

# Profile-based optimal matchings in the Student/Project Allocation problem\*

Augustine Kwanashie<sup>1</sup>, Robert W. Irving<sup>1</sup>,  
David F. Manlove<sup>1,†</sup> and Colin T.S. Sng<sup>2,‡</sup>

<sup>1</sup> School of Computing Science, University of Glasgow, UK

<sup>2</sup> Amazon.com, Inc., Texas, USA

## Abstract

In the *Student / Project Allocation problem* (SPA) we seek to assign students to individual or group projects offered by lecturers. Students provide a list of projects they find acceptable in order of preference. Each student can be assigned to at most one project and there are constraints on the maximum number of students that can be assigned to each project and lecturer. We seek matchings of students to projects that are optimal with respect to *profile*, which is a vector whose  $r$ th component indicates how many students have their  $r$ th-choice project. We present an efficient algorithm for finding a *greedy maximum matching* in the SPA context – this is a maximum matching whose profile is lexicographically maximum. We then show how to adapt this algorithm to find a *generous maximum matching* – this is a matching whose reverse profile is lexicographically minimum. Our algorithms involve finding optimal flows in networks. We demonstrate how this approach can allow for additional constraints, such as lecturer lower quotas, to be handled flexibly. Finally we present results obtained from an empirical evaluation of the algorithms.

**Keywords:** Greedy maximum matching; Generous maximum matching; Matching profile; Augmenting path

## 1 Introduction

In most academic programmes students are usually required to take up individual or group projects offered by lecturers. Students may be required to rank a subset of the projects they find acceptable in order of preference. Each project is offered by a unique lecturer who may also be allowed to rank the projects she offers or the students who are interested in taking her projects in order of preference. Each student can be assigned to at most one project and there are usually constraints on the maximum number of students that can be assigned to each project and lecturer. The problem then is to assign students to projects in a manner that satisfies these capacity constraints while taking into account the preferences of the students and lecturers involved. This problem has been described in the literature as the *Student-Project Allocation problem* (SPA) [4, 21, 5, 16]. Variants of SPA also exist in which *lower quotas* are assigned to projects and/or lecturers. These

\*A preliminary version of this paper appeared in the proceedings of IWOCA 2014: the 25th International Workshop on Combinatorial Algorithms.

<sup>†</sup>Supported by Engineering and Physical Sciences Research Council grant EP/K010042/1. Corresponding author. Email [david.manlove@glasgow.ac.uk](mailto:david.manlove@glasgow.ac.uk).

<sup>‡</sup>Work done while at the School of Computing Science, University of Glasgow.

lower quotas indicate the minimum number of students to be assigned to each project and lecturer.

Although described in an academic context, applications of SPA need not be limited to assigning students to projects but may extend to other scenarios, such as the assignment of employees to posts in a company where available posts are offered by various departments. Applications of SPA in an academic context can be found at the University of Glasgow [29], the University of York [7, 18, 27], the University of Southampton [6, 10] and the Geneva School of Business Administration [28]. As previously stated, it is widely accepted that matching problems (like SPA) are best solved by centralised matching schemes where agents submit their preferences and a central authority computes an optimal matching that satisfies all the specified criteria [9]. Moreover the potentially large number of students and projects involved in these schemes motivates the need to discover efficient algorithms for finding optimal matchings.

### 1.1 Two-sided preferences and stability

In SPA, students are always required to provide preference lists over projects. However, variants of the problem may be defined depending on the presence and nature of lecturer preference lists. Some variants of SPA require both students and lecturers to provide preference lists. These variants include: (i) the *Student/Project Allocation problem with lecturer preferences over Students* (SPA-S) [4] which requires each lecturer to rank the students who find at least one of her offered projects acceptable, in order of preference, (ii) the *Student/Project Allocation problem with lecturer preferences over Projects* (SPA-P) [21, 16] which involves lecturers ranking the projects they offer in order of preference and (iii) the *Student/Project Allocation problem with lecturer preferences over Student-Project pairs* (SPA-(S,P)) [4, 5] where lecturers rank student-project pairs in order of preference. These variants of SPA have been studied in the context of the well-known *stability* solution criterion for matching problems [9]. The general stability objective is to produce a matching  $M$  in which no student-project pair that are not currently matched in  $M$  can simultaneously improve by being paired together (thus in the process potentially abandoning their partners in  $M$ ). A full description of the results relating to these SPA variants can be found in [20].

### 1.2 One-sided preferences and profile-based optimality

In many practical SPA applications it is considered appropriate to allow only students to submit preferences over projects. When preferences are specified by only one set of agents in a two-sided matching problem, the notion of stability becomes irrelevant. This motivates the need to adopt alternative solution criteria when lecturer preferences are not allowed. In this subsection we describe some of these solution criteria and briefly present results relating to them. These criteria consider the size of the matchings produced as well as the satisfaction of the students involved.

When preference lists of lecturers are absent, the SPA problem becomes a two-sided matching problem with one-sided preferences. We assume students' preference lists can contain ties in these SPA variants. Various optimality criteria for such problems have been studied in the literature [20]. Some of these criteria depend on the *profile* or the *cost* of a matching. In the SPA context, the *profile* of a matching is a vector whose  $r$ th component indicates the number of students obtaining their  $r$ th-choice project in the matching. The *cost* of a matching (w.r.t. the students) is the sum of the ranks of the assigned projects in the students' preference lists (that is, the sum of  $rx_r$  taken over all components  $r$  of the profile, where  $x_r$  is the  $r$ th component value).

students' preferences:	lecturers' offerings:
$s_1 : p_1 \quad p_2 \quad p_3$	$l_1 : \{p_1, p_2\}$
$s_2 : p_1$	$l_2 : \{p_3\}$
$s_3 : p_2 \quad p_3$	project capacities: $c_1 = 1, c_2 = 1, c_3 = 1$
	lecturer capacities: $d_1 = 2, d_2 = 1$

Figure 1: A SPA instance  $I$

A *minimum cost maximum matching* is a maximum cardinality matching with minimum cost. A *rank-maximal matching* is a matching that has lexicographically maximum profile [15, 13]. That is the maximum number of students are assigned to their first-choice project and subject to this, the maximum number of students are assigned to their second choice project and so on. However a rank maximal matching need not be a maximum matching in the given instance (see, e.g., [20, p.43]). Since it is usually important to match as many students as possible, we may first optimise the size of the matching before considering student satisfaction. Thus we define a *greedy maximum matching* [14, 22, 11] to be a maximum matching that has lexicographically maximum profile. The intuition behind both rank-maximal and greedy maximum matchings is to maximize the number of students matched with higher ranked projects. This may lead to some students being matched to projects that are relatively low on their preference lists. An alternative approach is to find a *generous maximum matching* which is a maximum matching in which the minimum number of students are matched to their  $R$ th-choice project (where  $R$  is the maximum length of any students' preference list) and subject to this, the minimum number of students are matched to their  $(R - 1)$ th-choice project and so on. Greedy and generous maximum matchings have been used to assign students to projects in the School of Computing Science, and students to elective courses in the School of Medicine, both at the University of Glasgow, since 2007. Figure 1 shows a sample SPA instance with greedy and generous maximum matchings, namely  $M_1 = \{(s_1, p_3), (s_2, p_1), (s_3, p_2)\}$  and  $M_2 = \{(s_1, p_2), (s_2, p_1), (s_3, p_3)\}$  respectively.

A special case of SPA, where each project is offered by a unique lecturer with an infinite upper quota and zero lower quota, can be modelled as the *Capacitated House Allocation problem with Ties* (CHAT). This is a variant of the well-studied *House Allocation problem* (HA) [12, 30] which involves the allocation of a set of indivisible goods (which we call houses) to a set of applicants. In CHAT, each applicant is required to rank a subset of the houses in order of preference with the houses having no preference over applicants. The applicants play the role of students and the houses play the role of projects and lecturers. As in the case of SPA, we seek to find a many-to-one matching comprising applicant-house pairs. Efficient algorithms for finding profile-based optimal matchings in CHAT have been studied in the literature [11, 14, 25, 22]. The most efficient of these is the  $O(R^*m\sqrt{n})$  algorithm for finding rank-maximal, greedy maximum and generous maximum matchings in CHAT problems due to Huang et al. [11] where  $R^*$  is the maximum rank of any applicant in the matching,  $m$  is the sum of all the preference list lengths and  $n$  is the total number of applicants and houses. These models however fail to address the issue of load balancing among lecturers. In order to keep the assignment of students fair each lecturer will typically have a minimum (lower quota) and maximum (capacity/upper quota) number of students they are expected to supervise. These numbers may vary for different lecturers according to other administrative and academic commitments.

The CHAT algorithms mentioned above are based on modelling the problem in terms of a bipartite graph with the aim of finding a matching in the graph which satisfies the stated

criteria. However a more flexible approach would be to model the problem as a network with the aim of finding a flow that can be converted to a matching which satisfies the stated criteria. SPA has also been investigated in the network flow context [2, 29] where a *minimum cost maximum flow algorithm* is used to find a minimum cost maximum matching and other profile-based optimal matchings. The model presented in [29] allows for lower quotas on lecturers and projects as well as alternative lecturers to supervise each project. By an appropriate assignment of edge weights in the network it is shown that a minimum cost maximum flow algorithm (due to Orlin [23]) can find rank maximal, generous maximum and greedy maximum matchings in a SPA instance. This takes  $O(m \log n(m + n \log n))$  time in the worst case, where  $m$  and  $n$  are the number of vertices and edges in the network respectively. In the SPA context this takes  $O(m_2^2 \log n_1 + m_2 n_1 \log^2 n_1)$  time where  $n_1$  is the numbers of students and  $m_2$  is the sum of all the students' preference list lengths. However this approach involves assigning exponentially large edge weights (see, e.g., [20, p.405]), which may be computationally infeasible for larger problem instances due to floating point inaccuracies in dealing with such high numbers. For example given a large SPA instance involving say,  $n_1 = 100$  students each ranking  $R = 10$  projects in order of preference, edge weights could potentially be of the order  $n_1^R = 100^{10} = 10^{20}$  (and arithmetic involving such weights could easily require more than the 15-17 significant figures available in a 64-bit double-precision floating representation). Since the flow algorithms involve comparing these edge weights, floating point precision errors could easily cause them to fail in practice. Moreover using the standard assumption that arithmetic on numbers of magnitude  $O(n_1)$  takes constant time, arithmetic on edge weights of magnitude  $O(n_1^R)$  would add an additional factor of  $O(R)$  onto the running time of Orlin's algorithm.

### 1.3 Other SPA models and approaches

The variants of SPA already discussed above have been motivated by both practical and theoretical interests. These variants are usually distinguished by the (i) feasibility and (ii) optimality criteria specific to them. In this section, we discuss some more SPA models found in the literature as well as other approaches that have been used to solve these problems. The techniques employed include *Integer Programming (IP)* [6, 28, 24, 17], [24, 17], *Constraint Programming (CP)* [7, 27], and others [26, 10, 19].

In [24], an IP model for SPA was presented with the aim of optimising the overall satisfaction of the students and the lecturers offering the projects (i.e., minimising the overall cost on both sides). In [6] an IP model was presented for SPA problems involving individual and group projects. Various objective functions were also employed (often in a hierarchical manner). These include minimising the cost, balancing the work-load among lecturers, maximising the number of students assigned and maximising the number of first-choice assignments (w.r.t. student preferences). In [28] a more general IP model for SPA which allows project lower quotas was also presented. However none of these models simultaneously consider profile-based optimality as well as upper and lower quota constraints.

### 1.4 Our contribution

In Section 2 we formally define the SPA model. In Section 3 we present an  $O(n_1^2 R m_2)$  time algorithm for finding a greedy maximum matching given a SPA instance and prove its correctness. The algorithm takes lecturer upper quotas into consideration. In Section 4 we show how this algorithm can be modified in order to find a generous maximum matching. Section 5 introduces lecturer lower quotas to the SPA model and shows how our algorithm can be modified to handle this variant. In Section 6 we present results from an empirical

evaluation of the algorithms described. We conclude the paper in Section 7 by presenting some open problems.

## 2 Preliminary definitions

An instance  $I$  of the SPA problem consists of a set  $\mathcal{S}$  of students, a set  $\mathcal{P}$  of projects and a set  $\mathcal{L}$  of lecturers. Each student  $s_i$  ranks a set  $A_i \subseteq \mathcal{P}$  of projects that she considers acceptable in order of preference. This *preference list* of projects may contain ties. Each project  $p_j \in \mathcal{P}$  has an upper quota  $c_j$  indicating the maximum number of students that can be assigned to it. Each lecturer  $l_k \in \mathcal{L}$  offers a set of projects  $P_k \subseteq \mathcal{P}$  and has an upper quota  $d_k^+$  indicating the maximum number of students that can be assigned to  $l_k$ . Unless explicitly mentioned, we assume that all lecturer lower quotas are equal to 0. The sets  $\{P_1, \dots, P_k\}$  partition  $\mathcal{P}$ . If project  $p_j \in P_k$ , then we denote  $l_k = l(p_j)$ .

An *assignment*  $M$  in  $I$  is a subset of  $\mathcal{S} \times \mathcal{P}$  such that:

1. Student-project pair  $(s_i, p_j) \in M$  implies  $p_j \in A_i$ .
2. For each student  $s_i \in \mathcal{S}$ ,  $|\{(s_i, p_j) \in M : p_j \in A_i\}| \leq 1$ .

If  $(s_i, p_j) \in M$  we denote  $M(s_i) = p_j$ . For a project  $p_j$ ,  $M(p_j)$  is the set of students assigned to  $p_j$  in  $M$ . Also if  $(s_i, p_j) \in M$  and  $p_j \in P_k$  we say student  $s_i$  is assigned to project  $p_j$  and to lecturer  $l_k$  in  $M$ . We denote the set of students assigned to a lecturer  $l_k$  as  $M(l_k)$ . A *matching* in this problem is an assignment  $M$  that satisfies the capacity constraints of the projects and lecturers. That is,  $|M(p_j)| \leq c_j$  for all projects  $p_j \in \mathcal{P}$  and  $|M(l_k)| \leq d_k^+$  for all lecturers  $l_k \in \mathcal{L}$ .

Given a student  $s_i$  and a project  $p_j \in A_i$ , we define  $rank(s_i, p_j)$  as  $1 +$  the number of projects that  $s_i$  prefers to  $p_j$ . Let  $R$  be the maximum rank of a project in any student's preference list. We define the *profile*  $\rho(M)$  of a matching  $M$  in  $I$  as an  $R$ -tuple  $(x_1, x_2, \dots, x_R)$  where for each  $r$  ( $1 \leq r \leq R$ ),  $x_r$  is the number of students  $s_i$  assigned in  $M$  to a project  $p_j$  such that  $rank(s_i, p_j) = r$ . Let  $\alpha = (x_1, x_2, \dots, x_R)$  and  $\sigma = (y_1, y_2, \dots, y_R)$  be any two profiles. We define the *empty profile*  $O_R = (o_1, o_2, \dots, o_R)$  where  $o_r = 0$  for all  $r$  ( $1 \leq r \leq R$ ). We also define the *negative infinity profile*  $B_R^- = (b_1, b_2, \dots, b_R)$  where  $b_r = -\infty$  ( $1 \leq r \leq R$ ) and the *positive infinity profile*  $B_R^+ = (b_1, b_2, \dots, b_R)$  where  $b_r = \infty$  ( $1 \leq r \leq R$ ). We define the sum of two profiles  $\alpha$  and  $\sigma$  as  $\alpha + \sigma = (x_1 + y_1, x_2 + y_2, \dots, x_R + y_R)$ . Given any  $q$  ( $1 \leq q \leq R$ ), we define  $\alpha + q = (x_1, \dots, x_{q-1}, x_q + 1, x_{q+1}, \dots, x_R)$ . We define  $\alpha - q$  in a similar way.

We define the total order  $\succ_L$  on profiles as follows. We say  $\alpha$  *left dominates*  $\sigma$ , denoted by  $\alpha \succ_L \sigma$  if there exists some  $r$  ( $1 \leq r \leq R$ ) such that  $x_{r'} = y_{r'}$  for  $1 \leq r' < r$  and  $x_r > y_r$ . We define *weak left domination* as follows. We say  $\alpha \succeq_L \sigma$  if  $\alpha = \sigma$  or  $\alpha \succ_L \sigma$ . We may also define an alternative total order  $\prec_R$  on profiles as follows. We say  $\alpha$  *right dominates*  $\sigma$  ( $\alpha \prec_R \sigma$ ) if there exists some  $r$  ( $1 \leq r \leq R$ ) such that  $x_{r'} = y_{r'}$  for  $r < r' \leq R$  and  $x_r < y_r$ . We also define *weak right domination* as follows. We say  $\alpha \preceq_R \sigma$  if  $\alpha = \sigma$  or  $\alpha \prec_R \sigma$ .

The SPA problem can be modelled as a network flow problem. Given a SPA instance  $I$ , we construct a flow network  $N(I) = \langle G, c \rangle$  where  $G = (V, E)$  is a directed graph and  $c$  is a non-negative capacity function  $c : E \rightarrow \mathbb{R}^+$  defining the maximum flow allowed through each edge in  $E$ . The network consists of a single source vertex  $v_s$  and sink vertex  $v_t$  and is constructed as follows. Let  $V = \{v_s, v_t\} \cup \mathcal{S} \cup \mathcal{P} \cup \mathcal{L}$  and  $E = E_1 \cup E_2 \cup E_3 \cup E_4$  where  $E_1 = \{(v_s, s_i) : s_i \in \mathcal{S}\}$ ,  $E_2 = \{(s_i, p_j) : s_i \in \mathcal{S}, p_j \in A_i\}$ ,  $E_3 = \{(p_j, l_k) : p_j \in \mathcal{P}, l_k = l(p_j)\}$  and  $E_4 = \{(l_k, v_t) : l_k \in \mathcal{L}\}$ . We set the capacities as follows:  $c(v_s, s_i) = 1$  for all  $(v_s, s_i) \in E_1$ ,  $c(s_i, p_j) = 1$  for all  $(s_i, p_j) \in E_2$ ,  $c(p_j, l_k) = c_j$  for all  $(p_j, l_k) \in E_3$  and  $c(l_k, v_t) = d_k^+$  for all  $(l_k, v_t) \in E_4$ .

We call a path  $P'$  from  $v_s$  to some project  $p_j$  a *partial augmenting path* if  $P'$  can be extended adding the edges  $(p_j, l(p_j))$  and  $(l(p_j), v_t)$  to form an augmenting path with respect to flow  $f$ . Given a partial augmenting path  $P'$  from  $v_s$  to  $p_j$ , we define the *profile* of  $P'$ , denoted  $\rho(P')$ , as follows:

$$\rho(P') = O_R + \sum \{ \text{rank}(s_i, p_j) : (s_i, p_j) \in P' \wedge f(s_i, p_j) = 0 \} - \sum \{ \text{rank}(s_i, p_j) : (p_j, s_i) \in P' \wedge f(s_i, p_j) = 1 \}$$

where additions are done with respect to the  $+$  and  $-$  operations on profiles. Unlike the profile of a matching, the profile of an augmenting path may contain negative values. Also if  $P'$  can be extended to a full augmenting path  $P$  with respect to flow  $f$  by adding the edges  $(p_j, l(p_j))$  and  $(l(p_j), v_t)$  where  $v_s$  and  $p_j$  are the endpoints of  $P'$ , then we define the profile of  $P$ , denoted by  $\rho(P)$ , to be  $\rho(P) = \rho(P')$ . Multiple partial augmenting paths may exist from  $v_s$  to  $p_j$ , thus we define the *maximum profile of a partial augmenting path* from  $v_s$  to  $p_j$  with respect to  $\succ_L$ , denoted  $\Phi(p_j)$ , as follows:

$$\Phi(p_j) = \max_{\succ_L} \{ \rho(P') : P' \text{ is a partial augmenting path from } v_s \text{ to } p_j \}.$$

An augmenting path  $P$  is called a *maximum profile augmenting path* if

$$\rho(P) = \max_{\succ_L} \{ \Phi(p_j) : p_j \in \mathcal{P} \}.$$

Let  $f$  be an integral flow in  $N$ . We define the matching  $M(f)$  in  $I$  induced by  $f$  as follows:  $M(f) = \{ (s_i, p_j) : f(s_i, p_j) = 1 \}$ . Clearly by construction of  $N$ ,  $M(f)$  is a matching in  $I$ , such that  $|M(f)| = |f|$ . If  $f$  is a flow and  $P$  is an augmenting path with respect to  $f$  then  $\rho(M') = \rho(M) + \rho(P)$  where  $M = M(f)$ ,  $M' = M(f')$  and  $f'$  is the flow obtained by augmenting  $f$  along  $P$ . Also given a matching  $M$  in  $I$ , we define a flow  $f(M)$  in  $N$  corresponding to  $M$  as follows:

$$\begin{aligned} \forall (v_s, s_i) \in E_1, f(v_s, s_i) &= 1 \text{ if } s_i \text{ is matched in } M \text{ and } f(v_s, s_i) = 0 \text{ otherwise.} \\ \forall (s_i, p_j) \in E_2, f(s_i, p_j) &= 1 \text{ if } (s_i, p_j) \in M \text{ and } f(s_i, p_j) = 0 \text{ otherwise.} \\ \forall (p_j, l_k) \in E_3, f(p_j, l_k) &= c'_j \text{ where } c'_j = |M(p_j)| \\ \forall (l_k, v_t) \in E_4, f(l_k, v_t) &= d'_k \text{ where } d'_k = |M(l_k)| \end{aligned}$$

We define a student  $s_i$  to be *exposed* if  $f(v_s, s_i) = 0$  meaning that there is no flow through  $s_i$ . Similarly we define a project  $p_j$  to be *exposed* if  $f(p_j, l_k) < c_j$  and  $f(l_k, v_t) < d_k^+$  where  $l_k = l(p_j)$ .

Let  $M$  be a matching of size  $k$  in  $I$ . We say that  $M$  is a *greedy  $k$ -matching* if there is no other matching  $M'$  such that  $|M'| = k$  and  $\rho(M') \succ_L \rho(M)$ . If  $k$  is the size of a maximum cardinality matching in  $I$ , we call  $M$  a *greedy maximum matching* in  $I$ . Also we say that  $M$  is a *generous  $k$ -matching* if there is no other matching  $M'$  such that  $|M'| = k$  and  $\rho(M') \prec_R \rho(M)$ . If  $k$  is the size of a maximum cardinality matching in  $I$ , we call  $M$  a *generous maximum matching* in  $I$ . We also define the *degree* of a matching  $M$  to be the rank of one of the worst-off students matched in  $M$  or 0 if  $M$  is an empty set.

### 3 Greedy maximum matchings in SPA

In this section we present the algorithm **Greedy-max-spa** for finding a greedy maximum matching given a SPA instance. The algorithm is based on the general Ford-Fulkerson algorithm for finding a maximum flow in a network [8]. We obtain maximum profile augmenting paths by adopting techniques used in the bipartite matching approach for finding a greedy maximum matching in HA [14] and CHAT [25].

The **Greedy-max-spa** algorithm shown in Algorithm 1 takes in a SPA instance  $I$  as input and returns a greedy maximum matching  $M$  in  $I$ . A flow network  $N(I) = \langle G, c \rangle$  is constructed as described in Section 2. Given a flow  $f$  in  $N(I)$  that yields a greedy  $k$ -matching  $M(f)$  in  $I$ , if  $k$  is not the size of a maximum flow in  $N(I)$ , we seek to find a maximum profile augmenting path  $P$  with respect to  $f$  in  $N(I)$  such that the new flow  $f'$  obtained by augmenting  $f$  along  $P$  yields a greedy  $(k+1)$ -matching  $M(f')$  in  $I$ . Lemmas 3.1 and 3.2 show the correctness of this approach. We firstly show that if  $k$  is smaller than the size of a maximum flow in  $N(I)$  then such a path is bound to exist.

**Lemma 3.1.** *Let  $I$  be an instance of SPA and let  $\eta$  denote the size of a maximum matching in  $I$ . Let  $k$  ( $1 \leq k < \eta$ ) be given and suppose that  $M_k$  is a greedy  $k$ -matching in  $I$ . Let  $N = N(I)$  and  $f = f(M_k)$ . Then there exists an augmenting path  $P$  with respect to  $f$  in  $N$  such that if  $f'$  is the result of augmenting  $f$  along  $P$  then  $M_{k+1} = M(f')$  is a greedy  $(k+1)$ -matching in  $I$ .*

*Proof.* Let  $I' = C(I)$  be a new instance of SPA obtained from  $I$  as follows. Firstly we add all students in  $I$  to  $I'$ . Next, for every project  $p_j \in \mathcal{P}$ , we add  $c_j$  clones  $p_j^1, p_j^2, \dots, p_j^{c_j}$  to  $I'$  each of capacity 1. We then add all lecturers in  $I$  to  $I'$ . If  $p_j \in A_i$  in  $I$ , we add  $(s_i, p_j^r)$  to  $I'$  for all  $r$  ( $1 \leq r \leq c_j$ ). If  $p_j \in P_k$  is in  $I$ , we add  $(p_j^r, l_k)$  to  $I'$  for all  $r$  ( $1 \leq r \leq c_j$ ). Also if  $\text{rank}(s_i, p_j) = t$ , we set  $\text{rank}(s_i, p_j^r) = t$  for all  $r$  ( $1 \leq r \leq c_j$ ). Let  $G'$  be the underlying graph in  $I'$  involving only the student and project clones. With respect to the matching  $M_k = M(f)$ , we construct a cloned matching  $C(M_k)$  in  $I'$  as follows. If project  $p_j$  is assigned  $x_j$  students  $s_{q,1}, s_{q,2}, \dots, s_{q,x_j}$  in  $M_k$  we add  $(s_{q,r}, p_j^r)$  to  $C(M_k)$  for all  $1 \leq r \leq x_j$ . Hence  $C(M_k)$  is a greedy  $k$ -matching in  $I'$ .

Let  $M'_{k+1}$  be a greedy  $(k+1)$ -matching in  $I$  (this exists because  $k < \eta$ ). Then  $C(M'_{k+1})$  is a greedy  $(k+1)$ -matching in  $I'$ . Let  $X = C(M_k) \oplus C(M'_{k+1})$ . Then each connected component of  $X$  is either (i) an alternating cycle, (ii) an even-length alternating path or (iii) an odd-length alternating path in  $G'$  (with no restrictions on which matching the end edges belong to). The aim is to show that, by eliminating a subset of  $X$ , we are left with a set of connected components which can be transformed into a single augmenting path with respect to  $f(C(M_k))$  in  $N(I')$  and subsequently a single augmenting path with respect to  $f(M_k)$  in  $N(I)$ .

**Eliminating connected components of  $X$ :** Suppose  $D \subseteq X$  is a type (i) connected component of  $X$  or a type (ii) connected component of  $X$  whose end vertices are students (we may call this a type (ii)(a) component). Suppose also that  $\rho(D \cap C(M'_{k+1})) \succ_L \rho(D \cap C(M_k))$ . A new matching  $C(M'_k)$  in  $G'$  of cardinality  $k$  can be created from  $C(M_k)$  by replacing all the  $C(M_k)$ -edges in  $D$  with the  $C(M'_{k+1})$ -edges in  $D$  (i.e. by augmenting  $C(M_k)$  along  $D$ ). Since the upper quota constraints of the lecturers involved are not violated after creating  $C(M'_k)$  from  $C(M_k)$ , it follows that  $C(M'_k)$  is also a valid SPA matching in  $I'$ . Moreover  $\rho(C(M'_k)) \succ_L \rho(C(M_k))$  which is a contradiction to the fact that  $C(M_k)$  is a greedy  $k$ -matching in  $I'$ . A similar contradiction (to the fact that  $C(M'_{k+1})$  is a greedy  $(k+1)$ -matching in  $I'$ ) exists if we assume  $\rho(D \cap C(M_k)) \succ_L \rho(D \cap C(M'_{k+1}))$ . Thus  $\rho(D \cap C(M'_{k+1})) = \rho(D \cap C(M_k))$ .

Form the argument above, no type (i) or type (ii)(a) connected component of  $X$  contributes to a change in the size or profile as we augment from  $C(M_k)$  to  $C(M'_{k+1})$  or vice versa. In fact, this is true for any even-length connected component of  $X$  which does not cause lecturer upper quota constraints to be violated as we augment from  $C(M_k)$  to  $C(M'_{k+1})$  or vice versa. The claim can further be extended to certain groups of connected components which, when considered together, (i) have equal numbers of  $C(M_k)$  and  $C(M'_{k+1})$  edges and (ii) do not cause lecturer upper quota constraints to be violated

as we augment from  $C(M_k)$  to  $C(M'_{k+1})$  or vice versa. In all these cases, it is possible to *eliminate* such components (or groups of components) from consideration. Using the above reasoning, we begin by eliminating all type (i) and type (ii)(a) connected components of  $X$ .

Let  $\mathcal{D}$  be the union of all the edges in type (i) and type (ii)(a) connected components of  $X$ . Let  $X' = X \setminus \mathcal{D}$ . Then it follows that  $X' = C(M_k) \oplus C(M''_{k+1})$  for some greedy  $(k+1)$ -matching  $C(M''_{k+1})$  in  $I'$  which can be constructed by augmenting  $C(M'_{k+1})$  along all type (i) and type (ii)(a) components of  $X$ . Thus  $X'$  contains (1) even-length alternating paths whose end vertices are project clones (we call these type (ii)(b) paths), (2) odd-length alternating paths whose end edges are in  $C(M_k)$  (we call these type (iii)(a) paths) and (3) odd-length alternating paths whose end edges are in  $C(M''_{k+1})$  (we call these type (iii)(b) paths). Although these alternating paths are vertex disjoint, there are special cases where two alternating paths in  $X'$  may be *joined together* by pairing their end project clone vertices.

**Joining alternating paths:** Consider some lecturer  $l_q$  and project  $p_j \in P_q$ . We extend the notation  $l(p_j)$  to include all clones of  $p_j$  (i.e.  $l(p_j^r) = l_q$  for all  $r$  ( $1 \leq r \leq c_j$ )). Let

$$X_q = \{(s_i, p_j^r) \in C(M_k) : l_q = l(p_j^r) \wedge p_j^r \text{ is unmatched in } C(M''_{k+1})\} \text{ and } x_q = |X_q|$$

Thus  $X_q$  is the set of end edges incident to project clones belonging to a subset of the type (ii)(b) and type (iii)(a) paths in  $X'$ . Let

$$Y_q = \{(s_i, p_j^r) \in C(M''_{k+1}) : l_q = l(p_j^r) \wedge p_j^r \text{ is unmatched in } C(M_k)\} \text{ and } y_q = |Y_q|$$

Thus  $Y_q$  is the set of end edges incident to project clones belonging to a subset of the type (ii)(b) and type (iii)(b) paths in  $X'$ . Also let

$$Z_q = \{p_j^r : l_q = l(p_j^r) \wedge p_j^r \text{ is matched in } C(M''_{k+1}) \wedge p_j^r \text{ is matched in } C(M_k)\} \text{ and } z_q = |Z_q|$$

Thus  $d_q = v_q + x_q + z_q$  and  $d'_q = v'_q + y_q + z_q$  where  $v_q$  and  $v'_q$  are the number of unassigned positions that  $l_q$  has in  $C(M_k)$  and  $C(M''_{k+1})$  respectively.

Note that  $v_q \geq y_q$  if and only if  $v'_q \geq x_q$ . If  $v_q \geq y_q$  then all the paths with end edges in  $Y_q$  can be considered as valid alternating paths in  $C(M_k)$  (i.e. if they are used to augment  $C(M_k)$ ,  $l_q$ 's upper quota will not be violated in the resulting matching). Since  $v'_q \geq x_q$  then all the paths with end edges in  $X_q$  can be considered as valid alternating paths in  $C(M''_{k+1})$  (i.e. if they are used to augment  $C(M''_{k+1})$ ,  $l_q$ 's upper quota will not be violated in the resulting matching).

On the other hand, assume  $y_q > v_q$ . Then  $x_q > v'_q$ . Let  $Y'_q \subseteq Y_q$  be an arbitrary subset of  $Y_q$  of size  $v_q$  and let  $X'_q \subseteq X_q$  be an arbitrary subset of  $X_q$  of size  $v'_q$ . Thus all paths with end edges in  $X'_q$  and  $Y'_q$  can be considered as valid alternating paths in  $C(M''_{k+1})$  and  $C(M_k)$  respectively. Also  $|Y_q \setminus Y'_q| = y_q - v_q = |X_q \setminus X'_q| = x_q - v'_q$ . We can thus form a 1 – 1 correspondence between the edges in  $|Y_q \setminus Y'_q|$  and those in  $|X_q \setminus X'_q|$ . Let  $(s_i, p_j^r) \in Y_q \setminus Y'_q$  and  $(s_{i'}, p_{j'}^{r'}) \in X_q \setminus X'_q$  be the end edges of two alternating paths in  $X'$ . The paths can be joined together by *pairing* the clones of both end projects thus forming a *project pair*  $(p_j^r, p_{j'}^{r'})$  at  $l_q$ . These project pairs can be formed from all edges in  $Y_q \setminus Y'_q$  and  $X_q \setminus X'_q$ .

In the cases where project pairs are formed, the resulting path (which we call a *compound path*) may be regarded as a single path along which  $C(M_k)$  or  $C(M''_{k+1})$  may be augmented. In some cases, the two projects being paired may be end vertices of a single (or compound) alternating path. Thus pairing them together will form a cycle. Since the cycle is of even length and the lecturer's upper quota will not be violated if it is used to



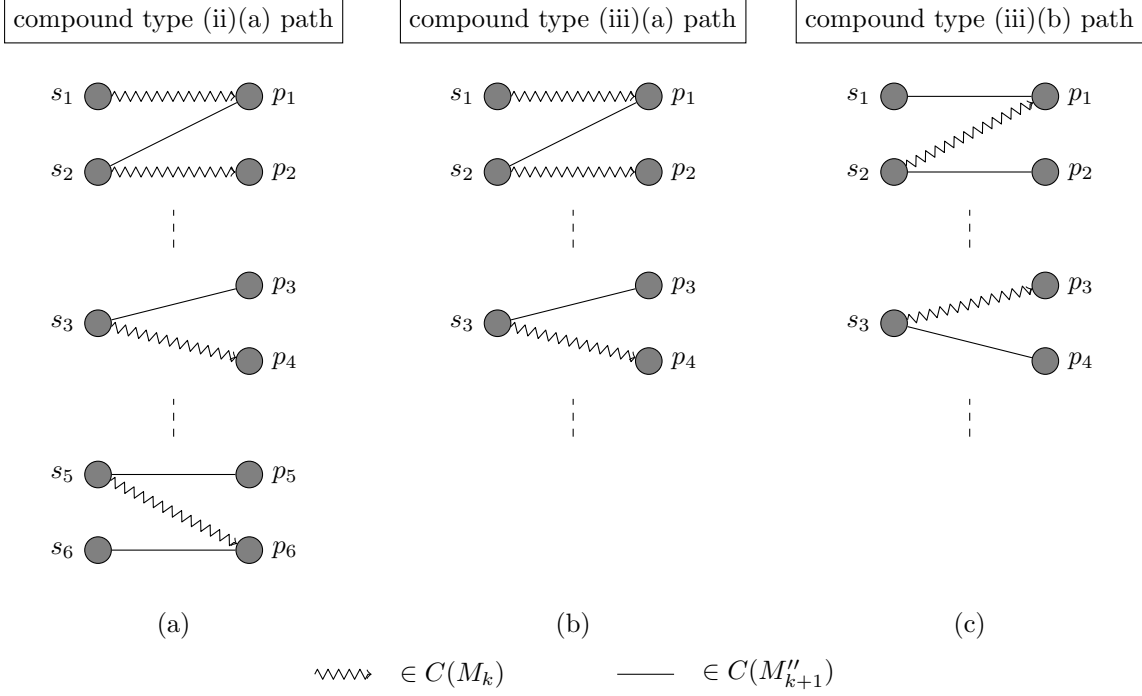


Figure 2: Some types of compound path in  $X'$

augment  $C(M_k)$  or  $C(M''_{k+1})$  it can be eliminated right away. For each lecturer  $l_q \in \mathcal{L}$ , once the pairings between alternating paths in  $Y_q \setminus Y'_q$  and  $X_q \setminus X'_q$  have been carried out (where applicable) and any formed cycles have been eliminated, we are left with a set of single or compound alternating paths of the following types (for simplicity we call all remaining alternating paths *compound paths* even though they may consist of only one path).

1. A *compound type (ii)(a) path* - a compound path with an even number of edges with both end vertices being students. This path will contain a type (iii)(a) path at one end, and a type (iii)(b) path at the other end with zero or more type (ii)(b) paths in between (See Figure 2(a)). Such a path can be eliminated from consideration.
2. A *compound type (ii)(b) path* - a compound path with an even number of edges with both end vertices being project clones. This path will contain one or more type (ii)(b) paths joined together. Such a path can also be eliminated from consideration as its end edges are incident to exposed project clones.
3. A *compound type (iii)(a) path* - a compound path with an odd number of edges with both end edges being matched in  $C(M_k)$ . This path will contain a type (iii)(a) path at one end with zero or more type (ii)(b) paths joined to it (See Figure 2(b)). We will consider these paths for elimination later in this proof.
4. A *compound type (iii)(b) path* - a compound path with an odd number of edges with both end edges being matched in  $C(M''_{k+1})$ . This path will contain a type (iii)(b) path at one end with zero or more type (ii)(b) paths joined to it (See Figure 2(c)). We will consider these paths for elimination later in this proof.

**Eliminating compound paths:** At this stage we are left with only compound type (iii)(a) and compound type (iii)(b) paths in  $X'$ . These paths, if considered independently decrease and increase the size of  $C(M_k)$  by 1 respectively. Since  $|C(M''_{k+1})| = |C(M_k)| + 1$

then there are  $q$  type (iii)(a) paths and  $(q + 1)$  type (iii)(b) paths. Consider some compound type (iii)(b) path  $D'$  and some compound type (iii)(a) path  $D''$ . Then we can consider the combined effect of augmenting  $C(M_k)$  or  $C(M''_{k+1})$  along  $D' \cup D''$ . Suppose that  $\rho((D' \cup D'') \cap C(M''_{k+1})) \succ_L \rho((D' \cup D'') \cap C(M_k))$ . A new matching  $C(M''_k)$  in  $G'$  of cardinality  $k$  can be created by augmenting  $C(M_k)$  along  $D' \cup D''$ . Since the upper quota constraints on the lecturers involved are not violated after creating  $C(M''_k)$  from  $C(M_k)$ , then  $C(M''_k)$  is also a valid SPA matching in  $I'$ . Thus  $\rho(C(M''_k)) \succ_L \rho(C(M_k))$  which is a contradiction to the fact that  $C(M_k)$  is a greedy  $k$ -matching in  $I'$ . A similar contradiction (to the fact that  $C(M''_{k+1})$  is a greedy  $(k + 1)$ -matching in  $I'$ ) exists if we assume  $\rho((D' \cup D'') \cap C(M_k)) \succ_L \rho((D' \cup D'') \cap C(M''_{k+1}))$ . Thus  $\rho((D' \cup D'') \cap C(M''_{k+1})) = \rho((D' \cup D'') \cap C(M_k))$ . It follows that, considering  $D'$  and  $D''$  together, the size and profile of the matching is unaffected as augment from  $C(M_k)$  to  $C(M''_{k+1})$  or vice versa and so both  $D'$  and  $D''$  can be eliminated from consideration.

**Generating an augmenting path in  $N(I)$ :** Once all these eliminations have been done, since  $|C(M''_{k+1})| = |C(M_k)| + 1$  it is easy to see that there remains only one path  $P'$  left in  $X'$  which is a compound type (iii)(b) path. The path  $P'$  can then be transformed to a component  $D$  in  $G(I)$  (where  $G(I)$  is basically the undirected counterpart of  $N(I)$  without capacities) by replacing all the project clones  $p_j^r$  ( $1 \leq r \leq c_j$ ) in  $P'$  with the original project  $p_j$  and, for every joined pair of project clones  $(p_j^r, p_j^{r'})$ , adding the lecturer  $l(p_j^r) = l(p_j^{r'})$  in between them. Thus a project may now appear more than once in  $D$ . A lecturer may also appear more than once in  $D$ .

Consider some project  $p_j \in D$  that appears more than once. Then let  $P'' \subset P'$  be the path consisting of edges between the first and last occurrence of the  $p_j$  clones in  $P'$  ( $P''$  corresponds to a collection of cycles belonging to  $D$  in  $G(I)$  involving  $p_j$ ). Thus  $P''$  is of even length and both end projects of  $P''$  are clones of the same project. Augmenting  $C(M_k)$  or  $C(M''_{k+1})$  along  $P''$  will not violate the lecturer upper quota constraints or affect the size or profile of the matching obtained (again using the same arguments presented above). Thus  $P''$  can be eliminated from consideration. Although this potentially breaks  $P'$  into two separate paths in  $G(I')$  it still remains connected in  $G(I)$ . Similarly consider some lecturer  $l_k \in D$  that appears more than once. Then let  $P''' \subset P'$  be the path consisting of edges between the first and last occurrence of the  $l_k$  clones in  $P'$  ( $P'''$  corresponds to a collection of type (ii)(b) paths with project clones offered by  $l_k$ ). Thus augmenting  $C(M_k)$  or  $C(M''_{k+1})$  along  $P'''$  will not violate the lecturer upper quota constraints or affect the size or profile of the matching obtained (again using the same arguments presented above). Thus  $P'''$  can be eliminated from consideration. Doing the above steps continually for all projects and lecturers that occur more than once in  $D$  eventually yields a valid path in  $G(I)$  in which all nodes are visited only once.

Finally we describe how the path  $D$  in  $G(I)$ , obtained after removing duplicate projects and lecturers, can be transformed to an augmenting path  $P$  in  $N(I)$  (i.e. we establish the direction of flow from  $v_s$  to  $v_t$  through  $P$  in  $N(I)$ ). Firstly we add the edge  $(v_s, s_i)$  to  $P$  where  $s_i$  is the exposed student in  $D$ . Next for every edge  $(s_{i'}, p_{j'}) \in M''_{k+1} \cap D$  we add a forward edge  $(s_{i'}, p_{j'})$  to  $P$ . Also for every edge  $(s_{i''}, p_{j''}) \in M_k \cap D$  we add a backward edge  $(p_{j''}, s_{i''})$  to  $P$ . Finally we add the edges  $(p_j, l(p_j))$  and  $(l(p_j), v_t)$  to  $P$  where  $p_j^r$  is the end project vertex in  $D$ . Thus  $P$  is an augmenting path with respect to  $f = f(M_k)$  in  $N(I)$  such that if  $f'$  is the flow obtained when  $f$  is augmented along  $P$  then  $M(f')$  is a greedy  $(k + 1)$ -matching in  $N(I)$ .  $\square$

**Lemma 3.2.** *Let  $f$  be a flow in  $N$  and let  $M_k = M(f)$ . Suppose that  $M_k$  is a greedy  $k$ -matching. Let  $P$  be a maximum profile augmenting path with respect to  $f$ . Let  $f'$  be the*

---

**Algorithm 1** Greedy-max-spa

---

**Require:** SPA instance  $I$ ;  
**Ensure:** return matching  $M$ ;  
1: define flow network  $N(I) = \langle G, c \rangle$ ;  
2: define empty flow  $f$ ;  
3: **loop**  
4:    $P = \text{Get-max-aug}(N(I), f)$ ;  
5:   **if**  $P \neq \text{null}$  **then**  
6:     augment  $f$  along  $P$ ;  
7:   **else**  
8:     return  $M(f)$ ;  
9:   **end if**  
10: **end loop**

---

flow obtained by augmenting  $f$  along  $P$ . Now let  $M_{k+1} = M(f')$ . Then  $M_{k+1}$  is a greedy  $(k+1)$ -matching.

*Proof.* Suppose for a contradiction that  $M_{k+1}$  is not a greedy  $(k+1)$ -matching. By Lemma 3.1, there exists an augmenting path  $P'$  with respect to  $f$  such that if  $f'$  is the result of augmenting  $f$  along  $P'$  then  $M'_{k+1} = M(f')$  is a greedy  $(k+1)$ -matching. Hence  $\rho(M'_{k+1}) \succ_L \rho(M_{k+1})$ . Since  $\rho(M'_{k+1}) = \rho(M) + \rho(P')$  and  $\rho(M_{k+1}) = \rho(M) + \rho(P)$ , it follows that  $\rho(P') \succ_L \rho(P)$ , a contradiction to the assumption that  $P$  is a maximum profile augmenting path.  $\square$

The **Get-max-aug** algorithm shown in Algorithm 2 accepts a flow network  $N(I)$  and flow  $f$  as input and finds an augmenting path of maximum profile relative to  $f$  or reports that none exists. The latter case implies that  $M(f)$  is already a greedy maximum matching. The method consists of three phases: an initialisation phase (lines 1 -15), the main phase which is a loop containing two other loops (lines 16 - 43) and a final phase (lines 44 - 52) where the augmenting path is generated and returned.

For each project  $p_j$  the **Get-max-aug** method maintains a variable  $\rho(p_j)$  describing the profile of a partial augmenting path  $P'$  from some exposed student to  $p_j$ . It also maintains, for every project  $p_j \in \mathcal{P}$ , a pointer  $\text{pred}(p_j)$  to the student or lecturer preceding  $p_j$  in  $P'$ . For every lecturer  $l_k \in \mathcal{L}$  a pointer  $\text{pred}(l_k)$  is also used to refer to any project preceding  $l_k$  in  $P'$ . Thus the final augmenting path produced will pass through each lecturer or project at most once. The initialisation phase of the method involves setting all  $\text{pred}$  pointers to **null** and  $\rho$  profiles to  $B_R^-$ . Next, the method seeks to find, for each project  $p_j$ , a partial augmenting path  $((v_s, s_i), (s_i, p_j))$  from the source, through an exposed student  $s_i$  to  $p_j$  should one exist. In the presence of multiple paths satisfying this criterion, the path with the best profile (w.r.t.  $\succ_L$ ) is selected. The variables  $\text{pred}(p_j)$  and  $\rho(p_j)$  are updated accordingly. Thus at the end of this phase  $\rho(p_j)$  indicates the maximum profile of an augmenting path of length 2 via some exposed student to  $p_j$  should one exist. If such a path does not exist then  $\rho(p_j)$  and  $\text{pred}(p_j)$  remain  $B_R^-$  and **null** respectively.

In the main phase, the algorithm then runs  $|f|$  iterations, at each stage attempting to increase the quality (w.r.t.  $\succ_L$ ) of the augmenting paths described by the  $\rho$  profiles. Each iteration runs two loops. Each loop identifies cases where the flow through one edge in the network can be reduced in order to allow the flow through another to be increased while improving the profile of the projects involved. In both loops, the decision on whether to switch the flow between candidate edges is made based on an edge relaxation operation similar to that used in the Bellman-Ford algorithm for solving the single source shortest path problem in which edge weights may be negative. In the first loop, we seek to evaluate the gain that may be derived from switching the flow through a student from one project

to another. Given an edge  $(s_i, p_k)$  with a flow of 1 in  $f$  and edge  $(s_i, p_j)$  with no flow in  $f$ , we define  $\sigma$  to be the resulting profile of  $p_j$  if the partial augmenting path ending at  $p_k$  is to be extended (via  $s_i$ ) to  $p_j$ . Thus  $\sigma$  will become the new value of  $\rho(p_j)$  should this extension take place. If  $\sigma \succ_L \rho(p_j)$  (i.e. if the proposed profile is better than the current one), we extend the augmenting path to  $p_j$  and update  $\rho(p_j) = \sigma$  and  $\text{pred}(p_j) = s_i$ .

In the second loop, we seek to evaluate the gain that may be derived from switching flow to some lecturer from one project to another. Given a lecturer  $l_k$ , let  $P'_k \subseteq P_k$  be the set of projects offered by  $l_k$  with positive outgoing flow and  $P''_k \subseteq P_k$  be the set of projects offered by  $l_k$  that are undersubscribed in  $M(f)$ . Then we seek to determine if an improvement can be obtained by switching a unit of flow from some project  $p_j \in P'_k$  to some other project  $p_m \in P''_k$ . This is achieved by comparing the  $\rho(p_j)$  and  $\rho(p_m)$  profiles and updating  $\rho(p_j) = \rho(p_m)$ ,  $\text{pred}(p_j) = l_k$  and  $\text{pred}(l_k) = p_m$  if  $\rho(p_m) \succ_L \rho(p_j)$  where  $\rho(p_m)$  represents the profile of a partial augmenting path that does not already pass through  $l_k$  (i.e.,  $\text{pred}(p_m) \neq l_k$ ). This means that the partial augmenting path ending at  $p_m$  can be extended further (via  $l_k$ ) to  $p_j$  while improving its profile. The intuition is that, after augmenting along such a path,  $p_m$  gains an extra student while  $p_j$  loses one.

During the final phase, we iterate through all exposed projects and find the one with the largest profile with respect to  $\succ_L$  (say  $p_q$ ). An augmenting path is then constructed through the network using the  $\text{pred}$  values of the projects and lecturers and the matched edges in  $M(f)$  starting from  $p_q$ . The generated path is returned to the calling algorithm. If no exposed project exists, the method returns **null**. We next show that **Get-max-aug** method produces such a maximum profile augmenting path in  $N$  with respect to  $f$  should one exist.

**Lemma 3.3.** *Given a SPA instance  $I$ , let  $f$  be a flow in  $N = N(I)$  where  $k = |f|$  is not the size of a maximum matching in  $I$  and  $M(f)$  is a greedy  $k$ -matching in  $I$ . Algorithm **Get-max-aug** finds a maximum profile augmenting path in  $N$  with respect to  $f$ .*

*Proof.* Consider some project  $p_j$  in  $\mathcal{P}$ . For any  $q$  ( $0 \leq q \leq k$ ) and for any  $r$  ( $0 \leq r \leq k$ ), we define  $\Phi_{2q+1,2r}(p_j)$  to be the maximum profile of any partial augmenting path with respect to  $f$  in  $N$  that starts at an exposed student, ends at  $p_j$ , and involves at most  $2q + 1$  student-project edges and at most  $2r$  project-lecturer edges. We represent the length of such a path using the pair  $(2q + 1, 2r)$ . Thus  $\Phi_{2k+1,2k}(p_j)$  gives the maximum profile of any partial augmenting path starting at an exposed student and ending at  $p_j$ . If such a path does not exist then  $\Phi_{2k+1,2k}(p_j) = B_R^-$ . Firstly we seek to show that after  $q$  iterations of the main loop of **Get-max-aug** where  $0 \leq q \leq k$ ,  $\rho_q(p_j) \succeq_L \Phi_{2q+1,2q}(p_j)$  for every project  $p_j \in \mathcal{P}$  where  $\rho_q(p_j)$  is the profile computed at  $p_j$  after  $q$  iterations of the main loop.

We prove this inductively. For the base case, let  $q = 0$ . Then  $\Phi_{1,0}(p_m)$  is the maximum profile of any partial augmenting path of length  $(1, 0)$  from an exposed student to project  $p_m$ . Hence, from the initialisation phase of **Get-max-aug**,  $\rho_0(p_m) = \Phi_{1,0}(p_m)$  and thus  $\rho_0(p_m) \succeq_L \Phi_{1,0}(p_m)$ . For the inductive step, assume  $1 \leq q \leq k$  and that the claim is true after the  $(q - 1)^{\text{th}}$  iteration (i.e.  $\rho_{q-1}(p_m) \succeq_L \Phi_{2q-1,2q-2}(p_m)$  for any  $p_m \in \mathcal{P}$ ). We will show that the claim is true for the  $q^{\text{th}}$  iteration (i.e.  $\rho_q(p_m) \succeq_L \Phi_{2q+1,2q}(p_m)$ ).

For each project  $p_m \in \mathcal{P}$  let  $S'_m = \{s_i \in \mathcal{S} : (s_i, p_m) \in E \wedge f(s_i, p_m) = 0\}$  and for each lecturer  $l_k \in \mathcal{L}$  let  $P'_k = \{p_m \in \mathcal{P} : l_k = l(p_m) \wedge f(p_m, l_k) < c_m\}$ . For each iteration of the main loop, we perform a relaxation step involving some student-project pair  $(s_i, p_m)$  where  $s_i \in S'_m$  and/or a relaxation step involving some project-lecturer pair  $(p_m, l_k)$  where  $p_m \in P'_k$ . Consider some project  $p_m$ . If there does not exist a partial augmenting path from an exposed student to  $p_m$ , of length  $\leq (2q + 1, 2q - 2)$  and with a better profile than  $\Phi_{2q-1,2q-2}(p_m)$ , then  $\Phi_{2q+1,2q-2}(p_m) = \Phi_{2q-1,2q-2}(p_m)$ . Otherwise there exists a partial

---

**Algorithm 2** Get-max-aug (method for Greedy-max-spa)

---

**Require:** flow network  $N(I) = \langle G, c \rangle$  where  $G = (V, E)$ , flow  $f$  where  $M(f)$  is a greedy  $|f|$ -matching;

```
1: /* initialisation */
2: for project  $p_j \in \mathcal{P}$  do
3:    $\rho(p_j) = B_R^-$ ;
4:    $pred(p_j) = \text{null}$ ;
5:   for all exposed student  $s_i \in S$  such that  $p_j \in A_i$  do
6:      $\sigma = O_R + rank(s_i, p_j)$ ;
7:     if  $\sigma \succ_L \rho(p_j)$  then
8:        $\rho(p_j) = \sigma$ ;
9:        $pred(p_j) = s_i$ ;
10:    end if
11:  end for
12: end for
13: for lecturer  $l_k \in \mathcal{L}$  do
14:    $pred(l_k) = \text{null}$ ;
15: end for
16: /* main phase */
17: for  $1 \dots |f|$  do
18:   /* first loop */
19:   for all  $(s_i, p_j) \in E$  where  $f(s_i, p_j) = 0$  and  $f(s_i, p_k) = 1$  for some  $p_k \in A_i$  do
20:      $\sigma = \rho(p_k) - rank(s_i, p_k) + rank(s_i, p_j)$ ;
21:     if  $\sigma \succ_L \rho(p_j)$  then
22:        $\rho(p_j) = \sigma$ ;  $pred(p_j) = s_i$ ;
23:     end if
24:   end for
25:   /* second loop */
26:   for all lecturer  $l_k \in \mathcal{L}$  do
27:      $\sigma = B_R^-$ ;
28:      $p_z = \text{null}$ ;
29:     for all project  $p_m \in P_k$  such that  $l(p_m) = l_k \wedge f(p_m, l_k) < c_m$  do
30:       if  $\rho(p_m) \succ_L \sigma$  then
31:          $\sigma = \rho(p_m)$ ;
32:          $p_z = p_m$ ;
33:       end if
34:     end for
35:     if  $p_z \neq \text{null}$  then
36:       for all project  $p_j \in P_k$  such that  $l(p_j) = l_k \wedge f(p_j, l_k) > 0 \wedge p_j \neq p_z$  do
37:          $\rho(p_j) = \sigma$ ;
38:          $pred(p_j) = l_k$ ;
39:          $pred(l_k) = p_z$ ;
40:       end for
41:     end if
42:   end for
43: end for
```

---

---

```

44: /* final phase */
45:  $\rho = \max_{\succ_L}(\{B_R^-\} \cup \{\rho(p_j) : p_j \in P \text{ is exposed}\})$ ;
46: if  $\rho \succ_L B_R^-$  then
47:    $p_q = \arg \max_{\succ_L}(\{B_R^-\} \cup \{\rho(p_j) : p_j \in P \text{ is exposed}\})$ ;
48:    $Q = \text{path obtained by following } pred \text{ values and matched edges in } M(f) \text{ from } p_q \text{ to an}$ 
      $\text{exposed student}$ ;
49:    $\text{return } \langle v_s \rangle ++ reverse(Q) ++ \langle l(p_q), v_t \rangle$ ; /*++ denotes concatenation*/
50: else
51:    $\text{return null}$ ;
52: end if

```

---

augmenting path from an exposed student to  $p_m$  of length at least  $(2q+1, 2q-2)$  with a better profile than  $\Phi_{2q-1, 2q-2}(p_m)$ . Such a path must contain a partial augmenting path from an exposed student to some project  $p_{m'}$  such that:

$$\Phi_{2q+1, 2q-2}(p_m) = \Phi_{2q-1, 2q-2}(p_{m'}) + rank(s_i, p_m) - rank(s_i, p_{m'}).$$

where  $s_i \in S'_m$  and  $f(s_i, p_{m'}) = 1$ . Thus we note the following identity involving  $\Phi_{2q+1, 2q-2}(p_m)$ :

$$\begin{aligned} \Phi_{2q+1, 2q-2}(p_m) = \max_{\succ_L} \{ & \Phi_{2q-1, 2q-2}(p_m), \\ & \{\Phi_{2q-1, 2q-2}(p_{m'}) + rank(s_i, p_m) - rank(s_i, p_{m'}) : s_i \in S'_m \wedge f(s_i, p_{m'}) = 1\} \}. \end{aligned} \quad (1)$$

Let  $\rho'_q(p_m)$  be the profile computed at  $p_m$  after the first sub-loop during the  $q^{th}$  iteration of the main loop of the **Get-max-aug** algorithm (i.e. at Line 25 during the  $q^{th}$  iteration). Then

$$\begin{aligned} \rho'_q(p_m) = \max_{\succ_L} \{ & \rho_{q-1}(p_m), \\ & \{\rho_{q-1}(p_{m'}) + rank(s_i, p_m) - rank(s_i, p_{m'}) : s_i \in S'_m \wedge f(s_i, p_{m'}) = 1\} \}. \end{aligned} \quad (2)$$

By the induction hypothesis,  $\rho_{q-1}(p_m) \succeq_L \Phi_{2q-1, 2q-2}(p_m)$ . Thus:

$$\begin{aligned} \rho'_q(p_m) &= \max_{\succ_L} \{ \rho_{q-1}(p_m), \{ \rho_{q-1}(p_{m'}) + rank(s_i, p_m) - rank(s_i, p_{m'}) : \\ & \quad s_i \in S'_m \wedge f(s_i, p_{m'}) = 1 \} \}. \text{ (by equation 2)} \\ &\succeq_L \max_{\succ_L} \{ \Phi_{2q-1, 2q-2}(p_m), \{ \Phi_{2q-1, 2q-2}(p_{m'}) + rank(s_i, p_m) - rank(s_i, p_{m'}) : \\ & \quad s_i \in S'_m \wedge f(s_i, p_{m'}) = 1 \} \}. \text{ (by the inductive hypothesis)} \\ &= \Phi_{2q+1, 2q-2}(p_m). \text{ (by equation 1)} \end{aligned}$$

Therefore:

$$\rho'_q(p_m) \succeq_L \Phi_{2q+1, 2q-2}(p_m). \quad (3)$$

Again, if there does not exist a partial augmenting path from an exposed student to  $p_m$ , of length  $\leq (2q+1, 2q)$  and with a better profile than  $\Phi_{2q+1, 2q-2}(p_m)$ , then  $\Phi_{2q+1, 2q}(p_m) = \Phi_{2q+1, 2q-2}(p_m)$ . Otherwise there exists a partial augmenting path from an exposed student to  $p_m$  of length  $(2q+1, 2q)$  with a better profile than  $\Phi_{2q+1, 2q-2}(p_m)$ . We can therefore note the following identity involving  $\Phi_{2q+1, 2q}(p_m)$ :

$$\begin{aligned} \Phi_{2q+1, 2q}(p_m) &= \max_{\succ_L} \{ \Phi_{2q+1, 2q-2}(p_m), \\ & \{ \Phi_{2q+1, 2q-2}(p_{m'}) : l_k = l(p_m) \wedge p_{m'} \in P'_k \wedge f(p_m, l_k) > 0 \wedge f(l_k, v_t) = d_k^+ \} \}. \end{aligned} \quad (4)$$

After the  $q^{th}$  iteration of the main loop has completed, we have:

$$\begin{aligned} \rho_q(p_m) &= \max_{\succ_L} \{ \rho'_q(p_m), \\ & \{ \rho'_q(p_{m'}) : l_k = l(p_m) \wedge p_{m'} \in P'_k \wedge f(p_m, l_k) > 0 \wedge f(l_k, v_t) = d_k^+ \} \}. \end{aligned} \quad (5)$$

We observe that the extra condition ( $\text{pred}(p_m) \neq l_k$ ) in Line 29 of the second loop, does not affect the correctness of equation 5. Suppose  $\text{pred}(p_m) = l_k$ , then  $\rho(p_m)$  must have been updated during the  $q^{\text{th}}$  iteration of the second loop (or during a previous iteration and has remained unchanged) by some project profile  $\rho(p'_j)$ . Thus setting  $\rho(p_j) = \rho(p_m)$  and  $\text{pred}(l_k) = p_m$  would be incorrect as  $p'_j$  is now the source of  $\rho(p_m)$  and not  $p_m$ . Moreover if indeed  $\rho(p_m) = \rho(p'_j) \succ_L \rho(p_j)$  then  $p'_j$  would be encountered later on during the iteration of the second loop.

$$\begin{aligned} \rho_q(p_m) &= \max_{\succ_L} \{ \rho'_q(p_m), \\ &\quad \{ \rho'_q(p_{m'}) : l_k = l(p_m) \wedge p_{m'} \in P'_k \wedge f(p_m, l_k) > 0 \wedge f(l_k, v_t) = d_k^+ \} \}. \\ &\succeq_L \max_{\succ_L} \{ \Phi_{2q+1, 2q-2}(p_m), \{ \Phi_{2q+1, 2q-2}(p_{m'}) : l_k = l(p_m) \wedge \\ &\quad p_{m'} \in P'_k \wedge f(p_m, l_k) > 0 \wedge f(l_k, v_t) = d_k^+ \} \} \text{ (from equation 3)} \\ &= \Phi_{2q+1, 2q}(p_m) \text{ (by equation 4)}. \end{aligned}$$

Therefore:

$$\rho_q(p_m) \succeq_L \Phi_{2q+1, 2q}(p_m).$$

But any partial augmenting path from an exposed student to  $p_j$  with respect to flow  $f$  can have length at most  $(2k+1, 2k)$ . Thus  $\rho(p_j) = \Phi_{2k+1, 2k}(p_j)$  after  $k$  iterations of the main loop.

Finally we show that a partial augmenting path  $P'$  (and subsequently a full augmenting path) can be constructed by following the  $\text{pred}$  values of projects and lecturers and the matched edges in  $M(f)$  starting from some exposed project  $p_j$  with the maximum  $\rho(p_j)$  profile, and ending at some exposed student (i.e. we show that such a path is continuous and contains no cycle).

Suppose for a contradiction that such a path  $P'$  contained a cycle  $C$ . Then at some step  $X$  during the execution of the algorithm,  $C$  would have been formed when, for some project  $p_j$ , either (i)  $\text{pred}(p_j)$  was set to some student  $s_i$  or (ii)  $\text{pred}(p_j)$  was set to some lecturer  $l_k$ . Let  $P''$  be any path in  $N(I)$ . We may extend our definitions for the profile of a matching and a partial augmenting path to cover the profile of any path in  $N(I)$  as follows:

$$\begin{aligned} \rho(P'') &= O_R + \sum \{ \text{rank}(s_i, p_j) : (s_i, p_j) \in P'' \cap E_2 \wedge f(s_i, p_j) = 0 \} - \\ &\quad \sum \{ \text{rank}(s_i, p_j) : (p_j, s_i) \in P'' \cap E_2 \wedge f(s_i, p_j) = 1 \}. \end{aligned}$$

Considering case (i) let  $p_m = M(s_i)$ . Also let  $\rho'(p_j)$  and  $\rho(p_j)$  be the profiles of partial augmenting paths from some exposed student to  $p_j$  before and after step  $X$  respectively. Then  $\rho(p_j) \succ_L \rho'(p_j)$ . Also  $\rho(p_j) = \rho(p_m) + \text{rank}(s_i, p_j) - \text{rank}(s_i, p_m)$ , i.e.,  $\rho(p_j) = \rho(p_m) + \rho(P'')$  where  $P'' = \{(s_i, p_j), (s_i, p_m)\}$ . Since we can also trace a path through all the other projects in  $C$  (using  $\text{pred}$  values and matched edges) from  $p_m$  to  $p_j$ , it follows that  $\rho(p_m) = \rho'(p_j) + \rho(C \setminus \{(s_i, p_j), (s_i, p_m)\})$ . Thus  $\rho(p_j) = \rho'(p_j) + \rho(C)$ . Note that  $\rho(C) = \rho(C' \setminus M) - \rho(C' \cap M)$  and  $C' = C \cap E_2$  is the set of edges in  $C$  involving only students and projects. As  $\rho(p_j) \succ_L \rho'(p_j)$ , it follows that  $\rho(C' \setminus M) \succ_L \rho(C' \cap M)$ . But since  $|C' \setminus M| = |C' \cap M|$ , and lecturer capacities are clearly not violated by the algorithm, a new matching  $M' = M \oplus C'$  can be generated such that  $\rho(M') \succ_L \rho(M)$  and  $|M'| = |M| = |f|$ , a contradiction to the fact that  $M$  is a greedy  $|f|$ -matching in  $I$ .

Considering case (ii) let  $p_m = \text{pred}(l_k)$ . As before let  $\rho'(p_j)$  and  $\rho(p_j)$  be the profiles of partial augmenting paths from some exposed student to  $p_j$  before and after step  $X$  respectively. Then  $\rho(p_j) \succ_L \rho'(p_j)$ . Also  $\rho(p_j) = \rho(p_m)$ . Since we can also trace a path through all the other projects in  $C$  (using  $\text{pred}$  values and matched edges) from  $p_m$  to  $p_j$ , it follows that  $\rho(p_m) = \rho'(p_j) + \rho(C \setminus \{(p_j, l_k), (p_m, l_k)\}) = \rho'(p_j) + \rho(C)$ . Thus  $\rho(p_j) = \rho'(p_j) + \rho(C)$ . Note that  $\rho(C) = \rho(C' \setminus M) - \rho(C' \cap M)$  and  $C' = C \cap E_2$  is

the set of edges in  $C$  involving only students and projects. As  $\rho(p_j) \succ_L \rho'(p_j)$ , it follows that  $\rho(C' \setminus M) \succ_L \rho(C' \cap M)$ . A similar argument to the one presented above shows a contradiction to the fact that  $M$  is a greedy  $|f|$ -matching in  $I$ .  $\square$

From Lemmas 3.1, 3.2 and 3.3, we can conclude that the algorithm **Greedy-max-spa** finds a greedy maximum matching given a SPA instance. Concerning the complexity of the algorithm, the main loop calls **Get-max-aug**  $\eta$  times where  $\eta$  is the size of a maximum cardinality matching in  $I$ . The first phase of **Get-max-aug** performs  $O(m_2)$  profile comparison operations and  $O(n_3)$  initialisation steps for the lecturer  $pred$  values where  $m_2 = |E_2|$ ,  $n_3 = |\mathcal{L}|$ , and each profile comparison step requires  $O(R)$  time. The loop in the main phase of **Get-max-aug** runs  $k$  times where  $k$  is the value of the flow obtained at that time. The first and second loops perform  $O(m_2)$  and  $O(n_2)$  relaxation steps respectively where  $n_2 = |\mathcal{P}|$  and each relaxation step requires  $O(R)$  time to compare profiles. The final phase of the algorithm performs  $O(n_2)$  profile comparisons, each also taking  $O(R)$  time. Thus the overall time complexity of the **Get-max-aug** method is  $O(m_2 R + n_3 + kR(m_2 + n_2) + n_2 R) = O(kR(m_2))$ . Thus the overall time complexity of the **Greedy-max-spa** algorithm is  $O(n_1^2 R m_2)$ .

When considering the additional factor of  $O(R)$  due to arithmetic on edge weights of  $O(n_1^R)$  size, Orlin's algorithm runs in  $O(R m_2^2 \log(n_1 + n_2) + R m_2 (n_1 + n_2) \log^2(n_1 + n_2))$  time. Suppose  $n_1 \geq n_2$ . Then Orlin's algorithm runs in  $O(R m_2^2 \log n_1 + n_1 R m_2 \log^2 n_1)$  time. If the first term of Orlin's runtime is larger than the second then our algorithm is slower by a factor of  $\frac{n_1^2}{m_2 \log n_1} \leq \frac{n_1}{\log n_1}$  as  $m_2 \geq n_1$ . If the second term of Orlin's runtime is larger than the first then our algorithm is slower by a factor of  $\frac{n_1}{\log^2 n_1} \leq \frac{n_1}{\log n_1}$ .

Now suppose  $n_2 > n_1$ . Then Orlin's algorithm runs in  $O(R m_2^2 \log n_2 + n_2 R m_2 \log^2 n_2)$  time. If the first term of Orlin's runtime is larger than the second then our algorithm is slower by a factor of  $\frac{n_1^2}{m_2 \log n_2} \leq \frac{n_1}{\log n_2} \leq \frac{n_1}{\log n_1}$  as  $m_2 \geq n_1$  and  $n_2 > n_1$ . If the second term of Orlin's runtime is larger than the first then our algorithm is slower by a factor of  $\frac{n_1^2}{n_2 \log^2 n_2} \leq \frac{n_1}{\log^2 n_2} \leq \frac{n_1}{\log n_1}$  as  $n_2 > n_1$ .

So our algorithm is slower than Orlin's by a factor of  $\frac{n_1}{\log n_1}$  in all cases. A straightforward refinement of our algorithm can be made by observing that if no profile is updated during an iteration of the main loop, then no further profile improvements can be made and we can terminate the main loop at this point. We conclude with the following theorem.

**Theorem 3.4.** *Given a SPA instance  $I$ , a greedy maximum matching in  $I$  can be obtained in  $O(n_1^2 R m_2)$  time.*

## 4 Generous maximum matchings in SPA

Analogous to the case for greedy maximum matchings, generous maximum matchings can also be found by modelling SPA as a network flow problem. Given a SPA instance  $I$  we define the following terms relating to partial augmenting paths in  $N(I)$ . For each project  $p_j \in \mathcal{P}$ , we define the *minimum profile of a partial augmenting path* from  $v_s$  through an exposed student to  $p_j$  with respect to  $\prec_R$ , denoted  $\Phi'(p_j)$ , as follows:

$$\Phi'(p_j) = \min_{\prec_R} \{\rho(P') : P' \text{ is a partial augmenting path from } v_s \text{ to } p_j\}.$$

If a partial augmenting path  $P'$  ending at project  $p_j$  can be extended to an augmenting path  $P$  by adding edges  $(p_j, l(p_j))$  and  $(l(p_j), v_t)$  then such an augmenting path is called a *minimum profile augmenting path* if  $\rho(P) = \min_{\prec_R} \{\Phi'(p_j) : p_j \in \mathcal{P}\}$ . A similar approach to that used to find a greedy maximum matching can be adopted in order to find a generous maximum matching. The main **Greedy-max-spa** algorithm will remain unchanged (we



will call it **Generous-max-spa** for convenience) as the intuition remains to successively find larger generous  $k$ -matchings until a generous maximum matching is obtained. We however make slight changes to the **Get-max-aug** algorithm in order to find a minimum profile augmenting path in the network should one exist (the resulting algorithm is then known as **Get-min-aug**). The changes are as follows. (i) We replace all occurrences of left domination  $\succ_L$  with right domination  $\prec_R$ . (ii) We also replace all occurrences of negative infinity profile  $B_R^-$  with a positive infinity profile  $B_R^+$ . (iii) Finally we replace both max functions (in lines 45 and 47) with the min function. Analogous statements and proofs of Lemmas 3.1, 3.2 and 3.3 exist in this context. Thus we may conclude with the following theorem concerning the **Generous-max-spa** algorithm.

**Theorem 4.1.** *Given a SPA instance  $I$ , a generous maximum matching in  $I$  can be obtained in  $O(n_1^2 R m_2)$  time.*

## 5 Lecturer lower quotas

In SPA problems it is often required that the workload of supervising student projects is evenly spread across the lecturing staff (i.e., that project allocations are *load-balanced* with respect to lecturers). This is important because any project allocation should be seen by lecturers to be fair. Moreover a lecturer's workload may have an effect on her performance in other academic and administrative duties. One way of achieving some notion of load-balancing with respect to lecturers is to introduce lower quotas. A lower quota on lecturer  $l_k$  is the minimum number of students that must be assigned to  $l_k$  in any feasible solution. We call this extension the *Student/Project Allocation problem with Lecturer lower quotas* (SPA-L). In an instance  $I$  of SPA-L, each lecturer  $l_k$  has an upper quota  $d_k(I)^+ = d_k^+$  and now additionally has a lower quota  $d_k^-(I)$  (it will be helpful to indicate specific instances to which these lower quotas refer within the notation). We assume that  $d_k^-(I) \geq 0$  and  $d_k^+(I) \geq \max\{d_k^-(I), 1\}$ . In the SPA-L context, our definition of a matching as presented in Section 2 needs to be tightened slightly. A *constrained matching* is a matching  $M$  in the SPA context with the additional property that, for each lecturer  $l_k$ ,  $|M(l_k)| \geq d_k^-(I)$ . A *constrained maximum matching* is a maximum matching taken over the set of constrained matchings in  $I$ . Suppose that  $L$  is the sum of the lecturer lower quotas in  $I$  (i.e.  $L = \sum_{l_k \in \mathcal{L}} d_k^-(I)$ ) and  $\eta$  is the size of a maximum matching in  $I^1$ . For some  $k$  in  $(L \leq k \leq \eta)$ , let  $\mathcal{M}'_k$  denote the set of constrained matchings of size  $k$  in  $I$ . A matching  $M \in \mathcal{M}'_k$  is a *constrained greedy  $k$ -matching* if  $M$  has lexicographically maximum profile, taken over all matchings in  $\mathcal{M}'_k$ . An analogous definition for a *constrained generous  $k$ -matching* can be made.

Due to the introduction of these lecturer lower quotas, instances of SPA-L are not guaranteed to admit a feasible solution. Thus given an instance  $I$  of SPA-L, we seek to find a constrained greedy or a constrained generous maximum matching should one exist. We therefore present results analogous to Lemmas 3.1, 3.2 and 3.3. Firstly however, we make the following observations.

**Proposition 5.1.** *Given an SPA-L instance  $I$ , the size of a constrained maximum matching (should one exist) in  $I$  is equal to the size of a maximum matching in the underlying SPA instance in  $I$ .*

*Proof.* Assume  $I$  admits a constrained matching. Then, by dropping the upper quota of each lecturer  $l_q \in \mathcal{L}$  from  $d_q^+(I)$  to  $d_q^-(I)$ , and finding a saturating flow in the network obtained from the resulting instance, we can obtain a matching  $M_k$  of size  $k$  where

---

<sup>1</sup>We will prove that  $\eta$  is equal to the size of a maximum constrained matching in Proposition 5.1

students' preferences:	lecturers' offerings:
$s_1 : p_1 \quad p_2$	$l_1 : \{p_1\}$
$s_2 : p_3 \quad p_2$	$l_2 : \{p_2\}$
$s_3 : p_3$	$l_3 : \{p_3\}$

$$c_1 = c_3 = 1, d_1^+ = d_3^+ = 1 \text{ and } c_2 = d_2^+ = 2$$

$$d_1^- = d_3^- = 0 \text{ and } d_2^- = 2$$

Figure 3: A SPA-L instance  $I$

$k = \sum_{l_q \in \mathcal{L}} d_q^-$ . By returning the lecturer upper quotas to their original values and then successively finding and satisfying standard augmenting paths (starting from  $f(M_k)$ ) we are bound to obtain a constrained maximum matching as lecturers do not lose any assigned students in the process. The absence of an augmenting path relative to the final flow is proof that the flow (and resulting constrained matching) is maximum.  $\square$

We also observe that a constrained greedy  $k$ -matching  $M_k$  in  $I$  need not be a greedy  $k$ -matching in  $I$ . That is, there may exist a matching  $M'_k$  of size  $k$  in  $I$  such that  $M'_k$  violates some of its lecturer lower quotas (i.e.  $M'_k$  is not a constrained matching) and  $\rho(M'_k) \succ_L \rho(M_k)$ . Figure 3 shows a SPA-L instance whose unique constrained greedy maximum matching is  $M = \{(s_1, p_2), (s_2, p_2), (s_3, p_3)\}$  and a greedy maximum matching  $M' = \{(s_1, p_1), (s_2, p_2), (s_3, p_3)\}$  such that  $\rho(M') \succ_L \rho(M)$ . However it is sufficient to show that, starting from  $M_k$ , we can successively identify and augment (w.r.t. the incumbent flow) maximum profile augmenting paths in  $N(I)$  until a constrained greedy maximum matching is found. Next we show that such augmenting paths exist.

**Lemma 5.2.** *Let  $I$  be an instance of SPA-L and let  $\eta$  denote the size of a constrained maximum matching in  $I$ . Let  $k$  ( $1 \leq k < \eta$ ) be given and suppose that  $M_k$  is a constrained greedy  $k$ -matching in  $I$ . Let  $N = N(I)$  and  $f = f(M_k)$ . Then there exists an augmenting path  $P$  with respect to  $f$  in  $N$  such that if  $f'$  is the result of augmenting  $f$  along  $P$  then  $M_{k+1} = M(f')$  is a constrained greedy  $(k+1)$ -matching in  $I$ .*

*Proof.* The proof is analogous to that presented for Lemma 3.1. We show that considering constrained matchings does not affect most of the arguments presented in the proof of Lemma 3.1. We will deal with the cases where considering constrained matchings may affect the arguments presented in the proof of Lemma 3.1. Firstly we observe that after cloning the projects in  $I$  to form a SPA-L instance  $I' = C(I)$ , the process of converting matchings in  $I$  to  $I'$  and vice versa is unaffected when the matchings considered are constrained. Thus since  $M_k$  is a constrained greedy  $k$ -matching in  $I$ ,  $C(M_k)$  is a constrained greedy  $k$ -matching in  $I'$ .

Let  $M'_{k+1}$  be a constrained greedy  $(k+1)$ -matching in  $I$  (this exists because  $k < \eta$ ). Then  $C(M'_{k+1})$  is a constrained greedy  $(k+1)$ -matching in  $I'$ . Let  $X = C(M_k) \oplus C(M'_{k+1})$ . Then each connected component of  $X$  is either (i) an alternating cycle, (ii)(a) an even-length alternating path whose end vertices are students, (ii)(b) an even-length alternating path whose end vertices are projects, (iii)(a) an odd-length alternating path whose end edges are in  $C(M_k)$  or (iii)(b) an odd-length alternating path whose end edges are in  $C(M'_{k+1})$ . We firstly show that the procedures used to “join” and “eliminate” these connected components in Lemma 3.1 are unaffected when  $C(M_k)$  and  $C(M'_{k+1})$  are constrained matchings. The even-length components that we firstly consider are:

1. type (i) and type (ii)(a) alternating paths.
2. compound type (ii)(a) paths.

When considering the elimination of these even-length components (or compound paths), the requirement that the upper quotas of the lecturers involved must not be violated still holds even if the matchings considered are constrained. Moreover the number of students assigned to each lecturer never drops when considering the elimination of these even-length components (or compound paths).

Let  $C(M''_{k+1})$  be the constrained greedy  $(k+1)$ -matching obtained from augmenting  $C(M'_{k+1})$  along all these even-length paths. Then  $X' = C(M_k) \oplus C(M''_{k+1})$  consists of a set of compound type(ii)(b) paths, compound type (iii)(a) and compound type (iii)(b) paths. These paths, if considered independently, may lead to some lecturer losing an assigned student when they are used to augment  $C(M_k)$  or  $C(M''_{k+1})$ . Thus the elimination argument, as presented in the proof of Lemma 3.1, does not hold. We modify this argument slightly as follows in the case of constrained matchings.

We firstly observe that  $C(M_k)$  and  $C(M''_{k+1})$  are constrained matchings. Thus augmenting  $C(M_k)$  or  $C(M''_{k+1})$  along  $X'$  leads to a constrained matching. When all the elements in  $X'$  are considered together, no lecturer violates her lower quota. If some lecturer loses a student due to some component of  $X'$  and drops below her lower quota, she must gain an extra student due to another component in  $X'$ . But since  $|C(M''_{k+1})| = |C(M_k)| + 1$  there are  $q$  compound type (iii)(a) paths and  $(q+1)$  compound type (iii)(b) paths in  $X'$  for some integer  $q$ . Compound type (ii)(b) components do not affect the size of the matchings.

We claim that there exists some compound type (iii)(b) path  $P'$  in  $X'$  such that when considering all the other components in  $X'$  (i.e.  $X' \setminus P'$ ), lecturer upper and lower quotas are not violated and the size of the matchings are unchanged. Thus the elimination arguments presented in the proof of Lemma 3.1 can be applied to  $X' \setminus P'$ .  $P'$  can be extended to end with edge  $(l_p, v_t)$  such that  $|M''_{k+1}(l_p)| > |M_k(l_p)| \geq d_p^-$ . If  $C(M''_{k+1})$  is the constrained greedy  $(k+1)$ -matching obtained from augmenting  $C(M'_{k+1})$  along  $X' \setminus P'$ , then  $C(M''_{k+1}) \oplus C(M_k) = P'$ . If such a path  $P'$  does not exist then  $|M''_{k+1}(l_p)| \leq |M_k(l_p)|$  for all  $l_p \in \mathcal{L}$ , a contradiction.

The rest of the proof for Lemma 3.1, involving the generation of an augmenting path, follows through.  $\square$

**Lemma 5.3.** *Let  $f$  be a flow in  $N$  and let  $M_k = M(f)$ . Suppose that  $M_k$  is a constrained greedy  $k$ -matching. Let  $P$  be a maximum profile augmenting path with respect to  $f$ . Let  $f'$  be the flow obtained by augmenting  $f$  along  $P$ . Now let  $M_{k+1} = M(f')$ . Then  $M_{k+1}$  is a constrained greedy  $(k+1)$ -matching.*

*Proof.* The proof for Lemma 3.2 holds even if  $M(f)$  and  $M(f')$  are constrained matchings as the number of students assigned to a lecturer never reduces as we augment  $f$  along  $P$ .  $\square$

**Lemma 5.4.** *Given an SPA-L instance  $I$ , let  $f$  be a flow in  $N(I)$  where  $k = |f|$  is not the size of a constrained maximum matching in  $I$  and  $M(f)$  is a constrained greedy  $k$ -matching in  $I$ . Algorithm **Get-max-aug** finds a maximum profile augmenting path in  $N(I)$  with respect to  $f$ .*

*Proof.* We observe that the proof presented for Lemma 3.3 also holds in this case even if  $M(f)$  is a constrained greedy  $k$ -matching.

The first part of the proof shows that after  $q$  iterations of the main loop of **Get-max-aug** where  $0 \leq q \leq k$ ,  $\rho(p_j) \succeq_L \Phi_{2q+1, 2q}(p_j)$  for every project  $p_j \in \mathcal{P}$  where  $\Phi_{2q+1, 2q}(p_j)$  is the maximum profile of any partial augmenting path of length  $\leq (2q+1, 2q)$  from an exposed student to  $p_j$ . By inspection, we observe that this argument remains unchanged even if  $M(f)$  is a constrained matching in  $I$ .

---

**Algorithm 3** Greedy-max-spa-1

---

**Require:** SPA-L instance  $I$ ;  
**Ensure:** return a matching  $M$  if one exists or **null** otherwise;  
1: copy  $I$  to from new instance  $I'$ ;  
2: **for all** lecturer  $l_k \in I'$  **do**  
3:   set  $d_k^+(I') = d_k^-(I)$ ;  
4:   set  $d_k^-(I') = 0$ ;  
5: **end for**  
6:  $\{I'$  becomes a SPA instance $\}$   
7:  $M' = \text{Greedy-max-spa}(I')$ ;  
8: **if**  $|M'| = \sum_{l_k \in \mathcal{L}} d_k^-(I)$  **then**  
9:   copy  $f(M')$  in  $N(I')$  into  $f$  in  $N(I)$ ;  
10:   **loop**  
11:      $P = \text{Get-max-aug}(N(I), f)$ ;  
12:     **if**  $P \neq \text{null}$  **then**  
13:       augment  $f$  along  $P$ ;  
14:     **else**  
15:       return  $M(f)$ ;  
16:     **end if**  
17:   **end loop**  
18: **else**  
19:   return **null**;  
20: **end if**

---

The second part of the proof shows that a partial augmenting path  $P'$  (and subsequently a full augmenting path) can be constructed by following the *pred* values of projects and lecturers and the matched edges in  $M(f)$  starting from some exposed project  $p_j$  with the maximum  $\rho(p_j)$  profile going through some exposed student and ending at the source  $v_s$ . That is, we show that such a path is continuous and contains no cycle. We prove this by demonstrating that, should a cycle  $C$  exist, then augmenting  $f$  along  $C$  would yield a flow of the same size  $f'$  such that  $M(f') \succ_L M(f)$  which is a contradiction to the fact that  $M(f)$  is a greedy  $k$ -matching. This result also holds in the case where  $M(f)$  is a constrained matching as any cycle found will not cause a lecturer to lose any assigned students and so the above arguments can still be made.  $\square$

Given Lemmas 5.2, 5.3 and 5.4, the **Greedy-max-spa** algorithm can be employed as part of an algorithm to find a constrained greedy maximum matching in a SPA-L instance should one exist. This new algorithm (which we call **Greedy-max-spa-1**) is presented in Algorithm 3. The algorithm takes an SPA-L instance  $I$  as input and returns a constrained greedy maximum matching  $M$ , should one exist, or **null** otherwise. A SPA instance  $I'$  is constructed from  $I$  by setting  $d_k^-(I') = 0$  and  $d_k^+(I') = d_k^-(I)$  for each lecturer  $l_k$ . Next we find a greedy maximum matching  $M'$  in  $I'$  using the **Greedy-max-spa** algorithm. If  $f' = f(M')$  is not a saturating flow (i.e., one in which all edges  $(l_k, v_t) \in E_4$  are saturated), then  $I$  admits no constrained matching and we return **null**. Otherwise we augment flow  $f$  in  $N(I)$  by calling the **Get-max-aug** algorithm, where  $f$  is the flow in  $N(I)$  obtained from cloning  $f'$  in  $N(I')$ . We continuously augment the flow until no augmenting path exists. The matching  $M = M(f)$  obtained from the resulting flow  $f$  is a greedy maximum constrained matching in  $I$ . Constrained generous maximum matchings can also be found in a similar way. We conclude with the following theorem.

**Theorem 5.5.** *Given a SPA-L instance  $I$ , a constrained greedy maximum matching and a constrained generous maximum matching in  $I$  can be obtained, should one exist, in  $O(n_1^2 R m_2)$  time.*

*Proof.* Firstly we show that the matching  $M'$  obtained in Line 7 of the **Greedy-max-spa-1** algorithm is a constrained greedy  $|f|$ -matching in  $I$ . Suppose otherwise and some other constrained matching  $M''$  of the same size exists in  $I$  such that  $\rho(M'') \succ_L \rho(M')$ . Then since  $|f| = \sum_{l_k \in \mathcal{L}} d_k^-(I)$ , every lecturer has exactly the same number of assigned students in  $M'$  and  $M''$  so  $M''$  is a valid matching in  $I'$ . This contradicts the fact that  $M'$  is a greedy maximum matching in  $I'$ .

Lemmas 5.2, 5.3 and 5.4 prove that once we obtain a constrained greedy  $|f|$ -matching in  $I$  (should one exist), the rest of the algorithm finds a maximum constrained greedy maximum matching in  $I$ .

For finding a constrained generous maximum matching we simply replace the call to **GreedyMaxSpa** in Line 7 and the call to **Get-max-aug** in Line 11 of the **Greedy-max-spa-1** algorithm with a call to the **Generous-max-spa** and the **Get-min-aug** algorithms respectively as described in Section 4.  $\square$

## 6 Empirical evaluation

### 6.1 Introduction

The **Greedy-Max-Spa** and **Generous-Max-Spa** algorithms were implemented in Java and evaluated empirically. In this section, we present results from empirical evaluations carried out on the algorithm implementations using both real-world and randomly-generated data. Results from the implemented algorithms were compared with those produced by an IP model of SPA in order to improve our confidence in the correctness of both implementations. We also investigate the feasibility issues that will be faced if a Min-Cost-Max-Flow (MCMF) approach (as suggested in [29]) is to be used when solving instances of SPA involving large numbers of students and projects or where students have long preference lists. Other experiments carried out involve varying certain properties of the randomly-generated SPA instances while measuring the runtime of the algorithms and the size, degree and cost of the matchings produced.

An instance generator was used to construct random SPA instances which served as input for the algorithm implementations. This generator can be configured to vary certain properties of the SPA instances produced as follows:

1. The number of students  $n_1$  (with a default value of  $n_1 = 100$ ). The number of projects and lecturers are set to  $n_2 = 0.3n_1$  and  $n_3 = 0.3n_1$  respectively.
2. The minimum  $R_{min}$  and maximum  $R_{max}$  length of any student's preference list (with default values  $R_{min} = R_{max} = 10$ ).
3. The popularity  $\lambda$  of the projects, as measured by the ratio between the number of students applying for one of the most popular projects and the number of students applying for one of the least popular projects (default value of 5).
4. The total capacity of the projects  $C_{\mathcal{P}}$  and lecturers  $C_{\mathcal{L}}$ . These capacities were not divided evenly amongst the projects and lecturers involved (default values are  $C_{\mathcal{P}} = 1.2n_1$  and  $C_{\mathcal{L}} = 1.2n_1$ ).
5. The tie density  $t_d$  ( $0 \leq t_d \leq 1$ ) of the students' preference list. This is the probability that some project is tied with the one preceding it on some student's preference list (default value is  $t_d = 0$ ).

6. The total project and lecturer lower quotas  $L_{\mathcal{P}}$  and  $L_{\mathcal{L}}$  respectively. These lower quotas were divided evenly amongst the projects and lecturers involved (default values are  $L_{\mathcal{P}} = L_{\mathcal{L}} = 0$ ).

We also created SPA instances from anonymised data obtained from previous runs of the student-project allocation scheme at the School of Computing Science, University of Glasgow and solved them using the implemented algorithms. We measured the runtime taken by the algorithms as well as the size, cost and degree of the matchings obtained. Experiments were carried out on a Windows machine with 4 Intel(R) Core(R) i5-2400 CPUs at 3.1GHz and 8GB RAM.

In the following subsections we present results obtained from the empirical evaluations carried out. In Section 6.2 we present the results of correctness tests carried out by comparing results obtained from IP models of SPA and implemented algorithms. In Section 6.3 we demonstrate when the MCMF approach becomes infeasible in practice. In Section 6.4 we present results from running the algorithms against real-world SPA instances. In Section 6.5 we vary certain properties of randomly generated SPA instances while measuring the runtime of the algorithms and the size, degree and cost of the matchings produced. We make some concluding remarks in Section 6.6.

## 6.2 Testing for correctness

Although the **Greedy-Max-Spa** and **Generous-Max-Spa** algorithms have been proven to be correct (See Theorems 3.4 and 4.1), bugs may still exist in the implementations. In order to improve our confidence in any empirical results obtained as part of an experimental evaluation of the algorithms' performance, we compared results from the implemented algorithms with those obtained from IP models of SPA. For each value of  $n_1$  in the range  $n_1 \in \{20, 40, 60, \dots, 200, 300, 400, \dots, 1000\}$ , 10,000 random SPA instances were generated and solved using both methods. For each SPA instance generated,  $R_{min} = R_{max} = 10$  (henceforth we refer to  $R_{min} = R_{max}$  as  $R$ ). The profiles of the resulting matchings were then compared and observed to be identical for all the instances generated. The resulting matchings were also tested to ensure they obeyed all the upper quota constraints for lecturers and projects. These correctness tests show that our implementations are likely to be correct.

## 6.3 Feasibility analysis of the MCMF approach

We implemented an algorithm for finding a minimum cost maximum flow in a given network. As stated in [2, 29], by the appropriate assignment of edge costs/weights in the underlying network  $N(I)$  of a SPA instance  $I$ , a minimum cost maximum flow algorithm can be used to find greedy and generous maximum matchings in  $I$ . We argued that this approach (as described in [2, 29]) would be infeasible due to the floating-point inaccuracies caused by the assignment of exponentially large edge costs/weights in the network. In this section we investigate this claim experimentally and demonstrate the feasibility issues that arise when using various Java data types to represent these edge weights.

Firstly we describe the cost functions required by a minimum cost maximum flow algorithm to find greedy and generous maximum matchings. For finding greedy maximum matchings we set the cost of an edge between a student  $s_i$  and a project  $p_j$  as  $n_1^{R-1} - n_1^{R-k}$  where  $k = \text{rank}(s_i, p_j)$ . For finding generous maximum matchings we set the cost of an edge between a student  $s_i$  and a project  $p_j$  as  $n_1^{k-1}$  where  $k = \text{rank}(s_i, p_j)$ . The cost for all other edges in the network are set to 0.

For the MCMF approach, we define an instance as infeasible if the matching produced is not optimal with respect to the greedy or generous criteria (when compared

with optimal results produced by the **Greedy-Max-Spa** and **Generous-Max-Spa** algorithms and CPLEX). We also consider an instance infeasible if the JVM runs out of memory when using the MCMF algorithm but does not when using the **Greedy-Max-Spa** and **Generous-Max-Spa** algorithms.

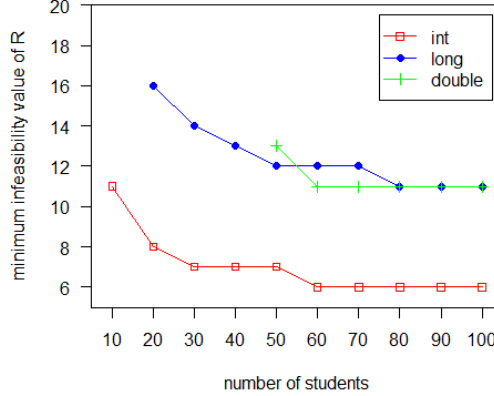


Figure 4: MCMF feasibility results

Figure 4 shows the feasibility results using three Java data types. For each value of  $n_1$  (number of students) in the range  $n_1 \in \{10, 20, 30, \dots, 100\}$  and for each value of  $R$  (length of each student's preference list) in the range  $R \in \{5, 6, \dots, \min\{20, 1.2n_1\}\}$ , we generated 1000 random SPA instances and solved them using the MCMF approach and the **Greedy-Max-Spa** algorithm. The graph shows the value of  $R$  at which infeasible solutions were first encountered. As expected, this number drops as we increase the instance size. Due to their greater precision than the `long` and `double` data types (when compared with `int`), we see that they handle much larger instances before encountering infeasibility issues. All instances tested for  $n_1 = 10$  when using `long` and  $n_1 \in \{10, 20, 30, 40\}$  when using `double` produced optimal matchings. This is probably because we do not yet encounter range errors (in the case of `long`) and precision errors (in the case of `double`) when solving these instances. The relatively low values of  $R$  and  $n_1$  observed where infeasibility prevails (e.g.,  $n_1 = 60, R = 6$  for the `int` type) reinforces our argument that approaches based on MCMF which employ these exponentially large edge weights are not scalable.

## 6.4 Real-world data

SPA instances derived from anonymised data obtained from previous runs of the student-project allocation scheme at the School of Computing Science, University of Glasgow were created and solved using the **Greedy-max-spa** algorithm. This section discusses some of the results obtained. Table 1 shows the properties of the generated SPA instances (with lecturer capacities not being considered in the 07/08 and 08/09 sessions) and Table 2 shows details of various profile-based optimal matchings found.

The results demonstrate a drawback in adopting the greedy optimisation criterion, namely that some students may have projects that are far down their preference lists. In all but the 2011/2012 session, at least one student had her worst-choice project in a greedy maximum matching. In the 2013/2014 session the number of students with their worst-choice project is reasonably high and so the greedy maximum matching would probably not be selected for that year.

The degree of generous maximum matchings are usually less than the others (obviously they are never greater). This is usually an attractive property in such matching schemes. In all the years considered apart from the 2013/2014 session all students got their third

Session	$n_1$	$n_2$	$n_3$	$R$	$C_{\mathcal{P}}$	$C_{\mathcal{L}}$
14/15	51	147	37	6	147	80
13/14	51	155	40	5	155	77
12/13	38	133	34	5	133	63
11/12	31	103	26	5	103	62
10/11	34	63	29	5	63	66
09/10	32	102	28	5	102	72
08/09*	37	56	-	5	56	56
07/08*	35	61	-	5	61	61

Table 1: Real-world SPA instances

Session	$ M $	Greedy		Generous		Min-Cost	
		Profile	Cost	Profile	Cost	Profile	Cost
14/15	51	(30, 7, 1, 5, 5, 3)	110	(16, 16, 9, 6, 4)	119	(28, 11, 3, 5, 2, 2)	101
13/14	51	(26, 7, 4, 6, 8)	116	(15, 18, 9, 6, 3)	117	(23, 12, 5, 6, 5)	111
12/13	38	(26, 6, 3, 2, 1)	60	(21, 13, 4)	59	(23, 11, 3, 1)	58
11/12	31	(22, 6, 2, 1)	44	(20, 9, 2)	44	(20, 9, 2)	44
10/11	34	(25, 4, 3, 1, 1)	51	(21, 9, 4)	51	(24, 5, 4, 1)	50
09/10	32	(23, 4, 2, 2, 1)	50	(19, 10, 3)	48	(20, 9, 2, 1)	48
08/09*	37	(26, 6, 2, 1, 2)	58	(23, 11, 3)	54	(23, 11, 3)	54
07/08*	35	(20, 9, 5, 0, 1)	58	(17, 14, 4)	57	(17, 14, 4)	57

Table 2: Real-world SPA results

choice project or better in the generous maximum matchings produced. However in the 2013/2014 session applying the generous optimality criterion did not improve on the degree of the matchings produced.

One of the major advantages of the minimum cost maximum matching optimality criterion is that in a certain sense it is more “egalitarian”. Minimising the overall cost of the matchings produced is also a very natural objective. It may be considered a disadvantage if matchings obtained by adopting the profile-based optimality criteria have significantly larger costs than the minimum obtainable cost. However, from the results obtained on these real-world datasets, there is very little difference between the costs of the greedy and generous maximum matchings and the minimum obtainable costs (except, once again, for the 2013/2014 session). Thus we can choose one of the profile-based optimal matchings with some confidence that it is “almost” of minimum cost. In Section 6.5 we consider these differences on multiple randomly generated SPA instances.

## 6.5 Randomly-generated instances

### 6.5.1 Introduction

This section discusses some of the results obtained by varying certain properties of the randomly generated SPA instances and measuring the cost, size and degree of the matchings produced. For each instance generated we found a greedy maximum matching, a generous maximum matching and a minimum cost maximum matching.



### 6.5.2 Varying the number of students

Keeping  $R$  constant, we investigated the effects of increasing the number of students  $n_1$  (and by implication  $n_2$ ,  $n_3$ ,  $C_P$  and  $C_L$  using the default dependencies listed in Section 6.1) on the degree, cost and size of the matchings produced as well as the time taken to find these matchings. For each value of  $n_1$  in the range  $n_1 \in \{100, 200, 300, \dots, 700\}$  we generated and solved 100 random SPA instances.

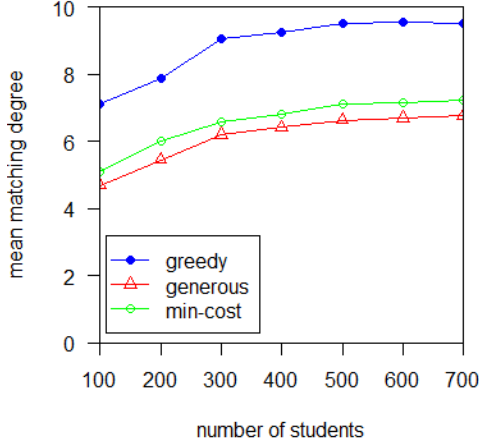


Figure 5: Mean matching degree vs  $n_1$

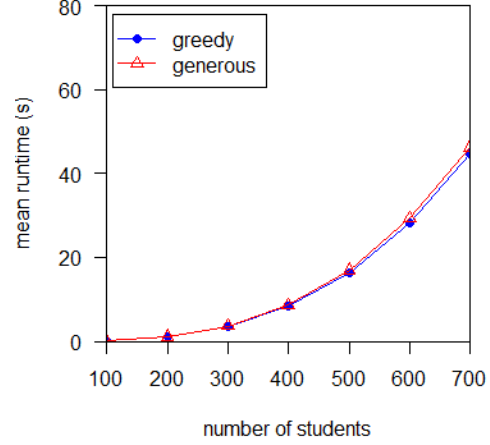


Figure 6: Mean algorithm runtime vs  $n_1$

Figure 5 shows the way the mean degree varies as we increase the number of students. The mean degrees of the greedy maximum matchings are the highest of the three with mean values  $\geq 8$  for  $n_1 > 200$ . As expected generous maximum matchings have the smallest degree, which rises slowly from about 4.8 to 6.5. An interesting observation is that the mean degree does not steeply rise as we increase the number of students. Also the mean degree for the minimum cost maximum matching is closer to the generous maximum matching degree than that of the greedy maximum matching. This is probably due to the fact that the cost function (*rank* in this case) is greater for higher degrees than lower ones, so, in some way, by minimising the cost, we are also seeking matchings with fewer students matched to projects that are father down their preference lists (i.e. have higher ranks).

Figure 6 shows how long it takes to find both profile-based optimal matchings. The main observation is that both **Greedy-max-spa** and **Generous-max-spa** algorithms are scalable and can handle decent-sized instances in reasonable times.

Figure 7 shows how the cost of the matchings generated vary with the number of students. The cost seems to grow proportionally with the number of students. We observe that greedy maximum matchings have larger costs than generous and minimum cost maximum matchings. This corresponds to the mean degree curves shown in Figures 5 where greedy maximum matchings tend to match some students to projects further down their preference list thus adding to the cost of the matching. The average size of the matchings produced was very close to  $n_1$  for all values of  $n_1$  tested.

### 6.5.3 Varying preference list length

The length of students' preference list is one property that can be varied easily in practice (in the SPA context, it is often feasible to ask students to rank more projects if required). So, will increasing the length of the preference lists affect the quality of the matchings produced or the time taken to find them? For each value of  $R$  in the range  $R \in \{1, 2, 3, \dots, 10\}$  we tested this by varying the preference list lengths of 1,000 randomly generated SPA

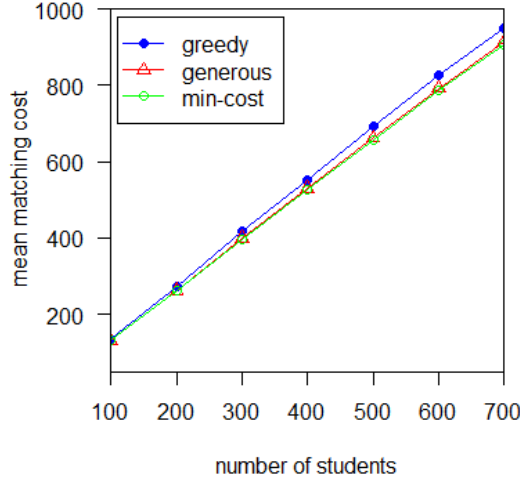


Figure 7: Mean matching cost vs  $n_1$

instances. Each instance had  $n_1 = 100$  students (with  $n_2$ ,  $n_3$ ,  $C_P$  and  $C_L$  all assigned their default values).

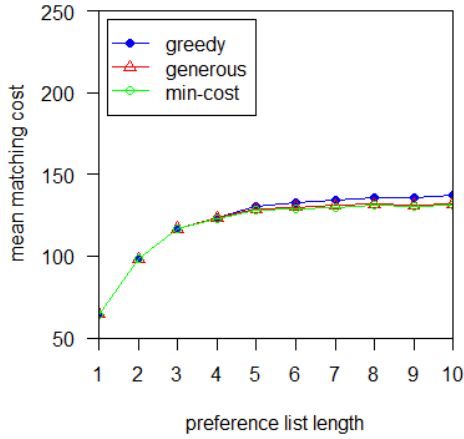


Figure 8: Mean matching cost vs  $R$

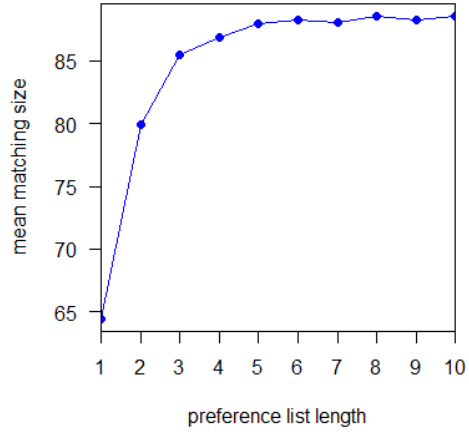


Figure 9: Mean matching size vs  $R$

Figure 8 shows how the mean cost of the matchings obtained varied as we increased the preference list lengths. For the profile-based optimal matchings, the mean cost rises steeply from  $R = 1$  to  $R = 4$  but seems to level off beyond that. We observe that the overall cost of the matchings produced does not significantly change for  $R > 5$ . Thus asking students to submit preference lists greater than 5 will not significantly affect the overall quality of the generous and minimum cost maximum matchings obtained. Once again we observe a difference between the cost of the greedy maximum matchings and the other two.

Figure 9 also shows an important trend as it highlights the value of  $R$  beyond which there is little increase in the mean matching size of profile-based optimal matchings. For the instances generated in this experiment, that value is  $R = 5$ . Thus asking students to submit preference lists of length greater than 5 will not significantly affect the overall size of maximum matchings obtained. Figure 10 shows how the mean degree of the matchings varied as we increased preference list length. For values of  $R \leq 3$  all matchings have the same mean degree as it is likely that some student gets her 3rd choice in each of these matchings. The curve for minimum cost maximum matchings is closer (with respect to degree) to that of generous maximum matchings (obviously generous maximum matchings

have lower degrees in general). They both seem to rise steeply for  $R \leq 5$  and then level off at  $R = 7$  and beyond. Thus asking students to submit preference lists greater than 7 will not significantly affect the overall degree of generous and minimum cost maximum matchings obtained. As expected, greedy maximum matchings had the highest degrees. For  $R > 5$ , the mean degree for greedy maximum matchings does not level off but continues to grow fairly steeply.

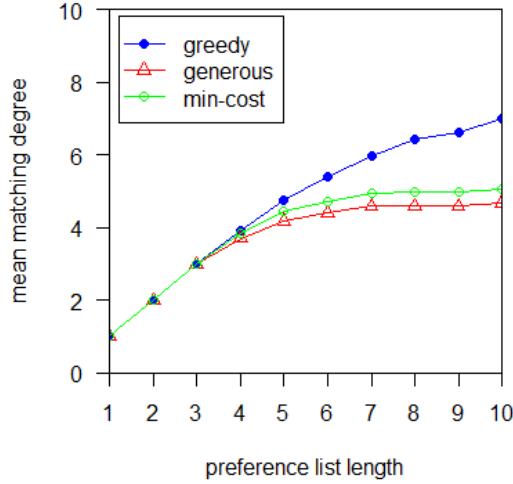


Figure 10: Mean matching degree vs  $R$

Finally we consider how long it takes for the implemented algorithms to find their solutions. In general, the algorithms all seem to handle SPA instances with relatively long preference lists ( $R = 10$ ) in reasonable time ( $< 1.5s$ ).

#### 6.5.4 Varying project popularity

Not all projects will be equally popular and so it is worth investigating the effects the relative popularity  $\lambda$  of the projects may have on the size and quality of the matchings produced. For these experiments, we set  $n_1 = 100$  (with all the other default values) and varied the popularity of the projects involved from 0 to 9 in steps of 1, generating 1,000 random instances for each popularity value. From Figure 11 we see that the cost of the matchings produced gradually increases as we increase the popularity ratio with the cost of the greedy maximum matching being slightly higher than the others (in line with other observations). From Figure 12 we observe no clear trend in the size of the matchings produced as we vary the popularity ratio.

Figure 13 shows the gaps between the mean degree of matchings produced using the various algorithms. Once again we see the mean degrees for the minimum cost and generous maximum matchings being considerably lower than that of the greedy maximum matchings as the popularity ratio increases. Runtimes for the **Greedy-max-spa** and **Generous-max-spa** algorithms were less than 0.25s.

## 6.6 Concluding remarks

Table 3 gives a breakdown of the profiles of 10,000 randomly generated SPA instances of size  $n_1 = 100$  with preference list length  $R = 10$ . It shows the percentage of students with their first choice projects, second choice projects, and so on for greedy, generous and minimum cost maximum matchings (represented by  $M_1$ ,  $M_2$  and  $M_3$  respectively). Although the choice of which profile-based optimal matching is best will, in practice, be problem-specific, the results (as presented in Sections 6.4 and 6.5) give us a general idea

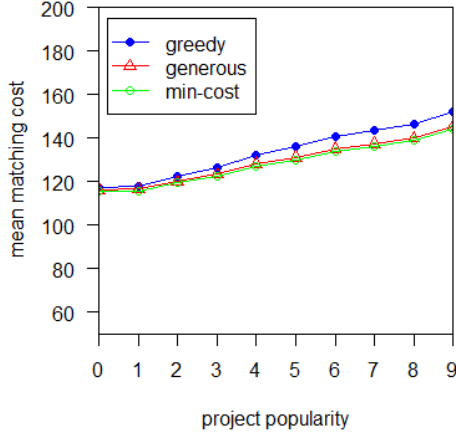


Figure 11: Mean matching cost vs popularity

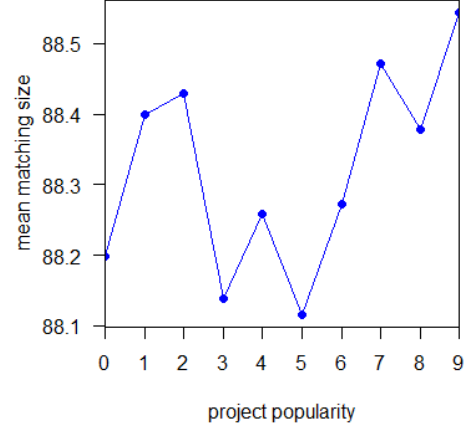


Figure 12: Mean matching size vs popularity

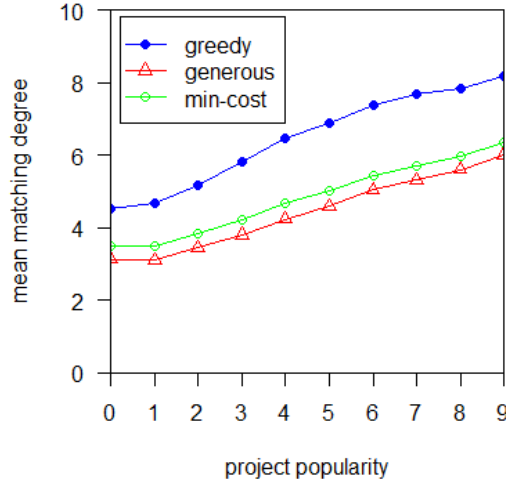


Figure 13: Mean matching degree vs popularity

of the strengths and weaknesses of the various optimality criteria. We summarise these points below.

	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Cost
$M_1$ :	67.94	18.16	7.00	3.14	1.63	0.90	0.55	0.34	0.21	0.14	161.23
$M_2$ :	53.07	38.24	8.68	0.00	0.00	0.00	0.00	0.00	0.00	0.00	155.62
$M_3$ :	61.88	26.84	9.31	1.85	0.13	0.00	0.00	0.00	0.00	0.00	151.51

Table 3: Mean matching profile and cost

With greedy maximum matchings we increase the percentage of students that are *happy* with their assigned projects (i.e., obtain their first choice). A rough estimate of how much better a greedy maximum matching is compared with other profile-based optimal matchings is the difference in the number of first-choice projects. Table 3 shows that the percentage of students with their first-choice project is higher when compared with minimum cost maximum matchings (by 6.06%) and significantly higher when compared with greedy maximum matchings (by 14.87%). However this is achieved at the risk of also increasing the percentage of students who are *disappointed* with their assigned projects

(we say a student  $s_i$  is disappointed with  $p_j = M(s_i)$  if  $\text{rank}(s_i, p_j) > \lceil R/3 \rceil$ ).

With generous maximum matchings we reduce the percentage of disappointed students. A rough estimate of how much better off a student is in a generous maximum matching compared with a greedy maximum matching, is the difference in the *degree* of the matchings. Table 3 shows a significant improvement in the degree as we move from greedy maximum matchings (with some matchings having a degree of 10) and generous maximum matchings (with all matchings having a degree  $\leq 3$ ). Although this is usually a very attractive property, this is achieved without considering the percentage of students who are happy with their assignments. Interestingly the generous criterion will continue to attempt to minimise the number of students matched to their  $n$ th choice project even as  $n$  tends to 1. This motivates a hybrid version of profile-based optimality where we initially adopt the generous criterion and, at some point (say for  $r$ th choice projects where  $r \leq 3$ ), switch to the greedy criterion.

Often the profile of a minimum cost maximum matching lies “in between” the two extremes given by a greedy maximum and generous maximum matching. This can be seen in terms of both the percentage of students with first-choice projects and the degree of the matchings. In terms of the percentage of students with first-choice projects, the results show that minimum cost maximum matchings lie almost halfway between greedy and generous maximum matching percentages. In terms of the degree of the matchings, it seems that minimum cost maximum matchings are a lot closer to generous than greedy maximum matchings. This is usually seen as a desirable property.

## 7 Conclusion

In this paper we investigate the Student / Project Allocation problem in the context of profile-based optimality. We showed how greedy and generous maximum matchings can be found efficiently using network flow techniques. We also presented a range of empirical results obtained from evaluating these efficient algorithms. An obvious question to ask at this stage relates to which other extensions of SPA of practical relevance or theoretical significance can be investigated. These include:

1. Can we improve on the  $O(n_1^2 R m_2)$  algorithm for finding greedy and generous maximum matchings in SPA? One approach would be to determine whether there are faster ways of finding maximum profile augmenting paths in the underlying network than that presented in Algorithm 2. Another approach may be perhaps to abandon the network flow method and consider adopting other techniques used for solving similar problems in the CHAT context [14, 22, 11].
2. The notion of *Pareto optimality* has been well studied in the HA context [1, 3]. It is easy to see that the profile-based optimality criteria defined here imply Pareto optimality. However studying Pareto optimality in its own right is of theoretical interest. Since Pareto optimal matchings in CHAT can be of varying sizes, this extends to SPA. Given a SPA instance we may seek to find a maximum Pareto optimal matching in time faster than  $O(n_1^2 R m_2)$ .

## References

- [1] A. Abdulkadiroğlu and T. Sönmez. Random serial dictatorship and the core from random endowments in house allocation problems. *Econometrica*, 66(3):689–701, 1998.

- [2] D. J. Abraham. Algorithmics of two-sided matching problems. Master’s thesis, University of Glasgow, Department of Computing Science, 2003.
- [3] D. J. Abraham, K. Cechlárová, D. F. Manlove, and K. Mehlhorn. Pareto optimality in house allocation problems. In *Proceedings of ISAAC 2004: the 15th Annual International Symposium on Algorithms and Computation*, volume 3341 of *Lecture Notes in Computer Science*, pages 3–15. Springer, 2004.
- [4] D. J. Abraham, R.W. Irving, and D. F. Manlove. Two algorithms for the Student-Project allocation problem. *Journal of Discrete Algorithms*, 5(1):79–91, 2007.
- [5] A. H. Abu El-Atta and M. I. Moussa. Student project allocation with preference lists over (student,project) pairs. In *Proceedings of ICCEE 09: the Second International Conference on Computer and Electrical Engineering*, pages 375–379. IEEE, 2009.
- [6] A.A. Anwar and A.S. Bahaj. Student project allocation using integer programming. *IEEE Transactions on Education*, 46(3):359–367, 2003.
- [7] J. Dye. A constraint logic programming approach to the stable marriage problem and its application to student-project allocation. BSc Honours project dissertation, University of York, Department of Computer Science, 2001.
- [8] L.R. Ford and D.R. Fulkerson. *Flows in Networks*. Princeton University Press, 1962.
- [9] D. Gusfield and R.W. Irving. *The Stable Marriage Problem: Structure and Algorithms*. MIT Press, 1989.
- [10] P.R. Harper, V. de Senna, I.T. Vieira, and A.K. Shahani. A genetic algorithm for the project assignment problem. *Computers and Operations Research*, 32:1255–1265, 2005.
- [11] C.-C. Huang, T. Kavitha, K. Mehlhorn, and D. Michail. Fair matchings and related problems. In *Proceedings of FSTTCS 2013: the 33rd International Conference on Foundations of Software Technology and Theoretical Computer Science*, volume 24, pages 339–350. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2013.
- [12] A. Hylland and R. Zeckhauser. The efficient allocation of individuals to positions. *Journal of Political Economy*, 87(2):293–314, 1979.
- [13] R.W. Irving. Greedy matchings. Technical Report TR-2003-136, University of Glasgow, Department of Computing Science, 2003.
- [14] R.W. Irving. Greedy and generous matchings via a variant of the Bellman-Ford algorithm. Unpublished manuscript, 2006.
- [15] R.W. Irving, T. Kavitha, K. Mehlhorn, D. Michail, and K. Paluch. Rank-maximal matchings. *ACM Transactions on Algorithms*, 2(4):602–610, 2006.
- [16] K. Iwama, S. Miyazaki, and H. Yanagisawa. Improved approximation bounds for the student-project allocation problem with preferences over projects. *Journal of Discrete Algorithms*, 13:59–66, 2012.
- [17] B. A. Kassa. A linear programming approach for placement of applicants to academic programs. *SpringerPlus*, 2(1):1–7, 2013.
- [18] D. Kazakov. Co-ordination of student-project allocation. Manuscript, University of York, Department of Computer Science, 2002.

- [19] G. Han L. Pan, S. C. Chu and J. Z. Huang. Multi-criteria student project allocation: A case study of goal programming formulation with dss implementation. In *Proceedings of ISORA 2009: The Eighth International Symposium on Operations Research and Its Applications, Zhangjiajie, China*, pages 75–82, 2009.
- [20] D. F. Manlove. *Algorithmics of Matching Under Preferences*. World Scientific, 2013.
- [21] D. F. Manlove and G. O’Malley. Student project allocation with preferences over projects. *Journal of Discrete Algorithms*, 6:553–560, 2008.
- [22] K. Mehlhorn and D. Michail. Network problems with non-polynomial weights and applications. Unpublished manuscript, 2006.
- [23] J.B. Orlin. A faster strongly polynomial minimum cost flow algorithm. *Operations Research*, 41(2):338–350, 1993.
- [24] H. M. Saber and J. B. Ghosh. Assigning students to academic majors. *Omega*, 29(6):513 – 523, 2001.
- [25] C.T.S. Sng. *Efficient Algorithms for Bipartite Matching Problems with Preferences*. PhD thesis, University of Glasgow, Department of Computing Science, 2008.
- [26] C. Y. Teo and D. J. Ho. A systematic approach to the implementation of final year project in an electrical engineering undergraduate course. *IEEE Transactions on Education*, 41(1):25–30, 1998.
- [27] M. Thorn. A constraint programming approach to the student-project allocation problem. BSc Honours project dissertation, University of York, Department of Computer Science, 2003.
- [28] S. Varone and D. Schindl. Course opening, assignment and timetabling with student preferences. In *Proceedings of ICORES: International Conference on Operations Research and Enterprise Systems*, 2013.
- [29] M. Zelvyte. The student-project allocation problem using network flow. BSc Honours project dissertation, University of Glasgow, School of Mathematics and Statistics, 2014.
- [30] L. Zhou. On a conjecture by Gale about one-sided matching problems. *Journal of Economic Theory*, 52(1):123–135, 1990.