Random matrices, phase transitions & queuing theory

Raj Rao Nadakuditi

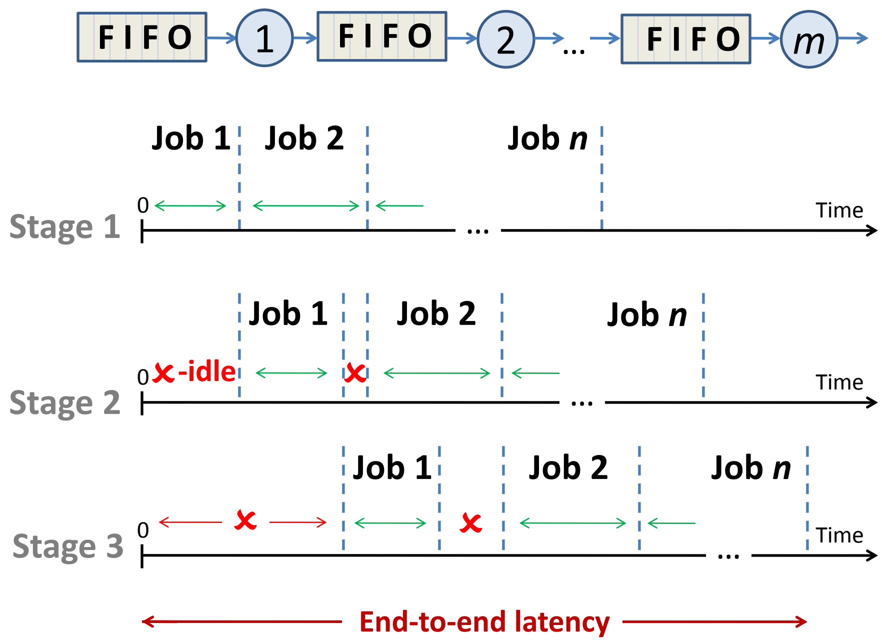
Dept. of Electrical Engg. & Computer Science

`http://www.eecs.umich.edu/~rajnrao`

Joint works with Jinho Baik, Igor Markov and Mingyan Liu

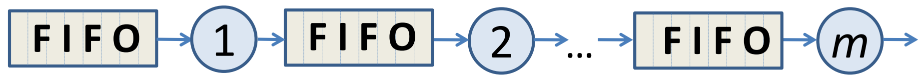
Queuing theory

A fundamental problem in queuing theory



- $m = \#$ servers, $n = \#$ customers (or jobs)
- Objective: Characterize $L(m, n) =$ exit time for n-th customer from m-th queue
  - Model for production systems, multi-hop networks, pipelined computation

Why is characterizing latency important?

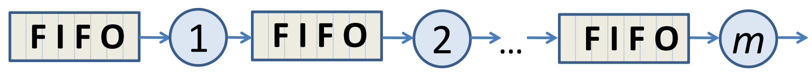


- Many existing applications are delay-sensitve
  - Production systems, Streaming audio and video - particularly audio
  - $\Rightarrow$ Optimal scheduling/provisioning $\Leftrightarrow$ delay-throughput tradeoff

- Emerging applications envision control and inference over large networks
  - Telemedicine, sensor networks and distributed computation
  - $\Rightarrow$ Quality of Service (QoS) guarantees important

- Network topology design
  - Ad-hoc, multi-hop networks prevalent (e.g. deliver interet to rural areas)
  - Optimal placement of hops? Remote diagnosis of service bottlenecks?
  - $\Rightarrow$ Statistical characterization of delay important

A basic model



Notation:

- S_i = Server $i \in \{1, \ldots, m\}$
- C_j = Customer $j \in \{1, \ldots, n\}$
- $w(i, j)$ = Service time for C_j at S_i

Assumptions:

- Infinitely long buffer
- Arrival process is Poissonian with rate α
- $w(i, j) \overset{ind.}{\sim} \exp(1/\mu_i) \Leftrightarrow$ M/M/m queue

Question: Average Delay?

Little's Law and average delay

Informally:

$$\boxed{\text{Avg. Time in System} = \frac{\text{Avg. \# of Cust.}}{\text{Eff. Arrival Rate}}}$$

By Burke's Theorem:

$$\mathbb{P}(\#\text{Cust. in Queue } i = k) = \left(1 - \frac{\mu_i}{\alpha}\right)^k \left(\frac{\mu_i}{\alpha}\right) \text{ for } k = 0, 1, \ldots$$

Consequently:

$$\boxed{\Rightarrow \text{Avg.\# Cust. in System} = \sum_{i=1}^{m} \frac{\mu_i}{\mu_i - \alpha_i}}$$

What Little's law says and does not say

$$\boxed{\text{Avg.\# Time in System} = \frac{1}{\alpha}\sum_{i=1}^{m}\frac{\mu_i}{\mu_i - \alpha_i}}$$

Mathematically:

$$\boxed{\text{Avg.\# Time in System} = \lim_{t\to\infty}\frac{\sum_{i=0}^{\alpha(t)} \text{Time spent by Customer } i}{\alpha(t)}}$$

- $\alpha(t) = \#$ Customers who arrived in the interval $[0, t]$
- No insights on: variance, pdf, bottleneck behavior, etc.
- Contrast with $L(m, n) =$ exit time for Customer n from Server m
  - Transient-like statistic! Computable?

What Little's law says and does not say

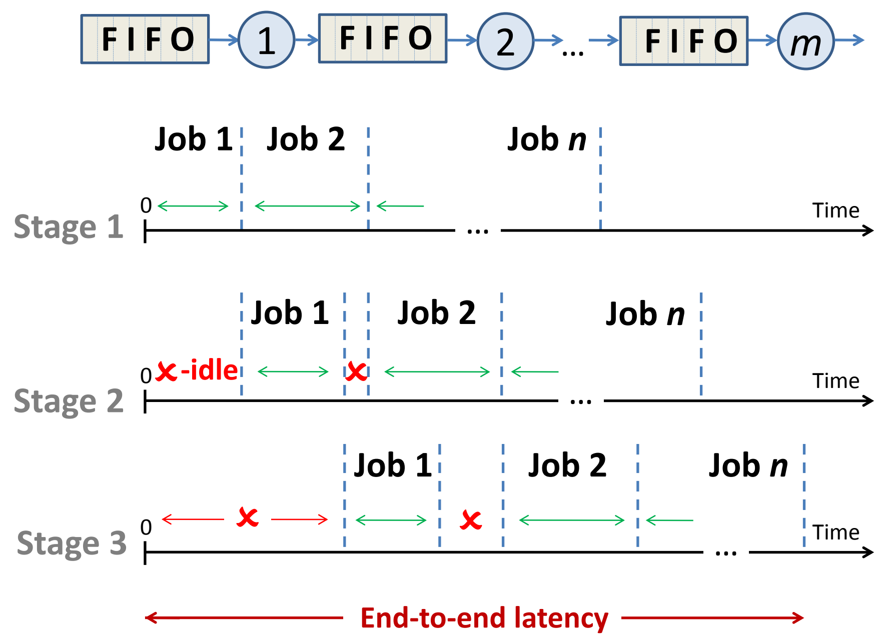
$$\boxed{\text{Avg.\# Time in System} = \frac{1}{\alpha}\sum_{i=1}^{m}\frac{\mu_i}{\mu_i-\alpha_i}}$$

Mathematically:

$$\boxed{\text{Avg.\# Time in System} = \lim_{t\to\infty}\frac{\sum_{i=0}^{\alpha(t)}\text{Time spent by Customer } i}{\alpha(t)}}$$

- $\alpha(t) = \#$ Customers who arrived in the interval $[0, t]$
- No insights on: variance, pdf, bottleneck behavior, etc.
- Contrast with $L(m, n) =$ exit time for Customer n from Server m
  - Transient-like statistic! Computable?
  - **Yes!** - Using random matrix theory!

What makes the problem difficult?



- $m = \#$ servers, $n = \#$ customers (or jobs)
- Objective: Characterize $L(m, n)$ = exit time for n-th customer from m-th queue
  - **Strong** interaction between arrival and departure process $\Rightarrow$ no independence

Main message

New insights beyond Little's Law:

- Latency mean and variance can be explicitly computed! ✓
- Analysis reveals emergence of **phase transitions** ✓
- **Rigorous** basis for statistical anomaly testing ✓
- Can show that $O(n^{1/3})$ jobs have statistically independent latencies ✓
- Extends easily to quasi-reversible networks (thanks Demos!) ✓
- Analysis of queue-state dependent servicing (inspired by backpressure algorithms) ✓
- Results appear to hold even for non-exponential service times ✓
  - Universality conjecture!

All made possible due to connection with random matrix theory!

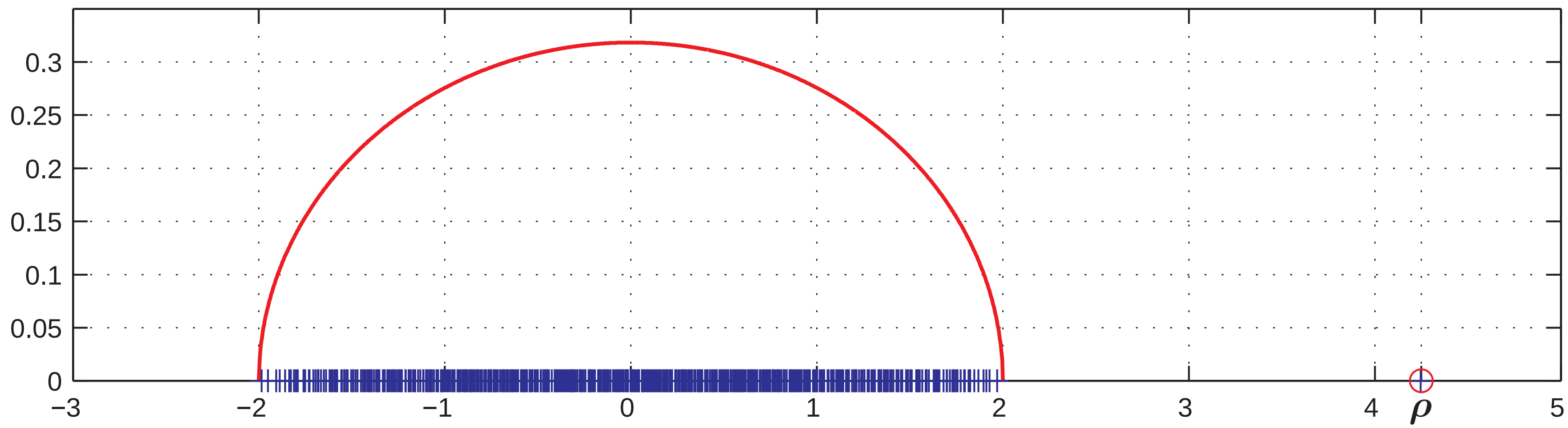
Phase transitions

A numerical example

- G = Gaussian random matrix
  - `G = randn(n,n)` or `G = sign(randn(n,n))`
- $X_n = \dfrac{G + G'}{\sqrt{2n}}$
- $\widetilde{X}_n = X_n + P_n$
  - $P_n = \theta\, u\, u'$
  - u is a fixed, non-random unit norm vector
  - X_n has i.i.d. zero mean, variance $1/2n$ entries (on off-diagonal)

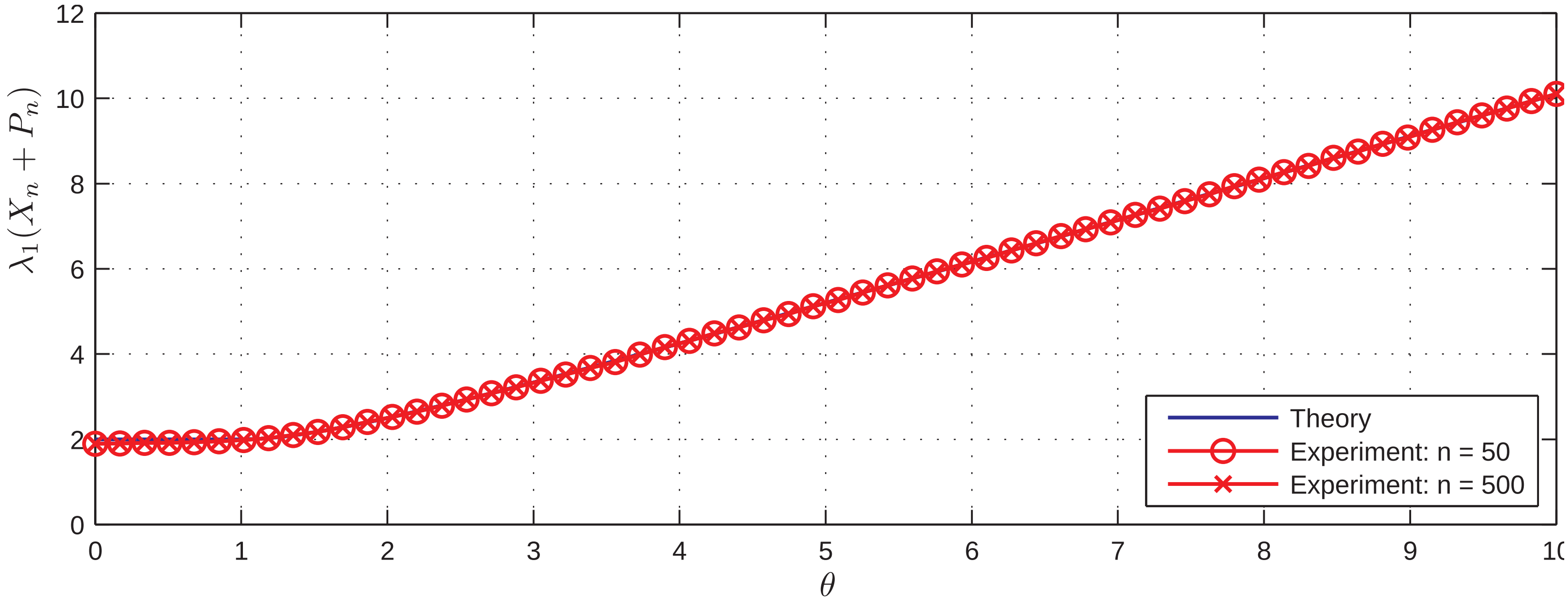
<u>Question</u>: Largest eigenvalue? Variation with θ?

One experimental realization



- $\theta = 4$, $n = 500$
- Bulk obeys semi-circle law on $[-2, 2]$
- Largest eig. ≈ 4.2

An eigenvalue phase transition



- Clear phase transition @ $\theta = 1$ with increasing n

Phase transition prediction

Theorem: Consider $\widetilde{X}_n = X_n + \theta u u'$

$$\widetilde{\lambda}_1 \xrightarrow{\text{a.s.}} \begin{cases} \theta + \dfrac{1}{\theta}, & \theta > 1 \\ 2, & \text{otherwise} \end{cases}$$

$$|\langle \widetilde{u}_1, u \rangle|^2 \xrightarrow{\text{a.s.}} \begin{cases} \left(1 - \dfrac{1}{\theta^2}\right), & \theta > 1 \\ 0, & \text{otherwise} \end{cases}$$

- Eigenvalue result first due to Peche (2006), Peche-Feral (2007)
- Eigenvector result new (and derived by us)
- Eigenvalues and eigenvectors are biased

Phase transitions & Random matrix theory

or

What theory predicts the phase transition?

Definitions and assumptions

Spectral measure: Eigenvalues of X_n are $\lambda_1, \ldots, \lambda_n$:

$$\mu_{X_n} = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i}$$

Assumptions:

1. $\mu_{X_n} \xrightarrow{\text{a.s.}} \mu_X$
2. $\operatorname{supp} \mu_X$ compactly supported on $[a, b]$
3. $\max$(eig) $\xrightarrow{\text{a.s.}}$ to b

A basic signal-plus-noise model

$$\boxed{\widetilde{X}_n = \theta u u^H + X_n}$$

Assumptions:

- X_n is symmetric with n real eigenvalues
- $\theta_1 > \ldots > \theta_k > 0$
- $X_n = Q\Lambda Q'$ where Q is a Haar distributed unitary (or orthogonal) matrix
- u is a unit-norm vector
- $X_n = GG^*$ will satisfy conditions

Phase transition in the eigenvalues

Theorem [Benaych-Georges and N.]: As $n \longrightarrow \infty$,

$$\boxed{\lambda_1(\widetilde{X}_n) \xrightarrow{\text{a.s.}} \begin{cases} G_\mu^{-1}(1/\theta_i) & \text{if } \theta > \theta_c := 1/G_\mu(b^+), \\ b & \text{otherwise,} \end{cases}}$$

- Critical threshold depends explicitly on spectral measure of "noise"

Cauchy transform of μ:

$$G_\mu(z) = \int \frac{1}{z-y} d\mu(y) \quad \text{for } z \notin \operatorname{supp} \mu_X.$$

Phase transition of eigenvectors

Theorem [Benaych-Georges and N.]: As $n \longrightarrow \infty$, for $\theta > \theta_c$:

$$\boxed{|\langle \widetilde{u}_1, u\rangle|^2 \xrightarrow{\text{a.s.}} -\frac{1}{\theta_i^2 G'_\mu(\rho)}}$$

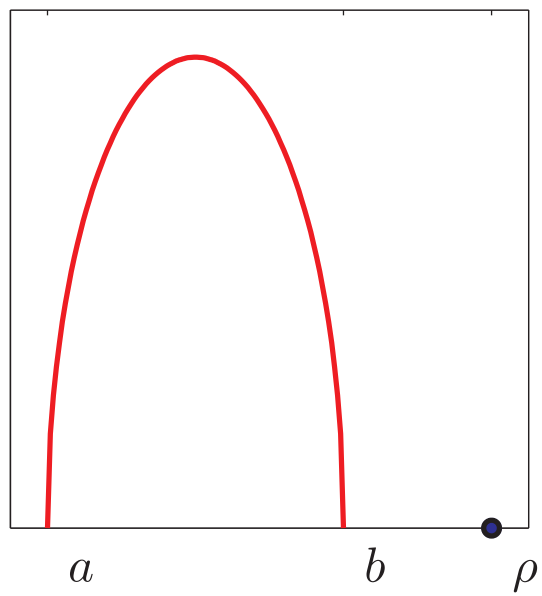
- $\rho = G_\mu^{-1}(1/\theta_i)$ is the corresponding eigenvalue limit

Theorem: As $n \longrightarrow \infty$, for $\theta \leq \theta_c$:

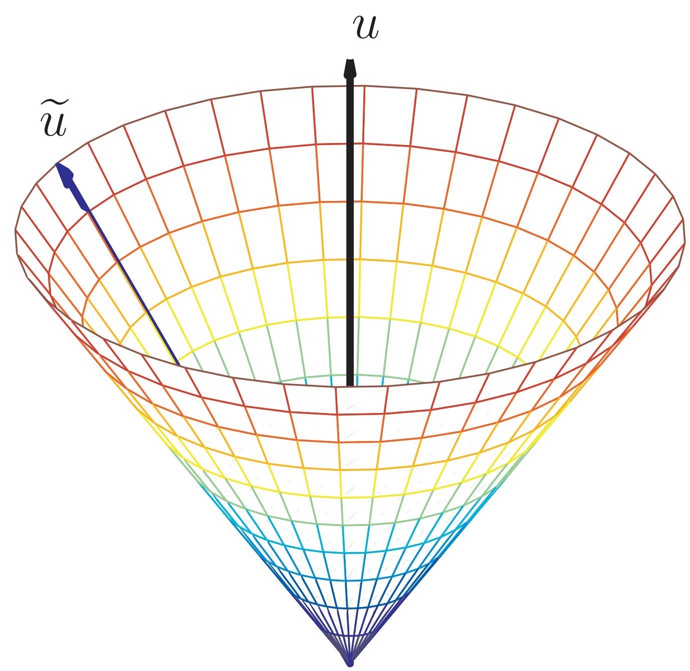
$$\boxed{\langle \widetilde{u}_1, u\rangle \xrightarrow{\text{a.s.}} 0}$$

- Eigenvalue density at edge needed of form $(x-b)^\alpha$ with $\alpha \in (0, 1]$

Above phase transition

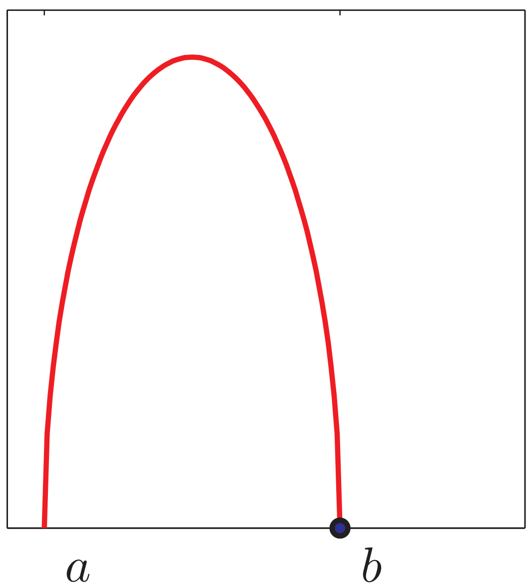


(a) Eigenvalue: $\theta > \theta_c$

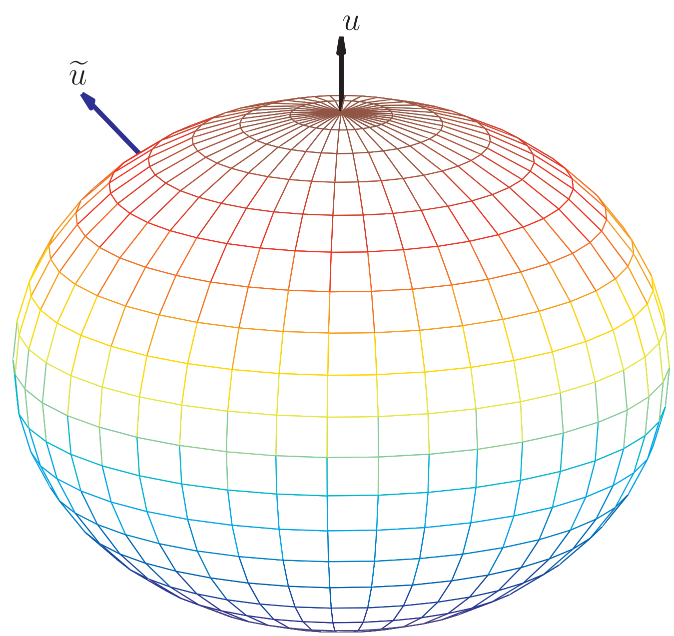


(b) Eigenvector: $\theta > \theta_c$

Below phase transition



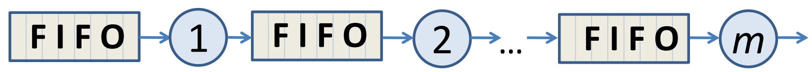
(c) Eigenvalue: $\theta \leq \theta_c$



(d) Eigenvector: $\theta \leq \theta_c$

The queuing theory connection

Problem setup



Assumptions:

- Infinitely long buffer
- Arrival process is Poissonian with rate α
- $w(i,j) \overset{ind.}{\sim} \exp(1/\mu_i) \Leftrightarrow$ M/M/m queue

Objective: Compute $L(m,n) =$ exit time for batch of n customers when

- Queues are in equilibrium before the batch of n customers arrive

The random matrix connection

Theorem [Baik & N., 2012]:

$$\boxed{L(m,n) \stackrel{\mathcal{D}}{=} \lambda_1(W)}$$

- $W = \Gamma^{1/2} g g^* \Gamma^{1/2} + \Sigma^{1/2} G G^* \Sigma^{1/2}$
- G is an $m \times (n-1)$ matrix of i.i.d. $\mathbb{CN}(0,1)$ entries
- g is an $m \times 1$ vector of i.i.d $\mathbb{CN}(0,1)$ entries
- $\Sigma = \mathrm{diag}(1/\mu_1, \ldots, 1/\mu_m)$
- $\Gamma = \mathrm{diag}\left(1/(\mu_1 - \alpha), \ldots, 1/(\mu_m - \alpha)\right)$
- Sanity check: $\alpha = 0$, $n = 1$, $L(m,n) = \sum_i 1/\mu_i |g_i|^2$
  - $|g_i|^2$ is chi-squared with 2 d.o.f. $\Leftrightarrow$ Exponential!

The random matrix connection

Theorem [Baik & N., 2012]:

$$\boxed{L(m,n) \stackrel{\mathcal{D}}{=} \lambda_1(W)}$$

- $W = \Gamma^{1/2} g g^* \Gamma^{1/2} + \Sigma^{1/2} G G^* \Sigma^{1/2}$
  - Rank-one-signal plus noise $\Rightarrow$ expect phase transition!
- G is an $m \times (n-1)$ matrix of i.i.d. $\mathbb{CN}(0,1)$ entries
- g is an $m \times 1$ vector of i.i.d $\mathbb{CN}(0,1)$ entries
- $\Sigma = \mathrm{diag}(1/\mu_1, \ldots, 1/\mu_m)$
- $\Gamma = \mathrm{diag}\left(1/(\mu_1 - \alpha), \ldots, 1/(\mu_m - \alpha)\right)$

New insight: phase transitions in queuing behavior

Recall:

- Arrival process is Poissonian with rate $\alpha < \mu_i$
- $w(i,j) \overset{ind.}{\sim} \exp(1/\mu_i) \Leftrightarrow$ M/M/m queue

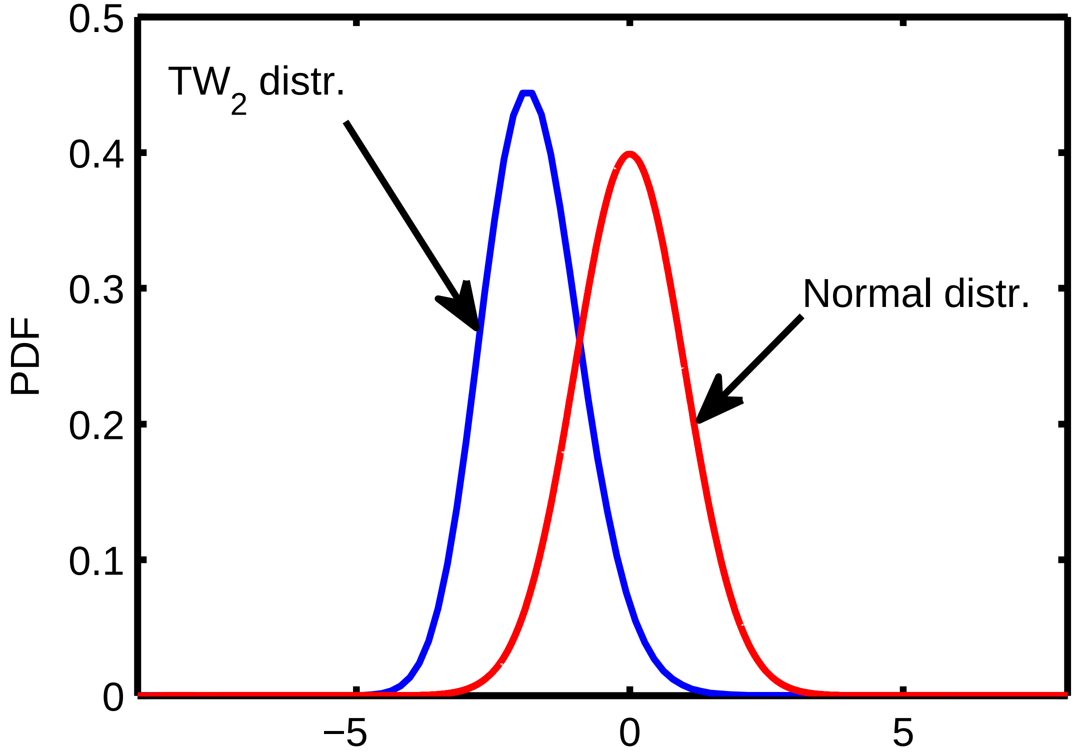
A critical rate:

$$l_{\text{crit}} = z \text{ such that} \sum_i \frac{1}{(\mu_i - z)^2} - \frac{n}{z^2} = 0, \qquad z = l_{\text{crit}} \in (0, \mu_{\min})$$

Theorem [Baik & N. , 2012]:

- Case 1: $0 < l_{\text{crit}} < \alpha \Leftrightarrow$ arrival rate is faster than critical rate
  - $L(m,n)$ is **normally** distributed: mean $O(n)$, variance $O(m)$
- Case 2: $l_{\text{crit}} > \mu_{\min} \Leftrightarrow$ slowest server is slower than critical rate
  - $L(m,n)$ is **normally** distributed: mean $O(n)$, variance $O(m)$
- Case 3: $\alpha < l_{\text{crit}} < \mu_{\min} \Leftrightarrow$ slowest server fast enough, arrival rate slow enough
  - $L(m,n)$ is **Tracy-Widom** distributed: mean $O(n)$ and variance $O(m^{2/3})$

Tracy-Widom versus Normal



New insights: phase transitions and more

A critical rate:

$$l_{\text{crit}} = z \text{ such that } \sum_i \frac{1}{(\mu_i - z)^2} - \frac{n}{z^2} = 0, \qquad z = l_{\text{crit}} \in (0, \mu_{\min})$$

Theorem [Baik & N. , 2012]:

- Case 1: $0 < l_{\text{crit}} < \alpha \Leftrightarrow$ arrival rate is faster than critical rate
  - $L(m,n)$ is **normally** distributed: mean $O(n)$, variance $O(m)$
- Case 2: $l_{\text{crit}} > \mu_{\min} \Leftrightarrow$ slowest server is slower than critical rate
  - $L(m,n)$ is **normally** distributed: mean $O(n)$, variance $O(m)$
- Case 3: $\alpha < l_{\text{crit}} < \mu_{\min} \Leftrightarrow$ slowest server fast enough, arrival rate slow enough
  - $L(m,n)$ is **Tracy-Widom** distributed: mean $O(n)$ and variance $O(m^{2/3})$

The importance of the variance scaling result

An elementary bound:

$$\mathrm{var}\max X_i \leq \sum_i \mathrm{var}\, X_i$$

Upper-bounding latency:

$$\mathrm{var}\, L(m, n) \leq O(n)$$

- Insight 1: Upper bound matched only when there is a bottleneck!
- Insight 2: Realized variance is much less than upper bound!
  - $\Rightarrow$ Service prov. due to upp. bound **very** conservative
  - Opportunity for perf. gains or relax system specs to meet existing QoS reqs!
    - $*$ Work with Mingyan Liu on optimal file-split.in multi-route, multi-hop ntwk

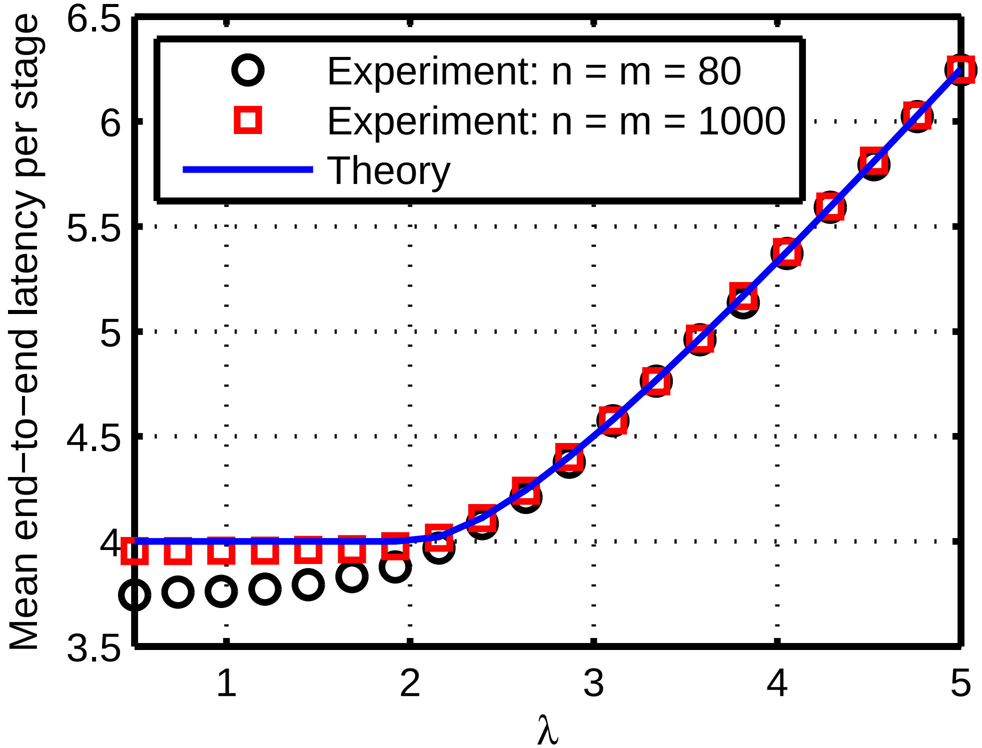
Numerical results

m	n	Mean		Variance	
		Experiment	Theory	Experiment	Theory
5	5	13.1024	12.3685	9.4351	15.0981
10	10	30.9954	30.3849	18.6033	23.9668
20	20	68.3172	67.8858	33.0268	38.0449
40	40	145.0274	144.7371	55.1251	60.3926
80	80	300.9902	300.7699	90.0644	95.8673
160	160	615.9515	615.7717	148.8302	152.1799
320	320	**1249**.4124	**1249**.4742	236.0294	241.5705
480	480	**1885**.7545	**1885**.0567	311.7331	316.5469
640	640	**2521**.6221	**2521**.5399	374.6064	383.4693
1000	1000	**3955**.4348	**3955**.3710	506.5496	516.3498

Empirical mean and variance of compared to theoretical predictions.

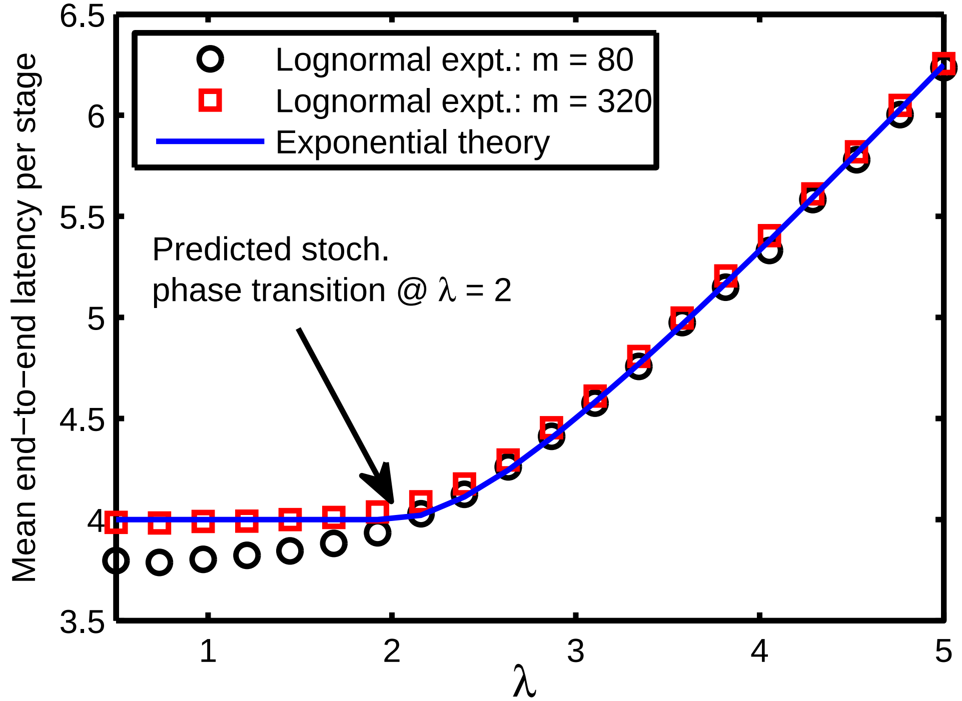
- Here $\mu_1 = \ldots = \mu_m = 1$
- "$8 = \infty$"

Numerical results



- Here $n = m$, $\mu_1 = \ldots = \mu_{m-1} = 1$, $\mu_m = 1/\lambda$; exponential service time
- Regime where the bottleneck does not affect distribution!

Numerical results



- Here $n = m$, $\mu_1 = \ldots = \mu_{m-1} = 1$, $\mu_m = 1/\lambda$; **lognormal** service time
- Conjecture: Distribution-independent limiting distribution

A fundamental recursion

Notation:

- S_i = Server $i \in \{1, \ldots, m\}$
- C_j = Customer $j \in \{1, \ldots, n\}$
- $w(i, j)$ = Service time for C_j at S_i
- $L(i, j)$ = Exit time for C_j from S_i

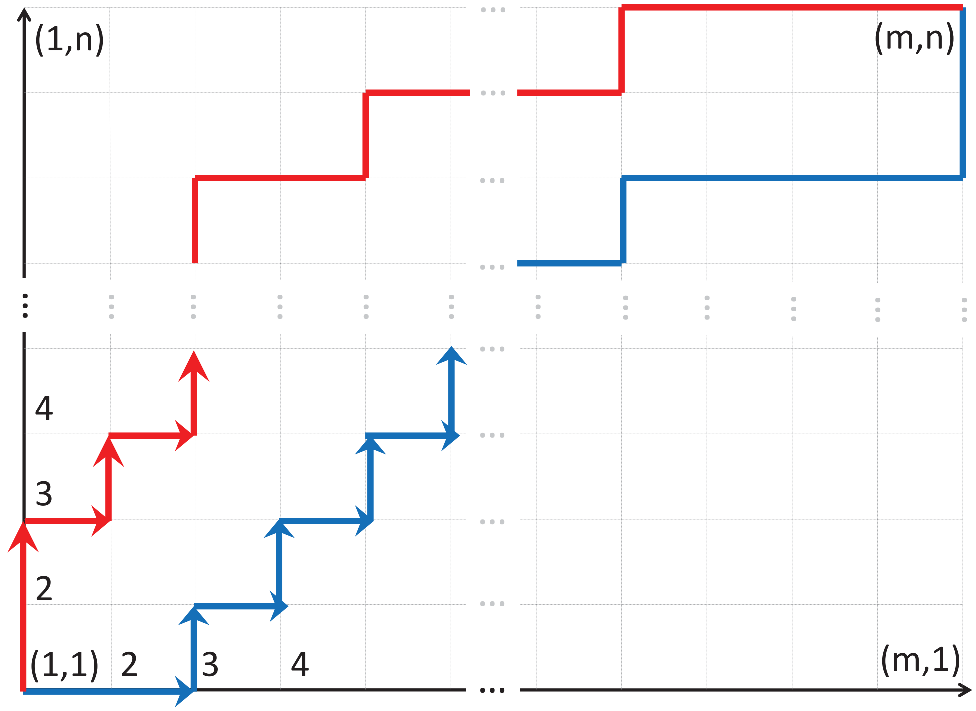
Fact (Glynn & Whitt, Tembe & Wolff):

$$L(i,j) = w(i,j) + \begin{cases} L(i-1,j) & \text{when } L(i,j-1) < L(i-1,j), \\ L(i,j-1) & \text{when } L(i,j-1) > L(i-1,j). \end{cases}$$

Equivalently,

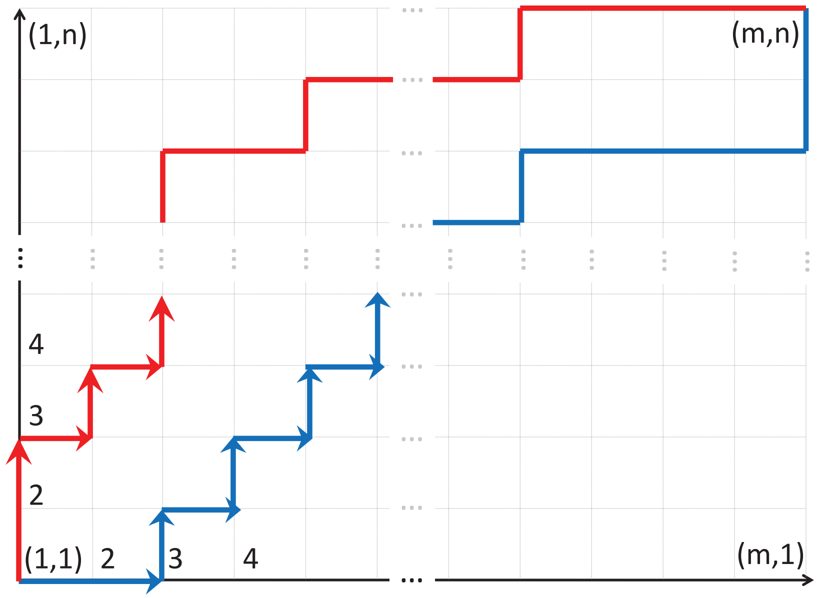
$$\boxed{L(i,j) = \max\{L(i-1,j), L(i,j-1)\} + w(i,j)}$$

The directed last-passage percolation problem



$$L(m,n) = \max\{L(m-1,n), L(m,n-1)\} + w(m,n)$$

The directed last-passage percolation problem



$$\boxed{L(m,n) = \max_{\pi \in P(m,m)} \left(\sum_{(k,\ell)\in\pi} w(k,\ell) \right)}$$

- $P(m,n)$ is the set of 'up/right paths' ending at (m,n)

The random matrix connection

$$\boxed{L(m,j) = \max_{\pi \in P(m,n)} \left(\sum_{(k,\ell)\in\pi} w(k,\ell) \right)}$$

<u>Theorem</u> [Borodin & Peché]: Assume

- $w(i,j) \sim \exp(1/(a_i + b_j))$
- $X_{ij} \sim \mathcal{CN}\left(0, \frac{1}{a_i+b_j}\right)$

$$\boxed{\Rightarrow L(m,n) \stackrel{\mathcal{D}}{=} \lambda_1(XX^*)}$$

- Related work by Johansson (2000)
- Result easily extended to Poissonian (discrete) random variables

The percolation mapping for our problem

C_6	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7
C_5	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7
C_4	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7
C_3	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7
C_2	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7
C_1	$\mu_1-\alpha$	$\mu_2-\alpha$	$\mu_3-\alpha$	$\mu_4-\alpha$	$\mu_5-\alpha$	$\mu_6-\alpha$	$\mu_7-\alpha$
	S_1	S_2	S_3	S_4	S_5	S_6	S_7

- Note that queues are in equilibrium before first customer enters
- Queue lengths are random and have (shifted) geometric distribution
- $\Rightarrow$ First customer served at S_i with rate $\mu_i - \alpha$, rest with μ_i
  - PASTA property = Poissonian Arrivals See Time Averages

Ergo the random matrix connection

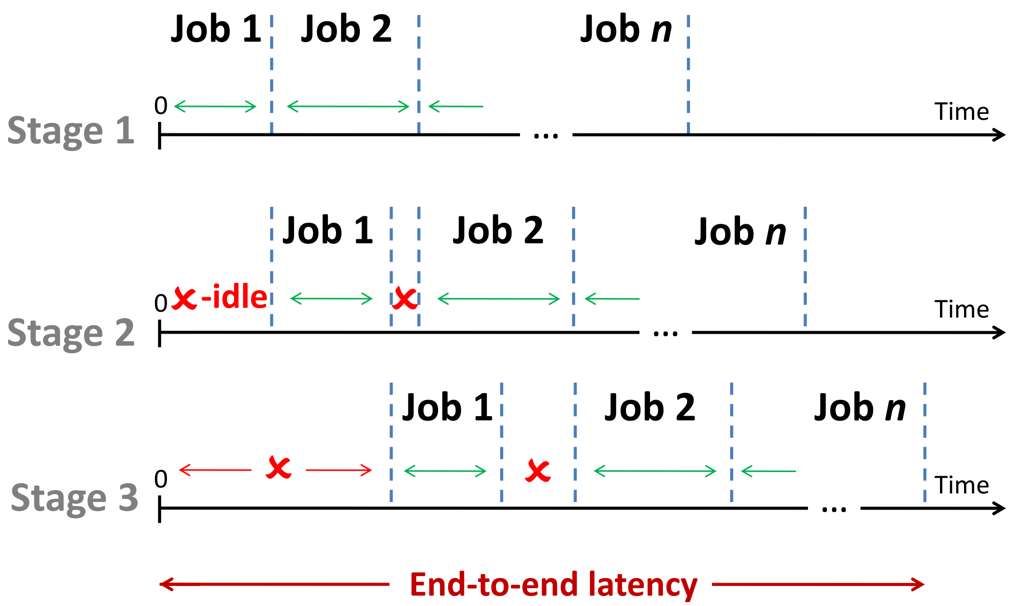
Theorem [Baik & N., 2012]:

$$\boxed{L(m,n) \stackrel{\mathcal{D}}{=} \lambda_1(W)}$$

- $W = \Gamma^{1/2} g g^* \Gamma^{1/2} + \Sigma^{1/2} G G^* \Sigma^{1/2}$
  - Rank-one-signal plus noise $\Rightarrow$ expect phase transition!
- G is an $m \times (n-1)$ matrix of i.i.d. $\mathbb{CN}(0,1)$ entries
- g is an $m \times 1$ vector of i.i.d $\mathbb{CN}(0,1)$ entries
- $\Sigma = \mathrm{diag}(1/\mu_1, \ldots, 1/\mu_m)$
- $\Gamma = \mathrm{diag}\,(1/(\mu_1 - \alpha), \ldots, 1/(\mu_m - \alpha))$

Why the random matrix connection?

Where are the non-intersecting random walks?

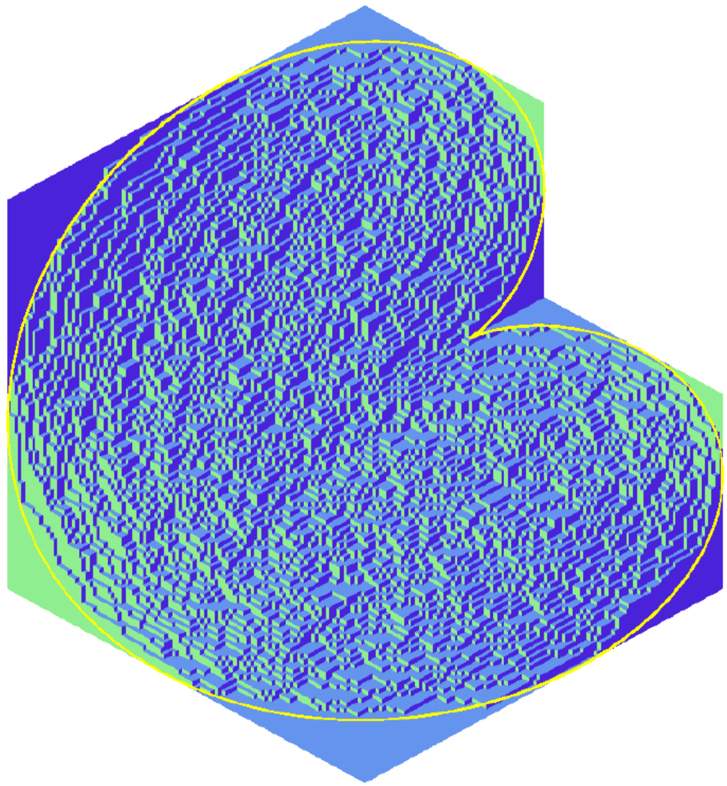


- FIFO protocol means exit time trajectories do not intersect
- Mathematics of random walks $\Leftrightarrow$ classical probability theory
- Mathematics of random walks <u>conditioned not to intersect</u> $\Leftrightarrow$ random matrix theory

Bijection with TASEP & corner growth model

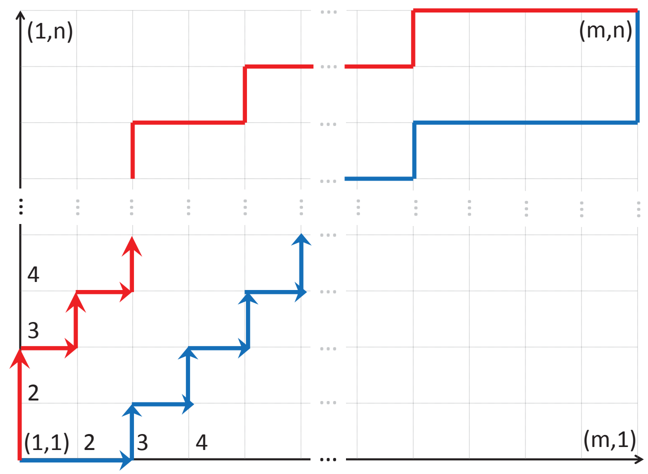
http://www-wt.iam.uni-bonn.de/~ferrari/animations/ContinuousTASEP.html

Non-interesecting random walks are everywhere!



- Taken from Andrei Okounkov's 2006 Fields Medal Citation

The traveling salesman problem



$$\boxed{L(m,n) = \max_{\pi \in P(m,n)} \left(\sum_{(k,\ell)\in\pi} w(k,\ell) \right)}$$

- Fix $w(k,l)$, what order of processing minimizes delay?
  - Limits of scheduling? Application-motivated extensions of RMT!

Main message

New insights beyond Little's Law:

- Latency mean and variance can be explicitly computed! ✓
- Analysis reveals emergence of **phase transitions** ✓
- **Rigorous** basis for statistical anomaly testing ✓
- Can show that $O(n^{1/3})$ jobs have statistically independent latencies ✓
- Extends easily to quasi-reversible networks (thanks Demos!) ✓
- Analysis of queue-state dependent servicing (inspired by backpressure algorithms) ✓
- Results appear to hold even for non-exponential service times ✓
  - Universality conjecture!

All made possible due to connection with random matrix theory!