Fluctuation Scaling in Large Service Systems

Xiaowei Zhang

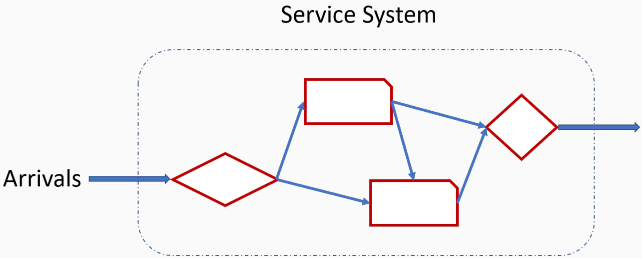
Qingdao, June 5, 2018

Joint work with L. Jeff Hong (CityUHK) and Jiheng Zhang (HKUST)

Outline

Introduction

Service Systems

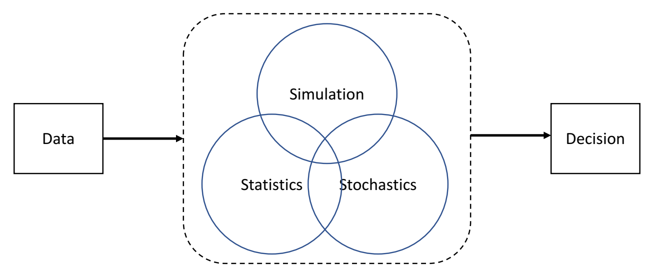


call centers, hospitals, inventory, ride-sharing, etc.

Optimizing Performance $\approx$ Managing Fluctuations

- Service operations, capacities, schedules: largely controllable
- Arrivals: exogenous, represent a primary source of uncertainty

Data-to-Decision



- Use data and statistical tools as black magic (✗)
- Modeling should be driven by both data and domain knowledge (✓)

Standard Approach to Modeling Arrivals

1. Collect arrival data
2. Compute inter-arrival times
3. Fit a probability distribution from popular families (Exp, Gamma, Weibull, etc.)
4. Perform goodness-of-fit test

- Poisson process, renewal process, and their time-varying extensions
- $M/\cdot/\cdot$
- $G/\cdot/\cdot$
- $M(t)/\cdot/\cdot$
- $G(t)/\cdot/\cdot$
- and so on...

Does it really matter (that much)?

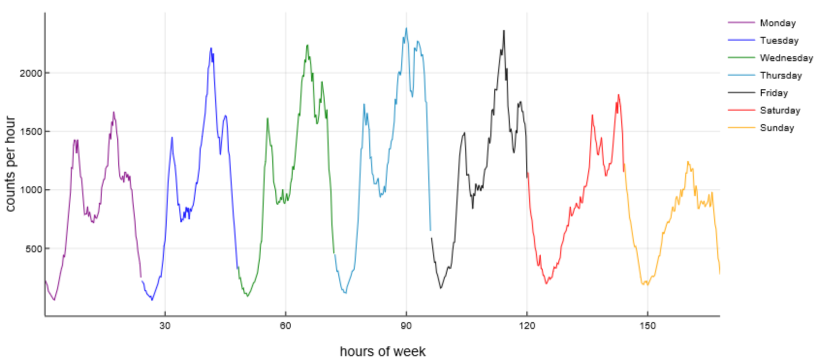
Data-Driven Modeling with Domain Knowledge

Spoiler

- Poisson microstructure: inter-arrival times are indeed exponential
- Microstructure does **NOT** matter much for typical service decisions

Arrivals to a Ride-sharing Platform

- Uber Pickups (NYC): daily volume 10,000∼40,000 in 2014

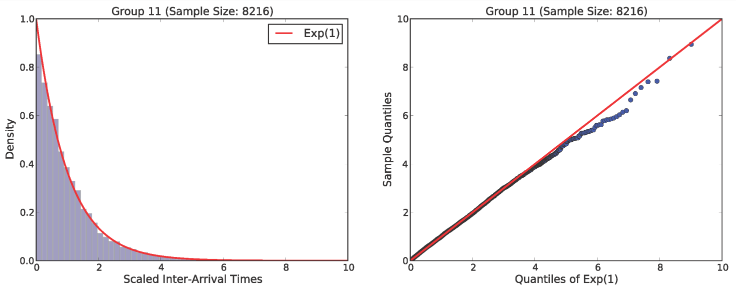


Hypothesis Test for Exponentiality of Inter-arrival Times

Lemma

If $N(t)$ is an inhomogeneous Poisson process with arrival rate $\lambda(t)$, then $N(\Lambda^{-1}(t))$ is a standard Poisson process, where $\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s$.

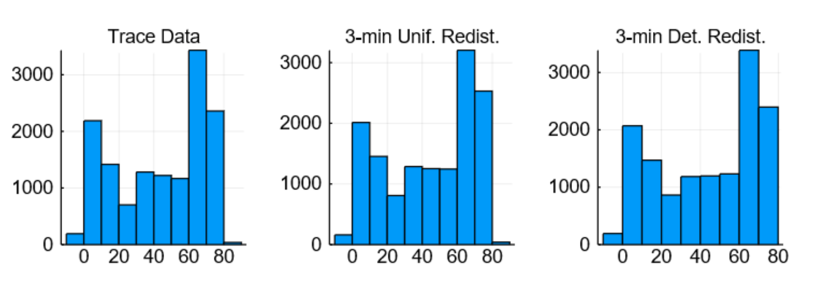
- **Null hypothesis**: arrivals follow an inhomogeneous Poisson process
- Under the null, the time-changed inter-arrival times are i.i.d. exponential
- See also Brown et al. (2005) and Kim and Whitt (2014)



Impact on Performance?

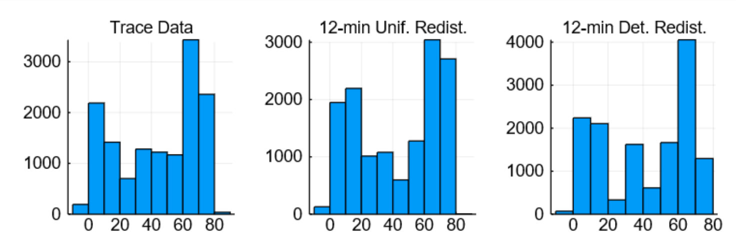
- Zheng, Zhang, and Glynn (2018)
- Run three different arrival sequences through the same service system
  1. Real arrival data
  2. Split the real arrivals into intervals of length x minutes; redistribute them randomly within each interval
  3. Split the real arrivals into intervals of length x minutes; redistribute them equally spaced within each interval
- Compare performance using synchronized service times

Distribution of Waiting Times ($x = 3$)



almost identical

Distribution of Waiting Times ($x = 12$)

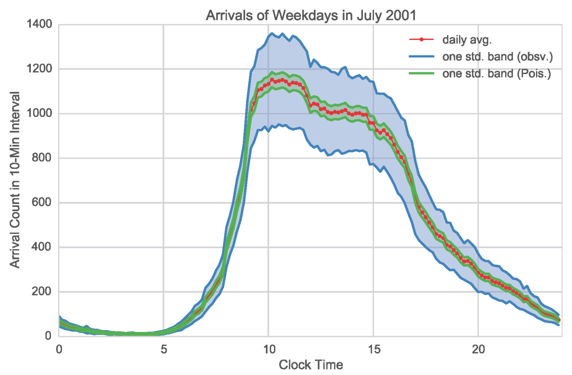


noticeably different

Look at the Bigger Picture

- Microstructure does not seem to have much impact on performance
- Should focus on the stochastic behavior over the time scale that is compatible with the service time and matters to decisions

Known Fact: Overdispersion



- Jongbloed and Koole (2001), Avramidis et al. (2004), Oreshkin et al. (2016)

Why Important?

- More uncertainty in demand, more requirement for supply
  - Newsvendor: if $D \sim \mathcal{N}(m, \sigma^2)$, then $Q^* = m + \beta\sigma$, where β represents the service level
  - Base stock policy under periodic review has a similar formula
- Square-root staffing rule for large service systems

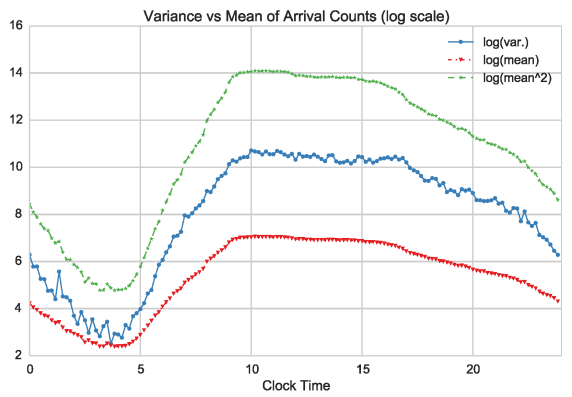
$$n \approx \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}}$$

- The term $\sqrt{\lambda}$ comes from
  - $\mathrm{Var}[A(t)] = \mathbb{E}[A(t)] = \lambda t$ (Poisson arrivals)
  - $\mathrm{Var}[A(t)] \sim \mathcal{O}(\lambda t)$ as $\lambda \to \infty$ (renewal arrivals)

How does fluctuation scale up?

Fluctuation Scaling

New Finding: Overdispersion is Amplified by Heavy Traffic

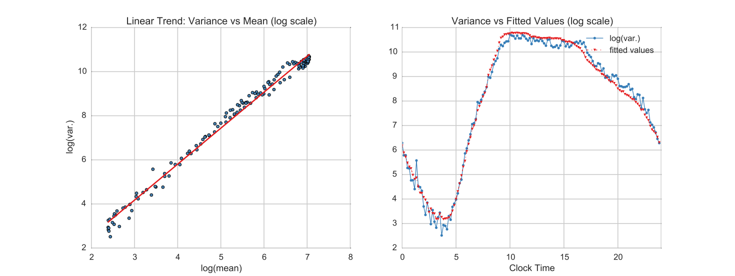


$$\log(\text{Mean}) < \log(\text{Var.}) < 2\log(\text{Mean})$$

Power Law Relationship

- Assume a linear relationship at the logarithmic scale

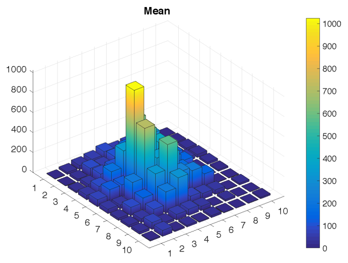
$$\log(\text{Var.}) = p \log(\text{Mean}) + c$$

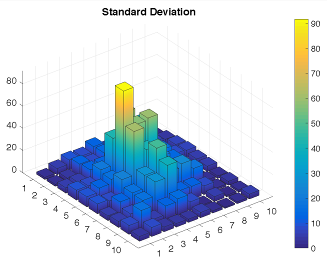


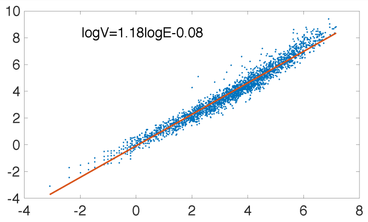
- $p \approx 1.6$ and $R^2 = 0.99$
- Taylor's law in ecology (Taylor, 1960), also found in physics, finance, etc.
- Conjecture: change safety margin of the staffing rule

$$\mathcal{O}(\lambda^{1/2}) \longmapsto \mathcal{O}(\lambda^{p/2})$$

- Ride information of DiDi Chengdu in Nov. 2016: time and location
- Partition the city into 10×10 grid, partition one day into 24 hours







Desirable Features for Arrival Model

- Overdispersion
- Fluctuation scaling with power law
- Poisson microstructure
- Analytical tractability

Typical Treatment for Overdispersion

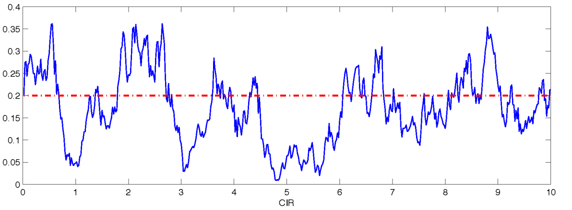
- Doubly stochastic Poisson process (DSPP): stochastic arrival rate
- Whitt (1999): $A(t) = N(\lambda G t)$
  - G is a random variable with $\mathbb{E}(G) = 1$, capturing day-to-day random variation, i.e. "busyness of the day"
- This model implies $\mathrm{Var.} \sim \mathcal{O}(\lambda^2)$

$$\mathrm{Var}(N(\lambda G t)) = \lambda t + \lambda^2 t^2 \mathrm{Var}(G)$$

  - Overestimate overdispersion: $p \in (0, 1)$
  - Lack of flexibility

- $A(t)$: a DSPP with arrival rate $X(t)$

$$\mathrm{d}X(t) = \kappa(\lambda - X(t))\,\mathrm{d}t + \sigma\lambda^{\alpha}\sqrt{X(t)}\,\mathrm{d}B(t)$$



- $X(t) > 0$
- Equilibrium of $X(t)$ is λ
- Volatility term: $\sigma\lambda^{\alpha}\sqrt{X(t)} \sim \mathcal{O}(\lambda^{\alpha+1/2})$
- So $\mathrm{Var}[A(t)] \sim \mathcal{O}(\lambda^{2\alpha+1})$: $p = 2\alpha + 1$

Limit Theorem

Theorem

Suppose $X(t)$ is initialized with its stationary distribution π and $\alpha \in (0, \frac{1}{2})$. Then, for any given $t > 0$,

$$\frac{A_\lambda(t) - \lambda t}{\lambda^{\alpha + \frac{1}{2}}} \Rightarrow \int_0^t U(s)\,\mathrm{d}s,$$

as $\lambda \to \infty$, where $U(t)$ is an Ornstein-Uhlenbeck (OU) process

$$\mathrm{d}U(t) = -\kappa U(t)\,\mathrm{d}t + \sigma \mathrm{d}B(t),$$

with initial distribution being its unique stationary distribution, i.e. normal distribution with mean 0 and variance $\frac{\sigma^2}{2\kappa}$.

- Critical for deriving the staffing rule

Staffing Rule

Service Quality v.s. Utilization

- Large queueing system: both λ and number of servers are large
- Service quality is measured by delay probability
- Goal: find minimum number of servers to make delay probability $\leq \epsilon$
- Key: distribution of queue length

Infinite-server Queue Approximation

- Consider an infinite-server queue with exponential service times
- Let $Q_\lambda(t)$ denote the number of customers in the system
- Consider the scaled number of customers

$$\tilde{Q}_\lambda = \frac{Q(t) - \lambda/\mu}{\lambda^{\alpha+\frac{1}{2}}}$$

- Show $\tilde{Q}_\lambda(t)$ converges a non-degenerate limit $\tilde{Q}_\infty(t)$ as $\lambda \to \infty$
- Compute the stationary distribution of $\tilde{Q}_\infty(t)$ as $t \to \infty$, denoted by $\tilde{Q}_\infty$
- n: number of servers in the many-server queue

$$\epsilon \approx \mathbb{P}(Q_\lambda(t) \geq n) = \mathbb{P}\left(\tilde{Q}_\lambda(t) \geq \frac{n - \lambda/\mu}{\lambda^{\alpha+\frac{1}{2}}}\right) \approx \mathbb{P}\left(\tilde{Q}_\infty \geq \frac{n - \lambda/\mu}{\lambda^{\alpha+\frac{1}{2}}}\right)$$

- Let β solves $\mathbb{P}(\tilde{Q}_\infty \geq \beta) = \epsilon$, then

$$n^* \approx \frac{\lambda}{\mu} + \beta \cdot \lambda^{\alpha+\frac{1}{2}}$$

- β can be computed explicitly

- By virtue of the exponential service assumption,

$$Q_\lambda(t) = Q_\lambda(0) + A_\lambda(t) - N'\left(\mu \int_0^t Q_\lambda(s)\,\mathrm{d}s\right),$$

where $N'(\cdot)$ is an independent Poisson process with unit rate

$$\begin{aligned} Q_\lambda(t) - \frac{\lambda}{\mu} &= Q_\lambda(0) - \frac{\lambda}{\mu} + A_\lambda(t) - \lambda t - \mu \int_0^t Q_\lambda(s)\mathrm{d}s - \lambda t \\ &\quad - N'\left(\mu \int_0^t Q_\lambda(s)\mathrm{d}s\right) - \mu \int_0^t Q_\lambda(s)\mathrm{d}s \end{aligned}$$

- Scaled by $\lambda^{\alpha+\frac{1}{2}}$ and letting $\lambda \to \infty$

$$\tilde{Q}_\infty(t) = \tilde{Q}_\infty(0) + \int_0^t U(s)\,\mathrm{d}s - \mu \int_0^t \tilde{Q}_\infty(s)\,\mathrm{d}s - 0$$

- Solve the equation

$$\tilde{Q}_\infty(t) = \int_0^t U(s) e^{\mu(t-s)} \, \mathrm{d}s$$

- The Laplace transform of $\tilde{Q}_\infty(t)$ can be computed analytically
- Sending $t \to \infty$ yields the Laplace transform of $\tilde{Q}_\infty$
  - normal distribution
  - parameters can be calculated analytically
- Easy to compute β in the staffing rule

$$n \approx \frac{\lambda}{\mu} + \beta \cdot \lambda^{\alpha + \frac{1}{2}}$$

by $\mathbb{P}(\tilde{Q}_\infty \geq \beta) = \epsilon$

Performance in Practice

- Call center of a U.S. bank
- Use customer arrivals in weekdays in 2002
- A constant staffing level for each 30-min time period
  - 48 staffing levels for a day in total
  - *Static* staffing: no day-to-day adjusting
- Simulate the system with real customer arrivals and exponential service

Target Quality of Service	Our Staffing Rule	Square-root Staffing Rule
0.20	0.230	0.537
0.10	0.137	0.469
0.05	0.084	0.439

Concluding Remarks

Conclusions

- Modeling should be driven by both data and decisions
    - Arrivals' microstructure is Poisson but has little impact on performance
    - Stochastic behavior at longer time scale matters, e.g., overdispersion
- Overdispersion is amplified by heavy traffic
    - Fluctuation scales up following a power law
- Proposed New tractable arrival model to capture the power law
- Developed the associated staffing rule with safety margin $\mathcal{O}(\lambda^{\alpha+1/2})$

Questions?

References

A. N. Avramidis, A. Deslauriers, and P. L'Ecuyer. Modeling daily arrivals to a telephone call center. *Manag. Sci.*, 50(7):896–908, 2004.

L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.*, 100(1):36–50, 2005.

G. Jongbloed and G. Koole. Managing uncertainty in call centers using Poisson mixtures. *Appl. Stoch. Model. Bus. Ind.*, 17:307–318, 2001.

S.-H. Kim and W. Whitt. Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manuf. Serv. Oper. Manag.*, 16(3): 464–480, 2014.

B. N. Oreshkin, N. Régnard, and P. L'Ecuyer. Rate-based daily arrival process models with application to call centers. *Oper. Res.*, 64(2):510–527, 2016.

L. R. Taylor. Aggregation, variance and the mean. *Nature*, 189(4766):732–735, 1960.

W. Whitt. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Oper. Res. Lett.*, 24:205–212, 1999.