A Data-Based Science for Service Engineering and Management, or

"Empirical Adventures in Call-Centers and Hospitals"

Avi Mandelbaum

Technion, Haifa, Israel

http://ie.technion.ac.il/serveng

LOIS Lecture, Eindhoven, October 2010

Research Partners

- **Students**:
  Aldor*, Baron*, Carmeli, Feldman, Garnett*, Gurvich*, Khudiakov*, Maman*, Marmor*, Reich, Rosenshmidt*, Shaikhet*, Senderovic, Tseytlin*, Yom-Tov*, Zaied, Zeltyn*, Zohar*, Zviran, . . .
- **Empirical/Statistical Analysis**:
  Feigin; Brown, Gans, Zhao; Shen; Ritov, Goldberg; Allon, Bassamboo, Gurvich; Armony, . . .
- **Theory**:
  Armony, Atar, Gurvich, Jelenkovic, Kaspi, Massey, Momcilovic, Reiman, Shimkin, Stolyar, Wasserkrug, Whitt, Zeltyn, . . .
- **Industry**:
  IBM Research (OCR: Carmeli, Vortman, Wasserkrug, Zeltyn), Rambam Hospital, Hapoalim Bank, Mizrahi Bank, Pelephone Cellular, . . .
- **Technion SEE Center / Labaratory**:
  Feigin; Trofimov, Nadjharov, Gavako, Kutsyy; Liberman, Koren, Rom, Plonsky; Research Assistants, . . .

History, Resources (Downloadable)

- Math. + C.S. + Stat. + O.R. + Mgt. $\Rightarrow$ **IE** ($\geq$ 1990)
- **"Service-Engineering" Course** ($\geq$ 1995):
  http://ie.technion.ac.il/serveng - website
  http://ie.technion.ac.il/serveng/References/**teaching**_paper.pdf

History, Resources (Downloadable)

- Math. + C.S. + Stat. + O.R. + Mgt. $\Rightarrow$ **IE** ($\geq$ 1990)
- **"Service-Engineering" Course** ($\geq$ 1995):
  http://ie.technion.ac.il/serveng - website
  http://ie.technion.ac.il/serveng/References/**teaching**_paper.pdf
- Search Google-Scholar for <**Call Centers**>
- **SEELab** ($\geq$ 2007), following StatLab ($\geq$ 2000):
  Data Repositories for Research & Teaching; Reports, Tutorials:
  http://ie.technion.ac.il/Labs/Serveng

History, Resources (Downloadable)

- Math. + C.S. + Stat. + O.R. + Mgt. $\Rightarrow$ **IE** ($\geq$ 1990)
- **"Service-Engineering" Course** ($\geq$ 1995):
  http://ie.technion.ac.il/serveng - website
  http://ie.technion.ac.il/serveng/References/**teaching**_paper.pdf
- Search Google-Scholar for <**Call Centers**>
- **SEELab** ($\geq$ 2007), following StatLab ($\geq$ 2000):
  Data Repositories for Research & Teaching; Reports, Tutorials:
  http://ie.technion.ac.il/Labs/Serveng
- **OCR** Project ($\geq$ 2008): **Hospitals**
  IBM Research + Rambam Hospital + Technion IE&M
  http://ie.technion.ac.il/Labs/Serveng/closed/OCR_Documents.php

The Case for Service Science / Engineering

- **Service Science / Engineering** (vs. Management) are emerging **Academic Disciplines**. For example, universities (world-wide), IBM (SSME, a là Computer-Science), USA NSF (SEE), Germany IAO (ServEng), ...

The Case for Service Science / Engineering

- **Service Science / Engineering** (vs. Management) are emerging **Academic Disciplines**. For example, universities (world-wide), IBM (SSME, a là Computer-Science), USA NSF (SEE), Germany IAO (ServEng), ...

- Models that explain fundamental phenomena, which are **common** across applications:
  - **Call Centers**
  - **Hospitals**
  - **Transportation**
  - Justice, Fast Food, Police, Internet, . . .

- Simple models at the Service of Complex Realities (Human)
  Note: Simple yet rooted in **deep analysis**.

The Case for Service Science / Engineering

- **Service Science / Engineering** (vs. Management) are emerging **Academic Disciplines.** For example, universities (world-wide), IBM (SSME, a là Computer-Science), USA NSF (SEE), Germany IAO (ServEng), ...

- Models that explain fundamental phenomena, which are **common** across applications:
  - **Call Centers**
  - **Hospitals**
  - **Transportation**
  - Justice, Fast Food, Police, Internet, . . .

- Simple models at the Service of Complex Realities (Human)
  Note: Simple yet rooted in **deep analysis**.

- Mostly **What Can Be Done** vs. **How To**

Title: Expands the Scientific Paradigm

Physics, Biology, . . . : Measure, Model, Experiment, Validate, Refine.

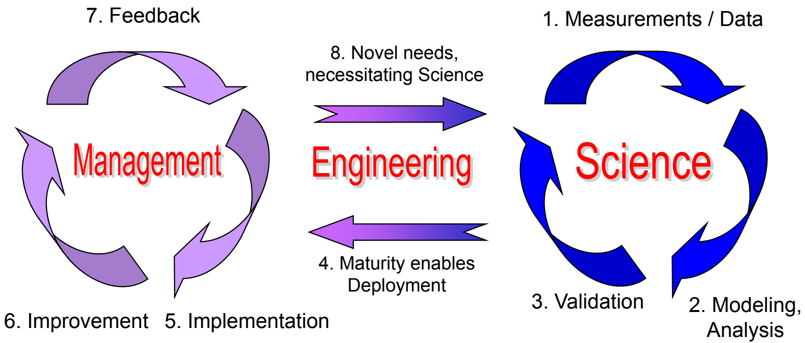
Human-complexity triggered above in Transportation, Economics.

Title: Expands the Scientific Paradigm

Physics, Biology, . . . : Measure, Model, Experiment, Validate, Refine.

Human-complexity triggered above in Transportation, Economics.

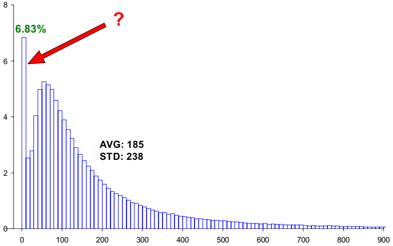
Starting with **Data**, expand to:



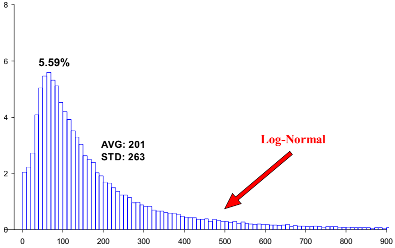
Beyond Averages: The Human Factor

Histogram of Service-Time in a (Small Israeli) Bank, 1999

January-October



November-December

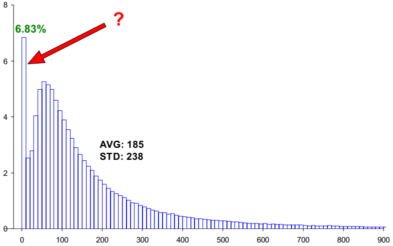


- 6.8% Short-Services:

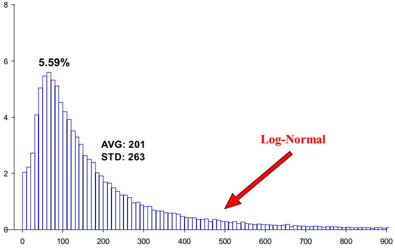
Beyond Averages: The Human Factor

Histogram of Service-Time in a (Small Israeli) Bank, 1999

January-October



November-December

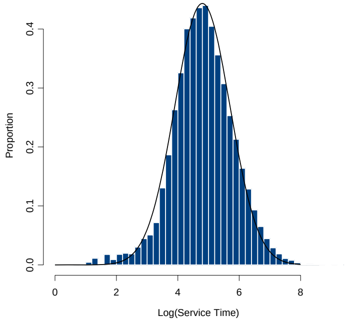


- **6.8% Short-Services:** Agents' "Abandon" (improve bonus, rest), (mis)lead by **incentives**
- **Distributions** must be measured (in **seconds** = **natural scale**)
- **LogNormal** service times common in call centers

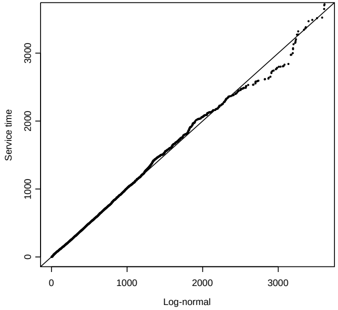
Validating LogNormality of Service-Times

Israeli Call Center, Nov-Dec, 1999

Log(Service Times)



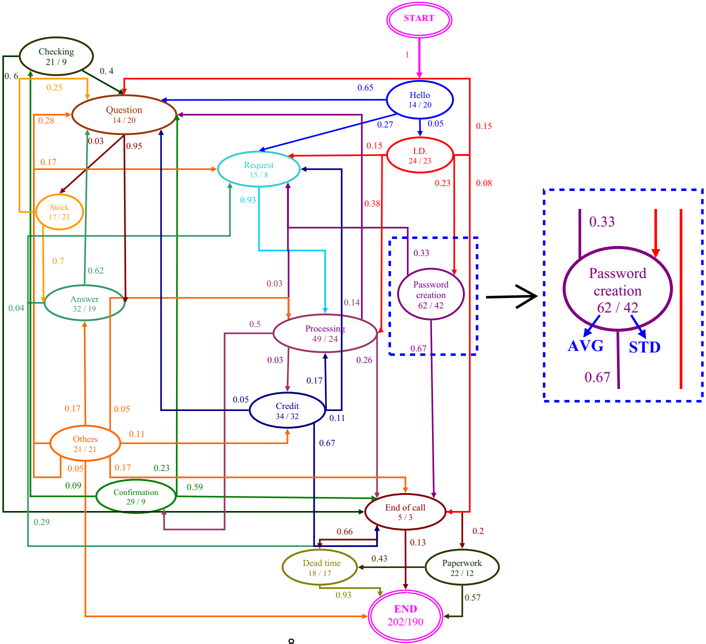
LogNormal QQPlot



- **Practically Important**: (mean, std)(log) characterization
- **Theoretically Intriguing**: Why LogNormal ? Naturally multiplicative but, in fact, also **Infinitely-Divisible** (Generalized Gamma-Convolutions)
- Simple-model of a complex-reality? The **Service Process:**

(Telephone) Service-Process = “Phase-Type" Model

Retail Service (Israeli Bank)



Statistics OR IE

Why Bother?

In large banking call centers:
+One Second to Service-Time implies **+Millions** in costs, annually

$\Rightarrow$ **Time and "Motion" Studies** (**Classical IE** with New-age IT)

Why Bother?

In large banking call centers:
+One Second to Service-Time implies **+Millions** in costs, annually

$\Rightarrow$ **Time and "Motion" Studies** (**Classical IE** with New-age IT)

- **Service-Process Model**: Customer-Agent Interaction
  - **Work Design** (w/ **Khudiakov**)
    eg. **Cross-Selling**: higher profit vs. (costlier) longer services;
    Analysis yields (congestion-dependent) cross-selling protocol
  - **“Worker" Design** (w/ **Gans & Shen**)
    eg. **Learning, Forgetting,** . . . : Predict individual performance;
    Important in high-turnover environments

Why Bother?

In large banking call centers:
+One Second to Service-Time implies **+Millions** in costs, annually

⇒ **Time and "Motion" Studies** (**Classical IE** with New-age IT)

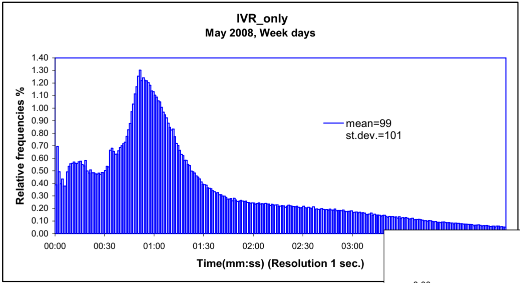
- **Service-Process Model**: Customer-Agent Interaction
  - **Work Design** (w/ **Khudiakov**)
    eg. **Cross-Selling**: higher profit vs. (costlier) longer services;
    Analysis yields (congestion-dependent) cross-selling protocol
  - **"Worker" Design** (w/ **Gans & Shen**)
    eg. **Learning, Forgetting,** . . . : Predict individual performance;
    Important in high-turnover environments

- **IVR-Process Model**: Customer-Machine Interaction
  75% services, poor design, yet scarce research;
  Same approach, automatic (easier) data

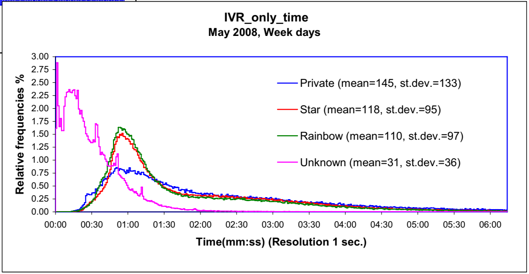
IVR-Time: Histograms

Israeli Bank: Served only by IVR, May 2008

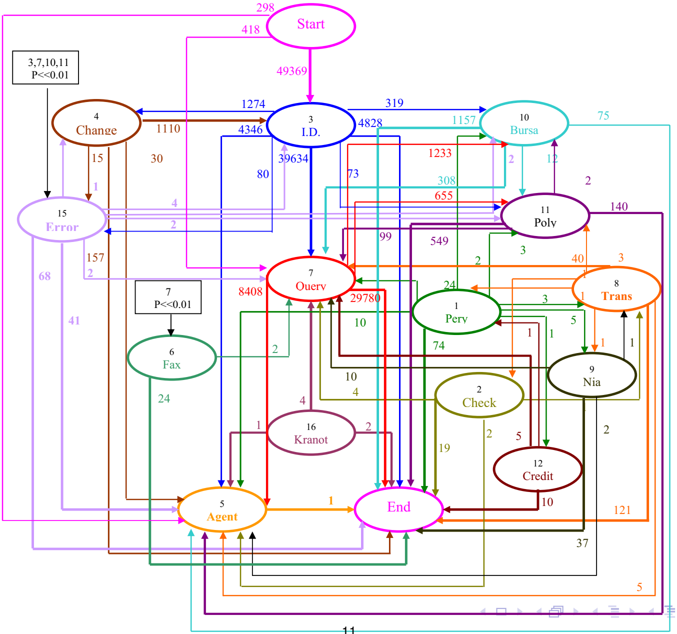
All Customers



By Service Type

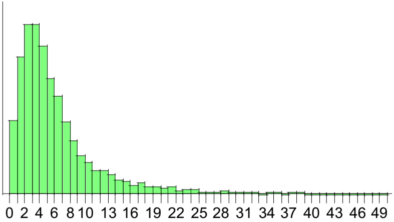


IVR-Process: “Phase-Type" Model



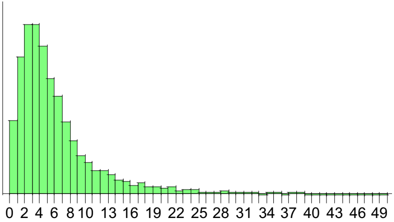
Beyond Averages: Length-of-Stay in a Hospital

Israeli Hospital, in Days: LN

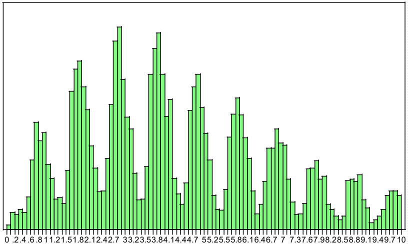


Beyond Averages: Length-of-Stay in a Hospital

Israeli Hospital, in Days: LN

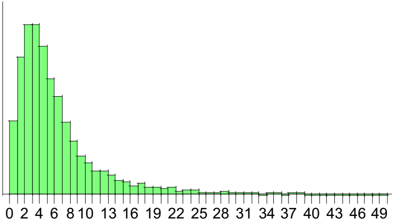


Israeli Hospital, in Hours

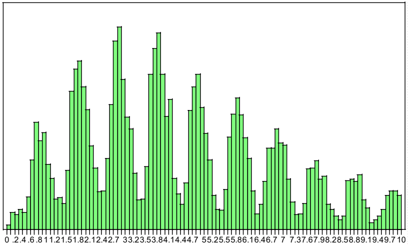


Beyond Averages: Length-of-Stay in a Hospital

Israeli Hospital, in Days: LN



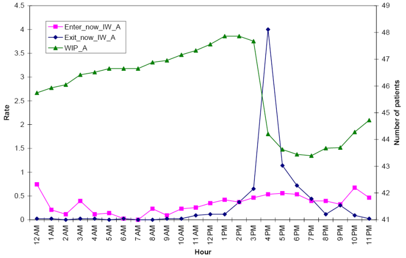
Israeli Hospital, in Hours



"Explanation": Patients released around **3pm** (1pm in Singapore)

Why Bother ?

Staffing, Bed Management, . . .



Started with Call Centers, Expanded to Hospitals

Call Centers - U.S. (Netherlands) Stat.

- \$200 – \$300 billion annual expenditures (0.5)
- 100,000 – 200,000 call centers (1500-2000)
- “Window" into the company, for better or worse
- Over 3 million agents = **2% – 4% workforce** (100K)

Started with Call Centers, Expanded to Hospitals

Call Centers - U.S. (Netherlands) Stat.

- \$200 – \$300 billion annual expenditures (0.5)
- 100,000 – 200,000 call centers (1500-2000)
- “Window" into the company, for better or worse
- Over 3 million agents = **2% – 4% workforce** (100K)

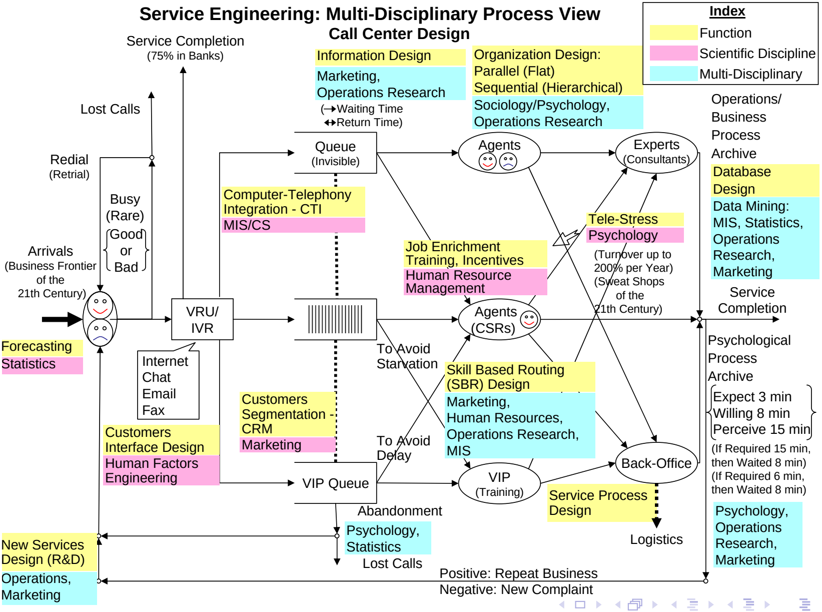
Healthcare - similar and unique challenges:

- Cost-figures far more staggering
- Risks much higher
- ED (initial focus) = hospital-window
- Over 3 million nurses

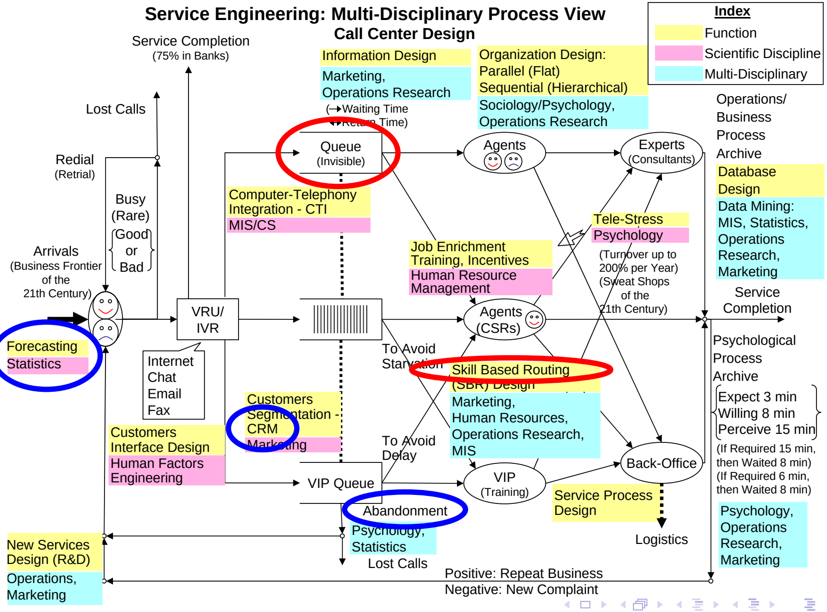
Call-Center Environment: Service Network

Call-Centers: “Sweat-Shops of the 21st Century"

Call-Center Network: Gallery of Models



Call-Center Network: Gallery of Models



Skills-Based Routing in Call Centers

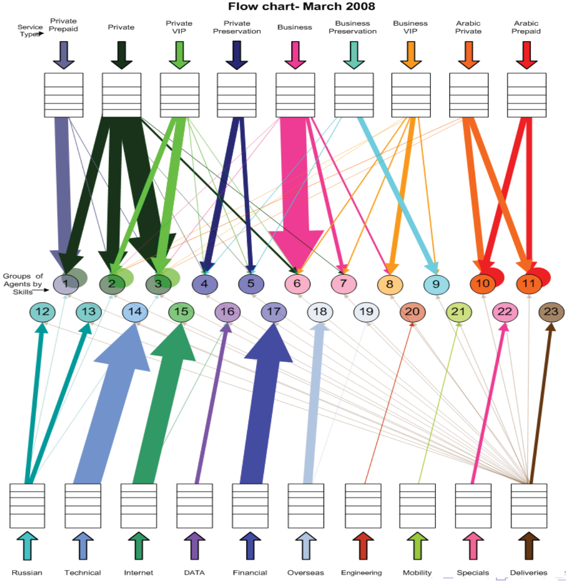
EDA and OR, with I. Gurvich and P. Liberman

Mktg. ⇒

OR ⇒

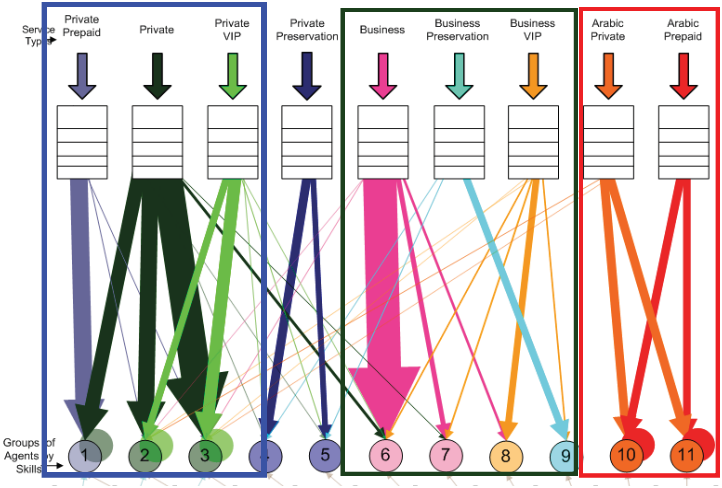
HRM ⇒

MIS ⇒



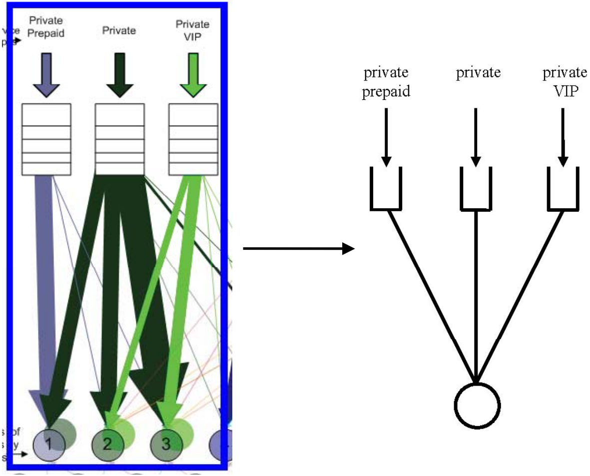
SBR Topologies: I; V, Reversed-V; N, X; W, M

Israeli Cellular, March 2008



SBR: Class-Dependent Services

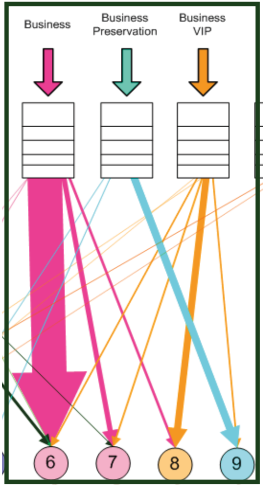
"Reduction" to V-Topology, with **R. Atar and G. Shaikhet**

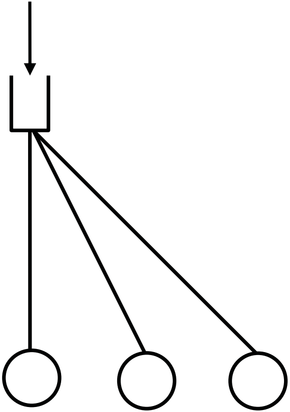


Reduction in the sense of **equivalent Brownian Control Problems**

SBR: Pool-Dependent Services

"Reduction" to Reversed-V and I, with **R. Atar and G. Shaikhet**

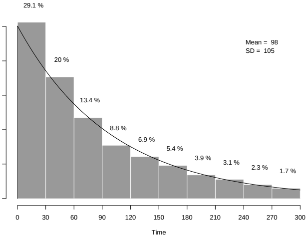




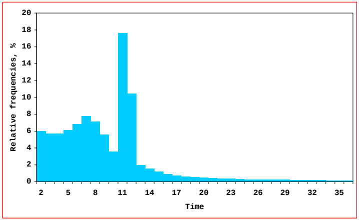
Reduction in the sense of **equivalent Brownian Control Problems**

Beyond Averages: Waiting Times in a Call Center

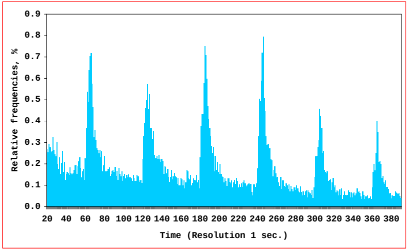
Small Israeli Bank



Large U.S. Bank



Medium Israeli Bank, in Seconds (Recall Hospital LOS, Hours)



ER / ED Environment: Service Network

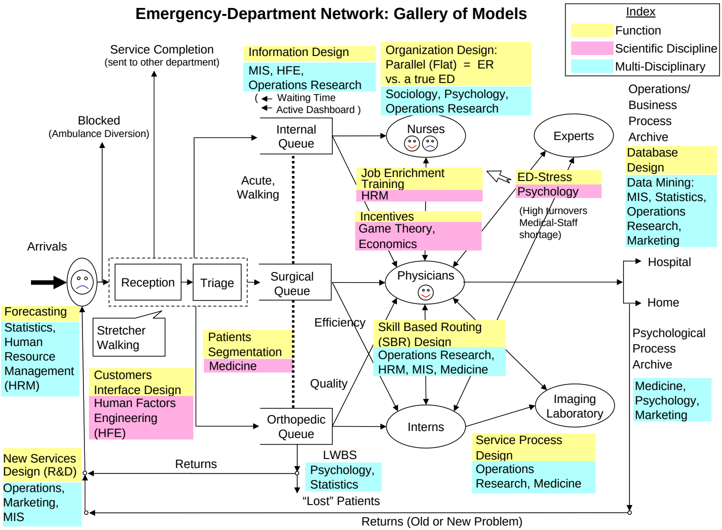
Acute (Internal, Trauma)

Walking

Multi-Trauma

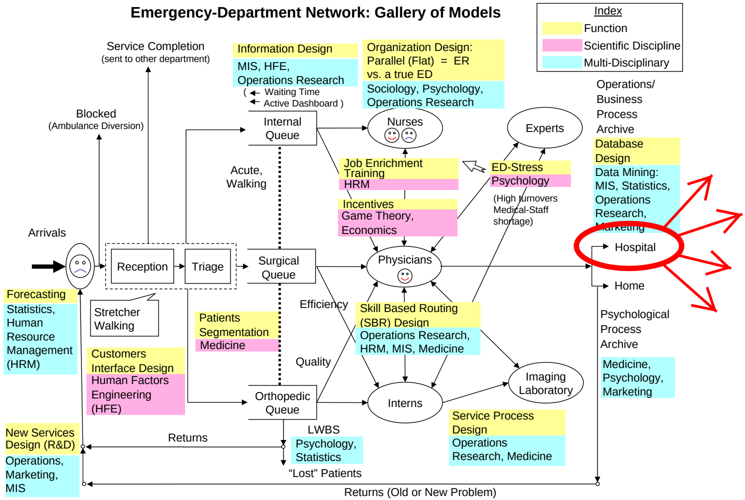
Queueing in a “Good" Beijing Hospital, at 6am

Emergency-Department Network: Gallery of Models



- **Forecasting**, SBR ≈ **Triage**, Abandonment = **LWBS**

Emergency-Department Network: Gallery of Models



- **Fork-Join** Q's, eg. After Physician: Nurse, Lab-Tests, X-Ray
- **Synchronization** Control, with **R. Atar and A. Zviran**
- **ED-to-IW** Routing

ED-to-IW Routing: A Hospital Bottleneck

Israeli Large Hospital (1/5/06 to 30/10/08, excluding 1-3/07)

	Ward A	**Ward B**	Ward C	Ward D
ALOS (days)	6.37	**4.47**	5.36	5.56
Avg Occupancy Rate	97%	**95%**	86%	92%
Avg # Patients per Month	206	187	210	210
Standard bed capacity	45	30	44	42
Avg # Patients /Bed/Month	4.57	**6.25**	4.77	4.77
Returns (within 3 months)	15.4%	15.6%	16.2%	14.8%

ED-to-IW Routing: A Hospital Bottleneck

Israeli Large Hospital (1/5/06 to 30/10/08, excluding 1-3/07)

	Ward A	**Ward B**	Ward C	Ward D
ALOS (days)	6.37	**4.47**	5.36	5.56
Avg Occupancy Rate	97%	**95%**	86%	92%
Avg # Patients per Month	206	187	210	210
Standard bed capacity	45	30	44	42
Avg # Patients /Bed/Month	4.57	**6.25**	4.77	4.77
Returns (within 3 months)	15.4%	15.6%	16.2%	14.8%

- The “**fastest" Ward B** is subject to highest **workload** = **bed-occupancy, bed-turnover (flux)**, yet clinically apt: **unfair!**

ED-to-IW Routing: A Hospital Bottleneck

Israeli Large Hospital (1/5/06 to 30/10/08, excluding 1-3/07)

	Ward A	**Ward B**	Ward C	Ward D
ALOS (days)	6.37	**4.47**	5.36	5.56
Avg Occupancy Rate	97%	**95%**	86%	92%
Avg # Patients per Month	206	187	210	210
Standard bed capacity	45	30	44	42
Avg # Patients /Bed/Month	4.57	**6.25**	4.77	4.77
Returns (within 3 months)	15.4%	15.6%	16.2%	14.8%

- The "**fastest" Ward B** is subject to highest **workload** = **bed-occupancy, bed-turnover (flux)**, yet clinically apt: **unfair!**
- With **P. Momcilovic and Y. Tseytlin**: Routing based on **Idleness-Ratios** (# idle beds in ward / # idle-beds in total), such that the "faster" the ward:
  - **Fairness**: the lower the bed-occupancy (**nurses happy**)
  - **Efficiency**: the higher the bed-turnover (**managers happy**)

ED-to-IW Routing: A Hospital Bottleneck

Israeli Large Hospital (1/5/06 to 30/10/08, excluding 1-3/07)

	Ward A	**Ward B**	Ward C	Ward D
ALOS (days)	6.37	**4.47**	5.36	5.56
Avg Occupancy Rate	97%	**95%**	86%	92%
Avg # Patients per Month	206	187	210	210
Standard bed capacity	45	30	44	42
Avg # Patients /Bed/Month	4.57	**6.25**	4.77	4.77
Returns (within 3 months)	15.4%	15.6%	16.2%	14.8%

- The “**fastest" Ward B** is subject to highest **workload** = **bed-occupancy, bed-turnover (flux)**, yet clinically apt: **unfair!**
- With **P. Momcilovic and Y. Tseytlin**: Routing based on **Idleness-Ratios** (# idle beds in ward / # idle-beds in total), such that the “faster" the ward:
  - **Fairness**: the lower the bed-occupancy (**nurses happy**)
  - **Efficiency**: the higher the bed-turnover (**managers happy**)
- **Reversed-V**: **Queue** = ED, **Servers** in Pool = Beds in Ward (**10's**)
- **Information** Analysis: **QED/Sub-Diffusion Approx.** (**Natural**)

Prerequisite I: Data

Averages Prevalent (and could be useful / interesting).

But I need data at the level of the **Individual Transaction**:
For each service transaction (during a phone-service in a call center, or a patient's visit in a hospital, or browsing in a website, or . . .), its **operational history** = time-stamps of events.

Prerequisite I: Data

Averages Prevalent (and could be useful / interesting).

But I need data at the level of the **Individual Transaction**:
For each service transaction (during a phone-service in a call center, or a patient's visit in a hospital, or browsing in a website, or . . .), its **operational history** = time-stamps of events.

Sources: **"Service-floor"** (vs. Industry-level, Surveys, . . .)

- Administrative (Court, via "paper analysis")
- Face-to-Face (Bank, via bar-code readers)
- **Telephone** (Call Centers, via ACD / CTI, IVR/VRU)
- **Hospitals** (Emergency Departments, . . .)

Prerequisite I: Data

Averages Prevalent (and could be useful / interesting).

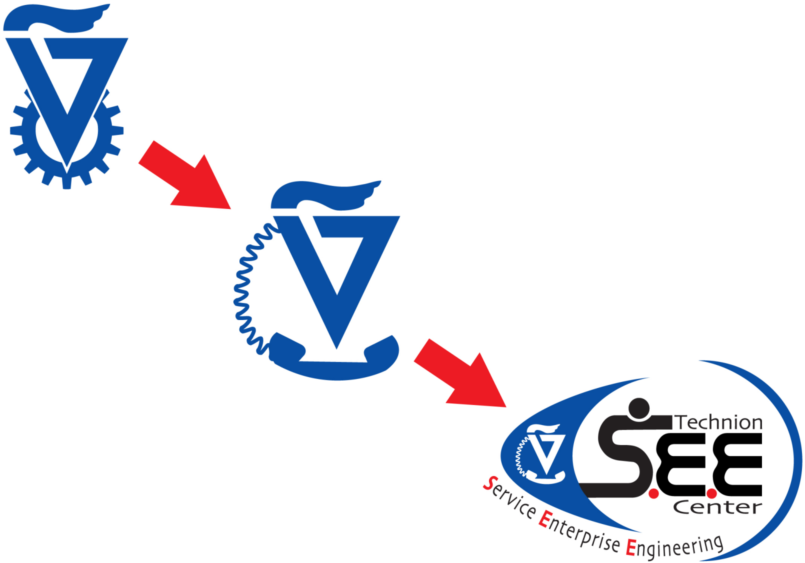
But I need data at the level of the **Individual Transaction**:
For each service transaction (during a phone-service in a call center, or a patient's visit in a hospital, or browsing in a website, or . . .), its **operational history** = time-stamps of events.

Sources: **"Service-floor"** (vs. Industry-level, Surveys, . . .)

- Administrative (Court, via "paper analysis")
- Face-to-Face (Bank, via bar-code readers)
- **Telephone** (Call Centers, via ACD / CTI, IVR/VRU)
- **Hospitals** (Emergency Departments, . . .)
- Expanding:
  - Hospitals, via **RFID**, with **I. Cohen, S. Israelit (MD), Y. Marmor**
  - Operational + Financial + Contents (Marketing, Clinical)
  - Internet, Chat (multi-media)

Pause for a Commercial:

Pause for a Commercial: The Technion SEE Center



Technion SEE = Service Enterprise Engineering

SEELab: Data-repositories for research and teaching

- For example:
  - Bank Anonymous: **1 years, 350K calls by 15 agents** - in 2000.
  - U.S. Bank: **2.5 years, 220M calls, 40M by 1000 agents**.
  - Israeli Cellular: **2.5 years, 110M calls, 25M calls by 750 agents**.
  - Israeli Bank: **from January 2010, daily-deposit at a SEESafe**.
  - Israeli Hospital: **4 years, 1000 beds; 8 ED's- Sinreich's data**.

Technion SEE = Service Enterprise Engineering

SEELab: Data-repositories for research and teaching

- For example:
  - Bank Anonymous: **1 years, 350K calls by 15 agents** - in 2000.
  - U.S. Bank: **2.5 years, 220M calls, 40M by 1000 agents**.
  - Israeli Cellular: **2.5 years, 110M calls, 25M calls by 750 agents**.
  - Israeli Bank: **from January 2010, daily-deposit at a SEESafe**.
  - Israeli Hospital: **4 years, 1000 beds; 8 ED's- Sinreich's data**.

SEEStat: Environment for graphical EDA in real-time

- **Universal Design, Internet Access, Real-Time Response.**

Technion SEE = Service Enterprise Engineering

SEELab: **Data-repositories for research and teaching**

- For example:
  - Bank Anonymous: **1 years, 350K calls by 15 agents** - in 2000.
  - U.S. Bank: **2.5 years, 220M calls, 40M by 1000 agents**.
  - Israeli Cellular: **2.5 years, 110M calls, 25M calls by 750 agents**.
  - Israeli Bank: **from January 2010, daily-deposit at a SEESafe**.
  - Israeli Hospital: **4 years, 1000 beds; 8 ED's- Sinreich's data**.

SEEStat: **Environment for graphical EDA in real-time**

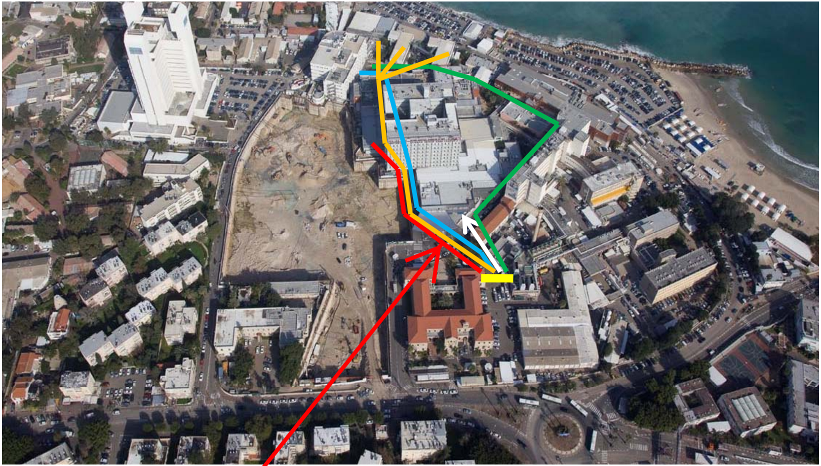
- **Universal Design, Internet Access, Real-Time Response**.

SEEServer: **Free for academic use**

Register, then access (presently) U.S. Bank and Small Israeli Bank.

eg. RFID-Based Data: Mass Casualty Event (MCE)

Drill: Chemical MCE, Rambam Hospital, May 2010



Focus on **severely wounded** casualties (≈ 40 in drill)

Note: 20 observers support real-time control (will help validation)

Data Cleaning: MCE with RFID Support

Data-base				Company report		comment
Asset id	order	**Entry** date	**Exit** date	Entry date	Exit date	
4	1	1:14:07 PM		1:14:00 PM		
6	1	12:02:02 PM	12:33:10 PM	12:02:00 PM	12:33:00 PM	
8	1	11:37:15 AM	12:40:17 PM	11:37:00 AM		**exit is missing**
10	1	12:23:32 PM	12:38:23 PM	12:23:00 PM		
12	1	12:12:47 PM	12:35:33 PM		12:35:00 PM	**entry is missing**
15	1	1:07:15 PM		1:07:00 PM		
16	1	11:18:19 AM	11:31:04 AM	11:18:00 AM	11:31:00 AM	
17	1	1:03:31 PM		1:03:00 PM		
18	1	1:07:54 PM		1:07:00 PM		
19	1	12:01:58 PM		12:01:00 PM		
20	1	11:37:21 AM	12:57:02 PM	11:37:00 AM	12:57:00 PM	
21	1	12:01:16 PM	12:37:16 PM	12:01:00 PM		
22	1	12:04:31 PM	12:20:40 PM			first customer is missing
22	2	12:27:37 PM		12:27:00 PM		
25	1	12:27:35 PM	1:07:28 PM	12:27:00 PM	1:07:00 PM	
27	1	12:06:53 PM		12:06:00 PM		
28	1	11:21:34 AM	11:41:06 AM	11:41:00 AM	11:53:00 AM	**exit time instead of entry time**
29	1	12:21:06 PM	12:54:29 PM	12:21:00 PM	12:54:00 PM	
31	1	11:40:54 AM	12:30:16 PM	11:40:00 AM	12:30:00 PM	
31	2	12:37:57 PM	12:54:51 PM	12:37:00 PM	12:54:00 PM	
32	1	11:27:11 AM	12:15:17 PM	11:27:00 AM	12:15:00 PM	
33	1	12:05:50 PM	12:13:12 PM	12:05:00 PM	12:15:00 PM	wrong exit time
35	1	11:31:48 AM	11:40:50 AM	11:31:00 AM	11:40:00 AM	
36	1	12:06:23 PM	12:29:30 PM	12:06:00 PM	12:29:00 PM	
37	1	11:31:50 AM	11:48:18 AM	11:31:00 AM	11:48:00 AM	
37	2	12:59:21 PM		12:59:00 PM		

Imagine **“Cleaning" 60,000+ customers per day** (call centers) !

Prerequisite II: Models (Fluid Q's)

"Laws of Large Numbers" capture **Predictable** Variability

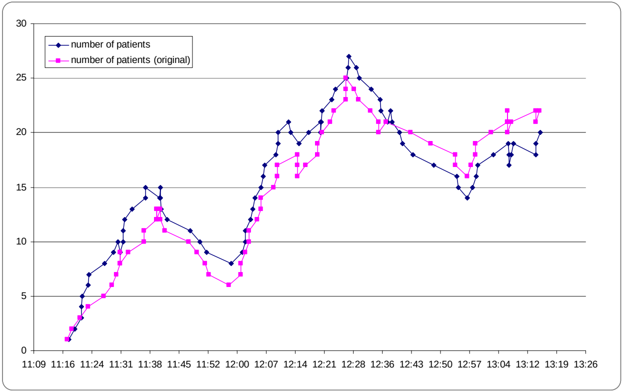
Deterministic Models: Scale Averages-out **Stochastic Individualism**

Prerequisite II: Models (Fluid Q's)

"Laws of Large Numbers" capture **Predictable** Variability

Deterministic Models: Scale Averages-out **Stochastic Individualism**

$\#$ **Severely-Wounded Patients, 11:00-13:00**



- Paths of doctors, nurses, patients (100+, **1 sec.** resolution) eg. Help predict "What if **150+ casualties** severely wounded ?"
- **Transient** Q's, where **Service-Process** = **Needy-Content Cycles** (with **G. Yom-Tov**, PhD)

Prerequisite II: Models (Diffusion/QED's Q's)

Traditional Queueing Theory predicts that **Service-Quality** and **Servers' Efficiency must** be traded off against each other.

For example, **M/M/1** (single-server queue): **91%** server's utilization goes with

$$\textbf{Congestion Index} = \frac{E[Wait]}{E[Service]} = \textbf{10},$$

and only 9% of the customers are served immediately upon arrival.

Prerequisite II: Models (Diffusion/QED's Q's)

Traditional Queueing Theory predicts that **Service-Quality** and **Servers' Efficiency** **must** be traded off against each other.

For example, **M/M/1** (single-server queue): **91%** server's utilization goes with

$$\textbf{Congestion Index} = \frac{E[Wait]}{E[Service]} = \mathbf{10},$$

and only 9% of the customers are served immediately upon arrival.

Yet, **heavily-loaded** queueing systems with **Congestion Index = 0.1** (Waiting one order of magnitude less than Service) are prevalent:

- **Call Centers**: Wait **"seconds"** for **minutes** service;
- **Transportation**: Search **"minutes"** for **hours** parking;
- **Hospitals**: Wait **"hours"** in ED for **days** hospitalization in IW's;

Prerequisite II: Models (Diffusion/QED's Q's)

Traditional Queueing Theory predicts that **Service-Quality** and **Servers' Efficiency** **must** be traded off against each other.

For example, **M/M/1** (single-server queue): **91%** server's utilization goes with

$$\textbf{Congestion Index} = \frac{E[Wait]}{E[Service]} = \mathbf{10},$$

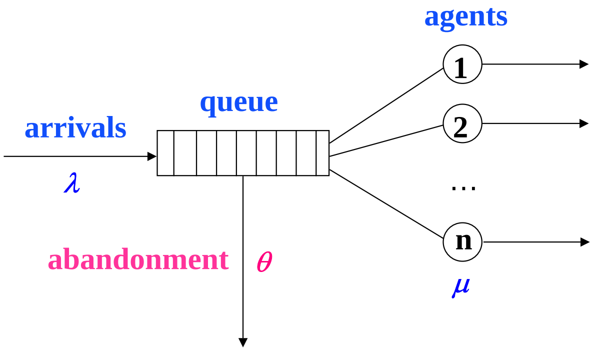
and only 9% of the customers are served immediately upon arrival.

Yet, **heavily-loaded** queueing systems with **Congestion Index = 0.1** (Waiting one order of magnitude less than Service) are prevalent:

- **Call Centers**: Wait **"seconds"** for **minutes** service;
- **Transportation**: Search **"minutes"** for **hours** parking;
- **Hospitals**: Wait **"hours"** in ED for **days** hospitalization in IW's;

and, moreover, a significant fraction are not delayed in queue. (For example, in well-run call-centers, **50%** served "immediately", along with over **90%** agents' utilization, is not uncommon) **?** **QED**

The Basic Staffing Model: Erlang-A (M/M/N + M)



Erlang-A **(Palm 1940's)** = **Birth & Death Q, with parameters:**

- λ – **Arrival** rate (Poisson)
- μ – **Service** rate (Exponential; $E[S] = \frac{1}{\mu}$)
- θ – **Patience** rate (Exponential, $E[\text{Patience}] = \frac{1}{\theta}$)
- n – Number of **Servers** (Agents).

Testing the Erlang-A Primitives

- **Arrivals**: Poisson?
- **Service-durations**: Exponential?
- **(Im)Patience**: Exponential?

Testing the Erlang-A Primitives

- **Arrivals**: Poisson?
- **Service-durations**: Exponential?
- **(Im)Patience**: Exponential?

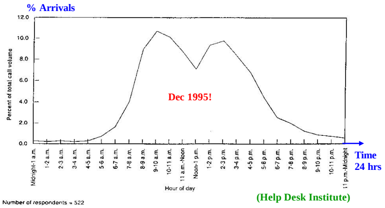
- Primitives independent (eg. Impatience and Service-Durations)?
- Customers / Servers Homogeneous?
- Service discipline FCFS?
- . . . ?

Validation: Support? Refute?

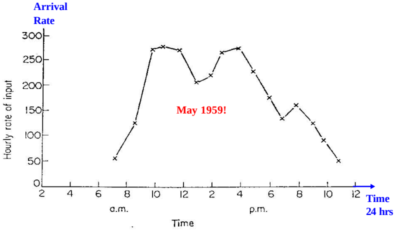
Arrivals to Service

Arrival-Rates to Three Call Centers

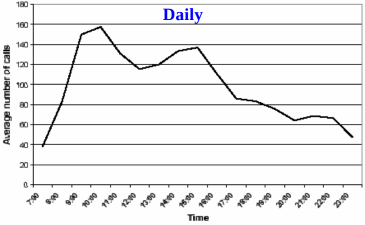
Dec. **1995** (U.S. 700 Helpdesks)



May **1959** (England)



November **1999** (Israel)



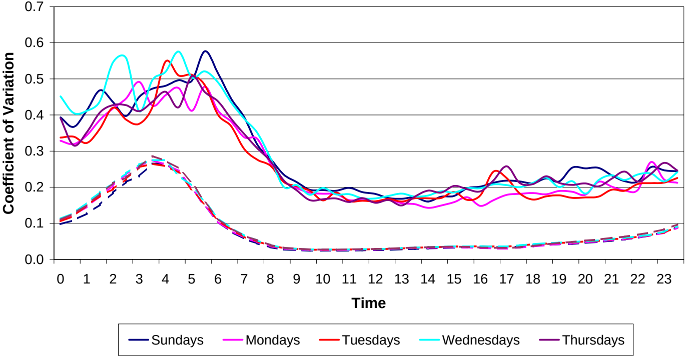
Random Arrivals "must be" (Axiomatically)

Time-Inhomogeneous Poisson

Arrivals to Service: only Poisson-Relatives

Arrival-Counts: Coefficient-of-Variation (CV), per 30 min.

Israeli-Bank Call-Center, 263 regular days (4/2007 - 3/2008)

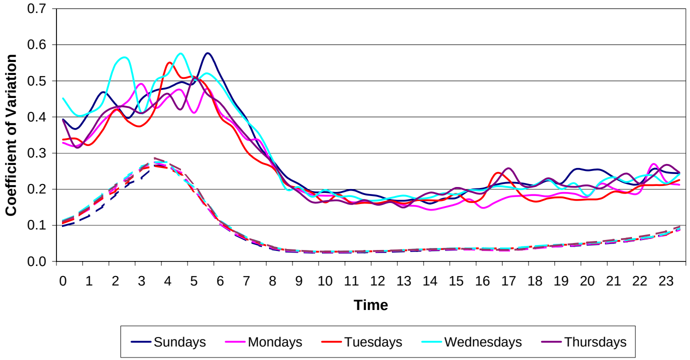


- **Poisson CV** (Dashed Line) = $1/\sqrt{\text{mean arrival-rate}}$
- Poisson CV's $\ll$ **Sampled CV's** (Solid) $\Rightarrow$ **Over-Dispersion**

Arrivals to Service: only Poisson-Relatives

Arrival-Counts: Coefficient-of-Variation (CV), per 30 min.

Israeli-Bank Call-Center, 263 regular days (4/2007 - 3/2008)

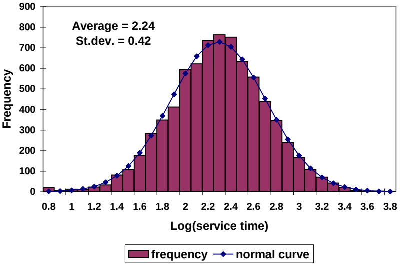


- **Poisson CV** (Dashed Line) = $1/\sqrt{\text{mean arrival-rate}}$
- Poisson CV's $\ll$ **Sampled CV's** (Solid) $\Rightarrow$ **Over-Dispersion**

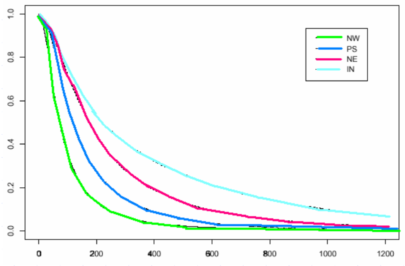
$\Rightarrow$ **Modeling** (Poisson-Mixture) of and **Staffing** ($> \sqrt{\cdot}$) against **Time-Varying Over-Dispersed** Arrivals (with **S. Maman and S. Zeltyn**)

Service Durations: LogNormal Prevalent

Israeli Bank Log-Histogram



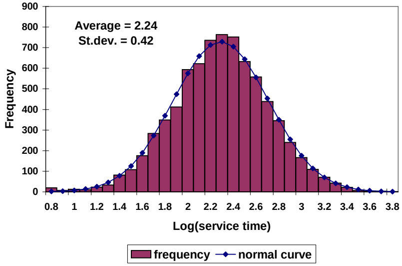
Service-Classes Survival-Functions



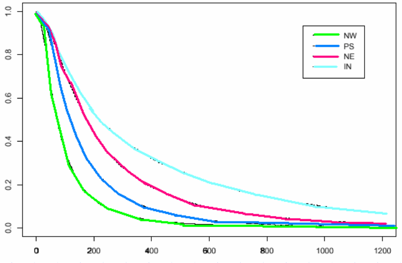
- **New** Customers: **2** min (NW);
- **Regulars**: **3** min (PS);
- **Stock**: **4.5** min (NE);
- Tech-Support: **6.5** min (IN).

Service Durations: LogNormal Prevalent

Israeli Bank Log-Histogram



Service-Classes Survival-Functions

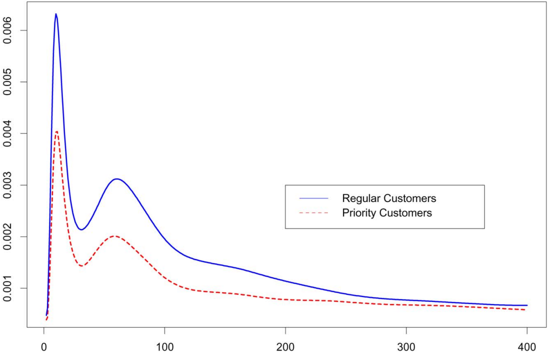


- **New** Customers: **2** min (NW);
- **Regulars**: **3** min (PS);
- **Stock**: **4.5** min (NE);
- Tech-Support: **6.5** min (IN).

▶ Service Durations are **LogNormal (LN)** and **Heterogeneous**

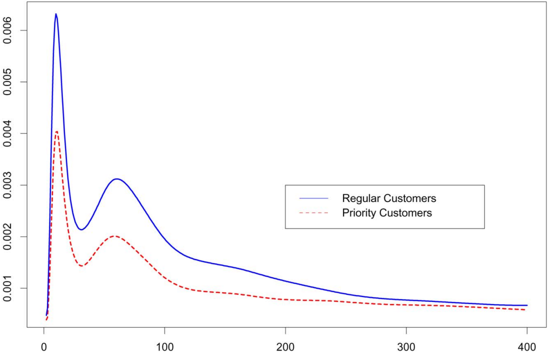
(Im)Patience while Waiting (Palm 1943-53)

Hazard Rate of (Im)Patience Distribution $\propto$ Irritation
Regular over VIP Customers – Israeli Bank



(Im)Patience while Waiting (Palm 1943-53)

Hazard Rate of (Im)Patience Distribution $\propto$ Irritation
Regular over VIP Customers – Israeli Bank



- **VIP** Customers are **more Patient** (Needy)
- **Peaks** of abandonment at times of **Announcements**
- Stat. Challenge: **Un-Censoring** - requires **Call-by-Call Data**

Erlang-A: Practical Relevance?

Experience:

- Arrival process **not pure Poisson** (time-varying, σ^2 too large)
- Service times **not Exponential** (typically close to LogNormal)
- Patience times **not Exponential** (various patterns observed).

Erlang-A: Practical Relevance?

Experience:

- Arrival process **not pure Poisson** (time-varying, σ^2 too large)
- Service times **not Exponential** (typically close to LogNormal)
- Patience times **not Exponential** (various patterns observed).

- Building Blocks need **not be independent** (eg. long wait associated with long service; with **M. Reich and Y. Ritov**)
- Customers and Servers **not homogeneous** (classes, skills)
- Customers return for service (after busy, abandonment; with **M. Gorfine and P. Khudiakov**)
- $\cdots$, and more.

Erlang-A: Practical Relevance?

Experience:

- Arrival process **not pure Poisson** (time-varying, σ^2 too large)
- Service times **not Exponential** (typically close to LogNormal)
- Patience times **not Exponential** (various patterns observed).

- Building Blocks need **not be independent** (eg. long wait associated with long service; with **M. Reich and Y. Ritov**)
- Customers and Servers **not homogeneous** (classes, skills)
- Customers return for service (after busy, abandonment; with **M. Gorfine and P. Khudiakov**)
- $\cdots$, and more.

Question: **Is Erlang-A Practically Relevant?**

Answer, via **Fitting a Simple Model to a Complex Reality**

Erlang-A: Simple, but Not Too Simple

Natural Questions:

1. Fitting Erlang-A (with **O. Plonsky and S. Zeltyn**).
2. Why does it practically work? justify **robustness**.
3. When does it fail? chart **boundaries**.
4. Generalize essential features.

Erlang-A: Simple, but Not Too Simple

Natural Questions:

1. Fitting Erlang-A (with **O. Plonsky and S. Zeltyn**).
2. Why does it practically work? justify **robustness**.
3. When does it fail? chart **boundaries**.
4. Generalize essential features.

Answers via **Asymptotic Analysis**, as load- and staffing-levels increase, which reveals model-essentials:

- **E**fficiency-**D**riven (**ED**) regime: Fluid models (deterministic)
- **Q**uality- and **E**fficiency-**D**riven (**QED**): Diffusion refinements.

Erlang-A: Simple, but Not Too Simple

Natural Questions:

1. Fitting Erlang-A (with **O. Plonsky and S. Zeltyn**).
2. Why does it practically work? justify **robustness**.
3. When does it fail? chart **boundaries**.
4. Generalize essential features.

Answers via **Asymptotic Analysis**, as load- and staffing-levels increase, which reveals model-essentials:

- **E**fficiency-**D**riven (**ED**) regime: Fluid models (deterministic)
- **Q**uality- and **E**fficiency-**D**riven (**QED**): Diffusion refinements.

Motivation: Moderate-to-large service systems (**100's - 1000's** servers), notably **Call-Centers**.

Results turn out **accurate** enough to also cover **<10** servers:

- Practically Important: Relevant to **Healthcare** (F. de Véricourt and O. Jennings; with **G. Yom-Tov; with Y. Marmor, S. Zeltyn**)
- Theoretically Justifiable: Gap-Analysis by **B. Zhang, J. van Leeuwaarden, B. Zwart**.

Operational Regimes: Conceptual Framework

$\boldsymbol{R}$: Offered Load

Def. $\boldsymbol{R}$ = Arrival-rate $\times$ Average-Service-Time = $\frac{\lambda}{\mu}$

eg. $\boldsymbol{R}$ = 25 calls/min. $\times$ 4 min./call = **100**

$\boldsymbol{N}$ = #Agents **?**

Operational Regimes: Conceptual Framework

***R*: Offered Load**

Def. $\boldsymbol{R}$ = Arrival-rate $\times$ Average-Service-Time = $\frac{\lambda}{\mu}$

eg. $\boldsymbol{R}$ = 25 calls/min. $\times$ 4 min./call = **100**

$\boldsymbol{N}$ = #Agents **?**

QD Regime: $\boxed{\boldsymbol{N} \approx \boldsymbol{R} + \delta \boldsymbol{R}}$, $\quad \mathbf{0.1} < \delta < \mathbf{0.25}$ $\quad$ (eg. $\boldsymbol{N} = \mathbf{115}$)

- Framework developed in **O. Garnett**'s MSc thesis
- Rigorously: $(\boldsymbol{N} - \boldsymbol{R})/\boldsymbol{R} \to \delta$, as $\boldsymbol{N}, \lambda \uparrow \infty$, with μ fixed.
- Performance: Delays are rare events

Operational Regimes: Conceptual Framework

R: Offered Load

Def. R = Arrival-rate $\times$ Average-Service-Time = $\frac{\lambda}{\mu}$

eg. R = 25 calls/min. $\times$ 4 min./call = **100**

N = #Agents **?**

QD Regime: $N \approx R + \delta R$, $\quad 0.1 < \delta < 0.25 \quad$ (eg. $N = 115$)

- Framework developed in **O. Garnett**'s MSc thesis
- Rigorously: $(N - R)/R \to \delta$, as $N, \lambda \uparrow \infty$, with μ fixed.
- Performance: Delays are rare events

ED Regime: $N \approx R - \gamma R$, $\quad 0.1 < \gamma < 0.25 \quad$ (eg. $N = 90$)

- Essentially **all** customers are delayed
- Wait same order as service-time; γ% Abandon (10-25%).

Operational Regimes: Conceptual Framework

R: Offered Load

Def. $\boldsymbol{R}$ = Arrival-rate $\times$ Average-Service-Time = $\frac{\lambda}{\mu}$

eg. $\boldsymbol{R}$ = 25 calls/min. $\times$ 4 min./call = **100**

$\boldsymbol{N}$ = #Agents **?**

QD Regime: $\boldsymbol{N \approx R + \delta R}$, $\quad \mathbf{0.1} < \delta < \mathbf{0.25} \quad$ (eg. $\boldsymbol{N} = \mathbf{115}$)

- Framework developed in **O. Garnett**'s MSc thesis
- Rigorously: $(\boldsymbol{N} - \boldsymbol{R})/\boldsymbol{R} \longrightarrow \delta$, as $\boldsymbol{N}, \boldsymbol{\lambda} \uparrow \infty$, with μ fixed.
- Performance: Delays are rare events

ED Regime: $\boldsymbol{N \approx R - \gamma R}$, $\quad \mathbf{0.1} < \gamma < \mathbf{0.25} \quad$ (eg. $\boldsymbol{N} = \mathbf{90}$)

- Essentially **all** customers are delayed
- Wait same order as service-time; γ% Abandon (10-25%).

QED Regime: $\boldsymbol{N \approx R + \beta\sqrt{R}}$, $\quad -\mathbf{1} < \beta < +\mathbf{1} \quad$ (eg. $\boldsymbol{N} = \mathbf{100}$)

- Erlang 1913-24, **Halfin & Whitt** 1981 (for Erlang-C)
- %Delayed between 25% and 75%
- E[Wait] $\propto \frac{1}{\sqrt{N}} \times$ E[Service] (**sec vs. min**); 1-5% Abandon.

Operational Regimes: Rules-of-Thumb, with **S. Zeltyn**

Constraint	P{Ab}		E[W]		P{W > T}	
Offered Load	Tight 1-10%	Loose $\geq 10\%$	Tight $\leq 10\%\mathrm{E}[\tau]$	Loose $\geq 10\%\mathrm{E}[\tau]$	Tight $0 \leq T \leq 10\%\mathrm{E}[\tau]$ $5\% \leq \alpha \leq 50\%$	Loose $T \geq 10\%\mathrm{E}[\tau]$ $5\% \leq \alpha \leq 50\%$
Small (10's)	QED	QED	QED	QED	QED	QED
Moderate-to-Large (100's-1000's)	QED	ED, QED	QED	ED, QED if $\tau \stackrel{d}{=} \exp$	QED	ED+QED

Operational Regimes: Rules-of-Thumb, with **S. Zeltyn**

Constraint	P{Ab}		E[W]		P{W > T}	
Offered Load	Tight 1-10%	Loose $\geq 10\%$	Tight $\leq 10\%\mathrm{E}[\tau]$	Loose $\geq 10\%\mathrm{E}[\tau]$	Tight $0 \leq T \leq 10\%\mathrm{E}[\tau]$ $5\% \leq \alpha \leq 50\%$	Loose $T \geq 10\%\mathrm{E}[\tau]$ $5\% \leq \alpha \leq 50\%$
Small (10's)	QED	QED	QED	QED	QED	QED
Moderate-to-Large (100's-1000's)	QED	ED, QED	QED	ED, QED if $\tau \overset{d}{=} \exp$	QED	ED+QED

ED: $N \approx R - \gamma R \quad (0.1 \leq \gamma \leq 0.25\,).$

QD: $N \approx R + \delta R \quad (0.1 \leq \delta \leq 0.25\,).$

QED: $N \approx R + \beta\sqrt{R} \quad (-1 \leq \beta \leq 1\,).$

ED+QED: $N \approx (1 - \gamma)R + \beta\sqrt{R} \quad (\gamma, \beta$ as above).

Operational Regimes: Rules-of-Thumb, with **S. Zeltyn**

Constraint	P{Ab}		E[W]		P{W > T}	
Offered Load	Tight 1-10%	Loose $\geq 10\%$	Tight $\leq 10\%\mathrm{E}[\tau]$	Loose $\geq 10\%\mathrm{E}[\tau]$	Tight $0 \leq T \leq 10\%\mathrm{E}[\tau]$ $5\% \leq \alpha \leq 50\%$	Loose $T \geq 10\%\mathrm{E}[\tau]$ $5\% \leq \alpha \leq 50\%$
Small (10's)	QED	QED	QED	QED	QED	QED
Moderate-to-Large (100's-1000's)	QED	ED, QED	QED	ED, QED if $\tau \stackrel{d}{=} \exp$	QED	ED+QED

ED: $N \approx R - \gamma R \quad (0.1 \leq \gamma \leq 0.25\,)$.

QD: $N \approx R + \delta R \quad (0.1 \leq \delta \leq 0.25\,)$.

QED: $N \approx R + \beta\sqrt{R} \quad (-1 \leq \beta \leq 1\,)$.

ED+QED: $N \approx (1 - \gamma)R + \beta\sqrt{R} \quad (\gamma, \beta$ as above).

WFM: How to determine specific staffing level N ? e.g. β.

QED Theory (Erlang '13; Halfin-Whitt '81; Garnett MSc; Zeltyn PhD)

Consider a sequence of **steady-state** M/M/N + G queues, $N = 1, 2, 3, \ldots$

Then the following points of view are **equivalent**, as $N \uparrow \infty$:

- **QED** $\quad \%\{\text{Wait} > 0\} \approx \alpha, \qquad 0 < \alpha < 1$;
- **Customers** $\quad \%\{\text{Abandon}\} \approx \dfrac{\gamma}{\sqrt{N}}, \qquad 0 < \gamma$;
- **Agents** $\quad \text{OCC} \approx 1 - \dfrac{\beta + \gamma}{\sqrt{N}} \qquad -\infty < \beta < \infty$;
- **Managers** $\quad N \approx R + \beta\sqrt{R}, \qquad R = \lambda \times \text{E(S)}$ not small;

QED Theory (Erlang '13; Halfin-Whitt '81; Garnett MSc; Zeltyn PhD)

Consider a sequence of **steady-state** M/M/N + G queues, $N = 1, 2, 3, \ldots$
Then the following points of view are **equivalent**, as $N \uparrow \infty$:

- **QED** $\quad \%\{\text{Wait} > 0\} \approx \alpha, \qquad 0 < \alpha < 1\,;$
- **Customers** $\quad \%\{\text{Abandon}\} \approx \dfrac{\gamma}{\sqrt{N}}, \qquad 0 < \gamma\,;$
- **Agents** $\quad \text{OCC} \approx 1 - \dfrac{\beta + \gamma}{\sqrt{N}} \qquad -\infty < \beta < \infty\,;$
- **Managers** $\quad N \approx R + \beta\sqrt{R}\,, \quad R = \lambda \times \text{E(S)}$ not small;

- **QED performance**: **Laplace Method** (asymptotics of integrals).
- **Parameters**: Arrivals and Staffing - β, Services - μ, (Im)Patience - $g(0)$ = **patience density at the origin**.

QED Approximations: Some Examples

G – patience distribution,

g_0 – patience density at origin $\quad (g_0 = \theta,$ if $\exp(\theta))$.

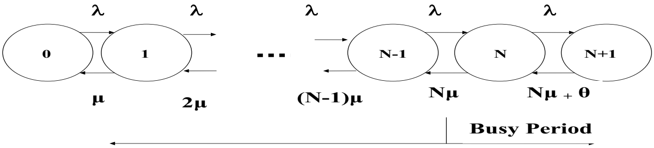
$$N = \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda}), \qquad -\infty < \beta < \infty .$$

$$\begin{aligned}
\mathrm{P}\{\mathrm{Ab}\} &\approx \frac{1}{\sqrt{N}} \cdot [h(\hat\beta) - \hat\beta] \cdot \left[\sqrt{\frac{\mu}{g_0}} + \frac{h(\hat\beta)}{h(-\beta)}\right]^{-1}, \\
\mathrm{P}\left\{W > \frac{T}{\sqrt{N}}\right\} &\approx \boxed{\left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat\beta)}{h(-\beta)}\right]^{-1}} \cdot \frac{\bar\Phi\left(\hat\beta + \sqrt{g_0\mu}\cdot T\right)}{\bar\Phi(\hat\beta)}, \\
\mathrm{P}\left\{\mathrm{Ab} \,\middle|\, W > \frac{T}{\sqrt{N}}\right\} &\approx \frac{1}{\sqrt{N}} \cdot \sqrt{\frac{g_0}{\mu}} \cdot \left[h\left(\hat\beta + \sqrt{g_0\mu}\cdot T\right) - \hat\beta\right].
\end{aligned}$$

Here

$$\begin{aligned}
\hat\beta &= \beta\sqrt{\frac{\mu}{g_0}} \\
\bar\Phi(x) &= 1 - \Phi(x), \\
h(x) &= \phi(x)/\bar\Phi(x), \quad \text{hazard rate of } N(0,1).
\end{aligned}$$

QED Intuition via Excursions: Busy-Idle Cycles



$Q(0) = N$: all servers busy, no queue.

Let $T_{N,N-1}$ = E[Busy Period] down-crossing $N \downarrow N-1$

$T_{N-1,N}$ = E[Idle Period] up-crossing $N-1 \uparrow N$)

Then $P(\text{Wait} > 0) = \frac{T_{N,N-1}}{T_{N,N-1} + T_{N-1,N}} = \left[1 + \frac{T_{N-1,N}}{T_{N,N-1}}\right]^{-1}$.

QED Intuition via Excursions: Asymptotics

Calculate $$T_{N-1,N} = \frac{1}{\lambda_N E_{1,N-1}} \sim \frac{1}{N\mu \times h(-\beta)/\sqrt{N}} \sim \frac{1}{\sqrt{N}} \cdot \frac{1/\mu}{h(-\beta)}$$

$$T_{N,N-1} = \frac{1}{N\mu\pi_+(0)} \sim \frac{1}{\sqrt{N}} \cdot \frac{\beta/\mu}{h(\delta)\,/\delta}, \qquad \delta = \beta\sqrt{\mu/\theta}$$

Both apply as $\sqrt{N}\,(1-\rho_N) \to \beta,\ -\infty < \beta < \infty.$

Hence, $$P(Wait > 0) \sim \left[1 + \frac{h(\delta)/\delta}{h(-\beta)/\beta}\right]^{-1}.$$

Process Limits (Queueing, Waiting)

- $\hat{Q}_N = \{\hat{Q}_N(t), t \geq 0\}$: stochastic process obtained by centering and rescaling:

$$\hat{Q}_N = \frac{Q_N - N}{\sqrt{N}}$$

- $\hat{Q}_N(\infty)$: stationary distribution of $\hat{Q}_N$

- $\hat{Q} = \{\hat{Q}(t), t \geq 0\}$: process defined by: $\hat{Q}_N(t) \xrightarrow{d} \hat{Q}(t)$.

$$\begin{tikzcd}
\hat{Q}_N(t) \arrow[r, "t \to \infty"] \arrow[d, "N \to \infty"'] & \hat{Q}_N(\infty) \arrow[d, "N \to \infty"] \\
\hat{Q}(t) \arrow[r, "t \to \infty"'] & Q(\infty)
\end{tikzcd}$$

Approximating (Virtual) Waiting Time

$$\hat{V}_N = \sqrt{N}\, V_N \Rightarrow \hat{V} = \left[\frac{1}{\mu}\,\hat{Q}\right]^+$$

Back to “Why does Erlang-A Work?"

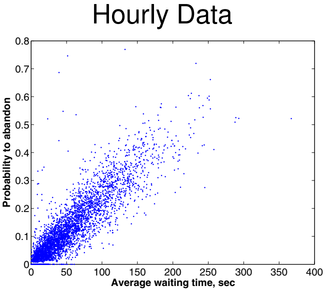
Theoretical Answer:

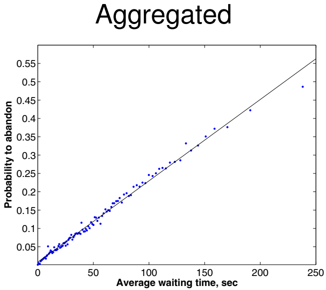
$$M_t^{?,J}/G/N_t + G \overset{d}{\approx} (M/M/N + M)_t \ , \quad t \geq 0.$$

- **General Patience**: Behavior at the origin is all that matters.
- **General Services**: Empirical insensitivity beyond the mean.
- **Time-Varying Arrivals**: Modified Offered-Load approximations.
- **Over-Dispersed Arrivals**: c-Staffing ($c > 1/2$).
- **Heterogeneous Customers**: 1-D state-collapse.

"Why does Erlang-A Work?" General Patience

Israeli Bank: Yearly Data





Theory:

Erlang-A: $\mathrm{P}\{\mathrm{Ab}\} = \theta \cdot \mathrm{E}[W_q]$;

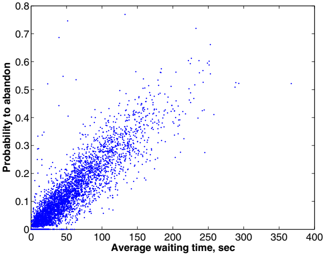
M/M/*N*+G: $\mathrm{P}\{\mathrm{Ab}\} \approx g(0) \cdot \mathrm{E}[W_q]$.

$g(0)$ = Patience-density at origin

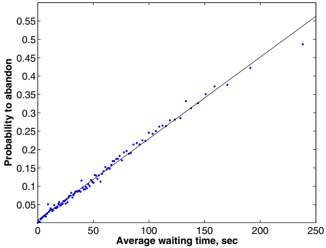
“Why does Erlang-A Work?" General Patience

Israeli Bank: Yearly Data

Hourly Data



Aggregated



Theory:

Erlang-A: $\mathrm{P}\{\mathrm{Ab}\} = \theta \cdot \mathrm{E}[W_q]$;

M/M/*N*+G: $\mathrm{P}\{\mathrm{Ab}\} \approx g(0) \cdot \mathrm{E}[W_q]$.

$g(0)$ = Patience-density at origin

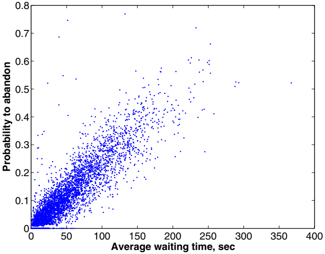
Recipe:

In both cases, use Erlang-A, with $\hat{\theta} = \widehat{\mathrm{P}\{\mathrm{Ab}\}} / \widehat{\mathrm{E}[W_q]}$ (slope above).

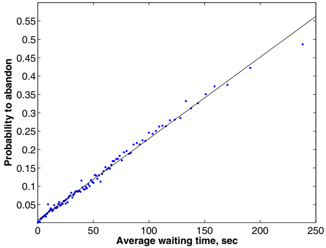
“Why does Erlang-A Work?" General Patience

Israeli Bank: Yearly Data

Hourly Data



Aggregated



Theory:

Erlang-A: $\mathrm{P}\{\mathrm{Ab}\} = \theta \cdot \mathrm{E}[W_q]$;

M/M/N+G: $\mathrm{P}\{\mathrm{Ab}\} \approx g(0) \cdot \mathrm{E}[W_q]$.

$g(0)$ = Patience-density at origin

Recipe:

In both cases, use Erlang-A, with $\hat{\theta} = \widehat{\mathrm{P}\{\mathrm{Ab}\}} / \widehat{\mathrm{E}[W_q]}$ (slope above).

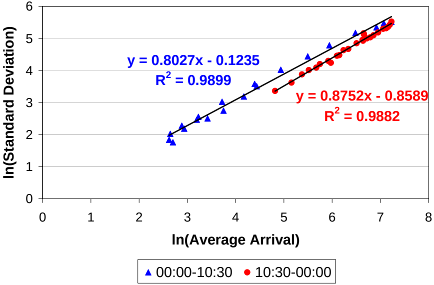
References on $g(0)$:

- Stationary M/M/N+GI, with **S. Zeltyn**
- Process G/GI/N+GI, with **P. Momcilovic**

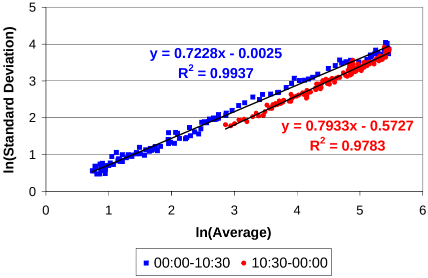
"Why does Erlang-A Work?" Over-Dispersion

ln(*STD*) vs. ln(*AVG*) (Israeli Bank, 4/2007-3/2008)

Tue-Wed, 30 min resolution



Tue-Wed, 5 min resolution



Significant linear relations (with **S. Aldor & P. Feigin**):

$$\ln(\mathrm{STD}) = c \cdot \ln(\mathrm{AVG}) + a$$

(Poisson: $\mathrm{STD} = \mathrm{AVG}^{1/2}$, hence $c = 1/2$, $a = 0$.)

Over-Dispersion: Random Arrival-Rates

Linear relation between ln(STD) and ln(AVG) gives rise to:

Poisson-Mixture (Doubly-Poisson, Cox) model for Arrivals:
Poisson(Λ) with **Random-Rate** of the form

$$\Lambda \;=\; \lambda \;+\; \lambda^{c} \cdot X, \quad c \leq 1\,;$$

Over-Dispersion: Random Arrival-Rates

Linear relation between ln(STD) and ln(AVG) gives rise to:

Poisson-Mixture (Doubly-Poisson, Cox) model for Arrivals:
Poisson(Λ) with **Random-Rate** of the form

$$\boldsymbol{\Lambda \;=\; \lambda \;+\; \lambda^{c} \cdot X, \quad c \leq 1\,;}$$

- c determines magnitude of over-dispersion (λ^c)
  $c = 1$, proportional to λ; $c \leq 1/2$, Poisson-level;
  - In **Call Centers**: $\boldsymbol{c \approx 0.75 - 0.85}$ (significant over-dispersion).
  - In **Emergency Departments**, $\boldsymbol{c \approx 0.5}$ (Poisson).

Over-Dispersion: Random Arrival-Rates

Linear relation between ln(STD) and ln(AVG) gives rise to:

Poisson-Mixture (Doubly-Poisson, Cox) model for Arrivals:
Poisson(Λ) with **Random-Rate** of the form

$$\Lambda = \lambda + \lambda^{c} \cdot X, \quad c \leq 1 ;$$

- c determines magnitude of over-dispersion (λ^{c})
  $c = 1$, proportional to λ; $c \leq 1/2$, Poisson-level;
  - In **Call Centers**: $c \approx 0.75 - 0.85$ (significant over-dispersion).
  - In **Emergency Departments**, $c \approx 0.5$ (Poisson).
- X random-variable with $E[X] = 0$ ($E[\Lambda] = \lambda$), capturing the magnitude of **stochastic deviation** from mean arrival-rate: under conventional Gamma prior (λ large), X can be taken Normal with std. derived from the intercept.

Over-Dispersion: Random Arrival-Rates

Linear relation between ln(STD) and ln(AVG) gives rise to:

Poisson-Mixture (Doubly-Poisson, Cox) model for Arrivals:
Poisson(Λ) with **Random-Rate** of the form

$$\Lambda \;=\; \lambda \;+\; \lambda^{c} \cdot X, \qquad c \leq 1\,;$$

- c determines magnitude of over-dispersion (λ^c)
  $c = 1$, proportional to λ; $c \leq 1/2$, Poisson-level;
  - In **Call Centers**: $c \approx 0.75 - 0.85$ (significant over-dispersion).
  - In **Emergency Departments**, $c \approx 0.5$ (Poisson).
- X random-variable with $E[X] = 0$ ($E[\Lambda] = \lambda$), capturing the magnitude of **stochastic deviation** from mean arrival-rate: under conventional Gamma prior (λ large), X can be taken Normal with std. derived from the intercept.

QED-c Regime: Erlang-A, with Poisson(Λ) arrivals, amenable to asymptotic analysis (with **S. Maman & S. Zeltyn**)

Over-Dispersion: The QED-c Regime

QED-c Staffing: Under offered-load $\boldsymbol{R} = \lambda \cdot \boldsymbol{E[S]}$,

$$\boldsymbol{N \;=\; R \;+\; \beta \cdot R^{c}} \;, \qquad \boldsymbol{0.5 < c < 1}$$

Performance measures (M/M/N + G):

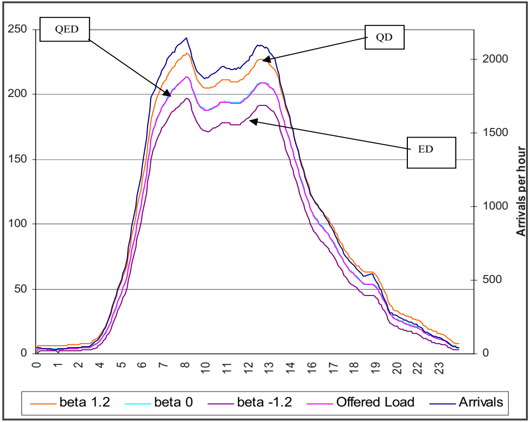
- Delay probability: $P\{W_q > 0\} \;\sim\; 1 - G(\beta)$
- Abandonment probability: $P\{Ab\} \;\sim\; \dfrac{E[X-\beta]_+}{n^{1-c}}$
- Average offered wait: $E[V] \;\sim\; \dfrac{E[X-\beta]_+}{n^{1-c} \cdot g_0}$
- Average actual wait: $E_{\Lambda,N}[W] \;\sim\; E_{\Lambda,N}[V]$

Why Does Erlang-A Work? Time-Varying Arrival Rates

Square-Root Staffing: $N_t = R_t + \beta\sqrt{R_t}\,, \quad -\infty < \beta < \infty$

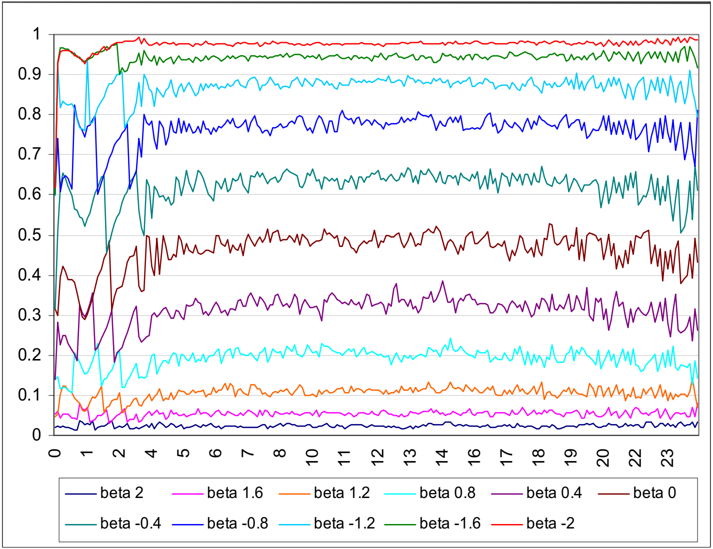
What is R_t, the **Offered-Load** at time t ? $\quad (\, R_t \neq \lambda_t \times \mathrm{E}[S] \,)$

Arrivals, Offered-Load and Staffing



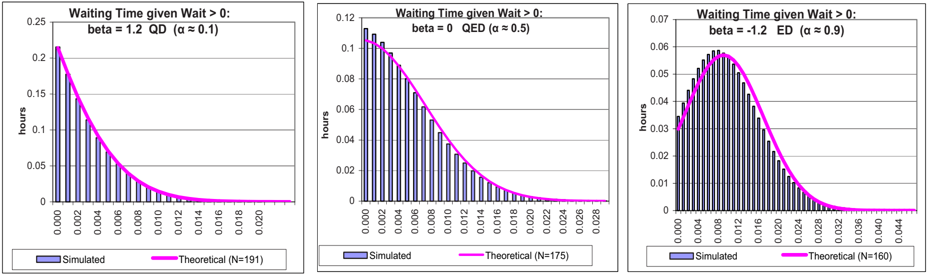
Time-Stable Performance of Time-Varying Systems

Delay Probability = As in the **Stationary Erlang-A** (Garnett)



Time-Stable Performance of Time-Varying Systems

Waiting Time, Given Waiting: Empirical vs. Theoretical Distribution



- **Empirical**: Simulate **time-varying** $M_t/M/N_t + M \;\; (\lambda_t, N_t = R_t + \beta\sqrt{R_t})$
- **Theoretical**: Naturally-corresponding **stationary** Erlang-A, with QED β-staffing
- **Generalizes** up to a station within a complex network (eg. Doctors in an Emergency Department).

What is the Offered-Load $R(t)$? Time-Varying Little

For $\boldsymbol{M_t/GI/N_t + GI}$, the **Offered-Load function**, $\{\boldsymbol{R(t),\ t \geq 0}\}$, is the **average number of customers (= busy servers)**, in a naturally corresponding $\boldsymbol{M_t/GI/\infty}$ queue (MOL = Modified Offered Load).

What is the Offered-Load $R(t)$? Time-Varying Little

For $\boldsymbol{M_t/GI/N_t + GI}$, the **Offered-Load function**, $\{\boldsymbol{R(t),\ t \geq 0}\}$, is the **average number of customers (= busy servers)**, in a naturally corresponding $\boldsymbol{M_t/GI/\infty}$ queue (MOL = Modified Offered Load).

Four (all useful) representations, capturing **"work before t"**:

$$R(t) = E[L(t)] = \int_{-\infty}^{t} \lambda(u) \cdot P(S > t - u)du = E\Big[A(t) - A(t - S)\Big] =$$

$$= E\left[\int_{t-S}^{t} \lambda(u)du\right] = E[\lambda(t - S_e)] \cdot E[S] \,.$$

- $\{L(t),\ t \geq 0\}$ is the number of customers (= busy-servers) in the above-mentioned $M_t/GI/\infty$ queue (hence **time-varying Little**);
- $\{A(t),\ t \geq 0\}$ is the Arrival-Process;
- $\boldsymbol{S}$ ($\boldsymbol{S_e}$) is a generic Service-Time (Residual Service-Time).

What is the Offered-Load $R(t)$? Time-Varying Little

For $\boldsymbol{M_t/GI/N_t + GI}$, the **Offered-Load function**, $\{\boldsymbol{R(t),\ t \geq 0}\}$, is the **average number of customers (= busy servers)**, in a naturally corresponding $\boldsymbol{M_t/GI/\infty}$ queue (MOL = Modified Offered Load).

Four (all useful) representations, capturing **"work before t"**:

$$R(t) = E[L(t)] = \int_{-\infty}^{t} \lambda(u) \cdot P(S > t-u)du = E\Big[A(t) - A(t-S)\Big] =$$
$$= E\left[\int_{t-S}^{t} \lambda(u)du\right] = E[\lambda(t - S_e)] \cdot E[S] .$$

- $\{L(t),\ t \geq 0\}$ is the number of customers (= busy-servers) in the above-mentioned $M_t/GI/\infty$ queue (hence **time-varying Little**);
- $\{A(t),\ t \geq 0\}$ is the Arrival-Process;
- $\boldsymbol{S}$ ($\boldsymbol{S_e}$) is a generic Service-Time (Residual Service-Time).

- **Stationary models**: $\lambda(t) \equiv \lambda$ then $R(t) \equiv \boldsymbol{\lambda/\mu}$.

- **QED-c**: $\boldsymbol{N_t = R_t + \beta R_t^c}$, $1/2 < c < 1$; ($c = 1$ separate analysis).

The Technion SEE Center / Laboratory

Data-Based Service Science / Engineering

Technion
S.E.E
Center
Service Enterprise Engineering

Technion SEE = Service Enterprise Engineering

SEELab: **Hub for data-based research and teaching**

- **History**: I.E. Dean, **B. Golany**, recruited **Hal and Inge Marcus**.
  - **Technion (parallel to Penn)**: In 2007, w/ **P. Feigin, V. Trofimov**.
  - **Wharton**: L. Brown, N. Gans, H. Shen (UNC).
  - **industry**

Technion SEE = Service Enterprise Engineering

SEELab: Hub for data-based research and teaching

- **History**: I.E. Dean, **B. Golany**, recruited **Hal and Inge Marcus**.
  - **Technion (parallel to Penn)**: In 2007, w/ **P. Feigin, V. Trofimov**.
  - **Wharton**: L. Brown, N. Gans, H. Shen (UNC).
  - **industry** (partial list):
    - U.S. Bank: **2.5 years, 220M calls, 40M by 1000 agents**.
    - Israeli Cellular: **2.5 years, 110M calls, 25M calls by 750 agents**.
    - Israeli Bank: **from January 2010, daily-deposit at a SEESafe**.
    - Israeli Hospital: **4 years, 1000 beds; 8 ED's - Sinreich's data**.

Technion SEE = Service Enterprise Engineering

SEELab: **Hub for data-based research and teaching**

- **History**: I.E. Dean, **B. Golany**, recruited **Hal and Inge Marcus**.
  - **Technion (parallel to Penn)**: In 2007, w/ **P. Feigin, V. Trofimov**.
  - **Wharton**: L. Brown, N. Gans, H. Shen (UNC).
  - **industry** (partial list):
    - U.S. Bank: **2.5 years, 220M calls, 40M by 1000 agents**.
    - Israeli Cellular: **2.5 years, 110M calls, 25M calls by 750 agents**.
    - Israeli Bank: **from January 2010, daily-deposit at a SEESafe**.
    - Israeli Hospital: **4 years, 1000 beds; 8 ED's - Sinreich's data**.

SEEStat: **Environment for graphical EDA in real-time**

- **Universal Design, Universal Access, Real-Time Response**.

Technion SEE = Service Enterprise Engineering

SEELab: Hub for data-based research and teaching

- **History**: I.E. Dean, **B. Golany**, recruited **Hal and Inge Marcus**.
  - **Technion (parallel to Penn)**: In 2007, w/ **P. Feigin, V. Trofimov**.
  - **Wharton**: L. Brown, N. Gans, H. Shen (UNC).
  - **industry** (partial list):
    - U.S. Bank: **2.5 years, 220M calls, 40M by 1000 agents**.
    - Israeli Cellular: **2.5 years, 110M calls, 25M calls by 750 agents**.
    - Israeli Bank: **from January 2010, daily-deposit at a SEESafe**.
    - Israeli Hospital: **4 years, 1000 beds; 8 ED's - Sinreich's data**.

SEEStat: Environment for graphical EDA in real-time

- **Universal Design, Universal Access, Real-Time Response**.
- **Clean DBs**: Operational-history of **individual transactions**.
- **Interface**: At varying resolutions (seconds, minutes, hours, days, months), graphically, in real-time.
- **Tools**: Classic Stat, and beyond (Survival Analysis, Distribution Fitting, Mixtures, Smoothing, . . .)

Technion SEE = Service Enterprise Engineering

SEELab: Hub for data-based research and teaching

- **History**: I.E. Dean, **B. Golany**, recruited **Hal and Inge Marcus**.
  - **Technion (parallel to Penn)**: In 2007, w/ **P. Feigin, V. Trofimov**.
  - **Wharton**: L. Brown, N. Gans, H. Shen (UNC).
  - **industry** (partial list):
    - U.S. Bank: **2.5 years, 220M calls, 40M by 1000 agents**.
    - Israeli Cellular: **2.5 years, 110M calls, 25M calls by 750 agents**.
    - Israeli Bank: **from January 2010, daily-deposit at a SEESafe**.
    - Israeli Hospital: **4 years, 1000 beds; 8 ED's - Sinreich's data**.

SEEStat: Environment for graphical EDA in real-time

- **Universal Design, Universal Access, Real-Time Response**.
- **Clean DBs**: Operational-history of **individual transactions**.
- **Interface**: At varying resolutions (seconds, minutes, hours, days, months), graphically, in real-time.
- **Tools**: Classic Stat, and beyond (Survival Analysis, Distribution Fitting, Mixtures, Smoothing, . . .)

SEEServer: Free for academic use

Register, then access (presently) U.S. Bank and Small Israeli Bank.