Stochastic Networks and Reflecting Brownian Motion: The Mathematics of Ruth Williams

Ioana Dumitriu, Todd Kemp, and Kavita Ramanan

Ioana Dumitriu is a professor of mathematics at the University of California, San Diego. Her email address is idumitriu@ucsd.edu.
Todd Kemp is a professor of mathematics at the University of California, San Diego. His email address is tkemp@ucsd.edu.
Kavita Ramanan is the Roland George Dwight Richardson University Professor of Applied Mathematics at Brown University. Her email address is kavita_ramanan@brown.edu.
Opening photo is Ruth Williams in 2016.
Communicated by Notices *Associate Editor Scott Sheffield.*
For permission to reprint this article, please contact: reprint-permission@ams.org.
DOI: https://doi.org/10.1090/noti2437

Introduction

When Ruth Williams enrolled at the University of Melbourne, Australia, as an undergraduate pursuing an honors BSc in mathematics, she launched a stellar mathematical career that spans five decades and is still going strong. After completing a second (research masters) degree in mathematics from Melbourne, she crossed the Pacific to begin her PhD studies at Stanford University. There were three women in her PhD cohort; by twist of fate, all three took an early reading course from Sam Karlin, and all three pursued dissertations related to probability theory.

Figure 1. Some members of the Bendigo Computer Club, circa 1970. Photo provided by Ruth Williams (center).

Williams took a number of probability courses from Kai Lai Chung at Stanford. He posed an open problem about stopped Feynman–Kac functionals and the reduced Schrödinger equation which he had solved in one dimension; she realized that, using methods from PDE, she could solve the higher-dimensional open question, which led to her first single-authored publication as a PhD student. Chung became her advisor. While she did not pursue further research in the direction of Schrödinger equations, she found stochastic processes particularly appealing: their study involves rigorous analysis, and they arise naturally in a wide range of applications. While taking a course from a young professor in the Stanford Business School, Michael Harrison, she learned about reflecting Brownian motion (RBM): a diffusion process constrained to stay inside a region by "reflecting" at the boundary, with many associated challenging open problems (at the time). During the remainder of her time at Stanford, and the following year (1983–84) as a Postdoctoral Visiting Member at the Courant Institute working with S. R. Srinivasa Varadhan, she worked on foundational theory for Brownian motion with oblique reflection in a wedge. This set the stage for the nature of much of her future work—development of rigorous theory motivated by applications.

She was recruited to UC San Diego by its historically very strong group in stochastic processes, anchored by Ron Getoor and Michael Sharpe.[1] She began her University of California career in a lively fashion: on her first campus visit, she was serendipitously "interviewed" by Paul Erdős (who spent a substantial amount of time in San Diego during this period).

Nearly four decades later, Ruth Williams is still at UC San Diego, where she is now a Distinguished Professor and holds the Charles Lee Powell Chair in Mathematics I. She is one of the most celebrated active probabilists in the world. Early recognition of her work came in the form of an Alfred P. Sloan Fellowship (1988–1992) and an NSF Presidential Young Investigator award (1987–1994), followed by an NSF Faculty Award for Women (1991–1997). Her fundamental contributions to 20th century probability theory—in particular stochastic processes—were honored with an invited talk at the 1998 International Congress of Mathematicians in Berlin, and with the prestigious Guggenheim Fellowship (2001–2002). She has had continuous NSF support since 1984.

Her accomplishments are so widely recognized that she has received highest honors from five professional associations: in addition to being an Inaugural Fellow of the AMS (2013), she is a Fellow of the Institute of Mathematical Statistics (1992), the American Association for the Advancement of Science (1995), the Institute for Operations Research and Management Sciences (2008), and the Society for Industrial and Applied Mathematics (2020). In 2016 she was awarded, jointly with Martin Reiman,[2] the highly prestigious John von Neumann Theory Prize from INFORMS, for seminal research contributions to the theory and applications of stochastic networks and their heavy traffic approximations.

Williams was elected to the American Academy of Arts and Sciences (2009), the National Academy of Sciences (2012), and was elected to be a Corresponding Member of the Australian Academy of Science (2018). She has been awarded honorary doctorates by the University of Melbourne and by La Trobe University, both in Melbourne, Australia.

Her research is interdisciplinary, involving the development of fundamental mathematical theory in order to provide insight into real-world phenomena from a variety of fields, ranging from communication networks to (more recently) systems biology. She has published more than eighty papers and two cornerstone books. Her 2006 textbook *Introduction to the Mathematics of Finance*, published by the AMS, is widely used in graduate courses in mathematical finance. Her 1983 book *Stochastic Integration* with Kai Lai Chung was, when it first appeared, the most comprehensive and comprehensible treatment of the subject, and it remains a highly regarded and widely used source today (the second author used the 2nd edition, published in 1990, as recently as 2019 as a primary source to teach a

[1] *UC San Diego seems to have made some excellent hires in 1983: Jim Agler and S. T. Yau also joined the faculty that year.*

[2] *Reiman was also a Stanford PhD student, supervised by Michael Harrison several years before Williams.*

popular advanced graduate course on stochastic differential equations). In addition to her transformative research, Ruth Williams is also widely known for the outstanding quality of her expository work: she has written several survey papers [Wil95, Wil16] that serve as introductions to the field and describe important open problems, and which have stimulated further research, including some early works by the third author. Williams is also a dynamic and highly skilled public speaker. In addition to her invited ICM address, she has given a long series of prestigious invited lectures such as a Plenary AMS Invited Address in 1994, the Markov Lecture of the Applied Probability Society in 2007, the Doob Lecture at the 2011 meeting on Stochastic Processes and their Applications, and the Le Cam Lecture at the IMS annual meeting in Vilnius, Lithuania in 2018.

Given the breadth and depth of Ruth Williams' work, along with the fact that she continues to be very active, it would be impossible to provide an exhaustive overview in this (or indeed any) article. Here, we will present the main themes of her prodigious career and highlight some of her most influential contributions. In the broadest possible terms, Ruth Williams has made foundational contributions to the understanding of *reflecting diffusions* and, using these as tools, has dramatically advanced the scientific understanding of a wide array of *stochastic networks* experiencing *heavy traffic*, by approximating them using rigorous probabilistic scaling limits.

Scaling limits. Probability theory has enjoyed over a century of remarkable success in analyzing and predicting the behavior of very complex systems. One reason is the central idea of scaling limits: encoding main parameters of a random system and scaling them together (in different proportions) to identify more tractable limit objects that can be analyzed more directly, and which then serve as useful approximations of the original system. This paradigm shows up in the first major theorems that are the capstone of any introductory probability course at the graduate or undergraduate level: the Strong Law of Large Numbers and the Central Limit Theorem. In their simplest form, these state that the *empirical average* of n independent identically distributed (i.i.d.) random variables $\{X_i\}_{i\in\mathbb{N}}$ converges, with probability one, to their common deterministic mean m (whenever the latter is well-defined) as $n \to \infty$, while in the case when X_i has finite variance, the *centered and rescaled sum* (divided by $\sqrt{n}$ instead of n) converges to an object that is still random: a normal random variable. In other words, taken together, these classical limit theorems show that empirical averages concentrate about their deterministic common mean, and have Gaussian fluctuations on the scale $n^{-1/2}$.

There are two versions of these core theorems that give different perspectives, which are relevant to the present story. Instead of averaging random variables X_i directly, consider their *empirical distribution* $\nu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$, which is a (random) probability measure-valued statistic that places equal mass at each point X_i. If the common law of the random variables is μ, then the Strong Law of Large Numbers tells us that ν_n converges *weakly in distribution* to μ almost surely, meaning that for each real-valued test function $f : \mathbb{R} \to \mathbb{R}$, $\int f\, d\nu_n$ converges to $\int f\, d\mu$ with probability one. The Central Limit Theorem then states that the fluctuations $\int f\, d\nu_n - \int f\, d\mu$ are of order $n^{-1/2}$, and $n^{1/2}[\int f\, d\nu_n - \int f\, d\mu]$ converges to a centered normal random variable (with variance $\int f^2\, d\mu - (\int f\, d\mu)^2$). (The original statements of these limit theorems mentioned above correspond to the special case $f(x) = x$.)

The above limit theorems concern real- or measure-valued random variables; we can consider analogous scaling limits for random elements of more exotic state spaces, such as paths (with some regularity, like continuity or at least right continuity with finite left limits) in some metric space. Such continuous-time stochastic processes model the evolution of dynamical systems that can have random influences. In this context, there are *functional* laws of large numbers and central limit theorems, with the latter also being referred to as *invariance principles*.

Let $\{X_i\}_{i\in\mathbb{N}}$ be i.i.d. *standardized* random variables (having mean zero and variance one), and denote $S(n) = X_1 + \cdots + X_n$. We can connect the dots (linear interpolation from $S(n-1)$ to $S(n)$) to create a piecewise affine random path $(S(t))_{t\geq 0}$. In this formulation, the two functional limit theorems can be phrased in terms of *rescaling space and time* in different proportions. For the Strong Law of Large Numbers, the statement is simply that, for any $t > 0$, $\lim_{r\to\infty} S(rt)/r = 0$ with probability one. (Had we not centered the random variables, the limit here would be the deterministic drift process $t \cdot m$ where m is the common mean of the random variables X_i.) The Central Limit Theorem in this context, known as Donsker's invariance principle, uses the different scaling $S^{(r)}(t) = S(rt)/\sqrt{r}$; here, the stochastic processes $S^{(r)}$ converge (weakly in distribution) as $r \to \infty$, to Brownian motion $B = (B(t))_{t\geq 0}$, the central Gaussian object in stochastic processes.

Brownian motion has quadratic scaling: for any $r > 0$, the new process $B^{(r)}(t) = B(rt)/\sqrt{r}$ is also a Brownian motion (it has the same law on path space). For this reason, one could just as well do the scaling $S(r^2t)/r$ in Donsker's theorem, going further out in time. There is a subtle but important difference when taken in concert with the law of large numbers, however: the pair $(S(rt)/r, S(rt)/\sqrt{r})$ scales differently from the pair $(S(rt)/r, S(r^2t)/r)$. In this example, it doesn't matter. In more complex examples the choice of whether to contract the space scaling or accelerate the time scaling can, in some circumstances, yield different results;

moreover, the latter can sometimes be more useful. This was a key insight, originally due to Michael Harrison, that played an important role in some of Ruth Williams' scaling limit theorems described below.

For more general stochastic processes (not arising from i.i.d. data), the law of large numbers kind of limit where space and time are scaled at the same rate is called a *fluid* or *hydrodynamic* limit, while scaling time with the square of the spatial rate is often referred to as a *diffusion* limit. A major theme of Ruth Williams' research program throughout her illustrious career has revolved around fluid and diffusion limits of a class of stochastic processes (discrete, continuous, or even measure-valued) that model *multiclass queueing networks* and more general *stochastic processing networks*. In order to provide some context for her work, we start by describing multiclass queueing networks, heavy traffic limits, and reflecting Brownian motions (RBMs).

Multiclass queueing networks. Queueing systems arise as models in a variety of applications, including computer systems, communication networks, transportation, service systems, and complex manufacturing systems. More recently, they have also been used by Ruth Williams in systems biology, for example as models of enzymatic processing. A *multiclass queueing network* consists of a fixed set of nodes (or stations), at which there are entities or jobs, which could represent customers or packets of data to be processed, and a server (or a pool of statistically homogeneous servers) capable of processing those jobs. The jobs at each node may belong to one of a finite number of types or classes depending on their arrival characteristics, service requirements, and routing needs, all of which may be random. A node is sometimes also referred to as a queue, which comprises the server(s), the jobs being processed, and the jobs awaiting processing at that node. If there is more than one class of job at a node, the node is called a multiclass queue; otherwise, it is called a single-class queue. Similarly, if there are multiple servers at a node, then it is referred to as a many-server queue; if not, it is said to be a single-server queue. When a job has finished service at a node, it either departs the system or changes class via a routing mechanism, which may be probabilistic. Networks in which jobs eventually leave the system are referred to as open networks. Networks can also differ in terms of the "service discipline" or protocol used by a server to process entities at its node. For example, under a head-of-the-line (or HL) protocol, entities of the same class that are awaiting service at a node are processed in the order in which they arrived to that node.

Quantities of interest in such systems include conditions for stability of the network dynamics, statistics of queue lengths of different classes of jobs at different nodes, the workload at each node (which is the amount of server effort required to serve all of the jobs at that node),

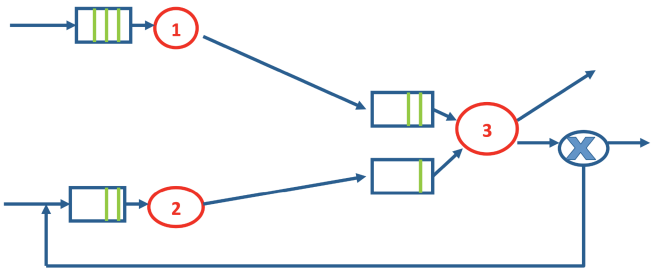

Figure 2. An example of a multiclass queueing network.

probabilities of critical rare events, and steady state or equilibrium distributions of these quantities. Starting with the Danish engineer A. K. Erlang in 1917, early work in queueing theory focused on exact closed form expressions for various statistics related to single-class queues. The first general results for networks were obtained by Jackson for open networks of single-server single-class HL queues with Markovian routing, where exogenous arrivals to each node are described by independent Poisson processes, jobs have independent exponentially distributed service times, and the service rate at each node is a function of the queue size. In particular, in 1963, Jackson showed that the equilibrium distribution of such networks has an explicit product form, which implies that in equilibrium, the numbers of jobs in distinct queues are independent. This was later generalized by several authors including Baskett et al. (1975) and Kelly (1979), who identified special classes of multiclass queueing networks that also have product form stationary distributions.

Heavy traffic limits and RBMs. However, beyond these special cases, typically it is not possible to compute performance measures of even HL multiclass queueing networks with general arrival processes and service distributions exactly. A particular regime of interest from an operations point of view is the so-called *heavy traffic regime*, where networks are congested or near capacity in the sense that the rate at which work is input to the system is approximately balanced by the capacity of the system to process that work. At such near-equilibrium regimes, performance can be strongly influenced by stochastic variability. Although early work of Kingman, Borovkov, and Prohorov in the early 1960s established approximations for steady-state distributions or finite-dimensional distributions of single-class queues, Iglehart and Whitt (1970) were the first to consider a functional heavy traffic approximation for a HL single-class (multi-server) queue, showing that a suitably rescaled job count process converges in distribution to a diffusion limit that is a so-called *reflecting Brownian motion* (RBM).

Standard Brownian motion takes both positive and negative values almost surely, and so it is not a good limit model for any random quantity that is by definition positive (like a queue length or workload process). Instead,

reflecting Brownian motion is a process whose increments coincide with that of Brownian motion on intervals when the process is positive, but is then modified when it hits zero. In fact, as shown by Skorokhod in 1961, one-dimensional reflecting Brownian motion Z can be represented as

$$Z(t) = B(t) + L(t), \tag{1}$$

where B is a standard Brownian motion and $L(t) := \sup_{s\in[0,t]} \max(-B(s), 0)$ is (proportional to) the so-called Brownian *local time*, which characterizes the amount of time Brownian motion spends near zero.

The construction in (1) yields the reflecting Brownian motion Z as a continuous function of the driving Brownian motion B. In the case where B is a standard one-dimensional Brownian motion, by a theorem of Lévy, the process Z has the same distribution as the "reflected" or absolute value process $|B|$; this is where the terminology "reflecting Brownian motion" comes from. This equivalence no longer holds true for Brownian motion with a drift, and in many higher-dimensional contexts, although the name is still used. As will be evident from the more precise definition given below, it is more accurate to think of a RBM or more general reflecting stochastic process as a process whose increments behave like those of the original process on the time intervals when the reflecting process lies in the interior of the state space, but is then suitably constrained to live within (the closure of) a domain (which is the non-negative reals in the one-dimensional case).

Skorokhod's idea was extended by Harrison and Reiman to study heavily loaded networks of single-class queues. In this case, each coordinate of the limit process represents the queue length at a node, and so the limit process must lie in the positive orthant. In 1981, Harrison and Reiman developed a multi-dimensional analog of the Skorokhod map in the positive orthant, and subsequently, Reiman exploited its continuity properties to show that the heavy traffic limit of open single-class HL queueing networks (with generally distributed interarrival and service times with finite moment conditions) is a reflecting Brownian motion in the orthant. Furthermore, their definition guaranteed that the process is a semimartingale, which means that it admits a decomposition as the sum of a (local) martingale and an adapted process that is (locally) of bounded variation. The semimartingale property is useful because it allows an easy application of stochastic calculus to study the evolution of sufficiently regular functionals of the process. The Skorokhod map is useful in that it is pathwise and, when continuous, it defines what is known as a strong solution to the corresponding stochastic differential equation with reflection (which means that the solution is measurable with respect to the filtration generated by the driving Brownian motion). However, it turns out that the Skorokhod map may fail to be well-defined or continuous for data associated with multiclass queueing networks and more general stochastic processing networks. An alternative is to consider distributional, rather than pathwise, limits and to characterize RBMs using the so-called submartingale problem introduced by Stroock and Varadhan in the 1970s to study (weak solutions to) stochastic differential equations with reflection in smooth domains with smooth boundary conditions.

Figure 3. Ruth Williams, Michael Harrison, and Jim Dai, at a conference in honor of Michael Harrison, 2009.

Ruth Williams' Contributions

Ruth Williams' mathematical career has centered on developing methodologies for the analysis of stochastic processing networks, proving hydrodynamic and heavy traffic limit theorems that yield fluid and diffusion approximations, and analyzing these approximations. Ruth Williams' most influential early work [VW85, Wil87, RW88, TW93, DW94] focused on developing the foundations of RBM in the orthant with discontinuities in the oblique reflection field at the boundary interfaces. At the time, there was limited theory for such non-smooth, non-symmetric situations, where novel behavior such as hitting corners can occur (in contrast to Brownian motion which hits individual points with probability zero in dimensions greater than one). In applications to queueing networks, the oblique reflection directions arise from routing in the network, and the orthant state space represents the fact that queue lengths are always non-negative with intersections of faces corresponding to several queues being empty simultaneously. This work on RBMs is beautifully summarized in the survey paper [Wil95], in which Williams succinctly defines RBMs in such domains, and discusses existence and uniqueness in law and characterizations of stationary distributions.

After establishing the foundations for these RBMs and the appearance in the early 1990s of surprising examples showing that the stability and heavy traffic behavior of multiclass queueing networks are more intricate than that

of single-class queueing networks, Ruth Williams turned to establishing invariance principles and heavy traffic limit theorems for multiclass queueing networks. An excellent short survey is in [Wil98a] (the paper accompanying her 1998 ICM talk), which describes the general modular framework she developed (with Maury Bramson) for establishing sufficient conditions for heavy traffic limit theorems for HL multiclass queueing networks.

Subsequently, at the turn of the century, she started analyzing more general stochastic processing networks, including those with resource sharing, such as in processor sharing and bandwidth sharing networks [GPW02, KW04, KKLW09, MPW19, PW16, FW21]. Often in resource sharing, service is shared amongst all entities and one needs to keep track of more information than queuelengths to describe the dynamics; also, these are non-head-of-the-line (non-HL) networks. This presents new mathematical challenges, which Williams overcame by introducing measure-valued stochastic processes to represent the dynamics of these networks and by developing new techniques for proving hydrodynamic and heavy traffic limit theorems for them.

Over the last fifteen years, catalyzed by participation in a meeting at the Institute for Mathematics and its Applications (IMA), Williams has also expanded her research to include applications in systems biology [MHTW10, LW19, AHLW19].

In what follows, we describe some of her research contributions in greater detail.

(i) Reflecting Brownian motion (RBM). We start by defining reflecting Brownian motions in domains with piecewise smooth boundaries [KW07]. Let $\{G_i\}_{i\in\ell}$ be a finite collection of open subsets of $\mathbb{R}^d$, each with continuously differentiable boundary, and let $G = \bigcap_{i\in\ell} G_i$. Fix vector fields $\gamma^i : \mathbb{R}^d \to \mathbb{R}^d$, $i \in \ell$. A *semimartingale reflecting Brownian motion* (SRBM) Z is a stochastic process on a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geq 0}, \mathbb{P})$ taking values only in $\overline{G}$, which has a decomposition of the form

$$Z(t) = X(t) + \sum_{i\in\ell} \int_{(0,t]} \gamma^i(Z(s))\, dY_i(s),$$

where X is a Brownian motion in $\mathbb{R}^d$ with respect to the filtration $(\mathscr{F}_t)_{t\geq 0}$ (with initial distribution supported in $\overline{G}$, and some fixed drift and covariance), and each Y_i is a continuous, adapted, non-decreasing process that only increases at times s when $Z(s) \in \partial G \cap \partial G_i$ (i.e., Z lies on the corresponding part of the boundary of the domain). In the one-dimensional setting where $G = (0, \infty)$, the process $Y = Y_1$ is the Brownian local time L and the vector field is simply $\gamma^1(x) = 1$ pointing into the region. In general, there is no reason to assume that the vector field γ^i is a normal vector field on ∂G_i. In particular, the geometry of the directions of reflection that arise in heavy traffic limit theorems for queueing networks is dictated by the routing structure in the network, and generally leads to *obliquely* reflecting Brownian motions. Such reflecting Brownian motions are related to elliptic PDE with oblique derivative boundary conditions in much the same way that Brownian motion is related to the Laplace equation.

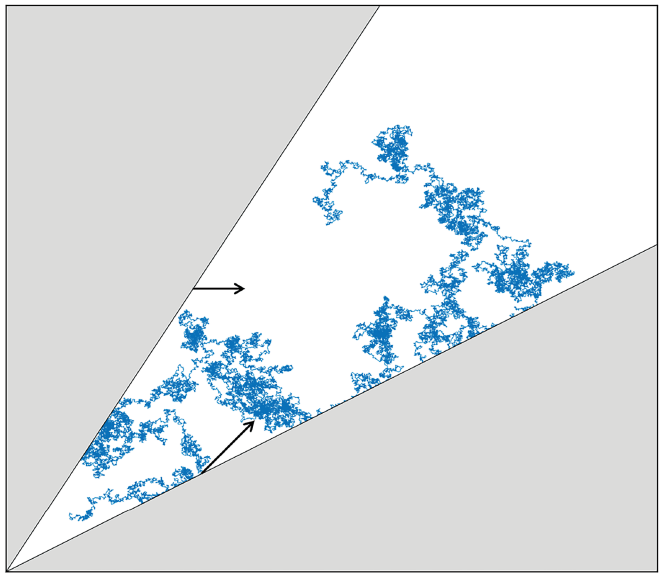

Figure 4. A simulation of a two-dimensional RBM (reflecting Brownian motion) in a wedge, with oblique reflection field. This simulation was provided by Prof. Xinyun Chen at the School of Data Science (SDS) in the Chinese University of Hong Kong, Shenzhen.

It is far from obvious that SRBMs should exist, and indeed some natural conditions on the domain G and the vector fields γ^i are required. The vector fields must be sufficiently regular and must, in a general sense, "point inward" on the boundary to have any chance of pushing the Brownian motion back into G when it tries to escape. More precisely, at each point $x \in \partial G$, some convex combination of the vectors $\gamma^i(x)$ for $i \in \ell$ such that $x \in \partial G \cap \partial G_i$ should point inward into G. In addition, the vector fields should not stray "too far" from the unit normal field, in a broad sense—this is to guarantee that the process does not oscillate too wildly near boundary intersections to be reflected in a meaningful way.

A special case of broad interest is when G is a polyhedral domain in the positive orthant, and the vector fields are constant on each face. In that case, the definition becomes somewhat simpler: $Z = X + RY$ where $Y = [Y_1, \ldots, Y_d]^\top$ as above and R is a $d \times d$ matrix, called the *reflection matrix* (or, more accurately, constraint matrix), whose columns are the (constant) vector fields. When G is the entire positive orthant, Ruth Williams and her student Lisa Taylor [TW93] identified sufficient conditions on the vector fields

for the (weak) existence and uniqueness in law of SRBMs, and showed that the conditions were also necessary in a paper with Martin Reiman [RW88]. The technical conditions on the vector fields can be stated succinctly in an algebraic form that R is a *completely-$\mathcal{S}$ matrix*, which means that for every principal submatrix $\tilde{R}$ of R, there is a vector $\tilde{y}$ in the positive orthant for which $\tilde{R}\tilde{y}$ lies in the positive orthant. This is a subtle result. Firstly, it should be noted that this condition is only necessary for a *semimartingale* RBM to exist; one can still have a well-posed RBM that is not a semimartingale when the completely-$\mathcal{S}$ condition fails and such non-semimartingale RBMs can also arise as heavy traffic limits of multiclass queueing networks. Moreover, the RBM constructed here is what is known as a weak solution to the stochastic differential equation with reflection. A longstanding open question that is still unresolved is whether strong solutions also exist under this condition.

Constraining Brownian motion to stay in a region by pushing it in the allowed "reflection" or constraint directions at the boundary can be thought of as a stochastic control problem, with highly singular controls. Proving that such processes exist and are unique in law is highly nontrivial. In the general piecewise smooth boundary case covered in [KW07], the proof of existence was tied together with the other side of the story: an invariance principle describing when an SRBM (or rather an *extended* SRBM, consisting of the triple (X, Y, Z)) arises as the diffusion scaling limit of a system $(X^{(r)}, Y^{(r)}, Z^{(r)})$ that only satisfies the boundary control approximately. (Again, the main technical hurdle is controlling oscillations at the boundary; achieving this even locally turns out to be enough to guarantee the requisite tightness for the diffusion limit to emerge.) Kang and Williams then proved the existence of such general SRBMs by exhibiting approximate extended systems and constructing the SRBM as their diffusion scaling limit.

(ii) Stationary distributions of RBMs. Williams simultaneously also initiated the study of the stationary distributions of SRBMs, which are Markov processes; this was natural given the importance of stationary measures for stochastic networks and Reiman's (1982) result on RBMs in the orthant arising as heavy traffic limits of open HL single-class networks. With Paul Dupuis [DW94], she obtained a general sufficient condition for the positive recurrence (or ergodicity) of SRBMs in the orthant, which reduced the problem to studying the long-time behavior of a deterministic constrained dynamical system (the "fluid" model) in the orthant. Next, in view of the fact that one-dimensional RBM is well-known to have a stationary distribution of exponential form and Jackson's result on product-form stationary distributions for a special class of open single-class networks, she set out to identify when SRBMs also exhibit analogous product-form or exponential stationary distributions.

In [HW87], Michael Harrison and Ruth Williams first studied this question for obliquely reflecting SRBMs on bounded domains with smooth boundaries governed by a smooth, possibly oblique, but non-tangential "reflection" vector field γ. Existence and uniqueness in law of such SRBMs follows from classical results of Stroock and Varadhan, and drawing on the classical connection between stationary distributions of reflected processes and elliptic PDE with (oblique) derivative boundary conditions, they showed that the RBM has an explicit stationary density with an exponential product form if and only if γ satisfies the following *skew-symmetry condition*: for all $x, \tilde{x} \in \partial G$,

$$\langle n(x), \gamma(\tilde{x}) - n(\tilde{x})\rangle + \langle \gamma(x) - n(x), n(\tilde{x})\rangle = 0,$$

where n denotes the inward normal vector field on the boundary ∂G, and $\gamma - n$ is the tangential part of γ.

For the case of RBMs (with covariance equal to the identity matrix) in a polyhedral domain in $\mathbb{R}^d$ with normal vector field n^i and a constant reflection vector field γ^i on the ith face, they also studied the formal analogue of the analytical PDE characterization of the stationary density that arises in the smooth case, dubbed it the *basic adjoint relation (BAR)*, and showed that the solution p of the BAR has the form

$$p(x) = \prod_{i=1}^{d} \exp(c_i x_i), \quad i = 1, \dots, d$$

for suitable real-valued constants $c_i, i = 1, \dots, d$, if and only if for any two distinct faces of G, labelled i and j,

$$\langle n^i, \gamma^j - n^j\rangle + \langle n^j, \gamma^i - n^i\rangle = 0.$$

In the particular case where the domain is the orthant and the associated reflection matrix R is normalized to have 1's along its diagonal, this reduces to saying that the matrix $R - I$ is skew-symmetric; here I is the $d \times d$ identity matrix. In all cases, the explicit dependence of the vector $c = (c_1, \dots, c_d)$ on the drift of the RBM was identified. In a separate paper [Wil87], Williams justified that the solution of the BAR is indeed the stationary density of the corresponding RBM. This was done by approximating the SRBMs in polyhedral domains with piecewise constant reflection vector fields by RBMs in certain approximating smooth domains with smooth vector fields, and showing that under the skew-symmetry condition, the RBM does not reach the non-smooth parts of the boundary. The latter property is of independent interest and in general fails when the skew-symmetry condition does not hold. These general results for RBMs have also been used in other applications, such as Atlas models in finance, which describe the evolution of equity markets in terms of rank-based stochastic differential equations.

(iii) Heavy traffic limits and multiplicative state space collapse. Although it was well-known that not all multiclass queueing networks could be approximated by SRBMs, Williams and Maury Bramson [Wil98a] laid out a modular approach to identifying classes of networks for which such an approximation is possible. Specifically, they identified general sufficient conditions under which such a limit theorem holds: first, the reflection matrix describing the SRBM must satisfy the completely-$\mathcal{S}$ condition and second, one must check whether a certain multiplicative state space collapse condition holds. As mentioned above, the completely-$\mathcal{S}$ condition guarantees well-posedness of the associated SRBM, and also enables proof of an invariance principle [Wil98b] that shows convergence of approximating processes to the SRBM under general conditions. The second condition is to establish what is known as *multiplicative state space collapse*, which is a generalization of the notion of state space collapse first considered by Reiman in 1984 and later used by Peterson in his 1991 work on heavy traffic limits for feedforward networks. Loosely speaking, state space collapse holds if, in diffusion scale, the job count process can be approximately recovered from the (typically lower-dimensional) workload process and the precision of this approximation becomes exact in the heavy traffic limit. The multiplicative version introduced by Bramson involves a normalization by the amount of work in the system. It is often easier to verify and can frequently be shown to imply state space collapse. Bramson and Williams also verified the sufficient conditions for several classes of networks including first-in-first-out (FIFO) networks of so-called Kelly type and networks with a HL proportional processor sharing service discipline. Taken together, these results represent a culmination of one and a half decades of focused effort by Ruth Williams to develop the requisite mathematical theory to identify and rigorously justify heavy traffic approximations of several families of multiclass queueing networks.

(iv) Resource sharing in stochastic processing networks. Having brought some measure of order to the understanding of HL multiclass queueing networks, at the turn of the century Ruth Williams started studying resource sharing problems in more general stochastic processing networks. With her student Steven Bell, Williams considered dynamic scheduling (or control) for parallel server systems with HL scheduling policies, and then later shifted her focus to the study of the non-HL Processor Sharing (PS) scheduling policy, and more general bandwidth sharing networks. The PS protocol, in which each server at any time divides its processing capacity equally amongst all jobs present in the queue at that time, seeks to provide an egalitarian allocation of a scarce resource among competing users and is an idealization of the round-robin protocol in time sharing computer systems.

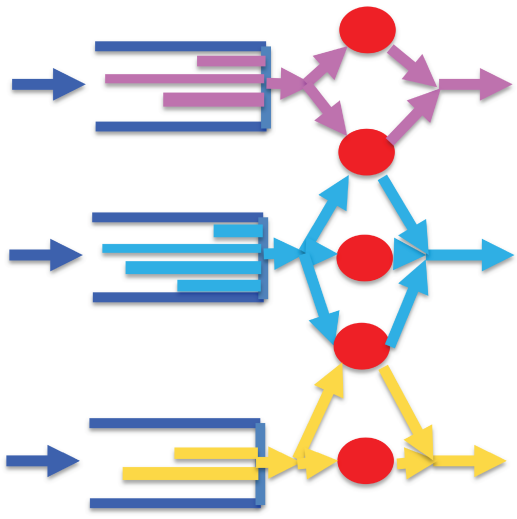

Figure 5. An example of a bandwidth sharing stochastic processing network. The split arrows indicate simultaneous resource possession.

There is a large body of literature on PS queues. However, with only rare exceptions, most of the literature imposes the stringent parametric assumptions of Poisson arrivals and/or exponential service requirements. Under these assumptions, the queue process is a Markov process in the sense that its instantaneous evolution at any time depends only on its current state (and not on the history), which greatly simplifies the analysis. Unfortunately, these assumptions are typically not satisfied in real-world applications.

In [GPW02], another paper written with her postdoctoral fellow Amber Puha, and the PhD thesis of her student Christian Gromoll, fluid and diffusion approximations were developed for PS queues with arrivals that form what is known as a renewal process, and jobs that have independent and identically distributed general service requirements. Since, under these more general distributions, the queue process on its own need no longer be Markovian, they introduced a measure-valued state representation that at any time t has a mass at the residual (remaining processing) service time of each job, from which one can recover traditional performance measures such as queue length and workload. They then established fluid and diffusion limit theorems for this measure-valued process.

These three papers together garnered the authors a "Best Publication Award" from the INFORMS Applied Probability Society in 2007. The citation stated that these papers "solve outstanding difficult problems, which advance the state of the art of Applied Probability." Ruth Williams' commitment to mentorship is evident from the fact that she told the last author of the present article at the time that the best thing about the award was that it was given jointly with her mentees.

Figure 6. INFORMS Best Publication Award prize ceremony, 2007. From left to right: Amber Puha, Christian Gromoll, Ruth Williams, Jim Dai.

These papers also served as the starting point for the study of bandwidth sharing communication networks. In 2000, Massoulié and Roberts introduced a connection-level model of Internet congestion control that represents the randomly varying flows in a network where bandwidth is shared fairly between file transfers, with fairness modulated by a parameter α. With Poisson arrivals and exponentially distributed file sizes, this model can be phrased as a multi-dimensional Markov chain in which the transition rates are solutions of concave optimization problems. Conditions for stability (positive recurrence) of this Markov chain were established early on, but characterizing the heavy traffic behavior was more challenging, because these are stochastic processing networks with simultaneous resource possession in which processing of files uses capacity from multiple resources simultaneously.

In 2001-02, while visiting Stanford on her Guggenheim fellowship, Ruth Williams initiated a collaboration on this problem with Frank Kelly, who was also visiting Stanford that year. Their goal was to obtain heavy traffic diffusion approximations for α-fair bandwidth sharing models by extending to this more complicated setting the approach developed earlier by Ruth Williams and Maury Bramson for multiclass queueing networks. First, in the work [KW04] with Kelly, Ruth Williams established long-time convergence of critical fluid model solutions to the set of invariant states. Then in [KKLW09], with Kelly and Ruth Williams' PhD students Weining Kang and Nam Lee, she used the asymptotic behavior of the fluid model to establish a dimension reduction called *multiplicative state space collapse*. Furthermore, in the case $\alpha = 1$, which corresponds to the natural case of proportional fair sharing of bandwidth, the multiplicative state space collapse property was combined with an invariance principle Ruth Williams established with W. Kang in [KW07] and her previous results on well-posedness of reflected diffusions in polyhedral domains, to show that the heavy traffic limit is a reflected diffusion in a polyhedral cone. In this case, it can also be deduced from previous work of Ruth Williams with Michael Harrison [HW87] that the stationary distributions of the heavy traffic limit are explicit and of product-form.

It should be emphasized that these limit theorems do not merely yield mathematical statements, but actually shed insight into the qualitative phenomenon of entrainment in these networks, whereby congestion at some resources may prevent other resources from working at their full capacity. Ruth Williams continues to work on this problem, with the ultimate goal to generalize these results to cover a more realistic version of this model that has generally distributed file sizes. In this case, the dynamics are represented by measure-valued processes, where understanding long-time behavior is much more complicated. Building on related works with Justin Mulvany and Amber Puha for the processor sharing model [MPW19, PW16], Ruth Williams and her PhD student Yingjia Fu have made recent progress on this subject. Specifically, in [FW21], Fu and Williams construct Lyapunov functions based on f-divergence (a generalization of relative entropy) to understand the long-time behavior of critical (measure-valued) fluid models in the presence of general file size distributions.

Figure 7. Ruth Williams working with her student Yingjia Fu, 2019.

(v) Constrained Langevin approximations for biochemical reaction networks. Key processes in chemical and biological systems are described by complex networks of chemical reactions, which are frequently not amenable

to exact analysis. Classically, the evolution of molecular concentrations is often modelled by coupled systems of nonlinear differential equations, which can be justified via a functional law of large numbers, in the limit as the number of molecules of all species goes to infinity. However, in systems biology the concentrations of some constituent molecules can be low, and thus deterministic models are inadequate. A common stochastic model of chemical kinetics treats the system as a continuous time Markov chain that tracks the number of molecules of each chemical species, and quantities of interest are then approximated by Monte Carlo estimates using simulations of the sample paths. However, since each reaction is accounted for in this model, these simulations can become computationally prohibitive even for a modest number of species. When the number of molecules is moderately large (though still not sufficiently large to ignore stochastic fluctuations), this model is often replaced by solutions of associated stochastic differential equations (SDE), referred to as diffusion approximations, which can be simulated more efficiently. Two commonly used diffusion approximations are the so-called linear noise approximation, obtained by linearizing fluctuations around the deterministic approximation, and the chemical Langevin equation. However, both approximations have serious drawbacks. The linear approximation fails to capture fluctuations due to nonlinearities in the reaction rates and, unlike the Markov chain models, its solution can become negative, which is not physically meaningful. On the other hand, the Langevin equation is better at capturing nonlinearities and serves as a good approximation as long as it is valid, but since its coefficients involve square roots of the concentrations of the species, it is typically ill-posed beyond the first time any coordinate of the solution reaches zero.

Several alternative models to deal with this negativity issue were proposed, including other Langevin-type models as well as hybrid methods that tried to combine the accuracy and robustness of the Markov chain models with the computational efficiency of diffusion approximations. Ruth Williams realized that some of the fixes unnecessarily perturb the *global* dynamics to deal with what is inherently a *local* issue (near the boundary of the orthant); she instead proposed a *constrained Langevin approximation*, which is an obliquely reflected diffusion in the orthant satisfying the non-negativity constraints of the component processes [AHLW19, LW19]. She presented preliminary results on this work as part of her Kolmogorov lecture at the World Congress in Probability and Statistics in July 2016. As demonstrated there, this approximation agrees with the chemical Langevin approximation until the first time any component goes negative, but is well-defined for all time and performs better than the existing approximations. Subsequently, Ruth Williams and Saul Leite rigorously showed that this reflected diffusion process arises as the weak limit of a sequence of jump-diffusion Markov processes that mimic the Langevin system in the interior and behave like a scaled version of the Markov chain on the boundary [LW19], which in particular required generalizing previous results on well-posedness of reflecting diffusions.

Continuing Legacy

Through her extraordinary continuing career, Ruth Williams has left a large imprint on probability theory and on mathematics in general. Her influence has been felt not only through her groundbreaking research, but through her direct involvement in the community. She has advised eleven PhD students (all of whom graduated from UC San Diego) and she is currently advising three more. She has supervised many postdoctoral fellows, masters students, and undergraduates (at UC San Diego). The research work that she did with her advisees and mentees has earned many accolades, some of which were highlighted above, and others are too numerous to mention.

Another constant in Ruth Williams' career has been her unwavering commitment to supporting and promoting women and underrepresented minorities. From organizing and speaking at women-centered and AWM-sponsored mathematical conferences, to extensive mentorship of junior colleagues and involvement in university-wide postdoctoral initiatives at the University of California, she has always been a strong advocate for the advancement of underrepresented groups in mathematics and science. In recognition of her dedication to this cause, INFORMS presented her with the prestigious Award for the Advancement of Women in Operations Research and Management Sciences (2017).

In conjunction with her many research accomplishments and accolades, Williams has provided a truly astonishing array of service to her department, to UC San Diego, and to the international mathematics and scientific communities. A complete list would go on for pages; we mention only a few highlights here. She has devoted decades to editorial boards of highly respected journals such as *Annals of Applied Probability, Electronic Journal of Probability* and *Electronic Communications in Probability*, and *Mathematics of Operations Research*. She has served on the Council (2003–2006) and as President (2011–2012) of the IMS (Institute of Mathematical Statistics). As IMS President, she spearheaded the effort to become an Associate Member of ICIAM (the International Council for Industrial and Applied Mathematics) to foster stronger ties to the applied mathematics community. In order to be more welcoming to junior researchers, she also arranged for tutorials to be added to the annual SSP (Seminar on Stochastic

Processes); in particular, she created a subcommittee of the IMS New Researchers Committee to suggest speakers for the SSP tutorials.

She was on the Bernoulli Society Council (2001–2004) and on the Board of Governors for the Institute for Mathematics and its Applications (2003–2006). She served as Chair of the Joint Program Committee for the 7th World Congress in Probability and Statistics (2008). She helped found the Steering Committee of the *Stochastic Networks* conference series initiated by Peter Glynn, Thomas Kurtz, and Peter Ney. She was a member of the Governing Board for the Australian Mathematics Research Institute, MATRIX (2015–2020). She currently serves on the Governing Council and the Executive Committee of the National Academy of Sciences. She has profoundly broken the stereotype partitioning mathematicians into those who are talented at research and mentorship and those who are devoted to service; Ruth Williams is a paragon of the mentor-scholar-academic.

Figure 8. Group photo from a conference in honor of Ruth Williams at the IMA, 2016. `https://www.ima.umn.edu/2014-2015/SW6.25-27.15`.

In San Diego, Ruth Williams met and married Bill Helton: a fellow UC San Diego mathematician who, like her, straddles the divide between pure and applied mathematics. These days, they enjoy spending their leisure time outdoors, gardening or hiking. He has been her constant companion and, in recent years, occasional collaborator. As it happens, her initial forays into systems biology applications included her first paper coauthored with Gheorghe Craciun and Bill Helton, on homotopy methods for counting equilibria in dynamic models of chemical reaction networks.

Williams is as active as ever, finding new ways to use mathematics to explain the world around us. Her current major research interests include stochastic models in systems biology, and entropy methods in the analysis of stochastic processing networks. On the first front, she has collaborated with the biodynamics lab at UC San Diego, led by Jeff Hasty and Lev Tsimring, on enzymatic processing networks. In connection with this area, she has worked with a PhD student, David Lipshutz, on (stochastic) differential delay equations relating to delayed protein degradation. Stochastic modeling of genetic circuits holds the promise of new understanding in cellular and molecular biology, a rapidly expanding quantitative field. She is currently collaborating with Domitilla Del Vecchio and Ron Weiss at MIT on stochastic modeling of epigenetic cell memory.

On the second front, Williams' current work using entropy-like notions has been very fruitful in analyzing fluid limits of certain non-HL systems: bandwidth sharing networks. These constitute just one of a huge number of non-HL real world networks, and there are many reasons to believe the Lyapunov approach can help understand these. This has the potential to make a huge impact on the field, since the relationship between bandwidth sharing models and more general non-HL stochastic processing networks is analogous to the relationship between bananas and non-banana fruits. Ruth Williams will no doubt leave a lasting mark on these problems—as she is fond of saying, "I eat problems for breakfast."

References

[AHLW19] David F. Anderson, Desmond J. Higham, Saul C. Leite, and Ruth J. Williams, *On constrained Langevin equations and (bio)chemical reaction networks*, Multiscale Model. Simul. **17** (2019), no. 1, 1–30, DOI 10.1137/18M1190999. MR3895328

[DW94] Paul Dupuis and Ruth J. Williams, *Lyapunov functions for semimartingale reflecting Brownian motions*, Ann. Probab. **22** (1994), no. 2, 680–702. MR1288127

[FW21] Y. Fu and R. J. Williams, *Asymptotic behavior of a critical fluid model for bandwidth sharing with general file size distributions*, 2021. To appear.

[GPW02] H. Christian Gromoll, Amber L. Puha, and Ruth J. Williams, *The fluid limit of a heavily loaded processor sharing queue*, Ann. Appl. Probab. **12** (2002), no. 3, 797–859, DOI 10.1214/aoap/1031863171. MR1925442

[HW87] J. M. Harrison and R. J. Williams, *Brownian models of open queueing networks with homogeneous customer populations*, Stochastics **22** (1987), no. 2, 77–115, DOI 10.1080/17442508708833469. MR912049

[KKLW09] W. N. Kang, F. P. Kelly, N. H. Lee, and R. J. Williams, *State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy*, Ann. Appl. Probab. **19** (2009), no. 5, 1719–1780, DOI 10.1214/08-AAP591. MR2569806

[KW04] F. P. Kelly and R. J. Williams, *Fluid model for a network operating under a fair bandwidth-sharing policy*, Ann. Appl. Probab. **14** (2004), no. 3, 1055–1083, DOI 10.1214/105051604000000224. MR2071416

[KW07] W. Kang and R. J. Williams, *An invariance principle for semimartingale reflecting Brownian motions in domains with piecewise smooth boundaries*, Ann. Appl. Probab. **17** (2007), no. 2, 741–779, DOI 10.1214/105051606000000899. MR2308342

[LW19] Saul C. Leite and Ruth J. Williams, *A constrained Langevin approximation for chemical reaction networks*, Ann. Appl. Probab. **29** (2019), no. 3, 1541–1608, DOI 10.1214/18-AAP1421. MR3914551

[MHTW10] W. H. Mather, J. Hasty, L. S. Tsimring, and R. J. Williams, *Correlation resonance generated by coupled enzymatic processing*, Biophys. J. **99** (2010), 3172–81.

[MPW19] Justin A. Mulvany, Amber L. Puha, and Ruth J. Williams, *Asymptotic behavior of a critical fluid model for a multiclass processor sharing queue via relative entropy*, Queueing Syst. **93** (2019), no. 3-4, 351–397, DOI 10.1007/s11134-019-09629-8. MR4032930

[PW16] Amber L. Puha and Ruth J. Williams, *Asymptotic behavior of a critical fluid model for a processor sharing queue via relative entropy*, Stoch. Syst. **6** (2016), no. 2, 251–300, DOI 10.1214/15-SSY198. MR3633537

[RW88] M. I. Reiman and R. J. Williams, *A boundary property of semimartingale reflecting Brownian motions*, Probab. Theory Related Fields **77** (1988), no. 1, 87–97. MR921820

[TW93] L. M. Taylor and R. J. Williams, *Existence and uniqueness of semimartingale reflecting Brownian motions in an orthant*, Probab. Theory Related Fields **96** (1993), no. 3, 283–317, DOI 10.1007/BF01292674. MR1231926

[VW85] S. R. S. Varadhan and R. J. Williams, *Brownian motion in a wedge with oblique reflection*, Comm. Pure Appl. Math. **38** (1985), no. 4, 405–443, DOI 10.1002/cpa.3160380405. MR792398

[Wil16] R. J. Williams, *Stochastic processing networks*, Annual Review of Statistics and Its Application **3** (2016), 323–345.

[Wil87] R. J. Williams, *Reflected Brownian motion with skew symmetric data in a polyhedral domain*, Probab. Theory Related Fields **75** (1987), no. 4, 459–485, DOI 10.1007/BF00320328. MR894900

[Wil95] R. J. Williams, *Semimartingale reflecting Brownian motions in the orthant*, Stochastic networks, IMA Vol. Math. Appl., vol. 71, Springer, New York, 1995, pp. 125–137. MR1381009

[Wil98a] R. J. Williams, *Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse*, Queueing Systems Theory Appl. **30** (1998), no. 1-2, 27–88, DOI 10.1023/A:1019108819713. MR1663759

[Wil98b] R. J. Williams, *An invariance principle for semimartingale reflecting Brownian motions in an orthant*, Queueing Systems Theory Appl. **30** (1998), no. 1-2, 5–25, DOI 10.1023/A:1019156702875. MR1663755

Ioana Dumitriu

Todd Kemp

Kavita Ramanan

Credits

Opening photo is courtesy of Erik Jepsen/UC San Diego Publications.

Figures 1, 2, 3, and 5 are courtesy of Ruth Williams.

Figure 4 is courtesy of Xinyun Chen.

Figure 6 is courtesy of Marty Reiman.

Figure 7 is courtesy of UC San Diego Publications.

Figure 8 is courtesy of the Institute for Mathematics and its Applications.

Photo of Ioana Dumitriu is courtesy of Alex Matthews/UCSD.

Photo of Todd Kemp is courtesy of UC San Diego Publications.

Photo of Kavita Ramanan is courtesy of Kavita Ramanan.