CASPT 2018

A branch-and-price-and-cut method for train unit scheduling with complex minimum turnround time requirements

Zhiyuan Lin · Pedro J. Copado-Mendez ·
Raymond S. K. Kwan

Abstract We propose a branch-and-price-and-cut method with warm-start for solving the integer fixed-charge multicommodity flow (IFMCF) model for the network flow level of the train unit scheduling problem, in particular with complex minimum turnround time requirements. This problem is regarded to be difficult due to its nature in integer flows and fixed-charge variables. The key components of this method are a warm-start strategy and a dynamic cut generation scheme that will only add certain constraints when needed. We also study the theoretical perspective in finding strong valid inequalities associated with the dynamic cuts for preventing time allowance violations due to coupling/decoupling.

Keywords Train unit scheduling · Minimum turnround time · Fixed-charge network flow · Branch-and-price-and-cut · Valid inequalities

1 Introduction

A *train unit* is a set of train carriages (or *cars* for short) with its own built-in engine(s). Without a locomotive, it is able to move in both directions on its own. A train unit can also be coupled with other units of the same or similar types, which are distinguished by features such as power source, numbers of

Zhiyuan Lin
Alliance Manchester Business School, University of Manchester
Manchester, M13 9SS
United Kingdom
Tel.: +44 (0)161 306 3475
E-mail: zhiyuan.lin@manchester.ac.uk

Pedro J. Copado-Mendez · Raymond S. K. Kwan
School of Computing, University of Leeds
Leeds, LS2 9JT
United Kingdom
Tel.: +44 (0)113 343 5430
E-mail: {p.j.copado-mendez, r.s.kwan}@leeds.ac.uk

cars and speed. Train units are the most commonly used passenger railway rolling stock in the UK and many other countries. Normally a train operating company will possess a fleet of train unit of different types to achieve balanced operations. A timetable gives information on a collection of trips each with its origin, destination, intermediate stations, departure time and arrival time, plus some extra requirements that may not be known by general public, such as passenger capacity demand, compatible types of train unit and maximum formation lengths if units are coupled to serve a trip.

Given a train operator's fixed timetable and a fleet of train units, the Train Unit Scheduling Problem (TUSP) (Lin and Kwan 2014) aims to cover each timetabled trip service by a train unit or a formation of coupled units. From the perspective of a train unit, the problem assigns a sequence of trains to it as its daily workload. As mentioned earlier, a notable feature of the TUSP differentiating it from other kinds of transport vehicle scheduling (e.g. bus and aeroplane) is that more than one train unit can be coupled together to serve the same trip. Coupling/decoupling (sometimes shortened as "c/d" in this paper) could provide more flexibility in unit capacity per trip, especially under the fluctuation of passenger demands. In addition, it can also be used as a way of redistributing train unit resources across the rail network. Despite of its benefits for real-world operations, c/d in TUSP has brought some challenges for automated train unit scheduling based on network flow optimisation methods, such as dynamic minimum turnround time for connecting two trips due to time consumed by c/d operations and formation lengths, redundant c/d operations, unit type compatibility relations, and locations banned for c/d operations.

An integer fixed-charge multicommodity flow model for solving the train unit scheduling problem has been proposed in Lin and Kwan (2013, 2014), where the fixed-charge variables (also called block-arc variables in the train unit scheduling context) are important for calculating coupling/decoupling related values that are used for dealing with the aforementioned dynamic minimum turnround time requirements and redundant c/d operations. However, due to the nature of this formulation, it was difficult to solve medium to large sized instances, making it once impractical for real-world instances. By dropping the block-arc variables, in Lin and Kwan (2016) a branch-and-price approach is proposed where the requirements realised by the block-arc variables in the fixed-charge model are either achieved by customised branching strategies or left to a post-processing phase known as Phase-2 in Lin and Kwan (2014). Although shown to be workable in real-world instances, solving the problem sequentially by two stages will often lose the overall optimality. Moreover, it may significantly increase the difficulty in Phase-2, which is a complex problem itself. Therefore, the focus of our research has turned back to the IFMCF model.

In this paper, we present our recent further exploration on the IFMCF modeland propose a solution approach based on warm-start and dynamically generated cuts that will only be added when minimum turnround time violation is detected. This method is shown to be useful for solving medium sized instances from TransPennine Express (TPE). Some theoretical exploration on valid inequalities associated with the cuts will also be discussed. The following parts are organised as follows. § 2 surveys the previous work relevant to this topic. § 3 describes the complex problem of time allowance in connecting two trips by the same train unit subject to both static and dynamic minimum turnround time. § 5 proposes a branch-and-price-and-cut approach for dealing with the above challenge due to dynamic minimum turnround time. Finally § 6 concludes this paper and envisages future research directions.

2 Literature review

In many countries, train units are the most commonly used rolling stock type for passenger rail networks and we restrict our scope of literature survey to train unit scheduling only.

2.1 Rolling stock circulation problem

The rolling stock circulation problem (RSCP) has been studied extensively by the Dutch group focusing on the real-world tasks from the Dutch railway company Nederlandse Spoorwegen (NS) where train units are used for passenger railway. In the RSCP, the predecessor and successor of each trip are basically given as the input. An early work by Schrijver (1993) first uses an integer multicommodity flow model for this kind of rolling stock scheduling problem for a real-world instance in NS over a single line with two coupling compatible unit types. The goal is to minimize the fleet size. This approach does not include issues such as train composition, unit coupling/decoupling, depot parking nor unit blockage.

Alfieri et al (2006) later consider a similar problem to Schrijver (1993) with two richer models. The first one ignores unit order in a coupled formation and passenger demands and coupling upper bounds are realised directly by constraints. The second model considers unit order by introducing a transition graph technique, in conjunction with new decision variables representing unit compositions for each trip. An extension of the work of Alfieri et al (2006) is proposed by Fioole et al (2006), where considerations in combining and splitting of passenger trains are studied and a new mixed integer programming model is proposed to deal with it. Both a commercial solver and a heuristic based linear programming (LP) relaxation are used with computational results showing that good quality solutions can be found within hours.

Peeters and Kroon (2008) have further extended the problem scenario in NS to multiple lines with one family of compatible unit types. In addition, extra variables and constraints for describing unit inventories are added. By applying Dantzig-Wolfe decomposition, the master problem is decomposed with respect to trains. Branch-and-price is used to get integer solutions which can handle real-world instances of NS in a short time after fine-tuning.

2.2 Train unit assignment problem

The train unit assignment problem (TUAP) has similar definitions and settings with the TUSP, in particular that no predecessors or successors are fixed in advance. Cacchiani et al (2010) present an integer multicommodity flow model for the TUAP based on a directed acyclic graph (DAG). Since the maximum number of coupled units is 2, LP-relaxation can be strengthened in an exact way with respect to the knapsack constraint per trip (Cacchiani et al (2013a)). An LP-based heuristic is used for finding the integer solutions. Real-world instances of a regional train operator in Italy was solved, where the fleets had up to 10 distinct unit types and the timetables contained 528–660 trains. The heuristic is able to find solutions 10–20% better than the manual solutions in practice.

Later a fast and effective heuristic method based on Lagrangian relaxation was presented in Cacchiani et al (2013b) for the same TASP problem. The focus is on finding a good suboptimal solution fast for real-time scenarios. Computational experiments involved some larger instances of up to around 1000 trips.

2.3 Train unit scheduling problem

The train unit scheduling problem (TUSP) (Lin and Kwan, 2014, 2016a; Kwan et al, 2017) is similar to the TUAP (Cacchiani et al, 2010), with additional real-world requirements such as unit type coupling compatibility, locations banned for coupling/decoupling, combination-specific coupling upper bounds, and station-level unit blockage issues, as well as dynamic minimum turnround time to be focused in this paper.

In Lin and Kwan (2014), a two-phase approach is proposed for the TUSP where the first phase assigns and sequences train trips to train units temporarily ignoring some station infrastructure details, and the second phase focuses on satisfying the remaining station detail requirements. The issue of dynamic MTRT has been considered and dealt with by block-arc variables in a way that all of the corresponding constraints are included in the model, making the problem computationally difficult. In Lin and Kwan (2016a), a customised branch-and-price approach for solving the network flow level of the TUSP is presented. Local convex hulls (Lin and Kwan, 2016b) are used to strengthen weak LP-relaxation bounds. TUSP with bi-level capacity requirements is studied in Lin et al (2017). A heuristic branch-and-bound is designed in Lin and Kwan (2018) for removing redundant c/d operations without the need of block-arc variables, as well as a warm-start solution generator for the model with block-arc variables.

For larger and harder TUSP instances, a hybridized algorithm called size limited iterative method (SLIM) is developed in Copado-Mendez et al (2017). It drives the network flow integer multicommodity flow model as a core ILP solver with an iterative heuristic framework. Observing that the final optimal solution only contains a very small subset of the solution space (arcs) in terms of the original data, the proposed hybridized method combines heuristics and exact methods, trying to take advantage of both of an exact and a pure heuristic to achieve high quality near-optimal solutions. The computational experiments have shown promising results.

Train units have to be scheduled at both the network and station level. Usually, the two levels are treated as individual scheduling problems. Lei et al (2017) connects the two levels together and mainly focuses on the train unit shunting at station level to resolve network scheduling solution given by the two-phase approach, where two operational aspects are to be further determined at the station level: unit permutation in a train served by coupled units and the precise conflict-free shunting movements.

2.4 Train unit rescheduling

How to best reschedule a fleet of rolling stock units during a disruption is an optimization problem regularly faced by railway operators. Lusby et al (2017) propose a branch-and-price method based on a path formulation where near-optimal solutions can be found within a few seconds. Furthermore, they show that the proposed methodology can be used, with minor modification, on a tactical planning level, with near-optimal rolling stock schedules. In addition, a method integrating rolling stock scheduling with train unit shunting is proposed by Haahr and Lusby (2017) where high quality solutions for real-life instances shows the benefits from such an integration.

3 Problem description

3.1 Modelling TUSP by directed acyclic graph

The TUSP can be modelled on a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, as used in Cacchiani et al (2010) and Lin and Kwan (2014, 2016a). For a typical DAG, we define the node set $\mathcal{N} = N \cup \{s, t\}$, where N is the set of *trip nodes* representing timetabled trips, and s and t are the *source* and *sink node* conventionally used for a network flow model. The arc set is denoted by $\mathcal{A} = A \cup A_0$. A *connection arc* $a = (i, j) \in A$ links two trip nodes $i, j \in N$ if "the two trips can be consecutively served by the same train unit". Note that in fact this statement is only applicable under certain conditions and the conditions where it is not true are exactly what this paper focusses on. A *sign-on arc* $(s, j) \in A_0, j \in N$ indicates the start of a unit's workload and a *sign-off arc* $(j, t) \in A_0, j \in N$ indicates the end of a unit's daily task. When the arrival location of i is different from the departure location of j, an empty-running movement is needed from i to j and we assume that its empty-running time $\Delta T^E_{ij} > 0$ is known in advance. The time needed for other auxiliary activities such as re-platforming, shunting and depot-return can also be included into

$\varDelta T_{ij}^{E}$ as they can be regarded as a special kind of empty-running. If there is neither empty-running nor other auxiliary activities (meaning the train unit simply stays at the same platform without any shunting during turnround), we let $\varDelta T_{ij}^{E} = 0$. We use $E \subset A$ to denote the set of empty-running arcs.

Generally every train nodes has a sign-on arc and a sign-off arc associated with it. However not every pair of trips i and j will be given a connection arc (i, j). A common criterion for determining whether i can be connected to j is that T_j^{dep}, the departure time of trip j, is sufficiently later than T_i^{arr}, the arrival time of trip i. By "sufficient" it usually means the minimum turnround time (MTRT), i.e. the gap needed for a turnround action between trips i and j, should be respected. Often a *default* value on MTRT is associated with a pair of a location L and time band T, denoted as $\varDelta T_{LT}^{0}$. Note that an arc $(i, j) \in A \setminus E$ can be uniquely mapped to a location/time band pair $(L, T)_{ij}$[1] and thus the MTRT to be followed over (i, j) can be set as $\varDelta T_{ij}^{0} := \varDelta T_{LT_{ij}}^{0}$. As for empty-running arc $(i, j) \in E$ with two associated location/time band pairs $(L, T)_i$ and $(L, T)_j$, we set $\varDelta T_{ij}^{0} := \varDelta T_{LT_i}^{0} + \varDelta T_{LT_j}^{0}$. When creating $\mathcal{G}$, a connection arc (i, j) will be generated between $i, j \in N$ if the following is met:

$$\rho_{ij} = T_j^{\text{dep}} - T_i^{\text{arr}} - \varDelta T_{ij}^{E} - \varDelta T_{ij}^{0} \geq 0, \tag{1}$$

where ρ_{ij} is called the *residual time* between i and j. Each generated arc is given a cost c_{ij} representing costs such as carriage-kilometre, empty-running and other possible preferences.

Let P be the set of s-t paths in $\mathcal{G}$ such that each $p \in P$ represents a sequence of trips as a workload plan for a unit. Moreover, P_j and P_a are used to denote the set of paths passing through node j and arc a respectively. Let K be the set of unit types, corresponding to the commodities in a multicommodity flow model. Type-graphs $\mathcal{G}^k = (\mathcal{N}^k, \mathcal{A}^k)$ as sub-graphs of $\mathcal{G}$ are constructed with respect to each type $k \in K$. The components of $\mathcal{G}^k$ will also be denoted in a similar way, e.g. P^k represents the set of paths in $\mathcal{G}^k$.

3.2 Complex MTRT for TUSP

Condition (1) can only guarantee the default MTRT $\varDelta T_{ij}^{0}$, to be satisfied in common situations. In practice there are more demanding requirements on MTRT and we discuss them in the following parts.

3.2.1 Bi-level MTRT

The default MTRT is a fixed value for given a location at a given period of time. For instance, at London Liverpool Street Station, MTRT of 5 minutes during peak time and 10 minutes during off-peak time are applied. Generally, an operator would like to have a not-too-short turnround time (e.g. $\geq$ 10

[1] When the arrival time of i and departure time of j are from different time bands, the time band yielding a larger MTRT will be used.

minutes) for connecting two trains to let operations be more robust. However, during peak hours, this preference may have to be compromised to be shorter, such as ≥ 5 minutes. Another occasion for using a longer MRTR is that the arrival train is a long journey, giving less punctuality and robustness to the subsequent turnround period. Therefore, the default MTRT could be increased from the standard 5 minutes to 10 minutes.

At TPE, the above rules are made more flexible. There is a desirable (longer) MTRT that is to be met as much as possible, while a mandatory (shorter) MTRT is also imposed to keep the train connections feasible. Often, this requirement has to be balanced with other objectives. For instance, it may not be appropriate to satisfy a desirable MTRT at the price of increasing the number of used train units.

To accommodate the above, two kinds of MTRT can be set for some (L, T) pair: a desirable MTRT $\Delta T^D_{L,T}$ that is to be reached as much as possible and a mandatory MTRT $\Delta T^M_{L,T}$ that must be satisfied, where $\Delta T^M_{LT} \leq \Delta T^D_{LT}$. Similar to the case with a single level of MTRT, two levels of MTRT can be defined over arcs as ΔT^D_{ij} and ΔT^M_{ij}. We define the *slack time* between $i, j \in N$ as $\sigma_{ij} = T^{\text{dep}}_j - T^{\text{arr}}_i - \Delta T^E_{ij}$.

Preferences between ΔT^D_{ij} and ΔT^M_{ij} are realised by giving different weights in their corresponding arcs. An arc with slack time between ΔT^M_{ij} and ΔT^D_{ij} is less preferred than an arc with slack time longer than ΔT^D_{ij}. Let c_{ij} be the standard cost for (i, j) as mentioned earlier. Let $\gamma_{ij} > 1$ be a penalty weight for arc (i, j) that is undesirable. Now we have a modified method in constructing arcs in $\mathcal{G}$ with bi-level MTRT:

For every $(i, j) \in A$:

– If $\sigma_{ij} < \Delta T^M_{ij}$, no arc will be created;
– If $\Delta T^M_{ij} \leq \sigma_{ij} < \Delta T^D_{ij}$, create an arc (i, j) and assign it with a penalised cost $\gamma_{ij} \cdot c_{ij}$;
– If $\sigma_{ij} \geq \Delta T^D_{ij}$, create an arc (i, j) and assign it with a standard cost c_{ij}.

Finally note that the definition of the residual time ρ_{ij} under the bi-level MTRT case should be based on the mandatory MTRT, i.e. $\rho_{ij} := \sigma_{ij} - \Delta T^M_{ij}$.

3.2.2 Dynamic MTRT due to coupling/decoupling (MTRT-CD)

We call the default and bi-level MTRT as *static* as their values do not change once all factors determining them are given. Thus, such MTRT can be included into the problem requirement by solely correctly setting certain components in the underlying DAG $\mathcal{G}$ before carrying out the subsequent solution process.

When coupling/decoupling is involved during a turnround, the MTRT imposed in ensuring time allowance has to be increased accordingly, as c/d operations takes non-negligible time. It is possible that for a certain trip pair i and j with $\rho_{ij} \geq 0$, the default MTRT ΔT^0_{ij} is no longer big enough when some c/d is involved. On the other hand, when no c/d takes place over the same (i, j), ΔT^0_{ij} would remain valid again. In this case, the effective MTRT

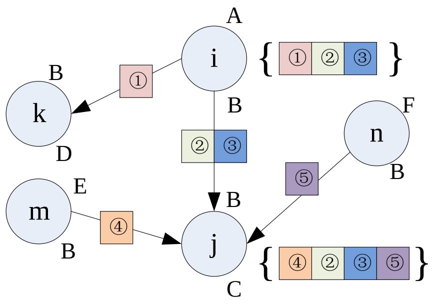

Fig. 1 Coupling/decoupling time calculation at (i, j)

of (i, j) varies depending on the specific situation of train units operated there and thus may not be totally determined beforehand as the static MTRT by setting up the DAG. We call this sort of MTRT as *dynamic*.

Figure 1 shows an example of time allowance calculation at a connection-arc $(i, j) \in A$. There are five train nodes i, j, k, m, n, four *used* connection-arcs $(i, j), (i, k), (m, j), (n, j)$ and five train units ① to ⑤. Train i is from location A to B (and so on for other trains) and the five units (can be of the same or different types) are all compatible. First, units ①,②,③ are coupled together to serve train i. Then at the arrival station B of i, unit ① is detached from the block and continues to serve train k while unit ② and ③ remain together at B. Meanwhile unit ④ arrives at B from E by train m and unit ⑤ arrives at B from F by train n, and both are attached to the unit block ②+③ to form a new block ④+(②+③)+⑤ to serve train j later. Note that, the permutation of coupled units (e.g. ④+(②+③)+⑤ against (②+③)+④+⑤ etc.) should also be carefully considered to avoid potential inoperable plans, which is left to some post-processing stage such as Phase-2 of Lin and Kwan (2014) and Lei et al (2017). We also assume that impractical operations such as units ② and ③ shortly decouple with each other and then couple with ④ and ⑤ to form a permutation "②+⑤+③+④" would be very unlikely to happen in practice. We make an assumption that it takes $n-1$ times of single coupling/decoupling operation when n blocks are coupled/decoupled, which is realistic since coupling/decoupling three or more units simultaneously will generally violate safety regulations. Note that we here only calculate the time taken solely by coupling/decoupling operations, because the time consumed by other auxiliary activities like shunting or empty-running has been already included into the connection information and subtracted from the time gap (as they can be determined a priori). As shown in Figure 1, there are one decoupling (①‖②+③) and two coupling (④ → ②+③, ②+③ ← ⑤) operations occurring at station B. Assuming that the slack time $\sigma_{ij} = 12$ minutes, and the default MTRT $\Delta T^0_{ij} = 5$ minutes, giving the residual time $\rho_{ij} = \sigma_{ij} - \Delta T^0_{ij} = 12 - 5 = 7$. Suppose each single coupling/decoupling operation takes 3 minutes, then the time needed for c/d operations is $3 \times 1 + 3 \times 2 = 9$, longer than the residual time. So there is no sufficient time to perform such a coupling/decoupling operation over connection (i, j).

Note that this conclusion cannot be drawn without a given solution over (i, j). Thus unlike static MTRT, it is difficult to include dynamic MTRT-CD operations beforehand while constructing the DAG $\mathcal{G}$.

3.2.3 Dynamic MTRT due to train formation length (MTRT-FL)

For some train operators such as TPE and Greater Anglia, there is another kind of dynamic MTRT, known as the MTRT due to train formation length (MTRT-FL) at some certain stations and/or time bands. Here, formation lengths are usually measured by the number of cars. The more cars a formation has, the longer the MTRT would be. The main reasons for this requirement are given in the following.

- Short-term preparation work such as seat/table/toilet cleaning and inserting tickets at seat backs are conducted during a turnround. The longer a formation is, the more time is needed to do the preparation.
- When coupling/decoupling is also involved, a drive needs to walk to the coupler to conduct certain operations and walk to the cab in the unit for the next trip.
- At stations with all dead-end platforms, a driver has to walk from one end to the other end of the train unit during a turnround to carry on the next trip unless the train unit will be taken over by another driver.

In the TUSP, the formation (or component) in terms of unit combination for each trip is not fixed but to be determined. Even if the requirements on the passenger capacity demand and the maximum number of coupled cars are satisfied, there may still be several possibilities of feasible formations for a trip, making this unknown in advance similar to the case the c/d operations. Therefore, this kind of MTRT has a dynamic nature as well.

Table 1 Dynamic MTRT due to formation length in Great Anglia network

Num of cars	MTRT (min)	Remarks
4	4	At Cromer MTRT for a 1 or 2 car diesel unit is 3 minutes Unit type 378 requires 5 minutes
8	6	Trains at Liverpool Street require 7 minutes.
12	7	Trains at Liverpool Street require 9 minutes.

For instance, in the Great Anglia rail network, the following MTRT rules in Table 1 are generally applied (without additional c/d operations). From the table it can be seen that MTRT increases as the number of cars in a train formation increases. In addition, while general rules applies, special condition also exists that MTRT-FL can also depend on factors such as location and unit type involved. We call the train unit types that are coupling compatible as being of the same *family*, and let $\mathcal{F}$ be the set of all unit families in fleet K. If special condition exists as in the Remarks area, train units from the same family will normally be grouped into the same MTRT category since they can be coupled with each other.

Formation length mixed with coupling/decoupling If no coupling/decoupling occurs during a turnround, the formation length remains unchanged from the arrival train to the departure train. When coupling/decoupling is involved, there are more than one formations (trains) during a turnround with different lengths. The operational rules in TPE states that in this mixed case, for arc (i, j), the longest formation between the arrival train i and departure train j should be used. Let n_i and n_j be the number of cars for train i and j. The formation length to be used over arc (i, j) should be $\max(n_i, n_j)$. Moreover, the MTRT required is set as the largest between the one due to coupling/decoupling and the one due to formation length, rather than the sum of them. For instance, when an 8 car formation splits into two 4 car formations during a turnround, the formation length of 8 cars should be used. Assuming that 8 cars require 7 minutes MTRT, and the corresponding decoupling requires 6 minutes of MTRT, the overall MTRT needed over (i, j) should therefore be $\max(6, 7) = 7$ minutes.

4 The ILP formulation

The objective of the TUSP is to minimise the total operational costs such as the number of used units, the carriage-kilometre and the number of c/d operations. See Lin and Kwan (2018) for details on the problem of redundant c/d and how to remove it by either heuristic based branch-and-bound or exact method with additional fixed-charge variables. We briefly list the major constraints to be satisfied in the following, and further details can be found in Lin and Kwan (2014, 2016a).

(i) For each unit type $k \in K$, there is a fleet size b_k that cannot be exceeded;
(ii) For each trip $j \in N$, a passenger demand should be satisfied;
(iii) A maximum number of coupled cars should not be exceeded, where its specific value may vary depending on the unit formation achieved at j;
(iv) When coupling two or more units of different types, compatibility relations should be satisfied;
(v) There are locations banned for performing c/d operations;
(vi) Time allowance in connecting two trips following both static and dynamic MTRT;
(vii) Station level refinement: unit blockage, unit order in coupled formations, re-linking

An integer fixed-charge formulation (IFC) developed for the TUSP is given:

$$(IFC) \ \min \ W_x \sum_{k \in K} \sum_{p \in P^k} c_p x_p + W_y \sum_{j \in N} \left(\sum_{a \in \delta_-(j)} y_a + \sum_{a \in \delta_+(j)} y_a \right) \tag{2}$$

subject to

$$\sum_{p \in P^k} x_p \leq b_k, \quad \forall k \in K \tag{3}$$

$$\sum_{k \in K} \sum_{p \in P^k_j} H^j_{f,k} x_p \leq h^j_f, \quad \forall f \in F_j, \forall j \in N \tag{4}$$

$$\sum_{k \in K} \sum_{p \in P^k_a} x_p \leq u_a y_a, \quad \forall a \in \mathcal{A} \tag{5}$$

$$x_p \in \mathbb{Z}_+, \quad \forall p \in P^k, \forall k \in K \tag{6}$$

$$y_a \in \{0, 1\}, \quad \forall a \in \mathcal{A} \tag{7}$$

In the above model, x_p gives the number of used units over path p in the DAG. The fixed-charge variable y_a, aka the block-arc variable in the TUSP context (Lin and Kwan, 2018), indicates whether arc a is used $(= 1)$ or not $(= 0)$. In the objective function, two weights W_x and W_y are assigned to two terms respectively. The first term minimises the operational costs that can be measured per path (unit diagram), including the number of used units, carriage-kilometre and the number of empty-running (dead-heading) trips and so on. Let $\delta_-(j), \delta_+(j) \in \mathcal{A}$ be the sets of arcs coming into and going out of trip j respectively. The second term is used for minimising the total number of coupling/decoupling operations, in order to remove the redundant ones (Lin and Kwan, 2018), where $\sum_{j \in N} \left(\sum_{a \in \delta_-(j)} y_a + \sum_{a \in \delta_+(j)} y_a \right)$ gives the total number of c/d operations over all trips.

As for the constraints, Constraints (3) ensure the number of used units for each type k does not exceed its fleet size b_k. Constraints (4) are the train convex hull constraints (Lin and Kwan, 2016b) representing the valid unit combinations as a result of target passenger capacity requirement, maximum number of coupled units and other implicit constraints, where F_j is the set of nonzero facets describing the convex hull associated with trip j. Constraints (5) are used for calculating the binary block-arc variables where u_a is the possible maximum flow amount over arc a. Finally Constraints (6)–(7) give the variable domains.

Two critical operational requirements are not directly included in (IFC), i.e. the coupling compatibility relations among different unit types and the locations banned for operating coupling/decoupling activities. They are satisfied by customised branching rules in the branch-and-price core-solver (Lin and Kwan, 2016a) without the need of block-arc variables y. The station level refinement process to ensure full operability is left to post-processing stages (Lin and Kwan, 2014; Lei et al, 2017).

As for Point (vi), the static MTRT can be directly satisfied while creating the DAG. The dynamic MTRT is a challenging issue. It is possible to leave them to Phase-2 as in the two-phase framework such that a model with only one kind of flow variable x_p is sufficient for the network flow level (Phase-1). However, solving the entire problem in a sequential manner by two stages will often lose the overall optimality. Moreover, it may significantly increase the difficulty in Phase-2, which is a complex problem per se. In Lin and Kwan (2014), constraints for removing the possibilities that violate MTRT-CD via fixed-charge variable y are used. For a used connection arc $(i, j) \in A$, if a time allowance violation such as the example in § 3.2.2 is found, the following constraint can cut-off such a violation:

$$\tau^D_{\text{arr}(i)} \left(\sum_{a \in \delta_+(i)} y_a - 1 \right) + \tau^C_{\text{dep}(j)} \left(\sum_{a \in \delta_-(j)} y_a - 1 \right) \le \rho_{ij}, \tag{8}$$

where $\tau^D_{\text{arr}(i)}$ and $\tau^C_{\text{dep}(j)}$ are the time consumption for a single decoupling operation at the arrival location of trip i and the a single coupling operation at the departure location of trip j respectively.

The dynamic MTRT-FL can be formally described as the following: For a formation up to n_r cars, the dynamic MTRT should be no more than τ^f_r minutes for unit family f, $r = 1, 2, \ldots, R$ and $f \in \mathcal{F}$, where $n_1 < n_2 < \cdots < n_R$ and $\tau^f_1 < \tau^f_2 < \cdots < \tau^f_R$. Note that when c/d is involved, the formation length should be based on the largest between the arrival and departure trains. Let $w^k_j = \sum_{p \in P^k_j} x_p$ be the number of units of type k used for trip $j \in N$. To satisfy MTRT-FL, the following constraints can be applied:

$$\sum_{k \in f} n_k w^k_i = \sum_{k \in f} \sum_{p \in P^k_i} n_k x_p \le n_r, \quad \forall (i, j) \in A^f_r, \forall f \in \mathcal{F}, r = 1, 2, \ldots, R \tag{9a}$$

$$\sum_{k \in f} n_k w^k_j = \sum_{k \in f} \sum_{p \in P^k_j} n_k x_p \le n_r, \quad \forall (i, j) \in A^f_r, \forall f \in \mathcal{F}, r = 1, 2, \ldots, R \tag{9b}$$

In the above, n_k is the number of cars for unit type k, and A^f_r is defined as

$$A^f_r = \begin{cases} \{a \in A : \tau^f_r \le \sigma_a < \tau^f_{r+1}\}, & \text{if } r \ne R \\ \{a \in A : \sigma_a \ge \tau^f_R\}, & \text{if } r = R \end{cases} \tag{10}$$

Taking the cases in Table 1 as an example. Assume there are two families, where Type 156 (with 3 cars) and 157 (having 5 cars) forms the first family f_1 and Type 378 (with 4 cars) forms the second family f_2. Also assume that the current arcs correspond to neither Liverpool Street nor Cromer. We thus have the following constraints derived from (9) shown in Table 2 (For simplicity, we use the node flow variable w^k_j).

In most cases, the constraints in (9) for $r = R$ are redundant as n_R is already the largest possible number of coupled cars and should be included

Table 2 Constraints for MTRT-FL corresponding to Table 1

Family	r	Constraints
$f_1 = \{156, 157\}$	1	$3w_q^{156} + 5w_q^{157} \leq 4, \quad q = i, j, \forall(i,j) : 4 \leq \sigma_{ij} < 6$
	2	$3w_q^{156} + 5w_q^{157} \leq 8, \quad q = i, j, \forall(i,j) : 6 \leq \sigma_{ij} < 7$
	3	$3w_q^{156} + 5w_q^{157} \leq 12, \quad q = i, j, \forall(i,j) : \sigma_{ij} \geq 7$
$f_2 = \{378\}$	1	$4w_q^{378} \leq 4, \quad q = i, j, \forall(i,j) : 5 \leq \sigma_{ij} < 6$
	2	$4w_q^{378} \leq 8, \quad q = i, j, \forall(i,j) : 6 \leq \sigma_{ij} < 7$
	3	$4w_q^{378} \leq 12, \quad q = i, j, \forall(i,j) : \sigma_{ij} \geq 7$

in Constraints (4). We therefore in principle no longer consider them in this paper. Preprocessing can be applied to strengthen some constraints in (9). For example, $3w_q^{156} + 5w_q^{157} \leq 4 \Leftrightarrow w_q^{156} \leq 1, w_q^{157} = 0$, and $3w_q^{156} + 5w_q^{157} \leq 8 \Leftrightarrow w_q^{156} \leq 1, w_q^{157} \leq 1$, most likely by explicitly computing the convex hulls in terms of variable w_q^k due to the small dimensions. It is also not difficult to include special cases as in Remarks in Table 1 at specific stations into (9), as locations can be implied by arcs.

Finally, when both c/d and formation length have to be taken into account as mentioned in § 3.2.3, the rule of taking the largest MTRT indicates that the two kinds of constraints (8) and (9) can be used simultaneously. In other words, when they are not synchronised with each other in terms of dynamic MTRT, the most constraining one will take effect and thus the largest MTRT value will be automatically applied.

Constraints (8) and (9) in theory can prevent time allowance violation with respect to MTRT-CD and formation length. However, since the number of such constraints can be large, in practice they will significantly slow down the solution process as observed from the experiments. In § 5, we further propose solution approaches in particular for dealing with the large number of Constraints (8) and (9).

5 A branch-and-price-and-cut approach

We propose a branch-and-price-and-cut to solve the IFCMF model (IFC) with block-arc variables based on the branch-and-price solver from Lin and Kwan (2016a). In addition, it is important to apply both a warm-start strategy and a dynamic cut generation method to ensure the model can be solved in a reasonable time in our tested instances.

Warm-start As aforementioned, one important task of the block-arc variables y is to minimise the total number of c/d operations, thus eliminating the redundant ones. We use the results from the heuristic approach given in Lin and Kwan (2018) from solving the model without block-arc variables to warm-start the IFCMF model (IFC). Simply speaking, the column generation process at the root node of the branch-and-bound tree is triggered by a feasible solution from the result of this heuristic, where promising arc flow assignments are branched to reduce redundant c/d within the involved arcs. It can often achieve high quality solutions with respect to minimising the total number of c/d operations, thus significantly speeds up the solution process for the IFCMF model.

We discuss the details for generating dynamic cuts corresponding to (8) and (9) in the next two subsections.

5.1 Dynamic cuts for MTRT due to formation length

We first discuss the generation of dynamic cuts corresponding to Constraints (9), as it does not involve the fixed-charge variable y. As aforementioned, only arcs in $A' = \{(i,j) \in A : \sigma_{ij} < \max_f \tau_R^f\}$ may have a potential in violating MTRT-FL. At the beginning of the branch-and-price, no cuts from (9) is included into Model (IFC). There are three branching rules in the branch-and-price solver in Lin and Kwan (2016a), i.e. train-family branching, banned location branching and integer branching. It is found that only starting the integer branching after both train-family and banned location branching are finished would most likely give the best efficiency on the branch-and-bound (BB) tree. Here we also follow this empirical convention and start generating dynamic cuts for MTRT-FL during the stage of integer branching where both train-family and banned location branching. We summarise the generation of cuts for MTRT-FL in Algorithm 1.

Algorithm 1 Generation of dynamic cuts for MTRT-FL at a BB tree node

```
• Given: A BB tree node $n$ to find its lower bound (e.g. by LP-relaxation)
• Let: $prt(n)$ be the parent node of $n$ and $A'_{prt(n)} \subseteq A'$ be the set of arcs whose cuts $\Pi_{prt(n)}$ in Constraints (9) have been generated in $prt(n)$
• Initialise: $\Pi_n := \Pi_{prt(n)}$, set of cuts used in reduced master problem (RMP) of $n$
• Initialise: $A'_n := A'_{prt(n)}$
• Solve the RMP at $n$ by column generation
repeat
    $ViolationFound = FALSE$
    for $(i,j) \in A' \setminus A'_n$ do
        if violation on MTRT-FL found at $(i,j)$ under family $f$ then
            $\Pi_n += \{\pi_{ij}^f\}$ // $\pi_{ij}^f : \sum_{k \in f} n_k w_q^k \leq n_r, q = i, j; r = 1, 2, \ldots, R-1$
            $A'_n += \{(i,j)\}$
            $ViolationFound = TRUE$
            Resolve RMP at $n$ by column generation, update LP-relaxation
            Break
        end if
    end for
until $ViolationFound = FALSE$
Based on last LP-relaxation solution, either cut-off $n$ or put $n$ into active queue
```

Let u_j be the maximum number of units that can be coupled for trip j. A predetermined order for the arcs in A' can be set in an descending way according to the value of $\frac{u_i+u_j}{\sigma_{ij}}$. Intuitively, the larger this value is, the more likely a violation on MTRT-FL will occur over (i, j). While doing the for loop as in Algorithm 1, following this order may help in speeding up the process.

5.2 Dynamic cuts for MTRT due to coupling/decoupling

Similar to MTRT-FL, not all connection arc will give a MTRT time violation due to c/d, as there is a maximum number of coupled units for a train formation, which may vary among trips and even for the same trip depending on the unit combination assign to it. Define $A^* = \{(i, j) \in A : \rho_{ij} < (u_i - 1)\tau_{\text{arr}(i)} + (u_j - 1)\tau_{\text{dep}(j)}\}$. Only arcs in A^* has a potential to be violated by MTRT-CD (We have assumed that impractical operations will not happen, see § 3.2.2).

5.2.1 Branch over block-arc variables violating MTRT-CD

Observed from numerical experiments, it is found that among all arcs in A^*, only a small proportion may actually violate Constraints (8) either in a final solution or during the branch-and-price process. Therefore, it is not necessary to impose Constraints (8) over all arcs in A^* throughout the branch-and-bound tree but only to add one when needed. This gives the basic idea of generating dynamic cuts to be used in conjunction with a branching scheme over the fixed-charge variables y.

This branching scheme over block-arc (fixed-charge) variable $y_a \in \mathcal{A}$ is based on the following principle:

(i) Only branch on fixed-charge y after all flow variables x are integral and no violation on formation length is found;
(ii) Variable selection when branching on y: First consider if an arc (i, j) violating MTRT-CD can be found, where the value of y_{ij} in the LP-relaxation can be either fractional or binary. If such a y_{ij} is found, branch on it that each of the two children will have a binary value on y_{ij} and the current violation over (i, j) is removed.
(iii) If such an arc cannot be found, find a fractional y_a and branch on it following a standard approach.

We elaborate this branching scheme in Algorithm 2.

A predetermined sorting for the arcs in A^* can also be set in a descending order according to the value of $\frac{|u_i-u_j|}{\sigma_{ij}}$. It is also based on the intuitively observation that the larger this value is, the more likely that $w_i \neq w_j$, thus increasing the possibility of a violation on MTRT-CD over (i, j). While doing the for loop as in Algorithm 2, following this order may help in speeding up the process.

Algorithm 2 Branching on y in conjunction with removing MTRT-CD

```
• Given: A BB tree node $n$ to find its lower bound (e.g. by LP-relaxation)
• Let: $prt(n)$ be the parent node of $n$ and $A^*_{prt(n)} \subseteq A^*$ be the set of arcs whose branching cuts/actions regarding MTRT-CD have been generated in $prt(n)$ and collected into $\Pi_{prt(n)}$
• Initialise: $\Pi_n := \Pi_{prt(n)}$, set of cuts/actions used in RMP of $n$
• Initialise: $A^*_n := A^*_{prt(n)}$
• Solve the RMP at $n$ by column generation
for $(i,j) \in A^* \setminus A^*_n$ do
    if Violation on MTRT-CD is found at $(i,j)$ then
        Form two (latent) branches:
        1. Fix $y_{ij}$ to 0, e.g. by deleting arc $a$. (Action recorded as $\pi^0_{ij}$)
        $\Pi_n + = \{\pi^0_{ij}\}$ and $A^*_n + = \{(i,j)\}$
        2. Fix $y_{ij}$ to 1 and add the cut
            $\pi^1_{ij} : \tau^D_{\mathrm{arr}(i)} \sum_{a \in \delta_+(i) \setminus \{(i,j)\}} y_a + \tau^C_{\mathrm{dep}(j)} \sum_{a \in \delta_-(j) \setminus \{(i,j)\}} y_a \le \rho_{ij}$
        $\Pi_n + = \{\pi^1_{ij}\}$ and $A^*_n + = \{(i,j)\}$
        Break
    end if
end for
if No violation on MTRT-CD found in $A^* \setminus A^*_n$ then
    Branch on fractional $y$ in a standard way
end if
```

5.3 Valid inequalities

It is well-known that fixed-charge network flow problem can suffer from weak LP-relaxation bounds. In this part we discuss several classes of valid inequalities to strengthen the bounds.

To begin with, the following valid inequalities are self-explanatory:

$$\sum_{a \in \delta_-(j)} y_a \ge 1; \sum_{a \in \delta_+(j)} y_a \ge 1, \quad \forall j \in N. \tag{11}$$

Next we focus on deriving certain classes of valid inequalities for the dynamic cuts used alongside with branching y into 1 as in Algorithm 2, i.e.

$$\tau^D_{\mathrm{arr}(i)} \sum_{a \in \delta_+(i) \setminus \{(i,j)\}} y_a + \tau^C_{\mathrm{dep}(j)} \sum_{a \in \delta_-(j) \setminus \{(i,j)\}} y_a \le \rho_{ij}. \tag{12}$$

For simplicity, since the overall maximum number of units that can be coupled at any trip can never exceed 3 in TPE, we take a more conservative step by setting $u_i = u_j = 3$ over arc (i,j). This implies that there can be at most two coupling and two decoupling operations during a turnround, and we thus assume $\rho_{ij} < 2\tau_{\mathrm{arr}(i)} + 2\tau_{\mathrm{dep}(j)}$. Other cases with u_i or $u_j < 3$ would make the problem simpler and can be obtained by analogy. We assume that $\tau_{\mathrm{arr}(i)}, \tau_{\mathrm{dep}(j)}$ and ρ_{ij} are all integers and the difference of $|\tau_{\mathrm{arr}(i)} - \tau_{\mathrm{dep}(j)}|$ is "sufficiently small", e.g. less than 3 minutes, which are common practices in railway companies. We also assume that $|\delta_+(i)| \gg 2$ and $|\delta_-(j)| \gg 2$ unless otherwise stated. Finally the following valid inequalities can be added while generating dynamic cuts (12) while fixing y_{ij} to 1 in Algorithm 2:

$$\sum_{a\in\delta_+(i)\backslash\{(i,j)\}} y_a \leq 2; \quad \sum_{a\in\delta_-(j)\backslash\{(i,j)\}} y_a \leq 2; \quad \sum_{a\in\delta_+(i)\cup\delta_-(j)\backslash\{(i,j)\}} y_a \leq 3. \tag{13}$$

5.3.1 Special case with equal coupling and decoupling time

Using simplified notations as $\tau^D_{\text{arr}(i)} = \tau_i$ and $\tau^C_{\text{dep}(j)} = \tau_j$ We first consider a special case where $\tau_i = \tau_j = \tau_{ij}$, which is commonly seen in TPE and other train operators. Letting $\varDelta_{ij} := \delta_+(i) \cup \delta_-(j) \setminus \{(i,j)\}$, Cut (12) then becomes $\tau_{ij} \sum_{a\in\varDelta_{ij}} y_a \leq \rho_{ij}$, where the range of ρ_{ij} is $0 \leq \rho_{ij} < 4\tau_{ij}$. Define the binary set for Cut (2) where y_{ij} is fixed to 1:

$$Y_{ij} = \left\{ y \in \{0,1\}^{\varDelta_{ij}} \middle| \tau_{ij} \sum_{a\in\varDelta_{ij}} y_a \leq \rho_{ij} \right\} \tag{14}$$

Let $C_h \in \mathscr{C}_h, h = 1,\ldots,4$ be the minimal covers of Y_{ij} when $(h-1)\tau_{ij} \leq \rho_{ij} < h\tau_{ij}$. For a fixed h, all C_h have the same cardinality $|C_h| = h$, giving the cover inequalities $\sum_{a\in C_h} y_a \leq h-1, \forall C_h \in \mathscr{C}_h$. To strengthen them, notice that the extended covers $E(C_h) = \varDelta_{ij}, \forall C_h \in \mathscr{C}_h$ and the fact that all coefficients of y in Y_{ij} are the same. From the classical conclusions (Nemhauser and Wolsey, 1988) , the following

$$\sum_{a\in\varDelta_{ij}} y_a \leq h-1, \quad h = 1,\ldots,4 \tag{15}$$

gives a facet of $\text{conv}(Y_{ij})$ when ρ_{ij} takes different integer values, which happens to be the same result if one simply applies basic rounding technique: $\sum_{a\in\varDelta_{ij}} y_a \leq \lfloor \frac{\rho_{ij}}{\tau} \rfloor = h-1$ for Y_{ij}.

5.3.2 General cases when $\tau_i \neq \tau_j$

Define $\varDelta_i = \delta_+(i)\setminus\{(i,j)\}$ and $\varDelta_j = \delta_-(j)\setminus\{(i,j)\}$. When $\tau_i \neq \tau_j$, the binary set becomes:

$$Y_{ij} = \left\{ y \in \{0,1\}^{\varDelta_{ij}} \middle| \tau_i \sum_{a\in\varDelta_i} y_a + \tau_j \sum_{a\in\varDelta_j} y_a \leq \rho_{ij} \right\}, \tag{16}$$

which is a very much simplified 0-1 knapsack set and classical results (Nemhauser and Wolsey, 1988) can be used to derive (strong) valid inequalities by finding minimal covers and lifting. Nonetheless, due to the simple structure of Y_{ij}, we here discuss some easily obtained strong valid inequalities to avoid the trouble of explicitly performing lifting procedures. Without loss of generality, we assume that $\tau_i < \tau_j$. First notice the obvious outcomes:

– If $\tau_i > \rho_{ij}, \tau_j > \rho_{ij}$, then Y_{ij} can be replaced by $x_a = 0, \forall a \in \varDelta_{ij}$

– If $\tau_i \leq \rho_{ij}, \tau_j > \rho_{ij}$, then $x_a = 0, \forall a \in \varDelta_j$ and the following cut is strong for Y_{ij}: $\sum_{a \in \varDelta_i} y_a \leq h-1$, when $(h-1)\tau_i \leq \rho_{ij} < h\tau_j, h = 2,3,4$. Note that in some cases the above cut can be dominated by (13).

When $\tau_i \leq \rho_{ij}, \tau_j \leq \rho_{ij}$, it implies $\rho_{ij} \geq 2$. A key step in getting strong cuts is to find as many minimal covers for Y_{ij} as possible. Here we show that all such minimal covers can be enumerated and some strong cuts can be derived under certain conditions without applying lifting procedures.

First there exists a family of minimal covers $C_i \in \mathscr{C}_i$ such that $C_i \subset \varDelta_i, \forall C_i \in \mathscr{C}_i$, since $|\delta_+(i)| \gg 2$, $|\delta_-(j)| \gg 2$, $|\tau_{\text{arr}(i)} - \tau_{\text{dep}(j)}|$ is "sufficiently small" and $2 \leq \rho_{ij} < 2\tau_i + 2\tau_j$. Indeed $\mathscr{C}_i$ is made of collections of any $\lfloor \frac{\rho_{ij}}{\tau_i} \rfloor + 1$ arcs in $\varDelta_i$. The extended cover of every C_i is the entire ground set: $E(C_i) = \varDelta_{ij}$. Let $C_i = \{a_1, a_2, \ldots, a_r\}$ be any minimal cover in $\mathscr{C}_i$. Now as long as $C_i \setminus \{a_1, a_2\} \cup \{q\}$, where $q \in \varDelta_j$ is the arc in $\varDelta_{ij}$ giving the largest coefficient, is independent (Nemhauser and Wolsey, 1988), the cut $\sum_{a \in \varDelta_{ij}} y_a \leq |C_i| - 1 = \lfloor \frac{\rho_{ij}}{\tau_i} \rfloor$ will define a facet of Y_{ij}. The set independence implies

$$\tau_i \left(\left\lfloor \frac{\rho_{ij}}{\tau_i} \right\rfloor - 1 \right) + \tau_j \leq \rho_{ij}. \tag{17}$$

Thus as long as (17) is satisfied, the cut based on extended cover $E(C_i) = \varDelta_{ij}$ will be facet-defining wrt Y_{ij}.

Second there exists a class of minimal covers $C_j \in \mathscr{C}_j$ such that $C_j \subset \varDelta_j, \forall C_j \in \mathscr{C}_j$, where each member of $\mathscr{C}_j$ is a collection of any $\lfloor \frac{\rho_{ij}}{\tau_j} \rfloor + 1$ arcs in $\varDelta_j$. Then we have $C_j \subset E(C_j) = \varDelta_j \subset \varDelta_{ij}$. Let an arbitrary $C_j = \{a_1, a_2, \ldots, a_r\}$. $\sum_{a \in \varDelta_j} y_a \leq \lfloor \frac{\rho_{ij}}{\tau_j} \rfloor$, the cut derived from $E(C_j)$ will be strong if the following two conditions are satisfied (Nemhauser and Wolsey, 1988):

(i) $C_j \setminus \{a_1, a_2\} \cup q$ is independent, where $q \in \varDelta_j$ is the arc giving the largest coefficient.

(ii) $C_j \setminus \{a_1\} \cup q$ is independent, where $q = \text{argmax}_a\{\tau_i : a \in \varDelta_i\}$.

(i) is always satisfied, while (ii) is satisfied if

$$\tau_i + \tau_j \left\lfloor \frac{\rho_{ij}}{\tau_j} \right\rfloor \leq \rho_{ij}. \tag{18}$$

There is a third class of minimal covers $C_{ij} \in \mathscr{C}_{ij}$ where $C_{ij} \cap \varDelta_i \neq \emptyset, C_{ij} \cap \varDelta_j \neq \emptyset$. The extended cover $E(C_{ij}) = C_{ij} \cup \varDelta_j$. When $|C_{ij} \cap \varDelta_j| = h, h = 1, 2, \ldots$, which implies $\rho_{ij} \geq h\tau_j$, any $C_{ij} \in \mathscr{C}_{ij}$ is made of h arcs from $\varDelta_j$ and $\lfloor \frac{\rho_{ij} - h\tau_j}{\tau_i} \rfloor + 1$ arcs from $\varDelta_i$. By applying the same criteria in Nemhauser and Wolsey (1988), the extended cover cut $\sum_{a \in C_{ij} \cup \varDelta_j} y_a \leq h + \lfloor \frac{\rho_{ij} - h\tau_j}{\tau_i} \rfloor$ will be strong if

$$\tau_i \left(\left\lfloor \frac{\rho_{ij} - h\tau_j}{\tau_i} \right\rfloor + 2 \right) + \tau_j(h-1) \leq \rho_{ij}. \tag{19}$$

Let $\{\cdot\}$ denote the fractional part of a number. The above is equivalent to

$$\left(2 - \left\{\frac{\rho_{ij} - h\tau_j}{\tau_i}\right\}\right)\tau_i \leq \tau_j \tag{20}$$

Note that since $\tau_i < \tau_j \leq \rho_{ij} < 2\tau_i + 2\tau_j$, actual possible values for h are here only limited to 1, 2 and 3.

Finally we finish this section by remarking that since all possible cases of the classes of minimal covers are found in $\mathscr{C} = \mathscr{C}_i \cup \mathscr{C}_j \cup \mathscr{C}_{ij}$, for each $C \in \mathscr{C}$, it is possible to derive an associated a strong cut for Y_{ij} following the results in Balas (1975) without actually applying explicit lifting, which can be rephrased with respect to (16) as the following:

Theorem 1 *(Balas (1975)) Let $C = \{a_1, a_2, \ldots, a_r\}$ be a minimal cover of Y_{ij} with $a_1 < a_2 < \cdots < a_r$ (if the arcs are indexed by positive integers) and $\tau(a_1) \geq \tau(a_2) \geq \cdots \geq \tau(a_r)$ where $\tau(a_q)$ is the coefficient of arc $a_q \in C$. Let $\mu_h = \sum_{q=1}^{h} \tau(a_q)$ for $h = 1, \ldots, r$; also let $\mu_0 = 0$ and $\lambda = \mu_r - \rho_{ij} \geq 1$. Every valid inequality of the form*

$$\sum_{a \in \Delta_{ij} \setminus C} \beta_a y_a + \sum_{a \in C} y_a \leq |C| - 1 \tag{21}$$

that represents a facet of $\text{conv}(Y_{ij})$ *satisfies the following conditions:*

(a) If $\mu_h \leq \tau(a) \leq \mu_{h-1} - \lambda$, then $\beta_a = h$.
(b) If $\mu_{h+1} - \lambda + 1 \leq \tau(a) \leq \mu_{h+1} - 1$, then $\beta_a = h + 1$.

For instance, for any $C \in \mathscr{C}_i$, for an arc $a \in \Delta_{ij} \setminus C \cap \Delta_i$, since $\tau(a) = \tau_i = \mu_1$ with $h = 1$, we should apply $\beta_a = h = 1$.

6 Conclusions and future research

A branch-and-price-and-cut approach is proposed for efficiently solving the IFCMF model for the network flow level of the TUSP. Two kinds of dynamic MTRT due to formation length and coupling/decoupling activities respectively are included into the model and customised solution approaches such as dynamic cuts are designed for them. Some theoretical analyses are given for deriving relatively simple strong valid inequalities without carrying out lifting.

Further research directions on the branch-and-price-and-cut method include more efficient cut generation strategies, computational experiments on TPE datasets and a possible integration with the hybrid heuristic method SLIM (Copado-Mendez et al, 2017) to solve even larger instances.

References

Alfieri A, Groot R, Kroon LG, Schrijver A (2006) Efficient circulation of railway rolling stock. Transp Sci 40(3):378–391

Balas E (1975) Facets of the knapsack polytope. Math Program 8(1):146–164

Cacchiani V, Caprara A, Toth P (2010) Solving a real-world train-unit assignment problem. Math Program Series B 124(1–2):207–231

Cacchiani V, Caprara A, Maróti G, Toth P (2013a) On integer polytopes with few nonzero vertices. Oper Res Lett 41(1):74–77

Cacchiani V, Caprara A, Toth P (2013b) A lagrangian heuristic for a train-unit assignment problem. Discrete Appl Math 161(12):1707–1718

Copado-Mendez P, Lin Z, Kwan R (2017) Size limited iterative method (SLIM) for train unit scheduling. Proceedings of the 12th Metaheuristics International Conference, Barcelona, Spain

Fioole PJ, Kroon L, Maróti G, Schrijver A (2006) A rolling stock circulation model for combining and splitting of passenger trains. Eur J Oper Res 174(2):1281–1297

Haahr J, Lusby RM (2017) Integrating rolling stock scheduling with train unit shunting. Eur J Oper Res 259(2):452 – 468

Kwan RSK, Lin Z, Copado-Mendez PJ, Lei L (2017) Multi-commodity flow and station logistics resolution for train unit scheduling. Multidisciplinary International Scheduling Conference, Kuala Lumpur, Malaysia

Lei L, Kwan RSK, Lin Z, Copado-Mendez PJ (2017) Station level refinement of train unit network flow schedules. In: 8th International Conference on Computational Logistics, 18-20 Oct 2017, Southampton, UK

Lin Z, Kwan RS (2016a) A branch-and-price approach for solving the train unit scheduling problem. Transp Res Part B 94:97–120

Lin Z, Kwan RS (2018) Redundant coupling/decoupling in train unit scheduling optimization. Electronic Notes in Discrete Mathematics 64:45 – 54

Lin Z, Kwan RSK (2014) A two-phase approach for real-world train unit scheduling. Public Transport 6(1):35–65

Lin Z, Kwan RSK (2016b) Local convex hulls for a special class of integer multicommodity flow problems. Computational Optimization and Applications 64(3):881–919

Lin Z, Barrena E, Kwan RSK (2017) Train unit scheduling guided by historic capacity provisions and passenger count surveys. Public Transport 9(1-2):137–154

Lusby RM, Haahr JT, Larsen J, Pisinger D (2017) A branch-and-price algorithm for railway rolling stock rescheduling. Transp Res Part B 99:228 – 250

Nemhauser G, Wolsey L (1988) Integer and Combinatorial Optimization. Wiley

Peeters M, Kroon LG (2008) Circulation of railway rolling stock: a branch-and-price approach. Computers & OR 35(2):538–556

Schrijver A (1993) Minimum circulation of railway stock. CWI Quarterly 6:205–217