FLUID AND DIFFUSION APPROXIMATIONS OF PROBABILISTIC MATCHING SYSTEMS

Burak Büke and Hanyi Chen

School of Mathematics, The University of Edinburgh

King's Buildings, Edinburgh, EH9 3FD, UK

E-mail:B.Buke@ed.ac.uk, H.Chen-29@sms.ed.ac.uk

Abstract

This paper focuses on probabilistic matching systems where two classes of users arrive at the system to match with users from the other class. The users are selective and the matchings occur probabilistically. Recently, Markov chain models are proposed to analyze these systems, however an exact analysis of these models to completely characterize the performance is not possible due to the probabilistic matching structure. In this work, we propose approximation methods based on fluid and diffusion limits using different scalings. We analyze the basic properties of these approximations and show that some performance measures are insensitive to the matching probability agreeing with the existing results. We also perform numerical experiments with our approximations to gain insight into probabilistic matching systems.

1 Introduction

The Internet has provided the society a new medium to carry out business and personal transactions. In this work, our goal is to provide tractable methods to analyze probabilistic matching systems introduced in Büke and Chen [3] to study the web portals that serve as a meeting point for suppliers and customers of a specific product or service. The examples of such systems include employment

and rental portals, matrimonial and dating websites and general purpose classified advertisement websites.

The users of a probabilistic matching system can be classified into two groups as customers (e.g. employers) and suppliers (e.g. employees). Customers arrive at the system according to a stochastic process. When a customer arrives at the system, she searches the list of suppliers to see if there is anybody selling the product (or the service) she demands. If she finds suppliers with suitable products, she buys a product choosing one uniformly at random and both the customer and the supplier leave the system together. If there are no suitable products available, then she posts an advertisement on the system indicating her demand and waits until a supplier with suitable product arrives at the system. The suppliers also exhibit a similar behavior.

The double-ended queue introduced in Kashyap [12] is a precursor for the matching systems and considers the queueing process of taxis and customers at a taxi stop. Taxis and customers arrive at the stop according to independent Poisson processes and if a taxi (customer) arrives when there are no customers (taxis) waiting at the stop, she waits until a customer (taxi) arrives. Recently, there has been a growing interest to study matching systems which can be perceived as generalizations of double-ended queues. For these systems each class of users has several subclasses, which we refer to as types, and these types determine whether users from different classes can match or not. Drawing an analogy between these matching systems and the taxi problem of Kashyap [12], in these systems there are different types of taxis each of which serves to a set of neighborhoods and a taxi accepts a customer, in other words matches with the customer, if and only if she is going to a neighborhood served by the type that the taxi belongs to. For these models, once the types of users are known, the matchings occur deterministically and the main goal is to devise policies to decide on which users should be matched with each other. Specifically, Adan and Weiss [1] and Caldentey et al. [5] focus on systems where multiple types of customers arrive at the system to be matched with multiple types of suppliers (servers). Caldentey et al. [5] focus on the ergodicity of these systems and analyze the matching rates between different types of users. Adan and Weiss [1] consider the first-come-first-serve matching policy and derive a product-form expression for the stationary probabilities when they exist. Bušić et al. [4] focus on computational complexity for

evaluating whether a matching system is stable under a given policy. In a recent work, Gurvich and Ward [10] study a similawr system where different types of users can match with each other and develop asymptotically optimal policies to minimize the holding cost.

The key feature which differentiates probabilistic matching systems from the conventional matching systems in the literature is the probabilistic nature of the matching process. When a customer arrives at the system, she checks the products of all the suppliers and may find each product suitable with a given probability independent of the others. Hence, with positive probability she may not find any suitable product, even if there are several suppliers offering a product in the system. To make this argument more concrete consider an employment portal as an example. An employer arriving at the employment portal first scans through the resumés of all the employees in the system and she may hire each potential employee with a given probability. There is a positive probability that she may not find any of the existing candidates suitable, in which case she posts a job advert and waits in the system until a suitable candidate arrives. Hence, unlike the double-ended queues, users from different classes can co-exist in the system when the matchings are probabilistic, which makes it essential to model the queueing system as a two dimensional stochastic process. Büke and Chen [3] study the effects of the matching probability on the performance of these systems using an exact analysis, show that if uncontrolled these systems are unstable and suggest admission control policies to stabilize these systems.

The probabilistic matching behavior complicates the analysis of these systems and renders a complete exact analysis intractable. Hence, in this work we propose approximation methods based on fluid and diffusion limits under two different scalings. Under our first scaling, we only scale time and space and keep the matching probability constant to obtain the limiting processes. We show that under this scaling both fluid and diffusion limits do not depend on the matching probability, which implies that the users from at most one class accumulate in the system and the probability of a user finding a match upon arrival approaches either zero or one.

To provide tools which address the matching probability explicitly, we propose a second scaling that also handles the abandonment of impatient users and scales the matching probability and the abandonment rate along with the time and space. The resulting fluid and diffusion limits under

this scaling involve differential equations which are not tractable analytically in the general case, although we can derive an analytical formula for the fluid limit when there are no abandonments. Büke and Chen [3] show that some performance measures, such as the difference between the average queue lenghths of different classes, are insensitive to the matching probability under certain control policies. Despite not imposing any control policy, similar to the results in [3] we show that the difference between queue lengths for different classes is also insensitive to the matching probability in the fluid limit.

In addition, we analyze the asymptotic behaviour of the fluid limits. We first compare the fluid limits under both scalings, i.e., limits with and without scaling the matching probability, and show that when the abandonment rate is zero, the fluid limits in both scaling regime agree with each other as time goes to infinity. Further, we show that for non-zero abandonment rates, the fluid limits converge to a unique fixed point, which is representative of the long run average number of users in the system. We prove that as the abandonment rate increases, the fixed point component for the class with lower arrival rate first experiences an increase and then decrease, while for the class with higher arrival rate it decreases monotonically. Finally we present numerical results of the fluid and diffusion limits in the second scaling regime.

There exists an extensive literature on fluid and diffusion approximations for Markovian systems with abandonments. Ward and Glynn [17] suggest diffusion approximations for the $M/M/1$ queue with exponential abandonments. They generalize these results to arrival, service and abandonment times with general distributions in [18]. Garnett et al. [9] consider $M/M/N$ queue with exponential abandonments and suggest diffusion approximations under Halfin-Whitt regime (see Halfin and Whitt [11]). Generalizing these results, Dai and He [7] and Mandelbaum and Momčilović [15] suggest diffusion approximations for many-server queues with general arrival, service and abandonment times. A recent work by Liu et al. [14] suggests diffusion approximations for the double sided queue where arrivals are renewal processes and customers abandon the system if they cannot find a match after an exponential time. This paper is closest to our work in nature and even though we restrict ourselves to Poisson arrival processes, our work extends [14] by assuming probabilistic matching structure.

2 The Probabilistic Matching Model

In this work, we study probabilistic matching systems introduced in Büke and Chen [3], where two classes of users, indexed by $i = 1, 2$ arrive at the system to be matched with users of the other class. We assume that class-i users arrive according to a Poisson process with rate λ_i. Any given pair of class-1 and class-2 users can match with each other with probability q independent of other users. Let $X_i(t)$ be the number of class-i users in the system at time t. When a class-1 user arrives at time t, she checks the class-2 queue to see if there exists any suitable users that she can match with. If she can find one or more suitable class-2 users to match with, she chooses one of them uniformly at random and they leave the system together. Otherwise, she joins the class-1 queue and waits in the system until she is picked by an arriving class-2 user. Due to the independence of matchings, a class-1 user finds a suitable class-2 user to match upon arrival with probability $1-(1-q)^{X_2(t)}$ and is not able to match with anyone with probability $(1-q)^{X_2(t)}$. For the analysis in Section 4 we also assume that the users are impatient and each user abandons the system without being matched after waiting an exponential time with rate $\gamma \geq 0$.

Under the assumption of Poisson arrivals, the number of users in a probabilistic matching system, $\{(X_1(t), X_2(t)), t \geq 0\}$, can be modelled as a continuous-time Markov chain (CTMC) on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the generator matrix

$$Q_{(i,j)(l,k)} = \begin{cases} \lambda_1(1-q)^j & \text{if } l = i+1 \text{ and } j = k, \\ \lambda_2(1-q)^i & \text{if } l = i \text{ and } k = j+1, \\ \lambda_1(1-(1-q)^j) + \gamma i & \text{if } l = i \text{ and } k = j-1 \geq 0, \\ \lambda_2(1-(1-q)^i) + \gamma j & \text{if } l = i-1 \geq 0 \text{ and } k = j, \\ -(\lambda_1 + \lambda_2 + \gamma(i+j)) & \text{if } l = i \text{ and } k = j, \\ 0 & \text{otherwise.} \end{cases}$$

The above model reduces to the one introduced in [3] if users do not abandon the system ($\gamma = 0$).

It is sometimes useful in our analysis to express the queue length processes, $X_i(t)$, as the difference of counting processes. We define $A_i(t)$ and $R_i(t)$ to be the number of arrivals and the

number of user abandonments from class-i up to time t, respectively. Similarly, defining $M(t)$ to be the number of matched pairs up to time t, we have the basic relation

$$X_i(t) = A_i(t) - M(t) - R_i(t) \text{ for all } t \geq 0 \text{ and } i = 1, 2.$$

The essential element distinguishing a probabilistic matching system from a conventional queuing system is the matching probability q. To see this, consider a probabilistic matching system with no abandonments ($\gamma = 0$). For systems with matching probability $q = 1$, class-1 and class-2 users cannot co-exist in the system at any time. Hence, the probabilistic matching system can be modeled as a continuous time random walk on integers $\{X(t), t \geq 0\}$ where $X(t) = k$ if $(X_1(t), X_2(t)) = (0, k))$ and $X(t) = -k$ if $(X_1(t), X_2(t)) = (0, k)$. Also when $q = 1$, the number of matched pairs up to time t is equal to the minimum of class-1 and class-2 arrivals. Hence,

$$X_i(t) = A_i(t) - M(t) = A_i(t) - \min\{A_1(t), A_2(t)\}, \text{ for all } t \geq 0 \text{ and } i = 1, 2.$$

However, when $0 < q < 1$, analyzing the matching process $M(t)$ is far more difficult. The one dimensional distribution of the matching process, $\mathbb{P}(M(t) = k)$ for a given $t \geq 0$ and $k \in \mathbb{N}$ is provided in [3] and its complicated nature indicates the difficulty in fully characterizing the law of the matching process. Hence, in this paper we propose fluid and diffusion approximations for probabilistic matching systems.

3 Fluid and Diffusion Approximations with Constant Matching Probability

In this section we focus on fluid and diffusion approximations for probabilistic matching systems obtained by only scaling time (or equivalently the arrival rates) and space while keeping the matching probability constant. This approach is especially useful in approximating systems where the probability that a given pair of users matches is high. For scalings with a constant matching probability, we assume that the users do not abandon the system without being matched, i.e., $\gamma = 0$.

3.1 Fluid Limits

We start with defining the scaled process $\{(\bar{X}_1^n(t), \bar{X}_2^n(t)), t \geq 0\}$ as $\bar{X}_i^n(t) = \frac{X_i(nt)}{n}, i = 1, 2$. In the rest of the paper, we use the notation $X(\omega, t)$ when we need to specify the sample path of the stochastic process $X(t)$ corresponding to a scenario $\omega \in \Omega$. For any $\omega \in \Omega$, we say that $\bar{X}_i^n(\omega, t)$ converges uniformly on compact sets (u.o.c.) to $\bar{X}_i(\omega, t)$ if $\sup_{0 \leq t \leq T} |\bar{X}_i^n(\omega, t) - \bar{X}_i(\omega, t)|$ converges to 0 for all $T > 0$ as $n \to \infty$. A direct application of the functional strong law of large numbers (see e.g. [2, 6, 19]) to Poisson arrival processes yields

$$\bar{A}_i^n(t) := \frac{A_i(nt)}{n} \xrightarrow{\text{a.s.}} \lambda_i t \text{ u.o.c. as } n \to \infty, i = 1, 2, \tag{1}$$

where a.s. indicates that the convergence is almost surely.

As users of a class accumulate in the system, the users of the other class are more likely to match upon their arrival. This implies that class-1 and class-2 users are unlikely to accumulate in the system at the same time. Lemma 1 formalizes this argument.

Lemma 1. *For any fixed $k > 0$, $\min\{\frac{X_1(nt)}{n^k}, \frac{X_2(nt)}{n^k}\} \xrightarrow{a.s.} 0$ u.o.c. as $n \to \infty$.*

Proof. If $q = 1$, since class-1 and class-2 do not co-exist in the system, for any $t \geq 0$, $\min\{X_1^n(t), X_2^n(t)\} = 0$, and hence the desired conclusion follows trivially. If $0 < q < 1$, to simplify the notation, define $I^{n,k}(t) := \min(\frac{X_1(nt)}{n^k}, \frac{X_2(nt)}{n^k})$, choose an $a \in (0, k)$ and let $\lambda = \lambda_1 + \lambda_2$. Then for $m \leq n^2 - 1 \in \mathbb{N}$, we have

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{0 \leq t \leq \frac{m+1}{n^2}} I^{n,k}(t) \geq n^{-a} \Big| \sup_{0 \leq t \leq \frac{m}{n^2}} I^{n,k}(t) < n^{-a}\Big) \\
&\qquad = \mathbb{P}\Big(\sup_{\frac{m}{n^2} \leq t \leq \frac{m+1}{n^2}} I^{n,k}(t) \geq n^{-a} \Big| \sup_{0 \leq t \leq \frac{m}{n^2}} I^{n,k}(t) < n^{-a}\Big) \\
&\qquad = \mathbb{P}\Big(\sup_{\frac{m}{n^2} \leq t \leq \frac{m+1}{n^2}} \min(X_1(nt), X_2(nt)) \geq n^{k-a} \Big| \sup_{0 \leq t \leq \frac{m}{n^2}} \min(X_1(nt), X_2(nt)) < n^{k-a}\Big) \\
&\qquad \leq \sum_{j=0}^{\infty} \frac{e^{-\frac{\lambda}{n}}(\frac{\lambda}{n})^j}{j!} j r^{n^{k-a}} \qquad (2) \\
&\qquad = \frac{\lambda}{n^2} r^{n^{k-a}}.
\end{aligned}$$

We see that the inequality (2) holds using the following argument. For both $X_1(nt)$ and $X_2(nt)$ to reach a level above n^{k-a} at some point during $[\frac{m}{n^2}, \frac{m+1}{n^2}]$, at least one of the arrivals occurring during $[\frac{m}{n^2}, \frac{m+1}{n^2}]$ should fail to match and stay in the system upon arrival when facing at least $[n^{k-a}]$ users from the other user queue. If we observe k arrivals during this time frame, the probability of this event is bounded by $jr^{n^{k-a}}$. Then, for any fixed $T > 0$,

$$\begin{aligned}
\mathbb{P}(\sup_{0\le t\le T} I^{n,k}(t) \ge n^{-a}) &= \mathbb{P}(\sup_{0\le t\le T} I^{n,k}(t) \ge n^{-a} | \sup_{0\le t\le T-\frac{1}{n^2}} B^n(t) < n^{-a}) \mathbb{P}(\sup_{0\le t\le T-\frac{1}{n^2}} I^{n,k}(t) < n^{-a}) \\
&\quad + \mathbb{P}(\sup_{0\le t\le T} I^{n,k}(t) \ge n^{-a} | \sup_{0\le t\le T-\frac{1}{n^2}} I^{n,k}(t) \ge n^{-a}) \mathbb{P}(\sup_{0\le t\le T-\frac{1}{n^2}} I^{n,k}(t) \ge n^{-a}) \\
&\le \mathbb{P}(\sup_{0\le t\le T} I^{n,k}(t) \ge n^{-a} | \sup_{0\le t\le T-\frac{1}{n^2}} I^{n,k}(t) < n^{-a}) \\
&\quad + \mathbb{P}(\sup_{0\le t\le T-\frac{1}{n^2}} I^{n,k}(t) \ge n^{-a}) \\
&\le \sum_{m=0}^{Tn^2} \mathbb{P}(\sup_{0\le t\le \frac{m+1}{n^2}} I^{n,k}(t) \ge n^{-a} | \sup_{0\le t\le \frac{m}{n^2}} I^{n,k}(t) < n^{-a}) \\
&\le \sum_{m=0}^{Tn^2} \frac{\lambda}{n^2} r^{n^{k-a}} \\
&= T\lambda r^{n^{k-a}}
\end{aligned}$$

For any $\epsilon > 0$, there exists an N_ϵ, such that for $n \ge N_\epsilon$ such that $\mathbb{P}(\sup_{0\le t\le T} I^{n,k}(t) \ge \epsilon) \le \epsilon$, which implies $I^{n,k}(t) \xrightarrow{\mathbb{P}} 0$ u.o.c. Furthermore,

$$\sum_{n=1}^{\infty} \mathbb{P}(\sup_{0\le t\le T} I^{n,k}(t) \ge n^{-a}) \le T\lambda \sum_{n=0}^{\infty} r^{n^{k-a}} < \infty.$$

For any $\epsilon > 0$ choosing $N \ge 1$, such that for $N^{-a} < \epsilon$, we get

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathbb{P}(\sup_{0\le t\le T} I^{n,k}(t) > \epsilon) &= \sum_{n=1}^{N-1} \mathbb{P}(\sup_{0\le t\le T} I^{n,k}(t) > \epsilon) + \sum_{n=N}^{\infty} \mathbb{P}(\sup_{0\le t\le T} I^{n,k}(t) > \epsilon) \\
&\le \sum_{n=1}^{N-1} \mathbb{P}(\sup_{0\le t\le T} I^{n,k}(t) > \epsilon) + \sum_{n=N}^{\infty} \mathbb{P}(\sup_{0\le t\le T} I^{n,k}(t) \ge n^{-a}) < \infty
\end{aligned}$$

Using Borel-Cantelli lemma we get

$$\mathbb{P}(\sup_{0\le t\le T} I^{n,k}(t) > \epsilon \text{ infinitely often}) = 0,$$

and $I^{n,k}(t) = \min(\frac{X_1(nt)}{n^k}, \frac{X_2(nt)}{n^k}) \xrightarrow{\text{a.s.}} 0$ u.o.c. □

Theorem 2. *$\bar{X_i^n}(t) \xrightarrow{a.s.} \bar{X}_i(t)$ u.o.c. as $n \to \infty$, where $\bar{X}_i(t) = \lambda_i t - \min(\lambda_1, \lambda_2)t$, $i = 1, 2$.*

Proof. Equation (1) and Lemma 1 imply that there exists a $\Omega' \subset \Omega$ with $\mathbb{P}(\Omega') = 1$ where for every $\omega \in \Omega'$,

$$\frac{A_i(\omega, nt)}{n} \to \lambda_i t \text{ u.o.c.}$$
$$\min(\frac{X_1(\omega, nt)}{n}, \frac{X_2(\omega, nt)}{n}) \to 0 \text{ u.o.c.}$$

for all $t \ge 0$ and $i = 1, 2$. Our first goal is to show $\bar{M}^n(\omega, t) := \frac{M(nt)}{n} \to \min\{\lambda_1 t, \lambda_2 t\}$ u.o.c. as $n \to \infty$ for all $t \ge 0$ and $\omega \in \Omega'$. Suppose that there exists some $\omega' \in \Omega'$ which this statement does not hold and without loss of generality assume $\lambda_1 \ge \lambda_2$. Also, we know that the number of matchings is always bounded by the number of arrivals as $M(\omega, t) < \min\{A_1(t), A_2(t)\}$ for all $t \ge 0$. These imply that there exists a $\delta > 0$, $N_\delta > 0$ sequences $n_j \to \infty$ as $j \to \infty$ and $0 \le t_j \le T$ such that $\lambda_2 t_j - \bar{M}^{n_j}(\omega, t_j) > \delta$ for all $j > N_\delta$. Boundedness of t_j and $M(\omega, 0) = 0$ also implies that there exists a subsequence $t_{j_k} \to t' > 0$. For any $\epsilon > 0$, we can choose N_ϵ such that for every $k > N_\epsilon$ we have $|\frac{A_i(n_{j_k} t_{j_k})}{n_{j_k}} - \lambda_i t_{j_k}| < \frac{\epsilon}{2}$ for $i = 1, 2$ and $|t_{j_k} - t'| < \frac{\epsilon}{2(\lambda_1 - \lambda_2)}$, which in turn implies

$$\begin{aligned}
\frac{A_1(n_{j_k} t_{j_k})}{n_{j_k}} - \frac{M(n_{j_k} t_{j_k})}{n_{j_k}} &= \frac{A_1(n_{j_k} t_{j_k})}{n_{j_k}} - \frac{M(n_{j_k} t_{j_k})}{n_{j_k}} - (\lambda_1 - \lambda_2)(t_{j_k} - t') + (\lambda_1 - \lambda_2)(t_{j_k} - t') \\
&> \lambda_2 t_{j_k} - \frac{M(n_{j_k} t_{j_k})}{n_{j_k}} + (\lambda_1 - \lambda_2)t' - \epsilon \\
&> \delta - \epsilon + (\lambda_1 - \lambda_2)t'.
\end{aligned}$$

Similarly, we also get

$$\frac{A_2(n_{j_k}t_{j_k})}{n_{j_k}} - \frac{M(n_{j_k}t_{j_k})}{n_{j_k}} > \delta - \epsilon.$$

Letting $\epsilon \to 0$, we get

$$\min\{\frac{X_1(\omega, n_{j_k}t_{j_k})}{n_{j_k}}, \frac{X_2(\omega, n_{j_k}t_{j_k})}{n_{j_k}}\} > \delta,$$

which contradicts with Lemma 1 and proves $\bar{M}^n(t) \xrightarrow{\text{a.s.}} \min\{\lambda_1 t, \lambda_2 t\}$ u.o.c. as $n \to \infty$ for all $t \geq 0$. Then the result follows using continuous mapping theorem ([6], Theorem 5.2). □

3.2 Diffusion Limits

Fluid limits provide useful approximations to determine how queue lengths grow, however they fail to represent the stochastic fluctuations. To understand the fluctuations of sample paths around the fluid limit, we now focus on diffusion approximations. A direct application of functional central limit theorem (see e.g. Theorem 5.7 in [6]) on Poisson arrival streams we get

$$\hat{A}_i^n(t) := \frac{A_i(nt) - n\bar{A}_i(t)}{\sqrt{n}} \Rightarrow \hat{A}_i(t), i = 1, 2, \tag{3}$$

where $\hat{A}_i = \sqrt{\lambda_i}B_i$, $B_i(t), i = 1, 2$, is independent one-dimensional standard Brownian motions and "$\Rightarrow$"denotes weak convergence. We define the process

$$\hat{X}_i^n(t) = \frac{X_i(nt) - \bar{X}_i(nt)}{\sqrt{n}}.$$

Now we are ready to state the diffusion limits for probabilistic matching systems when the matching probability is kept constant.

Theorem 3. *As $n \to \infty, \hat{X}_i^n \Rightarrow \hat{X}_i, i = 1, 2,$ where $\hat{X}_i$ is defined as:*

1. *If $\lambda_1 = \lambda_2$, $\hat{X}_i = \hat{A}_i - \min(\hat{A}_1, \hat{A}_2), i = 1, 2.$*

2. *If $\lambda_1 > \lambda_2, \hat{X}_1 = \hat{A}_1 - \hat{A}_2, \hat{X}_2 = 0.$*

Proof. We first consider the case when $\lambda_1 = \lambda_2 = \lambda$. Define $\hat{M}^n(t) := \frac{M(nt)-\lambda nt}{\sqrt{n}}$. Using Skorohod representation theorem (Theorem 5.1 in [6]) there exists versions of $A_i(t), \hat{A}_i(t)$ and $B_i(t), i = 1, 2$, which we denote $A_i'(t), \hat{A}_i'(t)$ and $B_i'(t), i = 1, 2$, and matching and scaled processes $M(t)$ and $\hat{A}_i^{n'}(t), i = 1, 2$ associated with these versions such that $\hat{A}_i^{n'}(t) \xrightarrow{\text{a.s.}} \hat{A}_i'(t) = \sqrt{\lambda}B_i'(t), i = 1, 2$. Lemma 1 implies $\min(\hat{A}_1^{n'}(t) - \hat{M}^{n'}(t), \hat{A}_2^{n'}(t) - \hat{M}^{n'}(t)) \xrightarrow{\text{a.s.}} 0$ u.o.c. Proceeding as in the same manner as in the proof of Theorem 2, we get $\hat{M}^{n'} \xrightarrow{\text{a.s.}} \min(\hat{A}_1', \hat{A}_2')$. Applying the continuous mapping theorem (Theorem 5.2 in [6]) the result follows for $\lambda_1 = \lambda_2 = \lambda$.

When $\lambda_1 > \lambda_2$, let $\tau_n = \inf\{t \geq 0 : A_2(t) \geq n\}$ and define a sequence of random variables $\{\xi_n\}_{n\geq 1}$ such that

$$\xi_n = \begin{cases} 1, & \text{the } n\text{-th arriving user-2 finds a match successfully upon her arrival,} \\ 0, & \text{otherwise.} \end{cases}$$

We have $\tau_n \to \infty$, as $n \to \infty$, and for any $n \geq 1$, $\sum_{n=1}^{A_2(t)} \xi_n \leq M(t)$. Generate a sequence of a uniform random variables $\{U_n\}_{n\geq 1}$ such that $U_n \sim U(0, 1)$, then assuming $0^0 = 1$, we have

$$\begin{aligned} \mathbb{P}(\xi_n = 0) &= \mathbb{P}(U_n < (1-q)^{X_1(\tau_n)}) \\ &\leq \mathbb{P}(U_n < (1-q)^{A_1(\tau_n)-n}) \\ &= \mathbb{E}[\mathbb{P}(U_n < (1-q)^{A_1(\tau_n)-n} | A_1(\tau_n))] \\ &= \mathbb{E}[((1-q)^{A_1(\tau_n)-n}) \wedge 1]. \end{aligned}$$

Next we show that there exists an $N > 0$ and $c > 0$ such that for any $n \geq N$,

$$\mathbb{E}[(1-q)^{A_1(\tau_n)-n}] < (1-q)^{cn}.$$

For any c_1 such that $1 < c_1 < \frac{\lambda_1}{\lambda_2}$ we have $\frac{A_1(t)}{t} - c_1\frac{A_2(t)}{t} \xrightarrow{\text{a.s.}} \lambda_1 - c_1\lambda_2$, i.e., there exists a $T > 0$, such that for any $t > T$, $A_1(t) - c_1A_2(t) > \frac{(\lambda_1 - c_1\lambda_2)t}{2}$ a.s. Since $\tau_n \to \infty$, there exists an $N > 0$

such that for any $n \geq N$, we have $\tau_n > T$ and

$$A_1(\tau_n) - c_1 A_2(\tau_n) = A_1(\tau_n) - c_1 n > \frac{(\lambda_1 - c_1\lambda_2)}{2}\tau_n > 0 \text{ a.s.}$$

Choosing $c = c_1 - 1$ we have $\mathbb{E}[(1-q)^{A_1(\tau_n)-n}] < (1-q)^{cn}$ and

$$\sum_{n=0}^{\infty} \mathbb{P}(\xi_n = 0) = \sum_{n=0}^{\infty} \mathbb{P}(U_n < r^{X_1(T_2(n))}) = \sum_{n=0}^{\infty} r^{cn} < \infty.$$

Using Borel-Cantelli Lemma, $\mathbb{P}(\xi_n = 1 \text{ infinitely often}) = 0$ which in turn implies

$$\hat{X}_2^n(t) = \frac{A_2(nt) - M(nt)}{\sqrt{n}} \xrightarrow{\text{a.s.}} 0.$$

Finally, we have

$$\begin{aligned}\hat{X}_1^n(t) &= \frac{A_1(nt) - M(nt)}{\sqrt{n}} - \frac{(\lambda_1 - \lambda_2)nt}{\sqrt{n}} \\ &= \frac{A_1(nt) - \lambda_1 nt}{\sqrt{n}} - \frac{A_2(nt) - \lambda_2 nt}{\sqrt{n}} - \frac{A_2(nt) - M(nt)}{\sqrt{n}}.\end{aligned}$$

Hence, the result follows from the continuous mapping theorem. □

We conclude that when the matching probability q is kept as a constant in the diffusion approximation, it is not present in both the fluid limits and the diffusion limits. Moreover, we can compare our results with those of an $M/M/1$ queue. When the arrival rates in probabilistic matching systems are not equal, the fluid and diffusion limits of queue length process i behaves in accordance with that in an M/M/1 queue with arrival rate λ_i and service rate λ_j (see Chen and Yao (2001) [6] for more details). When the arrival rates are identical, the diffusion limits are distinct from those of an M/M/1 queue, due to the fact that in a probabilistic matching system, the next arriving user i is possible to be matched immediately upon arrival which indicates that the accumulation of user j when no user i is at present would not be a "waste" unlike the service time generated in an empty M/M/1 queue. As a result, rather than having the one-sided regular function of the net-input process, we only have the positive sign of the difference between the arrival processes. We

suggest that this diffusion approximation would fit the system which has a relatively high matching probability of each pair of users and thus the probability of an arriving user getting matched increases significantly as the number of users from the other queue grows. However, the underlying assumption above does not hold in those systems which have a very small matching probability for each pair of users, because if q very close to 0, a user is not so likely to find a match upon arrival even when there are many users in the other queue.

4 Fluid and Diffusion Limits for Systems with Small Matching Probability

The matching probability disappears in the fluid and diffusion limits presented in Section 3 and this indicates that at most one class of users accumulate in the system and the systems with matching probability $0 < q < 1$ behave very similar to the systems with matching probability 1. However, in many real world problems the matching probability q is very small and we need tools that explicitly addresses the probabilistic nature of the matcings. In this section, we suggest a second type of diffusion approximation which scales q together with the space and time to get a better description of the dynamics of those systems with small matching probabilities.

We often observe that the users are impatient and may leave the system without being matched if they cannot match after waiting for sometime. We include this factor in the discussion of the queue length process in the new asymptotic regime, adopting a similar approach to that of Ward and Glynn [17]. We assume that each user has an exponentially distributed abandonment time with rate γ, $0 \leq \gamma < \infty$, independent of others, where $\gamma \ll \lambda_i, i = 1, 2$. Hence, as we scale space, time and the matching probability, we also let abandonment rate approach to zero.

4.1 Fluid Limits

Let $X_i^n(t)$ to be number of class-i users in a probabilistic matching system where class-i users arrive according to a Poisson process with λ_i, users abandon the system if they do not match after waiting an exponential time $\gamma^n = \frac{\gamma}{n}$, $(0 \leq \gamma < \infty)$, the matching probability is $q^n = \frac{q}{n}$, $0 < q < 1$. Then,

we define

$$\bar{X}^{s,n}(t) := \frac{X_i^n(nt)}{n}$$

to be the scaled system in this regime. Now our goal is to show that as $n \to \infty$, the scaled system approaches to the fluid limit $\bar{X}^s(t)$, which is the unique solution to the following ordinary differential equations (ODE):

$$\bar{X}_1^s(0) = \bar{X}_2^s(0) = 0, \tag{4}$$

$$\frac{d\bar{X}_1^s(t)}{dt} = \lambda_1 e^{-q\bar{X}_2^s(t)} - \lambda_2(1 - e^{-q\bar{X}_1^s(t)}) - \gamma\bar{X}_1^s(t), \tag{5}$$

$$\frac{d\bar{X}_2^s(t)}{dt} = \lambda_2 e^{-q\bar{X}_1^s(t)} - \lambda_1(1 - e^{-q\bar{X}_2^s(t)}) - \gamma\bar{X}_2^s(t). \tag{6}$$

The equations (5)-(6) are in the form $\frac{dx}{dt} = F(x) = (F_1(x), F_2(x))'$, where $F(\cdot)$ is Lipschitz and hence the initial value problem admits a unique solution. We first show that the solution $\bar{X}^s(t)$ is bounded when $\gamma > 0$.

Lemma 4. *Let $\bar{X}^s(t) = (\bar{X}_1^s(t), \bar{X}_2^s(t))$ be the unique solution to (4)-(6) and $\gamma > 0$, then*

$$\sup_{0 \le t < \infty} \bar{X}_i^s(t) < \lambda_i/\gamma, i = 1, 2$$

.

Proof. For any (x_1, x_2) such that $x_1 \ge \lambda_1/\gamma$, we have

$$F_1(x_1, x_2) = \lambda_1 e^{-qx_2} - \lambda_2(1 - e^{-qx_1}) - \gamma x_1 < \lambda_1 - \gamma x_1 \le 0.$$

Using (4) this implies that $\bar{X}_1^s(t) \le \lambda_1/\gamma$, for all t. Similar argument also holds for $\bar{X}_2^s(t)$. □

When the matching probability is scaled in a way that $q^n \to 0$, the techniques we use to derive fluid and diffusion limits differ from the ones used in Section 3. In particular, we appeal to the Laplace transform methods where a limiting kernel with the corresponding Laplace transform is identified (see e.g. [8] for a brief review of these methods). For this purpose, we need the Lévy

kernel for the Markov process $\bar{X}^{s,n}(t)$ defined as follows:

$$\begin{aligned}K^n(x,dy) := \lambda_1 n(1-\frac{q}{n})^{nx_2}\delta(y-(\frac{1}{n},0))dy + \lambda_2 n(1-\frac{q}{n})^{nx_1}\delta(y-(0,\frac{1}{n}))dy \\ + (\lambda_1 n(1-(1-\frac{q}{n})^{nx_2})+\gamma n x_2)\delta(y+(0,\frac{1}{n}))dy \\ + (\lambda_2(1-(1-\frac{q}{n})^{nx_1})+\gamma n x_1)\delta(y+(\frac{1}{n},0))dy,\end{aligned}$$

where $\delta(y)$ is the Dirac delta function. Then, we can define the Laplace transform of operator $K^n(x,dy)$ as

$$\begin{aligned}m^n(x,\theta) &= \int_{(0,\infty)\times(0,\infty)} e^{\langle\theta,y\rangle}K^n(x,dy) \\ &= \lambda_1 n(1-\frac{q}{n})^{nx_2}e^{\frac{\theta_1}{n}} + \lambda_2 n(1-\frac{q}{n})^{nx_1}e^{\frac{\theta_2}{n}} \\ &\quad + (\lambda_1 n(1-(1-\frac{q}{n})^{nx_2})+\gamma n x_2)e^{-\frac{\theta_2}{n}} + (\lambda_2 n(1-(1-\frac{q}{n})^{nx_1}+\gamma n x_1)e^{-\frac{\theta_1}{n}}. \qquad (7)\end{aligned}$$

Now, we are ready to state our result for convergence to the fluid limit.

Theorem 5. *For any $\delta > 0$ and $T > 0$,*

$$\limsup_{n\to\infty} n^{-1}\log\mathbb{P}(\sup_{0\le t\le T}|\bar{X}_i^{s,n}(t)-\bar{X}_i^s(t)|>\delta) < 0 \qquad (8)$$

and

$$\bar{X}_i^{s,n}(t) \xrightarrow{a.s.} \bar{X}_i^s(t) \ u.o.c.,$$

where $\bar{X}_i^s(t), i=1,2$ is the unique solution to the system of ODE given by (4)-(6).

Proof. If $\gamma = 0$, set $\mathbb{S} = \mathbb{R}_{\ge 0}\times\mathbb{R}_{\ge 0}$ and $T^n = T$, otherwise choose $C_i > \lambda_i/\gamma$ for $i=1,2$, and set $\mathbb{S} = [0,C_1]\times[0,C_2]$ and $T^n = \inf\{t\ge 0 : \bar{X}^{s,n}(t)\notin\mathbb{S}\}\wedge T$. Then, Proposition 5.1 in [8] implies

$$\limsup_{n\to\infty} n^{-1}\log\mathbb{P}(\sup_{0\le t\le T^n}|\bar{X}_i^{s,n}(t)-\bar{X}_i^s(t)|>\delta) < 0 \qquad (9)$$

if we can show that the following three conditions hold:

(i) There exists a $\eta_0 > 0$ such that

$$\sup_n \sup_{x\in\mathbb{S}} \sup_{|\theta|\leq\eta_0} \frac{m^n(x,n\theta)}{n} < \infty$$

.

(ii) $\sup_{x\in\mathbb{S}} \left| \frac{\partial m^n(x,\theta)}{\partial\theta}|_{\theta=0} - F(x) \right| \to 0.$

(iii) $\limsup_{n\to\infty} n^{-1} \log \mathbb{P}(|\bar{X}_i^{s,n}(0) - \bar{X}_i^s(0)| > \delta) < 0.$

The third condition is trivially satisfied as we assume that a probabilistic matching system is initially empty and we have $\bar{X}^{s,n}(0) = 0$ for all n. When $\gamma > 0$ the first condition follows as when $x \in \mathbb{S}$ for any $\eta_0 > 0$ and $\theta \leq \eta_0$ we have

$$\begin{aligned}
\frac{m^n(x,n\theta)}{n} &= \lambda_1(1-\frac{q}{n})^{nx_2}e^{\theta_1} + \lambda_2(1-\frac{q}{n})^{nx_1}e^{\theta_2} \\
&\quad + (\lambda_1(1-(1-\frac{q}{n})^{nx_2}) + \gamma x_2)e^{-\theta_2} + (\lambda_2(1-(1-\frac{q}{n})^{nx_1}) + \gamma x_1)e^{-\theta_1} \\
&\leq (\lambda_1+\lambda_2)e^{\eta_0} + \lambda_1 + \lambda_2 + \gamma(C_1+C_2).
\end{aligned}$$

Similarly, when $\gamma = 0$, the supremum can be bounded by $\leq (\lambda_1+\lambda_2)e^{\eta_0} + \lambda_1 + \lambda_2$. To prove the second condition we write

$$\begin{aligned}
\left.\frac{\partial m^n(x,\theta)}{\partial\theta_1}\right|_{\theta_1=0} &= \lambda_1(1-\frac{q}{n})^{nx_2} - (\lambda_2(1-(1-\frac{q}{n})^{nx_1}) + \gamma x_1), \\
\left.\frac{\partial m^n(x,\theta)}{\partial\theta_2}\right|_{\theta_2=0} &= \lambda_2(1-\frac{q}{n})^{nx_1} - (\lambda_1(1-(1-\frac{q}{n})^{nx_1}) + \gamma x_2).
\end{aligned}$$

Then it is easy to see pointwise convergence $\frac{\partial m^n(x,\theta)}{\partial\theta}|_{\theta=0} \to F(x)$ and the uniform convergence follows from continuity of the functions and compactness of the underlying set and (9) follows from Proposition 5.1 in [8]. When $\gamma = 0$, $T_n = T$ a.s., and when $\gamma > 0$ from (9), Lemma 4 and $C_i > \lambda_i/\gamma$ we conclude that $T^n \xrightarrow{\mathbb{P}} T$, which implies (8). The almost sure convergence is a simple application of Borel-Cantelli lemma. □

When there are abandonments ($\gamma > 0$), the right hand sides of (5) and (6) involve both e^{-qx}

and x terms which makes it difficult to obtain an analytical solution. However, when the customers do not abandon the system, the ODE can be solved analytically. Corollary 6 presents this special case.

Corollary 6. *When $\gamma = 0$, as $n \to \infty$,*

$$\bar{X}_i^{s,n}(t) \xrightarrow{a.s.} \frac{1}{q}\ln(e^{\lambda_1 qt} + e^{\lambda_2 qt} - 1) - \mathbb{I}_{\{i=2\}}\lambda_1 t - \mathbb{I}_{\{i=1\}}\lambda_2 t \ \ u.o.c., i = 1, 2. \tag{10}$$

Proof. Setting $\gamma = 0$ and taking the integral of (5) and (6), we see that

$$\bar{X}_1^s(t) + \lambda_2 t = \bar{X}_2^s(t) + \lambda_1 t =: y(t).$$

Then, we have

$$\frac{dy(t)}{dt} = e^{-qy(t)}(\lambda_1 e^{\lambda_1 qt} + \lambda_2 e^{\lambda_2 qt})$$

and $y(0) = 0$ which has the unique solution $y(t) = \frac{1}{q}\ln(e^{\lambda_1 qt} + e^{\lambda_2 qt} - 1)$ and the result follows. □

In [3], certain performance measures are proven to be independent of the matching probability q under some additional control policies. Specifically, Theorem 14 in [3] states that under an admission control policy where the difference between long run average queue lengths of class-1 and class-2 users does not depend on the matching probability q. The following corollary also indicates a similar property even under the presence of user abandonments.

Corollary 7. *When $\gamma > 0$, as $n \to \infty$, $\bar{X}_1^{s,n}(t) - \bar{X}_2^{s,n}(t) \xrightarrow{a.s.} \frac{\lambda_2 - \lambda_1}{\gamma}e^{-\gamma t} + \frac{\lambda_1 - \lambda_2}{\gamma}$.*

Proof. Applying Theorem 5 and the continuous mapping theorem, $\bar{X}_1^{s,n}(t) - \bar{X}_2^{s,n}(t)$ converges to the unique solution of

$$\frac{dx(t)}{dt} = \lambda_1 - \lambda_2 - \gamma x(t) \tag{11}$$

with initial condition $x(0) = 0$. Using integrating factors, the solution of this first order ODE can be obtained as $\bar{X}_1^s(t) - \bar{X}_2^s(t) = \frac{\lambda_2 - \lambda_1}{\gamma}e^{-\gamma t} + \frac{\lambda_1 - \lambda_2}{\gamma}$. □

Corollary 7 implies that when $\gamma > 0$, the matching probability q does not affect the difference between the numbers of class-1 and class-2 users in the system. As $t \to \infty$, this difference converges

to $\frac{\lambda_1-\lambda_2}{\gamma}$, which coincides with the results of [17] for M/M/1+M queue with has arrival rate λ_1, service rate λ_2 and abandonment rate $\gamma > 0$.

Next, we analyze the asymptotic behaviour of the fluid limits as time goes to infinity. Corollary 6 assumes γ to be 0 and allows us to compare $\bar{X}^s(t)$ with fluid limits $\bar{X}(t)$, given in Theorem 2. Different from $\bar{X}(t)$ which does not carry any information on the matching probability q, the fluid limits in Corollary 6 depends on q. When t is small, $\bar{X}_i^s(t)$ grows for both $i = 1$ and 2 as q increases. However, as t becomes larger, the influence of the matching probability becomes weaker. Proposition 8 shows that the fluid limits $\bar{X}^s(t)$ converges to $\bar{X}(t)$ as $t \to \infty$.

Proposition 8. *Suppose $\gamma = 0$, then as $t \to \infty$, $|\bar{X}_i(t) - \bar{X}_i^s(t)| \to 0, t \geq 0, i = 1, 2$.*

Proof. Without lost of generality, we assume that $\lambda_1 \geq \lambda_2$ and using Corollary 6 we get

$$\begin{aligned}\bar{X}_1^s(t) - \bar{X}_1(t) &= \frac{1}{q}\ln(e^{\lambda_1 qt} + e^{\lambda_2 qt} - 1) - \lambda_1 t \\ &= \ln(e^{\lambda_1 qt} + e^{\lambda_2 qt} - 1)^{\frac{1}{q}} - \lambda_1 t \\ &= \ln \frac{\sqrt[q]{e^{\lambda_1 qt} + e^{\lambda_2 qt} - 1}}{\sqrt[q]{e^{\lambda_1 qt}}}\end{aligned}$$

Since $\lambda_1 > \lambda_2$, as $t \to \infty$, $|\frac{\sqrt[q]{e^{\lambda_1 qt}+e^{\lambda_2 qt}-1}}{\sqrt[q]{e^{\lambda_1 qt}}}| \to 1$ and $|\bar{X}_1^s(t) - \bar{X}_1(t)| \to 0$. □

In other words, we can explain the dynamics of a probabilistic matching system in the following way: without considering the effect of user abandonments, if each pair of users gets harder to match with each other, we observe more users waiting in the system. However if we run the system long enough, the average of numbers of users in the system only depends on the arrival rates. Next we show that for general abandonment rate $\gamma \geq 0$, the fluid limits of the queue length processes converge to a fixed point as $t \to \infty$.

Proposition 9. *If $\gamma > 0$, the fluid limit $\bar{X}_i^s(t) \to x_i^*, i = 1, 2$ as $t \to \infty$, where $x_i^* \in \mathbb{R}$ satisfies the following set of equations*

$$\lambda_1 e^{-qx_2^*} - \lambda_2(1 - e^{-qx_1^*}) - \gamma x_1^* = 0, \tag{12}$$

$$\lambda_2 e^{-qx_1^*} - \lambda_1(1 - e^{-qx_2^*}) - \gamma x_2^* = 0. \tag{13}$$

Proof. First, we prove that Equations (12) and (13) have a unique solution. Subtracting the second equation from the first one $x_2^* = x_1^* + \frac{\lambda_2 - \lambda_1}{\gamma}$ and replacing this into (12) we get

$$\lambda_1 e^{-\frac{q(\lambda_2-\lambda_1)}{\gamma}} e^{-qx_1^*} - \lambda_2(1 - e^{-qx_1^*}) - \gamma x_1^* = 0.$$

The left hand side of the equation is decreasing in x_1^*, equals to $\lambda_1 e^{-\frac{q(\lambda_2-\lambda_1)}{\gamma}} > 0$ if $x_1^* = 0$ and goes to $-\infty$ as $x_1^* \to \infty$. Hence, using the intermediate value theorem we conclude that (12) and (13) have a unique solution and $x^* = (x_1^*, x_2^*)$ is the unique fixed point of the system of equations (4)-(6).

When $\lambda_1 \neq \lambda_2$, $\bar{X}^s(t)$ solving (4)-(6) converges to x^* as $t \to \infty$, if we can find a Lyapunov function $V(x)$ with the following properties (see e.g. Strogatz [16]):

1. $V(x) > 0$ for all $x \neq x^*$ and $V(x^*) = 0$.
2. $\frac{dV(\bar{X}^s(t))}{dt} < 0$ for all $x \neq x^*$.

Without loss of generality, we assume that $\lambda_1 > \lambda_2$ and define $V(x) = \lambda_1 - \lambda_2 + \gamma(x_2 - x_1)$. Writing $V(x)$ as $V(x) = \lambda_1 e^{-qx_2} - \lambda_2(1 - e^{-qx_1}) - \gamma x_1 - (\lambda_2 e^{-qx_1} - \lambda_1(1 - e^{-qx_2}) - \gamma x_2)$, we have $V(x^*) = 0$ and $V(x) \neq 0$ for all $x \neq x^*$. Applying Corollary 7 we have $x_1 - x_2 < \frac{\lambda_1 - \lambda_2}{\gamma}$ and hence, $V(x) > 0$. The second condition follows as

$$\frac{dV(\bar{X}^s(t))}{dt} = \gamma\left(\frac{d\bar{X}_2^s(t)}{dt} - \frac{d\bar{X}_2^s(t)}{dt}\right) = \lambda_2 - \lambda_1 + \gamma(\bar{X}_1^s(t) - \bar{X}_2^s(t)) = -V(\bar{X}^s(t)) < 0.$$

Therefore, x^* is globally asymptotically stable: for all initial conditions, $\bar{X}^s(t) \to x^*$ as $t \to \infty$.

When $\lambda_1 = \lambda_2 = \lambda$, Corollary 7 implies that $\bar{X}_1^s(t) = \bar{X}_2^s(t)$. Denoting $\tilde{X}(t) = \bar{X}_1^s(t) = \bar{X}_2^s(t)$ and $\tilde{x}^* = x_1^* = x_2^*$ we need to show that $\tilde{X}(t) \to \tilde{x}^*, t \to \infty$, where $\tilde{X}(t)$ and $\tilde{x}^*$ satisfy the following equations:

$$\frac{\tilde{X}(t)}{dt} = 2\lambda e^{-q\tilde{X}(t)} - \lambda - \gamma\tilde{X}(t), \tag{14}$$

$$0 = 2\lambda e^{-q\tilde{x}^*} - \lambda - \gamma\tilde{x}^* \tag{15}$$

The righthand side of (15) is a decreasing function of $\tilde{x}^*$ and can be seen to have a unique solution. Equation (14) defines a gradient system with potential function $U(x) = \lambda x + \frac{1}{2}\gamma x^2 + \frac{2\lambda}{q}e^{-qx}$, i.e., it can be written as $\frac{\tilde{X}(t)}{dt} = - \bigtriangledown U(\tilde{X}(t))$ where $U(x)$ is a continuously differentiable, single valued scalar function. Hence, using Theorem 7.2.1 in Strogatz [16] $\tilde{X}(t) \to \tilde{x}^*, t \to \infty$. □

The fixed point x^* in Proposition 9 can be thought of as the long run average numbers of users in the system. Now, we analyze how x^* behaves for different values of the abandonment rate γ. It is reasonable to expect that x^* should decrease as abandonment rate increases, which always holds for the user class with the higher arrival rate. However, Proposition 10 shows that for the class with lower arrival rate x^* first increases and then decreases as γ increases.

Proposition 10. *Suppose $\lambda_1 \geq \lambda_2$. Then the long run average number of user-1 x_1^* decreases as the abandonment rate γ increases, while the long run average number of user-2 x_2^* increases when $\frac{\lambda_1-\lambda_2}{\gamma} > \frac{\gamma x_1^*}{q\lambda_1(1-e^{-qx_2^*})+q\gamma x_2^*}$ and decrease when $\frac{\lambda_1-\lambda_2}{\gamma} < \frac{\gamma x_1^*}{q\lambda_1(1-e^{-qx_2^*})+q\gamma x_2^*}$.*

Proof. Manipulating Equation (12) to obtain x_2^*, substituting in Equation (13) and doing cancellations, we get

$$\ln(\lambda_2(1 - e^{-qx_1^*}) + \gamma x_1^*) = -qx_1^* - \frac{q(\lambda_2 - \lambda_1)}{\gamma} + \ln \lambda_1 \tag{16}$$

Taking the implicit derivative of x_1^* with respect to γ, we obtain

$$x_1^* + \gamma \frac{dx_1^*}{d\gamma} + \frac{\gamma}{q}\frac{d}{d\gamma}[\ln(\lambda_2(1 - e^{-qx_1^*}) + \gamma x_1^*)] + \frac{\ln(\lambda_2(1 - e^{-qx_1^*}) + \gamma x_1^*)}{q} - \frac{\ln \lambda_1}{q} = 0. \tag{17}$$

Letting $D_1 = \lambda_2(1 - e^{-qx_1^*}) + \gamma x_1^*, D_2 = \gamma\lambda_2 + \gamma^2 x_1^* + \frac{\gamma^2}{q}$ and substituting Equation (16) into Equation (17) to get rid of the logarithm terms, we get

$$\frac{dx_1^*}{d\gamma} = \frac{D_1}{D_2}(\frac{\lambda_2 - \lambda_1}{\gamma} - \frac{\gamma x_1^*}{qD_1}) \tag{18}$$

Since D_1 and D_2 are always positive, when $\lambda_1 \geq \lambda_2$, the right hand side of Equation (18) is always negative, and hence as γ increases x_1^* increases. Interchanging x_1^* and λ_1 with x_2^* and λ_2 the right hand side of Equation (18) is positive when $\frac{\lambda_1-\lambda_2}{\gamma} > \frac{\gamma x_2^*}{q\lambda_1(1-e^{-qx_2^*})+q\gamma x_2^*}$ and negative when $\frac{\lambda_1-\lambda_2}{\gamma} < \frac{\gamma x_2^*}{q\lambda_1(1-e^{-qx_2^*})+q\gamma x_2^*}$. Hence, the conclusion for x_2^* follows. □

Proposition 10 shows that as γ increases, the limiting number of users for the class with lower arrival rate first increases and then decreases and the limiting number of users for the class with higher arrival rate decreases monotonically, which coincides with the observation in Figure 3. This behavior can be explained as follows. As the abandonment rate increases, users from both classes tend to abandon the system a lot faster and hence the arriving users from the class with lower arrival rate are less likely find a match. The decrease in the number of matches is higher than the increase in the abandonments and as a result we observe a certain level of accumulation in the limit for users from the class with lower arrival rates.

4.2 Diffusion Limits

Now, we move to the discussion on the diffusion limits when the matching probability and the abandonment rate are both scaled to study the fluctuations of the queue lengths around the fluid limit $\bar{X}^s(t)$. We define

$$\hat{X}_i^{s,n}(t) = \frac{X_i^{s,n}(nt) - \bar{X}_i^s(nt)}{\sqrt{n}}$$

To prove weak convergence we again use convergence of generators utilizing the techniques in [8].

Theorem 11. *Suppose $\bar{X}^s(t) = (\bar{X}_1^s(t), \bar{X}_2^s(t))'$ is the unique solution to the system of ODEs given by (4)-(6). Denote*

$$\begin{aligned}
a_1(t) &= -q\lambda_2 e^{-q\bar{X}_1^s(t)}, \\
a_2(t) &= -q\lambda_1 e^{-q\bar{X}_2^s(t)}, \\
\sigma_1(t) &= \sqrt{\lambda_1 e^{-q\bar{X}_2^s(t)} + \lambda_2(1 - e^{-q\bar{X}_1^s(t)}) + \gamma\bar{X}_1^s(t)}, \\
\sigma_2(t) &= \sqrt{\lambda_2 e^{-q\bar{X}_1^s(t)} + \lambda_1(1 - e^{-q\bar{X}_2^s(t)}) + \gamma\bar{X}_2^s(t)},
\end{aligned}$$

and further define

$$\begin{aligned}
z(t) &= \int_0^t e^{\gamma s}\sigma_2(s)dB_2(s) - \int e^{\gamma s}\sigma_1(s)dB_1(s), \\
v(t) &= \int_0^t -a_2(s)z(s)e^{(a_1(s)+a_2(s))s}ds + \int_0^t e^{(\gamma+a_1(s)+a_2(s))s}\sigma_1(s)dB_1(s).
\end{aligned}$$

Then we have $\hat{X}^{s,n}(t) \Rightarrow \hat{X}^s(t)$*, where* $\hat{X}^s = (\hat{X}_1^s(t), \hat{X}_2^s(t))$

$$\hat{X}_1^s(t) = e^{-\gamma t} e^{(-a_1(t)-a_2(t))} v(t) \tag{19}$$

$$\hat{X}_2^s(t) = e^{-\gamma t}(e^{(-a_1(t)-a_2(t))} v(t) + z(t)). \tag{20}$$

Proof. We first show that if $\hat{X}^s(t)$ is the unique solution to the stochastic differential equation

$$d\hat{X}^s(t) = \sigma(\bar{X}^s(t))dB_t + \nabla F(\bar{X}^s(t))\hat{X}^s(t)dt, \tag{21}$$

starting from $\hat{X}^s(0) = (0,0)^*$, where $B = (B_1, B_2)^*$ is a two-dimensional standard Brownian motion,

$$\nabla F(x) = \begin{pmatrix} \frac{\partial F_1}{x_1} & \frac{\partial F_1}{x_2} \\ \frac{\partial F_2}{x_1} & \frac{\partial F_2}{x_2} \end{pmatrix} = \begin{pmatrix} -q\lambda_2 e^{-qx_1} - \gamma & -q\lambda_1 e^{-qx_2} \\ -q\lambda_2 e^{-qx_1} & -q\lambda_1 e^{-qx_2} - \gamma \end{pmatrix},$$

$$\sigma(x) = \begin{pmatrix} \sqrt{\lambda_1 e^{-qx_2} + \lambda_2(1 - e^{-qx_1}) + \gamma x_1} & 0 \\ 0 & \sqrt{\lambda_2 e^{-qx_1} + \lambda_1(1 - e^{-qx_2}) + \gamma x_2} \end{pmatrix},$$

and $\bar{X}^s(t) = (\bar{X}_1^s(t), \bar{X}_2^s(t))'$ is the unique solution to system of ODE given by (4)-(6). Then, we have $\hat{X}^{s,n}(t) \Rightarrow \hat{X}^s$. Defining $\mathbb{S}$ as in the proof of Theorem 5, the weak convergence follows from Lemma 5.5 in [8], if we can show the conditions below hold:

(a) $F(x)$ is continuously differentiable on $\mathbb{S}$,

(b) $\sup_{x\in\mathbb{S}} \sqrt{n} \left| \frac{\partial m^n(x,\theta)}{\partial \theta}|_{\theta=0} - F(x) \right| \to 0$,

(c) $\frac{\partial^2 m(x,\theta)}{\partial \theta^2}|_{\theta=0}$ is Lipschitz continuous in x on $\mathbb{S}$, where $m(x,\theta)$ is defined by

$$m(x,\theta) = \lambda_1 e^{-qx_2} e^{\theta_1} + \lambda_2 e^{-qx_1} e^{\theta_2} + (\lambda_1(1 - e^{-qx_2}) + \gamma x_2) e^{-\theta_2} + (\lambda_2(1 - e^{-qx_1}) + \gamma x_1) e^{-\theta_1}.$$

Condition (a) is trivial and condition (b) reduces to showing $\sqrt{n}\left(\left(1 - \frac{q}{n}\right)^{nx} - e^{-qx}\right)$ converges

to 0, which is elementary calculus and hence (b) holds as well. Finally

$$\frac{\partial^2 m(x,\theta)}{\partial \theta^2}\Big|_{\theta=0} = \begin{pmatrix} \lambda_1 e^{-qx_2} + \lambda_2(1-e^{-qx_1}) + \gamma x_1 & 0 \\ 0 & \lambda_2 e^{-qx_1} + \lambda_1(1-e^{-qx_2}) + \gamma x_2 \end{pmatrix}$$

which is Lipschitz on $\mathbb{R}^2_{\geq 0}$. Using Lemma 5.5 in [8], $\hat{X}^n \Rightarrow \hat{X}^s$ as $n \to \infty$, where $\hat{X}^s(t)$ is the unique solution to the stochastic differential equation (21).

Next we show that (19) and (20) together is the unique solution to (21) which can be expressed as:

$$\begin{aligned} d\hat{X}^s(t)_1 &= (-a_1(t) - \gamma)\hat{X}^s_1(t)dt - a_2(t)\hat{X}^s_2(t)dt + \sigma_1(t)dB_1(t) \\ d\hat{X}^s(t)_2 &= -a_1(t)\hat{X}^s_1(t)dt - (a_2(t) - \gamma)\hat{X}^s_2(t)dt + \sigma_2(t)dB_2(t). \end{aligned}$$

Defining $z_i(t) = e^{\gamma t}\hat{X}^s_i(t), i = 1, 2$, we obtain

$$\begin{aligned} dz_1(t) &= e^{\gamma t}d\hat{X}^s_1(t) + \gamma e^{\gamma t}\hat{X}^s_1(t)dt \\ &= (-a_1(t) - \gamma)e^{\gamma t}\hat{X}^s_1(t)dt - e^{\gamma t}a_2(t)\hat{X}^s_2(t)dt + \gamma e^{\gamma t}\hat{X}^s_1(t)dt + e^{\gamma t}\sigma_1(t)dB_1(t) \\ &= -a_1(t)z_1(t)dt - a_2 z_2(t)dt + e^{\gamma t}\sigma_1(t)dB_1(t), \end{aligned} \tag{22}$$

and similarly $dz_2(t) = -a_1(t)z_1 dt - a_2(t)z_2(t)dt + e^{\gamma t}\sigma_2(t)dB_2(t)$. Furthermore, letting $z(t) = z_2(t) - z_1(t)$ we have

$$dz(t) = e^{\gamma t}(\sigma_2(t)dB_2(t) - \sigma_1(t)dB_1(t)). \tag{23}$$

Solving Equation (23) directly, we obtain that $z(t) = \int_0^t e^{\gamma s}\sigma_2(s)dB_2(s) - \int e^{\gamma s}\sigma_1(s)dB_1(s)$. Substituting that $z_2(t) = z(t) + z_1(t)$ into the Equation (22), we have

$$dz_1(t) = (-a_1(t) - a_2(t))z_1(t)dt - a_2(t)z(t)dt + e^{\gamma t}\sigma_1(t)dB_1(t).$$

Let $v(t) = e^{(a_1(t)+a_2(t))t} z_1(t)$ and we have

$$\begin{aligned} dv(t) &= (a_1(t) + a_2(t))e^{(a_1(t)+a_2(t))}tz_1(t)dt + e^{(a_1(t)+a_2(t))t}dz_1 \\ &= -a_2(t)z(t)e^{(a_1(t)+a_2(t))t}dt + e^{(\gamma+a_1(t)+a_2(t))t}\sigma_1(t)dB_1(t), \end{aligned}$$

which further implies that $v(t) = \int_0^t -a_2(s)z(s)e^{(a_1(s)+a_2(s))s}ds + \int_0^t e^{(\gamma+a_1(s)+a_2(s))s}\sigma_1(s)dB_1(s)$. Hence,

$$\begin{aligned} \hat{X}_1^s(t) &= e^{-\gamma t}z_1(t) = e^{-\gamma t}e^{(-a_1(t)-a_2(t))}v(t) \\ \hat{X}_2^s(t) &= e^{-\gamma t}(z_1(t) + z(t)) = e^{-\gamma t}(e^{(-a_1(t)-a_2(t))}v_1(t) + z(t)), \end{aligned}$$

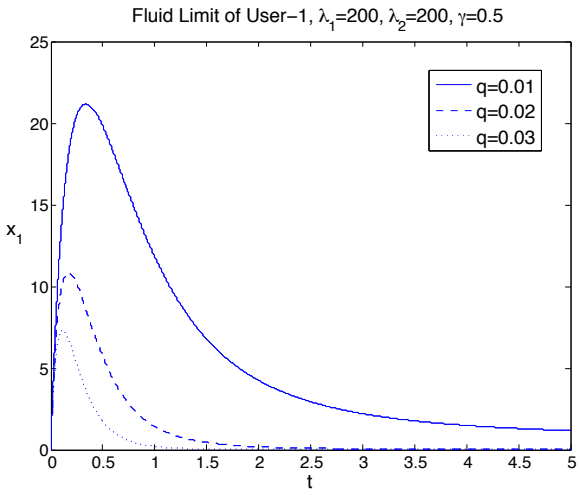
as desired. □

Theorem 11 indicates that if the fluid limit $\bar{X}^s(t)$ is given the diffusion limit can be fully characterized analytically. However, as we have seen in Section 4.1, it is not always possible to analytically solve the ODEs for the fluid limit. In the next section, we present numerical experiments to study fluid and diffusion limits presented in this section.

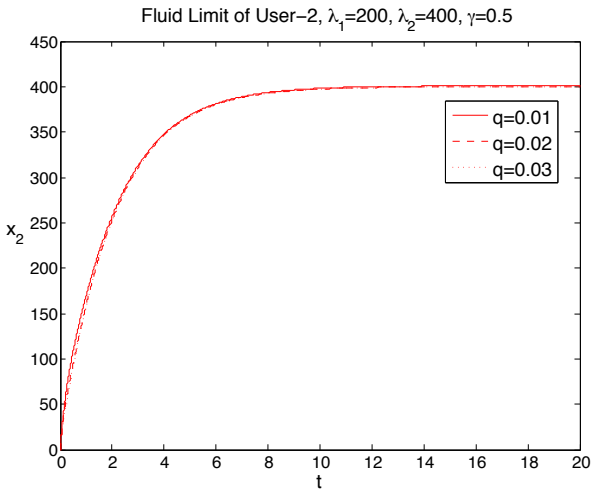
5 Numerical Experiments

In Section 4, we show that when the matching probability and abandonment rate are scaled to go to zero along with the time and space, the fluid and diffusion limits can be expressed as the unique solutions to some systems of ODEs and SDEs which do not have explicit solutions in general. To gain some insight into the solutions, we study numerical approximations in this section. We use Euler and Euler-Maruyama method to obtain numerical solutions of ODEs (4)-(6) and SDEs (21), respectively. (See Kloeden and Platen [13] for more details.)

To study the fluid limit which is the unique solution to the system of ODEs(4)-(6), we apply Euler method with step size $h = 10^{-6}$. In our first experiment, we test the effect of the matching probability q on the fluid limits. First we consider the case $\lambda_1 < \lambda_2$ by setting $\lambda_1 = 200, \lambda_2 = 400, \gamma = 0.5$ and compute the fluid limits for $q = 0.01, 0.02, 0.03$. The results are given in Figure 1.



(a) Fluid Limit of User 1 for different q



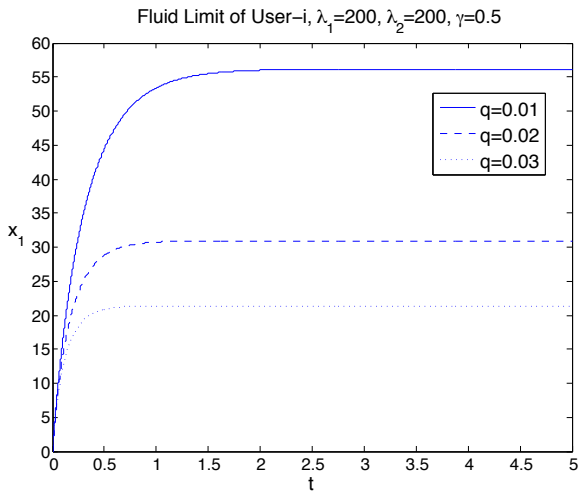
(b) Fluid Limit of User 2 for different q

Figure 1: Fluid Limits when $\lambda_1 < \lambda_2$ for various q

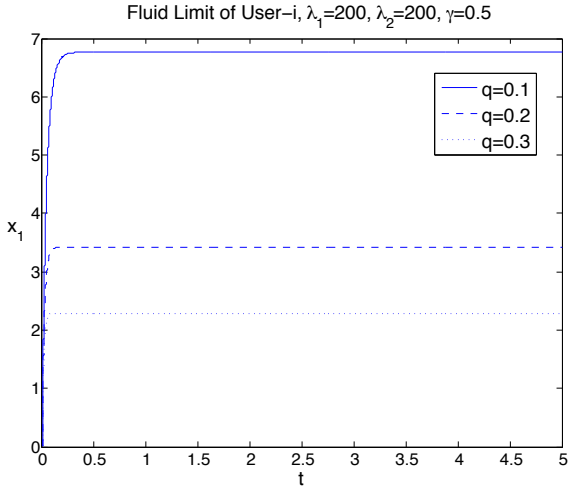
We observe that for the class with lower arrival rate, the number of users in the system demonstrates a very sharp increase at the beginning and then decreases approaching a limit as t goes to infinity. We see that there is a considerable difference between the number of users corresponding to different matcing probabilities for this class. On the other hand, the number of users for the class with higher arrival rate grows monotonically converging to its supremum as t goes to infinity. Surprisingly, the matching probability does not play a significant role for this class and the fluid limits corresponding to different matching probabilities are very close.

To test the case where $\lambda_1 = \lambda_2$, we performed the same experiment by taking $\lambda_1 = \lambda_2 = 200$. Figure 5 demonstrates that the number of users for both classes increase monotonically as t goes to infinity approaching to the supremum, which is very similar to the behavior of the class with higher arrival rate when the rates are not equal. However, in this case the matching probability has a major effect on the limiting number of users and as q increases the number of users in the system decreases. Also as q gets larger we see that the number of users increases to its supremum faster and the fluid limit is steeper.

Next we study how the effect of the abandonment rate γ on the number of users in the system. In this set of experiments, we set the arrival rates $\lambda_1 = 200, \lambda_2 = 400$ and the matching probability $q = 0.01$ and vary the abandonment rate. Figure 3 shows that the shape of fluid limits are not



(a) Fluid Limit of User 1 and 2 for different q

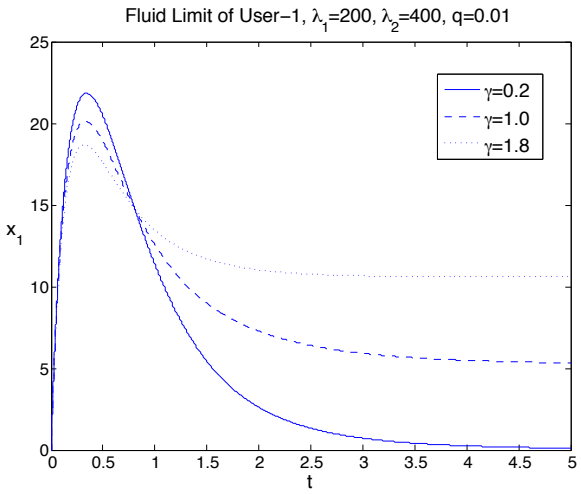


(b) Fluid Limit of User 1 and 2 for different q

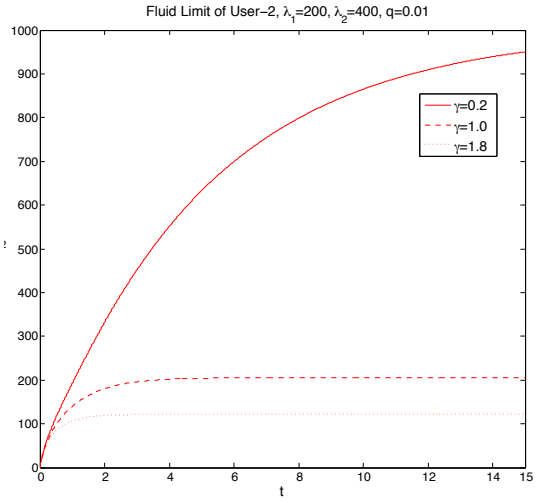
Figure 2: Fluid Limits when $\lambda_1 = \lambda_2$ for various q

affected by the changes in the abandonment rate, i.e., the number of users for the class with lower arrival rate first increases and then decreases and the number of users for the class with higher arrival rate decreases monotonically. We also see that when there are abandonments the number of users for the class with lower arrival rate does not converge to 0 as t goes to infinity. In agreement with Proposition 10, we see that the limiting number of users for the class with lower arrival rate increases in our experiments as the abandonment rate increases.

Now, we discuss numerical approximation to diffusion limit, which is the unique solution to the system of SDEs (21). In our experiments, we apply the Euler- Maruyama method with the step size $h = 10^{-6}$. We again start with the case when the arrival rates are not equal and set $\lambda_1 = 200, \lambda_2 = 400$. Figures 4 and 5 demonstrate some sample paths. We see that the fluctuations for the class with higher arrival rates are always bigger. When q is fixed we see that the changes in γ does not have a major effect on fluctuations. We also see that the fluctuations for the class with lower arrival rate diminish as t increases. As q increases the fluctuations diminish a lot faster.

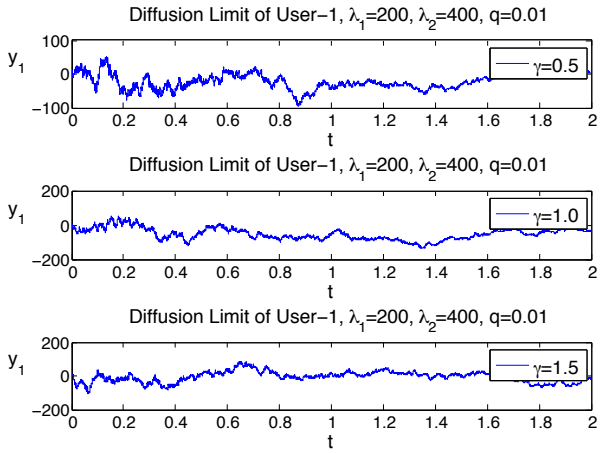


(a) Fluid Limit of User 1 for various γ

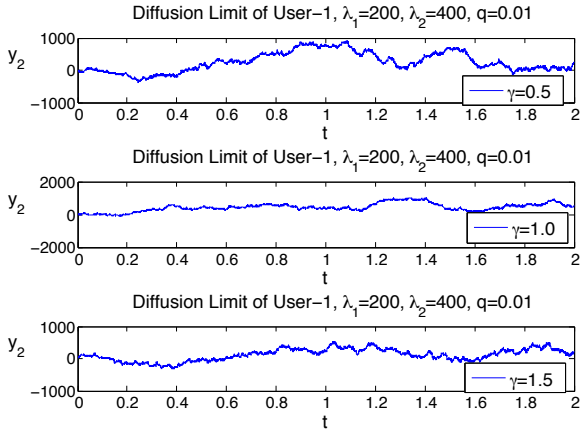


(b) Fluid Limit of User 2 for various γ

Figure 3: Fluid Limits when $\lambda_1 < \lambda_2$ for various γ

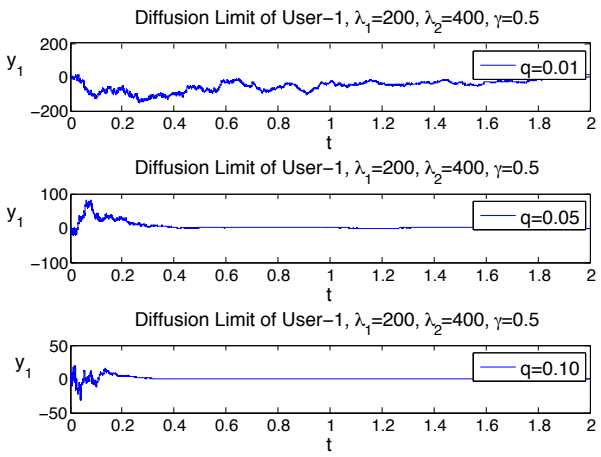


(a) Diffusion Limit of User 1 for various γ

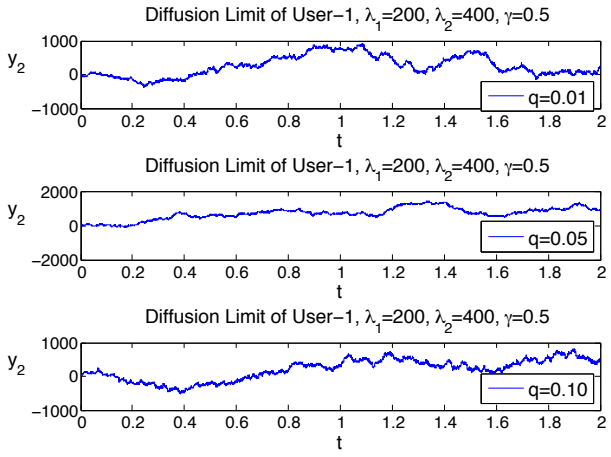


(b) Diffusion Limit of User 2 for various γ

Figure 4: Diffusion Limits when $\lambda_1 < \lambda_2$ for various γ

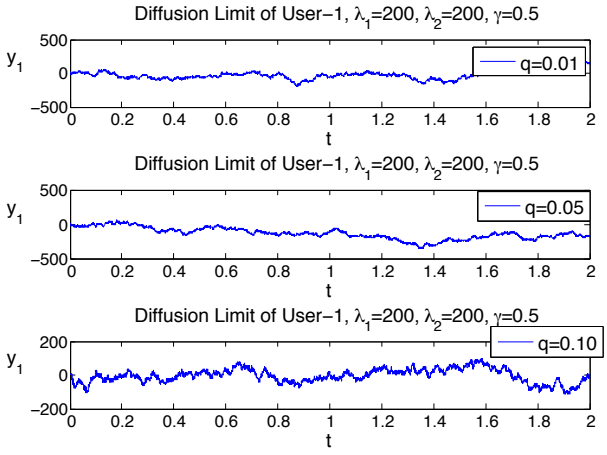


(a) Diffusion Limit of User 1 for various q

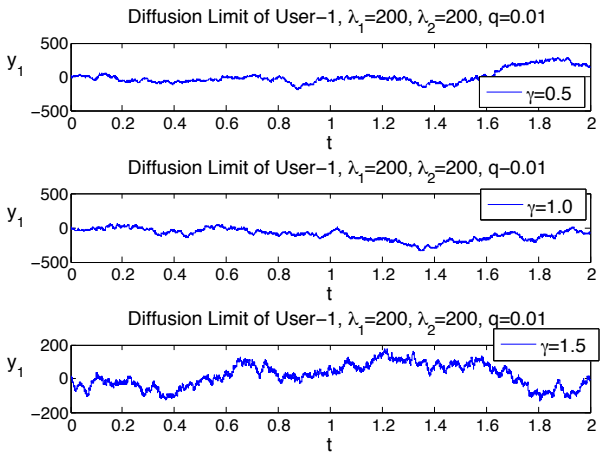


(b) Diffusion Limit of User 2 for various q

Figure 5: Diffusion Limits when $\lambda_1 < \lambda_2$ for various q



(a) Diffusion Limit of User 1 for various q



(b) Diffusion Limit of User 1 for various q

Figure 6: Diffusion Limits when $\lambda_1 = \lambda_2$ for various q and γ

6 Conclusion and Future Work

In this work, we propose two different scalings to obtain fluid and diffusion approximations to the queue length processes of probabilistic matching systems. For the first approach, the space and time are scaled while the matching probability is kept fixed. Under this scaling, the matching probability q does not play any role in the fluid limit and the minimum of the queue lengths converges to zero. We suggest that this scaling is used when the matching probability is considerably high.

The second scaling considers the systems in which the probability to match for each pair of users is small. The effect of abandonments is also taken into account and the matching probability and the departure rate are scaled along with time and space in this regime. The limiting processes enable us to address the matching probability explicitly. Unfortunately, the resulting system of ODE cannot be solved analytically in general, although, when there are no abandonments it is possible to obtain an analytical solution. In [3], some performance measures are shown to be insensitive to the matching probability under certain admission control policies. Using fluid limits, we show that the difference between the average queue lengths of different classes of users is also independent of the matching probability. We also analyze the asymptotic behaviour of the fluid limits in this scaling. First we show that when abandonment rate is zero, the two fluid limits, obtained with and without scaling the matching probability, converges to each other with time. We further show that when there are abandonments, the fluid limits converge to a unique fixed point, which represents the long run average number of users in the system. Conducting analysis on the fixed point, we reveal that as the abandonment rate increases, the number of users for the class with lower arrival rate first experiences an increase and then decrease while the number of users for the class with higher arrival rate decreases monotonically.

As analytical expressions are not available for fluid and diffusion limits, we resort to numerical methods to study the corresponding ODEs and SDEs. We see that for the class with higher arrival rate, the number of users in the system increases monotonically. On the other hand, the users from the class with lower arrival rate first tend to accumulate in the system and then decrease to a limit as time goes to infinity. This limit is different from zero and increases as the abandonment rate increases agreeing with our theoretical analysis. This indicates that there are always a significant

number of users waiting in the system from both classes.

The probabilistic matching systems exhibit many interesting properties and we believe the fluid and diffusion limits introduced in this work will be helpful in many directions. First, the approximations introduced here can be used to study the performance of admission control policies which are intractable using exact methods. Another promising research direction is to identify optimal and asymptotically optimal policies to maximize profit generated by charging users admission fees. The probabilistic matching systems studied in this work can also be extended to include different types of users within each class where each type has a different probability to match with users of other class.

References

[1] I. Adan and G. Weiss. Exact FCFS matching rates for two infinite muti-type sequences. *Operations Research*, 60(2):475–489, 2012.

[2] P. Billingsley. *Convergence of Probability Measures*. Wiley, New York, 1999.

[3] B. Büke and H. Chen. Stabilizing policies for probabilistic matching systems. *Submitted to Queueing Systems*, 2014.

[4] A. Bušić, V. Gupta, and J. Mairesse. Stability of the bipartite matching model. *Advances in Applied Probability*, 45(2):351–378, 2013.

[5] R. Caldentey, E. Kaplan, and G. Weiss. FCFS infinite bipartite matching of servers and customers. *Advances in Applied Probability*, 41(3):695–730, 2009.

[6] H. Chen and D. D. Yao. *Fundamentals of Queuing Networks, Performance, Asymptotics, and Optimization*. Springer, New York, 2001.

[7] J. G. Dai and S. He. Customer abandonment in many-server queues. *Mathematics of Operations Research*, 35(2):347–362, 2010.

[8] R. W. R. Darling and J. R. Norris. Structure of large random hypergraphs. *The Annals of Applied Probability*, 15(1A):125–152, 2005.

[9] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Operations Research*, 4(3):208–227, 2002.

[10] I. Gurvich and A. R. Ward. On the dynamic control of matching queues. *Stochastic Systems*, 4:1–45, 2014. Working paper.

[11] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

[12] B. R. K. Kashyap. The double-ended queue with bulk service and limited waiting space. *Operations Research*, 14(5):822–834, 1966.

[13] P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, 1999.

[14] X. Liu, Q. Gong, and V. G. Kulkarni. Diffusion models for double-ended queues with renewal arrival processes. 2014. Working Paper, arXiv:1401.5146.

[15] A. Mandelbaumi and P. Momčilović. Queues with many servers and impatient customers. *Mathematics of Operations Research*, 37(1):41–65, 2012.

[16] S. H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering (Studies in Nonlinearity)*. Westview Press, 1994.

[17] A. R. Ward and P. W. Glynn. A diffusion approximation for a markovian queue with reneging. *Queueing Systems*, 43(1/2):103–128, 2003.

[18] A. R. Ward and P. W. Glynn. A diffusion approximation for a $GI/GI/1$ queue with balking or reneging. *Queueing Systems*, 50(4):371–400, 2005.

[19] W. Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and their Application to Queues*. Springer-Verlag, Florham Park, NJ, USA, 2001.