Asymptotic Behavior of Heterogeneous TCP Flows and RED Gateway

by Peerapol Tinnakornsrisuphap, and Richard J. La

TR 2003-28

Asymptotic Behavior of Heterogeneous TCP Flows and RED Gateway

Peerapol Tinnakornsrisuphap and Richard J. La

Department of Electrical and Computer Engineering
and Institute for Systems Research,
University of Maryland, College Park
{peerapol,hyongla}@eng.umd.edu

August 4, 2003

Abstract

We introduce a stochastic model of a bottleneck ECN/RED gateway under a large number of heterogeneous TCP flows, *i.e.,* flows with diverse round-trips and session dynamics. We investigate the asymptotic behavior of the system and show that as the number of flows becomes large, the buffer dynamics and aggregate traffic simplify and can be accurately described by simple stochastic recursions independent of the number of flows, resulting in a scalable model. Based on the central limit theorem results presented in the paper we analyze the sources of fluctuations in queue size and describe the relationship between the packet marking function and variance of queue size.

1 INTRODUCTION

With the growing size and popularity of the Internet, there has been increasing interest in modeling and understanding Internet traffic. Accurate modeling of Internet traffic is also important from the perspective of deploying differentiated services since it is likely that best-effort traffic will comprise a significant portion of the Internet traffic in the foreseeable future.

Today Internet traffic consists of many heterogeneous traffic sources, the majority of which utilize the TCP congestion control mechanism [5]. Some applications, such as FTP, are relatively long-lived, while others are typically short-lived, *e.g.,* Web browsing. Characterizing and modeling TCP traffic yield an understanding of the interactions between the transport layer (TCP) and the network layer.

To this end researchers have proposed a number of different models - from detailed single flow models to predict the throughput of a single flow as a function of round-trip delay and packet loss probability [9] to linearized fluid models motivated by control theoretic point of view [3]. As a result, the behavior of a single long-lived TCP flow is relatively well understood. However, despite these efforts there remain several aspects of TCP that are not well understood, especially in the context of designing a proper active queue management (AQM) mechanisms that will interact with many TCP flows.

The introduction of AQM mechanisms adds additional complexity to accurate modeling of the interaction between TCP flows with network layer. The problem is further compounded by the heterogenous nature of flows, *i.e.,* different round-trip delays. In addition, much of modeling emphasis in the past was placed on understanding the behavior of long-lived TCP flows [9] and the role of session dynamics has been largely ignored in modeling the interaction of TCP flows with the network layer, *e.g.,* drop-tail gateways and AQM mechanisms.

Accurate modeling of individual TCP flows requires modeling of complex dynamics rising from additive-increase-multiplicative-decrease mechanism, session dynamics, and heterogeneous round-trip delays in conjunction with the underlying network layer. As the size of state space explodes with the number of sessions, this represents a major obstacle to modeling the interaction of many TCP flows in a realistic setting. For the same reason even numerical experiments become computationally prohibitive, and fail to provide an insight into the complex dynamics.

The existing literature on TCP traffic modeling usually skirts these major obstacles by relying on ad-hoc assumptions, which causes the model to be accurate only in certain regimes. Hollot *et al.* model short-lived TCP flows as exponential pulses that occur according to a Poisson process [4], and characterize their TCP windows through *shot noise* processes. This model implicitly assumes relatively low congestion levels in that short-lived flows last only a few round-trip times and do not experience packet drops or marks. Moreover, flows are always in either congestion avoidance (long-lived connections) or slow start (short-lived connections), and do not transit from one state to the other. In other words, the session dynamics are not modeled *explicitly* with connections arriving and leaving the network after transfers are completed. A similar approach to modeling short-lived flows is also taken in [8].

At the other end of the spectrum, Kherani and Kumar [7] suggest that as the capacity at the bottleneck queue serving TCP flows with *homogeneous* round-trip times (RTTs) becomes very small, this queue can be accurately described as a processor sharing queue. When the capacity is large, however, this processor sharing model becomes less accurate as newly arrived TCP flows cannot fully utilize their allocated bandwidth. In fact, in the large capacity regime these short-lived flows may terminate even before they can increase their transmission rates to fully utilize their allocated bandwidth due to slow start.

The shortcomings of these models suggest a need for a *unified* model that is accurate in *all* regimes, instead of being restricted to a specific regime. Recently, there has been an increasing interest in *macroscale* modeling of TCP flows. Unlike in microscale models where each individual TCP flow is modeled in details, macroscale models can be developed by systematically applying limit theorems to derive limiting traffic models. Since the number of connections that share a bottleneck link inside a network is likely to be large, especially in the core network or over a trans-continental link, such a macroscale model promises a potential to provide an accurate and yet scalable model without having to rely on any ad hoc assumptions. The potential benefits of such a model are three-fold. First, model simplifications (with the promise of scalability) typically occur when applying limit theorems, with irrelevant details filtered out. Second, due to a large number of results and techniques available on limit theorems in literature it is reasonable to expect the existence of suitable limit theorems (under very weak assumptions) which can be applied to the situation of interest. Finally, in the networking context, resource allocation problems are of the greatest interest in networks operating at high utilization, *e.g.,* when the number of users is large. In such a scenario, the limit behavior will become increasingly accurate as the number of users increases.

In this paper we build upon the model in [12] by incorporating (i) heterogeneous RTTs of TCP flows and (ii) an additional layer of dynamics, namely the session layer, with TCP (transport layer) and RED gateway [2] with ECN [1] capability (network layer). The model adopted in this paper is considerably more detailed and faithful than that in [12], and as such requires a more complicated analysis and provides deeper insight into the dynamics of the system. Using this detailed model, we establish various asymptotic and central limit theorem type results. They reveal several interesting observations on the dynamics of the system, which are summarized in Section 2.

The paper is organized as follows: An overview of the results in the paper is presented in Section 2. The model and the dynamics of network, transport, and session layers are described in Section 3. The asymptotic results are presented in Section 4, followed by central limit theorem results in Section 5. Section 6 gives a brief discussion of the results and a comparison with previously proposed models. Section 7 concludes the paper with suggestions for future work.

Some words on the notation in use: Equivalence in law or in distribution between random variables (rvs) is denoted by $=_{st}$. The indicator function of an event A is given by $\mathbf{1}\left[A\right]$, and we use $\xrightarrow{P}_n$ (resp. $\Longrightarrow_n$) to denote convergence in probability (resp. weak convergence or convergence in distribution) with n going to infinity. For scalars a and b we write $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$. We write $X^{(N)}$ to indicate the explicit dependence of the quantity X on the number N of sessions. An expectation of a rv X with a distribution function F is given by either $\mathbf{E}\left[X\right]$ or $\mathbf{E}\left[F\right]$. For simplicity, we introduce the notation $\mathbf{1}_X\left[x\right]$ and $\mathbf{P}_X\left[x\right]$ for $\mathbf{1}\left[X = x\right]$ and $\mathbf{P}\left[X = x\right]$, respectively.

2 Contributions

In this section we summarize the main results presented in this paper, using a detailed stochastic model of TCP flows with an RED gateway described in Section 3.

(i) We prove that as the number of sessions increases the queue size per session and the per-session workload brought in during a round-trip time converge to deterministic processes (a Law of Large Numbers or LLN). Moreover, the sessions become asymptotically independent, suggesting that the RED gateway does alleviate the synchronization problem among the connections observed with drop-tail gateways. The limiting model is shown to agree with the previously proposed models in [4] and [7] in their respective regimes, *i.e.,* when the capacity is very large or small.

(ii) We sharpen the LLN result with a Central Limit Theorem (CLT). A distributional recursion for buffer size is provided, which can be used for network provisioning/dimensioning. A closer inspection at the CLT results reveals that the sources of queue size fluctuation can be decomposed into (a) protocol structure, (b) fluctuation in the feedback information, (c) binary nature of the feedback information, and (d) variation in the RTTs and file size variation.

3 The Model

For each $N \in \{1, 2, 3, \cdots\}$ let $\mathcal{N} = \{1, \cdots, N\}$ be the set of sessions that share a bottleneck RED gateway. We assume that time is slotted into contiguous timeslots. Here the RTTs of TCP connections are approximated as integer multiples of timeslots, and a timeslot is the greatest common divisor of the RTTs of TCP flows. For our analysis we model three layers of dynamics - network, transport, and session layers - which interact with each other through mechanisms that will be specified shortly. At the lowest level, the network is simplified to be a single bottleneck router with an ECN/RED marking mechanism controlling the congestion level. The traffic injected into the network is controlled by TCP congestion control mechanism at the transport layer, which reacts to the marks from the network. Each TCP connection is initiated by a session. A session can be either active or idle. If a session is active, a file or an object is transferred through a TCP connection. A busy period of a session is defined to be the period from the time when the session receives a file to transfer till the time at which the TCP connection is torn down after completion of file transfer and the session goes idle. The duration of an idle period is random and represents the idle time between consecutive file transmissions. When a new file/object to be transferred arrives, the session becomes active again and sets up a new TCP connection. We now give detailed descriptions of the model for each layer and the interaction of these three layers.

3.1 Heterogeneous Round-trip Times

As mentioned earlier we approximate the RTTs of TCP connections as integer multiples of timeslots, and any congestion-control actions by TCP flows, *i.e.*, additive increase and multiplicative decrease, occur at the end of round-trip. The RTT of an active flow i at time t is denoted by $d_i^{(N)}(t) \in \mathcal{H} := \{2, 3, \cdots, D_{max}\}$.[1] We use $\beta_i^{(N)}(t+1)$ to denote the number of timeslots since the last action by an *active* flow i. Then, $\beta_i^{(N)}(t)$ evolves according to

$$\beta_i^{(N)}(t+1) = \left(1 + \beta_i^{(N)}(t)\mathbf{1}\left[\beta_i^{(N)}(t) < d_i^{(N)}(t)\right]\right)\mathbf{1}\left[X_i^{(N)}(t) > 0\right], \tag{1}$$

where $X_i^{(N)}(t)$ is the remaining workload (in packets) of the connection of session i at the beginning of timeslot $[t, t+1)$. [2] The rv $X_i^{(N)}(t) > 0$ only if session i is active in timeslot $[t, t+1)$. Hence, the last indicator function is one only if the connection is active. This will be explained further in the next subsection.

We use the following mapping $G_{i,t} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ to simplify our notation later.

$$G_{i,s}(a, b) = a \cdot \mathbf{1}\left[\beta_i^{(N)}(s) < d_i^{(N)}(s)\right] + b \cdot \mathbf{1}\left[\beta_i^{(N)}(s) \geq d_i^{(N)}(s)\right] \tag{2}$$

If we let $Y_i^{(N)}(t+1) = G_{i,t+1}(Y_i^{(N)}(t), Y_i^{new})$, then the values of $Y_i^{(N)}(t+1)$ will be updated to Y_i^{new} only at the end of round-trip, *i.e.*, $\beta_i^{(N)}(t+1) \geq d_i^{(N)}(t+1)$. Otherwise, $Y_i^{(N)}(t+1) = Y_i^{(N)}(t)$ since no action will be taken before the end of round-trip.

3.2 Session Dynamics

Each session $i \in \mathcal{N}$ is either active or idle. An idle session at the beginning of timeslot $[t, t+1)$ does not have any packets to transmit in the timeslot. An idle session in timeslot $[t, t+1)$ becomes active at the beginning of timeslot $[t+1, t+2)$ with probability $P_{ar} \in (0, 1)$ independently of the past. In other words, the duration of an idle period is geometrically distributed with parameter P_{ar} and has a mean of $1/P_{ar}$. This attempts to capture the dynamics of connection arrivals, where the interarrival times are reported to be exponentially distributed [10]. When N is large, the arrival rate of a new connection will be steady and the interarrival times will be given by a geometric rv.[3] Let $\{U_i(t), i \in \mathcal{N}; t = 0, 1, \cdots\}$ be a collection of i.i.d. rvs uniformly distributed on [0, 1] and $\mathbf{1}[U_i(t+1) < P_{ar}]$ be the indicator function of the event that a new file/object arrives in the timeslot $[t+1, t+2)$ for an idle session i.

Let $\{F_i(t), i \in \mathcal{N}; t = 0, 1, \cdots\}$ be a collection of i.i.d. non-negative integer-valued rvs with a general distribution function F. The workload of a connection of session i that becomes active at the beginning of timeslot $[t, t+1)$ is given by $F_i(t)$. This workload represents the *total* amount of workload a TCP connection brings in before it is torn down rather than workload brought in

[1] Although $\mathcal{H}$ does not include 1 in our model, it can be included at the price of more cumbersome proofs. Moreover, this does not cause any loss of generality of the model.

[2] We refer to a TCP connection of an active session i as connection i when there is no confusion.

[3] Recall that one can approximate an exponential rv X with parameter α with $\lceil X \rceil$, which is a geometric rv with parameter $p = 1 - e^{-\alpha}$.

by an object or a file. In other words, if the same TCP connection is used to transfer more than one object while it is alive, $F_i(t)$ represents the total amount of workload brought in by all objects during the duration of the TCP connection. The evolution of $X_i(t)$, which denotes the remaining workload, is given by the following:

$$\begin{aligned} X_i^{(N)}(t+1) = \mathbf{1}\left[X_i^{(N)}(t) > 0\right]\left(X_i^{(N)}(t) - A_i^{(N)}(t)\right) \\ +\mathbf{1}\left[X_i^{(N)}(t) = 0\right]\mathbf{1}\left[U_i(t+1) < P_{ar}\right]F_i(t+1)\ , \end{aligned} \tag{3}$$

where $A_i^{(N)}(t)$ denotes the number of packets transmitted by connection i at the beginning of timeslot $[t, t+1)$. This will be explained in the following subsection.

When a new connection arrives, its RTT is randomly selected, and the RTT of session i in timeslot $[t+1, t+2)$ is given by

$$d_i^{(N)}(t+1) = d_i^{(N)}(t)\mathbf{1}\left[X_i^{(N)}(t) > 0\right] + \mathbf{1}\left[X_i^{(N)}(t) = 0\right]\mathbf{1}\left[U_i(t+1) < P_{ar}\right]H_i(t+1)\ , \tag{4}$$

where $H_i(t+1)$'s are i.i.d. rvs that take values in $\mathcal{H}$ and determine the RTTs of newly arrived connections. The bound on the maximum RTT does not constrain the model because actual TCP flows also cannot have RTTs larger than the timeout value.

3.3 TCP Dynamics

For each $i \in \mathcal{N}$, let $W_i^{(N)}(t)$ be an integer-valued rv that encodes the congestion window size (in packets) at the beginning of timeslot $[t, t+1)$. We assume that the range of rv $W_i^{(N)}(t)$ is $\{0, 1, \cdots, W_{\max}\}$, where $W_{\max}$ is a finite integer representing the receiver advertised window size of the TCP connection. We assume that the congestion window size of an idle session is zero. When an idle session becomes active at the beginning of timeslot $[t, t+1)$, the congestion window size of TCP connection is set to one at the end of the first round-trip, *i.e.,* in timeslot $[t+d_i^{(N)}(t), t+d_i^{(N)}(t)+1)$, allowing a transmission of one packet. This models one RTT delay for three-way handshake. Here we describe how the congestion window sizes of active connections evolve.

Each TCP source transmits as many of the remaining data packets as allowed by its congestion window only at the end of each round-trip. We simplify the packet transmission in a round-trip and assume that the packets from a connection all arrive only in a single timeslot, rather than being spread out throughout the round-trip. Such simplification can be justified by the following:

(i) In the Internet, most of the packet arrivals at a bottleneck are usually compressed together due to the "ACK compression" phenomenon [13], which leads to bursty arrivals at the bottlenecks. Hence, modeling the packet arrivals over an RTT as a batch arrival in a single timeslot tends to be more accurate than modeling them as smooth arrivals throughout the RTT.

(ii) Aggregating a round-trip worth of packet arrivals into a single timeslot will result in burstier traffic from each flow. This will cause queue dynamics to fluctuate more than having a smooth arrival pattern. Therefore, the queue fluctuation in this model will provide an upper bound to an actual queue with smoother packet arrival patterns.

(iii) The information used for control action at the RED gateways is the average queue size. With the averaging mechanism with a long memory as in RED, the difference in the queue size due to our bursty packet arrivals will be smoothed out by the averaging mechanism of RED.

Suppose that connection i has $X_i^{(N)}(t)$ remaining packets (or workload) waiting to be transmitted at the beginning of timeslot $[t, t+1)$. The number of packets connection i transmits at the beginning of timeslot $[t, t+1)$, denoted by $A_i^{(N)}(t)$, is given by

$$A_i^{(N)}(t) = \min\left(W_i^{(N)}(t), X_i^{(N)}(t)\right) \mathbf{1}\left[\beta_i^{(N)}(t) \geq d_i^{(N)}(t)\right]. \tag{5}$$

Note from (5) that a connection transmits once per RTT.

The congestion control mechanism of TCP operates in two different modes: slow start (SS) and congestion avoidance (CA). A new TCP connection starts in SS. In SS, the congestion window size is doubled every RTT until one or more packets are marked. If a mark is received, then the congestion window size is halved and TCP switches to CA. The congestion window size is limited by the receiver advertised window size $W_{\max}$. Hence, the evolution of the congestion window of connection i in SS can be written as

$$\begin{aligned} W_{i,SS}^{(N)}(t+1) = G_{i,t+1}[W_i^{(N)}(t), &\min\left(2W_i^{(N)}(t) \vee 1, W_{\max}\right) M_i^{(N)}(t+1) \\ &+ \min\left(\lceil \frac{W_i^{(N)}(t)}{2} \rceil, W_{\max}\right)(1 - M_i^{(N)}(t+1))]\ , \end{aligned} \tag{6}$$

where $M_i^{(N)}(t+1)$ is an indicator function of the event that no marks have been received in the round-trip preceding the timeslot $[t, t+1)$, *i.e.*, $M_i^{(N)}(t+1) = 1$ when no packet from connection i is marked in the previous round-trip and $M_i^{(N)}(t+1) = 0$ when at least one packet is marked. The marking mechanism will be explained in more details in Subsection 3.4. From the definition of mapping in (2), the window size is updated only once per RTT. Throughout the paper we assume that window sizes are updated at the beginning of timeslot.

In CA, the congestion window size in the next timeslot is increased by 1 if no marks are received in a round-trip preceding the timeslot $[t, t+1)$, and if one or more packets are marked the congestion window is reduced by half. The congestion window size in CA can be described by the following:

$$\begin{aligned} W_{i,CA}^{(N)}(t+1) = G_{i,t+1}[W_i^{(N)}(t), &\min(W_i^{(N)}(t) + 1, W_{\max}) M_i^{(N)}(t+1) \\ &+ \min\left(\lceil \frac{W_i^{(N)}(t)}{2} \rceil, W_{\max}\right)(1 - M_i^{(N)}(t+1))]. \end{aligned} \tag{7}$$

We use $\{0,1\}$-valued rvs $\{S_i^{(N)}(t), i \in \mathcal{N}\}$ to encode the state of TCP connections. We interpret $S_i^{(N)}(t) = 0$ (resp. $S_i^{(N)}(t) = 1$) as connection i being in CA (resp. in SS) at the beginning of the timeslot $[t, t+1)$. Therefore, the complete recursion of the congestion window size can be written as

$$\begin{aligned} W_i^{(N)}(t+1) = \mathbf{1}&\left[X_i^{(N)}(t) - A_i^{(N)}(t) > 0\right] \\ &\times [S_i^{(N)}(t) W_{i,SS}(t+1) + (1 - S_i^{(N)}(t)) W_{i,CA}(t+1)]\ , \end{aligned} \tag{8}$$

where the first indicator function is used to reset the congestion window size to zero when Session i runs out of data to transmit and returns to its idle state.

Finally, the evolution of $S_i^{(N)}(t)$ is given by

$$\begin{aligned} S_i^{(N)}(t+1) = \mathbf{1}&\left[X_i^{(N)}(t) - A_i^{(N)}(t) \leq 0\right] \\ &+ \mathbf{1}\left[X_i^{(N)}(t) - A_i^{(N)}(t) > 0\right] S_i^{(N)}(t) M_i^{(N)}(t+1). \end{aligned} \tag{9}$$

This equation can be interpreted as follows. Connection i is in SS in timeslot $[t+1, t+2)$ if either (1) there is no packet left to transmit (so the connection resets) at the beginning of the timeslot or (2) the connection was active and in SS in timeslot $[t, t+1)$ and received no mark in the timeslot. From (9), we assume that a new TCP connection in SS is ready to be set up one timeslot after the previous connection is torn down after finishing its workload, and the new TCP connection becomes active when a new file/object arrives initiating a three-way handshake. We also assume that the SS/CA state is updated in the next timeslot following a transmission. However, the window size is updated one RTT after the transmission using the appropriate SS/CA state as in the correct operation of TCP.

3.4 Network Dynamics

In this subsection we explain how packets are marked to provide the congestion notification to the active TCP connections. The capacity of the bottleneck link is NC packets/slot for some positive constant C. The buffer size is assumed to be infinite so that no packets are dropped due to buffer overflow. Thus, congestion control is achieved solely through the random marking algorithm of the RED gateway.

Let $Q^{(N)}(t)$ denote the number of packets queued in the buffer at the beginning of timeslot $[t, t+1)$. Connection i injects $A_i^{(N)}(t)$ packets into the network, and they are put in the buffer at the beginning of timeslot $[t, t+1)$. Let the rv $A^{(N)}(t) := \sum_{i=1}^{N} A_i^{(N)}(t)$ denote the aggregate number of packets offered to the network by the N sessions at the beginning of timeslot $[t, t+1)$. Hence, $Q^{(N)}(t) + A^{(N)}(t)$ packets are available for transmission during that timeslot. Since the bottleneck link has a capacity of NC packets/timeslot, $\left[Q^{(N)}(t) + A^{(N)}(t) - NC\right]^+$ packets will not be served during timeslot $[t, t+1)$, and will remain in the buffer. Hence, their transmission is deferred to subsequent timeslots. The number of packets in the buffer at the beginning of timeslot $[t+1, t+2)$, $Q^{(N)}(t+1)$, is therefore given by

$$Q^{(N)}(t+1) = \left[Q^{(N)}(t) - NC + A^{(N)}(t)\right]^+ . \tag{10}$$

And, the average queue size $\hat{Q}^{(N)}(t)$ is given by

$$\hat{Q}^{(N)}(t+1) = (1-\alpha)\hat{Q}^{(N)}(t) + \alpha Q^{(N)}(t+1), \tag{11}$$

where $0 < \alpha \leq 1$ is the parameter of the exponential averaging mechanism.

Each incoming packet into the router in timeslot $[t, t+1)$ is marked with a probability $f^{(N)}\left(\hat{Q}^{(N)}(t)\right)$, depending on the average queue length at the beginning of the timeslot $[t, t+1)$. We represent this event using $\{0,1\}$-valued rvs $M_{i,j}^{(N)}(t+1)$ $(j = 1, ..., A_i^{(N)}(t))$ with the interpretation that $M_{i,j}^{(N)}(t+1) = 0$ (resp. $M_{i,j}^{(N)}(t+1) = 1$) if the jth packet from source i is marked (resp. not marked) in the RED buffer. To do so we introduce a collection of i.i.d. $[0,1]$-*uniform* rvs $\{V_{i,j}(t+1),\ i,j = 1, \cdots;\ t = 0, 1, \cdots\}$ that are assumed to be independent of other rvs. The process by which packets are marked is as follows. For each $i \in \mathcal{N}$ and $j = 1, 2, \ldots$, we define the marking rvs

$$M_{i,j}^{(N)}(t+1) = \mathbf{1}\left[V_{i,j}(t+1) > f^{(N)}(\hat{Q}^{(N)}(t))\right],$$

so that the rv $M_{i,j}^{(N)}(t+1)$ is the indicator function of the event that the jth packet from connection i is *not* marked in timeslot $[t, t+1)$. The indicator function of the event that no packets from connection i in timeslot $[t, t+1)$ are marked can now be written as $\prod_{j=1}^{A_i^{(N)}(t)} M_{i,j}^{(N)}(t+1)$. This information will be available to the TCP sender in the next timeslot. However, this information is used one RTT later to update the congestion window size as explained in Section 3.3, and $M_i^{(N)}(t+1)$ evolves according to

$$M_i^{(N)}(t+1) = G_{i,t}(M_i^{(N)}(t), M_{i,new}^{(N)}(t+1)), \tag{12}$$

and

$$M_{i,new}^{(N)}(t+1) = \prod_{j=1}^{A_i^{(N)}(t)} M_{i,j}^{(N)}(t+1)), \tag{13}$$

where we define $\prod_{j=1}^{0} M_{i,j}^{(N)}(t+1) = 1$. Notice that we use time parameter t for the mapping G to delay the change in the value of $M_i^{(N)}$ by one timeslot. Therefore, (6) and (7) evolve based on the markings in the previous round-trip as they should.

4 The Asymptotics - Law of Large Numbers

The first main result of the paper consists of the asymptotics for the normalized buffer content as the number of sessions becomes large. This result is discussed under the following Assumptions (A1)-(A2):
(A1) There exists a continuous function $f : \mathbb{R}_+ \to [0, 1]$ such that for each $N = 1, 2, \ldots,$

$$f^{(N)}(x) = f(N^{-1}x), \quad x \geq 0;$$

(A2) For each $N = 1, 2, \ldots,$ the initial conditions of rvs in (3), (4), (8), (9) and (10) are given by

$$Q^{(N)}(0) = W_i^{(N)}(0) = \beta_i^{(N)}(0) = d_i^{(N)}(0) = 0;$$
$$\text{and } S_i^{(N)}(0) = M_i^{(N)}(0) = 1; \quad i = 1, \ldots, N.$$

We denote the vector of state variables for user i in timeslot $[t, t+1)$ by

$$\mathbf{Y}_i^{(N)}(t) := (W_i^{(N)}(t), X_i^{(N)}(t), S_i^{(N)}(t), d_i^{(N)}(t), \beta_i^{(N)}(t), M_i^{(N)}(t)).$$

The rv $\mathbf{Y}_i^{(N)}(t)$ takes a value in $\mathcal{Y} := \{0,\ 1,\ldots, W_{max}\} \times \{0,\ 1,\ \ldots, X_{max}\} \times \{0,\ 1\} \times \{0,\ 2,\ 3,\ldots, D_{max}\} \times \{0,\ 1,\ \ldots, D_{max}\} \times \{0,\ 1\}$.

Assumption (A1) is a structural condition while (A2) is made essentially for technical convenience as it implies that for each N and all $t = 0, 1, \cdots$, the random vectors $\mathbf{Y}_1^{(N)}(t), \cdots, \mathbf{Y}_N^{(N)}(t)$ are *exchangeable*. Assumption (A2) can be omitted but at the expense of a more cumbersome discussion.

Theorem 1 *Assume that (A1)-(A2) hold. Then, for each $N = 1, 2, \ldots$ and $t = 0, 1, \ldots,$ there exist a (non-random) constant $q(t)$ and a $\mathcal{Y}$-valued rv $\mathbf{Y}(t) = (W(t), X(t), S(t), d(t), \beta(t), M(t))$ such that the following holds:*

(i) The following convergences take place:

$$\tfrac{Q^{(N)}(t)}{N} \xrightarrow{P}_N q(t) \text{ and } \tfrac{\hat{Q}^{(N)}(t)}{N} \xrightarrow{P}_N \hat{q}(t) \tag{14}$$

$$\mathbf{Y}_1^{(N)}(t) \Rightarrow_N \mathbf{Y}(t) \tag{15}$$

(ii) For any bounded function $g : \mathbb{N}_+^6 \to \mathbb{R}$

$$\tfrac{1}{N}\textstyle\sum_{i=1}^{N} g\left(\mathbf{Y}(t)\right) \xrightarrow{P}_N \mathbf{E}\left[g\left(\mathbf{Y}(t)\right)\right], \tag{16}$$

(iii) For any integer $I = 1, 2, \ldots,$ *the rvs* $\{\mathbf{Y}_i^{(N)}(t), i = 1, \ldots, I\}$ *become asymptotically independent as N becomes large, with*

$$\lim_{N\to\infty} \mathbf{P}[\mathbf{Y}_i^{(N)}(t) = \mathbf{y}_i, i = 1, \cdots, I] = \prod_{i=1}^{I} \mathbf{P}\left[\mathbf{Y}(t) = \mathbf{y}_i\right] \tag{17}$$

for any $\mathbf{y}_i \in \mathcal{Y}$, $i = 1, \ldots, I$.

In addition, with initial conditions $q(0) = W(0) = X(0) = d(0) = \beta(0) = 0,\ S(0) = M(0) = 1$, *it holds that*

$$q(t+1) = \left(q(t) - C + \mathbf{E}\left[A(t)\right]\right)^+ \tag{18}$$

$$\hat{q}(t+1) = (1-\alpha)\hat{q}(t) + \alpha q(t+1) \tag{19}$$

where $A(t) = \min\left(W(t), X(t)\right) \mathbf{1}\left[\beta(t) \geq d(t)\right]$. *Further, the recurrence*

$$\begin{aligned} \mathbf{Y}(t+1) &= (W(t+1), X(t+1), S(t+1), d(t+1), \beta(t+1), M(t+1)) \\ &=_{st} P(\mathbf{Y}(t)) := (P_1(\mathbf{Y}(t)), P_2(\mathbf{Y}(t)), \ldots, P_6(\mathbf{Y}(t))) \end{aligned}$$

holds in law, where the mappings $P_1, \ldots, P_6$ *are given in [11].*

A proof of Theorem 1 is given in [11].

5 Central Limit Theorem

In this section we sharpen the LLN results in Theorem 1 with a Central Limit Theorem (CLT) complement. The analysis is carried out under the same model. However, we need to strengthen Assumption (A1) to (A1b) and introduce Assumption (A3):
(A1b) Assumption (A1) holds with mapping $f : \mathbb{R}_+ \to [0, 1]$, which is continuously differentiable, *i.e.*, its derivative $f' : \mathbb{R}_+ \to \mathbb{R}$ exists and is continuous.
(A3) The workload of a new TCP connection is bounded, *i.e.*, there exists an integer X_{max} such that $F_i^{(N)}(t) \in \{1, \cdots, X_{max}\},\ i \in \mathcal{N},\ t = 0, 1, \ldots$ [4]

Fix $t = 0, 1, \ldots$. The following quantity plays a crucial role in the analysis:

$$K(t) := C - q(t) - \mathbf{E}\left[A(t)\right]. \tag{20}$$

We can interpret $K(t)$ as the asymptotic residual capacity per session in the timeslot $[t, t+1)$. Now define a collection of rvs that is integral to the analysis. For each $N = 1, 2, 3, \ldots$ and $\mathbf{y} = (w, x, s, d, b, m) \in \mathcal{Y}$, let

$$\begin{aligned} L_0^{(N)}(t) &= \frac{Q^{(N)}(t)}{N} - q(t), \\ \hat{L}_0^{(N)}(t) &= \frac{\hat{Q}^{(N)}(t)}{N} - \hat{q}(t), \end{aligned} \tag{21}$$

[4] The limit X_{max} can be lifted. This, however, results in a more complicated analysis. Nevertheless, such a restriction is necessary for the numerical calculation of the CLT on a computer.

and

$$L^{(N)}_{w,x,s,d,b,m}(t) := \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[w,x,s,d,b,m\right] - \mathbf{P}_{\mathbf{Y}(t)}\left[w,x,s,d,b,m\right] \tag{22}$$

We also use the notation $L^{(N)}_{\mathbf{y}}(t)$ interchangeably with $L^{(N)}_{w,x,s,d,b,m}(t)$.

Theorem 2 *Assume (A1b)-(A3) hold. Then, for each* $t = 0, 1, \ldots$, *there exists an* $\mathbb{R}^{|\mathcal{Y}|+2}$*-valued rv* $\mathbf{L}(t) = (L_0(t), \hat{L}_0(t), L_{\mathbf{y}}(t), \mathbf{y} \in \mathcal{Y})$ *such that the convergence*

$$\sqrt{N}\left(L_0^{(N)}(t), \hat{L}_0^{(N)}(t), L_{\mathbf{y}}^{(N)}(t),\ \mathbf{y} \in \mathcal{Y}\right) \Rightarrow_N \mathbf{L}(t) \tag{23}$$

holds. Moreover, the distributional recurrences

$$L_0(t+1) =_{st} \begin{cases} 0 & K(t) > 0 \\ L_0(t) + \bar{L}(t) & K(t) < 0 \\ (L_0(t) + \bar{L}(t))^+ & K(t) = 0 \end{cases} \tag{24}$$

and

$$\hat{L}_0(t+1) =_{st} (1-\alpha)\hat{L}_0(t) + \alpha L_0(t+1) \tag{25}$$

hold, where

$$\bar{L}(t) = \sum_{\mathbf{y}\in\mathcal{Y}} \min(w,x)\mathbf{1}\left[b \geq d\right] L_{\mathbf{y}}(t).$$

The distribution of the rv $L_{\mathbf{y}}(t)$, $\mathbf{y} \in \mathcal{Y}, t = 0, 1, \ldots$, *can be calculated recursively starting with* $t = 0$.

Finally, for any $t = 1, 2, \ldots$, *the rv* $L_0(t+1)$ *is Gaussian*[5] *if* $K(s) \neq 0$ *for all* $s < t$.

Proof. A complete proof is provided in Appendix A. ■

From the last statement of Theorem 2 a necessary condition for $L_0(t)$ not to be Gaussian is $K(s) = 0$ for some $s < t$. This is, however, a rare event because $K(t)$ is a real number. Therefore, in practice with a large number of flows the queue size distribution can be well approximated by a Gaussian rv.

It is shown in Appendix that the queue fluctuation $L_0(t)$ consists of four components :

(i) Fluctuation caused by the structure of protocols: This appears as a linear combination of several terms (through mappings $\Phi(\cdot)$, $\Psi(\cdot)$, and $\Gamma(\cdot)$ in (44), (45), and (46), respectively) that represent the fluctuations in the previous timeslot $[t-1, t)$.

(ii) Fluctuation caused by the discrepancy between the feedback information from RED to TCP sources $f^{(N)}(\hat{Q}^{(N)}(t))$ and the limiting feedback information $f(\hat{q}(t))$ (rvs $A^{(N)}_{\mathbf{y}}(t)$ in (60)): This uncertainty in feedback information can be explained by the following lemma (also known as the *Delta Method* [6, p. 214]). First define

$$\gamma(t) := f(\hat{q}(t)) \text{ and } \gamma^{(N)}(t) := f(\hat{Q}^{(N)}(t)/N).$$

[5] Here we intepret a constant as a Gaussian rv with standard deviation equals to zero.

Lemma 1 *If $f : \mathbb{R}_+ \to [0,1]$ is differentiable with a continuous derivative at $x = \hat{q}(t)$, then the convergence $\sqrt{N}(\frac{\hat{Q}^{(N)}(t)}{N} - \hat{q}(t)) \Rightarrow_N \hat{L}_0(t)$ implies $\sqrt{N}(\gamma^{(N)}(t) - \gamma(t)) \Rightarrow_N f'(\hat{q}(t))\hat{L}_0(t)$*

Note that as the slope of the feedback function increases, the magnitude of fluctuation due to this component increases as well. This verifies the observation that the magnitude of queue size oscillation at RED gateways increases with the slope of marking probability function of RED mechanism [12].

(iii) Binary nature of feedback information: The RED gateway either marks a packet or does not. This binary nature of feedback information poses limited feedback information granularity, and causes a fluctuation in queue size (rvs $B_{\mathbf{y}}^{(N)}(t)$ in (61)). Moreover, this fluctuation does not go away as the number of sessions increases, and can be well approximated by a Gaussian rv.

(iv) Fluctuation caused by the arrival of new TCP connections and the random idle periods (rvs $D_{\mathbf{y}}^{(N)}(t)$ in (66)): The larger the file size, *i.e.,* workload of a new TCP connection, and waiting time variances are, the larger the magnitude of this fluctuation is. This part of the fluctuation can also be described by a Gaussian rv.

5.1 Distributional Recursion of the CLT

In this section, we present a distributional relationship between $L_{\mathbf{y}}(t+1)$ and $L_{\mathbf{y}}(t),\ \mathbf{y} \in \mathcal{Y},\ t = 0, 1, \ldots$. The rv $L_{\mathbf{y}}(t+1)$ is equivalent in distribution to a summation of the following three distinct rvs: (i) a linear combination of $L_0(t)$ and $L_{\mathbf{y}'}(t),\ \mathbf{y}' \in \mathcal{Y}$; (ii) a rv representing the fluctuation caused by the difference between the sampled average of the packet marking rate (through the indicator functions) and its limiting marking probability in the time slot $[t, t+1)$; and (iii) a rv representing the fluctuation caused by newly arrived files from previously idle sessions in time slot $[t, t+1)$.

For part (i) we represent the linear combination of $L_{\mathbf{y}'}(t),\ \mathbf{y}' \in \mathcal{Y}$ through a mapping $\Phi(\,\cdot\,)$ (to be defined in (44) of Section 5.2.2). In part (ii) the fluctuation shows up as a linear combination of $J_{\mathbf{y}}(t)\hat{L}_0(t) + B_{\mathbf{y}}(t),\ \mathbf{y} \in \mathcal{Y}$. Here $\mathbf{B}(t) := (B_{\mathbf{y}}(t),\ \mathbf{y} \in \mathcal{Y})$ is a $|\mathcal{Y}|$-dimensional Gaussian random vector with mutually independent elements, and

$$J_{\mathbf{y}}(t) = \begin{cases} (w \wedge x)\gamma(t)^{(w\wedge x)-1} f'(\hat{q}(t)) \mathbf{P}_{\mathbf{Y}(t)}[\mathbf{y}], & \mathbf{y} = (w, x, s, d, b, m) \in \mathcal{Y}, w \wedge x \geq 1, \\ & b = d \\ 0, & \text{otherwise.} \end{cases}$$

For each $\mathbf{y} \in \mathcal{Y}$, $B_{\mathbf{y}}(t) \sim \mathcal{N}(0, R_{\mathbf{y}}(t))$ where

$$R_{\mathbf{y}}(t) = \mathbf{P}_{\mathbf{Y}(t)}[\mathbf{y}]\, \gamma(t)^{w\wedge x}(1 - \gamma(t)^{w\wedge x}). \tag{26}$$

The linear combination of $J_{\mathbf{y}}(t)\hat{L}_0(t) + B_{\mathbf{y}}(t),\ \mathbf{y} \in \mathcal{Y}$ is represented through a mapping $\Psi(\,\cdot\,)$ (to be defined in (45) of Section 5.2.2).

Finally, the fluctuation from the newly arrived files appears through the mapping $\Gamma(\,\cdot\,)$ (to be defined in (46) of Section 5.2.2) of the following Gaussian random vector:

$$\mathbf{D}(t) := (D_{\mathbf{y}}(t), \mathbf{y} \in \mathcal{Y}) \sim \mathcal{N}(\mathbf{0}, \mathbf{V}(t)),$$

where $\mathbf{V}(t) = [V_{\mathbf{y}\mathbf{y}'}(t),\ \mathbf{y}, \mathbf{y}' \in \mathcal{Y}]$. The elements $V_{\mathbf{y}\mathbf{y}'}(t)$ are given by

$$V_{\mathbf{y}\mathbf{y}'}(t) = \begin{cases} \mathbf{P}_{\mathbf{Y}(t)}\,[0,0,1,0,0,1]\, P_{ar}(1-P_{ar}), & \mathbf{y} = \mathbf{y}' = (0,0,1,0,0,1) \\ \mathbf{P}_{\mathbf{Y}(t)}\,[0,0,1,0,0,1]\, p_x q_d P_{ar}(1 - p_x q_d P_{ar}), & \mathbf{y} = \mathbf{y}' \in \mathcal{Y}' \\ -\mathbf{P}_{\mathbf{Y}(t)}\,[0,0,1,0,0,1]\, p_x q_d P_{ar}(1-P_{ar}), & \mathbf{y} = (0,0,1,0,0,1), \mathbf{y}' \in \mathcal{Y}' \\ -\mathbf{P}_{\mathbf{Y}(t)}\,[0,0,1,0,0,1]\, p_x p_{x'} q_d q_{d'} P_{ar}^2, & \mathbf{y}, \mathbf{y}' \in \mathcal{Y}' \\ 0, & \text{otherwise,} \end{cases} \tag{27}$$

where $\mathbf{y} = (w,x,s,d,b,m)$, $\mathbf{y}' = (w',x',s',d',b',m')$, $p_x = \mathbf{P}[F_i(t) = x]$, $x = 1, \ldots, X_{max}$, $q_y = \mathbf{P}[H_i(t) = y]$, $y = 2, \ldots, D_{max}$ and

$$\mathcal{Y}' = \{(0,x,1,d,0,1) |\ x \in \{1, \ldots, X_{max}\}, d \in \{2, \ldots, D_{max}\}\}, \tag{28}$$

i.e., , $(0,x,1,d,0,1) \in \mathcal{Y}'$ represents the state of a newly active session with new workload of x (packets) and RTT of d (timeslots).

The random vectors $\mathbf{L}(t) := (L_0(t), \hat{L}_0(t), L_{\mathbf{y}}(t), \mathbf{y} \in \mathcal{Y})$, $\mathbf{B}(t)$, and $\mathbf{D}(t)$ are mutually independent. Let

$$\mathbf{J}(t) := (J_{\mathbf{y}}(t), \mathbf{y} \in \mathcal{Y})$$

and

$$\mathbf{L}_{\mathcal{Y}}(t) := (L_{\mathbf{y}}(t), \mathbf{y} \in \mathcal{Y}).$$

We can now write

$$L_{\mathbf{y}}(t+1) = \Phi(\mathbf{L}_{\mathcal{Y}}(t), \gamma(t); \mathbf{y}) + \Psi(\hat{L}_0(t)\mathbf{J}(t) + \mathbf{B}(t); \mathbf{y}) + \Gamma(\mathbf{D}(t); \mathbf{y}). \tag{29}$$

5.2 State Space Partition and Mappings Φ, Ψ, and Γ

Before describing the mapping Φ, Ψ, and Γ, we first need to introduce the following partitions of the state space $\mathcal{Y}$ which the mappings will depend upon.

5.2.1 Partitions of state space $\mathcal{Y}$

We divide the state space $\mathcal{Y}$ into fourteen partitions, $\mathcal{Y}_1, \ldots, \mathcal{Y}_{14}$. Each partition represents a distinct state of the session dynamics, *e.g.,* if the state variable $\mathbf{y}$ of a session is in $\mathcal{Y}_9$ during timeslot $[t, t+1)$, then the session is idle in the timeslot $[t, t+1)$. These partitions are crucial to the decomposition in Appendix C, from which the mappings Φ, Ψ, and Γ will be derived.

Case (i): Connection is in CA and just increased its congestion window by one at the beginning of timeslot and the window size is strictly smaller than W_{max}.

$$\begin{aligned} \mathbf{y} \in \mathcal{Y}_1 \iff & w \in \{2, \ldots, W_{max} - 1\}, x \in \{1, \ldots, X_{max}\}, s = 0, \\ & d = b \in \{2, \ldots, D_{max}\}, m = 1 \end{aligned} \tag{30}$$

Case (ii): Connection just reduced its congestion window by half.

$$\begin{aligned} \mathbf{y} \in \mathcal{Y}_2 \iff & w \in \{1, \ldots, \lceil \frac{W_{max}}{2} \rceil\}, x \in \{1, \ldots, X_{max}\}, s = 0, \\ & d = b \in \{2, \ldots, D_{max}\}, m = 0 \end{aligned} \tag{31}$$

Case (iii): Connection just doubled its congestion window and the window size is strictly smaller than W_{max}.

$$\mathbf{y} \in \mathcal{Y}_3 \iff w \in \{2^n | n = 1, 2, \dots, \text{ and } 2^n < W_{\max}\}, x \in \{1, \dots, X_{max}\}, s = 1, \\ d = b \in \{2, \dots, D_{max}\}, m = 1 \tag{32}$$

Case (iv): Connection is in CA with congestion window size equal to W_{max} and transmits packets in the timeslot.

$$\mathbf{y} \in \mathcal{Y}_4 \iff w = W_{\max}, x \in \{1, \dots, X_{max}\}, s = 0, \\ d = b \in \{2, \dots, D_{max}\}, m = 1 \tag{33}$$

Case (v): Connection is in SS with congestion window size equal to W_{max} and transmits packets in the timeslot.

$$\mathbf{y} \in \mathcal{Y}_5 \iff w = W_{\max}, x \in \{1, \dots, X_{max}\}, s = 1, \\ b = d \in \{2, \dots, D_{max}\}, m = 1 \tag{34}$$

Case (vi): Connection is in the middle of a round-trip and only increases the counter β.

$$\mathbf{y} \in \mathcal{Y}_6 \iff w \in \{1, \dots, W_{\max}\}, x \in \{1, \dots, X_{max}\}, s \in \{0, 1\}, \\ b \in \{2, \dots, d - 1\}, d \in \{2, \dots, D_{max}\}, m \in \{0, 1\} \tag{35}$$

Case (vii): Connection transmitted packet(s) in the previous timeslot and none of the packet(s) was marked.

$$\mathbf{y} \in \mathcal{Y}_7 \iff w \in \{1, \dots, W_{\max}\}, x \in \{1, \dots, X_{max}\}, s \in \{0, 1\}, \\ b = 1, d \in \{2, \dots, D_{max}\}, m = 1 \tag{36}$$

Case (viii): Connection transmitted packet(s) in the previous timeslot and some of the packet(s) were marked.

$$\mathbf{y} \in \mathcal{Y}_8 \iff w \in \{1, \dots, W_{\max}\}, x \in \{1, \dots, X_{max}\}, s = 0, \\ b = 1, d \in \{2, \dots, D_{max}\}, m = 0 \tag{37}$$

Case (ix): Session is in an idle state.

$$\mathbf{y} \in \mathcal{Y}_9 \iff w = x = 0, s = 1, d = b = 0, m = 1 \tag{38}$$

Case (x): A new (TCP) connection is just initiated.

$$\mathbf{y} \in \mathcal{Y}_{10} \iff w = 0, x \in \{1, \dots, X_{max}\}, s = 1, \\ d \in \{2, \dots, D_{max}\}, b = 0, m = 1 \tag{39}$$

Case (xi): Connection is in three-way handshake.

$$\mathbf{y} \in \mathcal{Y}_{11} \iff w = 0, x \in \{1, \dots, X_{max}\}, s = 1, \\ b = \{1, \dots, d - 1\}, d \in \{2, \dots, D_{max}\}, m = 1 \tag{40}$$

Case (xii): Connection transmits a packet for the first time after three-way handshake.

$$\begin{aligned}\mathbf{y} \in \mathcal{Y}_{12} \quad \Longleftrightarrow \quad & w = 1, x \in \{1, \ldots, X_{max}\}, s = 1, \\ & b = d \in \{2, \ldots, D_{max}\}, m = 1\end{aligned} \tag{41}$$

Case (xiii): Connection completed transfer of packets in the previous timeslot and none of the packet(s) was marked. Session will become idle in the next timeslot.

$$\begin{aligned}\mathbf{y} \in \mathcal{Y}_{13} \quad \Longleftrightarrow \quad & w = \{1, \ldots, W_{\max} - 1\}, x = 0, s \in \{0, 1\}, \\ & b = 1, d \in \{2, \ldots, D_{max}\}, m = 1\end{aligned} \tag{42}$$

Case (xiv): Connection completed transfer of packets in the previous timeslot and some of the packet(s) were marked. Session will become idle in the next timeslot.

$$\begin{aligned}\mathbf{y} \in \mathcal{Y}_{14} \quad \Longleftrightarrow \quad & w = \{1, \ldots, W_{\max} - 1\}, x = 0, s = 0, \\ & b = 1, d \in \{2, \ldots, D_{max}\}, m = 0\end{aligned} \tag{43}$$

We are now ready to define the mappings Φ, Ψ, and Γ.

5.2.2 The mapping Φ, Ψ, and Γ

Let $\mathbf{Y}$ be an array of real numbers consisting of $Y_{\mathbf{y}'}, \mathbf{y}' = (a, b, c, d, e, f) \in \mathcal{Y}$. Let $a \in \mathbb{R}$. Define the following mappings:

$$\Phi(\mathbf{Y}, a; \mathbf{y}) = \begin{cases} Y_{w-1,x,0,d,d-1,1} & ;\ \mathbf{y} \in \mathcal{Y}_1 \\ Y_{2w-1,x,0,d,d-1,0} + Y_{2w,x,0,d,d-1,0} & ;\ \mathbf{y} \in \mathcal{Y}_2 \\ Y_{\frac{w}{2},x,1,d,d-1,1} & ;\ \mathbf{y} \in \mathcal{Y}_3 \\ Y_{w,x,0,d,d-1,1} + Y_{w-1,x,0,d,d-1,1} & ;\ \mathbf{y} \in \mathcal{Y}_4 \\ \sum_{w'=\lceil \frac{W_{\max}}{2} \rceil}^{W_{\max}} Y_{w',x,1,d,d-1,1} & ;\ \mathbf{y} \in \mathcal{Y}_5 \\ Y_{w,x,s,d,b-1,m} & ;\ \mathbf{y} \in \mathcal{Y}_6 \\ a^w (Y_{w,w+x,s,d,d,1} + Y_{w,w+x,s,d,d,0}) & ;\ \mathbf{y} \in \mathcal{Y}_7 \\ (1 - a^w)(Y_{w,w+x,0,d,d,0} + Y_{w,w+x,0,d,d,1} & \\ +Y_{w,w+x,1,d,d,0} + Y_{w,w+x,1,d,d,1}) & ;\ \mathbf{y} \in \mathcal{Y}_8 \\ (1 - P_{ar}) Y_{0,0,1,0,0,1} + \sum_{\mathbf{y}' \in \mathcal{Y}_{\mathbf{13}} \cup \mathcal{Y}_{\mathbf{14}}} Y_{\mathbf{y}'} & ;\ \mathbf{y} \in \mathcal{Y}_9 \\ P_{ar} p_x q_d Y_{0,0,1,0,0,1} & ;\ \mathbf{y} \in \mathcal{Y}_{10} \\ Y_{w,x,s,d,b-1,m} & ;\ \mathbf{y} \in \mathcal{Y}_{11} \\ Y_{0,x,s,d,b-1,m} & ;\ \mathbf{y} \in \mathcal{Y}_{12} \\ \sum_{w'=1}^{W_{\max}} \sum_{x'=1}^{w'} a^{w'} (Y_{w',x',s,d,d,1} + Y_{w',x',s,d,d,0}) & ;\ \mathbf{y} \in \mathcal{Y}_{13} \\ \sum_{w'=1}^{W_{\max}} \sum_{x'=1}^{w'} (1 - a^{w'})(Y_{w',x',0,d,d,0} + Y_{w',x',0,d,d,1} & \\ +Y_{w',x',1,d,d,0} + Y_{w',x',1,d,d,1}) & ;\ \mathbf{y} \in \mathcal{Y}_{14} \\ 0 & ;\ \text{otherwise} \end{cases} \tag{44}$$

$$\Psi(\mathbf{Y}; \mathbf{y}) = \begin{cases} Y_{w,w+x,s,d,d,1} + Y_{w,w+x,s,d,d,0} & ;\ \mathbf{y} \in \mathcal{Y}_7 \\ -Y_{w,w+x,0,d,d,0} - Y_{w,w+x,0,d,d,1} - Y_{w,w+x,1,d,d,0} - Y_{w,w+x,1,d,d,1} & ;\ \mathbf{y} \in \mathcal{Y}_8 \\ \sum_{w'=1}^{W_{\max}} \sum_{x'=1}^{w'} Y_{w',x',s,d,d,1} + Y_{w',x',s,d,d0} & ;\ \mathbf{y} \in \mathcal{Y}_{13} \\ \sum_{w'=1}^{W_{\max}} \sum_{x'=1}^{w'} -Y_{w',x',0,d,d,0} - Y_{w',x',0,d,d,1} - Y_{w',x',1,d,d,0} & \\ -Y_{w',x',1,d,d,1} & ;\ \mathbf{y} \in \mathcal{Y}_{14} \\ 0 & ;\ \text{otherwise} \end{cases} \tag{45}$$

$$\Gamma(\mathbf{Y};\mathbf{y}) \;=\; \begin{cases} Y_{\mathbf{y}} & ;\ \mathbf{y} \in \mathcal{Y}_9 \cup \mathcal{Y}_{10} \\ 0 & ;\ \text{otherwise} \end{cases} \tag{46}$$

These mappings then completely specify the distributional recursion of $L_{\mathbf{y}}(t+1)$ in (29) and also the recursion of $\mathbf{L}(t+1)$.

6 Discussion

Theorems 1 and 2 show that the dynamics of the queue at time t, denoted by $Q^{(N)}(t)$, can be approximated by $Nq(t) + \sqrt{N}L_0(t)$ with $q(t)$ determined via a simple deterministic recursion, which is independent of the number of sessions. Moreover, $L_0(t)$ is Gaussian (per remark following Theorem 2) and its variance can also be computed recursively. The offered traffic into the network during the timeslot, $A^{(N)}(t)$, can also be approximated by $N \cdot \mathbf{E}[A(t)]$. These approximations become more accurate as the number of sessions becomes large, and the computational complexity does not increase with N. The limiting model is therefore "scalable" as it does not suffer from the explosion of state space, nor does it require any ad-hoc assumptions in the analysis.

We briefly consider the resulting model from Theorem 1 when C is either very large or very small under the following assumption.

(A1c) The marking function $f : \mathbb{R} \to [0,1]$ in Assumption (A1) is monotonically increasing with $f(0) = 0$ and $\lim_{x\to\infty} f(x) = 1$.

First, suppose that $C \to \infty$. In this case, it is easy to see that $\lim_{C\to\infty} q(t) = 0$ for all $t = 0, 1, \ldots$, and hence the marking probability per flow also converges to zero from (A1c) for all t. Therefore, each incoming flow will always operate in SS and the resulting input traffic into the network can be approximated by the superposition of (discrete-time) Poisson arrival streams of random number of packets, each of which doubles its window size every round-trip. The aggregate input traffic is therefore similar to the time-reversed shot-noise processes, in agreement with [4].

On the other hand, if $C \simeq 0$, the queue will start building up, whence $\lim_{t\to\infty} q(t) = \infty$. Thus, for large t, all TCP flows (including incoming TCP flows) will experience marking probability close to one from Assumption (A1c). This implies that the congestion window size converges to one and each connection can transmit only one packet per round-trip. Since the bottleneck router will transmit packets non-selectively, any active flow will receive roughly equal throughput and hence the queue behavior approaches that of processor-sharing, assuming identical RTTs among all flows as claimed in [7].

The CLT analysis reveals the sources of fluctuation in the queue size. Components (i) and (iii) at the end of Section 5 are due to the protocols and cannot be mitigated without modifying protocols. In addition, component (iv) depends on the user behavior, and hence is beyond the control of network. Thus, network designers can only manipulate the slope of the feedback function to reduce oscillation of queue size. Although reducing the slope of the queue can decrease the magnitude of fluctuation, it also increases the average queue size as suggested by (36) in [11].

The aforementioned trade-off can be explored in the context of an optimization problem. Suppose that $q(t) \to_t q^\star$ and $L_0(t) \Rightarrow_t L_0^\star$.[6] Then, the steady-state queue distribution can be approximated by $Nq^* + \sqrt{N}L_0^\star$. For best-effort traffic, the marking function f should, for example, maximize $\mathbf{P}\left[0 < Nq^* + \sqrt{N}L_0^\star < NB\right]$, assuming that the buffer size is NB. Clearly, a more sophisticated performance metric can be adopted, which depends on the probabilities of an empty queue and buffer overflow. The availability of the queue distribution allows us to formulate such a problem as an optimization problem and propose a systematic solution.

[6] Note that these quantities depend on f.

7 Conclusion

In this paper we have developed a scalable model of a RED gateway under a large number of TCP flows. We have demonstrated several interesting behaviors of the limiting model and their implications on design of a good marking function. These results are further strengthened by our CLT results. These results provide us with a scalable tool that can be utilized for network dimensioning without suffering from the curse of dimensionality. Furthermore, their proofs provide valuable insights into the behavior of queue size and help us design better marking functions. Our model is shown to be consistent with other previously proposed models in their respective regime.

We are currently investigating the optimization problem described in Section 6 with various performance metrics. In addition, we are working on extending our model to cases where there are multiple bottlenecks in the network. We expect that similar results in such cases will provide us with additional insights into how different sets of flows traversing different bottlenecks affect the transient behavior of queue dynamics as well as steady-state queue sizes.

References

[1] Sally Floyd. TCP and explicit congestion notification. *Computer Communication Review*, 24(5):10–23, October 1994.

[2] Sally Floyd and Van Jacobson. Random early detection gateways for congestion avoidance. *IEEE/ACM Transactions on Networking*, 1(4):397–413, August 1995.

[3] C. Hollot, V. Misra, D. Towsley, and W. Gong. A control theoretic analysis of RED. In *Proceedings of IEEE INFOCOM*, 2001.

[4] C. V. Hollot, Yong Liu, Vishal Misra, and Don Towsley. Unresponsive flows and AQM performance. In *Proceedings of IEEE INFOCOM*, April 2003.

[5] Van Jacobson. Congestion avoidance and control. In *Proceedings of SIGCOMM'88 Symposium*, pages 314–332, August 1988.

[6] Alan F. Karr. *Probability.* Springer-Verlag, 1993.

[7] Arzad Kherani and Anurag Kumar. Stochastic models for throughput analysis of randomly arriving elastic flows in the Internet. In *Proceedings of IEEE INFOCOM*, New York City, NY, July 2002.

[8] M. Mellia, I. Stoica, and H. Zhang. TCP model for short lived flows. *IEEE Communications Letters*, 6(2):85–87, 2002.

[9] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose. Modeling TCP Reno performance: A simple model and its empirical validation. *IEEE/ACM Transactions on Networking*, April 2000.

[10] Vern Paxson and Sally Floyd. Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3:226–244, 1995.

[11] Peerapol Tinnakornsrisuphap and Richard J. La. Limiting model of ECN/RED under a large number of heterogeneous TCP flows. Technical report, Institute for Systems Research, University of Maryland, 2003.

[12] Peerapol Tinnakornsrisuphap and Armand M. Makowski. Limit behavior of ECN/RED gateways under a large number of TCP flows. In *Proceedings of IEEE INFOCOM*, San Francisco, CA, April 2003.

[13] L. Zhang, S. Shenker, and D. Clark. Observations on the dynamics of a congestion control algorithm: The effects of two-way traffic. In *Proceedings of ACM SIGCOMM*, pages 133–147, September 1991.

A A Proof of Theorem 2

As in the proof of the Weak Law of Large Numbers [11], we proceed by the induction on t of the following statement.

[E:t] The convergence (23) holds for some $\mathbb{R}^{|\mathcal{Y}|+2}$-valued rv

$$\mathbf{L}(t) := \left(L_0(t), \hat{L}_0(t), L_{\mathbf{y}}(t),\ \mathbf{y} \in \mathcal{Y}\right).$$

The proof of Theorem 2 utilizes the following propositions.

Proposition 1 *Under (A1b)–(A3), if* **[E:t]** *holds for some* $t = 0, 1, \ldots$, *then*

$$\sqrt{N} L_0^{(N)}(t+1) = \sqrt{N}\left(\frac{Q^{(N)}(t+1)}{N} - q(t+1)\right) \Rightarrow_N L_0(t+1) \tag{47}$$

for some $\mathbb{R}$*-valued rv* $L_0(t+1)$ *that satisfies the distributional relation in (24).*

Proposition 2 *Under (A1b)–(A3), if* **[E:t]** *holds for some* $t = 0, 1, \ldots$, *then* **[E:t+1]** *also holds.*

We will return to the proofs of Propositions 1 and 2 in Appendices B and D, respectively. Before doing so, we conclude this section with a proof of Theorem 2.

Proof of Theorem 2.
For $t = 0$ and $\mathbf{y} \in \mathcal{Y}$, $L_0^{(N)}(t) = \bar{L}^{(N)}(t) = L_{\mathbf{y}}^{(N)}(t) = 0$. Hence, the statement **[E:0]** holds trivially. Eq. (23) and (24) then hold for all t from the induction on t by Propositions 1 and 2.

The distributional recursion in (25) and (29) are established as byproducts in the proof of Proposition 2.

A closer inspection of the decomposition (29) reveals that $L_{\mathbf{y}}(t+1), \mathbf{y} \in \mathcal{Y}$, is Gaussian if $L_{\mathcal{Y}}(t)$ and $\hat{L}_0(t)$, *i.e.,* , all components of $\mathbf{L}(t)$, are either Gaussian or constant. This also implies that $\bar{L}(t+1)$ is either constant or Gaussian if all $L_{\mathbf{y}}(t+1), \mathbf{y} \in \mathcal{Y}$ are. From (25), we see that $\hat{L}_0(t)$ is neither Gaussian nor constant if either $L_0(t)$ or $\hat{L}_0(t-1)$ are not. By induction on (24) and (25), it is easy to see that the necessary condition for $\hat{L}_0(t)$ to be neither Gaussian nor constant is for $K(s) = 0$ for some $s < t$ because if $K(s) = 0$ then $L_0(s+1)$ will become truncated Gaussian and $\hat{L}_0(r), r = s+1, s+2, \ldots$ will no longer be Gaussian.

■

B A Proof of Proposition 1

Fix $t = 0, 1, \ldots$ and $N = 1, 2, \ldots$. we rewrite the limiting recursion in (18) in the following form:

$$q(t+1) = \left(q(t) - C + \mathbf{E}\left[A(t)\right]\right)^+ = (-K(t))^+ \tag{48}$$

with $K(t)$ given by (20). Combining this observation with the queue dynamics (48), let

$$\begin{aligned} \bar{L}^{(N)}(t) &= \tfrac{1}{N}\textstyle\sum_{i=1}^N A_i^{(N)}(t) - \mathbf{E}\left[A(t)\right] \qquad (49) \\ &= \tfrac{1}{N}\textstyle\sum_{i=1}^N \min(W_i^{(N)}(t), X_i^{(N)}(t))\mathbf{1}\left[\beta_i^{(N)}(t) \geq d_i^{(N)}(t)\right] \\ &\quad -\mathbf{E}\left[\min(W(t), X(t))\mathbf{1}\left[\beta(t) \geq d(t)\right]\right] \\ &= \tfrac{1}{N}\textstyle\sum_{i=1}^N \displaystyle\sum_{\mathbf{y}\in\mathcal{Y}} \min(w, x)\mathbf{1}\left[b \geq d\right]\left(\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[\mathbf{y}\right] - \mathbf{P}_{\mathbf{Y}(t)}\left[\mathbf{y}\right]\right) \\ &= \sum_{\mathbf{y}\in\mathcal{Y}} \min(w, x)\mathbf{1}\left[b \geq d\right] L_{\mathbf{y}}^{(N)}(t) \ . \end{aligned}$$

Then, we have

$$\begin{aligned} L_0^{(N)}(t+1) &= \left(\frac{Q^{(N)}(t)}{N} - C + \frac{1}{N}\sum_{i=1}^N A_i^{(N)}(t)\right)^+ - (-K(t))^+ \\ &= \max\left(L_0^{(N)}(t) + \frac{1}{N}\sum_{i=1}^N A_i^{(N)}(t) - \mathbf{E}\left[A(t)\right], K(t)\right) - K(t)^+ \\ &= \max\left(L_0^{(N)}(t) + \bar{L}^{(N)}(t), K(t)\right) - K(t)^+ \qquad (50) \end{aligned}$$

and

$$\sqrt{N}L_0^{(N)}(t+1) = \max\left(\sqrt{N}\left(L_0^{(N)}(t) + \bar{L}^{(N)}(t)\right), \sqrt{N}K(t)\right) - \sqrt{N}K(t)^+.$$

Under **[E:t]**, we can invoke the Continuous Mapping Theorem to conclude that

$$\sqrt{N}\left(L_0^{(N)}(t) + \bar{L}^{(N)}(t)\right) \Rightarrow_N L_0(t) + \bar{L}(t) \tag{51}$$

Three cases emerge depending on the sign of $K(t)$. If $K(t) = 0$, then (51) reduces to

$$\sqrt{N}L_0^{(N)}(t+1) \quad = \quad \left(\sqrt{N}\left(L_0^{(N)}(t) + \bar{L}^{(N)}(t)\right)\right)^+ \ . \tag{52}$$

Again by the Continuous Mapping Theorem (51) yields

$$\sqrt{N}L_0^{(N)}(t+1) \Rightarrow_N \left(L_0(t) + \bar{L}(t)\right)^+.$$

If $K(t) < 0$, then (51) reduces to

$$\sqrt{N}L_0^{(N)}(t+1) = \max\left(\sqrt{N}\left(L_0^{(N)}(t) + \bar{L}^{(N)}(t)\right), -\sqrt{N}|K(t)|\right)$$

and (51) gives us

$$\sqrt{N}L_0^{(N)}(t+1) \Rightarrow_N L_0(t) + \bar{L}(t)$$

since $|K(t)| > 0$ guarantees $\lim_{N\to\infty}\sqrt{N}|K(t)| = \infty$.

Finally, if $K(t) > 0$, then (51) reduces to

$$\sqrt{N} L_0^{(N)}(t+1) = \max\left(\sqrt{N}\left(L_0^{(N)}(t) + \bar{L}^{(N)}(t)\right) - \sqrt{N}K(t), 0\right)$$

and the convergence (51) yields $\sqrt{N} L_0^{(N)}(t+1) \Rightarrow_N 0$ since $\lim_{N\to\infty} \sqrt{N}K(t) = \infty$. This completes the proof of Proposition 1. ■

C A Key Decomposition

To facilitate the proof of Proposition 2, we introduce the following decomposition of $L_{\mathbf{y}}^{(N)}(t+1)$, $\mathbf{y} \in \mathcal{Y}, t = 1, \dots$ with fourteen different cases defined in Section 5.2.1.

Case (i): For $\mathbf{y} \in \mathcal{Y}_1$,

$$\begin{aligned} L_{\mathbf{y}}^{(N)}(t+1) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[w-1, x, 0, d, d-1, 1\right] \\ &\quad -\mathbf{P}_{\mathbf{Y}(t)}\left[w-1, x, 0, d, d-1, 1\right] \\ &= L_{\mathbf{y_1'}}^{(N)}(t), \end{aligned} \tag{53}$$

where $\mathbf{y_1'} = (w-1, x, 0, d, d-1, 1)$.
Case (ii): For $\mathbf{y} \in \mathcal{Y}_2$,

$$\begin{aligned} L_{\mathbf{y}}^{(N)}(t+1) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[2w-1, x, 0, d, d-1, 0\right] \\ &\quad -\mathbf{P}_{\mathbf{Y}(t)}\left[2w-1, x, 0, d, d-1, 0\right] \\ &\quad +\tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[2w, x, 0, d, d-1, 0\right] \\ &\quad -\mathbf{P}_{\mathbf{Y}(t)}\left[2w, x, 0, d, d-1, 0\right] \\ &= L_{\mathbf{y_{2a}'}}^{(N)}(t) + L_{\mathbf{y_{2b}'}}^{(N)}(t), \end{aligned} \tag{54}$$

where $\mathbf{y_{2a}'} = (2w-1, x, 0, d, d-1, 0)$ and $\mathbf{y_{2b}'} = (2w, x, 0, d, d-1, 0)$.
Case (iii): For $\mathbf{y} \in \mathcal{Y}_3$,

$$\begin{aligned} L_{\mathbf{y}}^{(N)}(t+1) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[\frac{w}{2}, x, 1, d, d-1, 1\right] \\ &\quad -\mathbf{P}_{\mathbf{Y}(t)}\left[\frac{w}{2}, x, 1, d, d-1, 1\right] \\ &= L_{\mathbf{y_3'}}^{(N)}(t), \end{aligned} \tag{55}$$

where $\mathbf{y_3'} = (\frac{w}{2}, x, 1, d, d-1, 1)$.
Case (iv): For $\mathbf{y} \in \mathcal{Y}_4$,

$$\begin{aligned} L_{\mathbf{y}}^{(N)}(t+1) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[W_{\max}, x, 0, d, d-1, 1\right] \\ &\quad -\mathbf{P}_{\mathbf{Y}_i^{(N)}(t)}\left[W_{\max}, x, 0, d, d-1, 1\right] \\ &\quad +\tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[W_{\max}-1, x, 0, d, d-1, 1\right] \\ &\quad -\mathbf{P}_{\mathbf{Y}_i^{(N)}(t)}\left[W_{\max}-1, x, 0, d, d-1, 1\right] \\ &= L_{\mathbf{y_{4a}'}}^{(N)}(t) + L_{\mathbf{y_{4b}'}}^{(N)}(t), \end{aligned} \tag{56}$$

where $\mathbf{y'_{4a}} = (W_{\max}, x, 0, d, d-1, 1)$ and $\mathbf{y'_{4b}} = (W_{\max}-1, x, 0, d, d-1, 1)$.
Case (v): For $\mathbf{y} \in \mathcal{Y}_5$,

$$\begin{aligned}
L_{\mathbf{y}}^{(N)}(t+1) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)} \left[2^{\lceil \log_2(W_{\max})\rceil - 1}, x, 1, d, d-1, 1\right] \\
&\quad -\mathbf{P}_{\mathbf{Y}(t)} \left[2^{\lceil \log_2(W_{\max})\rceil - 1}, x, 1, d, d-1, 1\right] \\
&\quad +\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)} \left[W_{\max}, x, 1, d, d-1, 1\right] \\
&\quad -\mathbf{P}_{\mathbf{Y}(t)} \left[W_{\max}, x, 1, d, d-1, 1\right] \\
&= L^{(N)}_{2^{\lceil \log_2(W_{\max})\rceil - 1}, x, 1, d, d-1, 1}(t) + L^{(N)}_{W_{\max}, x, 1, d, d-1, 1}(t). \qquad (57)
\end{aligned}$$

Case (vi): For $\mathbf{y} \in \mathcal{Y}_6$,

$$\begin{aligned}
L_{\mathbf{y}}^{(N)}(t+1) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)} [w, x, s, d, b-1, m] \\
&\quad -\mathbf{P}_{\mathbf{Y}(t)} [w, x, s, d, b-1, m] \\
&= L^{(N)}_{\mathbf{y'_6}}(t), \qquad (58)
\end{aligned}$$

where $\mathbf{y'_6} = (w, x, s, d, b-1, m)$.
Case (vii): For $\mathbf{y} \in \mathcal{Y}_7$, with some simple algebras, it can be shown that

$$\begin{aligned}
L_{\mathbf{y}}^{(N)}(t+1) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)} [w, w+x, s, d, d, 0]\, M^{(N)}_{i,new}(t+1) \\
&\quad -\mathbf{E}\left[\mathbf{1}_{\mathbf{Y}(t)} [w, w+x, s, d, d, 0]\, M_{new}(t+1)\right] \\
&\quad +\tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)} [w, w+x, s, d, d, 1]\, M^{(N)}_{i,new}(t+1) \\
&\quad -\mathbf{E}\left[\mathbf{1}_{\mathbf{Y}(t)} [w, w+x, s, d, d, 1]\, M_{new}(t+1)\right] \\
&= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)} [w, w+x, s, d, d, 0]\, (M^{(N)}_{i,new}(t+1) - \gamma^{(N)}(t)^w) \\
&\quad +\tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)} [w, w+x, s, d, d, 0]\, (\gamma^{(N)}(t)^w - \gamma(t)^w) \\
&\quad +\gamma(t)^w L^{(N)}_{w,w+x,s,d,d,0}(t) \\
&\quad +\tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)} [w, w+x, s, d, d, 1]\, (M^{(N)}_{i,new}(t+1) - \gamma^{(N)}(t)^w) \\
&\quad +\tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)} [w, w+x, s, d, d, 1]\, (\gamma^{(N)}(t)^w - \gamma(t)^w) \\
&\quad +\gamma(t)^w L^{(N)}_{w,w+x,s,d,d,1}(t) \\
&= (A+B)^{(N)}_{\mathbf{y'_{7a}}}(t) + \gamma(t)^w L^{(N)}_{\mathbf{y'_{7a}}}(t) \\
&\quad +(A+B)^{(N)}_{\mathbf{y'_{7b}}}(t) + \gamma(t)^w L^{(N)}_{\mathbf{y'_{7b}}}(t), \qquad (59)
\end{aligned}$$

where $\mathbf{y'_{7a}} = (w, w+x, s, d, d, 0)$, $\mathbf{y'_{7b}} = (w, w+x, s, d, d, 1)$ and

$$\begin{aligned}
A_{\mathbf{y}}^{(N)}(t) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)} [\mathbf{y}] \left(\gamma^{(N)}(t)^{w\wedge x} - \gamma(t)^{w\wedge x}\right) \qquad (60) \\
B_{\mathbf{y}}^{(N)}(t) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)} [\mathbf{y}] \left(M^{(N)}_{i,new}(t+1) - \gamma^{(N)}(t)^{w\wedge x}\right), \qquad (61)
\end{aligned}$$

with the notation $(A+B)_{\mathbf{y}}^{(N)}(t) = A_{\mathbf{y}}^{(N)}(t) + B_{\mathbf{y}}^{(N)}(t)$.

Case (viii): For $\mathbf{y} \in \mathcal{Y}_8$,

$$\begin{aligned} L_{\mathbf{y}}^{(N)}(t+1) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[w, w+x, s, d, d, 0\right](1 - M_{i,new}^{(N)}(t+1)) \\ &\quad -\mathbf{E}\left[\mathbf{1}_{\mathbf{Y}(t)}\left[w, w+x, s, d, d, 0\right](1 - M_{new}(t+1))\right] \\ &\quad +\tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[w, w+x, s, d, d, 1\right](1 - M_{i,new}^{(N)}(t+1)) \\ &\quad -\mathbf{E}\left[\mathbf{1}_{\mathbf{Y}(t)}\left[w, w+x, s, d, d, 1\right](1 - M_{new}(t+1))\right] \\ &= (1-\gamma(t)^w)L_{\mathbf{y}'_{\mathbf{8a}}}^{(N)}(t) - (A+B)_{\mathbf{y}'_{\mathbf{8a}}}^{(N)}(t) \\ &\quad +(1-\gamma(t)^w)L_{\mathbf{y}'_{\mathbf{8b}}}^{(N)}(t) - (A+B)_{\mathbf{y}'_{\mathbf{8b}}}^{(N)}(t), \end{aligned} \tag{62}$$

where $\mathbf{y}'_{\mathbf{8a}} = (w, w+x, s, d, d, 0)$ and $\mathbf{y}'_{\mathbf{8b}} = (w, w+x, s, d, d, 1)$.
Case (ix): For $\mathbf{y} \in \mathcal{Y}_9$,

$$\begin{aligned} L_{\mathbf{y}}^{(N)}(t+1) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[0,0,1,0,0,1\right]\mathbf{1}\left[U_i(t+1) > P_{ar}\right] \\ &\quad -(1-P_{ar})\mathbf{P}_{\mathbf{Y}(t)}\left[0,0,1,0,0,1\right] \\ &\quad +\tfrac{1}{N}\textstyle\sum_{i=1}^{N} \displaystyle\sum_{\mathbf{y}' \in \mathcal{Y}_{13}\cup\mathcal{Y}_{14}} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[\mathbf{y}'\right] - \mathbf{P}_{\mathbf{Y}(t)}\left[\mathbf{y}'\right] \\ &= D_{\mathbf{y}}^{(N)}(t) + (1-P_{ar})L_{0,0,1,0,0,1}^{(N)}(t) + \sum_{\mathbf{y}' \in \mathcal{Y}_{13}\cup\mathcal{Y}_{14}} L_{\mathbf{y}'}^{(N)}(t), \end{aligned} \tag{63}$$

where we define

$$D_{0,0,1,0,0,1}^{(N)}(t) = \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[0,0,1,0,0,1\right](\mathbf{1}\left[U_i(t+1) > P_{ar}\right] - (1-P_{ar})) \tag{64}$$

Case (x): For $\mathbf{y} \in \mathcal{Y}_{10}$,

$$\begin{aligned} &L_{\mathbf{y}}^{(N)}(t+1) \\ &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[0,0,1,0,0,1\right]\mathbf{1}\left[F_i(t+1) = x\right]\mathbf{1}\left[H_i(t+1) = d\right]\mathbf{1}\left[U_i(t+1) < P_{ar}\right] \\ &\quad -\mathbf{P}_{\mathbf{Y}(t)}\left[0,0,1,0,0,1\right]\mathbf{P}\left[F(t+1) = x\right]\mathbf{P}\left[H(t+1) = d\right]\mathbf{P}\left[U(t+1) < P_{ar}\right] \\ &= D_{\mathbf{y}}^{(N)}(t) + P_{ar}\mathbf{P}\left[F(t+1) = x\right]\mathbf{P}\left[H(t+1) = d\right]L_{0,0,1,0,0,1}^{(N)}(t), \end{aligned} \tag{65}$$

where we set

$$\begin{aligned} D_{\mathbf{y}}^{(N)}(t) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[0,0,1,0,0,1\right](\mathbf{1}\left[U_i(t+1) < P_{ar}\right]\mathbf{1}_{F_i(t+1)}\left[x\right]\mathbf{1}_{H_i(t+1)}\left[d\right] \\ &\quad -P_{ar}p_x q_d). \end{aligned} \tag{66}$$

for $\mathbf{y} \in \mathcal{Y}'$, where $\mathcal{Y}'$ is given in (28).
Case (xi): For $\mathbf{y} \in \mathcal{Y}_{11}$,

$$\begin{aligned} L_{\mathbf{y}}^{(N)}(t+1) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[w, x, s, d, b-1, m\right] \\ &\quad -\mathbf{P}_{\mathbf{Y}(t)}\left[w, x, s, d, b-1, m\right] \\ &= L_{\mathbf{y}'_{\mathbf{11}}}^{(N)}(t), \end{aligned} \tag{67}$$

where $\mathbf{y}'_{\mathbf{11}} = (w, x, s, d, b-1, m)$.

Case (xii): For $\mathbf{y} \in \mathcal{Y}_{12}$,

$$\begin{aligned} L_{\mathbf{y}}^{(N)}(t+1) &= \tfrac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[0,x,1,d,d-1,1\right] \\ &\quad -\mathbf{P}_{\mathbf{Y}(t)}\left[0,x,1,d,d-1,1\right] \\ &= L_{\mathbf{y'_{12}}}^{(N)}(t), \end{aligned} \tag{68}$$

where $\mathbf{y'_{12}} = (0,x,1,d,d-1,1)$.
Case (xiii): For $\mathbf{y} \in \mathcal{Y}_{13}$,

$$\begin{aligned} L_{\mathbf{y}}^{(N)}(t+1) &= \sum_{w'=1}^{W_{\max}} \sum_{x'=1}^{w'} \big(\tfrac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[w',x',s,d,d,0\right] M_{i,new}^{(N)}(t+1) \\ &\quad -\mathbf{E}\Big[\mathbf{1}_{\mathbf{Y}(t)}\left[w',x',s,d,d,0\right] M_{new}(t+1)\Big] \\ &\quad +\tfrac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[w',x',s,d,d,1\right] M_{i,new}^{(N)}(t+1) \\ &\quad -\mathbf{E}\Big[\mathbf{1}_{\mathbf{Y}(t)}\left[w',x',s,d,d,1\right] M_{new}(t+1)\Big]\big) \\ &= \sum_{w'=1}^{W_{\max}} \sum_{x'=1}^{w'} \big(\gamma(t)^{x'} (L_{w',x',s,d,d,0}^{(N)}(t) + L_{w',x',s,d,d,1}^{(N)}(t)) \\ &\quad +(A+B)_{w',x',s,d,d,0}^{(N)}(t) + (A+B)_{w',x',s,d,d,1}^{(N)}(t)\big) \ . \end{aligned} \tag{69}$$

Case (xiv): For $\mathbf{y} \in \mathcal{Y}_{14}$,

$$\begin{aligned} L_{\mathbf{y}}^{(N)}(t+1) &= \sum_{w'=1}^{W_{\max}} \sum_{x'=1}^{w'} \big(\tfrac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[w',x',0,d,d,0\right] (1 - M_{i,new}^{(N)}(t+1)) \\ &\quad -\mathbf{E}\Big[\mathbf{1}_{\mathbf{Y}(t)}\left[w',x',0,d,d,0\right] (1 - M_{new}(t+1))\Big] \\ &\quad +\tfrac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[w',x',0,d,d,1\right] (1 - M_{i,new}^{(N)}(t+1)) \\ &\quad -\mathbf{E}\Big[\mathbf{1}_{\mathbf{Y}(t)}\left[w',x',0,d,d,1\right] (1 - M_{new}(t+1))\Big] \\ &\quad +\tfrac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[w',x',1,d,d,0\right] (1 - M_{i,new}^{(N)}(t+1)) \\ &\quad -\mathbf{E}\Big[\mathbf{1}_{\mathbf{Y}(t)}\left[w',x',1,d,d,0\right] (1 - M_{new}(t+1))\Big] \\ &\quad +\tfrac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[w',x',1,d,d,1\right] (1 - M_{i,new}^{(N)}(t+1)) \\ &\quad -\mathbf{E}\Big[\mathbf{1}_{\mathbf{Y}(t)}\left[w',x',1,d,d,1\right] (1 - M_{new}(t+1))\Big]\big) \\ &= \sum_{w'=1}^{W_{\max}} \sum_{x'=1}^{w'} (1 - \gamma(t)^{x'})(L_{w',x',0,d,d,0}^{(N)}(t) + L_{w',x',0,d,d,1}^{(N)}(t) \\ &\quad +L_{w',x',1,d,d,0}^{(N)}(t) + L_{w',x',1,d,d,1}^{(N)}(t)) \\ &\quad -(A+B)_{w',x',0,d,d,0}^{(N)}(t) - (A+B)_{w',x',0,d,d,1}^{(N)}(t) \\ &\quad -(A+B)_{w',x',1,d,d,0}^{(N)}(t) - (A+B)_{w',x',1,d,d,1}^{(N)}(t). \end{aligned} \tag{70}$$

Denote

$$\mathbf{L}_{\mathcal{Y}}^{(N)}(t) = (L_{\mathbf{y}}^{(N)}(t), \mathbf{y} \in \mathcal{Y}) \ .$$

Then, from the above decompositions we can now write

$$L_{\mathbf{y}}^{(N)}(t+1) = \Phi(\mathbf{L}_{\mathcal{Y}}^{(N)}(t), \gamma(t); \mathbf{y}) + \Psi(\mathbf{A}^{(N)}(t) + \mathbf{B}^{(N)}(t); \mathbf{y}) + \Gamma(\mathbf{D}^{(N)}(t); \mathbf{y}), \tag{71}$$

for $\mathbf{y} \in \mathcal{Y}$.

D A Proof of Proposition 2

The proof of Proposition 2 is established with the help of the following technical results. Define the following random vectors:

$$\mathbf{A}^{(N)}(t) = \left(A_{\mathbf{y}}^{(N)}(t),\ \mathbf{y} \in \mathcal{Y}\right)$$
$$\mathbf{B}^{(N)}(t) = \left(B_{\mathbf{y}}^{(N)}(t),\ \mathbf{y} \in \mathcal{Y}\right),$$

where $A_{\mathbf{y}}^{(N)}(t), B_{\mathbf{y}}^{(N)}(t)$ are given in (60) and (61) for $\mathbf{y} \in \mathcal{Y}_7 \cup \mathcal{Y}_8 \cup \mathcal{Y}_{13} \cup \mathcal{Y}_{14}$. Otherwise, set them to zero.

Also,

$$\mathbf{D}^{(N)}(t) = \left(D_{\mathbf{y}}^{(N)}(t),\ \mathbf{y} \in \mathcal{Y}\right),$$

where $D_{0,0,1,0,0,1}^{(N)}(t)$ and $D_{\mathbf{y}}^{(N)}(t),\ \mathbf{y} \in \mathcal{Y}'$ are given in (64) and (66), respectively. Define $D_{\mathbf{y}}^{(N)}(t) = 0$ otherwise.

Let $\mathcal{F}_t$ denote the σ-field generated by the rvs $\{Q^{(N)}(0), W_i^{(N)}(0), X_i^{(N)}(0), U_i(s), F_i(s), H_i(s), V_i(s), V_{i,j}(s)\ ,\ i,j = 1,2,\dots\ ;\ s = 1,\dots,t\ \}$ with the rvs $Q^{(N)}(t)$ and $\mathbf{Y}_i^{(N)}(t)$ $(i = 1,\dots,N)$ being $\mathcal{F}_t$-measurable. Under the enforced independence assumptions

$$\mathbf{E}\left[M_{i,j}^{(N)}(t+1)|\mathcal{F}_t\right] = 1 - f^{(N)}(\hat{Q}^{(N)}(t)), \quad j = 1,2,\dots,A_i^{(N)}(t)$$

and

$$\mathbf{E}\left[M_{i,new}^{(N)}(t+1)|\mathcal{F}_t\right] = Z_i^{(N)}(t) \tag{72}$$

where

$$Z_i^{(N)}(t) = \left(1 - f^{(N)}(\hat{Q}^{(N)}(t))\right)^{A_i^{(N)}(t)} \tag{73}$$

by conditional independence.

We first show the marginal convergence of the random vectors $\mathbf{B}^{(N)}(t)$ and $\mathbf{D}^{(N)}(t)$ from the following results.

Proposition 3 *Under the conditions of Theorem 2, for any fixed t, the random vector*

$$\sqrt{N}\mathbf{B}^{(N)}(t) \Rightarrow_N \mathbf{B}(t) \tag{74}$$

where $\mathbf{B}(t) = (B_{\mathbf{y}}(t),\ \mathbf{y} \in \mathcal{Y})$ is a $|\mathcal{Y}|$-dimensional Gaussian random vector with mutually independent elements. For each $\mathbf{y} \in \mathcal{Y}$, $B_{\mathbf{y}}(t) \sim \mathcal{N}(0, R_{\mathbf{y}}(t))$ where $R_{\mathbf{y}}(t)$ is given in (26). Moreover, $(B_{\mathbf{y}}(t),\ \mathbf{y} \in \mathcal{Y})$ is independent of $\mathcal{F}_t$.

Proposition 4 *Assume (A1b)-(A3),*

$$\sqrt{N}\mathbf{D}^{(N)}(t) \Rightarrow_N \mathbf{D}(t)$$

where $\mathbf{D}(t)$ is a Gaussian random vector with zero mean and covariance matrix given in (27).

We now show the convergence in law of $\sqrt{N}A_{\mathbf{y}}^{(N)}(t), \mathbf{y} \in \mathcal{Y},\ w \wedge x \geq 1, b = d$.

Lemma 2 *Assume (A1b)-(A3),*

$$\sqrt{N}(A_{\mathbf{y}}^{(N)}(t)) \Rightarrow_N J_{\mathbf{y}}(t)L_0(t), \quad \mathbf{y} \in \mathcal{Y},\ w \wedge x \geq 1, b = d, \tag{75}$$

where $J_{\mathbf{y}}(t)$ is given in (26).

The following lemma establishes the joint convergence that is essential for the proof of Proposition 2.

Lemma 3 *Under Assumption (A1b)-(A3) and* **[E:t]**, *we have the following convergence.*

$$\begin{aligned}&\sqrt{N}\left(L_0^{(N)}(t+1), \hat{L}_0^{(N)}(t+1), L_{\mathbf{y}}^{(N)}(t), A_{\mathbf{y}}^{(N)}(t), B_{\mathbf{y}}^{(N)}(t), D_{\mathbf{y}}^{(N)}(t);\ \mathbf{y} \in \mathcal{Y}\right) \\ &\Rightarrow_N \left(L_0(t+1), \hat{L}_0(t+1), L_{\mathbf{y}}(t), J_{\mathbf{y}}(t)\hat{L}_0(t), B_{\mathbf{y}}(t), D_{\mathbf{y}}(t);\ \mathbf{y} \in \mathcal{Y}\right),\end{aligned} \tag{76}$$

where $D_{\mathbf{y}}(t), B_{\mathbf{y}}(t)$, and $J_{\mathbf{y}}(t)$ are the rvs given in Proposition 4, Proposition 3, and Lemma 2, respectively.

The proofs of Propositions 3, and 4, and Lemma 2 are given in Appendix E, while the proof of Lemma 3 is given in Appendix F.

We are now ready to present the proof of Proposition 2.

Proof of Proposition 2. By the decomposition in (71), for any given $\mathbf{y} \in \mathcal{Y}$ we have

$$L_{\mathbf{y}}^{(N)}(t+1) = \Phi(\mathbf{L}_{\mathcal{Y}}^{(N)}(t), \gamma(t); \mathbf{y}) + \Psi(\mathbf{A}^{(N)}(t) + \mathbf{B}^{(N)}(t); \mathbf{y}) + \Gamma(\mathbf{D}^{(N)}(t); \mathbf{y}).$$

The joint convergence

$$\begin{aligned}&\sqrt{N}\left(L_0^{(N)}(t+1), \hat{L}_0^{(N)}(t+1), L_{\mathbf{y}}^{(N)}(t+1),\ \mathbf{y} \in \mathcal{Y}\right) \\ \Rightarrow_N \quad &\left(L_0(t+1), \hat{L}_0(t+1), L_{\mathbf{y}}(t+1),\ \mathbf{y} \in \mathcal{Y}\right)\end{aligned} \tag{77}$$

is established through Lemma 3 and the continuous mapping theorem.

The distributional recursion of $L_{\mathbf{y}}(t+1),\ \mathbf{y} \in \mathcal{Y}$ is established from

$$\begin{aligned}\sqrt{N}L_{\mathbf{y}}^{(N)}(t+1) \quad &= \quad \sqrt{N}\left(\Phi(\mathbf{L}_{\mathcal{Y}}^{(N)}(t), \gamma(t); \mathbf{y}) + \Psi(\mathbf{A}^{(N)}(t) + \mathbf{B}^{(N)}(t); \mathbf{y}) + \Gamma(\mathbf{D}^{(N)}(t); \mathbf{y})\right) \\ &\Rightarrow_N \quad \Phi(\mathbf{L}_{\mathcal{Y}}(t), \gamma(t); \mathbf{y}) + \Psi(\hat{L}_0(t)\mathbf{J}(t) + \mathbf{B}(t)) + \Gamma(\mathbf{D}(t)).\end{aligned}$$

■

E Proof of Auxiliary Results

E.1 Proof of Proposition 3

The proof of Proposition 3 relies on the following three technical lemmas, whose proofs can be easily obtained and are omitted.

Lemma 4 *For any x in $\mathbb{R}$, the Taylor series expansion*

$$e^{jx} = 1 + jx - \frac{x^2}{2} + R(x) \tag{78}$$

holds, and the complex-valued remainder term $R(x)$ satisfies

$$|R(x)| \leq \frac{|x|^3}{6}. \tag{79}$$

Lemma 5 *[6, Thm. 5.20, p. 146] (Slutsky's theorem) If $X_N \Rightarrow_N X$ and $Y_N \Rightarrow_N y \in \mathbb{R}$, then $X_N Y_N \Rightarrow_N y \cdot X$ and $X_N + Y_N \Rightarrow_N X + y$.*

Lemma 6 *Consider the array of complex-valued rvs $\{C_{N,i},\ i = 1, \ldots, N;\ N = 1, 2, \ldots\}$ with $|C_{N,i}| < 1$ for $i = 1, \ldots, N$. If $\max_{i=1,\ldots,N} |C_{N,i}| \rightarrow_N 0$ a.s. and $\sum_{i=1}^N C_{N,i} \xrightarrow{P}_N \lambda$, then*

$$\prod_{i=1}^N (1 - C_{N,i}) \xrightarrow{P}_N e^{-\lambda}. \tag{80}$$

We now proceed with the proof of Proposition 3: Fix $N = 1, 2, \ldots$ and $\theta_{\mathbf{y}},\ \mathbf{y} \in \mathcal{Y}$, arbitrary in $\mathbb{R}$. It suffices to show that

$$\mathbf{E}\left[\exp\left(j\sqrt{N}\sum_{\mathbf{y}\in\mathcal{Y}} \theta_{\mathbf{y}} B_{\mathbf{y}}^{(N)}(t)\right) |\mathcal{F}_t\right] \xrightarrow{P}_N e^{-\frac{1}{2}\sum_{\mathbf{y}} \theta_{\mathbf{y}}^2 R_{\mathbf{y}}(t)} \tag{81}$$

By conditional independence, we find that

$$\begin{aligned}
&\mathbf{E}\left[\exp\left(j\sqrt{N}\sum_{\mathbf{y}\in\mathcal{Y}} \theta_{\mathbf{y}} B_{\mathbf{y}}^{(N)}(t)\right) |\mathcal{F}_t\right] \\
&= \prod_{i=1}^N \exp\left(-\frac{j}{\sqrt{N}}\sum_{\mathbf{y}\in\mathcal{Y}} \theta_{\mathbf{y}} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[\mathbf{y}\right] \gamma^{(N)}(t)^{w\wedge x}\right) \mathbf{E}\left[\exp\left(\frac{j}{\sqrt{N}}\sum_{\mathbf{y}\in\mathcal{Y}} \theta_{\mathbf{y}} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[\mathbf{y}\right] M_{i,new}^{(N)}(t+1)\right) |\mathcal{F}_t\right] \\
&= \prod_{i=1}^N \exp\left(-\frac{j}{\sqrt{N}}\sum_{\mathbf{y}\in\mathcal{Y}} \theta_{\mathbf{y}} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[\mathbf{y}\right] \gamma^{(N)}(t)^{w\wedge x}\right) \left[1 + \sum_{\mathbf{y}\in\mathcal{Y}} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[\mathbf{y}\right] \gamma^{(N)}(t)^{w\wedge x}(e^{\frac{j}{\sqrt{N}}\theta_{\mathbf{y}}} - 1)\right] \\
&= \prod_{i=1}^N \left[1 - \frac{j}{\sqrt{N}} C_i(t;\theta) - \frac{1}{2N} C_i(t;\theta)^2 + \alpha_i^{(N)}(t;\theta)\right] \left[1 + \frac{j}{\sqrt{N}} C_i(t;\theta) - \frac{1}{2N} C_i(t;\boldsymbol{\theta^2}) + \beta_i^{(N)}(t;\theta)\right],
\end{aligned} \tag{82}$$

by Lemma 4, where

$$\begin{aligned}
C_i(t;\theta) &= \sum_{\mathbf{y}\in\mathcal{Y}} \theta_{\mathbf{y}} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[\mathbf{y}\right] \gamma^{(N)}(t)^{w\wedge x} \\
C_i(t;\boldsymbol{\theta^2}) &= \sum_{\mathbf{y}\in\mathcal{Y}} \theta_{\mathbf{y}}^2 \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[\mathbf{y}\right] \gamma^{(N)}(t)^{w\wedge x}
\end{aligned}$$

and

$$\begin{aligned} \max(|\alpha_i^{(N)}(t,\theta)|, |\beta_i^{(N)}(t,\theta)|) &\leq \sum_{\mathbf{y}\in\mathcal{Y}} \frac{|\theta_{\mathbf{y}}|^3}{6\sqrt{N^3}} \\ &\leq |\mathcal{Y}| \max_{\mathbf{y}\in\mathcal{Y}} \frac{|\theta_{\mathbf{y}}|^3}{6\sqrt{N^3}} \ . \end{aligned} \tag{83}$$

It is easy to see that (82) becomes

$$\prod_{i=1}^{N}[1 - \frac{1}{2N}(C_i(t;\theta^{\mathbf{2}}) - C_i(t;\theta)^2) + \xi_i^{(N)}(t;\theta)] = \prod_{i=1}^{N}[1 - C_i^{(N)}(t)],$$

where we set $C_i^{(N)}(t) = \frac{1}{2N}(C_i(t;\theta) - C_i(t;\theta)^2) - \xi_i^{(N)}(t;\theta)$ and the remainder term $\xi_i^{(N)}(t;\theta)$ satisfies

$$\begin{aligned} \xi_i^{(N)}(t;\theta) \ &\leq \ \left|1 + \frac{j}{\sqrt{N}}C_i(t;\theta) - \frac{1}{2N}C_i(t;\theta^{\mathbf{2}})\right| \left|\alpha_i^{(N)}(t;\theta)\right| \\ &\quad + \left|1 - \frac{j}{\sqrt{N}}C_i(t;\theta) - \frac{1}{2N}C_i(t;\theta)^2\right| \left|\beta_i^{(N)}(t;\theta)\right| + \left|\alpha_i^{(N)}(t;\theta)||\beta_i^{(N)}(t;\theta)\right| \\ &\quad + \left|\frac{j}{2N^{3/2}}C_i(t;\theta^{\mathbf{2}})C_i(t;\theta) - \frac{j}{2N^{3/2}}C_i(t;\theta)^3 + \frac{1}{4N^2}C_i(t;\theta)^2C_i(t;\theta^{\mathbf{2}})\right|. \end{aligned}$$

Note that

$$\begin{aligned} & \max\Big(|C_i(t;\theta)|, |C_i(t;\theta^{\mathbf{2}})|, |C_i(t;\theta)^2|, |C_i(t;\theta^{\mathbf{2}})C_i(t;\theta)|, |C_i(t;\theta)^3|, |C_i(t;\theta)^2C_i(t;\theta^{\mathbf{2}})|\Big) \\ \leq \ & \max\left(\sum_{\mathbf{y}\in\mathcal{Y}}|\theta_{\mathbf{y}}|, \sum_{\mathbf{y}\in\mathcal{Y}}|\theta_{\mathbf{y}}|^2, (\sum_{\mathbf{y}\in\mathcal{Y}}|\theta_{\mathbf{y}}|)^2, \sum_{\mathbf{y}\in\mathcal{Y}}|\theta_{\mathbf{y}}| \cdot \sum_{\mathbf{y}\in\mathcal{Y}}|\theta_{\mathbf{y}}|^2, (\sum_{\mathbf{y}\in\mathcal{Y}}|\theta_{\mathbf{y}}|)^3, (\sum_{\mathbf{y}\in\mathcal{Y}}|\theta_{\mathbf{y}}|)^2 \cdot \sum_{\mathbf{y}\in\mathcal{Y}}|\theta_{\mathbf{y}}|^2\right) \\ := \ & C_{max}, \end{aligned}$$

which only depends on $\theta_{\mathbf{y}}, \ \mathbf{y} \in \mathcal{Y}$, and does not depend on N.

$$\begin{aligned} \xi_i^{(N)}(t;\theta) \ &\leq \ \left(\left(1 + \frac{C_{max}}{2N}\right)^2 + \left(\frac{C_{max}}{\sqrt{N}}\right)^2\right)^{\frac{1}{2}} (|\alpha_i^{(N)}(t;\theta)| + |\beta_i^{(N)}(t;\theta)|) \\ &\quad + \left(\left(\frac{C_{max}}{4N^2}\right)^2 + \left(\frac{2C_{max}}{2N^{3/2}}\right)^2\right)^{\frac{1}{2}} + |\alpha_i^{(N)}(t;\theta)||\beta_i^{(N)}(t;\theta)| \\ &\leq \ 2\left(1 + \frac{C_{max}}{N} + \left(\frac{1}{4N^2} + \frac{1}{N}\right)C_{max}^2\right) \max(|\alpha_i^{(N)}(t;\theta)|, |\beta_i^{(N)}(t;\theta)|) \\ &\quad + \left(\left(\frac{C_{max}}{4N^2}\right)^2 + \left(\frac{2C_{max}}{2N^{3/2}}\right)^2\right)^{\frac{1}{2}} + \max(|\alpha_i^{(N)}(t;\theta)|, |\beta_i^{(N)}(t;\theta)|)^2. \end{aligned} \tag{84}$$

The desired result (81) is now a simple consequence of Lemma 6 provided that the required conditions can be shown to hold, *i.e.*,

$$\lim_{N\to\infty} \max_{i=1,\ldots,N} |C_i^{(N)}(t)| = 0 \quad a.s. \tag{85}$$

and

$$\sum_{i=1}^{N} C_i^{(N)}(t) \xrightarrow{P}_N \sum_{\mathbf{y}\in\mathcal{Y}} \theta_{\mathbf{y}}^2 R_{\mathbf{y}}(t). \tag{86}$$

Condition (85) trivially holds while Condition (86) can be established from

$$\begin{aligned}\sum_{i=1}^{N} C_i^{(N)}(t) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(C_i(t;\theta^{\mathbf{2}}) - C_i(t;\theta)^2) - \tfrac{1}{N}\sum_{i=1}^{N}\xi_i^{(N)}(t;\theta) \\ &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\displaystyle\sum_{\mathbf{y}\in\mathcal{Y}} \theta_{\mathbf{y}}^2 \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}]\cdot\gamma^{(N)}(t)^{w\wedge x}(1-\gamma^{(N)}(t)^{w\wedge x}) \\ &\quad -\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\xi_i^{(N)}(t;\theta) \\ &\xrightarrow{P}_N \sum_{\mathbf{y}\in\mathcal{Y}} \theta_{\mathbf{y}}^2 R_{\mathbf{y}}(t), \end{aligned} \tag{87}$$

where the convergence follows from Theorem 1, Lemma 5 and

$$\lim_{N\to\infty} \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\xi_i^{(N)}(t;\theta) = 0 \quad a.s.$$

which is a simple consequence of (83) and (84). ■

E.2 Proof of Lemma 2

Lemma 2 is a corollary of Lemma 1.

For each $\mathbf{y}\in\mathcal{Y}, w\wedge x \geq 1, b = d$,

$$\begin{aligned}\sqrt{N}A_{\mathbf{y}}^{(N)}(t) &= \sqrt{N}\left(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}]\,(\gamma^{(N)}(t)^{w\wedge x} - \gamma(t)^{w\wedge x}\right) \\ &= \sqrt{N}\left(\gamma^{(N)}(t)^{w\wedge x} - \gamma(t)^{w\wedge x}\right)\cdot\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}] \\ &\Rightarrow_N (w\wedge x)\gamma(t)^{(w\wedge x)-1} f'(\hat{q}(t))\hat{L}_0(t)\cdot\mathbf{P}_{\mathbf{Y}(t)}[\mathbf{y}] \end{aligned} \tag{88}$$

where the convergence follows directly from Lemma 1 and Lemma 5. ■

E.3 Proof of Proposition 4

Denote $d_i^{(N)} = \mathbf{1}\left[\mathbf{Y}_i^{(N)}(t) = (0,0,1,0,0,1)\right]$

$$\begin{aligned}&\mathbf{E}\left[e^{j\sqrt{N}\cdot\frac{1}{N}\sum_{i=1}^{N}d_i^{(N)}(\theta_0(\mathbf{1}[U_i(t+1)>P_{ar}]-(1-P_{ar}))+\sum_{\mathbf{y}\in\mathcal{Y}'}\theta_{\mathbf{y}}(\mathbf{1}[U_i(t+1)<P_{ar}]\mathbf{1}_{F(t+1),H(t+1)}[x,d]-p_xq_dP_{ar}))}|\mathcal{F}_t\right] \\ &= \prod_{i=1}^{N}\exp(-\frac{j}{\sqrt{N}}d_i^{(N)}((1-P_{ar})\theta_0 + P_{ar}\sum_{\mathbf{y}\in\mathcal{Y}'}p_xq_d\theta_{\mathbf{y}})) \\ &\quad\times\mathbf{E}\left[e^{\frac{j}{\sqrt{N}}d_i^{(N)}(\theta_0\mathbf{1}[U_i(t+1)>P_{ar}]+\sum_{\mathbf{y}\in\mathcal{Y}'}\theta_{\mathbf{y}}\mathbf{1}_{F_i(t+1),H_i(t+1)}[x,d]\mathbf{1}[U_i(t+1)<P_{ar}])}|\mathcal{F}_t\right] \\ &= \prod_{i=1}^{N}\exp(-\frac{j}{\sqrt{N}}d_i^{(N)}((1-P_{ar})\theta_0 + P_{ar}\sum_{\mathbf{y}\in\mathcal{Y}'}p_xq_d\theta_{\mathbf{y}})) \\ &\quad\times[(1-P_{ar})e^{\frac{j}{\sqrt{N}}d_i^{(N)}\theta_0} + P_{ar}\sum_{\mathbf{y}\in\mathcal{Y}'}p_xq_de^{\frac{j}{\sqrt{N}}d_i^{(N)}\theta_{\mathbf{y}}}]\end{aligned} \tag{89}$$

Following a similar analysis to (82)-(85) in the proof of Proposition 3, we can show that (89) converges in probability as $N \to \infty$ to

$$e^{\left(-\frac{1}{2}\mathbf{P}_{\mathbf{Y}(t)}[0,0,1,0,0,1]\left[\left((1-P_{ar})\theta_0^2+P_{ar}\sum_{\mathbf{y}\in\mathcal{Y}'}p_x q_d \theta_{\mathbf{y}}^2\right)-\left((1-P_{ar})\theta_0+P_{ar}\sum_{\mathbf{y}\in\mathcal{Y}'}p_x q_d\theta_{\mathbf{y}}\right)^2\right]\right)} \tag{90}$$

which is a characteristic function of a Gaussian random vector with zero mean and covariance as described in (27). ■

F A Proof of Lemma 3

To establish the joint convergence of the random vector, we rely on the following lemma:

Lemma 7 *[6, p. 150] (Cramer-Wold device) The convergence in (76) holds if and only if for all choices of $\theta_0, \hat{\theta}_0, \theta_{\mathbf{y}}, \alpha_{\mathbf{y}}, \beta_{\mathbf{y}}, \omega_{\mathbf{y}},\ \mathbf{y} \in \mathcal{Y}$, we have*

$$\begin{aligned}&\sqrt{N}\left(\theta_0 L_0^{(N)}(t+1) + \hat{\theta}_0 \hat{L}_0^{(N)}(t+1) + \sum_{\mathbf{y}\in\mathcal{Y}} \theta_{\mathbf{y}} L_{\mathbf{y}}^{(N)}(t) + \alpha_{\mathbf{y}} A_{\mathbf{y}}^{(N)}(t) + \beta_{\mathbf{y}} B_{\mathbf{y}}^{(N)}(t) + \omega_{\mathbf{y}} D_{\mathbf{y}}^{(N)}(t)\right)\\ \Rightarrow_N\ \ &\theta_0 L_0(t+1) + \hat{\theta}_0 \hat{L}_0(t+1) + \sum_{\mathbf{y}\in\mathcal{Y}} \theta_{\mathbf{y}} L_{\mathbf{y}}(t) + \alpha_{\mathbf{y}} J_{\mathbf{y}}^{(N)}(t)\hat{L}_0(t) + \beta_{\mathbf{y}} B_{\mathbf{y}}(t) + \omega_{\mathbf{y}} D_{\mathbf{y}}(t).\end{aligned} \tag{91}$$

By Lemma 7, it suffices to show that

$$\begin{aligned}&\mathbf{E}\left[e^{j\sqrt{N}\left(\theta_0 L_0^{(N)}(t+1)+\hat{\theta}_0\hat{L}_0^{(N)}(t+1)+\sum_{\mathbf{y}\in\mathcal{Y}}\theta_{\mathbf{y}}L_{\mathbf{y}}^{(N)}(t)+\alpha_{\mathbf{y}}A_{\mathbf{y}}^{(N)}(t)+\beta_{\mathbf{y}}B_{\mathbf{y}}^{(N)}(t)+\omega_{\mathbf{y}}D_{\mathbf{y}}^{(N)}(t)\right)}\right]\\ &\to_N\ \ e^{-\frac{1}{2}\Omega'\mathbf{V}(t)\Omega}\cdot e^{-\frac{1}{2}\sum_{\mathbf{y}\in\mathcal{Y}}\beta_{\mathbf{y}}^2 R_{\mathbf{y}}(t)}\mathbf{E}\left[e^{j\left(\theta_0 L_0(t+1)+\hat{\theta}_0\hat{L}_0(t+1)+\sum_{\mathbf{y}\in\mathcal{Y}}\theta_{\mathbf{y}}L_{\mathbf{y}}(t)+\alpha_{\mathbf{y}}J_{\mathbf{y}}(t)\hat{L}_0(t)\right)}\right],\end{aligned} \tag{92}$$

where Ω is the properly enumerated vector of $\omega_{\mathbf{y}}, \mathbf{y} \in \mathcal{Y}$, and $\mathbf{V}(t)$ is given in (27) (also enumerated with the corresponding $\omega_{\mathbf{y}}$).

First note that

$$\begin{aligned}&\mathbf{E}\left[e^{j\sqrt{N}\left(\theta_0 L_0^{(N)}(t+1)+\hat{\theta}_0\hat{L}_0^{(N)}(t+1)+\sum_{\mathbf{y}\in\mathcal{Y}}\theta_{\mathbf{y}}L_{\mathbf{y}}^{(N)}(t)+\alpha_{\mathbf{y}}A_{\mathbf{y}}^{(N)}(t)+\beta_{\mathbf{y}}B_{\mathbf{y}}^{(N)}(t)+\omega_{\mathbf{y}}D_{\mathbf{y}}^{(N)}(t)\right)}\right]\\ &= \mathbf{E}\left[e^{j\sqrt{N}\left(\theta_0 L_0^{(N)}(t+1)+\hat{\theta}_0\hat{L}_0^{(N)}(t+1)+\sum_{\mathbf{y}\in\mathcal{Y}}\theta_{\mathbf{y}}L_{\mathbf{y}}^{(N)}(t)+\alpha_{\mathbf{y}}A_{\mathbf{y}}^{(N)}(t)\right)}\cdot\mathbf{E}\left[e^{j\sqrt{N}\sum_{\mathbf{y}\in\mathcal{Y}}\beta_{\mathbf{y}}B_{\mathbf{y}}^{(N)}(t)+\omega_{\mathbf{y}}D_{\mathbf{y}}^{(N)}(t)}|\mathcal{F}_t\right]\right]\\ &\to_N\ \ \exp\left(-\frac{1}{2}\Omega'\mathbf{V}(t)\Omega\right)\cdot\exp\left(-\frac{1}{2}\sum_{\mathbf{y}\in\mathcal{Y}}\beta_{\mathbf{y}}^2 R_{\mathbf{y}}(t)\right)\\ &\quad\cdot\mathbf{E}\left[\lim_{N\to\infty}\exp\left(j\sqrt{N}\left(\theta_0 L_0^{(N)}(t+1)+\hat{\theta}_0\hat{L}_0^{(N)}(t+1)+\sum_{\mathbf{y}\in\mathcal{Y}}\theta_{\mathbf{y}}L_{\mathbf{y}}^{(N)}(t)+\alpha_{\mathbf{y}}A_{\mathbf{y}}^{(N)}(t)\right)\right)\right],\end{aligned}$$

where the convergence follows from Propositions 3 and 4. The desired result (92) follows if we can establish the following joint convergence from Cramer-Wold device

$$\begin{aligned}
&\sqrt{N}\left(\theta_0 L_0^{(N)}(t+1)+\hat{\theta}_0\hat{L}_0^{(N)}(t+1)+\sum_{\mathbf{y}\in\mathcal{Y}}\theta_{\mathbf{y}}L_{\mathbf{y}}^{(N)}(t)+\alpha_{\mathbf{y}}A_{\mathbf{y}}^{(N)}(t)\right)\\
\Rightarrow_N\quad&\left(\theta_0 L_0(t+1)+\hat{\theta}_0\hat{L}_0(t+1)+\sum_{\mathbf{y}\in\mathcal{Y}}\theta_{\mathbf{y}}L_{\mathbf{y}}(t)+\alpha_{\mathbf{y}}J_{\mathbf{y}}(t)\hat{L}_0(t)\right). \qquad (93)
\end{aligned}$$

Notice that for all $y\in\mathcal{Y}$ such that $w\wedge x\geq 1$ and $b=d$,

$$\begin{aligned}
\sqrt{N}A_{\mathbf{y}}^{(N)}(t) &= \sqrt{N}\cdot\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[\mathbf{y}\right]\left(\gamma^{(N)}(t)^{w\wedge x}-\gamma(t)^{w\wedge x}\right)\\
&= \sqrt{N}\cdot\mathbf{P}_{\mathbf{Y}(t)}\left[\mathbf{y}\right]\left(\gamma^{(N)}(t)^{w\wedge x}-\gamma(t)^{w\wedge x}\right) \qquad (94)\\
&\quad+\sqrt{N}\left(\gamma^{(N)}(t)^{w\wedge x}-\gamma(t)^{w\wedge x}\right)\cdot\left(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[\mathbf{y}\right]-\mathbf{P}_{\mathbf{Y}(t)}\left[\mathbf{y}\right]\right)
\end{aligned}$$

Since $\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[\mathbf{y}\right]-\mathbf{P}_{\mathbf{Y}(t)}\left[\mathbf{y}\right]\xrightarrow{P}_N 0$ from Theorem 1, by applying Lemma 1 and Lemma 5, the second term converges to zero in distribution. Therefore, we can prove the proposition by showing instead

$$\begin{aligned}
&\sqrt{N}\left(\theta_0 L_0^{(N)}(t+1)+\hat{\theta}_0\hat{L}_0^{(N)}(t+1)+\sum_{\mathbf{y}\in\mathcal{Y}}\left(\theta_{\mathbf{y}}L_{\mathbf{y}}^{(N)}(t)+\alpha_{\mathbf{y}}\mathbf{P}_{\mathbf{Y}(t)}\left[\mathbf{y}\right]\left(\gamma^{(N)}(t)^{w}-\gamma(t)^{w}\right)\right)\right)\\
&\Rightarrow_N \theta_0 L_0(t+1)+\hat{\theta}_0\hat{L}_0(t+1)+\sum_{\mathbf{y}\in\mathcal{Y}}\left(\theta_{\mathbf{y}}L_{\mathbf{y}}(t)+\alpha_{\mathbf{y}}J_{\mathbf{y}}(t)\hat{L}_0(t)\right), \qquad (95)
\end{aligned}$$

From the proof of Proposition 1, we see that $\sqrt{N}L_0^{(N)}(t+1)$ can be written as a continuous map of $\sqrt{N}\left(L_0^{(N)}(t)+\sum_{\mathbf{y}\in\mathcal{Y}}\min(w,x)\mathbf{1}\left[b\geq d\right]L_{\mathbf{y}}^{(N)}(t)\right)$, regardless of the value of the residual capacity $K(t)$. Meanwhile, $\hat{L}_0^{(N)}(t+1)$ is a convex combination of $\hat{L}_0^{(N)}(t)$ and $L_0^{(N)}(t+1)$. Therefore, (93) follows directly from the continuous mapping theorem and **[E:t]**. ■