of arcs	Average degree per domain	Network density
Mat20	20	74	3.7	19.5 %
Mat30	30	116	3.9	13.3 %
Mat53	53	292	5.5	10.6 %
Clim	68	81	1.2	1.8 %
Pov	59	84	1.4	2.5%
Hiv	55	57	1.0	1.9 %
Rnd20	20	86	4.3	22.6%
Rnd30	30	179	6.0	20.6%
Rnd40	40	112	2.8	7.5%
Rnd50	50	250	5.0	7.6%
Rnd60	60	210	3.5	8.6%
Rnd150	150	435	2.9	1.8 %

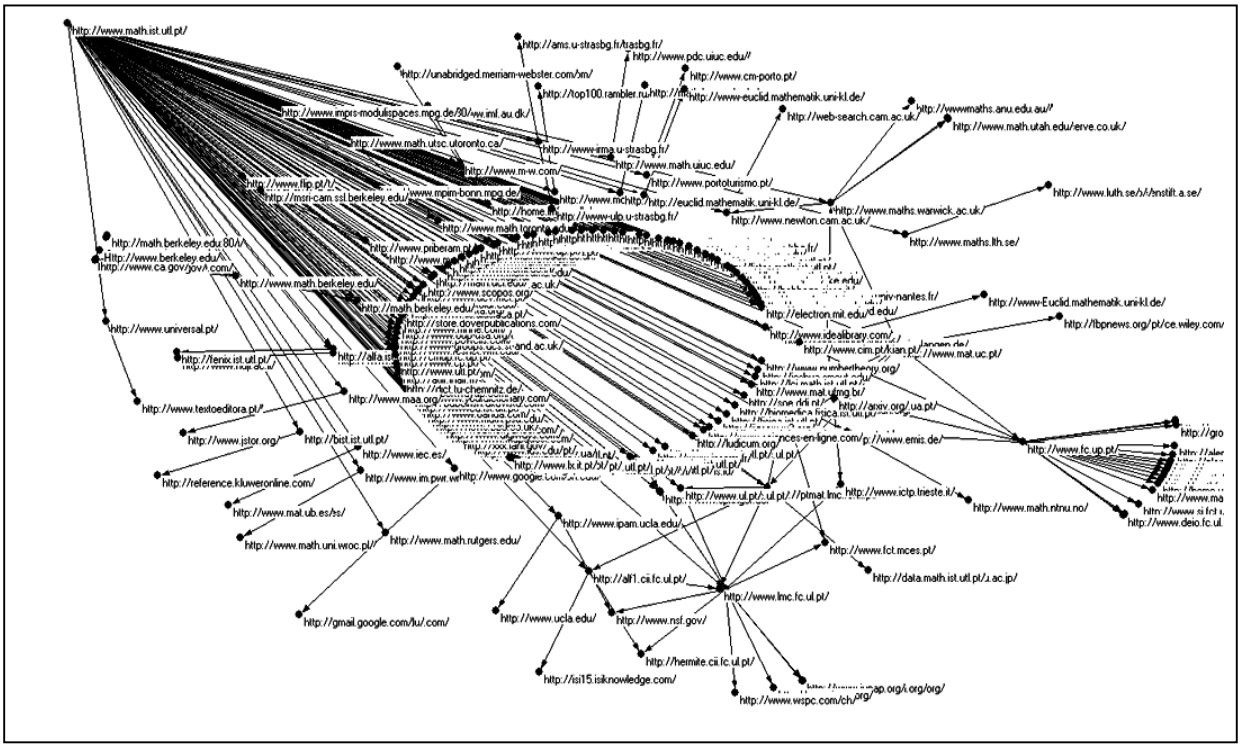

Figure 7. Ground web community on mathematics and mathematics education in portugal.

Should one want to minimise cost alone – the single goal corresponding to $b_1 = 0$ – the solution of the redesigning problem involves only two hubs plus 17 new arcs, the cost being equal to 2 + 17. As expected, the resulting structure is very unbalanced ($Y + T = 65 + 69$), whereas, the solution resulting from the balancing clusters objective alone – the single goal corresponding to $b_0 = 0$ – is a balanced structure, with maximum indegree equal to nine and maximum outdegree equal to 10 ($Y + T = 9 + 10$). As for the cost, this is an expensive solution with nine hubs and 55 extra arcs, thus totalling 64. Finally, Figure 8 below illustrates another solution obtained by taking into account minimisation of the two aspects, costs, with penalisations 0.6 and 0.4, respectively. As one can see, this restructuring of the community is far more equilibrated in terms of achievement of both optimisation objectives: the maximum indegree and outdegree are both 14, whereas the number of hubs is six and new arcs 26, leading to a cost equal to 32. The results obtained for this particular case, involving comparison of the web community Mat20 before and after redesigning and using social network analysis software UCINET 6.0 (Borgatti et al. (1999)), reveal that the principal goals are achieved. For example, the number of pairs disconnected in one way decreased from 88 to 0, the average distance between pairs of nodes decreased from 2.1 to 1.9 and the network diameter was equal to 5 prior to redesigning and afterwards equal to 3.

6.3 General results and discussion

The computational experiments performed with the conceptual web communities referred to in Section 6.1 will enable one to make a comparative evaluation of the proposed metaheuristics for the WBP. Taking into account preliminary tests performed with different weights, the values of $b_0 = 0.6$ and $b_1 = 0.4$ defining the objective

function seem to be the most adequate ones because the choice corresponds to solutions which are fairly equilibrated for both objectives. The results obtained from the Global GRASP are reported in Table 2 where α_1, α_2 take different values and, according to the above mentioned preliminary tests, *maxit* = 3, *maxitcons* = 3 and *maxitexch* = 40. The Global GRASP procedure ran five times with each WBP instance, and the table displays the worst, best and median values for each instance after the five runs. This table presents the number of nodes of each web community (column (1)), the parameters' values (column (2)), the objective function values in columns (3), (4), and (5) relative to the worst, median and best values found. The total CPU time of the five runs, given in seconds, is shown in column (6).

In Table 2 the best values are shown in bold characters. One may see that the best median and worst solutions are attained when parameters $\alpha_1 = \alpha_2 = 0.35$, in all but two cases. This means that the best choice for nodes when building a feasible solution through the constructive procedure, Greedy-Randomised Constructive, lies in the low randomising version ($\alpha_1 = \alpha_2 = 0.35$). The computing time is low and, in most cases, increases as α_1 and α_2 increase, that is, in terms of time, the randomised versions are heavier. Results from the Global Tabu heuristic are summarised in Table 3, where columns (1) and (3) to (5) have the same meaning as in the previous table. The following threshold algorithm parameters were chosen from the preliminary computational studies: $t = 5$ (number of initial solutions), *maxd* = 4, *maxitassi* = 10, *maxitloca* = 20 and all tabu lists with a fixed length equal to 10. The Global Tabu algorithm ran three times with no Strategic Oscillation procedure (SO) and three times with Strategic Oscillation, where *maxrepsol* = 3 (see column (2)). It was decided to run the Global Tabu only three times as, for our purpose, here the random factor is not as relevant as it is in the GRASP. Hence, column (6) represents the total CPU for three runs.

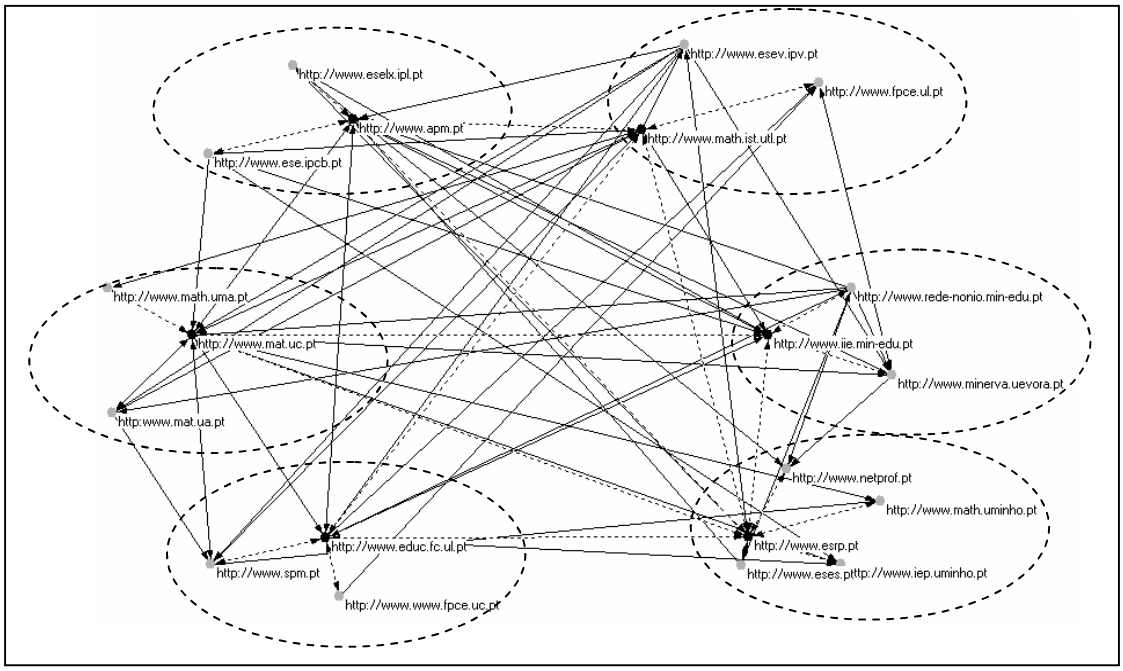

Figure 8. Solution from redesigning Mat20 with the WBP model.

Table 2. Computational results of Global GRASP

(1) Web community/n	(2) α_1, α_2	(3) Worst value	(4) Median value	(5) Best value	(6) Total CPU time (sec)
Mat20/20	0.0	40.8	40.8	40.8	0.8
	0.35	**30.6**	**30.6**	**30.6**	0.8
	0.5	**30.6**	**30.6**	**30.6**	0.9
	0.85	36.8	36.6	35.6	1.1
	1.0	36.8	36.2	33.4	1.4
Mat30/30	0.0	39.4	39.4	39.4	3.9
	0.35	**38.6**	**38.4**	**38.4**	3.8
	0.5	38.8	38.6	**38.4**	3.8
	0.85	44.0	42.8	41.8	3.1
	1.0	49.2	47.2	44.0	3.5

Table 2. Computational results of Global GRASP (contd.)

(1) Web community/n	(2) α_1, α_2	(3) Worst value	(4) Median value	(5) Best value	(6) Total CPU time (sec)
Mat53/53	0.0	70.2	70.2	70.2	17.4
	0.35	**68.8**	**67.6**	**67.2**	15.4
	0.5	69.2	68.6	67.6	14.2
	0.85	80.2	75.0	72.0	12.3
	1.0	91.8	87.4	80.4	14.0
Clim/68	0.0	106.0	106.0	106.0	8.2
	0.35	**103.4**	103.4	103.4	8.2
	0.5	105.4	**101.8**	**101.8**	8.8
	0.85	109.8	107.0	105.8	12.4
	1.0	120.6	118.6	117.2	15.2
Pov/59	0.0	97.8	97.8	97.8	9.7
	0.35	**91.6**	**91.6**	91.6	6.6
	0.5	97.0	95.6	91.6	6.7
	0.85	102.2	98.0	**89.8**	9.2
	1.0	110.2	106.6	100.4	10.8
Hiv/55	0.0	90.2	90.2	90.2	3.3
	0.35	88.0	**88.0**	88.0	4.1
	0.5	**87.0**	88.2	88.6	4.8
	0.85	93.0	89.8	**86.2**	7.3
	1.0	97.6	95.2	93.2	8.4
Rnd20/20	0.0	36.8	36.8	36.8	1.1
	0.35	**34.6**	**33.2**	**33.2**	1.1
	0.5	37.0	34.0	**33.2**	1.0
	0.85	44.2	40.8	39.8	1.2
	1.0	48.6	43.2	39.2	1.4
Rnd30/30	0.0	**68.0**	68.0	68.0	2.2
	0.35	**68.0**	**59.4**	**57.6**	2.3
	0.5	78.4	71.0	69.2	2.3
	0.85	74.6	73.8	63.4	2.8
	1.0	89.6	73.4	67.8	3.2
Rnd40/40	0.0	88.2	88.2	88.2	3.1
	0.35	**86.0**	**82.0**	**76.2**	2.5
	0.5	92.2	85.0	83.4	3.4
	0.85	92.2	86.8	82.6	4.1
	1.0	90.8	85.0	82.3	4.6
Rnd50/50	0.0	163.0	163.0	163.0	2.4
	0.35	118.4	114.6	109.8	4.5
	0.5	115.0	**110.4**	**99.2**	6.3
	0.85	**113.6**	118.0	124.6	8.0
	1.0	126.0	122.4	122.0	8.3
Rnd60/60	0.0	141.2	141.2	141.2	3.9
	0.35	**140.2**	**139.2**	137.2	6.4
	0.5	149.0	144.4	**134.4**	8.8
	0.85	151.8	146.4	144.4	11.7
	1.0	149.2	145.8	138.8	12.0
Rnd150/150	0.0	479.2	477.0	476.4	13.2
	0.35	**349.8**	**342.8**	**341.4**	78.5
	0.5	370.0	356.6	344.8	100.2
	0.85	483.4	353.6	346.8	119.7
	1.0	406.4	367.2	347.6	117.2

As expected, in most cases the version with SO takes more computing CPU time than the algorithm without SO (column (6)). The quality of the solution with SO is always better than or equal to the one obtained without SO forthe worst, median and best costs (column (5)). Finally, the two metaheuristics were combined using the Global GRASP

solutions (with $\alpha_1 = \alpha_2 = 0.35$) as input in Global Tabu version with Strategic Oscilation. This hybrid procedure ran three times for each instance. Table 4 displays the computational results obtained from the Global GRASP, the Global Tabu with SO, the hybrid heuristic and an exact method as well. The exact solutions came from the integer linear programming formulation (1)-(3), (5)-(7), (9), (10) and (13)-(17) given in Section 2, and solved with CPLEX 8.0. As above, column (6) refers to the total computing time of the three runs, whereas columns (3) to (5) present the results referring to the median solutions.

As illustrated in Table 4, the metaheuristics are always capable of finding feasible solutions, whereas the exact method could only find a solution, naturally the optimal one, for the instances with $n \leq 20$ domains. One should observe that the exact solution was found for the smaller instances Mat20 and Rnd20 only, as WBP presents a high computational complexity that induced the exact algorithm of the CPLEX optimiser to consume excessive time besides memory requirements. In the experiments undertaken more than 24 hours of CPU time was consumed to build the exact solution.

As the instance dimension increases, so does the heuristics' CPU time, which is to be expected. For all cases, the Global GRASP is the one requiring the lowest time and, in all but one case, Clim68, the hybrid heuristic was the best or one of the best. Note that, the number of hubs of each best solution found varies between 4 and 7 (column (3)) and in some cases, despite the equal objective function value, the solution may be different. Moreover, the hybrid heuristic also achieves the exact solution for the two lowest dimension instances, with $n = 20$.

Finally, as mentioned above, in most cases Global Tabu combined with Global GRASP provides better results as compared to the separate use of each heuristic. This means that the seed set solution K is important in improving the behaviour of tabu search.

On the other hand, for some instances the Global GRASP metaheuristic produces equal quality solutions, when compared with the hybrid method, and involves much lower CPU expenses.

7. CONCLUSIONS AND REMARKS

This paper has presented and discussed the Web Community Balancing Problem. Firstly, a network model was proposed, along with an integer linear programming formulation. Due to the high computational complexity of the problem and the fact that optimal approaches failed to find the optimal solution, apart from the smaller instances tested, GRASP and tabu search heuristics were then developed to obtain feasible solutions. The original contributions of these metaheuristics are related to the difficulty in maintaining feasible solutions for the WBP throughout the search process because the weakness constraint is extraordinarily heavy in terms of computational resources necessary to verify if a particular solution satisfies the constraint or not. Different versions

Table 3. Computational results of Global Tabu

(1) Web community/n	(2) Tabu search version	(3) Worst value	(4) Median value	(5) Best value	(6) Total CPU time (sec)
Mat20/20	without SO	37.0	36.4	36.2	37.5
	with SO	**36.4**	**34.4**	**31.0**	59.7
Mat30/30	without SO	39.2	**39.0**	39.0	78.4
	with SO	**39.0**	**39.0**	**38.8**	83.3
Mat53/53	without SO	68.0	67.6	**66.6**	198.8
	with SO	**67.4**	**67.2**	**66.6**	216.6
Clim/68	without SO	114.2	114.2	110.0	383.9
	with SO	**106.6**	**106.0**	**103.4**	1154
Pov/59	without SO	**97.8**	**97.8**	97.8	897.9
	with SO	**97.8**	**97.8**	**93.8**	860.6
Hiv/55	without SO	97.0	97.0	96.2	144.9
	with SO	**96.0**	**88.6**	**85.2**	359.2
Rnd20/20	without SO	52.0	52.0	52.0	34.9
	with SO	**34.8**	**34.8**	**34.8**	27.3
Rnd30/30	without SO	122.6	117.0	117.0	63.0
	with SO	**89.2**	**88.8**	**88.8**	84.0
Rnd40/40	without SO	107.2	107.2	107.2	213.7
	with SO	**75.8**	**75.8**	**74.6**	217.9
Rnd50/50	without SO	134.8	134.8	134.8	649.6
	with SO	**113.2**	**113.2**	**105.4**	843.2
Rnd60/60	without SO	149.2	149.2	149.2	2233.0
	with SO	**139.4**	**139.4**	**135.0**	2309.0
Rnd150/150	without SO	341.4	341.4	341.4	3686.0
	with SO	**340.8**	**340.2**	**339.6**	5098.6

Table 4. Comparative results of metaheuristics

(1) Web community/n	(2) Method	(3) Hubs + new arcs	(4) $Y + T$	(5) Objective function value	(6) Total CPU time (sec)
Mat20/20	exact	6 + 26	14 + 14	**30.4**	3863.4
	tabu search	6 + 32	15 + 14	34.4	59.7
	GRASP	6 + 25	14 + 16	30.6	**0.8**
	hybrid	6 + 26	14 + 14	**30.4**	10.1
Mat30/30	exact	-	-	-	-
	tabu search	6 + 29	23 + 22	39.0	83.3
	GRASP	6 + 28	23 + 22	**38.4**	**3.8**
	hybrid	6 + 28	23 + 22	**38.4**	8.6
Mat53/53	exact	-	-	-	-
	tabu search	6 + 48	44 + 43	**67.2**	216.9
	GRASP	8 + 56	37 + 36	67.6	**15.4**
	hybrid	6 + 48	44 + 43	**67.2**	86.8
Clim/68	exact	-	-	-	-
	tabu search	5 + 125	46 + 66	106.0	1154.0
	GRASP	4 + 115	42 + 34	**101.8**	**8.8**
	hybrid	5 + 118	42 + 32	103.4	697.0
Pov/59	exact	-	-	-	-
	tabu search	3 + 94	43 + 56	97.8	860.6
	GRASP	4 + 106	36 + 28	**91.6**	**1.3**
	hybrid	4 + 106	36 + 28	**91.6**	374.1
Hiv/55	exact	-	-	-	-
	tabu search	3 + 96	35 + 38	88.6	325.4
	GRASP	4 + 96	32 + 38	**88.0**	**4.1**
	hybrid	4 + 96	32 + 38	**88.0**	151.6
Rnd20/20	exact	6 + 28	15 + 17	**33.2**	49772.6
	tabu search	6 + 30	16 + 17	34.8	27.3
	GRASP	6 + 28	15 + 17	**33.2**	**1.1**
	hybrid	6 + 28	15 + 17	**33.2**	10.5
Rnd30/30	exact	-	-	-	-
	tabu search	6 + 42	81 + 69	88.8	84.0
	GRASP	9 + 53	29 + 28	59.4	**2.3**
	hybrid	7 + 44	30 + 28	**53.8**	50.8
Rnd40/40	exact	-	-	-	-
	tabu search	4 + 65	42 + 41	**75.8**	217.9
	GRASP	6 + 78	41 + 38	82.0	**2.5**
	hybrid	5 + 68	38 + 42	**75.8**	92.6
Rnd50/50	exact	-	-	-	-
	tabu search	6 + 82	94 + 57	113.2	843.2
	GRASP	6 + 93	68 + 70	114.6	**6.3**
	hybrid	5 + 83	69 + 59	**104.0**	290.8
Rnd60/60	exact	-	-	**-**	-
	tabu search	7 + 118	81 + 80	139.4	2309.0
	GRASP	4 + 98	96 + 99	139.2	**6.4**
	hybrid	5 + 103	80 + 75	**138.4**	408.8
Rnd150/150	exact	-	-	-	-
	tabu search	6 + 294	201 + 201	340.2	5098.6
	GRASP	6 + 294	203 + 204	342.8	78.5
	hybrid	4 + 282	206 + 213	**339.2**	17416.0

of the GRASP and tabu search had to be explored in order to tackle this specific WBP problem which, to the authors' knowledge, no one has studied. Computational experiments subsequently demonstrated that both procedures are capable of finding low cost and balanced web structures and, in the case of GRASP, within a short amount of computing time.

Table 5. Comparison of network indicators for Mat20 and Rnd150

(1) Web community/*n*	(3) Indicator	(4) Value from the initial community	(5) Value from the redesigned community
Mat20/20	flow betweeness	16.2%	26.2%
	incloseness average	19.4	53.0
	outcloseness average	24.1	52.8
Rnd150/150	flow betweeness	12.8%	43.3%
	incloseness average	2.5	41.6
	outcloseness average	5.7	41.6

Moreover, as shown in Table 5 above, for two illustrative web communities, Mat20 – small and real – and Rnd150 – big and random –, the flow betweenness and the closeness indicators (obtained with UCINET 6.0) increase after the web communities are redesigned, which is very encouraging for the methodology presented in this paper. In fact, the flow betweenness provides information about each domain's centrality in terms of the capacity that the domain has to contribute to the information flow among the community. The closeness indicators are weakness metrics calculated on the basis of the incoming (incloseness) and outgoing (outcloseness) arcs of a domain. A detailed study concerning social network indicators may be found in Wasserman and Faust (1994). The authors strongly believe that problems of diffusion and preservation of resources on the Web will definitely benefit from more balanced structures of the web communities, even if built above the naturally created forms of organization of the communities. On the other hand, this methodology can be used to compare the structure of different web communities or the same web community over different periods. In fact, from each structure the amount of redesigning needed can be calculated for comparison proposes.

Further work within this project will include experiments with other epistemic web communities and larger instances arising from Portugal's Ministry of Education project to update elementary education. Research work involving the study of a more robust comparison between web communities using data from recent projects undertaken at the Oxford Internet Institute on the World Wide Web of Science is also underway (Caldas et al. (2007b)).

Moreover, custom designed exact algorithms will also be required to obtain optimal solutions, at least for medium dimension instances of the Web Community Balancing Problem.

ACKNOWLEDGEMENT

Research was supported by PRODEP III (Medida 5, Acção 5.3) and POCTI-ISFL-1-152. The authors are also grateful for the referees' valuable comments.

REFERENCES

1. Abdinnour-Helm, S. (1998). A hybrid heuristic for the uncapacitated hub location problem. *European Journal of Operational Research*, 106: 489-499.
2. Abdinnour-Helm, S. and Venkataramanan, M. (1998). Solution approaches to hub location problems. *Annals of Operations Research*, 78: 31-50.
3. Batagelj, V. and Mrvar, A. (1998). Pajek, a program for large network analysis. *Connections*, 21(2): 47-57.
4. Borgatti, S., Everett, M.L., and Freeman, L.C. (Eds.) (1999). *UCINET 6.0 Version 1.00*, Analytic Technologies, Natick, Massachusetts.
5. Caldas, A. (2005). Galilei – A multi-agent system for the discovery of digital knowledge bases. Working paper, Oxford Internet Institute, University of Oxford, UK.
6. Caldas, A. (2007). Challenges in long-term archiving and preservation of government information. *Proceedings of the Workshop Preserving and Archiving Government Information in Digital Networks*, Oxford Internet Institute, University of Oxford, UK.
7. Caldas, A., Schroeder, R., Mesch, G., and Dutton, W. (2007a). The world wide web of science and democratisation of access to global sources of expertise. *Journal of Computer-Mediated Communication*, in press.
8. Caldas, A., Schroeder, R., Mesch, G., and Dutton, W. (2007b). Patterns of information search and access on the World Wide Web. *Journal of Computer-Mediated Communication*, in press.
9. Camargo, R., Miranda, G., and Lunac, H. (2008). Benders decomposition for the uncapacitated multiple allocation hub location problem. *Computers & Operations Research*, 35: 1047-1064.
10. Campbell, J. (1994). Integer programming formulations of discrete hub location problems. *European Journal of Operational Research*, 72: 387-405.
11. Correia, I. and Captivo, M. (2004). Bounds for the single source modular capacitated plant location problem. Working paper, Centro de Investigação Operacional, Universidade de Lisboa, Portugal.
12. Cortinhal, M. and Captivo, M. (2003). Upper and lower bounds for the single source capacitated location problem. *European Journal of Operational Research*, 151: 333-351.
13. Cplex (2002). *Using the Cplex Callable Library*, *Version 8.0*, ILOG, Inc.
14. Day, M. (2003). Preserving the fabric of our lives: A survey of Web preservation initiatives. *Proceedings of the*

7th European Conference on Research and Advanced Technology for Digital Libraries, Trondheim, Norway, pp. 17-22.

15. 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.
16. Ernst, A., Jiang, H., and Krishnamoorthy, M. (2005). Reformulations and computational results for uncapacitated single and multiple allocation hub covering problems. Working paper, CSIRO-Commonwealth Scientific and Industrial Research Organisation, Australia.
17. Flake, G., Lawrence, S., Giles, C., and Coetzee, F. (2002). Self-organization and identification of web communities. *Computer*, 35(3): 66-71.
18. Glover, F. (1989). Tabu search – Part I. *ORSA Journal on Computing*, 1(3): 190-206.
19. Glover, F. and Laguna, M. (1997). *Tabu Search*, Kluwer Academic Publishers, Boston.
20. Kara, B. and Tansel, B. (2000). On the single-assignment p-hub center problem. *European Journal of Operational Research*, 125: 648-655.
21. Kara, B. and Tansel, B. (2003). The single-assignment hub covering problem: Models and linearizations. *Journal of the Operational Research Society*, 54: 59-64.
22. Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). The web as a graph: Measurements, models, and methods. *Proceedings of the 5th International Conference on Computer and Combinatorics*, Tokyo, Japan, pp. 1-17.
23. Kleinberg, J. and Lawrence, S. (2001). The structure of the web. *Science*, 294: 1849-1850.
24. Klincewicz, J. (1991). Heuristics for the *p*-hub location problem. *European Journal of Operational Research*, 53: 25-37.
25. Klincewicz, J. (1996). A dual algorithm for the uncapacitated hub location problem. *Location Science*, 4(3): 173-184.
26. Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Trawling the web for emerging cyber-communities. *Computer Networks*, 31: 1481-1493.
27. O'Kelly, M. (1992). Hub facility location with fixed costs. *The Journal of the Regional Science Association International*, 71(3): 293-306.
28. Resende, M. and Ribeiro, C. (2002). Greedy randomized adaptive search procedures. In: F. Glover and G. Kochenberger (Eds.), *State-of-the-Art Handbook of Metaheuristics*, Kluwer, Norwell, pp. 219-249.
29. Rodríguez-Martín, I. and Salazar-González, J. (2008). Solving a capacitated hub location problem. *European Journal of Operational Research*, 184: 468-479.
30. Shorin-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.
31. Shorin-Kapov, D., Shorin-Kapov, J., and O'Kelly, M. (1996). Tight linear programming relaxations of uncapacitated *p*-hub median problems. *European Journal of Operational Research*, 94: 582-593.
32. 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.
33. Toyoda, M. and Kitsuregawa, M. (2001). A web community chart for navigating related communities. *Proceedings of the ACM Conference on Hypertext and Hypermedia*, Arhus, Denmark, pp. 103-112.
34. Tuzun, D. and Burke, L. (1999). A two-phase tabu search approach to the location routing problem. *European Journal of Operational Research*, 116: 87-89.
35. Wagner, B., (2004). Model formulations for hub covering problems. Working paper, Institute of Operations Research, Darmstadt University of Technology, Germany.
36. Wasserman, S. and Faust, K. (Eds.) (1994). *Social Network Analysis*, Cambridge University Press, Cambridge.