Adv. Appl. Prob. **41,** 695–730 (2009)
Printed in Northern Ireland

FCFS INFINITE BIPARTITE MATCHING OF SERVERS AND CUSTOMERS

RENÉ CALDENTEY,* *New York University*
EDWARD H. KAPLAN,** *Yale School of Management, Yale School of Medicine, and Yale School of Engineering and Applied Science*
GIDEON WEISS,*** *University of Haifa*

Abstract

We consider an infinite sequence of customers of types $\mathcal{C} = \{1, 2, \ldots, I\}$ and an infinite sequence of servers of types $\mathcal{S} = \{1, 2, \ldots, J\}$, where a server of type j can serve a subset of customer types $C(j)$ and where a customer of type i can be served by a subset of server types $S(i)$. We assume that the types of customers and servers in the infinite sequences are random, independent, and identically distributed, and that customers and servers are matched according to their order in the sequence, on a first-come–first-served (FCFS) basis. We investigate this process of infinite bipartite matching. In particular, we are interested in the rate $r_{i,j}$ that customers of type i are assigned to servers of type j. We present a countable state Markov chain to describe this process, and for some previously unsolved instances, we prove ergodicity and existence of limiting rates, and calculate $r_{i,j}$.

Keywords: Service systems; first-come–first-served; infinite bipartite matching; Markov chain

2000 Mathematics Subject Classification: Primary 90B22
Secondary 60J20; 68M20

1. Introduction

We consider a service system with $\mathcal{C} = \{1, 2, \ldots, I\}$ customer types and $\mathcal{S} = \{1, 2, \ldots, J\}$ server (service) types. A server of type j can serve a subset of customer types $C(j)$ and customers of type i can be served by a subset of server types $S(i)$. This service system can be represented by a bipartite graph $\mathcal{G} = (\mathcal{C} + \mathcal{S}, G)$, where the arc (i, j) connecting nodes $i \in \mathcal{C}$ and $j \in \mathcal{S}$ belongs to G if and only if $j \in S(i)$ (or, equivalently, $i \in C(j)$). We assume that this bipartite graph is connected. As an illustrating example, Figure 1 depicts a bipartite graph with three customer types, two server types, and customer-to-server matching $S(1) = \{1\}$, $S(2) = \{1, 2\}$, $S(3) = \{2\}$ and server-to-customer matching $C(1) = \{1, 2\}$, $C(2) = \{2, 3\}$; G consists of $\{(1, 1), (2, 1), (2, 2), (3, 2)\}$. This is the so-called 'W'-model in the taxonomy for routeing topologies in service networks proposed in [13].

Given an arbitrary service system $\mathcal{G} = (\mathcal{C} + \mathcal{S}, G)$, we will construct a stochastic first-come–first-served (FCFS) infinite bipartite matching based on $\mathcal{G}$. For this, we consider all infinite sequences of customers of varying types $\mathcal{C}^\infty = \{(c^n)_{n\geq 1} : c^n \in \mathcal{C},\ n \geq 1\}$, and similarly all

Received 15 August 2008; revision received 16 April 2009.
* Postal address: IOMS Department, Stern Business School, New York University, New York.
Email address: rcaldent@stern.nyu.edu
** Postal address: Yale School of Management, Box 208200, New Haven, CT 06520-8200, USA.
Email address: edward.kaplan@yale.edu
*** Postal address: Department of Statistics, The University of Haifa, Mount Carmel 31905, Israel.
Email address: gweiss@stat.haifa.ac.il

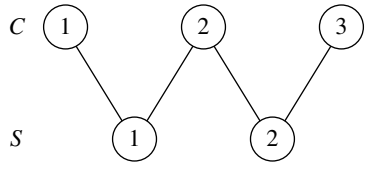

FIGURE 1: The bipartite graph for the 'W'-model.

infinite sequences of servers $\mathcal{S}^\infty = \{(s^n)_{n\geq 1} : s^n \in \mathcal{S}, n \geq 1\}$. For a given pair of sequences $(c^n, s^n)_{n\geq 1} \in \mathcal{C}^\infty \times \mathcal{S}^\infty$, the stream of customers is matched to servers (of appropriate types) according to the order in the sequence on an FCFS basis. Thus, the first server in the sequence of servers will pick the first customer in the sequence of customers which he can serve. In general, the nth server in the sequence will pick the first customer in the sequence of customers which he can serve and which has not been matched to one of the previous $n-1$ servers. The matching can also be regarded symmetrically from the point of view of customers. The first customer will be matched to the first server in the sequence of servers which can serve it, and the nth customer in the sequence will be matched to the first server in the sequence which can serve it, and which has not been matched to any of the previous $n-1$ customers. This introduces an infinite bipartite matching between the two infinite sequences, which is well defined and unique for each pair of sequences. Figure 2 shows an instance of this bipartite matching for the 'W'-model of Figure 1.

To model the stochastic nature of the FCFS infinite bipartite matching, we endow the product space $\mathcal{C}^\infty \times \mathcal{S}^\infty$ with the probability product measure P such that the nth coordinate of the pair of sequences $(c^n, s^n)_{n\geq 1} \in \mathcal{C}^\infty \times \mathcal{S}^\infty$ has probability distribution $\mathrm{P}((c^n, s^n) = (i, j)) = \alpha_i \beta_j$, for probability vectors $\alpha = (\alpha_i) \in \mathbb{R}^I$ and $\beta = (\beta_j) \in \mathbb{R}^J$. In other words, we assume that the sequences of customer and service types are random, independent, and identically distributed (i.i.d.) from the distributions given by α on $\mathcal{C}$ and β on $\mathcal{S}$.

Our purpose in this paper is to investigate these random infinite matchings. We wish to establish limiting results and in particular calculate the matching rates r_{ij} at which customers of type i are matched to servers of type j.

In Section 2 we motivate the research and survey some previous approaches. In Section 3 we make some definitions and obtain preliminary results. In Section 4 we formulate Markovian models to describe the matching. We conjecture that, under some natural assumptions, these

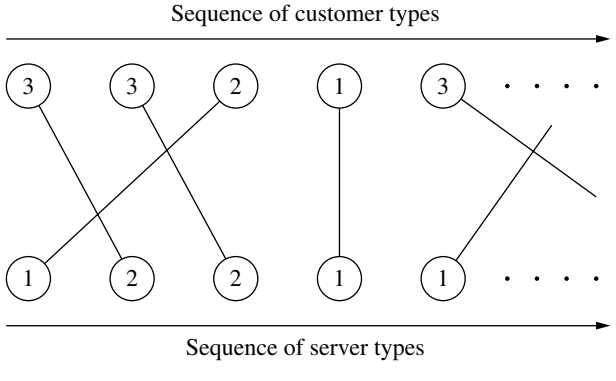

FIGURE 2: The FCFS infinite matching for the 'W'-model.

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

models are ergodic, which implies the existence of limiting matching rates. In Section 5 we analyze some simple cases, solve for the stationary distributions, and obtain the matching rates $r_{i,j}$. These lead to convergence results and formulae for the matching rates for every system with a bipartite graph which is almost complete, where each customer can be served by all except at most one server and vice versa. These results were previously unobtainable. In Section 6 we consider a more complex system, for which we are able to prove ergodicity and existence of limiting rates, by constructing an appropriate Lyapunov function. In Section 7 we give a more streamlined description of the Markov chains for systems with general bipartite graphs. In Section 8 we discuss two failed conjectured methods to calculate r_{ij}.

2. Motivation and background

Our approach to the bipartite matching of multiclass customers to multiclass servers appears to be new. The matching problem has previously been considered as a queueing model: customers arrive in independent Poisson streams of rates λ_i, and pools of servers serve them at combined rates μ_j, with the same bipartite matching of customer and server types. This is a complicated model, and its analysis under the FCFS policy seems intractable. We are aware of only one case of such a queueing system that has been analyzed in detail: the 'N'-model, which we introduce in Section 5.1, has been analyzed by Adan *et al.* [1], who obtained exact asymptotics for its stationary distributions.

Interest in the queueing model arises especially in the context of call centers, where various types of customer call, and are routed to various groups of skill-based servers [2], [13]. It is then often the case that the queueing system operates in balanced heavy traffic, where the sum of the λ_is equals the sum of the μ_js. If that is the case then we expect departures at the rates λ_i and service with no interruptions at the rates μ_j, but this is an unstable system, which is at best null recurrent. The heavy-traffic analysis of this system, under an FCFS policy, has so far been intractable. A more realistic heavy-traffic model has $\sum \lambda_i > \sum \mu_j$, and stability is achieved by reneging: customers of type i renege according to some patience time distribution F_i. This model was suggested and analyzed by Talreja and Whitt [20].

In the model of Talreja and Whitt the queues are stable, but reneging adds another level of complexity, and the stochastic model for this system still seems intractable. Whitt and Talreja considered the formal continuous and deterministic fluid model for this system. They derived the following important result. Under overload and reneging, the formal fluid system achieves global FCFS, and all classes of customers have the same waiting times. As a consequence, the rates of reneging as well as the actual carried load for each type can be calculated for the fluid model. From this, Whitt and Talreja derived equations which the matching rates $r_{i,j}$ need to satisfy. These were used to obtain the matching rates for two important cases, treelike graphs and complete graphs, as well as for graphs which are hybrids of the two. Two difficulties remain: (i) the solution of the equations for $r_{i,j}$ for general graphs may not be unique, and (ii) although simulations indicate that the stochastic matching in the queueing system converges to the rates obtained from the fluid model, proofs of convergence are currently unobtainable.

In view of the difficulty of the queueing models, it seems reasonable to look for simpler models. A first simplification which leads to our model is that there are no service times. Instead of servers being busy and providing service while they are busy at rates μ_j, there is an infinite sequence of servers, who are arriving at rates μ_j, and they are matched with the infinite sequence of customers, who arrive at rates λ_i. The matched customer and server simply leave the system, with 'service complete' at the instant of matching.

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

While this model may not be suitable for call centers, it is nevertheless useful to model several other applications. The following example arose in the Boston area public housing administration (some 20 years ago), and motivated Kaplan [16], [17]:

> Households applying for public housing are allowed to specify those housing projects in which they are willing to live; when a public housing unit becomes newly available, of those households willing to live in the associated housing project, the one that has been waiting the longest is offered the unit.

In this case households which apply for public housing and housing units which become available both arrive over time, and once a match is made, the housing service is complete. The housing unit is now occupied for a period which presumably is longer than the decision horizon. Thus, the housing units do not become available again, and correspond to our notion of an arrival stream of servers, rather than a pool of servers with service times, which become available again each time that they complete a service.

A further example also mentioned by Kaplan [6] is the matching of adoptive parents and adopted infants. Organ transplants, in particular kidney transplants, are another important example, though reneging is important in this context as many patients die while waiting for a suitable organ match [21]. Further applications of this type are the matchings of consumers and goods, matrimonial matches, etc. We believe there may also be applications in data switches and in peer-to-peer computer networks to which the model of infinite matching between two multiclass sequences is appropriate.

We make a second simplification in our model: by ignoring the arrival times, we simply consider the order of the customers and the order of the servers in the sequences of arrivals, and study the matches of the 1st, 2nd, 3rd, … customer, and the matches of the 1st, 2nd, 3rd … server. Thus, instead of Poisson arrivals of customers and servers with rates λ_i and μ_j, the only relevant parameters are the frequencies α_i and β_j. In particular, the temporal random interleaving of customers and servers is irrelevant.

To illustrate the significance of this second simplification, consider the following (single-class) model of a taxi rank. Passengers and taxis arrive at the rank in random independent Poisson streams of rates λ and μ, and are matched instantly. Thus, when some taxis are waiting at the rank, an arriving passenger will immediately depart with one of them, and when passengers are waiting at the rank, an arriving taxi will immediately depart with a passenger. Counting passengers as positive integers and taxis as negative integers, the contents of the taxi rank will form a Markovian continuous-time simple random walk process, with rate λ to move up and rate μ to move down. This however is transient for $\lambda \neq \mu$, and null recurrent for $\lambda = \mu$, and no long-term average analysis can be done. On the other hand, the matching here is trivial: the first taxi will take the first passenger, the second taxi will take the second passenger, etc.

With these two simplifications, no service periods, and no arrival times, we are at the bare bones of the FCFS multiclass infinite matching problem. If anything is tractable, this must be it. Surprisingly, the resulting problems are far from trivial, and the models which emerge seem very interesting, though extremely hard to analyze. We have made only limited progress, beyond formulation and solution of simple models. Our main contribution in this paper is the formulation of a Markovian framework in which the problem becomes meaningful. This enables us to formulate the question of existence of matching rates in the form of a conjecture about ergodicity of some Markov chains. Our main new discovery is the complete analysis of the almost-complete graph case, in Section 5, which enabled us to calculate the (previously unknown) matching rates for this case. We also prove in Section 6 the ergodicity of a more

complex model, in which the bipartite graph has each node connected to all except at most two nodes. We hope that this paper will serve as a challenge and motivation for further results.

3. Preliminaries

Definition 1. We say that the system has a balanced infinite matching if the fraction of customers of type i among the first n, who are matched by one of the first n servers, converges almost surely to α_i, and the fraction of servers of type j among the first n, who are matched by one of the first n customers, converges almost surely to β_j as $n \to \infty$.

We make two observations about balanced infinite matching. First, consider the pair of infinite sequences of customers and servers $(c^n, s^n)_{n\geq 1}$, and define $U_{m,n}$ for $0 \leq m < n$ as the *number of unmatched customers* in the FCFS matching of $(c^{m+1}, \ldots, c^n)$ and $(s^{m+1}, \ldots, s^n)$. This, of course, is also the number of unmatched servers.

Proposition 1. *$U_{m,n}$ is a stationary subadditive array of random variables, satisfying all the requirements of the subadditive ergodic theorem (cf. [9, Section 6.6, p. 361]).*

Proof. We check all four conditions (i)–(iv) of [9] for the subadditive ergodic theorem, beginning with conditions (ii)–(iv). (ii) $(U_{nk,(n+1)k})_{n\geq 1}$, the number of unmatched customers for nonoverlapping sections of length k, is a sequence of i.i.d. random variables; hence, it is clearly stationary and ergodic for each k. (iii) The distribution of the sequence $(U_{m,m+k})_{k\geq 1}$ does not depend on m. (iv) By definition, $0 \leq \mathrm{E}(U_{0,n}) \leq n$.

The main point to verify is the subadditivity (i). We need to show that $U_{0,m} + U_{m,n} \geq U_{0,n}$ for all $0 < m < n$. In words, we need to compare matching separately the first m and the remaining $n - m$, against matching the combined n, and show that the latter gets at least as many matches.

To prove this, we start from the matches obtained separately from $1, \ldots, m$ and from $m + 1, \ldots, n$, and construct the joint matching of $1, \ldots, n$. We perform an outer loop on the list of customers and servers $c^1, s^1, c^2, \ldots, c^n, s^n$ in that order, and adjust the matching. When the outer loop reaches c^k, if it is matched, we leave it alone and continue to the next in the outer loop. If c^k is unmatched, we look for a match, performing an inner loop on the servers $s^1, \ldots, s^n$ in that order. If $s^j \in S(c^k)$ and if s^j is unmatched, we add this match for a gain of one in the number of matches, and continue to the next in the outer loop. If $s^j \in S(c^k)$ and if s^j is currently matched to c^ℓ, where $\ell > k$, we cancel that match and replace it by the match of c^k and s^j, so the number of matches remains unchanged. We then continue to the next in the outer loop. In all other cases we go to the next server in the inner loop, until we reach s^n, at which point we leave c^k unmatched and go to the next in the outer loop. Clearly, this leads to the desired joint FCFS matching of $1, \ldots, n$, and the number of matches can only increase. This proves the subadditivity and completes the proof of the proposition.

We therefore have the following result.

Theorem 1. *The fraction of unmatched customers converges almost surely to*

$$\lim_{n\to\infty} \frac{U_{0,n}}{n} = \lim_{n\to\infty} \frac{\mathrm{E}(U_{0,n})}{n} = \inf_{n\geq 1} \frac{\mathrm{E}(U_{0,n})}{n} = \gamma.$$

The fraction of unmatched servers converges almost surely to the same constant.

Proof. This follows directly from the subadditive ergodic theorem.

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

We have shown that there are two possibilities. Either the system is balanced, as defined in Definition 1 (when the limit is almost surely $\gamma = 0$), or it almost surely has a fixed proportion of unmatched customers. Note, however, that at this stage we cannot say which types of customer will be unmatched, or indeed whether the makeup of the unmatched customers (and servers) obeys any rules. Also, at this stage we cannot calculate the value of γ, or say when the system is balanced with $\gamma = 0$.

Our second observation is a necessary condition for $\gamma = 0$. We introduce some notation. For $C \subseteq \mathcal{C}$, let $\alpha(C) = \sum_{i \in C} \alpha_i$, and similarly, for $S \subseteq \mathcal{S}$, let $\beta(S) = \sum_{j \in S} \beta_j$. For $C \subseteq \mathcal{C}$, let $S(C) = \bigcup_{i \in C} S(i)$, and similarly, for $S \subseteq \mathcal{S}$, let $C(S) = \bigcup_{j \in S} C(j)$.

Proposition 2. *A necessary condition for the system to be balanced, equivalently for $\gamma = 0$, is*

$$\alpha(C) \le \beta(S(C)), \qquad \beta(S) \le \alpha(C(S)), \qquad \text{for all } C \subseteq \mathcal{C},\ S \subseteq \mathcal{S}. \tag{1}$$

Proof. Assume that, for some set of customer types C, we have $\alpha(C) > \beta(S(C))$. Then clearly the fraction of servers in the sequence that can serve types in C converges almost surely to $\beta(S(C))$, and so the fraction of customers of types C in the sequence which are matched cannot exceed this, i.e. $\limsup_{n\to\infty}$ of the fraction served is less than or equal to $\beta(S(C))$. The total fraction of customers of types C in the sequence converges almost surely to $\alpha(C)$. Hence, the unmatched customers will comprise at least $\alpha(C) - \beta(S(C))$, so the fraction of unmatched customers of types C satisfies the condition that $\liminf_{n\to\infty}$ is greater than or equal to $\alpha(C) - \beta(S(C))$. Hence, the fraction of unmatched customers converges to $\gamma \ge \alpha(C) - \beta(S(C)) > 0$, and the system is not balanced. The necessity of $\beta(S) \le \alpha(C(S))$ is proved similarly.

Definition 2. For $(i, j) \in G$, we denote the fraction of i, j matches among $c^1, \dots, c^n$ and $s^1, \dots, s^n$ by $r^n_{i,j}$. If $r_{i,j} = \lim_{n\to\infty} r^n_{i,j}$ exists almost surely and is a constant, we call $r_{i,j}$ the i, j matching rate.

Our objective in this paper is to find conditions under which $r_{i,j}$ exists for all $(i, j) \in G$, and to calculate their values. We are still very far from a complete analysis of this problem.

Proposition 3. *Assume that the system is balanced and that the $r_{i,j}$ exist. Then they must satisfy the following* rate balance equations*:*

$$\sum_{i \in C(j)} r_{i,j} = \beta_j, \qquad j = 1, \dots, J,$$
$$\sum_{j \in S(i)} r_{i,j} = \alpha_i, \qquad i = 1, \dots, I. \tag{2}$$

Proof. This follows directly from the fact that the system is balanced.

Proposition 4. ([6].) *A necessary and sufficient condition for the existence of nonnegative solutions to (3) is condition (1).*

Proof. The proof is based on the *min-cut max-flow theorem* of network flows [12, Theorem 5.1, p. 11]. Consider the system graph and formulate the following max-flow problem. Direct the arcs of the graph from customer type to server type, and endow them with infinite capacity. Add node a and arcs from a to each customer type i with capacity α_i (customer arcs), and add node b with arcs from each server type j to b with capacity β_j (server arcs). We wish to find the maximal flow from a to b. Equivalently, we look for the minimal cut.

The customer arcs provide a cut with capacity 1 (as do the server arcs); hence, the maximal flow is less than or equal to 1. A nonnegative solution of (3) provides a flow of 1 and vice versa,

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

a flow of 1 defines nonnegative values of flow from i to j which solve (3). Hence, (3) has a nonnegative solution if and only if the minimal cut is 1.

A cut of capacity less than 1 will need to cut through a subset $C_1 \subset \mathcal{C}$ of the customer arcs, and through a subset $S_2 \subset \mathcal{S}$ of the server arcs, such that $\alpha(C_1) + \beta(S_2) < 1$. Let C_2 and S_1 be the complements of C_1 and S_2. The customer arcs C_1 and server arcs S_2 will be a cut if and only if there are no arcs $(i, j) \in G$ which go from C_2 to S_1. Equivalently, $C(S_1) \subseteq C_1$ and $S(C_2) \subseteq S_2$. Hence, for this cut,

$$\beta(S_1) = 1 - \beta(S_2) > \alpha(C_1) \geq \alpha(C(S_1)),$$

and (1) is violated. Conversely, if (1) is violated by $\beta(S_1) > \alpha(C(S_1))$ then customer arcs $C_1 = C(S_1)$ and server arcs S_2 form a cut with capacity less than 1.

When condition (1) fails, then by necessity $\gamma > 0$ and there is no nonnegative solution to (3). Consider then the minimum capacity cut as in the proof of Proposition 4. In this case the capacity of the cut is $\alpha(C_1) + \beta(S_2) = 1 - \gamma_0$, and $\beta(S_1) - \alpha(C_1) = \alpha(C_2) - \beta(S_2) = \gamma_0$. All the customers of S_1 are in C_1 and all the servers of C_2 are in S_2. It follows that the fractions of unmatched customers and of unmatched servers have to be at least $\gamma \geq \gamma_0$, and to include at least γ_0 out of C_2 and out of S_1. We conjecture that in this case $\gamma = \gamma_0$, and that limiting matching rates, summing up to $1 - \gamma$, will still exist.

We will not consider unbalanced cases any further, and assume from now on that condition (1) holds. This guarantees the existence of solutions to (3), but in general these solutions are not unique. Two cases that were discussed by Talreja and Whitt [20] are special.

The complete graph case. If the bipartite graph is complete (i.e. every server can serve every customer), then the FCFS infinite matching will match customer c^n with server s^n in the sequence, and clearly $r^n_{i,j}$ will converge almost surely to the matching rates $r_{i,j} = \alpha_i \beta_j$.

The tree graph case. When the bipartite system graph is a tree (i.e. it has no loops), then the solution to (3) is unique. Hence, if there exist any limiting values for $r^n_{i,j}$, they must equal that unique solution. In general, we do not yet have convergence proofs for this case.

4. A Markovian model

We introduce three Markov chains, each of which describes the process of FCFS matching. First we consider the process of matching the successive servers. Assume that the servers $s^1, \ldots, s^n$ have all been matched. In the process of matching them some customers have been considered in order, and not matched to any of them. Let X_n be the ordered list of these unmatched customers. Consider the matching of server s^{n+1}. It first examines X_n, and if there are any suitable customers in X_n, it will be matched to the first of them, which will be removed from the list. If none of the list X_n is compatible with s^{n+1}, it will examine the rest of the infinite sequence in order until it finds a match. In doing so, it will add a geometrically distributed greater than or equal to 0 number of ordered unmatched customers to the list. The resulting list is X_{n+1}, the ordered list of unmatched customers which were examined and left unmatched by $s^1, \ldots, s^n, s^{n+1}$. Here X_n is a Markov chain on the countable state space of finite words in the alphabet of the customer types $\mathcal{C}$.

Symmetrically, we can define the Markov chain Y_n, which describes the matching of successive customers. It moves on the state space of finite words in the alphabet $\mathcal{S}$, and Y_n consists of the ordered list of unmatched servers left over by customers $c^1, \ldots, c^n$.

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

More in keeping with our Definitions 1 and 2 of $U_{0,n}$ and $r^n_{i,j}$, we can also look at the matching of customers and servers, $c^1, \ldots, c^n$ with $s^1, \ldots, s^n$, and denote by $Z_n = (Z^c_n, Z^s_n)$ the pair of ordered lists of the leftover unmatched customers and servers.

These Markov chains, driven by the sequences of customers and servers, are of course closely related, as stated in the following proposition. For simplicity of notation, we denote the empty state by 0.

Proposition 5. (i) *Z^c_n is a prefix of X_n, and Z^s_n is a prefix of Y_n.*

(ii) *Let $|Z^c_n|$ and $|Z^s_n|$ denote the length of Z^c_n and Z^s_n. Then $|Z^c_n| = |Z^s_n|$.*

(iii) *There are no possible matches between Z^c_n and Z^s_n.*

(iv) *Let $X_n = 0$, $Y_n = 0$, $Z^c_n = 0$, and $Z^s_n = 0$ state that there are no unmatched customers or servers. Each one of these four statements implies all the others.*

(v) *The Markov chains are irreducible and a-periodic.*

(vi) *If one of the chains X_n, Y_n, Z_n is null recurrent or positive recurrent, so are the other two.*

Proof. (i) Let the latest of the matches of $s^1, \ldots, s^n$ be c^{n+L}, where $L \geq 0$. Then the list X_n contains L customers out of $c^1, \ldots, c^{n+L-1}$. Z^c_n consists of the intersection of the ordered list X_n with the ordered list $c^1, \ldots, c^n$. Similarly for Z^s_n.

The proofs of parts (ii) and (iii) are immediate.

(iv) All four statements say that all of $c^1, \ldots, c^n$ are matched with all of $s^1, \ldots, s^n$.

(v) Clearly, the state $Z^c_n = 0$, $Z^s_n = 0$ can be reached from every other state, and state 0 can be followed by state 0.

(vi) Follows from (iv), since if any of these irreducible chains is (positive) recurrent then 0 is a (positive) recurrent state in all of them (by (iv)), and, hence, all are (positive) recurrent.

The usefulness of these Markov chains hinges on the following crucial conjecture. While we have verified this conjecture for some special cases, we are still far from proving it in general.

Conjecture 1. *A sufficient condition for ergodicity of X, Y, Z is*

$$\alpha(C) < \beta(S(C)), \qquad \beta(S) < \alpha(C(S)), \quad \text{for all } C \subsetneq \mathcal{C},\ S \subsetneq \mathcal{S}. \tag{3}$$

Denote by A_{ij} the indicator of $(i, j) \in G$, where A is the $I \times J$ adjacency matrix of the bipartite system graph. Define the matrices L and M as follows:

$$L_{i,j} = \frac{\alpha_i A_{i,j}}{\alpha(C(j))}, \qquad M_{i,j} = \frac{\beta_j A_{i,j}}{\beta(S(i))}.$$

Here L_{ij} is the probability that a server of type j, searching a new stream of customers, will be matched with a customer of type i, and M_{ij} is the probability that a customer of type i, searching a new stream of servers, will be matched with a server of type j. Note that L' and M, and, hence, also ML' (of dimension $I \times I$) and $L'M$ (of dimension $J \times J$) are stochastic matrices (i.e. nonnegative elements with rows sum equal to 1).

Theorem 2. *If the Markov chains of the unmatched words are ergodic then $r^n_{i,j}$ converges almost surely as $n \to \infty$. Furthermore, the limiting values $r_{i,j}$ can be calculated from the steady-state distributions of these Markov chains. Let $\pi^X(\cdot)$, $\pi^Y(\cdot)$, $\pi^Z(\cdot)$, $\pi^{Z^c}(\cdot)$, and $\pi^{Z^s}(\cdot)$ be the steady-state probabilities of the Markov chains X_n, Y_n, and Z_n, and the marginal steady-state probabilities of the components of Z_n. Let $W^c(i, j)$ be the countable set of customer words*

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

in which the first customer that matches server type j is customer type i. Let $\overline{W}^c(j)$ be the countable set of customer words that do not contain any match for j. Let $W^s(i,j)$ and $\overline{W}^s(j)$ be sets of server words defined analogously. Then, for $(i,j) \in G$,

$$r_{i,j} = \beta_j \pi^X(W^c(i,j)) + \beta_j \pi^X(\overline{W}^c(j))L_{ij} \tag{4}$$

$$= \alpha_i \pi^Y(W^s(i,j)) + \alpha_i \pi^Y(\overline{W}^s(i))M_{ij} \tag{5}$$

$$= \alpha_i \pi^{Z^s}(W^s(i,j)) + \beta_j \pi^{Z^c}(W^c(i,j)) + \alpha_i \beta_j \pi^Z(\overline{W}^s(i) \cap \overline{W}^c(j)). \tag{6}$$

Proof. We discuss the convergence first. Assume that the processes X, Y, and Z are ergodic. Let $0 < n_1 < n_2 < \cdots$ be the successive times at which $X_{n_k} = Y_{n_k} = Z_{n_k} = 0$. They are regeneration points for the matching process. Let $T_k = n_k - n_{k-1}$, and let R_k be the count of (i,j) matches for $s^{n_{k-1}+1}, \ldots, s^{n_k}$ and $c^{n_{k-1}+1}, \ldots, c^{n_k}$. Then the T_k are i.i.d. and the R_k are i.i.d., and by the strong law of large numbers for renewal reward processes, almost surely we have

$$\lim_{n\to\infty} r^n_{i,j} = \lim_{k\to\infty} r^{n_k}_{i,j} = \lim_{k\to\infty} \frac{\sum_{l=1}^k R_l}{\sum_{l=1}^k T_l} = \frac{\mathrm{E}(R_1)}{\mathrm{E}(T_1)}.$$

So the limits exist.

We need the following two facts.

(i) Recall that r^n_{ij} was defined as the fraction of (i,j) matches for $s^1, \ldots, s^n$ with $c^1, \ldots, c^n$. However, the convergence proof holds also, with the same limit, if we count the fraction of (i,j) matches for $s^1, \ldots, s^n$ (with the entire $(c^m)_{m\geq 1}$ sequence), or if we count the fraction of (i,j) matches for $c^1, \ldots, c^n$ (with the entire $(s^m)_{m\geq 1}$ sequence).

(ii) Furthermore, since the $0 \leq r^n_{i,j} \leq 1$ are uniformly bounded, we have

$$r_{i,j} = \lim_{n\to\infty} r^n_{i,j} = \lim_{n\to\infty} \mathrm{E}(r^n_{i,j}).$$

We now look at $(i,j) \in G$, and prove (4)–(6).

To prove (4), consider the process X, and follow the matching of s^n. Let $I^s_n(i,j)$ be the indicator of the event that the match of s^n is an (i,j) match. This event is the union of two disjoint events: s^n is matched with one of the leftover customers, i.e. s^n is matched with $c^\ell \in X_{n-1}$, and this is an (i,j) match, or s^n finds no match in X_{n-1}, and is matched with a new customer, and the match is (i,j). Using the independence of s^n, X_{n-1}, and c^ℓ, $\ell > n + |X_{n-1}|$, we obtain

$$\begin{aligned}\mathrm{E}(I^s_n(i,j)) &= \mathrm{P}(s^n \text{ has an } (i,j) \text{ match}) \\ &= \beta_j \,\mathrm{P}(X_{n-1} \in W^c(i,j)) + \beta_j \,\mathrm{P}(X_{n-1} \in \overline{W}^c(j)) \frac{\alpha_i}{\alpha(C(j))}.\end{aligned}$$

As $n \to \infty$, this converges to the right-hand side of (4). By (i) and (ii),

$$r_{i,j} = \lim_{n\to\infty} \frac{\sum_{m=1}^n \mathrm{E}(I^s_m(i,j))}{n},$$

which completes the proof of (4).

The proof of (5) is analogous to (4).

To prove (6), we consider the process $Z = (Z^c, Z^s)$, and follow the matching of (c^n, s^n). We note that these two matches can include at most one (i,j) match, and this event can happen

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

in one of the following three mutually exclusive ways:

- $s^n = j$ and the first match for j in Z^c_{n-1} is i,
- $c^n = i$ and the first match for i in Z^s_{n-1} is j,
- $(c^n, s^n) = (i, j)$ and (Z^c_{n-1}, Z^s_{n-1}) have no match for c^n and no match for s^n.

These events are mutually exclusive because, by Proposition 5(iii), we cannot have $i \in Z^c_{n-1}$ and at the same time $j \in Z^s_{n-1}$. Define $I_n(i, j)$ as the indicator of the event that (c^n, s^n) introduce a new (i, j) match. Clearly, as $n = 1, 2, \ldots$, the indicators $I_n(i, j)$ count all the (i, j) matches, where the three mutually exclusive events correspond to an (i, j) match of c^m, s^n with $m < n$, $m > n$, or $m = n$. Then

$$\begin{aligned}\mathrm{E}(I_n(i, j)) &= \mathrm{P}((c^n, s^n) \text{ introduce an } (i, j) \text{ match}) \\ &= \beta_j \, \mathrm{P}(Z^c_{n-1} \in W^c(i, j)) + \alpha_i \, \mathrm{P}(Z^s_{n-1} \in W^s(i, j)) \\ &\quad + \alpha_i \beta_j \, \mathrm{P}(Z_{n-1} \in \overline{W}^c(j) \times \overline{W}^s(i)).\end{aligned}$$

As $n \to \infty$, this converges to the right-hand side of (6). The rest of the proof of (6) is the same as for (4).

The Markovian structure here is somewhat similar to random walks on free products of cyclic groups and to zero-automatic queues and product form recently discussed in [7], [8], [18], and [19].

5. Some simple examples

While we were unable to prove Conjecture 1 in general, we can try and verify it for specific models, and if possible we can then solve the balance equations and obtain the matching rates. This is what we do in this section. We start with two simple systems, the 'N'- and the 'W'-models. These are tree graphs, and so the matching rates are easy to obtain. However, the analysis of the Markov chains X, Y, and Z provides a proof for the almost-sure convergence to these rates. Next we consider the almost-complete graph case, in which each server type can be matched to all except at most one customer type (and vice versa). For these, we are able to prove Conjecture 1, solve the balance equations, and obtain the matching rates. These are the main new results in this paper.

5.1. Example 1: the 'N'-model

Customers are of types $\mathcal{C} = \{1, 2\}$ and servers are of types $\mathcal{S} = \{1, 2\}$, with $C(1) = \{1, 2\}$ and $C(2) = \{1\}$. The frequency of $c^n = 1$ is α, the frequency of $s^n = 1$ is β. This is the 'N'-model in the taxonomy of [13], as depicted in Figure 3. The queueing version of this system under an FCFS policy is the one analyzed in [1]. While the analysis of the queueing system is far from simple, the infinite matching problem here is quite easy. Conditions (3) for the 'N'-model are given by $\alpha + \beta > 1$.

We analyze the X_n process first. In this system only customers of type 2 may be left with no match, and so a complete description of the state of the system, as seen by each successive server, is given by the number of unmatched type-2 customers. After servers $s^1, s^2, \ldots, s^n$ have been matched, there are X_n type-2 customers which were left unmatched and are first in line for the next servers. When server s^{n+1} is of type 1, he will reduce X_n by 1. When server s^{n+1} is of type 2, he will increase X_n by a geometric greater than or equal to 0 number of type-2

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

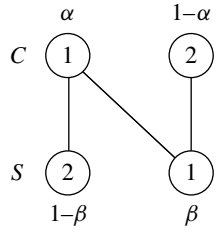

FIGURE 3: The bipartite graph and probabilities for the 'N'-model.

unmatched customers. It is not hard to see that $\{X_n\}_{n\geq 0}$ is a Markov chain starting at $X_0 = 0$ and with transition probabilities

$$\mathrm{P}(X_{n+1} = y \mid X_n = x) = \begin{cases} \beta + (1-\beta)\alpha & \text{if } x = y = 0, \\ \beta & \text{if } y = x - 1 \geq 0, \\ (1-\beta)\alpha(1-\alpha)^{y-x} & \text{if } y > x = 0 \text{ or } y \geq x \geq 1. \end{cases} \tag{7}$$

Proposition 6. *Under the condition that $\alpha + \beta > 1$, the Markov chain $\{X_n\}$ is ergodic with stationary distribution $\{\pi_x^X\}$ given by*

$$\pi_0^X = \frac{\alpha + \beta - 1}{\alpha\beta} \quad \textit{and} \quad \pi_x^X = (1-\beta)\left(\frac{1-\alpha}{\beta}\right)^x \pi_0^X, \quad x \geq 1. \tag{8}$$

Proof. As we shall see, this is a special case of Proposition 10, below. Our proof here serves as a first building block to the proof of Proposition 10.

Figure 4 depicts the states and the transitions of X_n for the 'N'-model. The balance equations for X_n are

$$\begin{aligned} \pi_0^X &= (\beta + (1-\beta)\alpha)\pi_0^X + \beta\pi_1^X, \\ \pi_x^X &= \sum_{y=0}^{x} (1-\beta)\alpha(1-\alpha)^{x-y}\pi_y^X + \beta\pi_{x+1}^X, \qquad x \geq 1. \end{aligned}$$

It is not difficult to check that the probabilities in (8) satisfy these equations. However, because this Markov chain involves jumps to distant states, it is not easy to derive the steady-state distribution directly from the balance equations.

For easier derivation, we consider a new Markov chain, $\tilde{X}_m$, which moves through states (x, a) and (x, b), $x = 0, 1, \ldots$, and is defined as follows. The visits of $\tilde{X}_m$ to the states (x, a) coincide with the sample path of X_n. The states (x, b) correspond to a server of type 2 searching for a match. In state $\tilde{X}_m = (x, a)$ assume that this corresponds to $X_n = x$. We then move

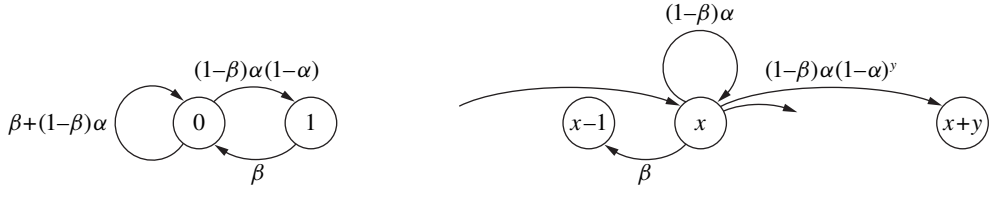

FIGURE 4: States and transitions for the X_n Markov chain of the 'N'-model.

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

to $\tilde{X}_{m+1} = (x-1, a)$ if server s^{n+1} is of type 1. If server s^{n+1} is of type 2, we move to $\tilde{X}_{m+1} = (x, b)$. In state $\tilde{X}_m = (x, b)$ we examine the next customer in line, which has not been considered for matching yet, say c^k. If it is of type 1, it will be matched with the server and we move to $\tilde{X}_{m+1} = (x, a)$, if it is of type 2 then this customer will join the other leftover unmatched type-2 customers, and we move to $\tilde{X}_{m+1} = (x+1, b)$. The advantage of the chain $\tilde{X}$ is that its transitions are only to neighboring states. The transition probabilities for the Markov chain $\tilde{X}_m$ are

$$\mathrm{P}(\tilde{X}_{m+1} = v \mid \tilde{X}_m = u) = \begin{cases} \beta & \text{if } u = (0,a) \text{ and } v = (0,a), \\ & \text{or } u = (x,a) \text{ and } v = (x-1,a),\ x > 0, \\ 1-\beta & \text{if } u = (x,a) \text{ and } v = (x,b),\ x \geq 0, \\ \alpha & \text{if } u = (x,b) \text{ and } v = (x,a),\ x \geq 0, \\ 1-\alpha & \text{if } u = (x,b) \text{ and } v = (x+1,b),\ x \geq 0. \end{cases} \tag{9}$$

Figure 5 depicts the states and the transitions of $\tilde{X}_m$ for the 'N'-model. Denote by a_x and b_x the steady-state probabilities of the states (x, a) and (x, b), respectively. The balance equations for $\tilde{X}_m$ are

$$\begin{aligned} (1-\beta)a_0 &= \alpha b_0 + \beta a_1, \\ b_0 &= (1-\beta)a_0, \\ a_x &= \alpha b_x + \beta a_{x+1}, \qquad x > 0, \\ b_x &= (1-\alpha)b_{x-1} + (1-\beta)a_x, \qquad x > 0. \end{aligned}$$

By considering the partition of states with first coordinates less than or equal to x and greater than or equal to $x+1$, we obtain the balance equation

$$(1-\alpha)b_x = \beta a_{x+1}, \qquad x \geq 0.$$

Hence,

$$b_x = \frac{\beta}{1-\alpha} a_{x+1}.$$

Substituting in the balance equation for $a_0,\ a_x, x > 0$, we obtain

$$\begin{aligned} (1-\beta)a_0 &= \alpha \frac{\beta}{1-\alpha} a_1 + \beta a_1 = \frac{\beta}{1-\alpha} a_1, \\ a_x &= \alpha \frac{\beta}{1-\alpha} a_{x+1} + \beta a_{x+1} = \frac{\beta}{1-\alpha} a_{x+1}, \qquad x > 0. \end{aligned}$$

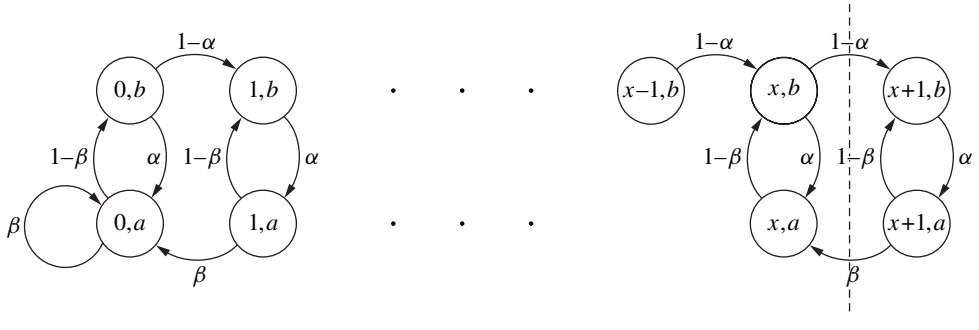

FIGURE 5: States and transitions for the $\tilde{X}_n$ Markov chain of the 'N'-model.

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

Hence,

$$a_1 = (1-\beta)\frac{1-\alpha}{\beta}a_0, \qquad a_x = \left(\frac{1-\alpha}{\beta}\right)^{x-1} a_1, \quad x > 0,$$

and so

$$a_x = (1-\beta)\left(\frac{1-\alpha}{\beta}\right)^{x} a_0, \qquad x > 0.$$

Since the visits of $\tilde{X}_m$ to states (x, a) coincide with visits of X_n to states x, we have $\pi_x^X = a_x/(\sum_{x=0}^{\infty} a_x)$ and the proposition follows.

Analogously, if we match successive customers, the leftover server words will consist of type-2 servers and the Markov chain which describes this, Y_n, will count the number of unmatched type-2 servers. Analogous to (8), the steady-state probabilities for Y_n are

$$\pi_0^Y = \frac{\alpha+\beta-1}{\alpha\beta} \quad \text{and} \quad \pi_x^Y = (1-\alpha)\left(\frac{1-\beta}{\alpha}\right)^{x} \pi_0^Y, \quad x \geq 1. \tag{10}$$

We now consider the process Z_n of the words of customers and servers that are left over after matching $c^1, \ldots, c^n$ with $s^1, \ldots, s^n$. As a rule, we write the state of Z as a pair of words, of the ordered leftover customers and ordered leftover servers. However, in the 'N'-model there will be only type-2 servers and customers left over, in equal number, so we let $Z_n = x$ denote that there are x type-2 customers and x type-2 servers left over, where $x = 0, 1, \ldots$.

Proposition 7. *Under the condition that $\alpha + \beta > 1$, the Markov chain $\{Z_n\}$ is ergodic with stationary distribution $\{\pi_x^Z\}$ given by*

$$\pi_x^Z = \frac{\alpha+\beta-1}{\alpha\beta}\left(\frac{(1-\alpha)(1-\beta)}{\alpha\beta}\right)^{x}, \qquad x \geq 0. \tag{11}$$

Proof. The transition probabilities are obtained as follows. For $x > 0$, if the next customer and server pair are of types 1, 1 then there will be two new matches (a (1,2) match and a (2,1) match) and we will have $x \to x-1$. If the pair are 1, 2 or 2, 1 then there will be one new match (a (1,2) match in the first instant, a (2,1) match in the second instant) and the new type 2 will be left over, so $x \to x$. If the pair are of types 2, 2 then both will be left over, and $x \to x+1$. In the case of $x = 0$ the only difference is that if the pair are of types 1, 1 then the two will match and the next state will still be 0. Hence,

$$\mathrm{P}(Z_{n+1} = y \mid Z_n = x) = \begin{cases} \alpha+\beta-\alpha\beta & \text{if } y = x = 0, \\ \alpha\beta & \text{if } y = x-1,\ x \geq 1, \\ \alpha(1-\beta)+\beta(1-\alpha) & \text{if } y = x,\ x \geq 1, \\ (1-\alpha)(1-\beta) & \text{if } y = x+1,\ x \geq 0. \end{cases} \tag{12}$$

It is immediate to obtain the steady-state distribution for this chain.

It is easy to check that calculation of the matching rates $r_{i,j}$ according to any of the three Markov chains X, Y, or Z as in (4)–(6) yields exactly the unique solution of (3), which is

$$r_{1,2} = 1-\beta, \qquad r_{1,1} = \alpha+\beta-1, \qquad r_{2,1} = 1-\alpha, \qquad r_{2,2} = 0. \tag{13}$$

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

5.2. Example 2: the 'W'- or 'M'-model

Customers are of types $\mathcal{C} = \{1, 2, 3\}$ and servers are of types $\mathcal{S} = \{1, 2\}$, with $C(1) = \{1, 2\}$ and $C(2) = \{2, 3\}$, and with $S(1) = \{1\}$, $S(2) = \{1, 2\}$, and $S(3) = \{2\}$; see Figure 1. The frequencies of the customers and servers are given by α_i, $i = 1, 2, 3$, β_j, $j = 1, 2$. Conditions (3) for the 'W'-model are $\beta_1 > \alpha_1$ and $\beta_2 > \alpha_3$. We will consider the processes X_n, Y_n, and Z_n.

The process of matching servers starts with X_0 empty. All type-2 customers are always matched, and because every server can serve either type-1 or type-3 customers, the unmatched customers left by $s^1, \dots, s^n$ are either all of type 1 or all of type 3. We let $X_n = (X_{n,1}, X_{n,3})$, where $X_{n,i}$ is the number of unmatched type-i customers, and where, for all $n \geq 1$, $X_{n,1}X_{n,3} = 0$. The transition probabilities of X_n are

$$\mathrm{P}(X_{n+1} = (z_1, z_3) \mid X_n = (x_1, x_3)) = \begin{cases} \alpha_1\beta_1 + \alpha_2 + \alpha_3\beta_2 & \text{if } z_1 = z_3 = x_1 = x_3 = 0, \\ \beta_1 & \text{if } z_1 = x_1 - 1,\ z_3 = x_3 = 0, \\ \beta_2 & \text{if } z_1 = x_1 = 0,\ z_3 = x_3 - 1, \\ \beta_2\alpha_1^{z_1 - x_1}(1 - \alpha_1) & \text{if } z_1 > x_1 = 0 \text{ or } z_1 \geq x_1 \geq 1 \text{ and } z_3 = x_3 = 0, \\ \beta_1\alpha_3^{z_3 - x_3}(1 - \alpha_3) & \text{if } z_1 = x_1 = 0 \text{ and } z_3 > x_3 = 0 \text{ or } z_3 \geq x_3 \geq 1. \end{cases}$$

Proposition 8. *If $\beta_1 > \alpha_1$ and $\beta_2 > \alpha_3$ then the Markov chain $\{X_n = (X_{n,1}, X_{n,3})\}$ of the 'W'-model is ergodic with stationary distribution $\pi^X = \{\pi^X_{x_1,x_3}\}$ such that $\pi^X_{x_1,x_3} = 0$ if $x_1x_3 > 0$ and*

$$\pi^X_{x_1,0} = (1 - \beta_1)\left(\frac{\alpha_1}{\beta_1}\right)^{x_1}\pi^X_{0,0}, \qquad \pi^X_{0,x_3} = (1 - \beta_2)\left(\frac{\alpha_3}{\beta_2}\right)^{x_3}\pi^X_{0,0}, \qquad x_1, x_3 \geq 1,$$

where

$$\pi^X_{0,0} = \frac{(\beta_1 - \alpha_1)(\beta_2 - \alpha_3)}{\beta_1\beta_2\alpha_2}.$$

Proof. This is also a special case of Proposition 10, below, and our proof here serves as a second building block to the proof of Proposition 10. Comparison with the 'N'-model shows that

$$\pi^X_{x,0} = \left(\frac{\alpha_1}{\beta_1}\right)^{x-1}\pi^X_{1,0}, \qquad \pi^X_{0,x} = \left(\frac{\alpha_3}{\beta_2}\right)^{x-1}\pi^X_{0,1}, \qquad x = 1, 2, \dots.$$

Consider the transitions from state $(0, 0)$ to states of the form $(x, 0)$, i.e. from no unmatched customers to unmatched customers of type 1, and back from $(1, 0)$ to $(0, 0)$, and similarly from $(0, 0)$ to $(0, x)$, with unmatched type-3 customers, and back from $(0, 1)$ to $(0, 0)$, as illustrated in Figure 6.

By equating the flow across the cuts we obtain the equations

$$\beta_1\pi^X_{1,0} = \beta_2\alpha_1\pi^X_{0,0}, \qquad \beta_2\pi^X_{0,1} = \beta_1\alpha_3\pi^X_{0,0},$$

from which we obtain

$$\pi^X_{x,0} = \beta_2\left(\frac{\alpha_1}{\beta_1}\right)^{x}\pi^X_{0,0}, \qquad \pi^X_{0,x} = \beta_1\left(\frac{\alpha_3}{\beta_2}\right)^{x}\pi^X_{0,0}, \qquad x = 1, 2, \dots.$$

The value of $\pi^X_{0,0}$ is obtained by summing all probabilities to 1.

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

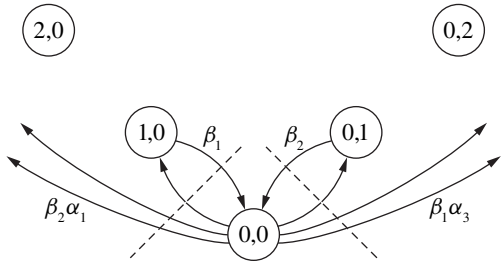

FIGURE 6: Transitions out of state (0, 0) for the X process of the 'W'-model.

The process of unmatched servers left by customers $c^1, \ldots, c^n$ is described by $Y_n = (Y_{n,1}, Y_{n,2})$, where $Y_{n,i}$ is the number of unmatched servers of type i, and $Y_{n,1}Y_{n,2} = 0$. The stationary distribution of Y_n is derived analogously to that of X_n: $\pi^Y_{y_1,y_2} = 0$ if $y_1 y_2 > 0$ and

$$\pi^Y_{y_1,0} = \alpha_3 \left(\frac{\beta_1}{1-\alpha_3}\right)^{y_1} \pi^Y_{0,0}, \qquad \pi^Y_{0,y_2} = \alpha_1 \left(\frac{\beta_2}{1-\alpha_1}\right)^{y_2} \pi^Y_{0,0}, \qquad y_1, y_2 \geq 1,$$

where

$$\pi^Y_{0,0} = \pi^X_{0,0} = \frac{(\beta_1 - \alpha_1)(\beta_2 - \alpha_3)}{\beta_1 \beta_2 \alpha_2}.$$

We now consider the process Z_n of the words of customers and servers that are left over after matching $s^1, \ldots, s^n$ with $c^1, \ldots, c^n$. As a rule, we write the state of Z as a pair of words, of the ordered leftover customers and the ordered leftover servers. However, here again in the 'W'-model this can be simplified. The possible situations are either no unmatched customers and servers, or unmatched customers of type 1 and servers of type 2, in equal number, or unmatched customers of type 3 and servers of type 1, in equal number. So we let $Z_n = (x, 0)$ denote that there are x type-1 customers and x type-2 servers left over, and we let $Z_n = (0, x)$ denote that there are x type-3 customers and x type-1 servers left over.

The transition rates for the Z process are simpler than for the X or Y process.

$$\mathrm{P}(Z_{n+1} = (u, v) \mid Z_n = (x, y)) = \begin{cases} \alpha_2 + \alpha_1\beta_1 + \alpha_3\beta_2 & \text{if } u = v = y = x = 0, \\ \alpha_1\beta_1 + (1-\alpha_1)\beta_2 & \text{if } u = x,\ v = y = 0,\ x \geq 1, \\ (1-\alpha_1)\beta_1 & \text{if } u = x-1,\ v = y = 0,\ x \geq 1, \\ \alpha_1\beta_2 & \text{if } u = x+1,\ v = y = 0,\ x \geq 1, \\ \alpha_3\beta_2 + (1-\alpha_3)\beta_1 & \text{if } v = y,\ u = x = 0,\ y \geq 1, \\ (1-\alpha_3)\beta_2 & \text{if } v = y-1,\ u = x = 0,\ y \geq 1, \\ \alpha_3\beta_1 & \text{if } v = y+1,\ u = x = 0,\ y \geq 1. \end{cases} \tag{14}$$

Proposition 9. *Under the conditions that $\beta_1 > \alpha_1$ and $\beta_2 > \alpha_3$, the Markov chain $\{Z_n\}$ is ergodic with stationary distribution $\{\pi^Z_{x,y}\}$ given by*

$$\pi^Z_{x,0} = \left(\frac{\alpha_1(1-\beta_1)}{\beta_1(1-\alpha_1)}\right)^x \pi^Z_{0,0}, \qquad \pi^Z_{0,x} = \left(\frac{\alpha_3(1-\beta_2)}{\beta_2(1-\alpha_3)}\right)^x \pi^Z_{0,0}, \qquad x \geq 1, \tag{15}$$

with

$$\pi^Z_{0,0} = \pi^Y_{0,0} = \pi^X_{0,0} = \frac{(\beta_1 - \alpha_1)(\beta_2 - \alpha_3)}{\beta_1 \beta_2 \alpha_2}.$$

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

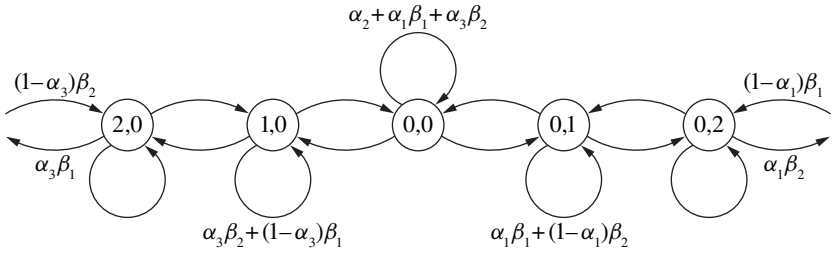

FIGURE 7: State transitions for the Z process of the 'W'-model.

Proof. Figure 7 shows the state transitions of the Z process of the 'W'-model. As can be seen, this has the transition mechanism of a simple birth-and-death process, moving on the whole integer line. Derivation of the steady-state probabilities is straightforward.

5.3. Almost-complete system graph

Examples 1 and 2 are special cases of a more general type of matching problem. Indeed, consider a model with $I+1$ customer types $\mathcal{C} = \{0, 1, \ldots, I\}$, and $J = I+1$ server types $\mathcal{S} = \{0, 1, \ldots, I\}$, such that $C(0) = \mathcal{C}$, $C(i) = \mathcal{C} \setminus \{i\}$, $i = 1, \ldots, I$ (equivalently, $S(0) = \mathcal{S}$, $S(i) = \mathcal{S} \setminus \{i\}$, $i = 1, \ldots, I$). In words, server type 0 and customer type 0 are 'universal' and can be matched with every type, while the other servers can serve all types of customer except one (for convenience, with the same index), and, similarly, all other customers can be served by all types of server except one type of server (with the same index). In terms of the system graph, it is *almost complete*: of the total $(I+1)^2$ possible arcs, the only arcs excluded are the I arcs (i, i), $i = 1, \ldots, I$.

For the almost-complete system graph model, we can define the Markov chain $\{X_n\}_{n\geq 1} \subseteq \mathbb{Z}_+^I$, where $X_{n,i}$ is the number of *unmatched* type-i customers left by servers $s^1, \ldots, s^n$. Clearly, there will never be any unmatched customer of type 0. Because each server can serve at least $I-1$ customer types among $1, \ldots, I$, it follows that $X_{n,i}X_{n,i'} = 0$ for all $1 \leq i \neq i' \leq I$.

Let $\mathbb{Z}_\perp^I \subseteq \mathbb{Z}_+^I$ be defined as follows: $(x_1, \ldots, x_I) \in \mathbb{Z}_\perp^I$ if and only if $x_i \geq 0$ for all $i = 1, \ldots I$ and $x_i x_{i'} = 0$ for all $i \neq i'$; in words, $\mathbb{Z}_\perp^I$ consists of the origin and the positive integer axes. Then the process X_n is a Markov chain on $\mathbb{Z}_\perp^I$ starting at $X_0 = 0$ and having transition probabilities

$$\mathrm{P}(X_{n+1} = y \mid X_n = x) = \begin{cases} 1 - \sum_{k=1}^I \alpha_k\beta_k & \text{if } y_k = x_k = 0 \text{ for all } k = 1, \ldots, I, \\ 1 - \beta_k & \text{if } y_k = x_k - 1 \geq 0 \text{ for some } k = 1, \ldots, I, \\ \beta_k(1-\alpha_k)\alpha_k^{y_k - x_k} & \text{if } 0 \leq x_k \leq y_k \neq 0 \text{ for some } k = 1, \ldots, I, \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 10. *Suppose that $\alpha_i + \beta_i < 1$ for all $i = 1, \ldots, I$. Then the Markov chain $\{X_n = (X_{n,1}, \ldots, X_{n,I})\}$ admits a stationary distribution $\pi^X = \{\pi_x^X : x \in \mathbb{Z}_\perp^I\}$ given by*

$$\pi_0^X = \left(1 + \sum_{i=1}^I \frac{\alpha_i\beta_i}{1-\alpha_i-\beta_i}\right)^{-1} \quad \text{and} \quad \pi_x^X = \beta_k\left(\frac{\alpha_k}{1-\beta_k}\right)^{x_k}\pi_0^X$$

for all $x \in \mathbb{R}_\perp^I$ such that $x_k \geq 1$.

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

Proof. Let us denote by $e_k \in \mathbb{Z}_+^I$ the kth unit vector in $\mathbb{Z}^I$ and by π^X the stationary distribution of $\{X_n\}$. Proceeding exactly as in the 'N'-model, we see that

$$\pi^X_{xe_k} = \pi^X_{e_k}\left(\frac{\alpha_k}{1-\beta_k}\right)^{x-1}.$$

Proceeding exactly as in the 'W'-model, we obtain the partial balance equation

$$\alpha_k\beta_k\pi^X_0 = (1-\beta_k)\pi^X_{e_k},$$

and, hence,

$$\pi^X_{xe_k} = \beta_k\left(\frac{\alpha_k}{1-\beta_k}\right)^{x}\pi^X_0.$$

Finally, equating the sum of all probabilities to 1, we obtain the expression for π^X_0.

The form of π^Y is symmetric to that of π^X, with the roles of the αs and βs reversed.

Now consider the Markov chain Z_n. Clearly, if $Z_n = (x, y) \neq 0$ then $x = y \in \mathbb{Z}_+^I$. Exactly as in the 'N'- and 'W'-models, we obtain

$$\pi^Z_{(e_kx,e_kx)} = \left(\frac{\alpha_k\beta_k}{(1-\alpha_k)(1-\beta_k)}\right)^{x}\pi^Z_0,$$

and, as always,

$$\pi^Z_0 = \pi^Y_0 = \pi^X_0 = \left(1+\sum_{i=1}^{I}\frac{\alpha_i\beta_i}{1-\alpha_i-\beta_i}\right)^{-1}.$$

We note an interesting product-form-like property here:

$$\pi^Z\{(e_kx, e_kx) \mid x > 0\} = \pi^X\{e_kx \mid x > 0\}\pi^Y\{e_kx \mid x > 0\}.$$

Note that we can have some server types j with $\beta_j = 0$ among the $I+1$ server types, or some customer types i with $\alpha_i = 0$ among the $I+1$ customer types. Indeed, in the 'W'-model we can add a third server type with $\beta_3 = 0$ whose customers are of types 1, 3, and universal server and customer types with $\alpha_0 = \beta_0 = 0$, which will make it into the almost-complete graph model with $I = 3$. The 'N'-model of course is the almost-complete graph model with $I = 1$.

5.4. Matching probabilities for the almost-complete graph model

For $I \geq 3$, the almost-complete system graph problem does not admit a unique solution to the rate balance equations (3). We now use the steady-state probabilities, π^X, π^Y, and π^Z, and (4)–(6) to calculate $r_{i,j}$.

Proposition 11. *The matching rates for the almost-complete graph model are given, for $i \neq 0$, $j \neq 0$, and $i \neq j$, by*

$$r_{i,j} = \frac{\alpha_i\beta_j[(1-\alpha_i)(1-\beta_j)-\alpha_j\beta_i]}{(1-\alpha_i-\beta_i)(1-\alpha_j-\beta_j)}\pi^Z_0, \tag{16}$$

$$r_{i,0} = \beta_0\frac{\alpha_i(1-\alpha_i)}{1-\alpha_i-\beta_i}\pi^Z_0, \qquad r_{0,j} = \alpha_0\frac{\beta_j(1-\beta_j)}{1-\alpha_j-\beta_j}\pi^Z_0, \tag{17}$$

and

$$r_{0,0} = \alpha_0\beta_0\pi^Z_0. \tag{18}$$

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

Proof. For $i \neq 0$, $j \neq 0$, and $i \neq j$, the three alternative formulae (4)–(6) read

$$\begin{aligned} r_{i,j} &= \beta_j \sum_{x=1}^{\infty} \pi^X_{xe_i} + \beta_j \sum_{x=0}^{\infty} \pi^X_{xe_j} \frac{\alpha_i}{1-\alpha_j} \\ &= \alpha_i \sum_{x=1}^{\infty} \pi^Y_{xe_j} + \alpha_i \sum_{x=0}^{\infty} \pi^Y_{xe_i} \frac{\beta_j}{1-\beta_i} \\ &= \alpha_i \sum_{x=1}^{\infty} \pi^Z_{xe_j} + \beta_j \sum_{x=1}^{\infty} \pi^Z_{xe_i} + \alpha_i \beta_j \pi^Z_0. \end{aligned}$$

Hence, using just the third equation,

$$\begin{aligned} r_{i,j} &= \alpha_i \sum_{x=1}^{\infty} \left(\frac{\alpha_j \beta_j}{(1-\alpha_j)(1-\beta_j)} \right)^x \pi^Z_0 + \beta_j \sum_{x=1}^{\infty} \left(\frac{\alpha_i \beta_i}{(1-\alpha_i)(1-\beta_i)} \right)^x \pi^Z_0 + \alpha_i \beta_j \pi^Z_0 \\ &= \alpha_i \frac{\alpha_j \beta_j}{1-\alpha_j-\beta_j} \pi^Z_0 + \beta_j \frac{\alpha_i \beta_i}{1-\alpha_i-\beta_i} \pi^Z_0 + \alpha_i \beta_j \pi^Z_0 \\ &= \frac{\alpha_i \beta_j [(1-\alpha_i)(1-\beta_j) - \alpha_j \beta_i]}{(1-\alpha_i-\beta_i)(1-\alpha_j-\beta_j)} \pi^Z_0. \end{aligned}$$

In addition, for $i \neq 0$, we obtain, from (6),

$$\begin{aligned} r_{i,0} &= \beta_0 \sum_{x=1}^{\infty} \pi^Z_{xe_i} + \beta_0 \alpha_i \pi^Z_0 \\ &= \beta_0 \sum_{x=1}^{\infty} \left(\frac{\alpha_i \beta_i}{(1-\alpha_i)(1-\beta_i)} \right)^x \pi^Z_0 + \beta_0 \alpha_i \pi^Z_0 \\ &= \beta_0 \frac{\alpha_i \beta_i}{1-\alpha_i-\beta_i} \pi^Z_0 + \beta_0 \alpha_i \pi^Z_0 \\ &= \beta_0 \frac{\alpha_i (1-\alpha_i)}{1-\alpha_i-\beta_i} \pi^Z_0, \end{aligned}$$

and similarly for $r_{0,j}$. Finally, from (6),

$$r_{0,0} = \alpha_0 \beta_0 \pi^Z_0.$$

6. Complete minus two bipartite system graphs

In this section we consider infinite matching when each of the $j = 1, \ldots, J$ server types can serve all except at most two of the customer types, and each of the $i = 1, \ldots, I$ customer types can be served by all except at most two of the server types. The Markov chains for these complete minus two systems are considerably more complicated than for the almost-complete graph systems. We analyze in some detail the simplest system of this form, the 'NN'-model. We study the Z Markov chain for this model and show that it is positive recurrent under the assumptions of Conjecture 1, thus proving the conjecture for this particular system. We prove this by using a Lyapunov function argument [3, pp. 167–173], [10, pp. 26–32]. For the general complete minus two graph case, we have so far only been able to describe the Markov chain.

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

6.1. The 'NN'-model

We consider the following 'NN'-model, described in Figure 8. Here $C(1) = \{2, 3\}$, $C(2) = \{1, 2\}$, and $C(3) = \{1\}$. This is the simplest system in which one server type has more than one excluded customer type. Servers of type 3 cannot serve any customers of type 2 or 3. Similarly, customers of type 3 cannot be served by servers of type 2 or 3.

The graph of this system is a tree, and we immediately see that the unique matching rates solving (3) are

$$r_{1,3} = \beta_3, \qquad r_{1,2} = \alpha_1 - \beta_3, \qquad r_{2,2} = 1 - \beta_1 - \alpha_1, \qquad r_{2,1} = \beta_1 - \alpha_3, \qquad r_{3,1} = \alpha_3.$$

From this we see that the necessary conditions in (1) for a balanced system and for the existence of matching rates are

$$\alpha_1 \geq \beta_3, \qquad \beta_1 \geq \alpha_3, \qquad \alpha_1 + \beta_1 \leq 1.$$

We now describe the Markovian model for this system. Clearly, the unmatched customers in X_n (or in Z_n^c) are either all of type 1 (left over by type-1 servers) or all of types 2 and 3 (left over by type-2 and type-3 servers). Consider the matching of s^{n+1} to X_n. If s^{n+1} is of type 1, it will be matched by the first customer in X_n if X_n consists of customers of types 2 and 3, and it will add a geometric greater than or equal to 0 number of customers of type 1 to X_n if X_n is empty or consists of type-1 customers. If s^{n+1} is of type 3, it will be matched by a type-1 customer if X_n consists of type-1 customers, otherwise it will add a geometric greater than or equal to 0 number of type-2 and type-3 customers, which will be mixed according to the conditional probabilities $\tilde{\alpha}_2 = \alpha_2/(\alpha_2 + \alpha_3)$ and $\tilde{\alpha}_3 = 1 - \tilde{\alpha}_2$. Finally, if s^{n+1} is of type 2 then it will be matched by a type-1 customer if X_n consists of type-1 customers or by a type-2 customer if X_n consists of type-2 and type-3 customers and there is at least one type-2 customer. Otherwise, it will add a geometric greater than or equal to 0 number of type-3 customers to X_n.

We can give a succinct description of the word of unmatched customers as follows. We write $X_n = 0$ if the word is empty, and we write $X_n = z$, $z = 1, 2, \ldots$, if it consists of $z > 0$ type-1 customers. We will write $X_n = (x, y)$, where $x = 0, 1, 2, \ldots$, $y = 0, 1, 2, \ldots$, and $x + y > 0$, to denote that X_n starts with x type-3 customers, and continues with y type-2 and type-3 customers, the first of which is of type 2 and the remaining are of unspecified type, each of them being of type 2 with probability $\tilde{\alpha}_2$ or of type 3 with probability $\tilde{\alpha}_3$, independent of all others.

Similarly, we can describe Y_n as $Y_n = z$ for empty or z type-1 server words, and $Y_n = (x, y)$ for a word starting with x type-3 servers and continuing with a mix of type-2 and type-3 servers, the first of which is type 2 and the remainder being unspecified, with probabilities $\tilde{\beta}_2 = \beta_2/(\beta_2 + \beta_3)$ and $\tilde{\beta}_3 = 1 - \tilde{\beta}_2$ for types 2 and 3.

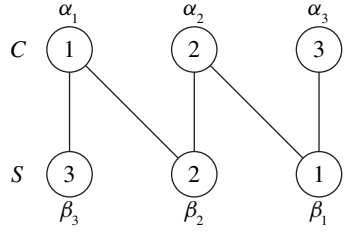

FIGURE 8: The bipartite graph for the 'NN'-model.

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

The lists of unmatched servers and customers left by the first n pairs of customers and servers, Z_n^c and Z_n^s, can be described exactly in the same way as X_n and Y_n, respectively. We note, however, that in the pair (Z_n^c, Z_n^s) one of the pair is always determined by the other. We can therefore combine them and describe Z_n as moving on $\mathbb{Z}^2$ with the following definition. For any integers $z > 0, x, y \geq 0, x + y > 0$,

$$Z_n = \begin{cases} (0,0) & \text{if both } Z_n^c \text{ and } Z_n^s \text{ are empty,} \\ (-z,0) & \text{if } Z_n^c = Z_n^s = z, \\ (x,y) & \text{if } Z_n^c = (x+y,0) \text{ and } Z_n^s = (x,y), \\ (x,-y) & \text{if } Z_n^c = (x,y) \text{ and } Z_n^s = (x+y,0). \end{cases}$$

Figure 9 depicts the states and transitions of the Markov chain Z. The figure may appear confusing at first sight, but we found it very useful for studying this system. The various states are drawn as open circles in the (x, y)-plane with 0 at the origin. Inside the circle we put the actual state, as the pair of $\begin{matrix} Z_n^c \\ Z_n^s \end{matrix}$ words. The order of servers and customers in the words Z_n^c and Z_n^s is left to right, and we use '$*$' to denote an unspecified type-2 or type-3 server or customer. Further explanations of the figure follow in the next paragraphs.

We now describe the Markov transitions for the Z chain. The process Z_n is a random walk in $\mathbb{Z}^2$ with $Z_{n+1} = Z_n + U_{n+1}$, where $\{U_n\}$ is a sequence of independent random variables and the distribution of U_{n+1} depends on Z_n in a very structured way: there are seven distributions, a distribution of U_{n+1} when Z_n is on the positive y-axis, on the negative y-axis, on the positive x-axis, in the positive quadrant, in the $x \geq 1$, $y \leq -1$ quadrant, on the negative x-axis, and at the origin. In addition, U_{n+1} is reflected at the x-axis in the sense that no transition is allowed

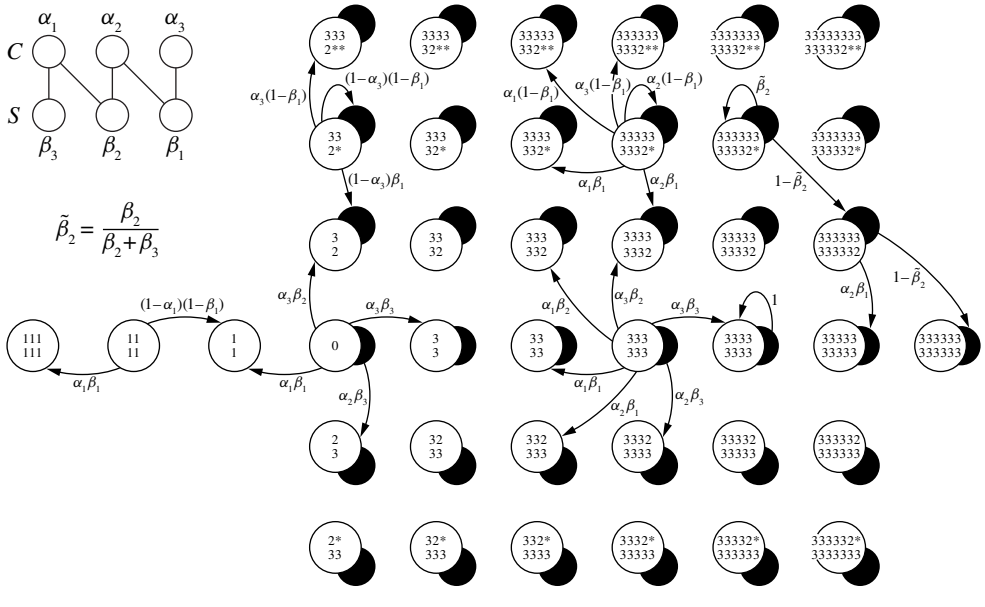

FIGURE 9: Nearest-neighbor transitions for the 'NN' model.

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

to cross the x-axis in a single step. We use the following notation to write the distribution of U_{n+1}:

$$\mathrm{P}(U_{n+1} = (i, j) \mid Z_n = (x, y)) = \mathbf{1}(x(x+i) \geq 0,\ y(y+j) \geq 0) \begin{cases} p^{+}_{i,j} & x, y \geq 1, \\ p^{-}_{i,j} & x \geq 1, y \leq -1, \\ p^{0}_{i,j} & x \geq 1, y = 0, \\ p'^{+}_{i,j} & x = 0, y \geq 1, \\ p'^{-}_{i,j} & x = 0, y \leq -1, \\ p^{00}_{i,j} & x = y = 0, \\ p^{-0}_{i,j} & x \leq -1, y = 0. \end{cases}$$

We find it convenient to use the indicator function $\mathbf{1}(\cdot)$ to capture boundary conditions (such as nonnegativities and the x-axis reflection described above) so that the remaining terms such as $p^{+}_{i,j}$ or $p^{-}_{i,j}$ are independent of x and y.

Most of the transitions are to adjacent states, i.e. $p_{i,j} = 0$ for $|i|, |j| > 1$, but if server s^{n+1} or customer c^{n+1} are of type 2 and $|y| > 1$ then $|y|$ can decrease to any value $0, \ldots, |y|$, and x will increase, with $x + |y|$ unchanged or decreased by 1.

On the positive quadrant, $x \geq 1$ and $y \geq 1$,

$$p^{+}_{0,0} = \alpha_3\beta_1 + \alpha_2\beta_2, \tag{19a}$$

$$p^{+}_{0,1} = \alpha_3(1-\beta_1), \tag{19b}$$

$$p^{+}_{-1,1} = \alpha_1(1-\beta_1), \tag{19c}$$

$$p^{+}_{-1,0} = \alpha_1\beta_1, \tag{19d}$$

$$p^{+}_{j,-j} = \alpha_2\beta_2\left(\frac{\beta_3}{1-\beta_1}\right)^{j}, \qquad j = 1, 2, \ldots, \tag{19e}$$

$$p^{+}_{j,-j-1} = \alpha_2\beta_2\frac{\beta_1}{1-\beta_1}\left(\frac{\beta_3}{1-\beta_1}\right)^{j}, \qquad j = 0, 1, 2, \ldots. \tag{19f}$$

The reflection at the x-axis is expressed by

$$\begin{aligned} \mathrm{P}(Z_{n+1} = (x+y, 0) \mid Z_n = (x, y)) &= \sum_{j=y}^{\infty} p^{+}_{j,-j} \\ &= \alpha_2\beta_2\left(\frac{\beta_3}{1-\beta_1}\right)^{y}\left(1 + \frac{\beta_3}{1-\beta_1} + \left(\frac{\beta_3}{1-\beta_1}\right)^{2} + \cdots\right) \\ &= \alpha_2(1-\beta_1)\left(\frac{\beta_3}{1-\beta_1}\right)^{y}, \end{aligned} \tag{20}$$

$$\mathrm{P}(Z_{n+1} = (x+y-1, 0) \mid Z_n = (x, y)) = \sum_{j=y-1}^{\infty} p^{+}_{j,-j-1} = \alpha_2\beta_1\left(\frac{\beta_3}{1-\beta_1}\right)^{y-1}.$$

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

Similarly, on the positive y-axis, when $x = 0$ and $y \geq 1$,

$$p'^{+}_{0,0} = \alpha_3\beta_1 + (1-\alpha_3)\beta_2, \tag{21a}$$

$$p'^{+}_{0,1} = \alpha_3(1-\beta_1), \tag{21b}$$

$$p'^{+}_{j,-j} = (1-\alpha_3)\beta_2\left(\frac{\beta_3}{1-\beta_1}\right)^j, \qquad j = 1, 2, \ldots, y-1, \tag{21c}$$

$$p'^{+}_{j,-j-1} = (1-\alpha_3)\beta_2\frac{\beta_1}{1-\beta_1}\left(\frac{\beta_3}{1-\beta_1}\right)^j, \qquad j = 0, 1, 2, \ldots, y-2, \tag{21d}$$

and the reflection at the x-axis is expressed by

$$\begin{aligned} \mathrm{P}(Z_{n+1} = (y, 0) \mid Z_n = (0, y)) &= (1-\alpha_3)(1-\beta_1)\left(\frac{\beta_3}{1-\beta_1}\right)^y, \\ \mathrm{P}(Z_{n+1} = (y-1, 0) \mid Z_n = (0, y)) &= (1-\alpha_3)\beta_1\left(\frac{\beta_3}{1-\beta_1}\right)^{y-1}. \end{aligned} \tag{22}$$

Transitions on the negative quadrant, $x \geq 1$ and $y \leq -1$, and on the negative y-axis are symmetric to $y > 0$ with an exchange of αs and βs. These transition probabilities and the remaining ones, for the x-axis, are shown in Figure 9, where we have marked the transitions according to five of the different distributions, on five states (open circles): $(-2, 0)$, $(0, 0)$, $(3, 0)$, $(0, 2)$, and $(3, 2)$. We did not draw the symmetric transitions for $y < 0$.

Instead of having transitions to nonadjacent states, we can resort to the same device which we used in the analysis of the 'N'-model in Section 5.1: we add shadow states, and instead of Z_n moving directly to Z_{n+1}, which is nonadjacent, we move from Z_n to a shadow state, and we then move along adjacent shadow states until we reach the state Z_{n+1}, where all the transitions out of shadow states are done in zero time. We represent this in Figure 9 as filled circles which are half hidden behind the open circles for all $x \geq 0$. The transitions which involve shadow nodes are as follows: for $y \geq 0$, we enter a shadow state (x, y) (filled circle) from a regular state (x, y) (open circle) with probability $\alpha_2(1-\beta_1)$ or from a regular state $(x, y+1)$ with probability $\alpha_2\beta_1$, or (if $x \geq 1$) from a shadow state $(x-1, y+1)$ with probability $\tilde{\beta}_3$. We leave a shadow state (x, y) to go to a regular state (x, y) with probability $\tilde{\beta}_2$ if $y > 0$, and with probability 1 if $y = 0$. Transitions for shadow states (x, y), $y \leq 0$, are symmetric.

The device of shadow states has some advantages: it allows for a neater Figure 9, it is better for describing the reflection at the x-axis, and it may simplify the solution of the balance equations. We have not however been able to solve the balance equations for the 'NN'-model so far.

6.2. Ergodicity of the 'NN'-model

We now show that Conjecture 1 holds for the 'NN'-model. We assume that condition (3) holds, that is,

$$\alpha_1 > \beta_3, \qquad \beta_1 > \alpha_3, \qquad \alpha_1 + \beta_1 < 1. \tag{23}$$

We then prove that the Z process of the 'NN'-model is ergodic. We consider Z as a Markov chain with states $(x, y) \in \mathbb{Z}^2$, as described in the previous section. We discuss first the states and the transitions for $x < 0$, where there are $-x$ unmatched type-1 customers and servers. The chain Z performs a simple random walk on the negative x-axis, moving towards 0 with probability $(1-\alpha_1)(1-\beta_1)$ and away from 0 with probability $\alpha_1\beta_1$. The chain will return to 0 in finite expected time if and only if $\alpha_1 + \beta_1 < 1$. We now ignore these states and consider the chain restricted to (x, y), $x \geq 0$.

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

We show that the Z process is ergodic by constructing a Lyapunov function and verifying the Foster Lyapunov criterion. We use the same type of Lyapunov function which was developed in the monograph of Fayolle *et al.* [10] for the homogeneous random walk on $\mathbb{Z}_+^2$. In fact, our proof is very close to the one given in [10, Section 3.3]. The random walks considered by Fayolle *et al.* have downward jumps that do not exceed 1 in any direction, so they have no 'reflection' at the axes, which makes them more homogeneous. This enabled them to find necessary and sufficient conditions for ergodicity. In our case there is reflection at the x-axis, so we need to work a little harder to adapt their method to our model. However, we already have necessary conditions for ergodicity, so we only need to prove the sufficient condition.

6.2.1. *Calculation of the first vector field.* Let the coordinates of $Z_{n+1} - Z_n$ be $(\theta_1(x, y), \theta_2(x, y))$, which depend on the location of $Z_n = (x, y)$, as described above. We now calculate the first vector field, $M(x, y)$, of the process Z_n. For this, we use the fact that $(\theta_1(x, y), \theta_2(x, y))$ depends on x only in so far as $x = 0$ or $x \geq 1$, and so we can write $M(x, y)$ as functions of y exclusively if we distinguish these two cases:

$$M(x, y) = \mathrm{E}(Z_{n+1} - Z_n \mid Z_n = (x, y)) = (\mathrm{E}(\theta_1(x, y)), \mathrm{E}(\theta_2(x, y)))$$
$$= \begin{cases} (M_1(y), M_2(y)), & x \geq 1, \\ (M_1'(y), M_2'(y)), & x = 0. \end{cases}$$

We start with calculation of $M_1(y)$ in the positive quadrant, $x \geq 1$ and $y \geq 1$. We use the transition probabilities (19a)–(19f) and (20):

$$\begin{aligned} M_1(y) &= \mathrm{E}(\theta_1(x, y)) \\ &= -1(p^+_{-1,0} + p^+_{-1,1}) + \sum_{j=1}^{y-1} j p^+_{j,-j} + y \sum_{j=y}^{\infty} p^+_{j,-j} \\ &\quad + \sum_{j=1}^{y-2} j p^+_{j,-j-1} + (y-1) \sum_{j=y-1}^{\infty} p^+_{j,-j-1} \\ &= -1(p^+_{-1,0} + p^+_{-1,1}) + \sum_{k=1}^{y} \sum_{j=k}^{\infty} p^+_{j,-j} + \sum_{k=1}^{y-1} \sum_{j=k}^{\infty} p^+_{j,-j-1} \\ &= -\alpha_1 + \sum_{k=1}^{y} \sum_{j=k}^{\infty} \alpha_2 \beta_2 \left(\frac{\beta_3}{1-\beta_1}\right)^j + \sum_{k=1}^{y-1} \sum_{j=k}^{\infty} \alpha_2 \beta_2 \frac{\beta_1}{1-\beta_1} \left(\frac{\beta_3}{1-\beta_1}\right)^j \\ &= -\alpha_1 + \sum_{k=1}^{y} \alpha_2 (1-\beta_1) \left(\frac{\beta_3}{1-\beta_1}\right)^k + \sum_{k=1}^{y-1} \alpha_2 \beta_1 \left(\frac{\beta_3}{1-\beta_1}\right)^k \\ &= -\alpha_1 + \alpha_2 \beta_3 + \sum_{k=1}^{y-1} \alpha_2 (\beta_1 + \beta_3) \left(\frac{\beta_3}{1-\beta_1}\right)^k \\ &= -\alpha_1 + \alpha_2 \beta_3 + \alpha_2 (\beta_1 + \beta_3) \frac{\beta_3}{\beta_2} \left(1 - \left(\frac{\beta_3}{1-\beta_1}\right)^{y-1}\right) \\ &= \frac{-\alpha_1 \beta_2 + \alpha_2 \beta_3}{\beta_2} - \frac{1-\beta_2}{\beta_2} \alpha_2 \beta_3 \left(\frac{\beta_3}{1-\beta_1}\right)^{y-1}. \end{aligned}$$

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

The calculation of $M_2(y)$ is similar, and so are the calculations of $M_1'(y)$ and $M_2'(y)$, from the transition probabilities (21a)–(21d) and (22). In summary, we obtain the first vector field, for $y \geq 1$:

$$M_1(y) = \frac{\beta_3\alpha_2 - \alpha_1\beta_2}{\beta_2} - \frac{1-\beta_2}{\beta_2}\alpha_2\beta_3\left(\frac{\beta_3}{1-\beta_1}\right)^{y-1}, \qquad \nearrow y, \tag{24a}$$

$$M_2(y) = \frac{(1-\beta_1)(\beta_2-\alpha_2)}{\beta_2} + \frac{1-\beta_2}{\beta_2}\alpha_2\beta_3\left(\frac{\beta_3}{1-\beta_1}\right)^{y-1}, \qquad \searrow y, \tag{24b}$$

$$M_1'(y) = \frac{(1-\alpha_3)\beta_3}{\beta_2} - \frac{1-\beta_2}{\beta_2}(1-\alpha_3)\beta_3\left(\frac{\beta_3}{1-\beta_1}\right)^{y-1}, \qquad \nearrow y, \tag{24c}$$

$$M_2'(y) = -\frac{(1-\beta_1)(\beta_1+\beta_3-\alpha_3)}{\beta_2} + \frac{1-\beta_2}{\beta_2}(1-\alpha_3)\beta_3\left(\frac{\beta_3}{1-\beta_1}\right)^{y-1}, \qquad \searrow y. \tag{24d}$$

As indicated, $M_1(y)$ and $M_1'(y)$ are increasing functions of y, while $M_2(y)$ and $M_2'(y)$ are decreasing functions of y. This is because the expected length of the nonadjacent diagonal moves is greater as y increases. All four quantities are bounded above and below for all y by the values at $y = 1$ or $y = \infty$ as follows:

$$\beta_3\alpha_2 - \alpha_1 \leq M_1(y) \leq \frac{\beta_3\alpha_2 - \alpha_1\beta_2}{\beta_2}, \tag{25a}$$

$$\frac{(1-\beta_1)(\beta_2-\alpha_2)}{\beta_2} \leq M_2(y) \leq 1 - \beta_1 - \alpha_2 - \alpha_2\beta_3, \tag{25b}$$

$$(1-\alpha_3)\beta_3 \leq M_1'(y) \leq \frac{(1-\alpha_3)\beta_3}{\beta_2}, \tag{25c}$$

$$-\frac{(1-\beta_1)(\beta_1+\beta_3-\alpha_3)}{\beta_2} \leq M_2'(y) \leq -(\beta_1 - \alpha_3 + \beta_3(1-\alpha_3)). \tag{25d}$$

We can repeat the calculations for $y \leq -1$, but this is unnecessary, as by the symmetry we can use the above formulae with αs and βs interchanged, y replaced by $|y|$, and the signs for M_2' and M_2 reversed.

For the following calculations, we need to keep track of the signs of the vector field coordinates. Let us first consider the case in which $y \geq 1$. Under condition (23), it is easy to verify that $M_1'(y) > 0$ and $M_2'(y) < 0$. On the other hand, the signs of $M_1(y)$ and $M_2(y)$ depend on the relationship between α_1 and β_2. If $\alpha_2 \leq \beta_2$ then it follows that $M_1(y) < 0$ and $M_2(y) > 0$. If $\alpha_2 > \beta_2$ then $M_1(1) < 0$, but it may change sign as y increases depending on the sign of $M_1(\infty)$. Also, if $\alpha_2 > \beta_2$ then $M_2(\infty) < 0$, but it may change sign as y decreases depending on the sign of $M_2(1)$. From the fact that $M_1(y) + M_2(y) = \alpha_3 - \beta_1$, we conclude that $M_1(y) + M_2(y) < 0$ for all y. In particular, this implies that there exist $1 \leq y_1 < \infty$ and $1 < y_2 \leq \infty$ with $y_1 < y_2$ such that $M_2(y)$ changes sign at y_1 and $M_1(y)$ changes sign at y_2 (if $y_1 = 1$ then $M_2(y) < 0$ for all y, and does not change sign, if $y_2 = \infty$ then $M_1(y) < 0$ for all y, and does not change sign).

Similar conclusions hold for $y \leq -1$, with the condition on α_2 and β_2 reversed, where we have $M_1(y) - M_2(y) < 0$, and y_3 and y_4 such that $-\infty \leq y_4 < -1$, $-\infty < y_3 \leq -1$, and $y_4 < y_3$ replacing y_1 and y_2. Table 1 summarizes the behavior of the vector field: the top half of the table corresponds to $y \geq 1$ with $1 \leq y_1 < y_2 \leq \infty$, the bottom half corresponds to $y \leq -1$ with $-1 \geq y_3 > y_4 \geq -\infty$. (See also Figure 10 in Section 6.2.3 for a drawing of the vector field.)

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

TABLE 1: Behavior of the first vector field of Z for the 'NN'-model.

	$\alpha_2 < \beta_2$			$\alpha_2 > \beta_2$		
$y \geq 1$	$M_1'(y) > 0 \nearrow$	$M_2'(y) < 0 \searrow$		$M_1'(y) > 0 \nearrow$	$M_2'(y) < 0 \searrow$	
	$M_1(y) < 0 \nearrow$	$M_2(y) > 0 \searrow$	$[1, y_1)$	$M_1(y) < 0$	$M_2(y) > 0$	
			$[y_1, y_2)$	$M_1(y) < 0 \nearrow$	$M_2(y) < 0 \searrow$	
			$[y_2, \infty)$	$M_1(y) > 0$	$M_2(y) < 0$	
$y \leq -1$	$M_1'(y) > 0 \nearrow$	$M_2'(y) > 0 \searrow$		$M_1'(y) > 0 \nearrow$	$M_2'(y) > 0 \searrow$	
	$(y_3, -1]$	$M_1(y) < 0$	$M_2(y) < 0$			
	$(y_4, y_3]$	$M_1(y) < 0 \nearrow$	$M_2(y) > 0 \searrow$	$M_1(y) < 0 \nearrow$	$M_2(y) < 0 \searrow$	
	$(-\infty, y_4]$	$M_1(y) > 0$	$M_2(y) > 0$			

6.2.2. *The Lyapunov function.* We will use a Lyapunov function of the form suggested by Fayolle *et al.* [10, pp. 39–56]. For $x \geq 0$ and $y \geq 0$, we take

$$\begin{aligned} Q(x, y) &= ux^2 + vy^2 + wxy, \\ f(x, y) &= Q^{1/2}(x, y), \\ \Delta f(x, y) &= Q^{1/2}(x + \theta_1(x, y), y + \theta_2(x, y)) - Q^{1/2}(x, y), \end{aligned}$$

where $u > 0$, $v > 0$, and $4uv > w^2$.

It is then shown in [10, Lemma 3.3.3] that, as $x^2 + y^2 \to \infty$,

$$\mathrm{E}(\Delta f(x, y)) = \frac{x[2u\,\mathrm{E}(\theta_1(x, y)) + w\,\mathrm{E}(\theta_2(x, y))] + y[w\,\mathrm{E}(\theta_1(x, y)) + 2v\,\mathrm{E}(\theta_2(x, y))]}{2f(x, y)} + o(1). \tag{26}$$

In other words, $\mathrm{E}(\Delta f(x, y))$ is almost constant along rays of constant y/x, if $x^2 + y^2$ is large. Strictly speaking, in [10, Lemma 3.3.3, pp. 42–43], $\theta_1(x, y)$ and $\theta_2(x, y)$ have just three distributions which are independent of x and y. In our case, there are seven possible distributions and some of them depend on y. However, because all of these distributions are stochastically bounded by a single random variable with finite expectation, the lemma still holds.

To verify the Foster Lyapunov criterion, we require that

$$\begin{aligned} 2u\,\mathrm{E}(\theta_1(x, y)) + w\,\mathrm{E}(\theta_2(x, y)) &< -\varepsilon, \qquad x \geq 1,\ y \geq 1, \\ w\,\mathrm{E}(\theta_1(x, y)) + 2v\,\mathrm{E}(\theta_2(x, y)) &< -\varepsilon, \qquad x \geq 0,\ y \geq 1, \end{aligned} \tag{27}$$

for some $\varepsilon > 0$. We will deal with $y < 0$ and with $y = 0$ later.

We will now choose appropriate u, v, and w so as to satisfy (27). In the next paragraphs, all the inequalities hold by the following argument: if $0 < A < B$ then $0 < (A - t)/(B - t) < 1$ and it is decreasing in t for $0 < t < A$.

Assume first that $\alpha_2 < \beta_2$. In this case (see Table 1) $M_1(y) < 0$, $M_2(y) > 0$, $M_1'(y) > 0$, and $M_2'(y) < 0$. The conditions on u, v, and w are as follows:

- for $x \geq 1$, to obtain $2u\,\mathrm{E}(\theta_1(x, y)) + w\,\mathrm{E}(\theta_2(x, y)) < 0$, we need

$$\frac{2u}{w} > \frac{M_2(y)}{-M_1(y)} \quad \text{or} \quad w < 0,$$

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

- for $x \geq 0$, to obtain $w\,\mathrm{E}(\theta_1(x, y)) + 2v\,\mathrm{E}(\theta_2(x, y)) < 0$, we need

$$\frac{2v}{w} < \frac{-M_1(y)}{M_2(y)}, \qquad \frac{2v}{w} > \frac{M_1'(\infty)}{-M_2'(\infty)}, \quad \text{and} \quad w > 0.$$

We see by (24a)–(24d), (25a)–(25d), and Table 1 that

$$\begin{aligned} K_1 &:= \frac{M_1'(\infty)}{-M_2'(\infty)} \\ &= \frac{M_1'(y)}{-M_2'(y)} + o(1) \quad \text{(as } y \text{ becomes large)} \\ &= \frac{(1-\alpha_3)\beta_3}{(1-\beta_1)(\beta_1+\beta_3-\alpha_3)} \\ &< 1, \end{aligned}$$

and also that

$$\begin{aligned} \frac{-M_1(\infty)}{M_2(\infty)} &\geq \frac{-M_1(y)}{M_2(y)} \\ &\geq \frac{-M_1(1)}{M_2(1)} \\ &= \frac{\alpha_1(1-\beta_1) + (1-\beta_2)\alpha_2\beta_3 - (1-\alpha_3)\beta_3}{\alpha_1(1-\beta_1) + (1-\beta_2)\alpha_2\beta_3 - (1-\beta_1)(\beta_1+\beta_3-\alpha_3)} \\ &= K_2 \\ &> 1. \end{aligned} \tag{28}$$

Hence, to satisfy (27), we need to impose $K_2 > 2v/w > K_1$ and $w/2u < K_2$ (this automatically includes the requirement that $w > 0$, since $K_1 > 0$ and $u, v > 0$).

Now assume that $\alpha_2 > \beta_2$, with y_1 and y_2 defined as above. Then

$$\begin{aligned} M_1(y) < 0 \quad &\text{and} \quad M_2(y) > 0 \quad \text{when} \quad 0 < y < y_1, \\ M_1(y) < 0 \quad &\text{and} \quad M_2(y) < 0 \quad \text{when} \quad y_1 < y < y_2, \\ M_1(y) > 0 \quad &\text{and} \quad M_2(y) < 0 \quad \text{when} \quad y_2 < y < \infty. \end{aligned}$$

In fact, there are four cases depending on whether $y_1 = 1$ or $y_1 > 1$ and on whether $y_2 < \infty$ or $y_2 = \infty$. We will consider the case in which $1 < y_1 < y_2 < \infty$, where (y_1, y_2) includes some integer values. All other cases are somewhat simpler.

For the range of values $1 \leq y \leq \lfloor y_1 \rfloor$, the calculations in (28) remain valid. Rewriting this for $1 \leq y \leq \lfloor y_1 \rfloor$, we have

$$\frac{-M_1(\lfloor y_1 \rfloor)}{M_2(\lfloor y_1 \rfloor)} \geq \frac{-M_1(y)}{M_2(y)} \geq \frac{-M_1(1)}{M_2(1)} =: K_2 > 1.$$

For the nonempty range of integer values of y such that $y_1 \leq y \leq y_2$, we have both $M_1(y) \leq 0$ and $M_2(y) \leq 0$, and because their sum is negative, at least one of $M_1(y)$ or $M_2(y)$ is not equal to 0 for every $y_1 \leq y \leq y_2$. We can therefore find a $\delta > 0$ such that

$$\min(M_1(y), M_2(y)) < -\delta \quad \text{for all } y_1 \leq y \leq y_2.$$

Therefore, for any $u, v, w > 0$, we can find ε small enough to satisfy (27).

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

For the remaining range of values, $\lceil y_2 \rceil \le y \le \infty$, we have $M_1(y) > 0$ and $M_2(y) < 0$, and so the conditions on u, v, and w are as follows:

- for $x \ge 1$, to obtain $2u\,\mathrm{E}(\theta_1(x, y)) + w\,\mathrm{E}(\theta_2(x, y)) < 0$, we need

$$\frac{2u}{w} < \frac{-M_2(y)}{M_1(y)} \quad \text{and} \quad w > 0,$$

- for $x \ge 0$, to obtain $w\,\mathrm{E}(\theta_1(x, y)) + 2v\,\mathrm{E}(\theta_2(x, y)) < 0$, we need

$$\frac{2v}{w} > \frac{M_1(y)}{-M_2(y)}, \qquad \frac{2v}{w} > \frac{M_1'(\infty)}{-M_2'(\infty)}, \quad \text{or} \quad w < 0.$$

We now see by (23), (25a)–(25d), and (28) that

$$\begin{aligned} 1 &> K_1 \\ &> K_3 := \frac{M_1(\infty)}{-M_2(\infty)} \\ &= \frac{\alpha_1(1-\beta_1) - (1-\alpha_3)\beta_3}{\alpha_1(1-\beta_1) - (1-\beta_1)(\beta_1+\beta_3-\alpha_3)} \\ &\ge \frac{M_1(y)}{-M_2(y)} \\ &\ge \frac{M_1(\lceil y_2 \rceil)}{-M_2(\lceil y_2 \rceil)}. \end{aligned} \tag{29}$$

So we need to impose $2v/w > K_3$ and $w/2u > K_3$ (automatically, $w > 0$).

6.2.3. *Verification of the Foster Lyapunov condition.* We now assume without loss of generality that $\alpha_2 < \beta_2$ (else we switch the roles of the αs and βs). Using the definitions of K_1, K_2, and K_3 above, we let

$$K_1^+ = \frac{(1-\alpha_3)\beta_3}{(1-\beta_1)(\beta_1+\beta_3-\alpha_3)},$$

$$K_2^+ = \frac{\alpha_1(1-\beta_1) + (1-\beta_2)\alpha_2\beta_3 - (1-\alpha_3)\beta_3}{\alpha_1(1-\beta_1) + (1-\beta_2)\alpha_2\beta_3 - (1-\beta_1)(\beta_1+\beta_3-\alpha_3)},$$

$$K_1^- = \frac{(1-\beta_3)\alpha_3}{(1-\alpha_1)(\alpha_1+\alpha_3-\beta_3)},$$

$$K_2^- = \frac{\beta_1(1-\alpha_1) + (1-\alpha_2)\beta_2\alpha_3 - (1-\beta_3)\alpha_3}{\beta_1(1-\alpha_1) + (1-\alpha_2)\beta_2\alpha_3 - (1-\alpha_1)\alpha_1+\alpha_3-\beta_3)}$$

if $M_2(-1) < 0$, and undefined otherwise,

$$K_3^- = \frac{(1-\beta_3)\alpha_3 - \beta_1(1-\alpha_1)}{(1-\alpha_1)(\alpha_1+\alpha_3-\beta_3) - \beta_1(1-\alpha_1)} \quad \text{if } M_1(-\infty) > 0, \text{ and equal to 0 otherwise.}$$

We will use the Lyapunov function

$$f(x, y) = \begin{cases} (u_1x^2 + v_1y^2 + w_1xy)^{1/2}, & y \ge 0, \\ (u_2x^2 + v_2y^2 - w_2xy)^{1/2}, & y < 0, \end{cases}$$

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

with the following choice of parameters:

$$u_1 = u_2 = 1, \qquad w_2 = K_1^- + \max(0, K_3^-), \qquad w_1 = \min\{1, 2 - w_2\},$$
$$v_1 = \tfrac{1}{2}w_1, \qquad v_2 = \tfrac{1}{2}w_2.$$

It is straightforward to check that

$$K_1^+ < \frac{2v_1}{w_1} < K_2^+, \qquad \frac{w_1}{2u_1} < K_2^+, \qquad 4u_1v_1 > w_1^2,$$
$$K_3^-, K_1^- < \frac{2v_2}{w_2} < K_2^-, \qquad K_3^- < \frac{w_2}{2u_2} < K_2^-, \qquad 4u_2v_2 > w_2^2,$$

where the requirement related to K_2^- is only imposed if K_2^- is well defined.

With these choices, we can find a ε such that

$$2u\,\mathrm{E}(\theta_x) + w\,\mathrm{E}(\theta_y) < -\varepsilon, \qquad w\,\mathrm{E}(\theta_x) + 2v\,\mathrm{E}(\theta_y) < -\varepsilon,$$

for all $x = 0, 1, 2, \ldots$ and all $y = \pm 1, \pm 2, \ldots$. We note that as a result, by (26),

$$\mathrm{E}(\Delta f(x, y)) < -\frac{x + |y|}{f(x, y)}\varepsilon < -\min\left(\frac{1}{u_1 + v_1 + w_1}, \frac{1}{u_2 + v_2 + w_2}\right)\varepsilon.$$

The final step is to check that this Lyapunov function works at $y = 0$ for large x. We note that, for large x and fixed A and B,

$$\begin{aligned} f(x + A, B) &= (u(x + A)^2 + w(x + A)B + vB^2)^{1/2} \\ &= (x + A)\left(u + \frac{wB}{x + A} + v\frac{B^2}{(x + A)^2}\right)^{1/2} \\ &= \sqrt{u}\left((x + A) + \frac{w}{2u}B\right) + o(1) \quad \text{as } x \text{ becomes large.} \end{aligned}$$

Hence, we need to evaluate

$$\begin{aligned} \mathrm{E}(\Delta f(x, 0)) &= \alpha_1\beta_1 - \alpha_1\beta_2 - \alpha_2\beta_1 + \alpha_3\beta_3 + \frac{w_1}{2}(1 - \alpha_2)\beta_2 + \frac{w_2}{2}\alpha_2(1 - \beta_2) + o(1) \\ &= \tfrac{1}{2}(\alpha_3 + \beta_3 - \alpha_1 - \beta_1) - \tfrac{1}{2}(1 - w_1)(1 - \alpha_2)\beta_2 \\ &\quad - \tfrac{1}{2}(1 - w_2)\alpha_2(1 - \beta_2) + o(1). \end{aligned}$$

Recall that $\alpha_3 < \beta_1$, $\beta_3 < \alpha_1$, and $\alpha_2 < \beta_2$. Since we have chosen

$$w_1 \le 1 \quad \text{and} \quad w_1 + w_2 \le 2,$$

we will have $\mathrm{E}(\Delta f(x, 0)) < -h < 0$ for some $h > 0$ and all large enough x.

This completes the proof that the 'NN'-model, under the required assumptions in (3), is ergodic.

Example. We take $\alpha = (\frac{7}{16}, \frac{3}{16}, \frac{3}{8})$ and $\beta = (\frac{5}{12}, \frac{1}{4}, \frac{1}{3})$. The Lyapunov function is

$$f(x, y) = \begin{cases} \left(x^2 + \frac{130}{207}xy + \frac{65}{207}\right)^{1/2}, & y \ge 0, \\ \left(x^2 - \frac{284}{207}xy + \frac{142}{207}\right)^{1/2}, & y < 0. \end{cases}$$

Figure 10 shows the first vector field and the contours of the Lyapunov function.

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

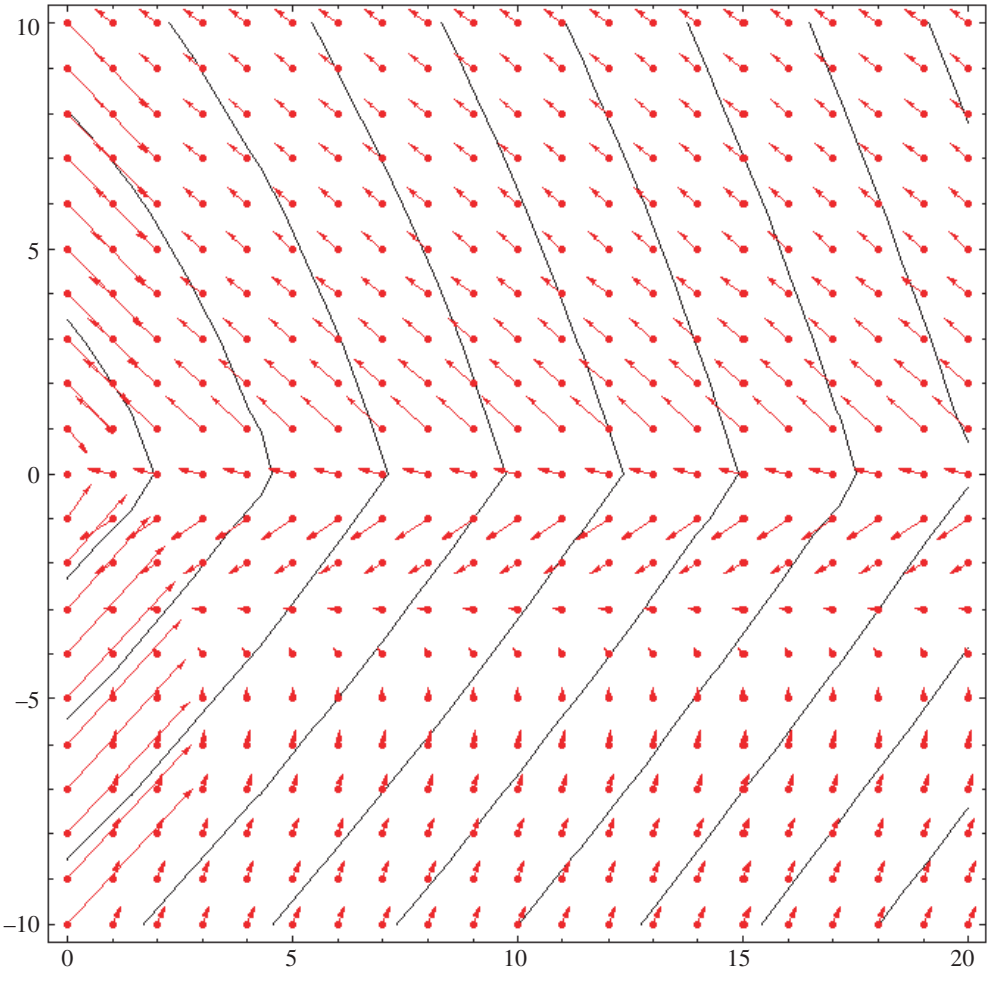

FIGURE 10: Vector field and Lyapunov function contours for the 'NN'-model example.

6.3. The general system with a complete minus two bipartite graph

Systems with complete minus two bipartite graphs are best described by partitioning all customer and server types into *incompatibility chains*. Each chain consists of alternating server and customer types (all different), so that each two successive types are incompatible. The first and last in the chain are compatible, with all the types not in the chain, and they are either incompatible, in which case we call the chain circular (it can be started at any member type), or they are incompatible only with their neighbor in the chain, in which case we call the chain linear. For the 'NN'-model, these incompatibility chains are (1, 1), which is linear, and (2, 3, 3, 2), which is circular.

Conditions (3) of Conjecture 1 for these systems are

$$\begin{aligned} \alpha_i + \beta_j &< 1 \quad \text{if } S(i) = \mathcal{S} \setminus j \text{ or } C(j) = \mathcal{C} \setminus i, \\ \alpha_i &< 1 - \beta_{j_1} - \beta_{j_2} \quad \text{if } S(i) = \mathcal{S} \setminus \{j_1, j_2\}, \\ \beta_j &< 1 - \alpha_{i_1} - \alpha_{i_2} \quad \text{if } C(j) = \mathcal{C} \setminus \{i_1, i_2\}. \end{aligned} \tag{30}$$

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

No further conditions are necessary, since $\beta(S(C)) = \alpha(C(S)) = 1$ for all subsets S and C which contain more than two types.

We now describe the Markov chains X, Y, and Z. The unmatched customers in the X process are all of one type or of two types, and the unmatched servers in the Y process are all of one type or all of two types. In the Z process, if Z^c consists of ℓ customers of two different types then Z^s must consist of ℓ servers of the unique type which is incompatible with both, and vice versa. The state space of Z can be described as a bundle of one-dimensional rays and of two-dimensional nonnegative quadrants, similar to the 'NN'-model. The rays correspond to types which have only a single incompatible type (those that start or end linear incompatibility chains) or to circular incompatibility chains of length 4, of the form (i_1, j_1, i_2, j_2), where i_1 and j_2 are incompatible. In the latter case, we can define the state of Z simply by the length of the Z^c and Z^s words, which is the total number of unmatched customers of type i_1 or i_2, and the total number of unmatched servers of type j_1 or j_2. The nonnegative quadrants are similar to those used in the description of the 'NN'-model. All the rays and quadrants intersect at 0. The difficulty is that the quadrants now intersect along both axes, some with reflection and some with crossover, as we now explain.

We denote by $\mathbb{Z}^+_{(i_1,i_2),j}$ the states in which we have unmatched servers of type j and unmatched customers of types i_1 and i_2, where the earliest unmatched customer is of type i_1. The state $(x, y) \in \mathbb{Z}^+_{(i_1,i_2),j}$, $x, y \geq 1$, consists of $x + y$ unmatched servers of type j and $x + y$ unmatched customers of types i_1 and i_2, where the first (earliest) x unmatched customers are of type i_1 and the last y unmatched customers begin with a customer of type i_2 followed by $y - 1$ customers whose type is unspecified. The unspecified customers are of types i_1 and i_2 with independent probabilities $\tilde{\alpha}_{i_1} = \alpha_{i_1}/(\alpha_{i_1} + \alpha_{i_2})$ and $\tilde{\alpha}_{i_2} = 1 - \tilde{\alpha}_{i_1}$. We define $\mathbb{Z}^+_{i,(j_1,j_2)}$ similarly.

Now consider part of an incompatibility chain of length 5, $(\ldots, i_1, j_1, i_2, j_2, i_3, \ldots)$. To avoid confusion, we will denote these types in that order as $(\ldots, c_2, s_2, c_1, s_3, c_3, \ldots)$. We will examine the quadrant $\mathbb{Z}^+_{c_1,(s_2,s_3)}$. From $Z_n = (x, y) \in \mathbb{Z}^+_{c_1,(s_2,s_3)}$, $x, y \geq 1$, we can move in one step to a state Z_{n+1} consisting of $x + y$ or $x + y - 1$ customers of type c_1 and servers of type s_2, on the boundary ray with the quadrant $\mathbb{Z}^+_{(c_1,c_2),s_2}$. This is exactly what happens in the 'NN'-model, where the two quadrants are $\mathbb{Z}^+_{3,(3,2)}$ and $\mathbb{Z}^+_{(3,2),3}$. However, in addition to this the quadrant $\mathbb{Z}^+_{c_1,(s_2,s_3)}$ borders on the quadrant $\mathbb{Z}^+_{c_1,(s_3,s_2)}$, and here there is the possibility of a crossover. From $Z_n = (1, y) \in \mathbb{Z}^+_{c_1,(s_2,s_3)}$, where $Z^s_n = s_2, s_3, *, \ldots, *$ and '$*, \ldots, *$' denotes servers of unspecified type s_2 or s_3, we can move to a state $Z^s_{n+1} = s_3, \ldots, s_3, s_2, *, \ldots, *$, which means that $Z_{n+1} \in \mathbb{Z}^+_{c_1,(s_3,s_2)}$.

The quadrants in which c_1 is involved are then, in order,

$$\mathbb{Z}^+_{(c_3,c_1),s_3} \Leftrightarrow \mathbb{Z}^+_{(c_1,c_3),s_3} \quad || \quad \mathbb{Z}^+_{c_1,(s_3,s_2)} \Leftrightarrow \mathbb{Z}^+_{c_1,(s_2,s_3)} \quad || \quad \mathbb{Z}^+_{(c_1,c_2),s_2} \Leftrightarrow \mathbb{Z}^+_{(c_2,c_1),s_2},$$

where '||' signifies a reflecting boundary and '$\Leftrightarrow$' signifies a crossover boundary. The first and last quadrants in turn have a reflecting boundary with quadrants involving the next members of the incompatibility chain.

Fayolle *et al.* [10] also characterized the ergodicity and transience of homogeneous random walks on complexes of two-dimensional quadrants. However, by assuming that all jumps downwards are of size no more than one in each direction, they did not allow for reflections and crossovers between neighboring quadrants. Thus, to show ergodicity for our complete minus two bipartite graph systems will require some significant modification of their technique. We do not pursue this any further in this paper.

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

7. Further specification of the Markov chains for systems with general bipartite graphs

We defined the processes X, Y, and Z as Markov chains moving in the countable state space of words in the alphabet of customer types for X, server types for Y, and pairs of words for Z. Following our more streamlined description of X, Y, and Z for almost-complete and complete minus two bipartite system graphs in Sections 5 and 6, we now give similar descriptions for systems with general bipartite graphs.

Consider the ordered finite sequence of unmatched customers left over after servers $s^1, \dots, s^n$ have been matched. We describe this sequence by pairs of the form $((i_1, x_1), (i_2, x_2), \dots, (i_\ell, x_\ell), \dots, (i_L, x_L))$, where $i_1, \dots, i_L$ are the types of the unmatched customers and $x_1, \dots, x_L$ are positive integers. Let $x_0^+ = 0$, let $x_\ell^+ = \sum_{j=1}^{\ell} x_j$, and let $\alpha_\ell^+ = \sum_{j=1}^{\ell} \alpha_{i_j}$. Then the earliest appearance of a customer of type i_ℓ is in place $x_{\ell-1}^+ + 1$. Following the first appearance of i_ℓ there will be $x_\ell - 1$ unmatched customers of types $i_1, \dots, i_\ell$. We leave the actual types of these unspecified in the definition of the state. Every customer of type $i_1, \dots, i_\ell$ which has joined the unmatched customers after the first customer of type i_ℓ has joined will not leave before the first customer of type $i_1, \dots, i_\ell$ leaves. Therefore, the $x_\ell - 1$ customers following the first appearance of i_ℓ will be of type i_k with probability $\alpha_{i_k}/\alpha_\ell^+$.

The transitions from $X_n = ((i_1, x_1), \dots, (i_L, x_L))$ to X_{n+1} are as follows: if server s^{n+1} is incompatible with all of $i_1, \dots, i_L$, there will be a geometric number of added customers of types incompatible with s^{n+1}. Conditional on the type of s^{n+1}, $X_{n+1} = ((i_1, x_1), \dots, (i_L, x_L + y), (i_{L+1}, x_{L+1}), \dots, (i_{L+K}, x_{L+K}))$. All the customer types $i_{L+1}, \dots, i_{L+K}$ are different from $i_1, \dots, i_L$, and incompatible with s^{n+1}. Here y and $x_{L+1} - 1, \dots, x_{L+K} - 1$ are all geometric greater than or equal to 0 random variables with appropriate distributions. The probability of this transition is

$$\alpha(C(s^{n+1}))(\alpha_L^+)^y \prod_{j=L+1}^{L+k} \alpha_{i_j} (\alpha_{i_j}^+)^{x_j - 1}.$$

If s^{n+1} is compatible with some of $i_1, \dots, i_L$ then it will be matched. Assume that i_ℓ is the first customer type in the list $(i_1, \dots, i_L)$ which is compatible with s^{n+1}. Then the first customer of type i_ℓ, in position x_ℓ^+ in the sequence, will be matched with s^{n+1} and removed from the sequence. To write the proper new state, the position of the first appearance of i_ℓ needs to be updated. If the second customer of a type i_ℓ in X_n is in position $x_{k-1}^+ + y$, where $k \geq \ell$ and $1 < y \leq x_k$, then the new state will be $X_{n+1} = ((i_1, x_1), \dots, (i_{\ell-1}, x_{\ell-1} + x_\ell - 1), \dots, (i_k, y - 1), (i_\ell, x_k - y + 1), \dots)$ (with an obvious modification if $k = \ell$), and if there are no more type-i_ℓ customers in the sequence, the new state will be $X_{n+1} = ((i_1, x_1), \dots, (i_{\ell-1}, x_{\ell-1} + x_\ell - 1), (i_{\ell+1}, x_{\ell+1}), \dots)$. Conditional on s^{n+1} being incompatible with $i_1, \dots, i_{\ell-1}$ and compatible with i_ℓ, the probabilities of these transitions are

$$\frac{\alpha_{i_\ell}}{\alpha_k^+} \left(1 - \frac{\alpha_{i_\ell}}{\alpha_k^+}\right)^{y-2} \prod_{j=\ell}^{k-1} \left(1 - \frac{\alpha_{i_\ell}}{\alpha_j^+}\right)^{x_j - 1}$$

and

$$\prod_{j=\ell}^{L} \left(1 - \frac{\alpha_{i_\ell}}{\alpha_j^+}\right)^{x_j - 1}.$$

The process Y is of course defined analogously. In the process $Z = (Z^c, Z^s)$, Z^c has the same state space as X and Z^s has the same state space as Y, with the added restriction that all

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

customer types in Z^c are incompatible with all server types in Z^s, and the lengths of the two are equal. It is not clear that Z has any advantage over X or Y in the general case.

It is not currently clear how useful these Markov chains may be in answering our main questions: how to calculate the $r_{i,j}$ and how to prove convergence. As we mentioned before (Section 4), the work of Mairesse *et al.* [7], [8], [18], [19] may be relevant here.

8. Some attempts to calculate the matching rates

Equations (3) require joint probabilities when the marginals α and β are given. There are in general many solutions to these equations. If we can obtain the stationary distributions of X, Y, and Z, we can calculate the $r_{i,j}$. However, finding these stationary probabilities is hard and seems like overkill. It would be better to find a direct way of calculating the matching rates $r_{i,j}$. In this case we could prove the convergence by showing that X, Y, and Z are ergodic without solving for the stationary distribution.

For completeness, we report here on two failed attempts: the algorithms are somewhat plausible, and yield the correct results for some models, but not for all.

8.1. An algorithm of Caldentey and Kaplan

The following algorithm was suggested by Caldentey and Kaplan [6] and verified by simulation for various models. Our presentation here is somewhat different from theirs.

Recall that for the complete bipartite system graph the matching rates are simply $r_{i,j} = \lim_{n\to\infty} r^n_{i,j} = \alpha_i\beta_j$, which corresponds to an independent joint distribution with the marginals α and β. We let $F^c_{i,j}$ denote these probabilities for the complete graph. Recall the definitions of the matrices A, L, and M in Section 4.

We wish to construct a matrix $F_{i,j}$ such that $F_{i,j} = 0$ wherever $A_{i,j} = 0$, and such that $F_{i,j}$ has the same marginals as $F^c_{i,j}$, namely $\sum_j F_{i,j} = \alpha_i$ and $\sum_i F_{i,j} = \beta_j$, i.e. $F_{i,j}$ is a solution to the equations in (3).

We start with an initial matrix F^0:

$$F^0_{i,j} = F^c_{i,j} A_{i,j}.$$

This has the correct pattern of zeros, but the marginals are wrong by the amounts

$$f^0_i = \alpha_i - \sum_j F^0_{i,j}, \qquad g^0_j = \beta_j - \sum_i F^0_{i,j}.$$

We let

$$\boldsymbol{F} = \sum_{i,j} F^0_{i,j} = \sum_i \alpha_i \beta(S(i)) = \sum_j \beta_j \alpha(C(j))$$

so that

$$\sum_i f^0_i = \sum_j g^0_j = 1 - \boldsymbol{F}.$$

We now recursively define the sequence of row I-vectors f^k, row J-vectors g^k, and matrices F^k as follows. For convenience, we define, with $f^{-1} = 0$,

$$f^{k+1} = g^k L', \qquad g^{k+1} = f^k M, \tag{31}$$

$$F^{k+1}_{i,j} = F^k_{i,j} + (-1)^{k-1} f^{k-1}_i M_{i,j} + (-1)^k g^k_j L_{i,j}. \tag{32}$$

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

Analogously, we define, with $G^0 = F^0$ and $g^{-1} = 0$, the sequence

$$G^{k+1}_{i,j} = G^k_{i,j} + (-1)^{k-1} g^{k-1}_j L_{i,j} + (-1)^k f^k_i M_{i,j}.$$

The idea here is as follows. Consider F^0. If we look at the row conditional probabilities $F^0_{i,j} / \sum_{j'} F^0_{i,j'}$, they equal M_{ij}, the probability that a customer of type i searching a sequence of new servers will be matched to server type j. If we look at the column conditional probabilities $F^0_{i,j} / \sum_{i'} F^0_{i',j}$, they equal L_{ij}, the probability that a server of type j searching a sequence of new customers will be matched to customer type i.

Here $F^0_{i,j}$ does not sum to 1, and has the shortfalls of f^0_i in row i and of g^0_j in column j. If we distribute the row shortfall f^0_i according to the row conditional probabilities $M_{i,j}$ and add this to row i, the row will be distributed like $M_{i,j}$ and it will sum to α_i, but the column sums will not equal β_j. This will give us F^1. If we distribute the column shortfall g^0_j according to $L_{i,j}$ and add this to column j, the column will be distributed like $L_{i,j}$ and it will sum β_j, but the row sums will not equal α_i. This will give us G^1. In either case the sum of all elements of the matrices is now 1.

If we add both compensations, we will overcompensate. The amount of overcompensation is f^1_i in row i and g^1_j in column j. We again distribute these according to $M_{i,j}$ and $L_{i,j}$, respectively.

After k steps,

$$(-1)^k f^k_i = \alpha_i - \sum_j \left(F^0_{i,j} + \sum_{\ell=0}^{k-1} ((-1)^\ell f^\ell_i M_{i,j} + (-1)^\ell g^\ell_j L_{i,j}) \right),$$

$$(-1)^k g^k_j = \beta_j - \sum_i \left(F^0_{i,j} + \sum_{\ell=0}^{k-1} ((-1)^\ell f^\ell_i M_{i,j} + (-1)^\ell g^\ell_j L_{i,j}) \right).$$

As $k \to \infty$, the quantities $f^k_i M_{i,j}$ and $g^k_j L_{i,j}$ converge to the same value of $(1 - \boldsymbol{F})\alpha_i \beta_j / \boldsymbol{F}$, and doing half the correction (adding just $(-1)^k g^k_j L_{i,j}$ to obtain $F^{k+1}_{i,j}$ or $(-1)^k f^k_i M_{i,j}$ to obtain $G^{k+1}_{i,j}$) gives almost the same result and converges to $F_{i,j}$. We now state this result without proof.

Proposition 12. *The sequences $F^k_{i,j}$ and $G^k_{i,j}$ converge as $k \to \infty$ to the same limit $F_{i,j}$, which is a solution of (3).*

8.2. A quasi-independence model for the matching rates

Kaplan considered a pipeline analogy to the matching problem, with the pipeline topology laid out as the bipartite graph $\mathcal{G}$ connecting feasible customer and server types. Fluid flow through a pipe is a function of the pressure exerted at each end of the pipe, and using the physical principles of fluid flow [15], Kaplan sought to develop a system of equations for the $I + J$ pressures that would produce the matched flow. For the limiting case of a non-Newtonian 'power-law' fluid, Kaplan found that the flow system reduced to a model known as quasi-independence.

The quasi-independence model appears in statistics in the analysis of two-dimensional contingency tables. A natural assumption to test is that the row factors are independent of the column factors, in other words, if α_i, $i = 1, \ldots, I$, and β_j, $j = 1, \ldots, J$, are the true row and column probability distributions of the two factors, then independence would imply that the true cell probabilities for the i, j cell would be $\alpha_i \beta_j$, and this hypothesis is then tested from

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

the observed cell frequencies in the sample. However, it is sometimes the case that some of the cells are missing, either because they could not be observed, or because they are impossible.

The assumption of quasi-independence is then as follows: for the given $\alpha_i,\ i = 1, \ldots, I$, and $\beta_j,\ j = 1, \ldots, J$, and for cells which are present in the table $(i, j) \in G$, we are looking for $x_i,\ i = 1, \ldots, I$, and $y_j,\ j = 1, \ldots, J$, such that

$$\sum_{\{j\,:\,(i,j)\in G\}} x_i y_j = \alpha_i, \quad i = 1, \ldots, I, \qquad \sum_{\{i\,:\,(i,j)\in G\}} x_i y_j = \beta_j, \quad j = 1, \ldots, J,$$
$$\text{and} \quad x_i, y_j \geq 0.$$

The concept of quasi-independence, the conditions for existence of such solutions, an algorithm to calculate the values of the x_i and y_j, and the statistical implications are derived and discussed in [4], [5], [11], and [14].

Quite clearly, the values of $h_{i,j} = x_i y_j,\ (i, j) \in G$, are a solution to the equations in (3). On the basis of the pipeline analogy, Kaplan hoped that these values would also prove to be the correct matching frequencies for our FCFS infinite matching model. Unfortunately, this turned out to be a false hope, as we will see in Section 8.3.

8.3. Counter examples to the algorithm of Caldentey and Kaplan and to the quasi-independence model

We consider the following two examples. Oddly, while the first example contradicts the quasi-independence model, it does agree with the Caldentey–Kaplan algorithm, and while the second example seems to agree with the quasi-independence model, it does not seem to agree with the Caldentey–Kaplan algorithm.

Example 1. We consider the network of the almost-complete graph of three customer and three server types. Here

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \qquad \alpha = \begin{pmatrix} 0.32 & 0.46 & 0.22 \end{pmatrix}, \qquad \beta = \begin{pmatrix} 0.29 & 0.24 & 0.47 \end{pmatrix}.$$

The correct matching frequencies can be calculated from (16). They agree with the values obtained from the Caldentey–Kaplan algorithm:

$$(r_{i,j}) = (F_{i,j}) = \begin{pmatrix} 0 & 0.1298 & 0.1902 \\ 0.1802 & 0 & 0.2798 \\ 0.1098 & 0.1102 & 0 \end{pmatrix},$$

while the quasi-independence model gives the values

$$(h_{i,j}) = \begin{pmatrix} 0 & 0.1285 & 0.1915 \\ 0.1815 & 0 & 0.2785 \\ 0.1085 & 0.1115 & 0 \end{pmatrix}.$$

Example 2. We consider the network of the almost-complete graph of three customer and three server types, and add another customer type which is served by only type-2 servers. Here

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \qquad \alpha = \begin{pmatrix} 0.4 & 0.2 & 0.2 & 0.2 \end{pmatrix}, \qquad \beta = \begin{pmatrix} 0.3 & 0.6 & 0.1 \end{pmatrix}.$$

https://doi.org/10.1239/aap/1253281061 Published online by Cambridge University Press

We do not know how to calculate the limiting matching rates for this case (nor have we proved convergence). Instead, we have simulated this example, by generating 10^6 pseudo random customer server pairs. The matching frequencies of the simulation were

$$(\hat{r}_{i,j}) = \begin{pmatrix} 0 & 0.398\,993 & 0 \\ 0.136\,61 & 0.063\,863 & 0 \\ 0 & 0.137\,29 & 0.062\,913 \\ 0.163\,824 & 0 & 0.036\,501 \end{pmatrix},$$

which seems to agree with the values given by the quasi-independence model, i.e.

$$(h_{i,j}) = \begin{pmatrix} 0 & 0.4000 & 0 \\ 0.1361 & 0.0639 & 0 \\ 0 & 0.1361 & 0.0639 \\ 0.1639 & 0 & 0.0361 \end{pmatrix},$$

but does not agree with the values from the Caldentey–Kaplan algorithm:

$$(F_{i,j}) = \begin{pmatrix} 0 & 0.4000 & 0 \\ 0.1444 & 0.0556 & 0 \\ 0 & 0.1444 & 0.0556 \\ 0.1556 & 0 & 0.0444 \end{pmatrix}.$$

Acknowledgements

We are grateful to Ward Whitt, Rishi Talreja, and Bob Foley for discussions of this problem. We wish to thank an anonymous referee for pointing out a major error in a previous version, and for many useful suggestions.

References

[1] Adan, I., Foley, R. D. and McDonald, D. R. (2007). Exact asymptotics of the stationary distribution of a Markov chain: a production model. EURANDOM Report 2008-036, Eindhoven, Netherlands. Available at http://www.eurandom.nl/EURANDOMreports.htm.

[2] Aksin, Z., Armony, M. and Mehrotra, V. (2007). The modern call-center: a multi-disciplinary perspective on operations management research. *Production Operat. Manag.* **16,** 665–688.

[3] Brémaud, P. (1999). *Markov Chains*. Springer, New York.

[4] Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *J. R. Statist. Soc. B* **25,** 220–233.

[5] Bishop, Y. M. M. and Fienberg, S. E. (1969). Incomplete two-dimensional contingency tables. *Biometrics* **25,** 119–128.

[6] Caldentey, R. A. and Kaplan, E. H. (2002). A heavy traffic approximation for queues with restricted customer-service matchings. Unpublished manuscript.

[7] Dao-Thi, T.-H. and Mairesse, J. (2006). Zero-automatic networks. In *Proc. VALUETOOLS* (Pisa, Italy), ACM, New York.

[8] Dao-Thi, T.-H. and Mairesse, J. (2007). Zero-automatic queues and product form. *Adv. Appl. Prob.* **39,** 502–536.

[9] Durrett, R. (1995). *Probability: Theory and Examples*, 2nd edn. Duxbury Press, Belmont, CA.

[10] Fayolle, G., Malyshev, V. A. and Menshikov, M. V. (1995). *Topics in the Constructive Theory of Countable Markov Chains*. Cambridge University Press.

[11] Fienberg, S. E. (1970). Quasi-independence and maximum likelihood estimation in incomplete contingency tables. *J Amer. Statist. Assoc.* **65,** 1610–1615.

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

[13] Garnett, O. and Mandelbaum, A. (2000). An introduction to skill-based routing and its operational complexities. Unpublished manuscript. Available at http://iew3.technion.ac.il/serveng/Lectures/SBR.pdf.

[14] Goodman, L. (1968). The analysis of cross classified data: independence, quasi-independence, and interactions in contingency tables with and without missing entries. *J. Amer. Statist. Assoc.* **63,** 1091–1131.
[15] Hwang, N. H. C. (1981). *Fundamentals of Hydraulic Engineering Systems*. Prentice Hall, Englewood Cliffs, NJ.
[16] Kaplan, E. H. (1984). Managing the demand for public housing. ORC Tech. Rep. 183, MIT.
[17] Kaplan, E. H. (1988). A public housing queue with reneging and task-specific servers. *Decision Sci.* **19,** 383–391.
[18] Mairesse, J. (2005). Random walks on groups and monoids with a Markovian harmonic measure. *Electron. J. Prob.* **10,** 1417–1441.
[19] Mairesse, J. and Mathéus, F. (2007). Random walks on free products of cyclic groups. *J. London Math. Soc.* **75,** 47–66. Appendix available at http://arxiv.org/abs/math/0509208.
[20] Talreja, R. and Whitt, W. (2007). Fluid models for overloaded multiclass many-service queueing systems with FCFS routeing. *Manag. Sci.* **54,** 1513–1527.
[21] Zenios, S. A. (1999). Modeling the transplant waiting list: a queueing model with reneging. *Queueing Systems* **31,** 239–251.