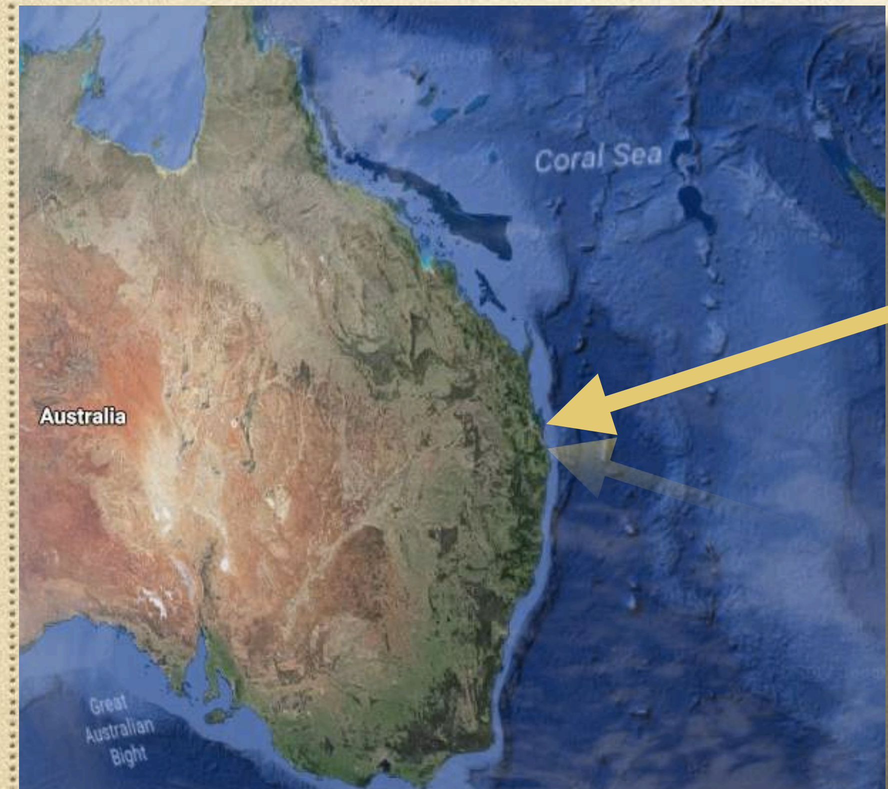
Parameter Estimation for Queues:
A Tale of Six Observations Schemes

Yoni Nazarathy, The University of Queensland,
Joint with Azam Asanjarani and Peter Taylor.

Research supported by ARC Grant, DP180101602.

INFORMS-APS 2019, Brisbane Australia

The 20th INFORMS Applied Probability Society Conference, July 2-4, 2019, Brisbane Australia.

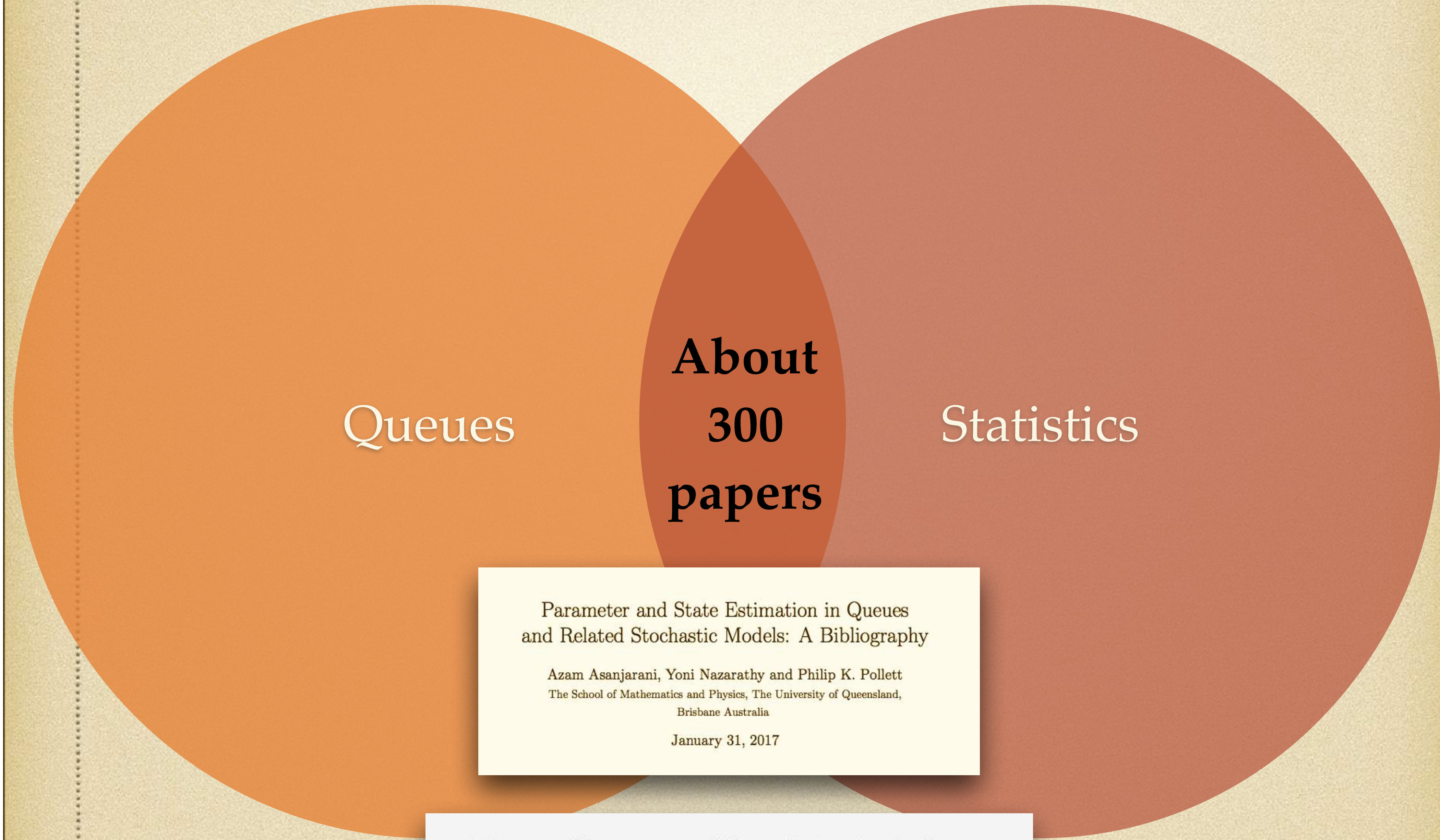


INFORMS-APS 2019, Brisbane Australia

The 20th INFORMS Applied Probability Society Conference, July 2-4, 2019, Brisbane Australia.

A number of related events are being held before and after the INFORMS-APS conference:

- Queues, Modelling, and Markov Chains: A Workshop Honouring Prof. Peter Taylor, June 28 - July 1.
- Applied^2 Probability, July 5 taking place at The University of Queensland, Brisbane. This is a satellite workshop dealing with concrete applications and is organized by CARM.
- 12th International Conference on Monte Carlo Methods and Applications (MCM2019), July 8 - 13 taking place in Sydney, Australia.



Parameter and State Estimation in Queues
and Related Stochastic Models: A Bibliography

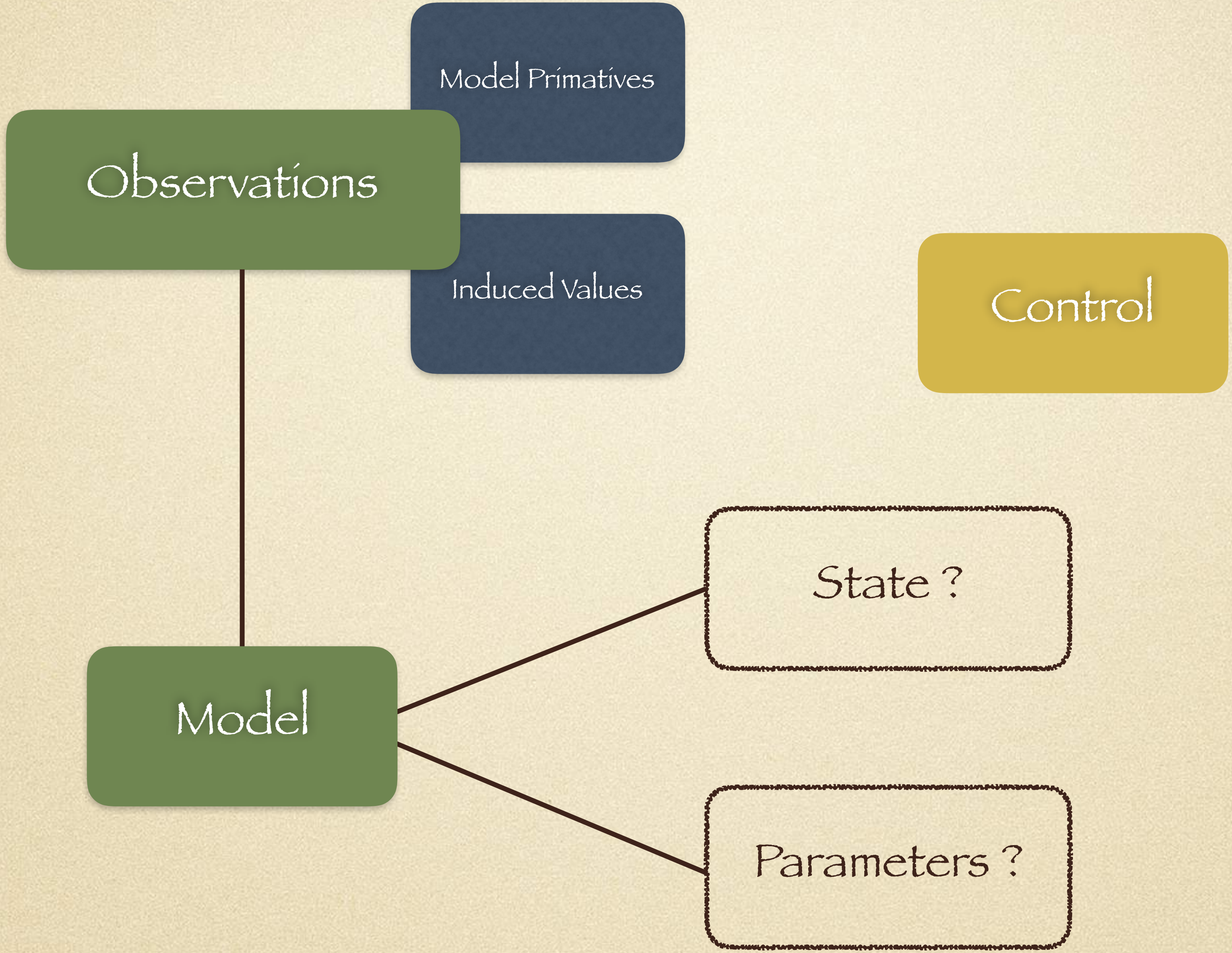
Azam Asanjarani, Yoni Nazarathy and Philip K. Pollett
The School of Mathematics and Physics, The University of Queensland,
Brisbane Australia

January 31, 2017

A Survey of Parameter and State Estimation in Queues

Azam Asanjarani, Yoni Nazarathy and Peter Taylor

June 8, 2018


Model Primatives
Observations
Induced Values
Control
State ?
Model
Parameters ?

Past Surveys

Chapter 13

Statistical analysis of queueing systems[1]

U. Narayan Bhat, Gregory K. Miller, and S. Subba Rao

ABSTRACT This paper provides an overview of the literature on the statistical analysis of queueing systems. Topics discussed include: model identification, parameter estimation using the maximum likelihood, method of moments and Bayesian frameworks, a discussion of covariance structure and autocorrelation in queueing systems, estimation from simulation experiments, hypothesis testing, and other related aspects. The bibliography, fairly exhaustive, should provide the reader with a source of articles that comprise the core of work done up to the present time.

CONTENTS

13.1	Introduction	351
13.2	Identification of models	352
13.3	Parameter estimation: The maximum likelihood method	354
13.4	Parameter estimation: The method of moments	360
13.5	Bayes estimation	361
13.6	Covariance structure, autocorrelation and mean value estimation	364
13.7	Estimation from special types of data	368
13.8	Estimation related to performance measures	372
13.9	Estimation from simulation of queueing processes	
13.10	Hypothesis testing	
13.11	Other related topics and future prospects	
	Bibliography	

Invited paper

STATISTICAL ANALYSIS OF QUEUEING SYSTEMS

U. NARAYAN BHAT

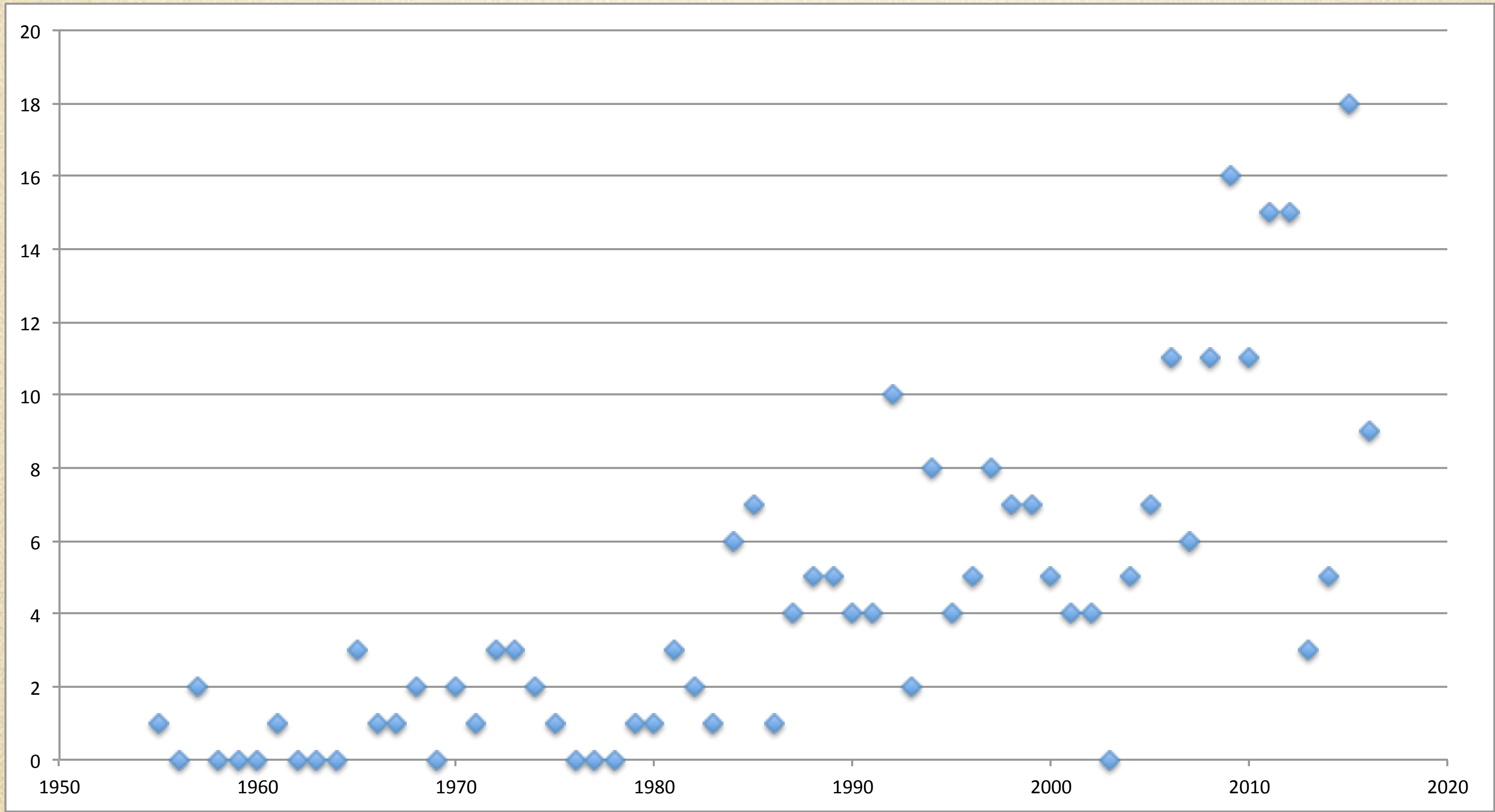
Department of Statistics, Southern Methodist University, Dallas, Texas 75275, USA

and

S. SUBBA RAO*

Department of Management and Systems, Washington State University, Pullman, Washington 99164, USA

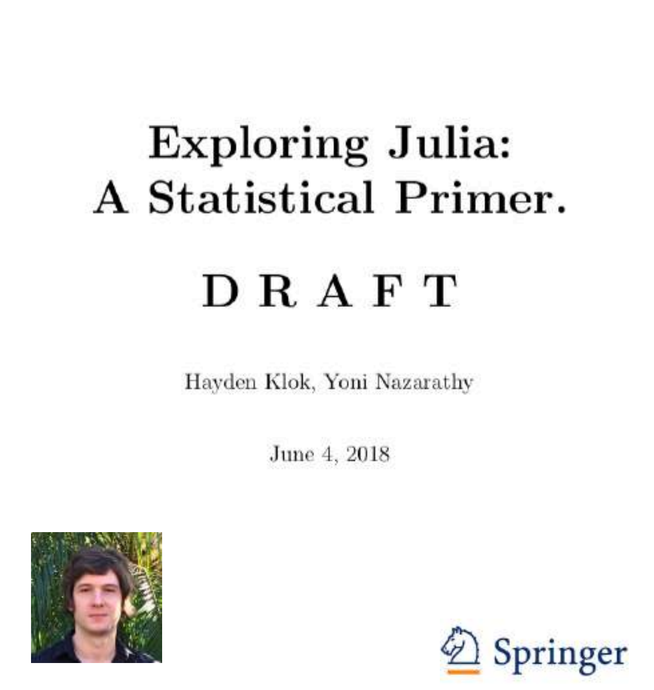
Number of Papers per year 1955 - 2016



Goal: Create a predictive model based on observations.

Setting for this talk: Single Pass Queue

Computation

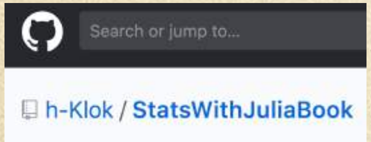

Exploring Julia:
A Statistical Primer.

DRAFT

Hayden Klok, Yoni Nazarathy

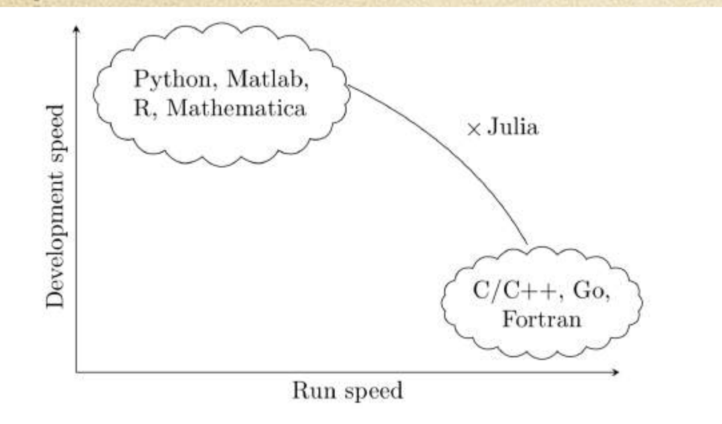
June 4, 2018

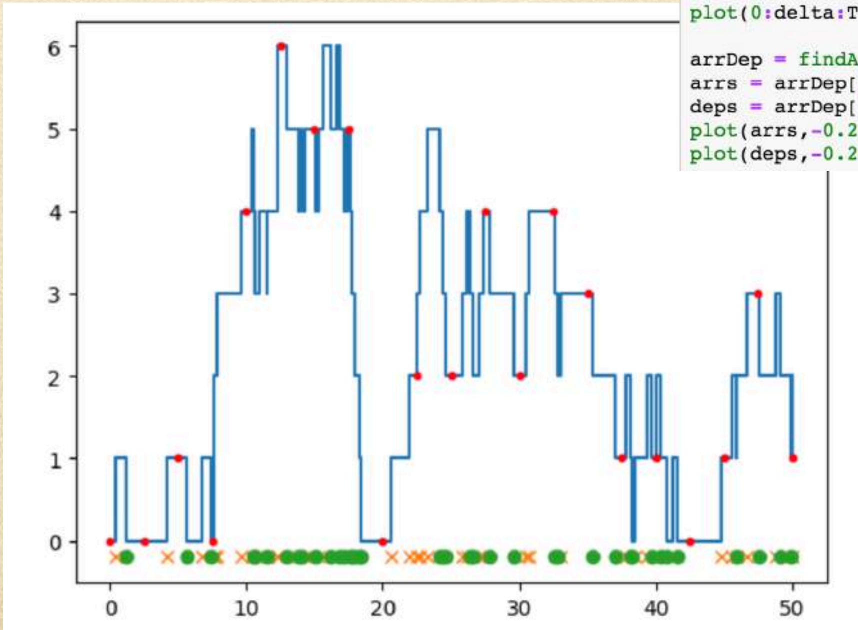
Springer



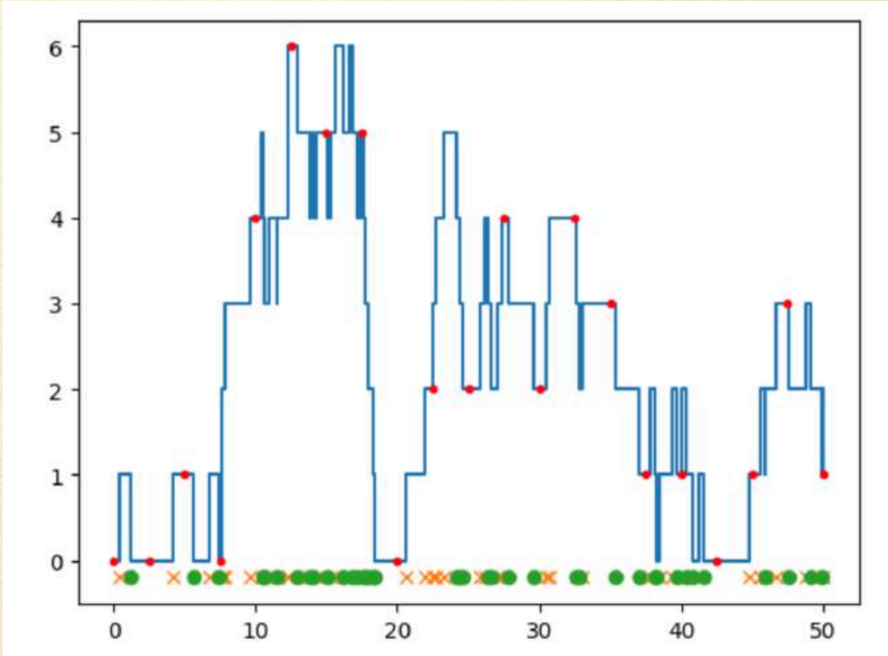

```julia
using DataStructures,Distributions

function simulateMM1(lambda,mu,Q0,T)
    t, Q = 0.0 , Q0
    tValues = [0.0]
    qValues = [Q0]
    while t<T
        if Q == 0 #arrival to an empty system
            t += rand(Exponential(1/lambda))
            Q = 1
        else  #change of state when system is not empty
            t += rand(Exponential(1/(lambda +mu)))
            Q += 2(rand() < lambda/(lambda+mu)) -1
        end
        push!(tValues,t)
        push!(qValues,Q)
    end
    return[tValues, qValues]
end
```

```julia
using PyPlot
T = 50
Q0 = 0
queueTraj = simulateMM1(0.9,1.0,Q0,T);
times = queueTraj[1]
qValues = queueTraj[2]
temp = stichSteps(times,qValues)
timesForPlot = temp[1]
qForPlot = temp[2]

delta = 2.5
qSampled = sampleQ(times,qValues,delta)
plot(timesForPlot,qForPlot)
plot(0:delta:T,qSampled,".",color="r")

arrDep = findArrDep(times,qValues)
arrs = arrDep[1]
deps = arrDep[2]
plot(arrs,-0.2*ones(length(arrs)),"x")
plot(deps,-0.2*ones(length(deps)),"o");
```





Six observation schemes



1. Full Observation

2. Discrete Intervals

3.Input and Output Process

4. Transactional Observations

5. Probing

6. Independent Primitives

1: Full Observation

$$\{Q(t) : t \in [0,T]\}$$

Some Problems of Statistical Analysis Connected with Congestion

D. R. COX, *Birkbeck College, University of London*

SUMM

Problem
previous worl
simpler new p
ing models is
information a
and b) service
count of prob

MAXIMUM LIKELIHOOD ESTIMATES IN A SIMPLE QUEUE

BY A. BRUCE CLARKE[1]

University of Michigan

0. Summary. The pro
the parameters involved
process is considered. A r
fies this problem to that
A useful and even simpl

1. Introduction. By a si
a negative exponential s

Received November 19,
[1] Research under contra
University.

Queueing Systems, 3 (1988) 289–304

LARGE SAMPLE INFERENCE FROM SINGLE SERVER QUEUES *

I.V. BASAWA
Department of Statistics, University of Georgia, Athens, GA 30602, U.S.A.

and

N.U. PRABHU
Mathematical Sciences Institute, Cornell University, Caldwell Hall, Ithaca, NY 14853, U.S.A.

Received: 16 April 1987
Revised: 17 March 1988

MAXIMUM LIKELIHOOD ESTIMATES IN A SIMPLE QUEUE

BY A. BRUCE CLARKE[1]

Sample until "busy time" reaches a pre-assigned value

ν - Initial queue size

m - Total departures

T - Time of last departure

n - Total arrivals

$$L(\lambda, \mu \;;\; \text{data}) = \left(1 - \frac{\lambda}{\mu}\right) e^{-\mu\tau - \lambda T} \mu^{m-\nu} \lambda^{n+\nu} K$$

Queueing Systems, 3 (1988) 289–304 289

LARGE SAMPLE INFERENCE FROM SINGLE SERVER QUEUES *

I.V. BASAWA
Department of Statistics, University of Georgia, Athens, GA 30602, U.S.A.

and

N.U. PRABHU
Mathematical Sciences Institute, Cornell University, Caldwell Hall, Ithaca, NY 14853, U.S.A.

Received: 16 April 1987
Revised: 17 March 1988

Rule 1: Observe the system until a fixed time t.
Rule 2: Observe the system until d departures.
Rule 3: Observe the system until m arrivals.
Rule 4: Stop after n transitions.

$$L_T(\theta, \phi) = \prod_{k=1}^{A(T)} f(u_k; \theta) \prod_{k=1}^{D(T)} g(v_k; \phi) \times \bar{F}_\theta(T - \sum_{k=1}^{A(T)} u_k)\bar{G}_\phi(T - \sum_{k=1}^{D(T)} v_k)$$

2: Discrete Intervals

$$\{Q(n\Delta), n = 1, \ldots, N\}$$

Statistical inference for discretely observed Markov jump processes.

Mogens Bladt
IIMAS–UNAM
A.P. 20-726
01000 Mexico, D.F.
Mexico
bladt@sigma.iimas.unam.mx

Michael Sørensen
Department of Applied Mathematics and Statistics
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen Ø
Denmark
michael@math.ku.dk

Estimation for queues from queue length data

J. V. Ross · T. Taimre · P. K. Pollett

Statistical inference for discretely observed Markov jump processes.

Mogens Bladt
IIMAS–UNAM
A.P. 20-726
01000 Mexico, D.F.
Mexico
bladt@sigma.iimas.unam.mx

Michael Sørensen
Department of Applied Mathematics and Statistics
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen Ø
Denmark
michael@math.ku.dk

$$L_\tau^{(c)}(Q) = \prod_{i=1}^{m} \prod_{j \neq i} q_{ij}^{N_{ij}(\tau)} e^{-q_{ij} R_i(\tau)}$$

An EM algorithm:

E-step, Calculate: $g : Q \to E_{Q_0}(\log L_\tau^{(c)}(Q) \mid Y = y)$

M-step, Calculate: $Q_0 = \mathrm{argmax}_Q\, g(Q)$

M/M/c

Let c grow…

Estimation for queues from queue length data

J. V. Ross · T. Taimre · P. K. Pollett

$$\lambda \sim \alpha c \qquad \frac{\lambda}{\mu c} \to x_0 < 1$$

Get an Ornstein-Uhlenbeck process…

$$E[Q(t) \mid Q(0)] \approx x_0 + e^{-\mu t}(Q(0) - x_0)$$

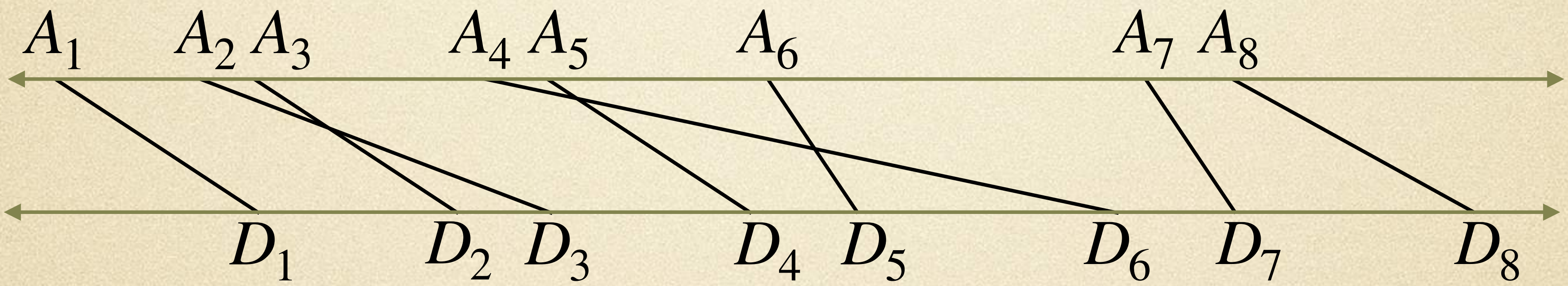
$$var\big(Q(t) \mid Q(0)\big) \approx \frac{\lambda}{\mu c^2}(1 - e^{-2\mu t})$$

$$cov(Q(s), Q(s+t)) \approx \frac{\rho}{c^2} e^{-\mu|t|}$$

Strategy: MLE for the OU process…

3: Input and Output Processes

$$\{A_1, \ldots, A_N\}, \{D_1, \ldots, D_N\}$$



Most Attention Given to M/G/∞

Why do we like Infinite Servers?

$$c(t) = Cov\big(Q(0), Q(t)\big) = \begin{cases} \dfrac{2\lambda(1-\lambda)}{\pi}\displaystyle\int_0^{\pi} \dfrac{(\sin\theta)^2 e^{-t(1+\lambda-2\sqrt{\lambda}\cos\theta)}}{(1+\lambda-2\sqrt{\lambda}\cos\theta)^3}\,d\theta & \text{for M/M/1,} \\ \dfrac{\lambda}{\mu^2}e^{-\mu t} & \text{for M/M/}\infty\,. \end{cases}$$

The Annals of Mathematical Statistics
1970, Vol. 41, No. 2, 651–654

AN $M/G/\infty$ ESTIMATION PROBLEM[1]

BY MARK BROWN
Cornell University

CDF of the time since the last arrival at a departure point: $H(x) = 1 - \left(1 - G(x)\right)e^{-\lambda x}$

J. Appl. Prob. **50,** 1044–1056 (2013)
Printed in England

SOJOURN TIME ESTIMATION IN AN M/G/∞ QUEUE WITH PARTIAL INFORMATION

NAFNA BLANGHAPS,*
YUVAL NOV * ** AND
GIDEON WEISS,* *University of Haifa*

Nonparametric fun

Ann. Inst. Statist. Math.
Vol. 51, No. 1, 71–97 (1999)

NON-PARAMETRIC ESTIMATION FOR THE $M/G/\infty$ QUEUE

N. H. BINGHAM[1] AND SUSAN M. PITTS[2]

[1] *Departme*

[2] *Department*

J. R. Statist. Soc. B (2004)
66, *Part* 4, *pp.* 861–875

Nonparametric inference about service time distribution from indirect measurements

Peter Hall

Australian National University, Canberra, Australia

and Juhyun Park

U

Adv. Appl. Prob. **48,** 1117–1138 (2016)
*doi:*10.1017/apr.2016.67

NONPARAMETRIC ESTIMATION OF THE SERVICE TIME DISTRIBUTION IN THE M/G/∞ QUEUE

ALEXANDER GOLDENSHLUGER,* *University of Haifa*

Bernoulli **24**(4A), 2018, 2531–2568
https://doi.org/10.3150/17-BEJ936

The $M/G/\infty$ estimation problem revisited

ALEXANDER GOLDENSHLUGER

Department of Statistics, University of Haifa, Haifa 31905, Israel. E-mail: goldensh@stat.haifa.ac.il

Output process only (of M/G/1):

Statistics and Computing **14**: 261–266, 2004

Filtering recursions for calculating likelihoods for queues based on inter-departure time data

PAUL FEARNHEAD

Department of Mathematics and Statistics, Lancaster University, LA1 4YF
p.fearnhead@lancaster.ac.uk

4: Transactional Observations

$$\{(D_1, \ldots, D_N), B\}$$

MANAGEMENT SCIENCE
Vol. 36, No. 5, May 1990
Printed in U.S.A.

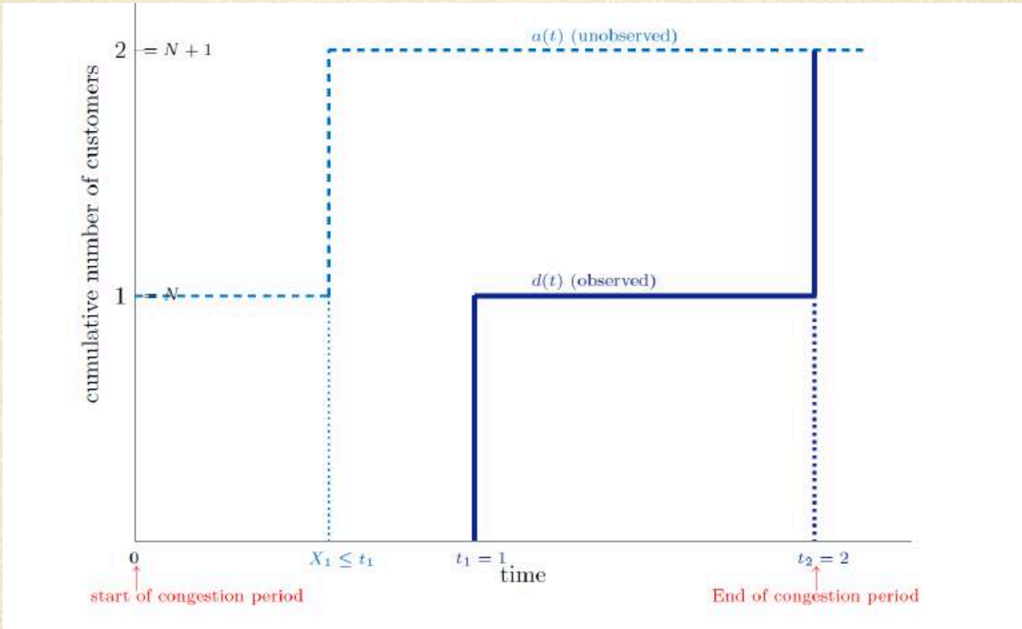
THE QUEUE INFERENCE ENGINE: DEDUCING QUEUE STATISTICS FROM TRANSACTIONAL DATA*

RICHARD C. LARSON

Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

The transactional data of a queueing system are the recorded times of service commencement and service completion for each customer served. With increasing use of computers to aid or even perform service one often has machine readable transactional data, but virtually no information about the queue itself. In this paper we propose a way to deduce the queueing behavior of Poisson arrival queueing systems from only the transactional data and the Poisson assumption. For each congestion period in which queues may form (in front of a single or multiple servers), the key quantities obtained are mean wait in queue, time-dependent mean number in queue, and probability distribution of the number in queue observed by a randomly arriving customer. The methodology builds on arguments of order statistics and usually requires a computer to evaluate a recursive function. The results are exact for a homogeneous Poisson arrival process (with unknown parameter) and approximately correct for a slowly time varying Poisson process. (QUEUES; INFERENCE; DATA ANALYSIS; POISSON)

Example: Consider a simple "transactional observation"



You observe:

- At time 0 server starts working
- At time 1 a customer departs (and server continues to work)
- At time 2 a second customer departs and server stops working

With Poisson arrivals:

$$E[n_Q(t)] = \begin{cases} t, & t \in [0,1), \\ 0, & t \in [1,2]. \end{cases}$$

Transactional data (QIE) fun

Operations Research Letters 38 (2010) 420–426

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

ELSEVIER

Queue inference from periodic reporting data

Jesse C. Frey[a,*], Edward H. Kaplan[b,c,d]

Inferring most likely queue length from transactional data

Dragomir D. Dimitrijevic*

Telogy Networks, Inc., 20250 Century Boulevard, German town, MD 20874, USA

Received 1 September 1992; revised 1 April 1996

Estimating characteristics of queueing networks using transactional data

Avi Mandelbaum and Sergey Zeltyn

Faculty of Industrial Engineering and Management, Technion, Haifa 32000, Israel

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

ELSEVIER

Analysis of an unobservable queue using arrival and departure times ☆

Jinsoo Park[a], Yun B. Kim[a,*], Thomas R. Willemain[b]

[a] *Department of Systems Management Engineering, Sungkyunkwan University, Suwon, South Korea*

[b] *Department of Decision Science and Engineering Systems, Rensselaer Polytechnic Institute, Troy, NY, USA*

DEDUCING QUEUEING FROM TRANSACTIONAL DATA: THE QUEUE INFERENCE ENGINE, REVISITED

DIMITRIS J. BERTSIMAS

Massachusetts Institute of Technology, Cambridge, Massachusetts

L. D. SERVI

GTE Laboratories, Incorporated, Waltham, Massachusetts

(Received April 1990; revision received February 1991; accepted July 1991)

Contents lists available at SciVerse ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

ELSEVIER

Remarks on queue inference from departure data alone and the importance of the queue inference engine

Lee K. Jones

Department of Mathematical Sciences, University of Massachusetts Lowell, 1 University Ave., Lowell MA 01854, United States

MOMENT ESTIMATION OF CUSTOMER LOSS RATES FROM TRANSACTIONAL DATA

D.J. DALEY[1]

Australian National University

School of Mathematical Sciences

Canberra ACT 0200 Australia

L.D. SERVI

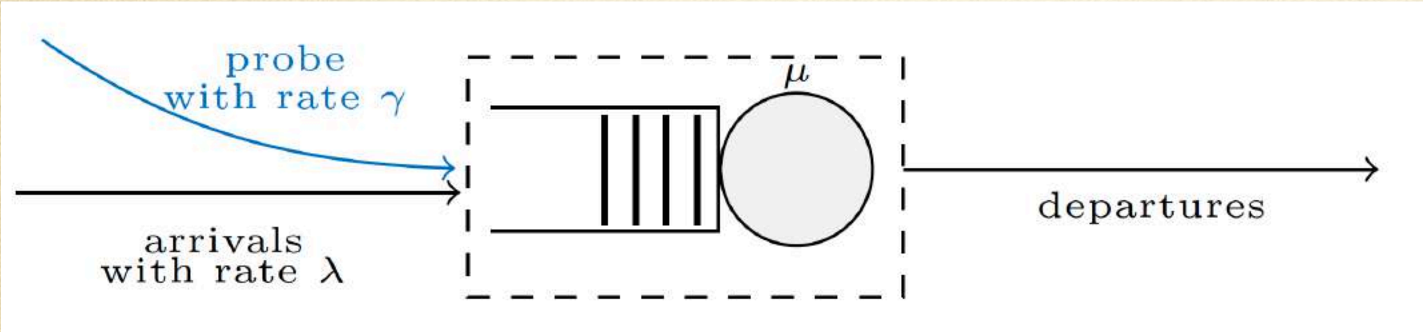
GTE Laboratories Incorporated

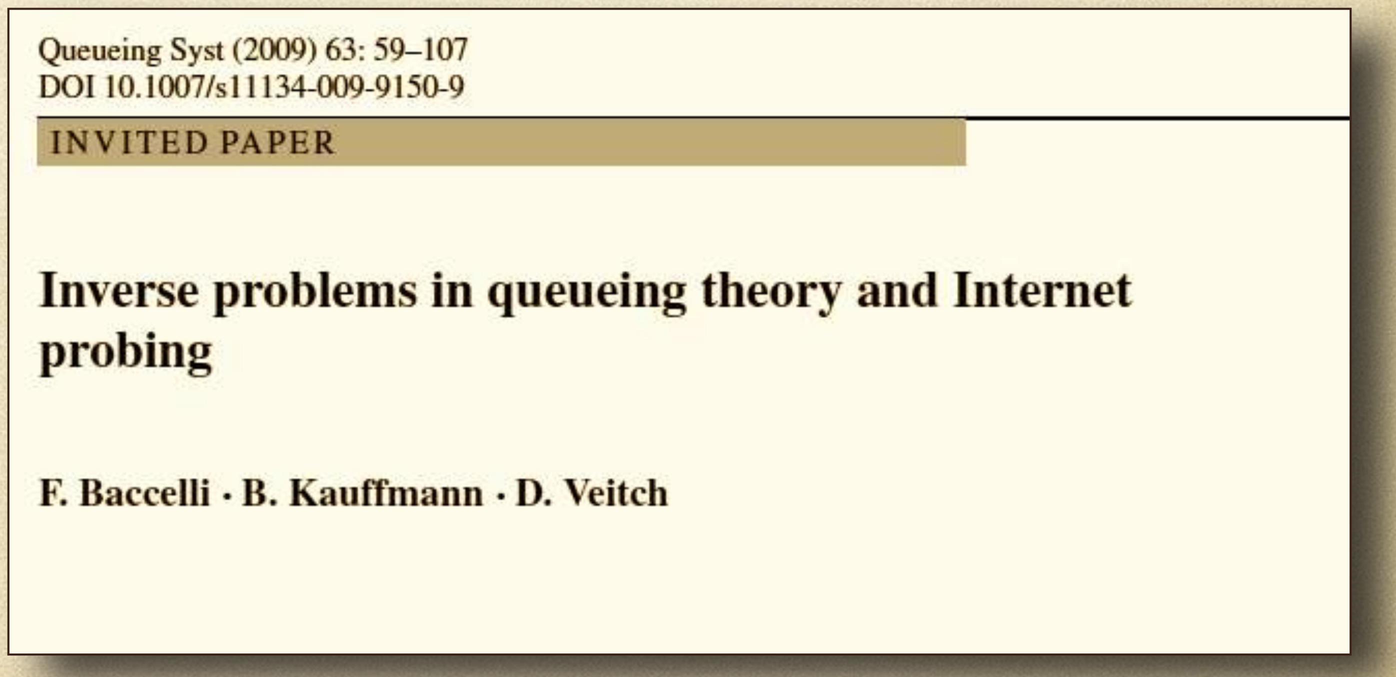
40 Sylvan Road

Waltham, MA 02254 USA

5: Probing

$$\{S_1^{\text{probe}}, \ldots, S_N^{\text{probe}}\}$$




Queueing Syst (2009) 63: 59–107
DOI 10.1007/s11134-009-9150-9

INVITED PAPER

Inverse problems in queueing theory and Internet probing

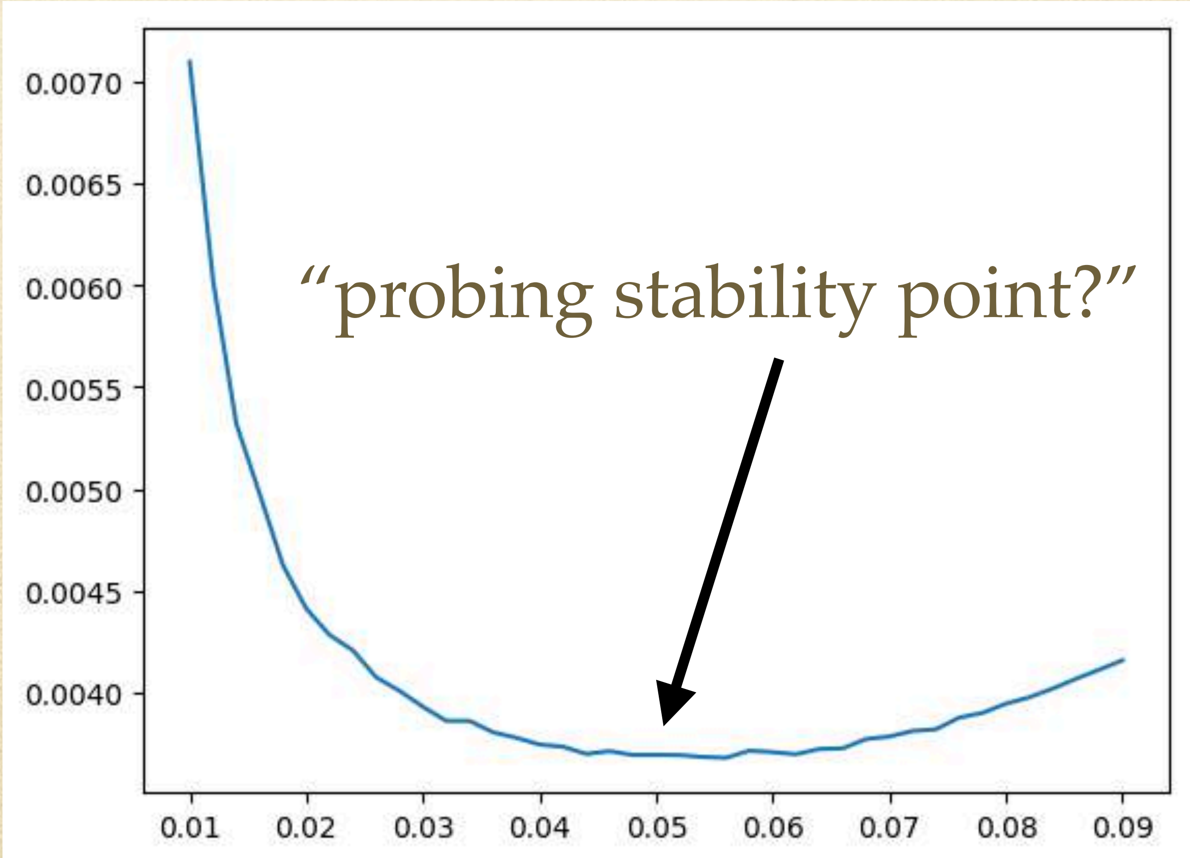
F. Baccelli · B. Kauffmann · D. Veitch

Best probe rate?

$$\hat{\lambda} = \mu - \lambda_{\text{probes}} - \frac{1}{\text{mean probe sojourn}}$$

E.g. T=1000 $\quad \lambda = 0.9 \quad \mu = 1$

MSE



λ_{probes}

Simulated 10^5 times
for each point on the grid

```julia
#This function returns a sequence of sojourn times of an M/M/1 queue
#that starts empty
#gamma is arrival of probes
#alpha is service rate

flip(p) = (rand() < p)

function simulateMM1SojournWithProbes(gamma,alpha,lambda,mu,T)

    pProbe = gamma/(gamma+lambda)

    tNextArr = rand(Exponential(1/(lambda+gamma)))
    tNextDep = Inf
    t = tNextArr

    waitingRoom = Queue(Tuple{Float64,Bool})
    arrTimeOfCustomerInService = NaN
    isCustomerInServiceProbe = false
    sojournTimes = Array{Float64}([])

    initQueue = rand(Geometric(lambda/mu))
    for i in 1:initQueue
        enqueue!(waitingRoom,(NaN,false))
    end

    while t<T
        if t == tNextArr
            if isnan(arrTimeOfCustomerInService)
                isCustomerInServiceProbe = flip(pProbe)
                tNextDep = t + rand(Exponential(1/(isCustomerInServiceProbe ? alpha : mu )))
                arrTimeOfCustomerInService = t
            else
                enqueue!(waitingRoom,(t,flip(pProbe)))
            end
            tNextArr = t + rand(Exponential(1/(lambda+gamma)))
        else #departure
            if length(waitingRoom) == 0
               tNextDep = Inf
               if isCustomerInServiceProbe
                   push!(sojournTimes, t-arrTimeOfCustomerInService)
               end
               arrTimeOfCustomerInService = NaN
            else
               if isCustomerInServiceProbe
                   push!(sojournTimes, t-arrTimeOfCustomerInService)
               end
               arrTimeOfCustomerInService,isCustomerInServiceProbe = dequeue!(waitingRoom)
               tNextDep = t + rand(Exponential(1/(isCustomerInServiceProbe ? alpha : mu)))

            end
        end
        t = min(tNextArr,tNextDep)
    end

    if length(sojournTimes) == 0
        sojournTimes = [1/mu] #QQQQ
    end
    return sojournTimes
end
```

Inferring Network Characteristics via Moment-Based Estimators

Sara Alouf, Philippe Nain

INRIA Sophia Antipolis, B.P. 93
06902, Sophia Antipolis Cedex, France
{salouf, nain}@sophia.inria.fr

Don Towsley

Dept. of Computer Science
University of Massachusetts
Amherst, MA 01003, USA
towsley@cs.umass.edu

The M+M/M/1/K Queue

The M+M/D/1/K Queue

Straightforward inverse estimation formulas

Parameter Estimation for Partially Observed Queues

Thomas M. Chen, *Member, IEEE,* Jean Walrand, *Fellow, IEEE,* and David G. Messerschmitt, *Fellow, IEEE*

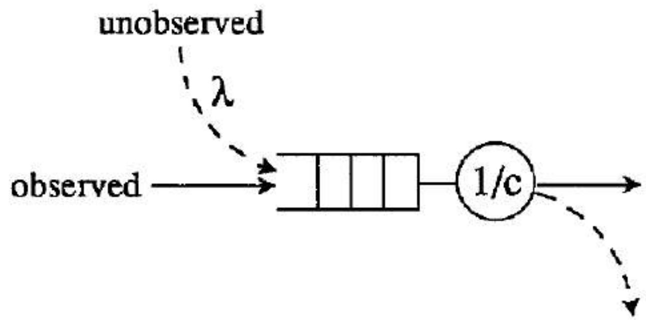


Fig. 2. Partially observed queue.

$$L_k(\lambda) = \prod_{n=2}^{k} p_n(\tau_n \mid \tau_{n-1}; \lambda)$$

$$F(x, t \mid x_0; \lambda) \equiv P(V(t) \leq x \mid V(0) = x_0)$$

$$p_n(x \mid y; \lambda) = f(x - c, a_n - a_{n-1} \mid y; \lambda)$$

6: Independent Primitives

$$\{(A_1, \ldots, A_N), (S_1, \ldots, S_M)\}$$

Volume 1, Number 2 OPERATIONS RESEARCH LETTERS April 1982

SOME CONSEQUENCES OF ESTIMATING PARAMETERS FOR THE M/M/1 QUEUE *

Lee SCHRUBEN and Radhika KULKARNI

Technical Note: Traffic Intensity Estimation

Peter C. Kiessler, Robert Lund

Department of Mathematical Sciences, Clemson University, Clemson, South Carolina 29634-0975

Volume 1, Number 2 OPERATIONS RESEARCH LETTERS April 1982

SOME CONSEQUENCES OF ESTIMATING PARAMETERS FOR THE M/M/1 QUEUE *

Lee SCHRUBEN and Radhika KULKARNI

Data: $\{(A_1, \ldots, A_N), (S_1, \ldots, S_M)\}$

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{m} A_i} \qquad \hat{\mu} = \frac{m}{\sum_{i=1}^{m} S_i}$$

$$\hat{\rho} = \frac{\hat{\lambda}}{\hat{\mu}}$$

If $\hat{\rho} \geq 1$ then resample.

$$\frac{\hat{\rho}}{\rho} \sim F(2m, 2n)$$

Problem: $$E\left[\frac{\hat{\rho}}{1-\hat{\rho}} \mathbf{1}\{\hat{\rho} < 1\}\right] = \infty$$

Technical Note: Traffic Intensity Estimation

Peter C. Kiessler, Robert Lund

Department of Mathematical Sciences, Clemson University, Clemson, South Carolina 29634-0975

$$V(t) = V(0) + \sum_{n=1}^{A(t)} S_n - \int_0^t \mathbf{1}\{V(\tau) > 0\}\, d\tau$$

$$\hat{\rho}_{\text{virtual}} = \frac{1}{t} \int_0^t \mathbf{1}\{V(\tau) > 0\}\, d\tau$$

Bias bounding:

$$\left| E[\hat{\rho}_{\text{virtual}}] - \rho \right| \leq \frac{c + V(0)}{t}$$

Bayesian Celebration...

The Annals of Applied Statistics
2010, Vol. 4, No. 3, 1533–1557
DOI: 10.1214/10-AOAS336

BAYESIAN INFERENCE FOR DOUBLE PARETO LOGNORMAL QUEUES

BY PEPA RAMIREZ-COBO[1], ROSA E. LILLO, SIMON WILSON AND MICHAEL P. WIPER

CNRS France, Universidad Carlos III de Madrid, Trinity College Dublin and Universidad Carlos III de Madrid

In this article we describe a method for carrying out Bayesian estimation for the double Pareto lognormal (*dPlN*) distribution which has been proposed as a model for heavy-tailed phenomena. We apply our approach to estimate the *dPlN*/*M*/1 and *M*/*dPlN*/1 queueing systems. These systems cannot be analyzed using standard techniques due to
a Laplace tran
some recent ap
rival distributi
trated with ap

Queueing Systems 15 (1994) 419–426

Short communication

Bayesian inference in Markovian queues

Carmen Armero

Department of Statistics and Operations Research, University of Valencia, Doctor Moliner 50, 46100 Burjassot, Valencia, Spain

Received 25 January 1992; revised 19 January 1993

+ Many Other Papers

In summary:

1. Full Observation

2. Discrete Intervals

3.Input and Output Process

4. Transactional Observations

5. Probing

6. Independent Primitives