Don't choose between workload balance and makespan minimization

Sébastien Deschamps[1], Frédéric Meunier[2]

[1] Saint-Gobain Recherche
sebastien.deschamps2@saint-gobain.com
[2] CERMICS, École des Ponts
frederic.meunier@enpc.fr

Keywords : *bipartite graph, load balancing, fairness.*

1 Balancing workload minimizes the makespan for free

Let $G = (U, W, E)$ be a bipartite graph. Assume given non-negative real numbers d_u for $u \in U$. Let $X \coloneqq \{\boldsymbol{x} \in \mathbb{R}^E_{\geqslant 0} \colon \sum_{e \in \delta(u)} x_e = d_u \ \forall u \in U\}$. Here, $\delta(v)$ denotes the edges incident to a vertex v. For $\boldsymbol{x} \in X$, denote by $\ell^{\max}(\boldsymbol{x})$ (resp. $\ell^{\min}(\boldsymbol{x})$) the quantity $\max_{w \in W} \sum_{e \in \delta(w)} x_e$ (resp. $\min_{w \in W} \sum_{e \in \delta(w)} x_e$). Then we have this fact, which might look surprising at first glance.

Proposition 1 *Over X, every $\boldsymbol{x}$ minimizing $\ell^{\max}(\cdot) - \ell^{\min}(\cdot)$ minimizes $\ell^{\max}(\cdot)$ and maximizes $\ell^{\min}(\cdot)$.*

What might make it counter-intuitive is that a same solution optimizes three criteria at the same time, while these criteria do not share the same monotonicity. An application is the following. Let U be tasks to be performed, W be workers, and d_u be the total time to be spent on performing task u ("demand"). Each worker w can only work on a subset of the tasks (the neighborhood of w in G). Given $\boldsymbol{x} \in X$, the quantity x_e is the time spent by the endpoint in W on performing the endpoint in U. With this interpretation, $\sum_{e \in \delta(w)} x_e$ is the total time worked by $w \in W$ ("load"), and $\ell^{\max}(\boldsymbol{x})$ (resp. $\ell^{\min}(\boldsymbol{x})$) is the maximal (resp. minimal) time worked among the workers. The proposition above translates into: *If you minimize the difference of working times between the most loaded worker and the least loaded one, then you also minimize the load of the most loaded one and maximize the load of the least loaded one.* In the case where the workers work in parallel and they do not stop working until they are done with their load, there is another interpretation: *If you minimize the difference of working times between the most loaded worker and the least loaded, then you also minimize the makespan.* (We remind the reader that the *makespan* is a standard criterion from industrial engineering, defined as the duration of a project, from its beginning to its end.)

Proposition 1 can easily be derived from standard results on lexicographically optimal bases of polymatroids (which go back to the works of Meggido [4] and Fujishighe [2]). For such results, the polymatroid structure looks crucial. Maybe more surprising is the fact that the proposition still holds if d takes integer values, and X is restricted to integer points; this is then a consequence of a recent theorem by Frank and Murota [1]. Special cases were obtained before; see, e.g., the work by Harvey et al. [3], which finds its motivation in load balancing as well.

2 Contributions

In the present work, we show that Proposition 1 above can be kept even without the structure of a polymatroid. Assume given in addition vectors $\boldsymbol{a}^v \in \mathbb{R}^{\delta(v)}_{>0}$ attached to the vertices v of G.

For a point $\boldsymbol{x}$ and a subset A of its indices, we denote by $\boldsymbol{x}_A$ the vector $(x_i)_{i\in A}$. We change now the definition of X to $X^{\boldsymbol{a}} \coloneqq \{\boldsymbol{x}\in\mathbb{R}^E_{\geqslant 0}\colon \boldsymbol{a}^u\cdot\boldsymbol{x}_{\delta(u)} = d_u\ \forall u\in U\}$ and the definition of $\ell^{\max}(\boldsymbol{x})$ and $\ell^{\min}(\boldsymbol{x})$ to $\ell^{\max}(\boldsymbol{x}) \coloneqq \max_{w\in W}\boldsymbol{a}^w\cdot\boldsymbol{x}_{\delta(w)}$ and $\ell^{\min}(\boldsymbol{x}) \coloneqq \min_{w\in W}\boldsymbol{a}^w\cdot\boldsymbol{x}_{\delta(w)}$. The starting setting is the special case when $\boldsymbol{a}^v$ is the all-one vector. Our first result is the following.

Theorem 1 *Suppose $\ell^{\max}(\boldsymbol{x}) > \ell^{\min}(\boldsymbol{x})$ for all $\boldsymbol{x}\in X^{\boldsymbol{a}}$. Then over $X^{\boldsymbol{a}}$, every $\boldsymbol{x}$ minimizing $\ell^{\max}(\cdot)-\ell^{\min}(\cdot)$ minimizes $\ell^{\max}(\cdot)$ and maximizes $\ell^{\min}(\cdot)$.*

By compactness, the existence of $\boldsymbol{x}\in X^{\boldsymbol{a}}$ minimizing $\ell^{\max}(\cdot)-\ell^{\min}(\cdot)$ is ensured as soon as $X^{\boldsymbol{a}}$ is non-empty. Theorem 1 is a generalization of Proposition 1, apart for unfortunate inputs where it is possible to have the quantities $\boldsymbol{a}^w\cdot\boldsymbol{x}_{\delta(w)}$ equal for all $w\in W$ ("it is possible to get all workers equally loaded"), something which is not expected to occur generically. There are actually examples showing that this condition is necessary. (Anyway, when it is possible to have the quantities $\boldsymbol{a}^w\cdot\boldsymbol{x}_{\delta(w)}$ equal for all $w\in W$, then Proposition 1 is immediate.) There are also examples showing that the theorem does not hold when d takes integer values and $X^{\boldsymbol{a}}$ is restricted to integer points.

Our second result shows that when the non-negativity constraint is dropped, we get somehow the reverse situation where minimizing $\ell^{\max}(\cdot)$ or maximizing $\ell^{\min}(\cdot)$ actually minimizes $\ell^{\max}(\cdot)-\ell^{\min}(\cdot)$. In this case, we are able to consider even more general dependencies of the loads to the values x_{uw}. Assume given maps $f_v\colon\mathbb{R}^{\delta(v)}\to\mathbb{R}$ attached to the vertices v of G. We extend the definitions of $\ell^{\max}(\cdot)$ and $\ell^{\min}(\cdot)$ to $\ell^{\max}(\boldsymbol{x}) \coloneqq \max_{w\in W} f_w(\boldsymbol{x}_{\delta(w)})$ and $\ell^{\min}(\boldsymbol{x}) \coloneqq \min_{w\in W} f_w(\boldsymbol{x}_{\delta(w)})$. Let $X^f \coloneqq \{\boldsymbol{x}\in\mathbb{R}^E\colon f_u(\boldsymbol{x}_{\delta(u)}) = d_u\ \forall u\in U\}$. Note that, contrary to what holds for X and $X^{\boldsymbol{a}}$, the $\boldsymbol{x}$ are not constrained to have non-negative components.

Let A be a finite set. To ease the statement of the theorem, we define a certain set $\mathcal{F}_A$ of maps $\mathbb{R}^A\to\mathbb{R}$. A map $f\colon\boldsymbol{x}_A\in\mathbb{R}^A\mapsto f(\boldsymbol{x}_A)\in\mathbb{R}$ belongs to $\mathcal{F}_A$ if it is a continuous increasing self-bijection of $\mathbb{R}$ when restricted to any component x_i (the components of $\boldsymbol{x}_{A\setminus\{i\}}$ being then fixed). Linear maps $\boldsymbol{x}_A\mapsto\boldsymbol{a}\cdot\boldsymbol{x}_A$ with $\boldsymbol{a}\in\mathbb{R}^A_{>0}$, as in Theorem 1, belong to $\mathcal{F}_A$. Moreover, the set $\mathcal{F}_A$ is stable by many binary operations (e.g., addition and maximum), which shows that it actually contains quite complicated maps.

Theorem 2 *Suppose that each f_v belongs to $\mathcal{F}_{\delta(v)}$ and that G is connected. Then, for every $\boldsymbol{x}\in X^f$ such that $\ell^{\max}(\boldsymbol{x}) > \ell^{\min}(\boldsymbol{x})$, there exists $\boldsymbol{x}'\in X^f$ such that $\ell^{\max}(\boldsymbol{x}) > \ell^{\max}(\boldsymbol{x}') = \ell^{\min}(\boldsymbol{x}') > \ell^{\min}(\boldsymbol{x})$.*

When we specialize the maps f_v to $\boldsymbol{x}_{\delta(v)}\mapsto\boldsymbol{a}^v\cdot\boldsymbol{x}_{\delta(v)}$ with $\boldsymbol{a}^v\in\mathbb{R}^{\delta(v)}_{>0}$, we are exactly in the setting of Theorem 1, except for the non-negativity constraint. Theorem 2 implies in particular that if there exists $\boldsymbol{x}\in X^f$ minimizing $\ell^{\max}(\boldsymbol{x})$, then this $\boldsymbol{x}$ is such that $\ell^{\max}(\boldsymbol{x}) = \ell^{\min}(\boldsymbol{x})$. This is thus the opposite phenomenon as that of Theorem 1. Besides, even if X^f is non-empty as soon as no vertex in U is isolated, the infimum of $\ell^{\max}(\boldsymbol{x})$ over X^f is actually not necessarily attained, as some easy examples show.

References

[1] András Frank and Kazuo Murota. Decreasing minimization on M-convex sets: background and structures. *Mathematical Programming*, 195(1-2):977–1025, 2022.

[2] Satoru Fujishige. Lexicographically optimal base of a polymatroid with respect to a weight vector. *Mathematics of Operations Research*, 5(2):186–196, 1980.

[3] Nicholas JA Harvey, Richard E Ladner, László Lovász, and Tami Tamir. Semi-matchings for bipartite graphs and load balancing. *Journal of Algorithms*, 59(1):53–78, 2006.

[4] Nimrod Megiddo. Optimal flows in networks with multiple sources and sinks. *Mathematical Programming*, 7:97–107, 1974.