Adv. Appl. Prob. **45,** 351–378 (2013)
Printed in Northern Ireland
© *Applied Probability Trust* 2013

STABILITY OF THE BIPARTITE MATCHING MODEL

ANA BUŠIĆ,[*] *INRIA and École Normale Supérieure*
VARUN GUPTA,[**] *Carnegie Mellon University*
JEAN MAIRESSE,[***] *CNRS and University Paris Diderot*

Abstract

We consider the bipartite matching model of customers and servers introduced by Caldentey, Kaplan and Weiss (2009). Customers and servers play symmetrical roles. There are finite sets C and S of customer and server classes, respectively. Time is discrete and at each time step one customer and one server arrive in the system according to a joint probability measure μ on $C \times S$, independently of the past. Also, at each time step, pairs of *matched* customers and servers, if they exist, depart from the system. Authorized *matchings* are given by a fixed bipartite graph $(C, S, E \subset C \times S)$. A *matching policy* is chosen, which decides how to match when there are several possibilities. Customers/servers that cannot be matched are stored in a buffer. The evolution of the model can be described by a discrete-time Markov chain. We study its stability under various admissible matching policies, including ML (match the longest), MS (match the shortest), FIFO (match the oldest), Random (match uniformly), and Priority. There exist natural necessary conditions for stability (independent of the matching policy) defining the maximal possible stability region. For some bipartite graphs, we prove that the stability region is indeed maximal for any admissible matching policy. For the ML policy, we prove that the stability region is maximal for any bipartite graph. For the MS and Priority policies, we exhibit a bipartite graph with a nonmaximal stability region.

Keywords: Markovian queueing theory; stability; bipartite matching

2010 Mathematics Subject Classification: Primary 60J10
Secondary 60K25; 68M20; 05C21

1. Introduction

In queueing theory, customers and servers play different roles. Customers arrive in the system, accumulate in a buffer, get served by a server, and eventually depart. Servers on the other hand alternate between idle and busy periods but remain forever in the system.

Within this framework, many variations and refinements are possible. For instance, we may consider a model with multiclass customers and flexible servers. A customer of a given class c must choose its server from a specified subset $S(c)$ of the servers. And, of course, the subsets $S(c)$ may intersect; see Figure 1.

Received 4 May 2010; revision received 28 August 2012.

[*] Postal address: INRIA, 23 Avenue d'Italie, CS 81321, 75214 Paris Cedex 13, France.
Email address: ana.busic@inria.fr
[**] Current address: Booth School of Business, University of Chicago, Chicago IL 60637, USA.
Email address: varun.gupta@chicagobooth.edu
[***] Postal address: University Paris Diderot, Sorbonne Paris Cité, LIAFA, UMR 7089 CNRS, Paris, France.
Email address: jean.mairesse@liafa.univ-paris-diderot.fr

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

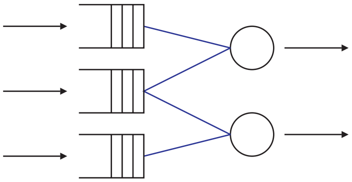

FIGURE 1: Queueing model of a call center.

In this paper, we consider a model with the same multiclass flavor, but in which, by contrast, customers and servers play completely identical roles. We now argue that this simple symmetry requirement leads in a natural and ineluctable way to the *bipartite matching model.*

By symmetry, both customers and servers should arrive into the system and depart from it. More specifically, upon completion of a service, both the customer and the server should depart simultaneously. To model arrivals, we have a priori more flexibility, but there is basically one nontrivial choice which is to assume that time is discrete and that customers and servers arrive in pairs.

Consider the simplest possible model with continuous-time arrivals: (i) there is only one class of customers and one class of servers; (ii) customers and servers arrive according to a Poisson process with respective rates λ and μ; (iii) services have duration 0. Let us describe the state by $Z = X - Y$, where X is the number of unmatched customers and Y is the number of unmatched servers. The process Z is a birth-and-death continuous-time Markov process on $\mathbb{Z}$ with drift $\lambda - \mu$. It is either transient (if $\lambda \neq \mu$) or null recurrent (if $\lambda = \mu$), but it is never positive recurrent.

Let us switch to discrete-time independent and identically distributed (i.i.d.) arrivals. At each time step, a batch of customers and a batch of servers arrive into the system. If the size of the batches are allowed to be different for customers and servers, then we are back to the continuous-time situation, and even the simplest model is never positive recurrent. Therefore, to get a nontrivial model, the natural assumption is that exactly one customer and one server arrive into the system at each time step. The resulting model is symmetric in another respect: both arrivals and departures occur in pairs.

For simplicity, we always assume that the service durations are null. So the model is specified by: (i) the finite set C of customer classes and the finite set S of server classes; (ii) the probability law μ on $C \times S$ for the arrivals in pairs; (iii) the bipartite graph $(C, S, E \subset C \times S)$ giving the possible matchings between customers and servers (hence, the possible departures in pairs); (iv) the matching policy to decide how to match when several choices are possible. We consider so-called *admissible* policies which depend only on the current state of the system. Under these assumptions, the buffer content evolves as a discrete-time Markov chain. We call this model the *bipartite matching* model.

In the bipartite matching model, there is an equal number of customers and servers at any time, but the matching constraints may result in instability with unmatched customers and servers accumulating. It turns out that proving stability, i.e. positive recurrence of the Markov chain, is nontrivial. Given a bipartite graph (C, S, E), there exist natural necessary conditions on μ for stability to hold. When these conditions are also sufficient, we say that the stability region is *maximal.*

Until now, the study of the bipartite matching model has focused on the FIFO (first-in–first-out) policy. The model was introduced by Caldentey *et al.* [3], under an additional assumption of independence between arriving customers and servers (for all c and s, $\mu(c, s) = \mu(c, S)\mu(C, s)$). In the paper, the authors conjectured that any bipartite graph

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

has a maximal stability region for the FIFO policy [3, Conjecture 4.2], and they explicitly treated some small models. In [1], Adan and Weiss proved the conjecture in a fascinating way. They showed that, for some intricate representation of the state space, the model exhibits an explicit 'product-form' stationary distribution.

In the present paper we consider the stability issue for various admissible matching policies: ML (match the longest), MS (match the shortest), FIFO (match the oldest), Random (match uniformly), and Priority. (The FIFO case was not completely solved in [1] since only the case of independent customers and servers was treated.) Even the irreducibility of the Markov chain describing the model is not obvious, and we first study this problem in detail (Section 4). Then we obtain the following results:

- sufficient conditions under which any admissible policy is stable (Section 5);
- the MS policy and some Priority policies do not have a maximal stability region for the simple NN model of Figure 2 (Section 6);
- for any bipartite graph, the ML policy has a maximal stability region (Section 7).

We conjecture that the stability region is always maximal for the FIFO and Random policies.

1.1. Related prior work

In the original paper [3], the authors mentioned several possible domains of applications for the bipartite matching (BM) model, ranging from call centers to public housing administration, and they quoted related references. The product-form result of [1] relied on product-form results for other multiclass queueing models; see [11] and the references therein. The BM model is also related to the constrained queueing network (CQN) model of Tassiulas and Ephremides [10]. However, there is a crucial difference: in the BM model, the edges that may be activated depend on the current state of the system (there should be nonempty buffers at both ends of the edges), which is not the case in [10]. On the other hand, the CQN model is multi-hop, while the BM model is single-hop. In [10], the max-weight policy is shown to have a maximal stability region, and our Theorem 7.1 is very reminiscent of this. Also, the proofs have the same flavor using a quadratic Lyapunov function. Among single-hop CQN models, input-queued crossbar switches have received a lot of attention [8]. The max-weight policy is known to have a maximal stability region under mild conditions on the arrival process [4]. Crossbar switches have a topology close to that of the BM model: a bipartite graph with disjoint arrival and departure nodes. Models for call centers with 'skills-based routeing', see, e.g. [7, Section 5], can be viewed as continuous-time versions of single-hop CQN models in which each arrival class may be served by a subset of the servers.

1.2. Notation

Denote by $\mathbb{N} = \{0, 1, 2, \dots\}$ the set of nonnegative integers. Let A^* be the free monoid generated by A. For any word $w \in A^*$ and any $B \subset A$, set $|w|_B = \#\{i \mid w_i \in B\}$, the number of occurrences in w of letters from B. For $B = \{b\}$, we shorten the notation to $|w|_b$. Furthermore, for any $w \in A^*$, set $[w] := (|w|_a)_{a \in A}$ (the commutative image of w).

2. The bipartite matching model

We now proceed to a more formal definition of the model.

Definition 2.1. A *bipartite matching structure* is a quadruple (C, S, E, F), where

- C is the nonempty and finite set of customer classes;

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

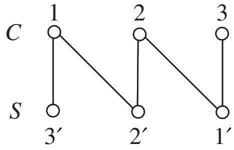

FIGURE 2: NN graph.

- S is the nonempty and finite set of server classes;
- $E \subset C \times S$ is the set of possible matchings;
- $F \subset C \times S$ is the set of possible arrivals.

The bipartite graph (C, S, E) is called the *matching graph*. It is assumed to be connected. The bipartite graph (C, S, F) is called the *arrival graph*. It is assumed to have no isolated vertices.

The two assumptions in Definition 2.1 are made without loss of generality; see Remark 2.1 and Remark 3.1 below.

In Figure 2 we give an example of a matching graph with three customer and three server classes, called the 'NN graph' in the following.

Customers and servers play symmetrical roles in the model. Also, E and F play dual roles. The graph (C, S, E) defines the pairs that may depart from the system, while the graph (C, S, F) defines the pairs that may arrive into the system.

Definition 2.2. A *bipartite matching model* is a triple $[(C, S, E, F), \mu, \textsc{Pol}]$, where

- (C, S, E, F) is a bipartite matching structure;
- μ is a probability measure on $C \times S$ satisfying

$$\operatorname{supp}(\mu) = F, \qquad \operatorname{supp}(\mu_C) = C, \qquad \operatorname{supp}(\mu_S) = S, \tag{2.1}$$

  where μ_C and μ_S are the C and S marginals of μ;
- $\textsc{Pol}$ is an admissible matching policy (to be defined in Section 2.2).

The notation will be simplified to $[(C, S, E), \mu, \textsc{Pol}]$, since this does not result in ambiguity. We say that the model $[(C, S, E), \mu, \textsc{Pol}]$ is *associated* with the structure (C, S, E, F).

Remark 2.1. For (2.1) to have solutions, (C, S, F) must be without isolated vertices, the assumption made in Definition 2.1. This is not a real restriction: if it is not satisfied, we can consider a new model without such customer or server classes.

A realization of the model is as follows. Consider an i.i.d. sequence of random variables of law μ, representing the arrival stream of customer/server pairs. A state of the buffer consists of an equal number of customers and servers with no possible matchings between the nonzero classes. Upon arrival of a new ordered pair (c, s), two situations may occur: if neither c nor s match with the servers/customers already present in the buffer, then c and s are simply added to the buffer; if c or s can be matched then it departs the buffer with its match. If several matchings are possible for c and s then it is the role of the matching policy to select one. An admissible policy selects according to the current state of the buffer (and not according to the whole history of the buffer contents for instance). The resulting evolution of the buffer is described by a discrete-time Markov chain.

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

2.1. State space description

Depending on the matching policy, we consider either a commutative (e.g. for Random) or a noncommutative (e.g. for FIFO) state space description. The different policies considered in the paper will be formally defined in Section 2.2.

Let us choose a matching graph (C, S, E). We introduce the following convenient notation: $C(s)$ is the set of customer classes that can be matched with an s-server; $S(c)$ is the set of server classes that can be matched with a c-customer:

$$S(c) = \{s \in S \colon (c, s) \in E\}, \qquad C(s) = \{c \in C \colon (c, s) \in E\}.$$

For any subsets $A \subset C$ and $B \subset S$, we define

$$S(A) = \bigcup_{c \in A} S(c), \qquad C(B) = \bigcup_{s \in B} C(s).$$

Commutative state space. A state of the system is given by (x, y), $x = (x_c)_{c \in C}$ and $y = (y_s)_{s \in S}$, where x_c denotes the number of customers of class c and y_s denotes the number of servers of class s. The *commutative state space* is

$$\mathcal{E} = \left\{(x, y) \in \mathbb{N}^C \times \mathbb{N}^S \colon \sum_{c \in C} x_c = \sum_{s \in S} y_s \text{ for all } (c, s) \in E,\ x_c y_s = 0\right\}. \tag{2.2}$$

Noncommutative state space. A state of the system is given by two finite words of the same size, $k \geq 0$, on the alphabets C and S, describing unmatched customers and servers. The *noncommutative state space* is

$$\mathcal{E} = \left\{(u, v) \in \bigcup_{k \geq 0} (C^k \times S^k) \colon ([u], [v]) \text{ belongs to (2.2)}\right\}.$$

Facet. Both the commutative and noncommutative state spaces can be decomposed into facets, defined only by the nonzero classes.

Definition 2.3. A *facet* is an ordered pair (U, V) such that $U \subset C$, $V \subset S$, and $U \times V \subset (C \times S - E)$. The *zero-facet*, $(\varnothing, \varnothing)$, is denoted by $\varnothing$.

For a facet $\mathcal{F} = (U, V)$, define

$$\begin{aligned}
C_\bullet(\mathcal{F}) &= U, & S_\bullet(\mathcal{F}) &= V,\\
C_\circledcirc(\mathcal{F}) &= C(V), & S_\circledcirc(\mathcal{F}) &= S(U),\\
C_\circ(\mathcal{F}) &= C - (C_\bullet(\mathcal{F}) \cup C_\circledcirc(\mathcal{F})), & S_\circ(\mathcal{F}) &= S - (S_\bullet(\mathcal{F}) \cup S_\circledcirc(\mathcal{F})).
\end{aligned}$$

The symbol '$\bullet$' stands for the nonzero classes, the symbol '$\circledcirc$' stands for the classes that are forced to be at zero (since they are matched with nonzero classes), and the symbol '$\circ$' stands for the classes that happen to be at zero.

The following notion will play an important role later on.

Definition 2.4. A facet $\mathcal{F}$ is called saturated if $C_\circ(\mathcal{F}) = \varnothing$ or $S_\circ(\mathcal{F}) = \varnothing$.

In Figure 3, the facet on the left is nonsaturated, while the facet on the right is saturated.

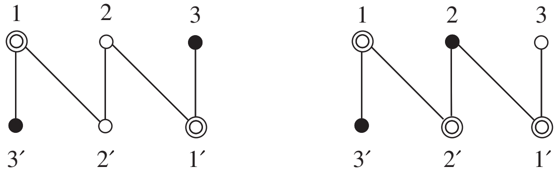

FIGURE 3: NN graph: facets ({3}, {3′}) and ({2}, {3′}).

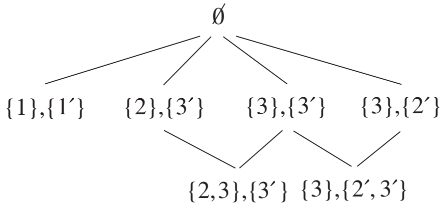

FIGURE 4: Facets for the NN graph.

Graphical convention. A facet $\mathcal{F}$ can be represented graphically by coloring the nodes of the bipartite graph according to the above convention (see Figure 3 for an illustration):

- nodes in $C_\bullet(\mathcal{F})$ and $S_\bullet(\mathcal{F})$ are represented as filled circles;
- nodes in $C_\circledcirc(\mathcal{F})$ and $S_\circledcirc(\mathcal{F})$ are represented as double circles;
- nodes in $C_\circ(\mathcal{F})$ and $S_\circ(\mathcal{F})$ are represented as simple circles.

In Figure 4 we have represented the facets of the NN graph. The more complex case of the NNN graph will be given in Section 5, Figure 14.

Algorithm 1 below takes as input a matching graph and returns as output the set of facets. The termination and correctness of the algorithm are easily proved.

Algorithm 1. (*Computation of the facets.*)
Data: a bipartite graph $G = (C, S, E)$. *Result:* FACETS, the set of all facets of G.

```
begin
    FACETS $\leftarrow \varnothing$; New $\leftarrow \varnothing$;
    foreach $(i, j) \in C \times S - E$ do New $\leftarrow$ New $\cup \{(\{i\}, \{j\})\}$;
    while New $\neq \varnothing$ do
        FACETS $\leftarrow$ FACETS $\cup$ New;
        Old $\leftarrow$ New; New $\leftarrow \varnothing$;
        forall $\mathcal{H}, \mathcal{K} \in$ Old such that $\mathcal{H} \neq \mathcal{K}$ do
            if $C_\bullet(\mathcal{H}) = C_\bullet(\mathcal{K})$ or $S_\bullet(\mathcal{H}) = S_\bullet(\mathcal{K})$ then
                $Z \leftarrow (C_\bullet(\mathcal{H}) \cup C_\bullet(\mathcal{K}), S_\bullet(\mathcal{H}) \cup S_\bullet(\mathcal{K}))$;
                New $\leftarrow$ New $\cup \{Z\}$;
    FACETS $\leftarrow$ FACETS $\cup \{\varnothing\}$;
    return FACETS;
end
```

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

2.2. Admissible matching policies

Informally, a matching policy is *admissible* if

- only the current state of the buffer is taken into account;
- priority is given to customers/servers that are already present in the buffer: if the state is (u, v) and the new arrival is $(c, s) \in E$, then c and s are matched if and only if there are no servers from $S(c)$ in v and no customers from $C(s)$ in u.

It follows from the first point that an admissible matching policy can be described as a mapping $\odot\colon \mathcal{E} \times (C \times S) \to \mathcal{E}$ which returns the new state of the system after an arrival. The second point is a restriction that we call the *buffer-first* assumption. Observe however that a matching policy that always gives priority to new arrivals can be seen as a special case of buffer-first with an arrival probability μ such that $\mu(E) = 0$.

We now define admissible policies formally, distinguishing between the noncommutative and commutative state spaces.

For a word $w \in A^k$ and $i \in \{1, \dots, k\}$, we denote by $w_{[i]} := w_1 \dots w_{i-1} w_{i+1} \dots w_k$ the subword of w obtained by deleting w_i.

Definition 2.5. (*Noncommutative case.*) A matching policy is *admissible* if there are functions Φ and Ψ such that

$$(u, v) \odot (c, s) = \begin{cases} (uc, vs) & \text{if } |u|_{C(s)} = 0,\ |v|_{S(c)} = 0,\ (c, s) \notin E, \\ (u, v) & \text{if } |u|_{C(s)} = 0,\ |v|_{S(c)} = 0,\ (c, s) \in E, \\ (u_{[\Phi(u,s)]}, v_{[\Psi(v,c)]}) & \text{if } |u|_{C(s)} \neq 0,\ |v|_{S(c)} \neq 0, \\ (u_{[\Phi(u,s)]}c, v) & \text{if } |u|_{C(s)} \neq 0,\ |v|_{S(c)} = 0, \\ (u, v_{[\Psi(v,c)]}s) & \text{if } |u|_{C(s)} = 0,\ |v|_{S(c)} \neq 0. \end{cases}$$

The FIFO and LIFO (last-in–first-out) policies are admissible matching policies with functions Φ and Ψ as follows.

FIFO: $\Phi(u, s) = \arg\min\{u_k \in C(s)\}$, $\Psi(v, c) = \arg\min\{v_k \in S(c)\}$.

LIFO: $\Phi(u, s) = \arg\max\{u_k \in C(s)\}$, $\Psi(v, c) = \arg\max\{v_k \in S(c)\}$.

For $c \in C$, let $e_c \in \mathbb{N}^C$ be defined by $(e_c)_c = 1$ and $(e_c)_d = 0$, $d \neq c$. For $s \in S$, let e_s be defined accordingly.

Definition 2.6. (*Commutative case.*) A matching policy is *admissible* if there are functions Φ and Ψ such that

$$(x, y) \odot (c, s) = \begin{cases} (x + e_c, y + e_s) & \text{if } x_{C(s)} = 0,\ y_{S(c)} = 0,\ (c, s) \notin E, \\ (x, y) & \text{if } x_{C(s)} = 0,\ y_{S(c)} = 0,\ (c, s) \in E, \\ (x - e_{\Phi(x,s)}, y - e_{\Psi(y,c)}) & \text{if } x_{C(s)} \neq 0,\ y_{S(c)} \neq 0, \\ (x - e_{\Phi(x,s)} + e_c, y) & \text{if } x_{C(s)} \neq 0,\ y_{S(c)} = 0, \\ (x, y - e_{\Psi(y,c)} + e_s) & \text{if } x_{C(s)} = 0,\ y_{S(c)} \neq 0. \end{cases}$$

The following commutative matching policies are admissible (for Random, ML, and MS policies, $\Phi(u, s)$ and $\Psi(v, c)$ are random variables).

Priority. For each customer class $c \in C$, we define a priority function $\alpha_c\colon S(c) \to \{1, \dots, |S(c)|\}$. Similarly, for each server class $s \in S$, we define $\beta_s\colon C(s) \to \{1, \dots, |C(s)|\}$.

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

In the case of several matching options, a customer/server is matched with the server/customer that has the highest priority (greatest value of the priority function). It is convenient to specify the priorities by two $|C| \times |S|$ matrices A and B, defined by

$$A_{cs} = \begin{cases} \alpha_c(s), & (c, s) \in E, \\ 0, & \text{otherwise}, \end{cases} \qquad B_{cs} = \begin{cases} \beta_s(c), & (c, s) \in E, \\ 0, & \text{otherwise}. \end{cases}$$

Then $\Phi(x, s) = \arg\max\{\beta_s(c) \colon c \in C(s),\ x_c > 0\}$ and $\Psi(y, c) = \arg\max\{\alpha_c(s) \colon s \in S(c),\ y_s > 0\}$.

Random. $\Phi(x, s)$ and $\Psi(y, c)$ are random variables respectively valued in $C(s)$ and $S(c)$, and respectively distributed as $(x_i / \sum_{j \in C(s)} x_j)_{i \in C(s)}$ and $(y_i / \sum_{j \in S(c)} y_j)_{i \in S(c)}$. Intuitively, the match is chosen uniformly among all possible ones.

ML. $\Phi(x, s)$ and $\Psi(y, c)$ are random variables uniformly distributed on $\arg\max\{x_i : i \in C(s)\}$ and $\arg\max\{y_i : i \in S(c)\}$, respectively.

MS. $\Phi(x, s)$ and $\Psi(y, c)$ are random variables uniformly distributed on $\arg\min\{x_i > 0 \colon i \in C(s)\}$ and $\arg\min\{y_i > 0 \colon i \in S(c)\}$, respectively.

3. Necessary conditions for stability

To introduce the main ideas, consider first a simpler finite and deterministic problem. Let (C, S, E) be a matching graph. Consider a batch of customers $x \in \mathbb{N}^C$ and a batch of servers $y \in \mathbb{N}^S$ of equal size: $\sum_c x_c = \sum_s y_s$. A *perfect matching* of x and y is a tuple $m \in \mathbb{N}^E$ such that

$$x_c = \sum_{s \in S(c)} m_{cs} \quad \text{for all } c \in C, \qquad y_s = \sum_{c \in C(s)} m_{cs} \quad \text{for all } s \in S.$$

By Hall's theorem (also known as the 'marriage theorem'), there exists a perfect matching if and only if

$$\sum_{c \in U} x_c \leq \sum_{s \in S(U)} y_s \quad \text{for all } U \subset C, \qquad \sum_{s \in V} y_s \leq \sum_{c \in C(V)} x_c \quad \text{for all } V \subset S. \tag{3.1}$$

A perfect matching, if there is one, can be obtained by restating the model as a flow network and by solving the maximum flow problem for which efficient algorithms exist [5], [6].

The bipartite matching model is much more complicated: first it is random, and second the matchings have to be performed on the fly, at each time step. However, the two ingredients of the simpler model will play an instrumental role in the analysis: (i) the conditions NCond, to be defined in (3.2), are related to (3.1); (ii) the restatement as a flow problem is used in most of the proofs.

Consider now a bipartite matching model $[(C, S, E), \mu, \text{Pol}]$. We identify the model with the Markov chain on the state space $\mathcal{E}$ describing the evolution of the buffer content.

Let P be the transition matrix of the Markov chain. A probability measure π on $\mathcal{E}$ is *stationary* if $\pi P = \pi$. It is *attractive* if, for any probability measure ν on $\mathcal{E}$, the sequence of Cesàro averages of νP^n converges weakly to π.

Definition 3.1. The model is said to be *stable* if the Markov chain has a unique and attractive stationary probability measure.

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

This implies in particular that the graph of the Markov chain has a unique terminal strongly connected component with all states leading to it. We return to this point in Section 4.

Let μ_C be a probability measure on C, and let μ_S be a probability measure on S. Define the following conditions on (μ_C, μ_S) (collectively referred to as NCond):

$$\begin{aligned} \mu_C(U) &< \mu_S(S(U)) \quad \text{for all } U \subsetneq C,\ U \neq \varnothing, \\ \mu_S(V) &< \mu_C(C(V)) \quad \text{for all } V \subsetneq S,\ V \neq \varnothing. \end{aligned} \tag{3.2}$$

The above conditions appear in [3]. They have a natural interpretation. Let μ_C and μ_S be the marginals of the arrival probability μ. Customers from U need to be matched with servers from $S(U)$. The first condition of NCond asks for strictly more servers in average from $S(U)$ than customers from U. The second condition has a dual interpretation.

Using the strong law of large numbers, we also see that the arrivals up to time n satisfy (3.1) for all values of n large enough if and only if NCond is satisfied.

Lemma 3.1. *The conditions* NCond *are necessary stability conditions: if the Markov chain is stable then the conditions* NCond *are satisfied by the marginals of* μ.

Proof. We suppose that the conditions NCond are not satisfied.

Assume first that there exists $U \subset C$ such that $\mu_C(U) > \mu_S(S(U))$. Let A_n and B_n be the total number of customers of class U and the total number of servers of class $S(U)$ to arrive in the system up to time n. Let X_n be the number of customers of class U present in the system at time n. By definition, $X_n \geq A_n - B_n$. By the strong law of large numbers, we have, almost surely,

$$\lim_n \frac{A_n}{n} = \mu_C(U), \qquad \lim_n \frac{B_n}{n} = \mu_S(S(U)), \qquad \lim_n \frac{X_n}{n} \geq \mu_C(U) - \mu_S(S(U)) > 0.$$

So the Markov chain is transient. Similarly, if there exists $V \subset S$ such that $\mu_S(V) > \mu_C(C(V))$, the model is unstable. (This part of the argument appears in [3, Proposition 3.4].)

Assume now that there exists $U \subset C,\ U \neq C$, such that

$$\mu_C(U) = \mu_S(S(U)). \tag{3.3}$$

Observe that $S(U) \neq S$; otherwise, we would have $\mu_C(U) = \mu_S(S) = 1$ which would contradict $U \neq C$. Set $V = S - S(U)$. Equation (3.3) is equivalent to $\mu_S(V) = \mu_C(C - U)$. The bipartite matching graph (C, S, E) is represented in Figure 5. By assumption we have $(U \times V) \cap E = \varnothing$.

We have

$$\begin{aligned} \mu(U \times V) &= \mu_C(U) - \mu(U \times S(U)), \\ \mu((C - U) \times S(U)) &= \mu_S(S(U)) - \mu(U \times S(U)). \end{aligned}$$

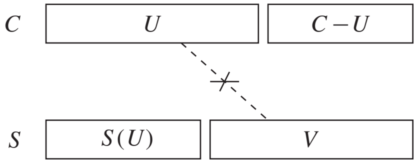

Figure 5: The bipartite graph (C, S, E).

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

Using (3.3), we obtain

$$\mu(U \times V) = \mu((C - U) \times S(U)). \tag{3.4}$$

Let D_n be the number of departures of type $(C - U) \times S(U)$ up to time n. Let A_n, B_n, and X_n be defined as above. Set $Z_n = A_n - B_n$. We have

$$X_n \geq A_n - (B_n - D_n) \geq Z_n. \tag{3.5}$$

For an arrival of type $U \times V$, the Z-process makes a +1 jump; for an arrival of type $(C - U) \times S(U)$, the Z-process makes a -1 jump; otherwise, the Z-process remains constant. We have the following two cases.

- If $\mu(U \times V) > 0$ then, according to (3.4), $(Z_n)_n$ is null recurrent. According to (3.5), $(X_n)_n$ is either null recurrent or transient.
- If $\mu(U \times V) = 0$ then X_n is increasing with n. Therefore, either $(X_n)_n$ goes to ∞ with some probability, or $(X_n)_n$ becomes ultimately constant with probability 1. In the second case, to see that the model cannot be stable, start from another initial condition $\widetilde{X}_0 > \lim_n X_n$.

Hence, in all cases, the model cannot be stable.

Remark 3.1. Consider a *nonconnected* matching graph (C, S, E). Consider a probability μ and an admissible matching policy such that the bipartite matching model is stable. Let (C', S', E') be a connected subgraph of (C, S, E). Following the exact same steps as in the proof of Lemma 3.1, we prove that

$$\mu_C(C') = \mu_S(S'), \qquad \mu(C' \times (S - S')) = 0, \qquad \mu((C - C') \times S') = 0$$

(otherwise, the Markov chain is either transient or null recurrent). Therefore, we can decompose the model into connected components and treat them separately. Hence, the assumption of connectedness of (C, S, E) in Definition 2.1 was made without loss of generality.

3.1. Complexity of verifying NCond

Let us fix (C, S, E) and the probability measures (μ_C, μ_S) such that $\operatorname{supp}(\mu_C) = C$ and $\operatorname{supp}(\mu_S) = S$. We want an efficient algorithm to decide if the conditions NCond are satisfied.

The number of inequalities in NCond is exponential in $|C| + |S|$. So, checking directly if all the inequalities are satisfied is a method whose time complexity is exponential in $|C| + |S|$. To go beyond, we need additional material.

We use the standard terminology of network flow theory; see, for instance, [6]. Consider the directed graph (see Figure 6)

$$\mathcal{N} = (C \cup S \cup \{i, f\}, E \cup \{(i, c),\ c \in C\} \cup \{(s, f),\ s \in S\}). \tag{3.6}$$

Endow the arcs of E with infinite capacity, an arc of type (i, c) with capacity $\mu_C(c)$, and an arc of type (s, f) with capacity $\mu_S(s)$.

Recall that a *cut* is a subset of the arcs whose removal disconnects i and f. The *capacity* of a cut is the sum of the capacities of the arcs. Set $A = E \cup \{(i, c),\ c \in C\} \cup \{(s, f),\ s \in S\}$. Recall that $T\colon A \to \mathbb{R}_+$ is a *flow* if (i) for all c, $T(i, c) = \sum_{s \in S(c)} T(c, s)$ and for all s, $\sum_{c \in C(s)} T(c, s) = T(s, f)$; (ii) for all $(x, y) \in E$, $T(x, y)$ is less than or equal to the capacity of (x, y). The *value* of T is $\sum_c T(i, c) = \sum_s T(s, f)$.

Let NCond$_{\leq}$ be the set of inequalities obtained from NCond by replacing the strict inequalities by large inequalities.

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

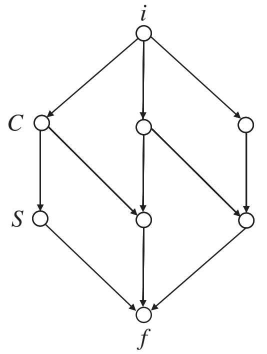

FIGURE 6: The graph $\mathcal{N}$ associated with the NN model of Figure 2.

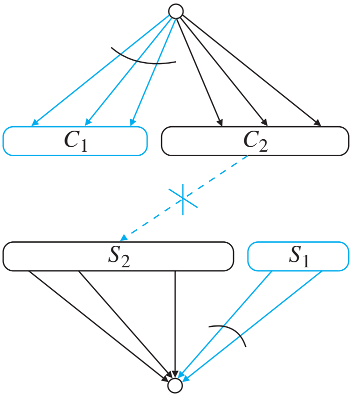

FIGURE 7: Illustration of the proof of Lemma 3.2.

Lemma 3.2. *There exists a flow of value* 1 *in* $\mathcal{N}$ *if and only if* (μ_C, μ_S) *satisfies* NCOND$_\leq$. *There exists a flow* T *of value* 1 *such that* $T(c, s) > 0$ *for all* $(c, s) \in E$ *if and only if* (μ_C, μ_S) *satisfies* NCOND.

The first part of Lemma 3.2 is proved in [3, Proposition 3.7]. We reproduce the argument for completeness.

Proof of Lemma 3.2. The celebrated max-flow min-cut Theorem [6] states that the maximal value of a flow is equal to the minimal capacity of a cut. Observe that the set of arcs $\{(i, c), c \in C\}$ forms a cut of capacity 1. Therefore, the maximal flow is less than or equal to 1, and it is 1 if and only if all cuts have a capacity greater than or equal to 1.

To be of finite capacity, a cut must contain only customer arcs $\{(i, c), c \in C\}$ and server arcs $\{(s, f), s \in S\}$. Consider a subset $\mathcal{A} = \{(i, c), c \in C_1\} \cup \{(s, f), s \in S_1\}$. Set $C_2 = C - C_1$ and $S_2 = S - S_1$ (see Figure 7). The set $\mathcal{A}$ is a cut if and only if $C_2 \times S_2 \cap E = \varnothing$, or, equivalently, if and only if $S(C_2) \subset S_1$ and $C(S_2) \subset C_1$. Also, the capacity of $\mathcal{A}$ is $\mu_C(C_1) + \mu_S(S_1)$.

Assume that the cut $\{(i, c), c \in C_1\} \cup \{(s, f), s \in S_1\}$ is of capacity strictly less than 1. We have

$$\mu_C(C_1) + \mu_S(S_1) < 1 \quad \Longleftrightarrow \quad \mu_C(C_1) < \mu_S(S_2).$$

But, $C(S_2) \subset C_1$, so, if NCOND$_\leq$ is satisfied, we must have

$$\mu_S(S_2) \leq \mu_C(C(S_2)) \leq \mu_C(C_1).$$

So we have proved that NCOND$_\leq$ is not satisfied.

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

Conversely, if $\text{NC}\textsc{ond}_{\leq}$ is not satisfied then there exist C_1, S_2, $C(S_2) = C_1$ such that $\mu_C(C_1) < \mu_S(S_2)$. Set $C_2 = C - C_1$ and $S_1 = S - S_2$. By definition, $C_2 \times S_2 \cap E = \varnothing$; therefore, $\{(i, c),\ c \in C_1\} \cup \{(s, f),\ s \in S_1\}$ is a cut. Its capacity is $\mu_C(C_1) + \mu_S(S_1) < 1$.

By contraposing the above, we obtain

$$\text{NC}\textsc{ond}_{\leq}\text{ satisfied} \quad\Longleftrightarrow\quad \text{all cuts have a capacity} \geq 1 \quad\Longleftrightarrow\quad \text{maximal flow is } 1.$$

We now prove the second part of the lemma. Assume that the conditions NCOND are not satisfied. If the conditions $\text{NC}\textsc{ond}_{\leq}$ are not satisfied either then, by the first part of the proof, there exists no flow of value 1. Assume now that the conditions $\text{NC}\textsc{ond}_{\leq}$ are satisfied. Then there exists $U \subset C$, $U \neq C$, such that $\mu_C(U) = \mu_S(S(U))$. Let T be any flow of value 1. Using the flow relation for U, we obtain

$$\sum_{(c,s)\in U\times S(U)} T(c, s) = \sum_{(i,c)\in\{i\}\times U} T(i, c) = \mu_C(U).$$

Using $\mu_C(U) = \mu_S(S(U))$ and the flow relation for $S(U)$, we deduce that

$$\sum_{(c,s)\in U\times S(U)} T(c, s) = \mu_S(S(U)) \quad\Longrightarrow\quad \sum_{(c,s)\in(C-U)\times S(U)} T(c, s) = 0.$$

Now it follows from the connectedness of (C, S, E) that $(C - U) \times S(U) \cap E \neq \varnothing$. We conclude that the flow T is such that $T(c, s) = 0$ for some $(c, s) \in E$.

Assume now that the conditions NCOND are satisfied. Fix η such that $0 < \eta < 1/|E|$. Consider the function $T_\eta\colon A \to \mathbb{R}_+$ defined by

$$T_\eta(x, y) = \begin{cases} \eta & \text{for } (x, y) = (c, s) \in E, \\ |S(c)|\eta & \text{for } (x, y) = (i, c), \\ |C(s)|\eta & \text{for } (x, y) = (s, f). \end{cases}$$

By construction, T_η is a flow. Set

$$\widetilde{\mu}_C(c) = \frac{\mu_C(c) - |S(c)|\eta}{1 - |E|\eta}, \qquad \widetilde{\mu}_S(s) = \frac{\mu_S(s) - |C(s)|\eta}{1 - |E|\eta}. \tag{3.7}$$

For small enough η, observe that $\widetilde{\mu}_C$ and $\widetilde{\mu}_S$ are probability measures on C and S, respectively. Choose η small enough such that $(\widetilde{\mu}_C, \widetilde{\mu}_S)$ satisfies NCOND. This is possible since the conditions NCOND are open conditions.

Consider the directed graph $\mathcal{N}$, see (3.6), with new capacities on the customer and server arcs defined by $\widetilde{\mu}_C$ and $\widetilde{\mu}_S$. By applying the first part of the proof, there exists a flow $\widetilde{T}\colon A \to \mathbb{R}_+$ of value 1. Define

$$T\colon A \to \mathbb{R}_+, \qquad T = T_\eta + (1 - |E|\eta)\widetilde{T}.$$

By construction, T is a flow for the graph $\mathcal{N}$ with the original capacity constraints (μ_C for the customer arcs and μ_S for the server arcs). The value of T is 1 and it satisfies $T(x, y) > 0$ for all $(x, y) \in E$. This completes the proof.

There exist algorithms to find the maximal flow which are polynomial in the size of the underlying graph, independent of the arc capacities. For instance, the classical 'augmenting path algorithm' of Edmonds and Karp [5] operates in $O((|C| + |S|)|E|^2)$ time, and there exist more sophisticated algorithms operating in $O((|C| + |S|)^3)$ time.

Take one of these polynomial algorithms, call it MAXFLOW and consider it as a blackbox. We build on this to design a polynomial algorithm to check NCOND. Let us detail the construction.

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

Lemma 3.3. *Define $(\widetilde{\mu}_C, \widetilde{\mu}_S)$ as in (3.7). The pair (μ_C, μ_S) satisfies* NCOND *if and only if the pair $(\widetilde{\mu}_C, \widetilde{\mu}_S)$ satisfies* NCOND *for η strictly positive and small enough.*

Proof. Assume that (μ_C, μ_S) satisfies NCOND. Since we are dealing with open conditions, any small enough perturbation of (μ_C, μ_S) still satisfies NCOND.

Assume now that (μ_C, μ_S) does not satisfy NCOND. There exists $U \subset C$, $U \neq C$, such that $\mu_C(U) \geq \mu_S(S(U))$. By using (3.7), we obtain

$$(1 - |E|\eta)\widetilde{\mu}_C(U) + \Bigg(\sum_{c\in U} |S(c)|\Bigg)\eta \geq (1 - |E|\eta)\widetilde{\mu}_S(S(U)) + \Bigg(\sum_{s\in S(U)} |C(s)|\Bigg)\eta,$$

$$\begin{aligned}(1 - |E|\eta)\widetilde{\mu}_C(U) + |E \cap (U \times S(U))|\eta \geq{} & (1 - |E|\eta)\widetilde{\mu}_S(S(U)) \\ & + |E \cap (C(S(U)) \times S(U))|\eta.\end{aligned}$$

By definition, we have $U \subset C(S(U))$. We conclude that $\widetilde{\mu}_C(U) \geq \widetilde{\mu}_S(S(U))$. So the pair $(\widetilde{\mu}_C, \widetilde{\mu}_S)$ does not satisfy NCOND.

Using Lemmas 3.2 and 3.3, NCOND is satisfied if and only if MAXFLOW$(\mathcal{N}, \widetilde{\mu}_C, \widetilde{\mu}_S)$ returns 1 for small enough η. So, the trick is to run MAXFLOW on the input $(\mathcal{N}, \widetilde{\mu}_C, \widetilde{\mu}_S)$ by considering η as a formal parameter made 'as small as needed'. When running MAXFLOW on $(\mathcal{N}, \widetilde{\mu}_C, \widetilde{\mu}_S)$, the algorithm deals with values of the type $x + y\eta$, and adds and compares them according to the rules below.

Consider quantities of the type $x + y\eta$ for $x, y \in \mathbb{R}$. Addition is defined as follows: if $x_1, x_2, y_1, y_2 \in \mathbb{R}$ then $(x_1 + y_1\eta) + (x_2 + y_2\eta) = (x_1 + x_2) + (y_1 + y_2)\eta$. Comparisons are defined as follows:

$$\begin{aligned}[x_1 + y_1\eta = x_2 + y_2\eta] \quad &\Longleftrightarrow \quad [x_1 = x_2,\ y_1 = y_2], \\ [x_1 + y_1\eta < x_2 + y_2\eta] \quad &\Longleftrightarrow \quad [(x_1 < x_2) \text{ or } (x_1 = x_2,\ y_1 < y_2)]. \end{aligned} \tag{3.8}$$

So, η should be considered small enough not to reverse any strict inequality. On any given input, the MAXFLOW algorithm will stop in finite time, so it will have performed only a finite number of operations (additions and comparisons). Therefore, it is possible, a posteriori, to assign to η a value which is small enough to enforce (3.1).

Algorithm 2 below returns 'yes' if NCOND is satisfied and 'no' otherwise. The termination is obvious and the correctness follows from Lemmas 3.2 and 3.3. As a consequence, we get the following result.

Algorithm 2. (*Checking* NCOND.)
Data: (C, S, E), (μ_C, μ_S) such that $\text{supp}(\mu_C) = C$, $\text{supp}(\mu_S) = S$. *Result:*'Yes' if NCOND, 'no' if ¬(NCOND).

```
begin
   Compute $\mathcal{N}, \widetilde{\mu}_C, \widetilde{\mu}_S$;
      if MAXFLOW($\mathcal{N}, \widetilde{\mu}_C, \widetilde{\mu}_S$) = 1 then
         RESULT ← yes;
      else
         RESULT ← no;
end
return RESULT;
```

Proposition 3.1. *Given a bipartite model $[(C, S, E), \mu]$, there exists an algorithm of time complexity $O((|C| + |S|)^3)$ to decide if* NCOND *is satisfied.*

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

4. Connectivity properties of the Markov chain

Define the following property for the transition graph of the Markov chain.

(UTC) A unique (terminal) strictly connected component with all states leading to it.

Property (UTC) is necessary for stability as defined in Definition 3.1. However, property (UTC) is not granted in bipartite matching models and counterexamples are given below (Examples 4.2 and 4.3). In fact, we will see that we are in an unusual situation: the necessary stability conditions NCond turn out to be sufficient conditions for the property (UTC) (Theorem 4.2)! Observe also that property (UTC) is weaker than irreducibility, and we will give an example of a model satisfying NCond and (UTC) without being irreducible (Example 4.4).

4.1. Stable structures

To establish property (UTC), we make a detour by introducing and studying a notion of independent interest: stable structures.

Definition 4.1. A bipartite matching structure (C, S, E, F) is *stable* if there exists a probability measure μ satisfying (2.1) and whose marginals μ_C and μ_S satisfy NCond.

The justification for this terminology will appear in Section 7: we prove there that, under the ML policy, any model satisfying NCond is stable. So a structure is stable if and only if there exists an associated model which is stable.

On the one hand, there exist stable structures. Consider the following example.

Example 4.1. Consider $(C, S, E, C\times S)$, where (C, S, E) is the NN bipartite graph of Figure 2. Let

$$\begin{aligned} \mu_C:&\quad \mu_C(1)=\mu_C(2)=\tfrac{2}{5}, &\quad \mu_C(3)=\tfrac{1}{5},\\ \mu_S:&\quad \mu_S(1')=\mu_S(2')=\tfrac{2}{5}, &\quad \mu_S(3')=\tfrac{1}{5}.\end{aligned}$$

The product measure $\mu=\mu_C\times\mu_S$ has marginals μ_C and μ_S, and we check that (μ_C, μ_S) satisfies NCond. (The probability μ is associated with the flow $T(c, s)=\frac{1}{5}$ for all $(c, s)\in E$; see Lemma 3.2.) Also, it is easily proved that, for any admissible matching policy, the graph of the Markov chain is irreducible.

On the other hand, there exist unstable structures. We illustrate this using two examples.

Example 4.2. Consider the structure (C, S, E, F), where (C, S, E) is the NN graph of Figure 2 and

$$F=\{(1,3'),(2,2'),(3,1')\}.$$

Consider any μ with $\operatorname{supp}(\mu)=F$. We have $\mu_C(1)=\mu_S(3')=\mu(1,3')$, which violates NCond for $V=\{3'\}$. We can also prove that property (UTC) is not satisfied. Consider a state of the type (x, y) with $x=y=(0,0,k)$ for some $k\geq 0$. Any one of the three possible arrivals leave the state unchanged. In particular, there is an infinite number of terminal components.

Example 4.3. Consider the bipartite matching structure defined in Figure 8. The left-hand diagram shows the graph (C, S, E) and the right-hand diagram shows the graph (C, S, F).

Consider any μ with $\operatorname{supp}(\mu)=F$. We have

$$\mu_S(\{1', 2'\})=\mu(3,1')+\mu(4,2')\leq\mu_C(\{3,4\}),$$

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

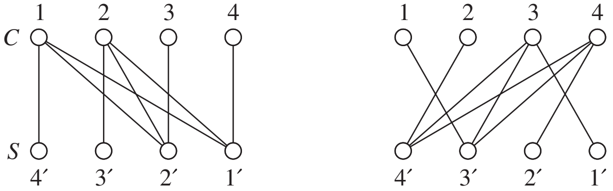

FIGURE 8: *Left*: the matching graph (C, S, E). *Right*: the arrival graph (C, S, F).

which contradicts NCOND for $U = \{3, 4\}$. We can also prove that property (UTC) is not satisfied. Consider a state (x, y) with $x_3 + x_4 = k > 0$. Reducing the number of customers of classes 3 and 4 would require an arrival of type $(1, 1')$ or $(1, 2')$ or $(2, 1')$ or $(2, 2')$. But none of these pairs belong to F. Therefore, it is impossible to reach a state (x', y') with $x'_3 + x'_4 < x_3 + x_4$. On the other hand, an arrival of type $(3, 3')$ or $(3, 4')$ or $(4, 3')$ or $(4, 4')$ strictly increases the number of customers of classes 3 and 4. Hence, all the states are transient, and there is no terminal strongly connected component.

Stability of a structure is a decidable property. There exists a probability measure μ with the requested properties if and only if the following system of linear inequalities in the indeterminates $\mu(c, s)$, $c \in C$, $s \in S$, have a solution:

$$\begin{aligned}
&\sum_{(c,s)\in C\times S} \mu(c, s) = 1,\\
&\mu(c, s) > 0 \quad \text{for all } (c, s) \in F,\\
&\mu(c, s) = 0 \quad \text{for all } (c, s) \in C \times S - F,\\
&\mu_C(c) = \sum_{s\in S} \mu(c, s) \quad \text{for all } c \in C,\\
&\mu_S(s) = \sum_{c\in C} \mu(c, s) \quad \text{for all } s \in S,\\
&\text{NCOND}.
\end{aligned}$$

However, the number of inequalities is exponential in $|C| + |S|$. We are going to propose a criterion which is much simpler, both conceptually and algorithmically.

Consider a bipartite matching structure (C, S, E, F). Define $\widetilde{F} = \{(s, c) \mid (c, s) \in F\}$. Associate with the structure the *directed* graph $(C \cup S, E \cup \widetilde{F})$; in other words, the nodes are $C \cup S$ and the arcs are

$$c \to s \quad \text{if } (c, s) \in E, \qquad s \to c \quad \text{if } (c, s) \in F.$$

We have represented in Figure 9 the directed graph associated with the structure of Example 4.3.

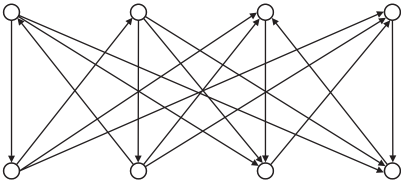

FIGURE 9: The directed graph associated with the structure of Figure 8.

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

The graph of Figure 9 is not strongly connected: the four nodes on the right form a strongly connected component. Similarly, the directed graph associated with the structure of Example 4.2 is not strongly connected. On the other hand, the directed graph associated with the structure of Example 4.1 is strongly connected. This is not a coincidence.

Theorem 4.1. *Let (C, S, E, F) be a bipartite matching structure. The following two properties are equivalent:*

- *(C, S, E, F) is a stable structure;*
- *$(C \cup S, E \cup \widetilde{F})$ is strongly connected.*

We are not aware of any result of this flavor in the literature. In particular, one can decide if a structure is stable by testing the strong connectivity of $(C \cup S, E \cup \widetilde{F})$. This can be done with time complexity $O(|E| + |F|)$ using a depth-first search algorithm.

Proof of Theorem 4.1. Assume that (C, S, E, F) is a stable structure. Let μ be a probability measure satisfying (2.1) and NCOND. Suppose that there exist $c \in C$ and $s \in S$ with no directed path from c to s in $(C \cup S, E \cup \widetilde{F})$. Let $\operatorname{succ}(c)$ be the set of nodes that can be reached starting from c in $(C \cup S, E \cup \widetilde{F})$. Set

$$C_1 = C \cap \operatorname{succ}(c), \qquad S_1 = S \cap \operatorname{succ}(c), \qquad C_2 = C - C_1, \qquad S_2 = S - S_1.$$

By assumption, $s \in S_2$. By construction, the following two properties hold:

$$\mu(C_2, S_1) = 0, \qquad (C_1 \times S_2) \cap E = \varnothing.$$

Using $\mu(C_2, S_1) = 0$, we obtain

$$\mu_S(S_1) = \mu(C_1, S_1) \leq \mu_C(C_1).$$

But, using $[(C_1 \times S_2) \cap E = \varnothing]$ and NCOND for $U = C_1$, we obtain

$$\mu_C(C_1) < \mu_S(S(C_1)) = \mu_S(S_1).$$

From this contradiction, we deduce that, for all $c \in C$ and $s \in S$, there exists a directed path from c to s in $(C \cup S, E \cup \widetilde{F})$. Similarly, we can prove that, for all $s \in S$ and $c \in C$, there exists a directed path from s to c in $(C \cup S, E \cup \widetilde{F})$.

Assume now that $(C \cup S, E \cup \widetilde{F})$ is strongly connected. We are going to construct an explicit probability μ satisfying NCOND, using Perron–Frobenius theory. Consider the matrices $A \in \mathbb{R}_+^{C \times S}$ and $B \in \mathbb{R}_+^{S \times C}$ defined by

$$A_{cs} = \begin{cases} 1/|S(c)| & \text{if } (c, s) \in E, \\ 0 & \text{otherwise,} \end{cases} \qquad B_{sc} = \begin{cases} 1/\#\{d, (s, d) \in \widetilde{F}\} & \text{if } (s, c) \in \widetilde{F}, \\ 0 & \text{otherwise.} \end{cases}$$

By construction, matrices A and B are stochastic. Consider the stochastic matrix $AB \in \mathbb{R}_+^{C \times C}$. By construction, we have $(AB)_{cd} > 0$ if and only if there is a path of length 2 from c to d in the graph $(C \cup S, E \cup \widetilde{F})$. Since $(C \cup S, E \cup \widetilde{F})$ is strongly connected, we deduce that AB is irreducible. Since AB is stochastic, its spectral radius is 1. Applying the Perron–Frobenius theorem [9], we obtain the existence of a line vector $x \in \mathbb{R}_+^C$ such that, for all $c, x_c > 0$, $\sum_c x_c = 1$, and $xAB = x$. Set $y = xA$. Define the probability measure μ on $C \times S$

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

by $\mu(c, s) = y_s B_{sc}$. By construction, we have $\mu_C = x$ and $\mu_S = y$. Also, by construction, $\mathrm{supp}(\mu) = F$. Define the function $T\colon A \to \mathbb{R}_+$ by

$$T(i, c) = x_c \quad \text{for all } c \in C, \qquad T(s, f) = y_s \quad \text{for all } s \in S,$$
$$T(c, s) = x_c A_{cs} \quad \text{for all } (c, s) \in E.$$

By construction, T is a flow of value 1 such that $T(c, s) > 0$ for all $(c, s) \in E$. Using Lemma 3.2, we find that $(x, y) = (\mu_C, \mu_S)$ satisfies NCond.

4.2. Back to property (UTC)

We now have all the ingredients needed to prove the following result.

Theorem 4.2. *Consider a bipartite matching model* $[(C, S, E), \mu, \text{Pol}]$. *Assume that the structure* (C, S, E, F) *is stable, or, equivalently, that* $(C \cup S, E \cup \widetilde{F})$ *is strongly connected. Then the transition graph of the Markov chain of the bipartite matching model satisfies property (UTC).*

Proof. We are going to prove that the empty state can be reached starting from any state. This is a sufficient condition for property (UTC) to hold. The unique terminal strongly connected component is the set of states that can be reached from the empty state.

We carry out the proof in the commutative case, but it works unchanged in the noncommutative case (the only information needed is the number of customers/servers of each class). Consider a general nonempty state (X, Y). Recall that $|X| = |Y|$. To prove that the empty state is reachable, it is sufficient to prove that we can always reach a state (X', Y') such that $|X'| < |X|$. To prove this, the argument is in relation to the facet to which (X, Y) belongs; see Definition 2.3.

Set $X = (x_c)_{c\in C}$ and $Y = (y_s)_{s\in S}$. Assume that (X, Y) belongs to the facet $\mathcal{F} = (U, V)$, i.e. $x_c > 0$ if and only if $c \in U$, and $y_s > 0$ if and only if $s \in V$. We lighten the notation by replacing $C_\bullet(\mathcal{F})$, $S_\bullet(\mathcal{F})$, $C_\circledcirc(\mathcal{F})$, etc., by $C_\bullet$, $S_\bullet$, $C_\circledcirc$, etc. If there exists $(c, s) \in C_\circledcirc \times S_\circledcirc$ such that $\mu(c, s) > 0$, then the proof is completed. Assume now that $\mu(C_\circledcirc \times S_\circledcirc) = 0$. Choose $(c, s) \in C_\circledcirc \times S_\circledcirc$. By assumption, there exists a path from s to c in $(C \cup S, E \cup \widetilde{F})$. Let us denote it by $(s = s_1, c_1, s_2, c_2, \dots, s_k, c_k = c)$; see Figure 10. Assume that $c_1, \dots, c_{k-1} \notin C_\circledcirc$ and $s_2, \dots, s_k \notin S_\circledcirc$. (If not, consider a subpath satisfying the two properties.) Then $c_1 \in C_\circ$. Indeed, if $c_1 \in C_\bullet$ then $(c_1, s_2) \in E$ and $s_2 \in S_\circledcirc$, which contradicts the above assumption. Now, since $(s_i, c_i) \in \widetilde{F}$ and $(c_i, s_{i+1}) \in E$ by construction and since $c_1 \in C_\circ$, we obtain $c_1, \dots, c_{k-1} \in C_\circ$ and $s_2, \dots, s_k \in S_\circ$.

By definition of the graph $(C \cup S, E \cup \widetilde{F})$, we have $\mu(c_i, s_i) > 0$ for all i. Choose the sequence of arrivals $(c_1, s_1), \dots, (c_k, s_k)$. Consider the effect of the arrival of (c_1, s_1). Since $s_1 \in S_\circledcirc$, it will be matched with a customer of $C_\bullet$ (and not with c_1, even if $(c_1, s_1) \in E$, since an admissible matching policy is always *buffer-first*; see Section 2.2). Since $c_1 \notin C_\circledcirc$, it will remain unmatched. Let $(X^{(1)}, Y^{(1)})$ be the new state. We have $|X^{(1)}| = |X|$. Also, in the new state, we have $c_1 \in C_\bullet^{(1)}$, which implies that $s_2 \in S_\circledcirc^{(1)}$. So we can repeat the argument inductively.

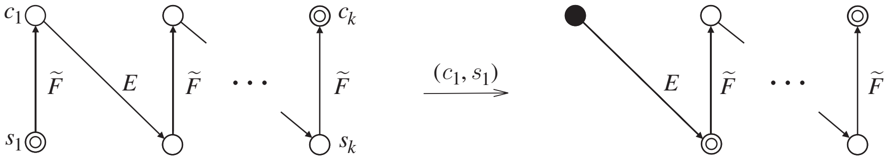

Figure 10: The path $(s = s_1, c_1, s_2, c_2, \dots, s_k, c_k = c)$ in $(C \cup S, E \cup \widetilde{F})$.

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

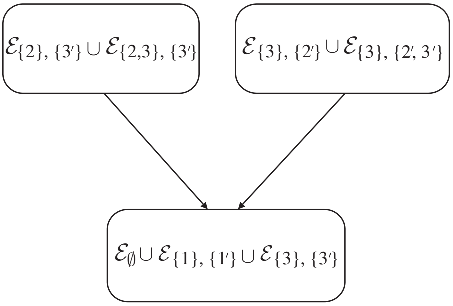
$\mathcal{E}_{\{2\},\{3'\}} \cup \mathcal{E}_{\{2,3\},\{3'\}}$
$\mathcal{E}_{\{3\},\{2'\}} \cup \mathcal{E}_{\{3\},\{2',3'\}}$
$\mathcal{E}_{\emptyset} \cup \mathcal{E}_{\{1\},\{1'\}} \cup \mathcal{E}_{\{3\},\{3'\}}$

FIGURE 11: Transition graph for the RANDOM policy.

After the arrivals of $(c_1, s_1), \dots, (c_{k-1}, s_{k-1})$, we are in a state $(X^{(k-1)}, Y^{(k-1)})$ satisfying

$$|X^{(k-1)}| = |X|, \qquad s_k \in S_{\circledcirc}^{(k-1)}, \ c_k \in C_{\circledcirc}^{(k-1)}.$$

Therefore, after the arrival of (c_k, s_k), we end up in a state $(X^{(k)}, Y^{(k)})$ such that $|X^{(k)}| = |X| - 1$. This completes the proof.

Example 4.4. Consider a bipartite matching model associated with the structure (C, S, E, F), where (C, S, E) is the NN graph of Figure 2 and

$$F = \{(1, 1'), (2, 2'), (3, 3')\}.$$

The graph $(C \cup S, E \cup \widetilde{F})$ is strongly connected. According to Theorem 4.2, the transition graph of the Markov chain satisfies property (UTC). However, it is not irreducible. Indeed, for any policy, it is impossible to reach the state $((0, 1, 0); (0, 0, 1))$ starting from the empty state. To further illustrate, let us give the precise structure of the transition graph for the RANDOM policy. Denote by $\mathcal{E}_{\mathcal{F}}$ the set of states belonging to the facet $\mathcal{F}$. The transition graph is represented in Figure 11.

Below, we study the stability of bipartite matching models. Therefore, we always assume that the necessary conditions NCOND are satisfied. So we obtain property (UTC) for the Markov chain as a consequence of Theorem 4.2.

5. Models that are stable for all admissible policies

Definition 5.1. Consider a bipartite graph (C, S, E) and an admissible matching policy POL. The *stability region* is the set of values of μ for which the bipartite matching model $[(C, S, E), \mu, \text{POL}]$ is stable.

The stability region is included in the polyhedron defined by NCOND. The stability region is *maximal* if it is equal to this polyhedron.

Let us introduce a new set of conditions, SCOND, that defines a polyhedron included in the stability region. These conditions are simply those that allow us to prove stability using a linear Lyapunov function (see the proof of Proposition 5.1).

Denote by $\mathfrak{F}$ the set of facets. Define the following conditions on μ (referred to as SCOND):

$$\mu_C(C_{\circledcirc}(\mathcal{F})) + \mu_S(S_{\circledcirc}(\mathcal{F})) > 1 - \mu(E \cap C_{\circ}(\mathcal{F}) \times S_{\circ}(\mathcal{F})) \quad \text{for all } \mathcal{F} \in \mathfrak{F} - \{\varnothing\}. \quad (5.1)$$

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

Let $\mathcal{F}$ be a saturated facet; see Definition 2.4. Assume for instance that $C_\circ(\mathcal{F}) = \varnothing$. Then $E \cap C_\circ(\mathcal{F}) \times S_\circ(\mathcal{F}) = \varnothing$ and $C_\circledcirc(\mathcal{F}) = C - C_\bullet(\mathcal{F})$. So (5.1) implies that

$$\mu_S(S_\circledcirc(\mathcal{F})) > \mu_C(C_\bullet(\mathcal{F})).$$

Since $S_\circledcirc(\mathcal{F}) = S(C_\bullet(\mathcal{F}))$, we recognize exactly (3.2) for $U = C_\bullet(\mathcal{F})$. Conversely, consider $U \subsetneq C$ and the associated condition in NCOND: $\mu_C(U) < \mu_S(S(U))$. Choose a state with a strictly positive number of customers/servers for the classes U and $S - S(U)$. Let $\mathcal{F}$ be the corresponding facet. The facet $\mathcal{F}$ is saturated: $S_\circledcirc(\mathcal{F}) = S(U)$, $S_\bullet(\mathcal{F}) = S - S(U)$, $S_\circ(\mathcal{F}) = \varnothing$. Let us apply (5.1) to the facet $\mathcal{F}$:

$$\mu_C(C_\circledcirc(\mathcal{F})) + \mu_S(S(U)) > 1,$$
$$\mu_S(S(U)) > \mu_C(C - C_\circledcirc(\mathcal{F})) \geq \mu_C(U).$$

To summarize, the subset of the inequalities (5.1) obtained by considering only the saturated facets gives precisely the inequalities NCOND.

We now show that the conditions SCOND are sufficient stability conditions.

Proposition 5.1. *A bipartite model with probability μ satisfying* SCOND *is stable under any admissible matching policy.*

Proof. Consider the linear Lyapunov function

$$L(u, v) = |u|, \qquad (u, v) \in \mathcal{E},$$

the number of unmatched customers (servers). Let $(U_n, V_n)_n$ be the Markov chain of the buffer content. Let $\mathcal{F} \neq \varnothing$ be an arbitrary and fixed facet. Then, for any $(u, v) \in \mathcal{F}$, we have (see Table 1)

$$\begin{aligned}
&\mathbb{E}[L(U_{n+1}, V_{n+1}) \mid (U_n, V_n) = (u, v)] - L(u, v) \\
&\quad = -\mu(C_\circledcirc(\mathcal{F}), S_\circledcirc(\mathcal{F})) + \mu(C_\circ(\mathcal{F}), S_\bullet(\mathcal{F})) + \mu(C_\bullet(\mathcal{F}), S_\bullet(\mathcal{F})) \\
&\qquad + \mu(C_\bullet(\mathcal{F}), S_\circ(\mathcal{F})) + \mu(C_\circ(\mathcal{F}) \times S_\circ(\mathcal{F}) \cap E^c) \\
&\quad = 1 - \mu_C(C_\circledcirc(\mathcal{F})) - \mu_S(S_\circledcirc(\mathcal{F})) - \mu(C_\circ(\mathcal{F}) \times S_\circ(\mathcal{F}) \cap E).
\end{aligned}$$

Inequality (5.1) implies directly that

$$\mathbb{E}[L(U_{n+1}, V_{n+1}) \mid (U_n, V_n) = (u, v)] - L(u, v) < \epsilon < 0.$$

By application of the Lyapunov–Foster theorem, see, for instance, [2, Section 5.1], we conclude that the model is stable.

Corollary 5.1. *Consider a bipartite graph in which any nonzero facet is saturated. For any admissible matching policy, the stability region is maximal.*

TABLE 1: Variation of the linear Lyapunov function.

	$C_\circledcirc$	$C_\circ$	$C_\bullet$
$S_\circledcirc$	-1	0	0
$S_\circ$	0	0 or 1	1
$S_\bullet$	0	1	1

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

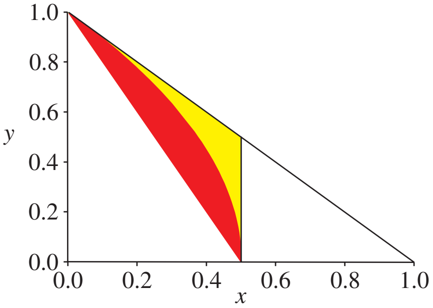

FIGURE 12: NCOND and SCOND for the NN graph with $\mu = \mu_C \times \mu_S$ and $\mu_C = \mu_S$.

The bipartite graph ($C = \{1, 2\},\ S = \{1', 2'\}, C \times S - \{(2, 2')\}$) is such that any nonzero facet is saturated. Therefore, its stability region is maximal for any admissible policy. The same is true for the 'almost complete graphs' ($C = \{1, \ldots, k\},\ S = \{1', \ldots, k'\}, C \times S - \{(i, i')$ for all $i\}$).

Example 5.1. Consider the NN graph from Figure 2. The graph has only one nonzero facet that is nonsaturated, facet ($\{3\}, \{3'\}$). For any admissible policy, the stability region is at least the polyhedron SCOND, Proposition 5.1, which is defined by the conditions NCOND and

$$\mu_C(1) + \mu_S(1') > 1 - \mu(2, 2').$$

Assume now that $\mu = \mu_C \times \mu_S$ and $\mu_C = \mu_S$. Set $x = \mu_C(1) = \mu_S(1')$ and $y = \mu_C(2) = \mu_S(2')$. Then

$$\begin{array}{lllll} \text{NCOND}: & x < 0.5 & \text{and} & 2x + y > 1, \\ \text{SCOND}: & \text{NCOND} & \text{and} & 2x + y^2 > 1. \end{array}$$

In Figure 12, the light-shaded region corresponds to SCOND, and the union of the light- and dark-shaded regions corresponds to NCOND.

Unfortunately, for some bipartite graphs, the polyhedron SCOND is empty. This is illustrated by the following example.

Example 5.2. Consider the NNN graph of Figures 13 and 14. The condition SCOND for facet ($\{1\}, \{4'\}$) gives

$$\mu_C(\{3, 4\}) + \mu_S(\{1', 2'\}) > 1 - \mu(2, 3'), \tag{5.2}$$

and, for facet ($\{4\}, \{1'\}$), gives

$$\mu_C(1) + \mu_S(4') > 1 - \mu(2, 2') - \mu(2, 3') - \mu(3, 3'). \tag{5.3}$$

Inequality (5.2) is equivalent to $\mu_C(1) + \mu_S(4') < 1 - \mu_C(2) - \mu(\{1, 3, 4\}, 3')$. Together with (5.3) this gives

$$\begin{aligned} \mu_C(1) + \mu_S(4') &< 1 - \mu_C(2) - \mu(\{1, 3, 4\}, 3') \\ &< 1 - \mu(2, 2') - \mu(2, 3') - \mu(3, 3') < \mu_C(1) + \mu_S(4'), \end{aligned}$$

which is impossible.

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

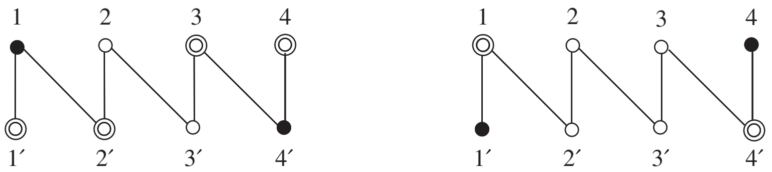

FIGURE 13: NNN graph: facets ({1}, {4′}) and ({4}, {1′}).

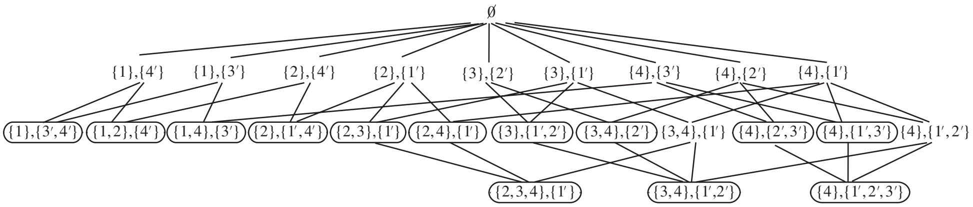

FIGURE 14: Facets for the NNN graph. Saturated facets are encircled (13 among 25 facets).

6. PRIORITIES and MS are not always stable

Consider the NN bipartite graph of Figure 2 and Example 5.1. For this model, Proposition 5.1 does not allow us to decide if the stability region is maximal (see Figure 12). In Figure 15, we give simulation results for the average buffer size up to time $n = 1\,000\,000$ for the NN graph with $\mu = \mu_C \times \mu_S$, $\mu_C = \mu_S$, and the MS policy. We can see that the average buffer size is growing rapidly near the $2x + y = 1$ line. This does not necessarily imply instability, as even for stable models we could have the mean stationary buffer size that is growing unboundedly as we approach the boundary of the stability region. In fact, we show below that, for the PRIORITY and MS matching policies, the stability region is not maximal.

Proposition 6.1. *Consider the NN model with either the* MS *policy or the* PRIORITY *policy, given by*

$$A = \begin{bmatrix} 0 & 2 & 1 \\ 2 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 2 & 1 \\ 2 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

For both policies, the stability region is not maximal.

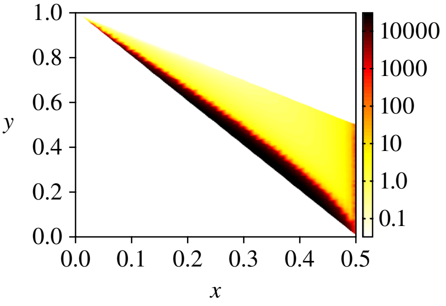

FIGURE 15: Average buffer size for the NN graph with $\mu = \mu_C \times \mu_S$, $\mu_C = \mu_S$, and MS policy; $x = \mu_C(1) = \mu_S(1')$ and $y = \mu_C(2) = \mu_S(2')$.

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

Proof. We carry out the proof for the PRIORITY policy. The idea of the proof is to use the interplay between two different Markov chains: one describing the evolution of the buffer content, and an auxiliary one which mimics the evolution of the customers/servers of classes $2/2'$ on some of the facets.

Consider an auxiliary Markov chain on $\mathbb{Z}$ with the following transition probabilities.

	$x \to x-1$	$x \to x$	$x \to x+1$
$x < 0$	a_{-1}	a_0	a_1
$x = 0$	b_{-1}	b_0	b_1
$x > 0$	c_{-1}	c_0	c_1

Assume that $a_{-1}, a_1, b_{-1}, b_1, c_{-1}, c_1$ are all different from 0. The chain is positive recurrent if and only if $[a_{-1} < a_1,\ c_1 < c_{-1}]$. The stationary distribution is then equal to

$$\pi(0) = \left(1 + \frac{b_{-1}}{a_1 - a_{-1}} + \frac{b_1}{c_{-1} - c_1}\right)^{-1},$$

$$\pi(x) = \pi(0)\frac{b_1}{c_1}\left(\frac{c_1}{c_{-1}}\right)^{x} \quad \text{if } x > 0,$$

$$\pi(x) = \pi(0)\frac{b_{-1}}{a_{-1}}\left(\frac{a_{-1}}{a_1}\right)^{|x|} \quad \text{if } x < 0.$$

Consider an NN model with a probability μ such that $\operatorname{supp}(\mu) = C \times S$. Let $(X, Y) = (X(n), Y(n))_n$ be the Markov chain of the buffer content, where $X(n) = (X_1(n), X_2(n), X_3(n))$ and $Y(n) = (Y_1(n), Y_2(n), Y_3(n))$. Assume without loss of generality that (X, Y) is given under the form of a stochastic recursive sequence, that is,

$$(X(n+1), Y(n+1)) = \Phi(X(n), Y(n), \theta_n),$$

where $(\theta_n)_n$ is an i.i.d. sequence of random variables distributed according to μ, and Φ is a deterministic function.

Consider now the process $(X_2(n) - Y_2(n))_n$. This is not a Markov chain. However, if $X_2(n) + X_3(n) > 0$ then ‘it becomes a Markov chain’. More precisely, if $X_2(n) + X_3(n) > 0$ then

$$X_2(n+1) - Y_2(n+1) = \Phi_2(X_2(n) - Y_2(n), \theta_n), \tag{6.1}$$

where Φ_2 is a deterministic function. This can be checked by direct inspection. Moreover, the transition kernel on $\mathbb{Z}$ defined by recursion (6.1) is of the same type as the above auxiliary chain with parameters

$$\begin{aligned}
a_1 &= \mu(1,1') + \mu(1,3') + \mu(2,1') + \mu(2,3'),\\
a_0 &= \mu(1,2') + \mu(2,2') + \mu(3,1') + \mu(3,3'),\\
a_{-1} &= \mu(3,2'),\\
b_1 &= \mu(2,1') + \mu(2,3'),\\
b_0 &= \mu(1,1') + \mu(1,3') + \mu(2,2') + \mu(3,1') + \mu(3,3'),\\
b_{-1} &= \mu(1,2') + \mu(3,2'),\\
c_1 &= \mu(2,3'),\\
c_0 &= \mu(1,3') + \mu(2,1') + \mu(2,2') + \mu(3,3'),\\
c_{-1} &= \mu(1,1') + \mu(1,2') + \mu(3,1') + \mu(3,2').
\end{aligned}$$

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

TABLE 2: Effect of arrivals.

Arrival	Possible matchings	Selected matchings	ΔX_2
$(1, 1')$	$(1, 3')$, $(2, 1')$, $(3, 1')$	$(1, 3')$, $(2, 1')$	-1
$(1, 2')$	$(1, 3')$, $(1, 2')$, $(2, 2')$	$(1, 3')$, $(2, 2')$	-1
$(2, 1')$	$(2, 1')$, $(3, 1')$	$(2, 1')$	0
$(2, 2')$	$(2, 2')$	$(2, 2')$	0
$(2, 3')$	$\varnothing$	$\varnothing$	$+1$
$(3, 2')$	$(2, 2')$	$(2, 2')$	-1
$(3, 1')$	$(2, 1')$, $(3, 1')$	$(2, 1')$	-1

Let us justify for instance the values of c_{-1}, c_0, c_1. We are in the case $X_2(n) + X_3(n) > 0$ and $X_2(n) - Y_2(n) > 0$, which implies that

$$X_1(n) = 0, \quad X_2(n) > 0, \qquad Y_1(n) = Y_2(n) = 0, \quad Y_3(n) > 0.$$

In Table 2 we show the effect of the different possible classes of arrivals, restricting to the ones which may affect X_2, i.e. when the customer class is 2 or the server class is $1'$ or $2'$. To simplify, we have assumed in Table 2 that $X_3 > 0$. For $X_3 = 0$, the 'Possible matchings' column would be affected, but not the 'Selected matchings' and 'ΔX_2' columns.

Let us comment on a couple of cases. If the arrival is of type $(1, 1')$ then the selected matching is $(2, 1')$ rather than $(3, 1')$ due to the PRIORITY policy ($B_{2,1'} > B_{3,1'}$). If the arrival is of type $(1, 2')$, the selected matching is $(2, 2')$ rather than $(1, 2')$ according to the buffer-first property of admissible policies; see Section 2.2. The other cases are argued similarly.

Let us introduce a new Markov chain $(W_n)_n$ on $\mathbb{Z}$ defined by

$$W_{n+1} = \Phi_2(W_n, \theta_n).$$

(The process $(W_n)_n$ is different from the process $(X_2(n) - Y_2(n))_n$. The former is always defined according to recursion (6.1), while the latter is defined according to (6.1) only for the ns such that $X_2(n) + X_3(n) > 0$. The former is Markovian while the latter is not.)

The condition $c_1 < c_{-1}$ becomes $\mu_C(2) < \mu_S(1') + \mu_S(2')$, and $a_{-1} < a_1$ becomes $\mu_S(2') < \mu_C(1) + \mu_C(2)$. Both conditions follow from NCOND. So the auxiliary chain $(W_n)_n$ is ergodic and its stationary distribution π satisfies

$$\pi(0) = \left(1 + \frac{b_{-1}}{a_1 - a_{-1}} + \frac{b_1}{c_{-1} - c_1}\right)^{-1}, \qquad \pi(\mathbb{Z}_+^*) = \pi(0)\frac{b_1}{c_{-1} - c_1},$$
$$\pi(\mathbb{Z}_-^*) = \pi(0)\frac{b_{-1}}{a_1 - a_{-1}},$$

where $\mathbb{Z}_+^* = \{1, 2, \ldots\}$ is the set of strictly positive integers and $\mathbb{Z}_-^* = \{-1, -2, \ldots\}$ is the set of strictly negative integers. From now on, we fix an initial condition W_0 satisfying

$$W_0 \sim \pi, \qquad W_0 \perp\!\!\!\perp (\theta_n)_n.$$

Let us switch back to the Markov chain (X, Y). Set

$$L(n) = X_2(n) + X_3(n), \qquad \Delta L(n) = L(n+1) - L(n).$$

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

If $L(n) > 0$ then we check by direct inspection that

$$\Delta L(n) = \Psi(X_2(n) - Y_2(n), \theta_n),$$

where Ψ is a deterministic function. We have, in particular,

$$\begin{aligned}\alpha &:= \mathbb{E}[\Delta L(n) \mid L(n) > 0,\ Y_2(n) > 0] = \mu(3, 2') + \mu(3, 3') - \mu(1, 1') - \mu(2, 1'),\\ \beta &:= \mathbb{E}[\Delta L(n) \mid L(n) > 0,\ X_2(n) = Y_2(n) = 0]\\ &= \mu(2, 3') + \mu(3, 2') + \mu(3, 3') - \mu(1, 1'),\\ \gamma &:= \mathbb{E}[\Delta L(n) \mid L(n) > 0,\ X_2(n) > 0] = \mu(2, 3') + \mu(3, 3') - \mu(1, 1') - \mu(1, 2').\end{aligned}$$

Let us turn again to the auxiliary chain $(W_n)_n$. By performing the computation, we obtain

$$\mathbb{E}[\Psi(W_n, \theta_n)] = \pi(\mathbb{Z}_-^*)\alpha + \pi(0)\beta + \pi(\mathbb{Z}_+^*)\gamma.$$

The ergodic theorem for Markov chains, see, for instance, [2, Section 3.4], gives

$$\lim_n \frac{1}{n}\sum_{i=0}^{n-1} \Psi(W_n, \theta_n) = \pi(\mathbb{Z}_-^*)\alpha + \pi(0)\beta + \pi(\mathbb{Z}_+^*)\gamma \quad \text{almost surely.}$$

Assume that $\pi(\mathbb{Z}_-^*)\alpha + \pi(0)\beta + \pi(\mathbb{Z}_+^*)\gamma > 0$. Then we have

$$\lim_n \sum_{i=0}^{n-1} \Psi(W_n, \theta_n) = +\infty \quad \text{almost surely.}$$

Therefore, for each $\varepsilon > 0$, there exists $K_\varepsilon \geq 0$ such that

$$\mathbb{P}\biggl(\min_{n\geq 1} \sum_{i=0}^{n-1} \Psi(W_n, \theta_n) > -K_\varepsilon\biggr) \geq 1 - \varepsilon. \tag{6.2}$$

Let us switch back to the Markov chain $(X(n), Y(n))_n$. Choose the initial condition $(X(0), Y(0))$ such that

$$X_2(0) - Y_2(0) = W_0, \qquad \min(X_3(0), Y_3(0)) = K_\varepsilon,$$

where K_ε is defined in (6.2). By construction, on the event

$$\mathcal{A} = \biggl\{\min_{n\geq 1} \sum_{i=0}^{n-1} \Psi(W_n, \theta_n) > -K_\varepsilon\biggr\},$$

we have

$$L(n) > 0, \qquad X_2(n) - Y_2(n) = W_n, \quad \text{for all } n.$$

So, on the event $\mathcal{A}$, we have

$$L(n) = K_\varepsilon + \sum_{i=0}^{n-1} \Delta L(i) = K_\varepsilon + \sum_{i=0}^{n-1} \Psi(W_i, \theta_i) \to +\infty.$$

We conclude that the Markov chain (X, Y) of the NN model is transient.

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

We now show that the stability region is not maximal, by giving an example such that $\pi(\mathbb{Z}_-^*)\alpha + \pi(0)\beta + \pi(\mathbb{Z}_+^*)\gamma > 0$. Consider $\mu_C = (\frac{1}{3}, \frac{2}{5}, \frac{4}{15})$, $\mu_S = \mu_C$, and $\mu = \mu_C \times \mu_S$. Thus, conditions NCOND are satisfied. However, we have

$$\begin{aligned} a_1 &= \tfrac{11}{25}, & a_0 &= \tfrac{34}{75}, & a_{-1} &= \tfrac{8}{75}, \\ b_1 &= \tfrac{6}{25}, & b_0 &= \tfrac{13}{25}, & b_{-1} &= \tfrac{6}{25}, \\ c_1 &= \tfrac{8}{75}, & c_0 &= \tfrac{34}{75}, & c_{-1} &= \tfrac{11}{25}, \end{aligned}$$

and

$$\pi(0) = \tfrac{25}{61}, \qquad \pi(\mathbb{Z}_+^*) = \tfrac{18}{61}, \qquad \pi(\mathbb{Z}_-^*) = \tfrac{18}{61}.$$

This gives $\alpha = -\frac{1}{15}$, $\beta = \frac{13}{75}$, $\gamma = -\frac{1}{15}$, and

$$\pi(\mathbb{Z}_-^*)\alpha + \pi(0)\beta + \pi(\mathbb{Z}_+^*)\gamma = \tfrac{29}{915} > 0.$$

This completes the proof.

Consider now the MS policy. Set $L(n) = \min(X_3(n) - X_2(n), Y_3(n) - Y_2(n))$. The initial distribution can be taken such that

$$X_2(0) - Y_2(0) \sim \pi, \qquad \begin{cases} X_3(0) - X_2(0) = K_\varepsilon & \text{if } X_2(0) > 0, \\ Y_3(0) - Y_2(0) = K_\varepsilon & \text{otherwise.} \end{cases}$$

Modulo these modifications, the proof carries over unchanged.

7. ML is always stable

In this section we show that the ML policy has a maximal stability region.

The idea of the proof is as follows. Consider the quadratic Lyapunov function

$$L(x, y) = \sum_{c \in C} x_c^2 + \sum_{s \in S} y_s^2, \qquad (x, y) \in \mathcal{E}. \tag{7.1}$$

Observe that the ML policy minimizes the value of this Lyapunov function at each step. We introduce an alternate policy that depends on the arrival distribution μ. For this policy, we manage to prove that the quadratic Lyapunov function has a negative drift outside a finite region.

Theorem 7.1. *For any bipartite graph, the ML policy has a maximal stability region.*

Proof. We introduce an alternate matching policy. This policy is admissible, corresponds to a commutative state space, but does not belong to the policies listed in Section 2.2. It is a probabilistic policy and its specificity is to be facet dependent.

Let us describe the alternate policy on a nonempty facet $\mathcal{F}$. Set $C_\bullet = C_\bullet(\mathcal{F})$, $S_\bullet = S_\bullet(\mathcal{F})$, $C_\circledcirc = C_\circledcirc(\mathcal{F})$, etc. To describe the matching policy, the only thing we have to describe is how to match an arriving customer of class $c \in C_\circledcirc$ or server of class $s \in S_\circledcirc$. Let us concentrate first on a server of class $s \in S_\circledcirc$.

From NCOND,

$$\mu_C(C_\bullet) < \mu_S(S_\circledcirc), \qquad \mu_S(S_\bullet) < \mu_C(C_\circledcirc).$$

We build a directed graph as in (3.6) but restricted to the nodes in $C_\bullet$ and $S_\circledcirc$. Formally,

$$\mathcal{N}_{\mathcal{F}} = (C_\bullet \cup S_\circledcirc \cup \{i, f\}, \{E \cap C_\bullet \times S_\circledcirc\} \cup \{(i, c), c \in C_\bullet\} \cup \{(s, f), s \in S_\circledcirc\}). \tag{7.2}$$

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

Endow the arcs of $E \cap C_\bullet \times S_\circledcirc$ with infinite capacity, an arc of type (i, c) with capacity $\mu_C(c)$, and an arc of type (s, f) with capacity $\mu_S(s)$.

As in Lemma 3.2, NCond implies that the minimal cut of $\mathcal{N}_\mathcal{F}$ has capacity $\mu_C(C_\bullet)$. Any maximal flow T is such that, for all $c \in C_\bullet$, $T(i, c) = \mu_C(c)$ and, for all $s \in S_\circledcirc$, $T(s, f) \leq \mu_S(s)$. Let us prove that there exists a maximal flow T such that

$$T(i, c) = \mu_C(c) \quad \text{for all } c \in C_\bullet, \qquad T(s, f) < \mu_S(s) \quad \text{for all } s \in S_\circledcirc.$$

Define $\widetilde{\mu}_S$ on $S_\circledcirc$ by $\widetilde{\mu}_S(s) = \mu_S(s) - \eta$. Here $\eta > 0$ is chosen to be small enough so that, for all $U \subset C_\bullet$, $\mu_C(U) < \widetilde{\mu}_S(S(U))$ and, for all $V \subset S_\circledcirc$, $\widetilde{\mu}_S(V) < \mu_C(C(V))$. This is possible since NCond are open conditions. Consider the same network as above but with the capacities $\widetilde{\mu}_S(s)$ on the arcs (s, f). The minimal cut still has capacity $\mu_C(C_\bullet)$. A maximal flow T is such that, for all $c \in C_\bullet$, $T(i, c) = \mu_C(c)$ and, for all $s \in S_\circledcirc$, $T(s, f) \leq \widetilde{\mu}_S(s) < \mu_S(s)$. Clearly, T is also a flow for the original network.

The server $s \in S_\circledcirc$ is matched to $c \in C_\bullet \cap C(s)$ randomly, independently of the past, with probability

$$P^\mathcal{F}_{sc} = \frac{1}{\mu_S(s)} \left[T(c, s) + \frac{\mu_S(s) - T(s, f)}{|C_\bullet \cap C(s)|} \right].$$

Let us check that this indeed defines a probability:

$$\begin{aligned}
\sum_{c \in C_\bullet \cap C(s)} P^\mathcal{F}_{sc} &= \frac{1}{\mu_S(s)} \left[\mu_S(s) - T(s, f) + \sum_{c \in C_\bullet \cap C(s)} T(c, s) \right] \\
&= \frac{1}{\mu_S(s)} \left[\mu_S(s) - T(s, f) + T(s, f) \right] \\
&= 1.
\end{aligned}$$

For $c \in C_\bullet$ and $s \in S(c)$, set $\varepsilon_{sc} = (\mu_S(s) - T(s, f))/|C_\bullet \cap C(s)|$. For $c \in C_\bullet$, set $\varepsilon_c = \sum_{s \in S(c)} \varepsilon_{sc}$. We have $\varepsilon_c > 0$. Observe that

$$\sum_{s \in S(c)} \mu_S(s) P^\mathcal{F}_{sc} = \mu_C(c) + \varepsilon_c \quad \text{for all } c \in C_\bullet. \tag{7.3}$$

Symmetrically, we define the directed graph of type (7.2) but on the nodes $C_\circledcirc$ and $S_\bullet$. We build a maximal flow on this new graph as above, and, based on this flow, we define the probability $P^\mathcal{F}_{cs}$ that a customer $c \in C_\circledcirc$ is matched to a server $s \in S_\bullet \cap S(c)$. For $s \in S_\bullet$, we define ε_{cs}, $c \in C(s)$, and ε_s accordingly. We have $\varepsilon_s > 0$.

Let $(X(n), Y(n))_n$ be the Markov chain of the buffer content of the model. Assume that $(X(n), Y(n)) = (x, y) \in \mathcal{F}$ and let $c \in C_\bullet$. We have the following cases.

(i) $X(n + 1)_c = X(n)_c - 1$ if and only if

- the arriving customer is not of class c;
- the arriving server is of class $s \in S(c)$;
- the arriving server is matched with c (probability $P^\mathcal{F}_{sc}$).

This case happens with probability $\alpha_c = \sum_{s \in S(c)} \mu(C - c, s) P^\mathcal{F}_{sc}$.

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

(ii) $X(n+1)_c = X(n)_c + 1$ if and only if

- the arriving customer is of class c;
- the arriving server is not matched with c. This may occur in two possible ways: either the arriving server is of class $s \notin S(c)$, or the arriving server is of class $s \in S(c)$ but is not matched with c (probability $1 - P^{\mathcal{F}}_{sc}$).

This case happens with probability $\beta_c = \mu(c, S - S(c)) + \sum_{s\in S(c)} \mu(c,s)(1 - P^{\mathcal{F}}_{sc})$.

(iii) $X(n+1)_c = X(n)_c$.

Using (7.3), we obtain

$$\begin{aligned}\sum_{s\in S(c)} \mu(C-c,s)P^{\mathcal{F}}_{sc} &= \sum_{s\in S(c)} \mu_S(s)P^{\mathcal{F}}_{sc} - \sum_{s\in S(c)} \mu(c,s)P^{\mathcal{F}}_{sc}\\ &= \mu_C(c) + \varepsilon_c - \sum_{s\in S(c)} \mu(c,s)P^{\mathcal{F}}_{sc}\\ &= \mu(c, S-S(c)) + \mu(c, S(c)) + \varepsilon_c - \sum_{s\in S(c)} \mu(c,s)P^{\mathcal{F}}_{sc}\\ &= \mu(c, S-S(c)) + \sum_{s\in S(c)} \mu(c,s)(1-P^{\mathcal{F}}_{sc}) + \varepsilon_c.\end{aligned}$$

Thus, $\alpha_c = \beta_c + \varepsilon_c$. Observe that $\beta_c < \mu_C(c)$. We obtain, for $(x,y) \in \mathcal{F}$ and $c \in C_\bullet$,

$$\begin{aligned}\mathbb{E}[X(n+1)^2_c - X(n)^2_c \mid (X(n), Y(n)) = (x,y)] &= \beta_c(2x_c+1) - \alpha_c(2x_c-1)\\ &= 2\beta_c - \varepsilon_c(2x_c - 1)\\ &< 2\mu_C(c) - \varepsilon_c x_c.\end{aligned}$$

Let L be the quadratic Lyapunov function (7.1). Define $\Delta L(n) = L(X_{n+1}, Y_{n+1}) - L(X_n, Y_n)$. Set $\varepsilon = \min_{v\in C_\bullet \cup S_\bullet} \varepsilon_v > 0$. Then (the first term in the sum takes care of the vertices in $C - C_\bullet$ and $S - S_\bullet$)

$$\begin{aligned}\mathbb{E}[\Delta L(n) \mid (X_n, Y_n) = (x,y)] &< 2 + \sum_{c\in C_\bullet}(2\mu_C(c) - \varepsilon_c x_c) + \sum_{s\in S_\bullet}(2\mu_S(s) - \varepsilon_s y_s)\\ &< 2 + 2\mu_C(C_\bullet) + 2\mu_S(S_\bullet) - \varepsilon\Bigl(\sum_{c\in C_\bullet} x_c + \sum_{s\in S_\bullet} y_s\Bigr)\\ &< 6 - 2\varepsilon \sum_{c\in C} x_c.\end{aligned}$$

Fix $\delta > 0$. If $\sum_{c\in C} x_c > (6+\delta)/2\varepsilon$ then $\mathbb{E}[\Delta L(n)] < -\delta$. There are finitely many facets, so there is a finite set $A \subset \mathcal{E}$ such that

$$\mathbb{E}[\Delta L(n)] < -\delta \quad \text{for all } (x,y) \notin A. \tag{7.4}$$

By the Lyapunov–Foster theorem, see, for instance, [2, Section 5.1], the alternate matching policy is stable.

Since the ML matching policy minimizes the value of the quadratic Lyapunov function, we have a fortiori that (7.4) holds for it. Therefore, the ML policy is also stable.

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press

7.1. Conclusion

Many open questions remain. First, we do not know if the stability region is always maximal for the FIFO and Random policies (for FIFO, the result of [1] covers only the 'independent' case, i.e. $\mu = \mu_c \times \mu_S$). Numerical experiments seem to indicate that it is indeed the case. Second, for the MS and Priority policies, we know that the stability region is not always maximal, but we do not know how to compute it. Last, we would like to obtain sufficient conditions for stability, valid for all admissible policies, and which are better than those of Section 5.

References

[1] Adan, I. and Weiss, G. (2012). Exact FCFS matching rates for two infinite multitype sequences. *Operat. Res.* **60,** 475–489.

[2] Brémaud, P. (1999). *Markov chains* (Texts Appl. Math. **31**). Springer, New York.

[3] Caldentey, R., Kaplan, E. H. and Weiss, G. (2009). FCFS infinite bipartite matching of servers and customers. *Adv. Appl. Prob.* **41,** 695–730.

[4] Dai, J. and Prabhakar, B. (2000). The throughput of data switches with and without speedup. In *Proc. INFOCOM 2000*, Vol. 2, pp. 556–564.

[5] Edmonds, J. and Karp, R. (1972). Theoretical improvements in algorithmic efficiency for network flow problems. *J. Assoc. Comput. Mach.* **19,** 248–264.

[6] Ford, L. R., Jr. and Fulkerson, D. (1962). *Flows in Networks*. Princeton University Press.

[7] Gans, N., Koole, G. and Mandelbaum, A. (2003). Telephone call centers: tutorial, review, and research prospects. *Manufacturing Service Operat. Manag.* **5,** 79–141.

[8] McKeown, N. W., Mekkittikul, A., Anantharam, V. and Walrand, J. C. (1999). Achieving 100% throughput in an input-queued switch. *IEEE Trans. Commun.* **47,** 1260–1267.

[9] Seneta, E. (1981). *Nonnegative Matrices and Markov Chains*. Springer, New York.

[10] Tassiulas, L. and Ephremides, A. (1992). Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Trans. Automatic Control* **37,** 1936–1948.

[11] Visschers, J., Adan, I. and Weiss, G. (2012). A product form solution to a system with a multi-type customers and multi-type servers. *Queueing Systems* **70,** 269–298.

https://doi.org/10.1239/aap/1370870122 Published online by Cambridge University Press