Characterization of Queue Fluctuations in Probabilistic AQM Mechanisms

Peerapol Tinnakornsrisuphap *
QUALCOMM, Inc.
5775 Morehouse Drive
San Diego, CA, 92121
peerapol@qualcomm.com

Richard J. La
Dept. of Electrical and Computer Engineering
and Institute for Systems Research
University of Maryland
College Park, MD, 20742
hyongla@eng.umd.edu

ABSTRACT

We develop a framework for studying the interaction of an active queue management (AQM) scheme with a generic end-user congestion-control mechanism. As the number of flows in the network increases, the queue dynamics can be accurately approximated by a simple deterministic process. In addition, we investigate the sources of queue fluctuations in a probabilistic AQM scheme interacting with responsive flows. We demonstrate that there are two distinct sources of queue fluctuations; one is the deterministic oscillations which can be predicted by the aforementioned deterministic process. The other source is the random fluctuations introduced by the probabilistic nature of the marking schemes. We discuss the relationship between these two types of fluctuations and provide insights into how to control them. Concrete examples in this framework are provided for several popular algorithms such as Random Early Detection, Random Early Markings and Transmission Control Protocols.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Modeling techniques; G.3 [**Probability and Statistics**]: Stochastic processes, Probablistic algorithms

General Terms

Algorithms, Performance, Theory

Keywords

Active Queue Management, queue fluctuations, Central Limit Theorem

*This work is performed while the author was with the Deparment of Electrical and Computer Engineering, University of Maryland.

SIGMETRICS/Performance'04, June 12–16, 2004, New York, NY, USA.

With the growing size and popularity of the Internet, there has been an increasing interest in modeling and understanding large-scale network traffic as accurate modeling of large-scale network traffic is critical to the problem of traffic management and control in best-effort networks.

Recently, Kelly [13] has suggested that the problem of rate allocation for elastic traffic can be posed as one of achieving maximum aggregate utility of the users and proposed an optimization framework for rate allocation in best-effort networks. Under this framework, he has shown that the system optimum is achieved at the equilibrium between the end users and resources. Based on this observation researchers have proposed various end user congestion control algorithms, *e.g.,* proportional-fair congestion-controller (PFCC), in conjunction with a variety of forms of active queue management (AQM) mechanisms, *e.g.,* Random Early Marking (REM) and Adaptive Virtual Queue (AVQ). They solve either the system optimization problem or its relaxation [1, 13, 15].

A network with an AQM mechanism can be viewed as a feedback system, where end users adjust their transmission rates based on the feedback from the AQM mechanism in the form of dropped or marked packets [6]. Existing literatures on probabilistic AQM traffic modeling typically focus on developing a detailed model for the interaction between the Transmission Control Protocol (TCP) [9] and Random Early Detection (RED), as they are the most deployed congestion-control mechanism and AQM in the Internet [2, 22]. However, to the best of our knowledge, the interaction of other newly proposed end-user algorithms and AQM mechanisms has been studied mostly in a control-theoretic framework, using a *deterministic* model [8, 15, 17]. In addition, most of the efforts were focused on the stability analysis of the deterministic model or the convergence of user rates to a desired equilibrium point.

Many of these AQM mechanisms, however, adopt a probabilistic marking mechanism, in which the marking probability of a packet depends on the current estimate of congestion level at the bottleneck link. Although the previously studied deterministic models may be a reasonable approximation of the system when the number of flows is large and the granularity of the feedback information is sufficiently fine, they cannot capture the detailed (packet-level) dynamics of the system or the probabilistic nature of the marking mechanism.

In this paper we develop a detailed discrete-time stochas-

tic model for studying the interaction of an AQM mechanism with a generic end-user congestion-control mechanism. In contrast to the existing models, our model captures the detailed packet-level dynamics of the interaction and the probabilistic nature of the marking mechanism. We use this model to investigate the behavior of the queue size at a bottleneck link and to identify the sources of queue fluctuations as the number of flows in the system becomes large. In general, the traffic modeling and resource allocation problems are more interesting when the resource is shared by many users.

It is worth noting that accurate modeling of a large number of flows requires modeling of complex dynamics rising from the details of protocols and the interaction of the end-user algorithms and the network layer, *i.e.,* AQM mechanisms. As the size of the state space required to model the system explodes with the number of flows, this represents a major obstacle to modeling the interaction of many 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.

We will show that as the number of flows becomes large, the AQM queue dynamics can be accurately approximated by a sum of a deterministic process and a stochastic process. Here the deterministic process represents the average or expected behavior of the queue, while the stochastic process captures the random fluctuations in the queue size. We demonstrate that the recursion of the deterministic process depends only on the capacity of the bottleneck link and the *expected* traffic arrival. Moreover, this average traffic arrival rate (with a large number of flows) is closely related to the average arrival rate of a single flow utilizing the same (and properly scaled) congestion-control mechanism. This justifies the use of a deterministic feedback system model to study the *expected* queue behavior. The fluctuation of the deterministic queue process can be predicted and analyzed using a control-theoretic approach (see [8] for example).

The characterization of the random fluctuation in the queue size, which cannot be captured by any of the deterministic models, reveals several interesting points. In contrast to the claim made in [17], although the stability of the deterministic system representing the deterministic process of the queue behavior does not depend on the details of the end user protocol, the magnitude of the random queue oscillation *does* depend on the details of the protocol. In addition, some of the variables used for computing the packet marking probability affect both the fluctuation of the deterministic process and the random fluctuation in queue size.

This random fluctuation, which can be well-approximated by a Gaussian random variable (rv) except in some rare cases, originates from the random marking mechanism in a probabilistic AQM and can be attributed to the following causes. The first cause is the nature of the feedback information; since the feedback information from the AQM mechanism is piggybacked in the packets/acknowledgements and the number of transmitted packets is finite over a measurement period, *e.g.,* a round-trip time (RTT), the granularity of the feedback information is limited by the number of packets acknowledged during the measurement period. The effects of this limitation cannot be investigated in the model without detailed packet level operation. To the best of our knowledge, this is the first analysis that captures such a fluctuation. The second major cause is the discrepancy of the actual feedback information and the limiting feedback information determined by the deterministic queue process. The magnitude of this fluctuation is influenced by the sensitivity of the marking probability function to the variations in the parameters on which the function depends.

This paper is organized as follows: First, Section 1 describes the generic models for end-user algorithms and AQM mechanisms in this framework. Section 2 presents our results on the behavior of the average queue size per flow, which is followed by the description of the random fluctuations in queue size in Section 3. Section 4 describes three main sources of the random fluctuations in queue size and their implications. We give an example of TCP-RED in Section 5 and also explain our results in the context. The paper then concludes in Section 6.

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 $\mathbf{1}[A]$, 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. We write $X^{(N)}$ to indicate the explicit dependence of the quantity X on the number N of flows. An expectation of a rv X is given by $\mathbf{E}[X]$. For simplicity, we introduce the notation $\mathbf{1}_X[x]$ and $\mathbf{P}_X[x]$ for $\mathbf{1}[X = x]$ and $\mathbf{P}[X = x]$, respectively. For some positive integers n, m, $\mathbf{y} = [y_1 \dots y_n]^T \in \mathbf{R}^n$ and a mapping $g : \mathbf{R}^n \to \mathbf{R}^m$, *i.e.,* $g(\mathbf{y}) = [g_1(\mathbf{y}) \dots g_m(\mathbf{y})]$, denote

$$\partial g(y)/\partial y^T = \begin{bmatrix} \frac{\partial g_1(y)}{\partial y_1} & \cdots & \frac{\partial g_1(y)}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial g_m(y)}{\partial y_1} & \cdots & \frac{\partial g_m(y)}{\partial y_n} \end{bmatrix},$$

given that all the partial derivatives exist. The term $\partial g(y)/\partial y^T$ is the Jacobian matrix representing the sensitivity of g to its parameters in the neighborhood of y.

1. THE MODEL

Time is assumed to be discrete and slotted into contiguous timeslots. While we assume in this paper that connections have the same RTT and a timeslot equals the RTT of the connections for simplicity of presentation, it is possible to extend the model and analysis to incorporate heterogeneous round-trip delays of connections into the model (see the approach taken in [21], for example). However, our previous work in [21] suggested that, in the case of TCP-RED, the heterogeneous delays of connections do not significantly change the *qualitative* results compared to the case with homogeneous round-trip delays discussed in [22]. This will be discussed in more details in Section 5.

We consider a simple network with a single bottleneck AQM gateway and N traffic flows. Each of these N traffic sources utilizes the same congestion-control mechanism (to be specified later). The congestion-control mechanism is assumed to be window-based and ECN-capable, *i.e.,* the transmission rate is controlled by the congestion window size which reacts to the ECN marks received in each round-trip. A rate-based congestion-control algorithm typically can also be approximated by a window-based algorithm and vice versa [14]. The capacity of this bottleneck link is NC packets/slot for some positive constant C.[1] The AQM buffer

[1]Notice that we are interested in the queue dynamics when

is modeled as an infinite queue, so that no packet losses occur due to buffer overflow, and congestion-control is achieved solely through the random marking algorithm at the AQM gateway.

1.1 Congestion-Control Mechanisms

The congestion window size of a connection can take values in a finite set $\mathcal{W} := \{w_1, w_2, \ldots, w_{max}\}$ where the elements $w_i \in [1, W_{\max}]$, $i = 1, 2, \ldots$, are not necessarily integers. The constant $W_{\max}$ is a finite integer representing the maximum window size (in packets) of the connection. Here the maximum window size could, for instance, represent the receiver's buffer size and is independent of the congestion-control mechanism.

Fix $i = 1, \ldots, N$. Let a $\mathcal{W}$-valued rv $W_i^{(N)}(t)$ encode the congestion window size of connection i at the beginning of timeslot $[t, t+1)$ and let $\mathbf{Y}_i^{(N)}(t)$ be a vector of the state variables of connection i (including $W_i^{(N)}(t)$) at the beginning of the timeslot $[t, t+1)$. We assume that $\mathbf{Y}_i^{(N)}(t)$ takes a value in a discrete, finite state space $\mathcal{Y}$.

We assume that the congestion window size and other state variables are updated once at the end of every round-trip. Consequently, $\mathbf{Y}_i^{(N)}(t+1)$ evolves according to a mapping $\Xi : \mathcal{Y} \times \{0, 1, \ldots, W_{\max}\} \to \mathcal{Y}$, *i.e.,*

$$Y_i^{(N)}(t+1) = \Xi\left(Y_i^{(N)}(t), M_i^{(N)}(t+1)\right) , \tag{1}$$

where $M_i^{(N)}(t+1)$ represents the number of marks received in the acknowledgments during the timeslot $[t, t+1)$. The rv $M_i^{(N)}(t+1)$ depends on the AQM mechanism implemented at the gateway (to be specified in Section 1.3).

For example, in the case of TCP Reno congestion-control mechanism [9], the state variable $\mathbf{Y}_i^{(N)}(t)$ consists only of the congestion window size $W_i^{(N)}(t)$. [2] Therefore, $\mathcal{Y}_{TCP} = \mathcal{W}_{TCP} = \{1, 2, \ldots, W_{\max}\}$ and $\mathbf{Y}_i^{(N)}(t) = W_i^{(N)}(t)$, $i = 1, \ldots, N$, $t = 0, 1, \ldots$. TCP Reno congestion-control mechanism can be described by the following window size updating rule

$$\Xi_{TCP}(w, m) = \min(w+1, W_{\max})\, \mathbf{1}[m = 0] + \lceil \frac{w}{2} \rceil \mathbf{1}[m > 0] , \tag{2}$$

for any $w \in \mathcal{W}_{TCP}$ and $m = 0, 1, \ldots, w$.

Equation (2) emulates the TCP congestion-control mechanism as follows: If no packet from source i is marked in the timeslot $[t, t+1)$ then the congestion window size in the next timeslot is increased by one packet. On the other hand, if one or more packets are marked in the timeslot $[t, t+1)$, then the congestion window size in the next timeslot is reduced by half. The size of the congestion window is limited by the maximum window size $W_{\max}$.

Another example of a popular congestion-control mechanism is the proportional-fair congestion-controller (PFCC) considered by Kelly [12] and later by Low [16] and Kunniyur and Srikant [14]. In this type of congestion-control mechanism, the size of the window in the next timeslot also depends on the actual number of marks received. A pseudo-code for such an algorithm is given by

```
every ACK do
   W <- W + k * U'(W)
every ECN mark do
   W <- W - k
```

where $U'(W)$ is the derivative of the user's utility function that represents the utility the user receives as a function of its rate. The corresponding mapping for the PFCC is approximately

$$\Xi_{PFCC}(w, m) = \arg\min_{l \in \mathcal{W}_{PFCC}} \left(|kwU'(w) - km - l|\right), \tag{3}$$

for any $w \in \mathcal{W}_{PFCC}$ and $m = 0, 1, \ldots, \lfloor w \rfloor$. For instance, w here represent the window size (in bytes) normalized by the packet size (in bytes).

Each connection transmits the data packets into the network at the beginning of timeslots. The number of packets connection i transmits at the beginning of timeslot $[t, t+1)$, denoted by $A_i^{(N)}(t)$, is a function of the state variables $\mathbf{Y}_i^{(N)}(t)$ and is determined by a mapping $\Lambda : \mathcal{Y} \to \{0, 1, \ldots, W_{\max}\}$, *i.e.,*

$$A_i^{(N)}(t) = \Lambda\left(\mathbf{Y}_i^{(N)}(t)\right). \tag{4}$$

In the case of TCP, for example, a connection transmits as many packets as allowed by its congestion window provided that it has enough data to transmit, *i.e.,*

$$A_i^{(N)}(t) = \lfloor W_i^{(N)}(t) \rfloor. \tag{5}$$

1.2 Network Dynamics

In this subsection we explain how packets are marked to provide the congestion notification to the connections. Let $Q^{(N)}(t)$ denote the number of packets queued in the buffer at the beginning of timeslot $[t, t+1)$. Each 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

$$A^{(N)}(t) := \sum_{i=1}^{N} A_i^{(N)}(t) \tag{6}$$

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

$$Q^{(N)}(t+1) = \left[Q^{(N)}(t) - NC + A^{(N)}(t)\right]^+. \tag{7}$$

Each incoming packet into the bottleneck gateway is marked according to a marking mechanism depending on the AQM mechanism implemented at the gateway. This mechanism will be specified in the next section. We represent the possibility of a packet being marked by the $\{0, 1\}$-valued rvs $M_{i,j}^{(N)}(t+1)$ $(j = 1, ..., A_i^{(N)}(t))$ with the interpretation that $M_{i,j}^{(N)}(t+1) = 1$ (resp. $M_{i,j}^{(N)}(t+1) = 0$) if the jth packet from source i is marked (resp. not marked) in the AQM

there exists N flows in the system and does not imply that the capacity of the bottleneck router changes with the number of flows.

[2] This ignores the session-level dynamics. See [21] for model of TCP with slow-start/congestion avoidance

buffer. The number of marks connection i receives in the timeslot can now be written as

$$M_i^{(N)}(t+1) = \begin{cases} \sum_{j=1}^{A_i^{(N)}(t)} M_{i,j}^{(N)}(t+1), & A_i^{(N)}(t) > 0 \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

This information will be available to the sender in the next timeslot.

1.3 AQM Mechanism

AQM gateways control their congestion level by randomly marking incoming packets to signal the traffic sources of the congestion level. In order to do so, an AQM mechanism calculates a marking probability in each timeslot depending on the current and past values of the queue and average queue sizes and the arrival rates. The average queue size in an AQM mechanism is assumed to be given by

$$\hat{Q}^{(N)}(t+1) = (1-\alpha)\hat{Q}^{(N)}(t) + \alpha Q^{(N)}(t+1), \quad (9)$$

where $0 < \alpha \leq 1$ is the parameter of the exponential weighted moving average (EWMA) mechanism.

The computation of the marking probability might involve some internal state variables which are updated recursively (see for example Random Early Marking mechanism (REM) [1]). To facilitate this, we introduce an $\mathbf{R}^{n_B}$-valued rv $B^{(N)}(t)$ for some positive integer n_B to represent this internal state variable. Its recursion is given by a $\mathcal{R} \to \mathbf{R}^{n_B}$ mapping $\Psi^{(N)}$, *i.e.,*

$$B^{(N)}(t+1) = \Psi^{(N)}(R^{(N)}(t)), \quad (10)$$

where $R^{(N)}(t)$ denote a vector of variables used for computing the marking probability which is assumed to be

$$R^{(N)}(t) = \left[B^{(N)}(t), Q^{(N)}(s), \hat{Q}^{(N)}(s), A^{(N)}(s),\ t-\tau_w \leq s \leq t\right],$$

for some constant positive integer τ_w. For each t, $R^{(N)}(t)$ takes a value in the state space $\mathcal{R} = \mathbf{R}^{n_B} \times \mathbf{R}_+^{\tau_w} \times \mathbf{R}_+^{\tau_w} \times \mathbb{Z}_+^{\tau_w}$. Here we implictly assume that the marking function depends on the past queue and average queue sizes and packet arrival rates over a finite horizon determined by τ_w. For instance, if $\tau_w = 0$, then the marking function would depend only on the current values.

The probability that the AQM marks an incoming packet in a timeslot is assumed to depend on $R^{(N)}(t)$ at the beginning of the timeslot. We represent this dependence through a mapping $f^{(N)} : \mathcal{R} \to [0,1]$.

In order to define the events of packet marking by the AQM mechanism 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 $i = 1,\ldots,N$ and $j = 1,2,\ldots$, we define the marking rvs

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

so that the rv $M_{i,j}^{(N)}(t+1)$ is the indicator function of the event that the jth packet from source i is marked in timeslot $[t, t+1)$.

In the case of RED the marking probability is computed based on the current average queue size, *i.e.,*

$$M_{i,j}^{(N)}(t+1) = \mathbf{1}\left[V_{i,j}(t+1) < f_{RED}^{(N)}(\hat{Q}^{(N)}(t))\right], \quad (11)$$

for some mapping $f_{RED}^{(N)} : \mathbf{R}_+ \to [0,1]$.

For REM, the marking probability depends on the aggregate arrival rate of flows and can be calculated from the following equations:

$$\begin{aligned} B^{(N)}(t+1) &= \left[B^{(N)}(t) + \kappa(A^{(N)}(t) - NC)\right]^+ \quad (12) \\ M_{i,j}^{(N)}(t+1) &= \mathbf{1}\left[V_{i,j}(t+1) < f_{REM}^{(N)}(B^{(N)}(t))\right], \end{aligned}$$

where $\kappa > 0$ is a constant parameter and $f_{REM}^{(N)}(x) = 1 - e^{-x/N}$.

2. THE AVERAGE QUEUE DYNAMICS

The first main result of the paper consists of the asymptotics for the normalized buffer content as the number of flows becomes large. This result is discussed under the following Assumptions (A1)-(A2):

(A1) There exist a continuous functions $f : \mathcal{R} \to [0,1]$ and $\Psi : \mathcal{R} \to \mathbf{R}^{n_B}$ such that for each $N = 1,2,\ldots$ and $x \in \mathcal{R}$:

$$f^{(N)}(x) = f(N^{-1}x) \text{ and } \Psi^{(N)}(x) = \Psi(N^{-1}x)$$

(A2) For each $N = 1,2,\ldots$, the initial conditions of rvs in (1), (7), (9) and (10) are given by

$$\begin{aligned} &Q^{(N)}(s) = \hat{Q}^{(N)}(s) = 0, \quad s \leq 0, \\ &\mathbf{Y}_i^{(N)}(0) = \mathbf{y}, \quad i = 1,\ldots,N \quad \text{and} \quad B^{(N)}(0) = \Psi^{(N)}(\mathbf{0}), \end{aligned}$$

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

Assumption (A1) is a structural condition. Since we are interested in the the dynamics when there exists N flows in the system, then f is just a surrogate function representing the average contribution that each flow has on the marking probability. More specifically, we scale other system parameters, *e.g.,* bottleneck link capacity, with N the congestion level at the bottleneck gateway, *e.g.,* queueing delay, should be measured as a function of $R^{(N)}(t)$ normalized by the number of flows N. For instance, in the case of REM, the congestion level is measured by the queueing delay at the gateway. Hence, if the link capacity scales with N, then the correct measure of the congestion level is the queue size per flow in our model.

Assumption (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.

Our results are summarized by the following theorem:

THEOREM 1. *Assume (A1)-(A2) to hold. Then, for each $t = 0,1,\ldots$, there exist a (non-random) constant $q(t), \hat{q}(t)$, $\mathcal{R}$-valued vector $r(t)$, $\mathbf{R}^{n_B}$-constant $b(t)$, and a $\mathcal{Y}$-valued rv $\mathbf{Y}(t)$ such that the following holds:*

(i) The convergence results

$$\begin{aligned} &\frac{Q^{(N)}(t)}{N} \xrightarrow{P}_N q(t), \quad \frac{\hat{Q}^{(N)}(t)}{N} \xrightarrow{P}_N \hat{q}(t), \\ &\frac{R^{(N)}(t)}{N} \xrightarrow{P}_N r(t), \quad \frac{B^{(N)}(t)}{N} \xrightarrow{P}_N b(t), \end{aligned}$$

and

$$\mathbf{Y}_1^{(N)}(t) \Longrightarrow_N \mathbf{Y}(t) \quad (13)$$

take place;

(ii) For any bounded function $g : \mathcal{Y} \to \mathbf{R}$,

$$\tfrac{1}{N}\sum_{i=1}^{N} g(\mathbf{Y}_i^{(N)}(t)) \xrightarrow{P}_N \mathbf{E}\left[g(\mathbf{Y}(t))\right]. \quad (14)$$

(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

$$\lim_{N\to\infty} \mathbf{P}\left[\mathbf{Y}_i^{(N)}(t) = \mathbf{y}_i,\ i = 1, \ldots, I\right] = \prod_{i=1}^{I} \mathbf{P}\left[\mathbf{Y}(t) = \mathbf{y}_i\right] \quad (15)$$

for any $\mathbf{y}_1, \ldots, \mathbf{y}_I$ in $\mathcal{Y}$

Moreover, with initial conditions $q(s) = \hat{q}(s) = 0,\ s \le 0$, $b(0) = \Psi(\mathbf{0})$, and $\mathbf{Y}(0) = \mathbf{Y}_1^{(N)}(0)$, it holds that, for $t \ge 0$,

$$\begin{aligned} q(t+1) &= (q(t) - C + \mathbf{E}\left[A(t)\right])^+, \quad (16)\\ \hat{q}(t+1) &= (1-\alpha)\hat{q}(t) + \alpha q(t+1),\\ r(t+1) &= \left[b(t), q(s), \hat{q}(s), \mathbf{E}\left[A(s)\right],\ t+1-\tau_w \le s \le t+1\right],\\ b(t+1) &= \Psi(r(t)), \end{aligned}$$

where $A(t) = \Lambda\left(\mathbf{Y}(t)\right)$ and

$$\mathbf{Y}(t+1) =_{st} \Xi\left(\mathbf{Y}(t), M(t+1)\right). \quad (17)$$

The limiting rv $M(t+1)$ is given by

$$M(t+1) = \begin{cases} \sum_{j=1}^{A(t)} M_j(t+1), & A(t) > 0\\ 0, & \text{otherwise.} \end{cases} \quad (18)$$

and

$$M_j(t+1) = \mathbf{1}\left[V_j(t+1) \le f(r(t))\right], \quad j = 1, \ldots, A(t), \quad (19)$$

for i.i.d. $[0,1]$-uniform rvs $\{V_j(t+1),\ t = 0, 1, \ldots,$ and $j = 1, 2, \ldots\}$.

The proof of Theorem 1 is omitted here for the brevity of the presentation. It generalizes the proof of Theorem 1 in [22]. Here we only discuss the implication of the results in Theorem 1.

The recursion of the asymptotic queue $q(t)$ depends only on the capacity of the queue and the *expected* amount of traffic injected into the network in each timeslot, *i.e.,* $\mathbf{E}\left[A(t)\right]$. The latter can be determined from the limiting rv $\mathbf{Y}(t)$, whose distributional recursion in (17) closely resembles the recursion formula of a single user in (1). Therefore, one can see that the average amount of traffic injected into the network in timeslot $[t+1, t+2)$ from (17) and (1) is the same if the limiting rv $\mathbf{Y}(t)$ and $\mathbf{Y}_i^{(N)}(t)$ are in the same state and the marking probability in timeslot $[t, t+1)$ is the same. This observation justifies the use of single flow model, where the aggregate behavior of a large set of flows is modeled using a deterministic model of a single flow (see [8] and subsequent work, for example).

Much of the research effort along this control-theoretic direction has been focused on studying the local or global stability of the controlled queue. This line of research effort has led to a set of sufficient conditions for the queue to be either locally or globally (asymptotically) stable, *i.e.,* the queue eventually settles to an equilibrium value. However, very little emphasis has been placed on understanding the queue behavior during a transient period. Our results suggest that, under a large number of flows, the same control-theoretic approach used in [8] provides reasonable prediction to the oscillatory behavior of the queue size during a transient period and also provides insights into how one may be able to control such queue fluctuations. For example, one can show that the slope of the feedback probability function in RED, which can be viewed as a feedback gain, can have a significant effect on both transient and steady-state behavior of the asymptotic queue $q(t)$.

3. THE RANDOM QUEUE FLUCTUATIONS

In this section, we complement the results in Theorem 1 with a Central Limit Theorem (CLT) result. While the limiting recursion in Theorem 1 captures the *expected* behavior of the systems, the Central Limit analysis captures the errors/uncertainty due to the randomness in the system. This uncertainty appears as a fluctuation in the queue and is absent from the deterministic models. The analysis is carried out under the same model. However, we need a strengthened Assumption (A1):

(A1bis) Assumption (A1) holds with mapping $f : \mathcal{R} \to [0,1]$ and $\Psi : \mathcal{R} \to \mathbf{R}^{n_B}$ that are (totally) differentiable.[3]

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

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

We can interpret $K(t)$ as the asymptotic residual capacity per user in the timeslot $[t, t+1)$.

Now define a collection of rvs that are integral to our analysis. For each $N = 1, 2, 3, \ldots$ and $\mathbf{y} \in \mathcal{Y}$, set

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

and

$$\begin{aligned} L_{\mathbf{y}}^{(N)}(t) &:= \tfrac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}\left[\mathbf{y}\right] - \mathbf{P}_{\mathbf{Y}(t)}\left[\mathbf{y}\right]\\ \mathbf{L}^{(N)}(t) &:= \Big[L_B^{(N)}(t), L_0^{(N)}(s), \hat{L}_0^{(N)}(s), L_{\mathbf{y}}^{(N)}(s);\\ &\qquad \mathbf{y} \in \mathcal{Y}, s = 0, \ldots, t\Big]^T \end{aligned}$$

THEOREM 2. *Assume (A1bis)-(A2) hold. Then, for each $t = 0, 1, \ldots$, there exists an $\mathbf{R}^{(|\mathcal{Y}|+2)\cdot(t+1)+n_B}$-valued rv*

$$\mathbf{L}(t) = \Big[L_B(t), L_0(s), \hat{L}_0(s), L_{\mathbf{y}}(s);\ \mathbf{y} \in \mathcal{Y}, s = 0, 1, \ldots, t\Big]^T$$

such that the convergence

$$\sqrt{N}\mathbf{L}^{(N)}(t) \Longrightarrow_N \mathbf{L}(t) \quad (22)$$

holds. Moreover, the distributional recurrence

$$L_0(t+1) =_{st} \begin{cases} 0 & K(t) > 0\\ L_0(t) + \bar{L}(t) & K(t) < 0\\ \left(L_0(t) + \bar{L}(t)\right)^+ & K(t) = 0 \end{cases} \quad (23)$$

$$\hat{L}_0(t+1) =_{st} (1-\alpha)\hat{L}_0(t) + \alpha L_0(t+1) \quad (24)$$

and

$$L_B(t+1) =_{st} \frac{\partial \Psi(r(t))}{\partial R^T} L_R(t) \quad (25)$$

[3] A sufficient condition for f to be totally differentiable at $r \in \mathcal{R}$ is that all partial derivatives $\partial f/\partial R_i$ exist and are continuous in a neighborhood of r. Also note that any real-valued function on $\mathcal{R}$ can be approximated with arbitrary small error by a totally differentiable function.

hold, where

$$\bar{L}(t) = \sum_{\mathbf{y}\in\mathcal{Y}} \Lambda(\mathbf{y}) L_{\mathbf{y}}(t) \ , \tag{26}$$

and

$$L_R(t) = [L_B(t), L_0(s), \hat{L}_0(s), \bar{L}_0(s),\ \tau_w - t \le s \le t]^T. \tag{27}$$

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 $t = 1, 2, \ldots$, if $K(s) \neq 0$ for all $s < t$, then the rv $L_0(t+1)$ is Gaussian.

A proof of Theorem 2 is provided in the appendix.

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 technical artifact of the limiting regime as in practice it is unlikely that the real-valued residual capacity would attain the value zero *exactly*. Therefore, in practice with a large number of flows the queue size distribution at any fixed time t can be well approximated by a Gaussian rv with a mean $N \cdot q(t)$.

4. DISCUSSION

Our results in Sections 2 and 3 tell us that when the number of flows N is large, the queue size can be approximated by $Q^{(N)}(t) \simeq N \cdot q(t) + \sqrt{N} L_0(t)$ in some distributional sense. Both the deterministic process $q(t)$ and the distribution of the rv $L_0(t)$ can be computed recursively. In addition, the recursional formulas do not depend on the number of flows N.

The CLT analysis reveals the sources of *random* fluctuations in the queue size which cannot be captured by the limiting model of the system described in Section 2. It is shown in the proof of Theorem 2 that the queue fluctuation $L_0(t+1)$ consists of three components :

(i) Fluctuation caused by the discrepancy between the limiting feedback information $f(r(t))$ and the feedback information $f^{(N)}(R^{(N)}(t))$ from the AQM to the end user congestion-control mechanism: This uncertainty in feedback information can be explained by the following lemma (also known as the *Delta Method*) [23, Thm. 3.1, p. 26].

LEMMA 1. *If $f : \mathcal{R} \to \mathbb{R}^m$ is (totally) differentiable in the neighborhood of $r(t)$, then the convergence $\sqrt{N}\left(\frac{R^{(N)}(t)}{N} - r(t)\right) \Longrightarrow_N L_R(t)$ implies*

$$\sqrt{N}\left(f\left(\frac{R^{(N)}(t)}{N}\right) - f(r(t))\right) \Longrightarrow_N \frac{\partial f(r(t))}{\partial R^T} R(t) \ . \tag{28}$$

The vector $\frac{\partial f(r(t))}{\partial R^T}$ represents the sensitivity of the probability function f to the fluctuations around the limiting parameter $r(t)$. For example, this term corresponds to the slope of the marking probability function in the case of RED. 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 [5].

In case of REM (in the regime where the mapping (12) is linear), it is easy to see that this random fluctuations component is proportional to $\kappa \bar{L}(t)$. In other words, the random fluctuations in the queue originate from the fluctuations in the arrival rates amplified by the gain κ.

(ii) The granularity of feedback information: The congestion control mechanism at the end users estimates the marking probability $f^{(N)}(R^{(N)}(t))$ from the number of marks received during an RTT. However, the number of marks is limited by the number of packets transmitted. This nature of feedback information poses limited feedback information granularity, and causes a fluctuation in the queue size. This fluctuation can be well-approximated by a Gaussian rv and cannot be captured without taking into account the detailed packet level operations of the congestion-control mechanism.

(iii) Fluctuation caused by the structure of protocols: The protocols adopted by the end users and gateway determine the mappings used for update rules and marking mechanism, and hence the evolution of the state variables of the connections and queue dynamics. It turns out the magnitude of the random queue oscillation depends on the detailed interaction of the end user protocol and AQM mechanism. This observation indicates that although the stability of the deterministic process may not depend on the details of end user algorithm as suggested in [17], the magnitude of the random queue oscillation does. Hence, in order to accurately model the queue dynamics and use it for, for instance, network provisioning, modeling the detailed dynamics of the interaction may be required.

Components (ii) and (iii) are due to the protocols and cannot be mitigated without protocol modifications. Thus, network designers can only manipulate the sensitivity of the feedback function, *e.g.,* its slope, to reduce oscillation of queue size. Although reducing the sensitivity of the feedback function can decrease the magnitude of fluctuation, it also typically results in an increase of the average queue size and/or a slower response time.

It is notable from the Central Limit analysis that a random marking mechanism in AQM always introduces random fluctuations in the queue. This is an intrinsic behavior of a random marking mechanism due to the limited granularity in the feedback information. Such a fluctuation can be reduced by increasing the quality of the feedback information either through an increase in the number of ECN bits or through an in-band signaling mechanism as suggested in [18]. Also, the granularity of the feedback information improves as the size of the window becomes large. There are, however, other means to alleviate such fluctuations. For example, TCP Vegas [3] and its variants such as TCP FAST [10] use delay information instead of marks to adjust their congestion window sizes. Given accurate timestamps in the packets, each flow can calculate the appropriate adjustment to its congestion window size, thus mitigating the uncertainty from the randomness in the marks.

It is worth noting that some of the system parameters affect both deterministic and random fluctuations. For instance, the slope of the marking probability function not only affects the magnitude of random queue size fluctuation, but also influences the settling time of the deterministic queue process during a transient period.

5. EXAMPLE OF TCP/RED

In this section, we provide an example of Random Early Detection [7] as AQM mechanism and TCP Reno [9] as the congestion-control mechanism. This example is the most widely studied case [8, 17]. A system comprised of persistent TCP flows with a homogeneous RTT can be described using (2) and (11). This model is similar to the model in [22]. Both the Monte-Carlo simulations and the NS-2 simulations in [22] suggest that the limiting behavior of the queue follows the results in Theorem 1. Furthermore, the rate of decrease of the standard deviation of the normalized queue size appears to be consistent with the prediction of the Central Limit Theorem, *i.e.,* the standard deviation at the steady-state decreases according to $\frac{1}{\sqrt{N}}$.

In the case of TCP-RED with session-layer dynamics and variable RTTs, similar results are derived in [20, 21]. It shows that there are additional sources of queue fluctuations in the system. These are the fluctuations caused by the arrivals of new TCP connections and the random idle periods: 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. Furthermore, there are also fluctuations introduced by the heterogeneous nature of the round-trip delays of the connections. The magnitude of these fluctuations varies with the variance of the round-trip delays among the connections.

One thing to note in the case of TCP-RED is that the granularity of the information used for window size update is very coarse ; connections only check whether packets are marked or not using ECN ECHO option at the TCP receivers [6].[4] Therefore, TCP-RED does not take full advantage of the improvement in feedback information granularity when the window size increases. One simple scheme to improve the feedback information granularity in TCP-RED is to increase the number of feedback information bits in the ECN mechanism. If the improved feedback information is properly utilized, the magnitude of queue fluctuation can be reduced. Multi-level ECN (MECN) [4] is an example of such a scheme.

6. CONCLUSIONS

In this paper, we have developed a stochastic model with detailed packet level operations of a probabilistic Active Queue Management scheme and generic end-user congestion-control mechanism. We classify the queue fluctuations in such system with a large number of flows into two distinct components, *i.e.,* the deterministic queue fluctuations and the random queue fluctuations. The deterministic queue fluctuation can be effectively modeled in a deterministic system and is predictable by control-theoretic analysis. On the other hand, the random queue fluctuations, which are well approximated by Gaussian processes, originate from the random marking mechanism in probabilistic AQM.

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.

[4]Recall that throughout the paper we assume that connections are ECN capable.

7. REFERENCES

[1] Sanjeewa Athuraliya, David Lapsley, and Steven Low. An enhanced random early marking algorithm for Internet flow control. In *Proceedings of IEEE INFOCOM*, 2000.

[2] Francois Baccelli, David R. McDonald, and Julien Reynier. A mean-field model for multiple TCP connections through a buffer implementing RED. Technical report, INRIA, April 2002.

[3] Lawrence S. Brakmo and Larry L. Peterson. TCP Vegas: end to end congestion avoidance on a global internet. *IEEE Journal on Selected Areas in Communications*, 13(8):1465–1480, October 1995.

[4] Arjan Durresi, Mukundan Sridharan, Chunlei Liu, Mukul Goyal, and Raj Jain. Multilevel early congestion notification. In *Proceedings of the 5th World Multiconference on Systemics, Cybernetics and Informatics*, pages 12–17, Orlando, FL, July 2001.

[5] Victor Firoiu and Marty Borden. A study of active queue management for congestion control. In *Proceedings of IEEE INFOCOM*, 2000.

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

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

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

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

[10] Cheng Jin, David X. Wei, and Steven H. Low. FAST TCP: motivation, architecture, algorithms, performance. Submitted for publication.

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

[12] Frank Kelly, A.K. Maulloo, and D.K.H. Tan. Rate control in communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49:237–252, 1998.

[13] Frank P. Kelly. Charging and rate control for elastic traffic. *European Transactions on Telecommunications*, 8:33–37, 1997.

[14] Srisankar Kunniyur and R. Srikant. End-to-end congestion control schemes: Utility functions, random losses and ECN marks. In *Proceedings of IEEE INFOCOM*, 2000.

[15] Srisankar Kunniyur and R. Srikant. A time scale decomposition approach to adaptive ECN marking. In *Proceedings of IEEE INFOCOM*, 2001.

[16] Steven H. Low. A duality model of TCP and queue management algorithms. In *Proceedings of ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management*, September 2000.

[17] Steven H. Low, Fernando Paganini, J. Wang, S. Adlakha, and John C. Doyle. Dynamics of TCP/RED and a scalable control. In *Proceedings of IEEE INFOCOM*, March 2002.

[18] Mayank Sharma, Dina Katabi, Balaji Prabhakar, and Rong Pan. A general multiplexed ECN channel and its

use for wireless loss notification. Submitted for Publication, February 2003.

[19] Peerapol Tinnakornsrisuphap. *Dynamics of Random Early Detection Gateway under a Large Number of TCP Flows*. PhD thesis, University of Maryland, May 2004.

[20] Peerapol Tinnakornsrisuphap and Richard J. La. Asymptotic behavior of heterogeneous TCP flows and RED gateways. Technical report, Institute for Systems Research, University of Maryland, 2003.

[21] Peerapol Tinnakornsrisuphap and Richard J. La. Limiting model of ECN/RED under a large number of heterogeneous TCP flows. In *Proceedings of IEEE Conference on Decision and Control*, Maui, Hawaii, December 2003.

[22] 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.

[23] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

APPENDIX

A. PROOF OF THEOREM 2

The proof of Theorem 2 is carried out by the induction on t of the following statement.

[E:t] The convergence (22) holds for some $\mathbf{R}^{(|\mathcal{Y}|+2)(t+1)+n_B}$-valued rv

$$\mathbf{L}(t) = (L_B(t), L_0(s), \hat{L}_0(s), L_{\mathbf{y}}(s),\ \mathbf{y} \in \mathcal{Y}, s = 0, 1, \ldots, t)$$

The main induction argument from **[E:t]** to **[E:t+1]** is established by the following proposition:

PROPOSITION 1. *Under (A1bis) and (A2), if* **[E:t]** *holds for some $t = 0, 1, \ldots$, then* **[E:t+1]** *also holds.*

The distributional recurrence in (23) follows from the following result:

PROPOSITION 2. *Under (A1bis) and (A2), if* **[E:t]** *holds for some $t = 0, 1, \ldots$, then*

$$\begin{aligned}\sqrt{N}L_0^{(N)}(t+1) &= \sqrt{N}\left(\frac{Q^{(N)}(t+1)}{N} - q(t+1)\right)\\ &\Longrightarrow_N \quad L_0(t+1) \end{aligned} \tag{29}$$

for some $\mathbf{R}$-valued rv $L_0(t+1)$ that satisfies the distributional relation in (23).

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

A.1 A Proof of Theorem 2

For $t = 0$ and $\mathbf{y} \in \mathcal{Y}$, $L_B^{(N)}(t) = L_0^{(N)}(t) = \bar{L}^{(N)}(t) = L_{\mathbf{y}}^{(N)}(t) = 0$. Hence, the statement **[E:0]** holds trivially. Eq. (22) and (23) then hold for all t from the induction on t by Propositions 1 and 2.

The distributional recursion in (24) and the distributional recursion of $L_B(t), L_{\mathbf{y}}(t),\ \mathbf{y} \in \mathcal{Y}$ are established as byproducts in the proof of Proposition 1.

It will be evident in the proof of Proposition 1 that the rvs involved in the distributional recursions up to time t are either Gaussian or constant except when there exists $K(s) = 0$ for some $s < t$. In that case, $L_0(s+1)$ will be truncated Gaussian from (23). The last statement in the Theorem 2 follows from this observation.

B. A PROOF OF PROPOSITION 2

The proof of Proposition 2 relies on two important steps. First, we rewrite $L_0^{(N)}(t+1)$ in terms of continuous maps of $\mathbf{L}^{(N)}(t)$, where the maps depend on the residual capacity per user $K(t)$. Equation (23) then follows from the continuous mapping theorem and **[E:t]**.

Fix $t = 0, 1, \ldots$ and $N = 1, 2, \ldots$. We rewrite the limiting recursion in (16) in the following form:

$$q(t+1) = (q(t) - C + \mathbf{E}\left[A(t)\right])^+ = (-K(t))^+ \tag{30}$$

with $K(t)$ given by (20). Combining this observation with the queue dynamics (30), let

$$\begin{aligned}\bar{L}^{(N)}(t) &= \tfrac{1}{N}\textstyle\sum_{i=1}^N A_i^{(N)}(t) - \mathbf{E}\left[A(t)\right]\\ &= \tfrac{1}{N}\textstyle\sum_{i=1}^N \Lambda(\mathbf{Y}_i^{(N)}(t)) - \mathbf{E}\left[\Lambda(\mathbf{Y}(t))\right]\\ &= \tfrac{1}{N}\textstyle\sum_{i=1}^N \sum_{\mathbf{y}\in\mathcal{Y}} \Lambda(\mathbf{y})\left(\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}] - \mathbf{P}_{\mathbf{Y}(t)}[\mathbf{y}]\right)\\ &= \sum_{\mathbf{y}\in\mathcal{Y}} \Lambda(\mathbf{y}) L_{\mathbf{y}}^{(N)}(t)\ . \end{aligned} \tag{31}$$

Then, we have

$$\begin{aligned}L_0^{(N)}(t+1) &= \left(\frac{Q^{(N)}(t)}{N} - C + \frac{1}{N}\sum_{i=1}^N A_i^{(N)}(t)\right)^+ - (-K(t))^+\\ &= \max\left(L_0^{(N)}(t) + \frac{1}{N}\sum_{i=1}^N A_i^{(N)}(t) - \mathbf{E}\left[A(t)\right], K(t)\right) - K(t)^+\\ &= \max\left(L_0^{(N)}(t) + \bar{L}^{(N)}(t), K(t)\right) - K(t)^+ \end{aligned} \tag{32}$$

and

$$\begin{aligned}&\sqrt{N}L_0^{(N)}(t+1)\\ &= \max\left(\sqrt{N}\left(L_0^{(N)}(t) + \bar{L}^{(N)}(t)\right), \sqrt{N}K(t)\right) - \sqrt{N}K(t)^+.\end{aligned} \tag{33}$$

Under **[E:t]** and decomposition (31), we can invoke the Continuous Mapping Theorem to conclude that

$$\sqrt{N}\left(L_0^{(N)}(t) + \bar{L}^{(N)}(t)\right) \Longrightarrow_N L_0(t) + \bar{L}(t). \tag{34}$$

Three cases emerge depending on the sign of $K(t)$. If $K(t) = 0$, then Equation (33) reduces to

$$\sqrt{N}L_0^{(N)}(t+1) = \left(\sqrt{N}\left(L_0^{(N)}(t) + \bar{L}^{(N)}(t)\right)\right)^+. \tag{35}$$

Again by the Continuous Mapping Theorem and (34), Equation (35) yields

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

If $K(t) < 0$, then Equation (33) 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 the convergence (34) gives us

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

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

Finally, if $K(t) > 0$, then Equation (33) 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 (34) yields $\sqrt{N}L_0^{(N)}(t+1) \Longrightarrow_N 0$ since $\lim_{N\to\infty}\sqrt{N}K(t) = \infty$. This completes the proof of Proposition 2.

C. SOME USEFUL FACTS AND RESULTS

In this appendix, we present some useful facts and results that will facilitate the presentation of the proof of Proposition 1. In Appendix C.1, we present simple facts on the conditional distributions of the number of marks received in timeslot $[t+1, t+2)$, given the complete history of events up to $[t, t+1)$. We then show in Appendix C.2 that for any $\mathbf{y} \in \mathcal{Y}$, $L_{\mathbf{y}}^{(N)}(t+1)$ can be written in terms of $L_{\mathbf{y}}^{(N)}(t)$ and rvs representing random fluctuations corresponding to components (i)-(ii) in the discussion in Section 4. Finally in Appendix C.3, we present the marginal convergences of the aforementioned fluctuations, and subsequently, show their joint convergence which is essential in establishing Proposition 1.

C.1 Simple Facts

Let $\mathcal{F}_t$ be a σ-field generated by

$$\{Q^{(N)}(0), \mathbf{Y}_i^{(N)}(0), V_{i,j}(s), V_j(s), \mathbf{Y}(0); \\ i,j = 1,2,\ldots,\ s = 0,\ldots,t\}.$$

Then $Q^{(N)}(t)$ and $\mathbf{Y}_i^{(N)}(t)$ are $\mathcal{F}_t$-measurable, *i.e.,* $Q^{(N)}(t)$ and $\mathbf{Y}_i^{(N)}(t)$ can be completely determined given the past history up until time t. As a result, for $m \le \Lambda(\mathbf{Y}_i^{(N)}(t))$, and $\Lambda(\mathbf{Y}_i^{(N)}(t)) \ge 1$, we have under Assumption (A1):

$$\begin{aligned}\mathbf{P}\left[M_i^{(N)}(t+1) = m|\mathcal{F}_t\right] &= \binom{\Lambda(\mathbf{Y}_i^{(N)}(t))}{m} f\left(\frac{R^{(N)}(t)}{N}\right)^m \\ &\quad\times\left(1 - f\left(\frac{R^{(N)}(t)}{N}\right)\right)^{\Lambda(\mathbf{Y}_i^{(N)}(t))-m} \\ &= \chi_{\mathbf{Y}_i^{(N)}(t),m}\left(\frac{R^{(N)}(t)}{N}\right), \qquad (36)\end{aligned}$$

where the mapping $\chi_{x,y}(z) : \mathcal{Y}\times\mathbb{Z}_+\times\mathbf{R}_+ \to [0,1]$ is given by

$$\chi_{x,y}(z) = \begin{cases} \binom{\Lambda(x)}{y} f(z)^y\left(1-f(z)\right)^{\Lambda(x)-y} & y \le \Lambda(x),\ \Lambda(x)\ge 1 \\ 0 & \text{otherwise}\end{cases}$$

Similarly, we also have for $\Lambda(\mathbf{Y}(t)) \ge 1$ and $m \le \Lambda(\mathbf{Y}(t))$

$$\begin{aligned}&\mathbf{P}\left[M(t+1) = m|\mathcal{F}_t\right] \\ &\quad= \binom{\Lambda(\mathbf{Y}(t))}{m} f(r(t))^m\left(1-f(r(t))\right)^{\Lambda(\mathbf{Y}(t))-m} \\ &\quad= \chi_{\mathbf{Y}(t),m}\left(r(t)\right). \qquad (37)\end{aligned}$$

C.2 A Key Decomposition

For any $\mathbf{y} \in \mathcal{Y}$, $N = 1,2,\ldots$, $i = 1,\ldots,N$ and $t = 0,1,\ldots$

$$\begin{aligned}\mathbf{1}_{\mathbf{Y}_i^{(N)}(t+1)}[\mathbf{y}] &= \sum_{\mathbf{y}_1\in\mathcal{Y}_{\mathbf{y},1}} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}_1] \qquad (38)\\ &\quad+\sum_{m=0}^{W_{\max}}\sum_{\mathbf{y}_2\in\mathcal{Y}_{\mathbf{y},2,m}} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}_2]\,\mathbf{1}_{M_i^{(N)}(t+1)}[m],\end{aligned}$$

where $\mathcal{Y}_{\mathbf{y},1} \subset \mathcal{Y}$ is the set of states from which a connection will always transition to the state $\mathbf{y}$ in the next timeslot, and $\mathcal{Y}_{\mathbf{y},2,m} \subset \mathcal{Y}$ is the set of states from which a connection will transition to the state $\mathbf{y}$ in the next timeslot only upon receiving exactly m marks in this timeslot.

The decomposition (38) leads to

$$\begin{aligned}L_{\mathbf{y}}^{(N)}(t+1) &= \tfrac{1}{N}\textstyle\sum_{i=1}^N \mathbf{1}_{\mathbf{Y}_i^{(N)}(t+1)}[\mathbf{y}] - \mathbf{P}_{\mathbf{Y}(t+1)}[\mathbf{y}] \\ &= \tfrac{1}{N}\textstyle\sum_{i=1}^N\sum_{\mathbf{y}_1\in\mathcal{Y}_{\mathbf{y},1}} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}_1] - \mathbf{P}_{\mathbf{Y}(t)}[\mathbf{y}_1] \\ &\quad+\tfrac{1}{N}\textstyle\sum_{i=1}^N\sum_{m=0}^{W_{\max}}\sum_{\mathbf{y}_2\in\mathcal{Y}_{\mathbf{y},2,m}} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}_2]\,\mathbf{1}_{M_i^{(N)}(t+1)}[m] \\ &\quad-\mathbf{E}\left[\mathbf{1}_{\mathbf{Y}(t)}[\mathbf{y}_2]\,\mathbf{1}_{M(t+1)}[m]\right]. \qquad (39)\end{aligned}$$

Note that

$$\mathbf{E}\left[\mathbf{1}_{\mathbf{Y}(t)}[\mathbf{y}_2]\,\mathbf{1}_{M(t+1)}[m]\right] = \mathbf{P}_{\mathbf{Y}(t)}[\mathbf{y}_2]\,\chi_{\mathbf{y}_2,m}(r(t)). \qquad (40)$$

Therefore, simple manipulations with (39) yields

$$\begin{aligned}&L_{\mathbf{y}}^{(N)}(t+1) \\ &= \sum_{\mathbf{y}_1\in\mathcal{Y}_{\mathbf{y},1}} L_{\mathbf{y}_1}^{(N)}(t) + \sum_{m=0}^{W_{\max}}\sum_{\mathbf{y}_2\in\mathcal{Y}_{\mathbf{y},2,m}} \Big\{\tfrac{1}{N}\textstyle\sum_{i=1}^N \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}_2] \\ &\quad\times\Big[\mathbf{1}_{M_i^{(N)}(t+1)}[m] - \chi_{\mathbf{y}_2,m}(\frac{R^{(N)}(t)}{N}) \\ &\quad+\chi_{\mathbf{y}_2,m}(\frac{R^{(N)}(t)}{N}) - \chi_{\mathbf{y}_2,m}(r(t))\Big] \\ &\quad+\tfrac{1}{N}\textstyle\sum_{i=1}^N\left(\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}_2] - \mathbf{P}_{\mathbf{Y}(t)}[\mathbf{y}_2]\right)\chi_{\mathbf{y}_2,m}(r(t))\Big\} \\ &= \sum_{\mathbf{y}_1\in\mathcal{Y}_{\mathbf{y},1}} L_{\mathbf{y}_1}^{(N)}(t) + \sum_{m=0}^{W_{\max}}\sum_{\mathbf{y}_2\in\mathcal{Y}_{\mathbf{y},2,m}} \upsilon_{\mathbf{y}_2,m}^{(N)}(t) \\ &\quad+\gamma_{\mathbf{y}_2,m}^{(N)}(t) + \chi_{\mathbf{y}_2,m}(r(t))L_{\mathbf{y}_2}^{(N)}(t), \qquad (41)\end{aligned}$$

where we define

$$\begin{aligned}&\upsilon_{\mathbf{y},m}^{(N)}(t) \qquad (42)\\ &= \tfrac{1}{N}\textstyle\sum_{i=1}^N \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}]\left(\chi_{\mathbf{y},m}\left(\frac{R^{(N)}(t)}{N}\right) - \chi_{\mathbf{y},m}(r(t))\right)\end{aligned}$$

and

$$\begin{aligned}&\gamma_{\mathbf{y},m}^{(N)}(t) \qquad (43)\\ &= \tfrac{1}{N}\textstyle\sum_{i=1}^N \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}]\left(\mathbf{1}_{M_i^{(N)}(t+1)}[m] - \chi_{\mathbf{y},m}\left(\frac{R^{(N)}(t)}{N}\right)\right).\end{aligned}$$

The collection of triangular arrays $\{\upsilon_{\mathbf{y},m}^{(N)}(t);\ \mathbf{y}\in\mathcal{Y}, m = 0,\ldots,W_{\max}\}$ represents component (i) in the discussion in Section 4, while $\{\gamma_{\mathbf{y},m}^{(N)}(t);\ \mathbf{y}\in\mathcal{Y}, m = 0,\ldots,W_{\max}\}$ represents component (ii). The sets $\mathcal{Y}_{\mathbf{y},1}$ and $\mathcal{Y}_{\mathbf{y},2,m}$, $m = 0,1,\ldots,W_{\max}$ are specified according the structure of the

protocols, *i.e.,* component (iii). In order to establish Proposition 1, the next subsection gives results on necessary convergences of the triangular arrays as N become large. These convergences can be established independently of the sets $\mathcal{Y}_{\mathbf{y},1}$ and $\mathcal{Y}_{\mathbf{y},2,m},\ m=0,1,\ldots,W_{\max}$.

C.3 Auxiliary Results

We first show the marginal convergence of the triangular arrays

$$\Upsilon^{(N)}(t) := \left[v^{(N)}_{\mathbf{y},m}(t);\ \mathbf{y}\in\mathcal{Y},\ m=0,1,\ldots,W_{\max}\right]^T$$

and

$$\Gamma^{(N)}(t) := \left[\gamma^{(N)}_{\mathbf{y},m}(t);\ \mathbf{y}\in\mathcal{Y},\ m=0,1,\ldots,W_{\max}\right]^T$$

from the following results.

PROPOSITION 3. *Under the conditions of Theorem 2, for any fixed t, the following convergence*

$$\sqrt{N}\mathbf{\Gamma}^{(N)}(t) \Longrightarrow_N \mathbf{\Gamma}(t) \tag{44}$$

holds, where $\mathbf{\Gamma}(t) = [\gamma_{\mathbf{y},m}(t);\ \mathbf{y}\in\mathcal{Y}, m=0,\ldots,W_{\max}]^T$ is a $|\mathcal{Y}|\cdot(W_{\max}+1)$-dimensional $(\mathbf{0},\mathbf{S}(t))$-Gaussian random vector.

The covariance matrix

$$\mathbf{S}(t) = [S_{(\mathbf{x},m),(\mathbf{y},n)}(t);\ \mathbf{x},\mathbf{y}\in\mathcal{Y},\ m,n=0,1,\ldots,W_{\max}]$$

is given by

$$\begin{aligned} & S_{(\mathbf{x},m),(\mathbf{y},n)}(t) \\ = \ & \mathbf{E}\left[\gamma_{\mathbf{x},m}(t)\gamma_{\mathbf{y},n}(t)\right] \\ = \ & \begin{cases} \mathbf{P}_{\mathbf{Y}(t)}[\mathbf{y}]\,\chi_{\mathbf{y},m}(r(t))(1-\chi_{\mathbf{y},m}(r(t))), & \mathbf{x}=\mathbf{y}, m=n \\ -\mathbf{P}_{\mathbf{Y}(t)}[\mathbf{y}]\,\chi_{\mathbf{x},m}(r(t))\chi_{\mathbf{y},n}(r(t)), & \mathbf{x}=\mathbf{y}, m\neq n \\ 0, & \mathbf{x}\neq\mathbf{y}. \end{cases} \end{aligned} \tag{45}$$

Moreover, $\Gamma(t)$ is independent of $\mathcal{F}_t$.

We now show the convergence of $\sqrt{N}\Upsilon^{(N)}(t)$ in distribution.

LEMMA 2. *Assume (A1bis), (A2), and* **[E:t]**, *then*

$$\sqrt{N}\Upsilon^{(N)}(t) \Longrightarrow_N \mathbf{J}(t)L_{f_R}(t), \tag{46}$$

where

$$\mathbf{J}(t) = \left[\begin{pmatrix}\Lambda(\mathbf{y})\\ m\end{pmatrix}\mathbf{P}_{\mathbf{Y}(t)}[\mathbf{y}]\,J_{\mathbf{y},m}(t);\ \mathbf{y}\in\mathcal{Y}, m=0,\ldots,W_{\max}\right]^T.$$

The constants $J_{\mathbf{y},m}(t)$ are given by

$$\begin{aligned} & J_{\mathbf{y},m}(t) \\ = \ & \begin{cases} mf(r(t))^{m-1}(1-f(r(t)))^{\Lambda(\mathbf{y})-m} & \\ -(\Lambda(\mathbf{y})-m)f(r(t))^m & \\ \cdot(1-f(r(t)))^{\Lambda(\mathbf{y})-m-1}, & m=1,\ldots,\Lambda(\mathbf{y})-1, \\ & \Lambda(\mathbf{y})\geq 1 \\ -\Lambda(\mathbf{y})(1-f(r(t)))^{\Lambda(\mathbf{y})-1}, & m=0, \Lambda(\mathbf{y})\geq 1 \\ \Lambda(\mathbf{y})f(r(t))^{\Lambda(\mathbf{y})-1}, & m=\Lambda(\mathbf{y})\geq 1 \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

and $L_{f_R}(t)$ is the limiting rv from the following convergence

$$\sqrt{N}\left(f\left(\frac{R^{(N)}(t)}{N}\right)-f(r(t))\right) \Longrightarrow_N L_{f_R}(t) := \frac{\partial f(r(t))}{\partial R^T}L_R(t),$$

where $\frac{\partial f(r(t))}{\partial R^T}$ represents the sensitivity of the marking function to the fluctuations in the parameters around the point $r(t)$.

The following lemma establishes the joint convergence that is essential for the proof of Proposition 1.

LEMMA 3. *Under Assumption (A1bis), (A2), and* **[E:t]**, *we have the following convergence:*

$$\begin{aligned} & \sqrt{N}\left(L^{(N)}_0(t+1), \mathbf{L}^{(N)}(t), \Upsilon^{(N)}(t), \Gamma^{(N)}(t)\right) \\ & \Longrightarrow_N \left(L_0(t+1), \mathbf{L}(t), \mathbf{J}(t)L_{f_R}(t), \Gamma(t)\right), \end{aligned} \tag{47}$$

where $\Gamma(t)$ and $\mathbf{J}(t)L_{f_R}(t)$ are given in Proposition 3 and Lemma 2, respectively.

The proofs of Propositions 3, Lemma 2 and 3 are given in Section E.

D. A PROOF OF PROPOSITION 1

The proof of Proposition 1 is established with the help of the decomposition in Appendix C.2 and the convergence results in Appendix C.3.

By the decomposition in (41), for any given $\mathbf{y}\in\mathcal{Y}$ we have

$$\begin{aligned} L^{(N)}_{\mathbf{y}}(t+1) = & \sum_{\mathbf{y}_1\in\mathcal{Y}_{\mathbf{y},1}} L^{(N)}_{\mathbf{y}_1}(t) \\ & + \sum_{m=0}^{W_{\max}}\sum_{\mathbf{y}_2\in\mathcal{Y}_{\mathbf{y},2,m}} v^{(N)}_{\mathbf{y}_2,m}(t)+\gamma^{(N)}_{\mathbf{y}_2,m}(t)+\chi_{\mathbf{y}_2,m}(r(t))L^{(N)}_{\mathbf{y}_2}(t). \end{aligned}$$

Thus, the joint convergence

$$\begin{aligned} & \sqrt{N}\left(L^{(N)}_0(t+1), L^{(N)}_{\mathbf{y}}(t+1), \mathbf{L}^{(N)}(t),\ \mathbf{y}\in\mathcal{Y}\right) \\ & \Longrightarrow_N \left(L_0(t+1), L_{\mathbf{y}}(t+1), \mathbf{L}(t),\ \mathbf{y}\in\mathcal{Y}\right) \end{aligned} \tag{48}$$

is established through Lemma 3, $[\mathbf{E}:\mathbf{t}]$, and the continuous mapping theorem.

To establish $[\mathbf{E}:\mathbf{t}+\mathbf{1}]$, observe that

$$\sqrt{N}L^{(N)}_B(t+1) \Longrightarrow_N \frac{\partial\Psi(r(t))}{\partial R^T}\cdot L_R(t), \tag{49}$$

from the Delta Method (Lemma 1), where $L_R(t)$ is given in (27) and its elements can be written as a linear combination of the elements in $\mathbf{L}(t)$. Therefore, $L_R(t)$ jointly converges with the rvs in (48). A closer inspection at the proof of the delta method (see Section 7.4 in [19] for example) reveals that the limiting rv in (49) is also jointly convergence in law with the limiting rvs in (48).

Meanwhile, $\hat{L}^{(N)}_0(t+1)$ is just a convex combination of $\hat{L}^{(N)}_0(t)$ and $L^{(N)}_0(t+1)$. Therefore, **[E:t+1]** follows from the continuous mapping theorem and these observations.

The distributional recursion of $L_{\mathbf{y}}(t+1),\ \mathbf{y}\in\mathcal{Y}$, is now easily established from (41), *i.e.,*

$$\begin{aligned} \sqrt{N}L^{(N)}_{\mathbf{y}}(t+1) \Longrightarrow_N & \sum_{\mathbf{y}_1\in\mathcal{Y}_{\mathbf{y},1}} L_{\mathbf{y}_1}(t) \\ & + \sum_{m=0}^{W_{\max}}\sum_{\mathbf{y}_2\in\mathcal{Y}_{\mathbf{y},2,m}} \upsilon_{\mathbf{y}_2,m}(t)+\gamma_{\mathbf{y}_2,m}(t)+\chi_{\mathbf{y}_2,m}(r(t))L_{\mathbf{y}_2}(t). \end{aligned}$$

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 $\mathbf{R}$, the Taylor series expansion*

$$e^{jx} = 1 + jx - \frac{x^2}{2} + R(x) \tag{50}$$

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

$$|R(x)| \leq \frac{|x|^3}{6}. \tag{51}$$

LEMMA 5. *[11, Thm. 5.20, p. 146] (Slutsky's theorem) If $X_N \Longrightarrow_N X$ and $Y_N \Longrightarrow_N y \in \mathbf{R}$, then $X_N Y_N \Longrightarrow_N y \cdot X$ and $X_N + Y_N \Longrightarrow_N X + y$.*

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*

$$\prod_{i=1}^N (1 - C_{N,i}) \xrightarrow{P}_N e^{-\lambda}. \tag{52}$$

We now proceed with the proof of Proposition 3: Fix $N = 1,2,\ldots$ and $\theta_{\mathbf{y},m} \in \mathbf{R},\ \mathbf{y} \in \mathcal{Y}, m = 0,\ldots,W_{\max}$. It suffices to show that

$$\begin{aligned} &\mathbf{E}\left[\exp\left(j\sqrt{N}\sum_{\mathbf{y},m}\theta_{\mathbf{y},m}\gamma_{\mathbf{y},m}^{(N)}(t)\right)|\mathcal{F}_t\right] \\ &\xrightarrow{P}_N \quad e^{-\frac{1}{2}\Theta_\Gamma^T \mathbf{S}(t)\Theta_\Gamma}, \end{aligned} \tag{53}$$

for $\Theta_\Gamma = [\theta_{\mathbf{y},m},\ \mathbf{y} \in \mathcal{Y}, m = 0,\ldots,W_{\max}]^T$. By conditional independence, we find that

$$\begin{aligned} &\mathbf{E}\left[\exp\left(j\sqrt{N}\sum_{\mathbf{y},m}\theta_{\mathbf{y},m}\gamma_{\mathbf{y},m}^{(N)}(t)\right)|\mathcal{F}_t\right] \\ &= \prod_{i=1}^N \exp\left(-\frac{j}{\sqrt{N}}\sum_{\mathbf{y},m}\theta_{\mathbf{y},m}\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}]\,\chi_{\mathbf{y},m}\left(\frac{R^{(N)}(t)}{N}\right)\right) \\ &\times\mathbf{E}\left[\exp\left(\frac{j}{\sqrt{N}}\sum_{\mathbf{y},m}\theta_{\mathbf{y},m}\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}]\,\mathbf{1}_{M_i^{(N)}(t+1)}[m]\right)|\mathcal{F}_t\right] \end{aligned} \tag{54}$$

$$\begin{aligned} &= \prod_{i=1}^N \exp\left(-\frac{j}{\sqrt{N}}\sum_{\mathbf{y},m}\theta_{\mathbf{y},m}\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}]\,\chi_{\mathbf{y},m}\left(\frac{R^{(N)}(t)}{N}\right)\right) \\ &\times\left[1 + \sum_{\mathbf{y},m}\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}]\,\chi_{\mathbf{y},m}\left(\frac{R^{(N)}(t)}{N}\right)(e^{\frac{j}{\sqrt{N}}\theta_{\mathbf{y},m}} - 1)\right] \\ &= \prod_{i=1}^N \left[1 - \frac{j}{\sqrt{N}}C_i(t;\theta) - \frac{1}{2N}C_i(t;\theta)^2 + r_{i,1}^{(N)}(t;\theta)\right] \\ &\quad\times\left[1 + \frac{j}{\sqrt{N}}C_i(t;\theta) - \frac{1}{2N}C_i(t;\theta^{\mathbf{2}}) + r_{i,2}^{(N)}(t;\theta)\right], \end{aligned} \tag{55}$$

by Taylor series expansion (Lemma 4), where

$$\begin{aligned} C_i(t;\theta) &= \sum_{\mathbf{y},m}\theta_{\mathbf{y},m}\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}]\,\chi_{\mathbf{y},m}\left(\frac{R^{(N)}(t)}{N}\right) \\ C_i(t;\theta^{\mathbf{2}}) &= \sum_{\mathbf{y},m}\theta_{\mathbf{y},m}^2\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}]\,\chi_{\mathbf{y},m}\left(\frac{R^{(N)}(t)}{N}\right) \end{aligned}$$

and

$$\begin{aligned} &\max(|r_{i,1}^{(N)}(t,\theta)|, |r_{i,2}^{(N)}(t,\theta)|) \\ &\leq \sum_{\mathbf{y},m}\frac{|\theta_{\mathbf{y},m}|^3}{6\sqrt{N^3}} \\ &\leq (W_{\max}+1)\cdot|\mathcal{Y}|\max_{\mathbf{y},m}\frac{|\theta_{\mathbf{y},m}|^3}{6\sqrt{N^3}}\ . \end{aligned} \tag{56}$$

It is easy to see that (55) can be rewritten as

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

where we set

$$C_i^{(N)}(t) = \frac{1}{2N}(C_i(t;\theta^{\mathbf{2}}) - C_i(t;\theta)^2) - \xi_i^{(N)}(t;\theta)$$

and the remainder term $\xi_i^{(N)}(t;\theta)$ can be shown to satisfy

$$\lim_{N\to\infty}\tfrac{1}{N}\textstyle\sum_{i=1}^N \xi_i^{(N)}(t;\theta) = 0 \quad a.s. \tag{57}$$

as a consequence of (56).

The desired result (53) is now a simple consequence of Lemma 6 provided that the required conditions can be shown to hold, *i.e.,*

$$\lim_{N\to\infty}\max_{i=1,\ldots,N}|C_i^{(N)}(t)| = 0 \quad a.s. \tag{58}$$

and

$$\sum_{i=1}^N C_i^{(N)}(t) \xrightarrow{P}_N \frac{1}{2}\Theta_\Gamma^T\mathbf{S}(t)\Theta_\Gamma. \tag{59}$$

Condition (58) trivially holds while Condition (59) can be established from

$$\begin{aligned} &\sum_{i=1}^N C_i^{(N)}(t) \\ &= \frac{1}{2N}\sum_{i=1}^N (C_i(t;\theta^{\mathbf{2}}) - C_i(t;\theta)^2) - \frac{1}{2N}\sum_{i=1}^N \xi_i^{(N)}(t;\theta) \\ &= \frac{1}{2N}\sum_{i=1}^N\sum_{\mathbf{y}}\mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}] \\ &\quad\times\left[\sum_m \theta_{\mathbf{y},m}^2\chi_{\mathbf{y},m}\left(\frac{R^{(N)}(t)}{N}\right)\left(1 - \chi_{\mathbf{y},m}\left(\frac{R^{(N)}(t)}{N}\right)\right)\right. \\ &\quad\left. - \sum_{k\neq n}\theta_{\mathbf{y},k}\theta_{\mathbf{y},n}\chi_{\mathbf{y},k}\left(\frac{R^{(N)}(t)}{N}\right)\chi_{\mathbf{y},n}\left(\frac{R^{(N)}(t)}{N}\right)\right] \\ &\quad - \frac{1}{2N}\sum_{i=1}^N \xi_i^{(N)}(t;\theta) \\ &\xrightarrow{P}_N \frac{1}{2}\sum_{\mathbf{y}}\mathbf{P}_{Y(t)}[\mathbf{y}]\left(\sum_m \theta_{\mathbf{y},m}^2\chi_{\mathbf{y},m}(r(t))(1 - \chi_{\mathbf{y},m}(r(t)))\right. \\ &\quad\left. - \sum_{k\neq n}\theta_{\mathbf{y},k}\theta_{\mathbf{y},n}\chi_{\mathbf{y},k}(r(t))\chi_{\mathbf{y},n}(r(t))\right), \end{aligned} \tag{60}$$

where the convergence follows from Theorem 1, Slutsky's Theorem (Lemma 5), and (57).

E.2 Proof of Lemma 2

Lemma 2 is a corollary of the Delta Method (Lemma 1). For $\mathbf{y} \in \mathcal{Y}$ such that $\Lambda(\mathbf{y}) \geq 1$ and $m = 0, 1, \ldots, \Lambda(\mathbf{y})$:

$$\begin{aligned}\sqrt{N} v_{\mathbf{y},m}^{(N)}(t) &= \sqrt{N}\left(\chi_{\mathbf{y},m}(\frac{R^{(N)}(t)}{N}) - \chi_{\mathbf{y},m}(r(t))\right) \\ &\quad \cdot \tfrac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}] \\ &= \begin{pmatrix} \Lambda(\mathbf{y}) \\ m \end{pmatrix} \sqrt{N} D_{\mathbf{y},m}^{(N)}(t) \cdot \tfrac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\mathbf{Y}_i^{(N)}(t)}[\mathbf{y}], \qquad (61)\end{aligned}$$

where we let

$$\begin{aligned}D_{\mathbf{y},m}^{(N)}(t) = f^{(N)}\left(R^{(N)}(t)\right)^m \left(1 - f^{(N)}\left(R^{(N)}(t)\right)\right)^{\Lambda(\mathbf{y})-m} \\ - f(r(t))^m \left(1 - f(r(t))\right)^{\Lambda(\mathbf{y})-m}. \qquad (62)\end{aligned}$$

The desired result follows directly by applying Slutsky's Theorem (Lemma 5) and the Delta Method (Lemma 1) to (61).

E.3 A Proof of Lemma 3

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

LEMMA 7. *[11, p. 150] (Cramer-Wold device) Let*

$$\begin{aligned}Z^{(N)}(t) &= \theta_0 L_0^{(N)}(t+1) + \Theta_{\mathbf{L}}^T \mathbf{L}^{(N)}(t) \\ &\quad + \Theta_{\Upsilon}^T \Upsilon^{(N)}(t) + \Theta_{\Gamma}^T \Gamma^{(N)}(t),\end{aligned}$$

for constant $\theta_0 \in \mathbf{R}$ and real-valued vectors (of appropriate dimension) $\Theta_{\mathbf{L}}, \Theta_{\Upsilon}, \Theta_{\Gamma}$.

The convergence in (47) holds if and only if for all choices of $\theta_0, \Theta_{\mathbf{L}}, \Theta_{\Upsilon}, \Theta_{\Gamma}$, we have

$$\begin{aligned}\sqrt{N} Z^{(N)}(t) \implies_N \theta_0 L_0(t+1) + \Theta_{\mathbf{L}}^T \mathbf{L}(t) \\ + \Theta_{\Upsilon}^T \mathbf{J}(t) L_{f_R}(t) + \Theta_{\Gamma}^T \Gamma(t).\end{aligned}$$

By Cramer-Wold device (Lemma 7), it suffices to show that

$$\begin{aligned}\mathbf{E}\left[e^{j\sqrt{N} Z^{(N)}(t)}\right] &\to_N e^{-\frac{1}{2}\Theta_{\Gamma}^T \mathbf{S}(t) \Theta_{\Gamma}} \\ &\quad \cdot \mathbf{E}\left[e^{j\left(\theta_0 L_0(t+1) + \Theta_{\mathbf{L}}^T \mathbf{L}(t) + \Theta_{\Upsilon}^T \mathbf{J}(t) L_{f_R}(t)\right)}\right]. \qquad (63)\end{aligned}$$

First note that

$$\begin{aligned}&\mathbf{E}\left[e^{j\sqrt{N} Z^{(N)}(t)}\right] \\ &= \mathbf{E}\left[e^{j\sqrt{N}\left(\theta_0 L_0^{(N)}(t+1) + \Theta_{\mathbf{L}}^T \mathbf{L}^{(N)}(t) + \Theta_{\Upsilon}^T \Upsilon^{(N)}(t)\right)}\right. \\ &\quad \left. \cdot \mathbf{E}\left[e^{j\sqrt{N}\Theta_{\Gamma}^T \Gamma^{(N)}(t)} | \mathcal{F}_t\right]\right] \\ &\to_N e^{-\frac{1}{2}\Theta_{\Gamma}^T \mathbf{S}(t) \Theta_{\Gamma}} \\ &\quad \cdot \mathbf{E}\left[\lim_{N\to\infty} e^{j\sqrt{N}\left(\theta_0 L_0^{(N)}(t+1) + \Theta_{\mathbf{L}}^T \mathbf{L}^{(N)}(t) + \Theta_{\Upsilon}^T \Upsilon^{(N)}(t)\right)}\right]\end{aligned}$$

where the convergence follows from Proposition 3. The desired result (63) follows if we can establish the following joint convergence from Cramer-Wold device

$$\begin{aligned}&\sqrt{N}\left(\theta_0 L_0^{(N)}(t+1) + \Theta_{\mathbf{L}}^T \mathbf{L}^{(N)}(t) + \Theta_{\Upsilon}^T \Upsilon^{(N)}(t)\right) \\ &\implies_N \left(\theta_0 L_0(t+1) + \Theta_{\mathbf{L}}^T \mathbf{L}(t) + \Theta_{\Upsilon}^T \mathbf{J}(t) L_{f_R}(t)\right). \qquad (64)\end{aligned}$$

From the proof of Proposition 2, we see that $\sqrt{N} L_0^{(N)}(t+1)$ can be written as a continuous map of

$$\sqrt{N}\left(L_0^{(N)}(t) + \sum_{\mathbf{y}\in\mathcal{Y}} \Lambda(\mathbf{y}) L_{\mathbf{y}}^{(N)}(t)\right),$$

regardless of the value of the residual capacity $K(t)$. Also note that $L_R(t)$ comprises only of components in $\mathbf{L}(t)$, and hence $L_{f_R}(t)$ is jointly convergence in law with $\mathbf{L}(t)$ by arguments similar to the one following (49). It is then easy to show that (64) holds from the continuous mapping theorem.