Scientific abstract – ***Appointment-driven Resource Networks (ARNets): Data-Based Modeling, Analysis and Design, with Applications to Healthcare and Judicial Services***

The proposed research focuses on complex service networks that are driven by appointments of their customers and servers, or more broadly by scheduling-via-appointments of network resources. I refer to such networks as ARNets (Appointment-driven Resource Networks); my primary motivation for studying them originated in the healthcare and judiciary systems, where most services are by appointment. Yet, despite appointment scheduling (algorithms and software) being prevalent operational practice, ARNets are still plugged with congestion at levels that cry for relief. This raises challenges and opportunities, which I propose to address by marrying mathematics with data through ARNet modelling and analysis; or specifically, by developing a theory that will narrow its gap from actual practice—a theory that must hence reach beyond and differ from the state-of-the art. To this end, I formed close data-based partnerships with a leading cancer center and with the Israeli judiciary, which will ensure that major practical aspects, specifically, complexity and uncertainty, will be accommodated by my ARNet models and their analysis; this grants me the courage to hope for a real positive impact on ARNets' operational practice.

Operationally, appointment scheduling aims at matching demand and capacity which, in reality, has taken a deterministic perspective. For example, schedules use average punctuality and service-durations and, in both supporting theory and its practice, they rely on perfect adherence to a deterministic schedule. However, service processes (recall our hospitals and courts) are often inherently complex and uncertain. Under such premises, scheduling problems are notoriously hard to solve and, moreover, existing algorithms (too) often do not scale to practically relevant problem sizes.

I propose to develop data-based mathematical models, enrich them with human factors, analyze/approximate them asymptotically and computationally (since such models are unlikely to be tractable as is), validate my models against the systems that originated them, and finally develop implementable scheduling algorithms that are capable of capturing ARNets' complexity and stochasticity. I shall focus, in particular, on asymptotic analysis and design principles of stochastic ARNets, and, to the extent possible, I shall incorporate, or at least touch on, central features such as time-variability, multiple resources per activity or multiple activities per resource, mix of heavy-traffic regimes and time-scales and fairness.

My analysis will be carried out at various levels of modeling-details: from crude fluid models to the finer diffusion approximations. An essential research aspect is empirical support, utilizing a vast collection of data repositories from hospitals, courts, transportation and online services. A unique data provider is a Real-Time Location System (RTLS), deployed at the Dana Farber Cancer Institute (DFCI) in Boston. Via 900 sensors across 8 floors, it tracks continuously and automatically the exact location of patients and providers (over 800 patients and 100s of providers daily). Moreover, this RTLS data is augmented by DFCI's EHR system and its appointments log. Our four years of data thus track operationally the planned vs. the actual, and we seek to improve the former and lessen its gap from the latter. No less importantly, such rich data enables theory-validation, reproducible-research and, finally, data mining of novel research problems and directions.

1 Scientific Background: Appointment Scheduling (AppS)

Motivation via Exploratory Data Analysis (EDA): Consider Figures 1 and 2 with their animations (that are accessible via the links within their captions). These figures depict *real* data of customer-flows in two service systems, healthcare and justice, which are driven by appointments. I shall refer to such systems as Appointment-driven Resource Networks (ARNets). Specifically, Figure 1 is a snapshot at 11:17 on day XX-YYY-2014, of all patient locations at the infusion unit of the Dana Farber Cancer Institute (DFCI), on the 9^{th}-floor of their Yawkey Center in Boston. Figure 2 is based on 2,000 civil cases that were filed at Israeli Magistrate and District courts and resolved between 2008-2012. One can in fact dig way deeper: Figure 3 is a snapshot of the *complete appointment-book* of patients at DFCI, created via their system of Appointment Scheduling (AppS). It covers around 900 patients daily, who are treated by about 350 medical personnel (in particular doctors and nurses) in 100(20) exam(consultation) rooms and over 150 infusion spaces. Yet, most significantly, our data in fact covers the *actual* precise locations of *all* patients and resources at DFCI during almost 4 years: this data has been gathered continuously and automatically, since November 2013, through a Real-Time Location System (RTLS), that collects data via 900 sensors scattered over 8 clinical floors. For example, Figure 4 complements Figure 1 by depicting jointly, in addition to patients, also hospital flows of doctors, nurses (infusion, blood-draw) and room-status.

Continuing with my EDA, Figure 5 exhibits schematically the most common daily care-paths at DFCI: they constitute all or some phases among blood-test (which is a multi-server queue), doctor-exam (single server) and infusion treatment (again multi-server). Several (or many) care-paths typically combine into a treatment plan: for example, Figure 6 records the exact location/status of a single patient, during her or his 61 DFCI visits within a 6-month period. Additional features of our complex ARNets include highly variable work-profiles of resources (see Figure 7 that also alludes to high Judicial Workload (JW), and Figure 8 that underscores the importance of measuring performance appropriately); or a resource (e.g. infusion nurse) can be simultaneously in charge of several activities; or several resources (doctor, nurse, equipment) must cater to a single activity (patient exam). Resources also operate over different time scales (e.g. nurse services take minutes while their patients occupy an infusion chair for hours). And there are often non-operational factors that must be accounted for in AppS (notably fairness in healthcare and justice).

Why start my proposal with this lengthy EDA? Because it reveals many characteristics of ARNets that are beyond existing research; and this gap between practical needs and theoretical offerings, with the importance and challenge to reduce it, motivated the current proposal. I shall now elaborate.

QED ARNets, ideally: The practice of AppS is prevalent in service systems. To wit, in healthcare, justice and ample other systems, the majority of services operate under scheduled appointments. Such ARNets thrive to be Quality- and Efficiency-Driven (QED) [3, 17, 19, 39] which entails that waiting times should be overall short: waiting of both customers to servers — namely high service-quality, and servers to customers — high resource-efficiency. This QED regime, when feasible, requires carefully matching supply and de-

mand, and this is the role of AppS systems. However, mismatches are common, as manifested through experiences that greatly deviate from schedule (e.g. long delays). This could have profound negative consequences for both customers (e.g. "justice delayed is justice denied") and service providers (e.g. taking it to the extreme: on February 8, 2011, a 54-year-old Jerusalem Municipal Court judge committed suicide, after leaving a note that he "could no longer cope with the immense workload").

State-of-the-art Theory: AppS problems are notoriously difficult and, consequently, solution approaches have so far been developed only for specialized cases; notably for settings where it is assumed that service durations are deterministic and punctuality is perfect [32, 36], or for settings where service durations are uncertain but where a single server is in operation [15]. Still the literature on AppS is vast and extensively surveyed (e.g. [1, 11, 13, 21]). In particular, outpatient scheduling is reviewed in [8], where it is remarked that, in practice, schedules are commonly based on mean procedure times. See also [45–47], then [16, 20, 36] that focus on cancer care, and [28] for recent relevant AppS research.

Themes of the following research are directly relevant to ours: [18] raises challenges in AppS of complex outpatient clinics (DFCI is such); [10] is a unique operations-research example of AppS for judges; and [26] develops a data-driven model for an appointment-generated arrival process. More specifically, my proposed research builds on three pillars. First, *data-based infinite-server heuristics* that continue [28] (see [43, 44] for previous successes of these heuristics). Second, *bottom-up QED asymptotics of many-server AppS-driven queues* [22, 39] (e.g infusion beds), in the steps of [2] that analyzed AppS for a single-server queue in conventional heavy-traffic. And third, a *top-down mathematical framework for ARNets*, which will flexibly accommodate their complex features (e.g. one/many to one/many matching of activities and resources). Depending on progress, my research might expand to touch upon issues beyond the purely operational. Two that stand out, given my focus on healthcare and justice, are AppS that are fair to or entertain preferences of its users (customers or serves); rare such examples are [14, 34] for fairness and [27] for preferences.

Why be optimistic about narrowing the gap between practice and theory? Prerequisites are relevant research and receptive industry. For the former, consider the six histograms in Figures 9a-9f: each fits some known distribution to a stochastic component of the DFCI ARNet, notably punctuality of patients and doctors. My take of this figure, plus accessibility to data in unprecedented quantity, quality and granularity, is the optimism that simple+relevant models can "serve well" the complex reality of ARNets. Regarding receptive industry, Healthcare OR/OM/IE has already been progressing in large fast strides. As for the less-researched "Production of Justice", scholars of the judiciary have started to argue that court congestion is mostly a consequence of mismanagement, outdated procedures and judicial passivity [9, 12, 30, 31, 33, 41]. Accordingly, budget increases must be guided by research to cure court congestion [7, 24, 25, 35, 48]. And such research is my goal here: its three successful starts (based on the above three pillars and further outlined below), together with the quality and diversity of my research parters, solidify my optimism.

2 Research Objectives & Expected Significance

To address the challenges and opportunities raised, one must advance state-of-the-art understanding – empirical and theoretical – of ARNets.[1] I propose to do so by marrying mathematics with data through models. My anchor point will be the high-resolution data repositories at the Technion SEELab (described in Appendix A of [3]), with special focus on data from DFCI (RTLS) and the Israeli Judiciary – both being appointment-driven. To start, I shall expand my EDA, which will improve the fit of my ARNet building-blocks to practice; this will guide the development and calibration of ARNets that support AppS; which will then enable the development of data-driven algorithms, and their validation under practical circumstances. More concretely, I propose data-based theoretical research that will take into account:

Uncertainty. The proposed models and algorithms will take into account the stochastic nature of ARNets: for example, in time-evolution of appointment books (changes, cancellations); through model building-blocks (e.g. punctuality and service-durations – recall Figures 9a-9f); or through customer flows/paths (e.g. long court delays could result in arbitration, or a negative blood-test could result in cancellation of chemotherapy).

Complexity. ARNet models will be gradually adapted to the complex reality of ARNet systems. An example is that several appointments are made simultaneously: one either schedules several future visits to a facility (infusion treatments or court hearings; recall Figure 6), or schedules several services during a single visit (the facility thus being a network of resources; recall Figures 1-4). Another instance is that one or several resources must be scheduled simultaneously to cater to a single or several services; e.g. a doctor-patient appointment at DFCI could require simultaneous availability of the doctor, the patient and the exam-room, but possibly also a nurse, a piece-of-equipment and an additional doctor – thus one activity requiring six ARNet resources. (Note that customers are viewed as resources – see §3.3 for the latter's definition.)

Realistic-scale. I aim to develop a scalable framework capable of dealing with the service environments encountered in practice, in particular those employing hundreds of servers and serving hundreds of customers daily (e.g. Figures 3-4).

Data. My models will be calibrated, and their findings validated and tested against SEELab's data repositories. Importantly, emphasis will be given to *reproducible* research, thus continuing the practice of [3, 29]. (See §6.2 in [3], on "data-based research" being a great opportunity but no less of a challenge, and the need for "Fundamental changes" – both within our IE/OR/OM communities as well as our potential data-partners.)

Methodologically, I shall integrate queueing theory (mostly asymptotic) and optimization to address the technical challenges underlying AppS. This will be representatively demonstrated in the next section through five examples: two encouraging theoretical *starts*, a promising mathematical *framework*, and (possibly) two plans for theory-based *pilots*. To be concrete:

[1] For terminological simplicity, I shall use ARNets to refer both to a model and to its originating real system.

Example 1 (§3.1) - EDA & Theory: infinite-server (IS) Models. In [28], a multi-server queueing dynamics (as in Figure 1) was approximated by an infinite-server model. This enabled significant improvements over existing AppS algorithms - theoretical and DFCI's - which is naturally triggering further research.

Example 2 (§3.2) - Theory: Many-server appointment queues in the asymptotic QED and NDS regimes. Following the pioneering [2], I shall outline here a preliminary analysis of a many-server appointment queue in the asymptotic QED (Quality- and Efficiency-Driven) regime, and then hint at an NDS (Non-Degenerate Slowdown) analogue [5]. The analysis, jointly with J. Huang and P. Momcilovic, is asymptotically stationary. It will then be generalized to acknowledge explicitly time-varying dynamics (e.g. in DFCI, service capacity changes over the day).

Example 3 (§3.3) - Mathematical resource-focused framework for ARNets. My starting point is the BSF proposal [4], jointly with M. Armony and P. Momcilovic. Its goal is to develop a flexible framework, in which customers and servers are modeled symmetrically - both viewed as *resources* that wait for each other (see Figure 6 in [6]). Here, I propose to add to this framework AppS of its resources, while taking advantage of its flexibility to incorporate some ARNets' complex features, as described above.

Example 4 (§3.4) - Pilot at DFCI, based on our theoretical research, existing and future, of offline and online AppS. Attention will be given to stochastic variability in punctuality and service durations [28], which is presently lacking from AppS at DFCI (and often more generally).

Example 5 (§3.4) - Pilot at the Israeli Judiciary, thus complementing ISF project 404/17 by Keren Weinshall-Margel. The theme is the "Production of Justice": it starts with understanding JW (and EDA at the SEELab), and it will culminate in judicial AppS - which is the research that I propose here.

Expected Significance: This proposal will gradually develop a data-based theoretical framework for stochastic ARNets and their AppS systems; and it will hopefully also affect the practice of AppS in ARNets. The courage behind this hope stems from the reasons that concluded the previous subsection ("Why be optimistic..."); plus an exploding trend of "data-analytics" in the service sector, that relies on cheaper and cheaper automatic data-gathering technologies (e.g. RTLS) that supply transaction-level bias-free data.

3 Detailed Description of the Proposed Research

3.1 Example 1: IS Models of Appointment-Sequencing/Scheduling under Uncertainty

I start with a brief summary of [28]. Its authors believe that it presents the first practical algorithm for offline sequencing and scheduling, of a stochastic ARNet with multiple (many) servers. In this **offline problem**, there are n given customers that are to be assigned to service during a time interval $[0, \bar{T}]$. The service facility has c_t available servers at time $t \in [0, \bar{T}]$. Regular business hours of the facility are $[0, T]$, $T \leq \bar{T}$; any work during $(T, \bar{T}]$ counts as overtime, which incurs a cost of γ per server per unit of time. The planner's problem is to choose (deterministic) appointment times a_i, for each customer $i = 1, \ldots, n$; $a_i \leq T$ must

hold (and additional constraints, e.g. deadlines, can readily be incorporated). Service time (duration) and punctuality of the ith customer are $D_i \geq 0$ and P_i, which are distributed F_i and G_i, respectively. All random variables are assumed mutually independent (which can be relaxed).

Customer i arrives to the system at time $t = a_i + P_i$; $P_i > 0$ implies that customer i is late, and conversely for $P_i < 0$. (No-shows are appointments with zero service requirements.) With S_i being the service start time of customer i, waiting time is $W_i = S_i - a_i - P_i \geq 0$, and the corresponding server is busy during $[S_i, S_i + D_i)$. Overtime work due to customer i is $O_i = (S_i + D_i - T)^+$. Customers are served by any available idle server in the order of their arrival, and they remain in service until completed. The appointment *sequencing* problem is to select the appointment times $a_1, \ldots, a_n$, so as to minimize the expected cost

$$\mathbb{E}\left[\sum_{i=1}^{n}(W_i + \gamma O_i)\right]. \tag{3.1}$$

Given a permutation π of $(1, \ldots, n)$, the appointment *scheduling* problem is to select appointment times $a_1, \ldots, a_n$, to minimize (3.1) subject to $a_{\pi_1} \leq a_{\pi_2} \leq \cdots \leq a_{\pi_n}$ (scheduling is easier than sequencing).

Results. A plethora of well-performing heuristics for "single-server" ($c_t \equiv 1$) scheduling is surveyed in §2 of [28]). None extends to multiple-server scheduling and sequencing, not to mention a stochastic environment. The challenge is to map a given schedule to delays and overtime, which we do in [28] via an IS model. Specifically, we first introduced an IS model that approximates *occupancy* in a multiple-server system, given a fixed appointment schedule. Second, we validated extensively our approximations against DFCI data. Third, we embedded the IS model in an optimization routine that allows one to produce well-performing schedules. Finally, we benchmarked our IS-based schedules against existing approaches. In §7 of [28], and for DFCI infusion operations (Figure 1), we compared our IS approach with means-based scheduling (used at DFCI and prevalent elsewhere [8]). Our algorithm performed significantly better, suggesting that deployment of the IS appointment sequencing approach at DFCI could decrease total cost (waiting plus overtime) by roughly 30%. We now describe briefly the IS framework.

Infinite-Server (IS) Model. Given n appointment times $a_1, \ldots, a_n$, our approximation assumes ample servers: customer i arrives at time $a_i + P_i$ and departs at $a_i + P_i + D_i$ (no waiting). Let $Z_i(t)$ be the indicator of customer i being present in the system at time $t \in \mathbb{R}$:

$$Z_i(t) := 1_{\{a_i + P_i \leq t < a_i + P_i + D_i\}}.$$

It follows that $\mathbb{E}Z_i(t) = \mathbb{E}\tilde{F}_i(t - a_i - P_i) =: \Omega_i(t)$, with $\tilde{F}_i(x) := 1_{\{x \geq 0\}}(1 - F_i(x))$. Let $Z(t) := \sum_{i=1}^{n} Z_i(t)$, $t \in \mathbb{R}$, denote the total number of customers in the IS system. Then $\mathrm{Var}(Z(t)) =: \sigma^2(t) = \sum_{i=1}^{n} \Omega_i(t)\ (1 - \Omega_i(t))$, by customers' independence. Note that $Z(t)$ is a lower bound to actual occupancy, yet it proves highly useful for AppS. Specifically, with c_t being the staffing level, and given a function $r : \mathbb{R} \to [0, \infty)$, define the cost as

$$Q(Z) := \int_{-\infty}^{\infty} r(Z(t) - c_t)\ \mathrm{d}t + \tilde{\gamma} \int_{T}^{\infty} Z(t)\ \mathrm{d}t,$$

for some $\tilde{\gamma} > 0$; the above two terms approximate costs of waiting time and overtime. Practically, we use a CLT-based approximation of $Z(t)$: $\tilde{Z}(t) := \mathbb{E}Z(t) + \xi(t)\,\sigma(t)$, where $\xi(t)$ is a standard normal random variable and $\sigma(t)$ is given above. Then $Q(\tilde{Z})$ approximates $Q(Z)$, and hence we seek to

$$\min_{\{a_i \in [0,T]\}_i} \mathbb{E}Q(\tilde{Z}). \tag{3.2}$$

The approximation $\tilde{Z}(t)$ is amenable to analysis as it is fully characterized by its mean and standard deviation. For example, if $r(x) = x^+$ and G_i is not varying with i, we have

$$\frac{\partial}{\partial a_i}\mathbb{E}Q(\tilde{Z}) = \int_{-\infty}^{\infty}\left[\frac{\Omega_i(t) - 1/2}{\sigma(t)}\varphi(\psi(t)) - \Phi(\psi(t))\right]\mathrm{d}\Omega_i(t) + \tilde{\gamma}\,\Omega_i(T).$$

As a concrete example (Figure 8 in [28]), we solved (3.2) for various values of $\tilde{\gamma}$; this revealed schedules close to the efficient frontier of "wait-time vs. overtime" costs, from which one selects the best schedule that trades off γ units of wait-time with one unit of overtime.

Research Agenda for Example 1. The above "vanilla" version of the sequencing problem is already interesting technically. We believe that it will be amenable to the following important extensions:

1. *Online algorithm.* It is of practical interest to consider an online AppS, where patients submit appointment requests asynchronously. In [28], we experimented with a myopic version of our offline algorithm, demonstrating a gap in performance of approximately 32%, between offline vs. myopic. This gap, due to the non-anticipatory nature of the myopic algorithm, is the potential that a "smart" online AppS can materialize, and which I propose to pursue.

2. *Appointment book process.* SEELab data from DFCI includes also all modifications/cancellations within the appointment book. This will reveal how AppS for a particular day evolves/changes (usually starting around one year earlier). It will yield models for appointment book evolution (a first, to our knowledge); it will shed light on "home-waiting" for an appointment (vs. intra-facility waiting); and, as importantly, it will enhance prediction of future "demand" for specific days, which will serve online AppS.

3. *Enhancing the basic IS*, with data-based features from the "Complexity" part of Section 2.

3.2 Example 2: Many-server appointment-queues in the asymptotic QED and NDS regimes

Problem description. It is now convenient to let n denote the number of servers, and $[0, H]$ be the time-interval. Customers require service of exponential duration with mean $1/\mu$. (No-shows can be easily accommodated.) Our planner is provided with an operational budget $\alpha\sqrt{n}H$ (QED budget), which bounds costs of waiting (customers) plus overtime (servers). The objective is to *maximize the number of appointments* $N(n, H)$, without exceeding budget, which entails also determining appointment times $\{t_i\}$ in $[0, H]$ (and which resolves the trade-off that many early (late) appointments incur high waiting (overtime) costs).

We aim at n large. Then accounting for service beyond time H as overtime would have overtime costs trivially dominate waiting costs. One way for a meaningful balance is to have a server that is busy

at time $H-\frac{1}{\mu}$ not count that service, beyond $H-\frac{1}{\mu}$, as overtime. Thus, overtime equals the total service requirements of only those customers who are *waiting* in queue at time $H-\frac{1}{\mu}$. (Think "law of large numbers" over many "days" $[0,H]$: the many customers in service at time $H-\frac{1}{\mu}$ will, "on average", complete their service by time H.) Formally, denote by $Q(t)$ the number of customers waiting in queue at time t. Then the average total overtime and total waiting time are given respectively by

$$\frac{1}{\mu}\times\mathbb{E}\left[Q\left(H-\frac{1}{\mu}\right)\right], \qquad \int_0^\infty Q(t)dt.$$

In the above circumstances, is it natural to **Conjecture** that one could optimally schedule all appointments before $H-\frac{1}{\mu}$. This gives rise to three intervals: 1. $[0,H-\frac{1}{\mu}]$; 2. $[H-\frac{1}{\mu},H-\frac{1}{\mu}+\xi]$ (here $H-\frac{1}{\mu}+\xi$ is the first time after $H-\frac{1}{\mu}$ when the queue empties, and must remain so thereafter); 3. $[H-\frac{1}{\mu}+\xi,H-\frac{1}{\mu}+\xi+\tau]$ (here $H-\frac{1}{\mu}+\xi+\tau$ is the first time after $H-\frac{1}{\mu}+\xi$ when the system becomes empty and remains so). Note that waiting costs accrue only during the first two intervals, and that all n servers are busy during the second interval. This and some calculations yield an expression for the expected total costs, given $N(n,H)$ and $\{t_i\}$:

$$C(N,\{t_i\},H)=c_w\left(\mathbb{E}\left[\int_0^{H-\frac{1}{\mu}}Q(s)ds\right]+\frac{3}{2n\mu}\left(\mathbb{E}[Q^2(H-\frac{1}{\mu})]+\mathbb{E}[Q(H-\frac{1}{\mu})]\right)\right)+\frac{c_o}{\mu}\times\mathbb{E}\left[Q\left(H-\frac{1}{\mu}\right)\right],$$

with c_w being unit waiting-cost and c_o is unit overtime-cost. Thus the planner's problem is to maximize $N(n,H)$, subject to $\frac{1}{H\sqrt{n}}C(N,\{t_i\},H)\le\alpha$, which I choose to represent in the following manner:

$$\max_{\{t_i\},N(n,H)}\sqrt{n}\left(\frac{N(n,H)}{nH}-\mu\right), \text{ subject to } \frac{1}{H\sqrt{n}}C(N,\{t_i\},H)\le\alpha.$$

The reason is that the problem as originally posed is intractable, and the above representation suggests how to pursue its asymptotic solution, as n and H grow large. Due to space constraints, however, and since there is way more to do than has already been done, I must now take a shortcut to the final conjecture, on what would constitute an asymptotically optimal strategy. To this end, I introduce the following 2 constants:

$$\beta^*=\text{argmin}_{\beta<0}\left[c_w\mathbb{E}[Q]-\frac{c_o}{\mu}\beta\right]=\text{argmin}_{\beta<0}\left[c_w\frac{\sigma}{|\beta|+|\beta|^2\frac{\Phi(|\beta|/\sigma)}{\phi(|\beta|/\sigma)}}+\frac{c_o}{\mu}\frac{|\beta|}{\sigma}\right],$$

$$\text{and}\quad \zeta^*=\alpha-\left[c_w\frac{\sigma}{|\beta^*|+|\beta^*|^2\frac{\Phi(|\beta^*|/\sigma)}{\phi(|\beta^*|/\sigma)}}+\frac{c_o}{\mu}\frac{|\beta^*|}{\sigma}\right].$$

Conjecture: As n and H grow indefinitely, and under some appropriate conditions, an asymptotically optimal AppS invites a total of $nH(\mu+\frac{1}{\sqrt{n}}\zeta^*)$ customers, with appointment times as follows:

- n customers have appointment time 0;
- Customer i, $i=n+1,\cdots,nH(\mu+\frac{1}{\sqrt{n}}\zeta^*)$, has appointment time

$$T_{i,H}=\min\left\{\frac{1}{n(\mu-\beta^*/\sqrt{n})}(i-n),H-\frac{1}{\mu}\right\}.$$

Comment: Example 1 in [28] has one schedule with exactly the above arrival shape: arrivals at the outset to have all servers start the day busy, and then all other arrivals spread uniformly over $[0, H-\frac{1}{\mu})$.

Here are some "hints" to how this conjecture was arrived at:

• When n and H are large, the term $\frac{1}{H} \times \frac{3}{2n\mu}\left(\mathbb{E}[Q^2(H-\frac{1}{\mu})]] + \mathbb{E}[Q(H-\frac{1}{\mu})]\right)$ is negligible. The cost constraint then simplifies to

$$\frac{1}{H}\left(c_w\mathbb{E}\left[\int_0^{H-\frac{1}{\mu}} \frac{Q(s)}{\sqrt{n}}ds\right] + \frac{c_o}{\mu} \times \mathbb{E}\left[\frac{Q\left(H-\frac{1}{\mu}\right)}{\sqrt{n}}\right]\right) \leq \alpha. \tag{3.3}$$

• $\frac{Q}{\sqrt{n}}$ can be approximated by $\widehat{X}^+$, where $\widehat{X}$ is the following diffusion process, over $[0, H-\frac{1}{\mu})$:

$$d\widehat{X}(t) = 1_{\{\widehat{X}(t)>0\}}\beta(t)dt - 1_{\{\widehat{X}(t)\leq 0\}}\mu\widehat{X}(t)dt + \frac{1}{\sqrt{2}}dB(t);$$

here $B(\cdot)$ is a standard Brownian motion, and the drift function $\beta(\cdot)$ is to be determined.

• Invitations at time $H-\frac{1}{\mu}$ are "translated" to a diffusion-jump at that time. The diffusion control problem is then to determine both the optimal drift-function $\beta(\cdot)$ and the optimal jump-size at time $H-\frac{1}{\mu}$.

• If appointment times are "translated" to a diffusion drift $\beta(t)$, then $(n\mu H + \beta(H)\sqrt{n})$ are invited during $[0, H-\frac{1}{\mu})$. This suggests $nH(\mu + \frac{1}{\sqrt{n}}\zeta)$ invitations in total, for an appropriate constant ζ, which leaves $H(\sqrt{n}\zeta - \sqrt{n}\frac{\beta(H)}{H})$ scheduled to arrive at time $H-\frac{1}{\mu}$; hence $\widehat{X}$ has a jump of size $\zeta - \frac{\beta(H)}{H}$ at time $H-\frac{1}{\mu}$.

• We have finally arrived at the following diffusion optimization-problem:

$$\max_{\zeta,\beta}\ \zeta\,, \qquad \text{subject to} \quad \limsup_{H\to\infty}\frac{1}{H}\left[c_w\mathbb{E}\left[\int_0^{H-\frac{1}{\mu}}\widehat{X}^{\beta,+}(s)ds\right] + \frac{c_o}{\mu}H(\zeta - \frac{\beta(H)}{H})\right] \leq \alpha.$$

• The last conjecture is that an optimal drift function for the above problem is linear: $\beta(t) = \beta t$, with $\beta < 0$. This brings us back to our conjectured asymptotically-optimal appointment schedule, in terms of β^* and ζ^*.

Relationship to [2], by Armony, Atar and Honnappa, who considered AppS for a single-server in conventional heavy-traffic: our overall approach is similar yet with some subtle complicating differences. Indeed, in [2], overtime is easy to model, the number of customers is a priori given, and the corresponding diffusion is the tractable reflected Brownian motion (ours is piecewise-OU). It is also worth mentioning that we expect the analysis in [2] to carry over, with some minor adaptations, to the many-server NDS regime [5].

Research Agenda for Example 2. The first goal is to complete the analysis of the above base-case model, which can be viewed as an AppS analogue of Halfin-Whitt [22]. Significantly, the simple shape of an asymptotically optimal AppS makes it attractive for additional applications, notably data-driven scheduling as in [38]. Further directions (straightforward to propose though not necessarily so to pursue) are:

1. *Enriched features:* no-shows by coin-flip (easy); punctuality, of both customers and servers (yet unclear); networks with both many-server and single-server nodes [2], all in heavy traffic (as in Figure 5).

2. *Time-varying models:* first a fluid model, which will capture the "shape" of the occupancy process (as opposed to approximating it through the IS process in Example 1, or the ARNet to be introduced in the next Example 3). Time-varying diffusion refinements are expected to be tractable only numerically.

3.3 Example 3: Framework for ARNets (static and dynamic, deterministic or stochastic)

In [4], a framework for processing/activity networks was developed, which was inspired by many-server queueing systems in heavy-traffic: in such systems, customers and servers are viewed symmetrically, and both are referred to as *resources* (e.g. customers could also be a bottleneck; or, in the QED regime, customers and servers, namely resources, should scarcely wait for each other). Figure 1 in [4], based on call-center data, illustrates well this symmetry. (Despite some similarity in appearance, that figure is different from the judiciary networks in our Figure 2.) The goal of [4] was to first develop algorithms that generate data-based models, starting from transaction-level data (e.g. from RTLS). These models would then be analyzed to support design, staffing and control; here I propose to add to them AppS, thus creating appointment-books to all resources *jointly*. I shall refer to the resulting models as CARNets (*Closed* ARNets).

Why Closed? The framework [4] follows the pioneering work of Harrison [23], who was inspired by conventional heavy-traffic. Harrison's models are *open*, which have in them an inherent asymmetry between servers and customers: customers have a finite sojourn time within the system while servers stay forever. Complete symmetry thus naturally calls for a *closed* model, where *both* customers and servers remain in the system forever. (Note that this is without loss of generality: the outside "world" in an open network can be made, or be approximated by, a node in a corresponding closed network). Moreover, the appointments in ARNets entail synchronization between resources. Hence the fact that [4] covers fork-join constructs renders natural the modelling of appointments where, say, a customer-exam requires a room, doctor, nurse and equipment; or where a nurse caters simultaneously to multiple infusion chairs.

Fluid CARNets: the static model. The most basic CARNet is a closed system with n different resource pools and m different activities. Resources engage in activities. A resource unit can be in several states (one at a time); we then refer to any pair *(resource, state)* as a *sub-resource*. Let k be the number of sub-resources in the system. Any two units of a given sub-resource are interchangeable. An $n \times k$ incidence matrix R, with values in $\{0, 1\}$, describes the relationship between resources and sub-resources: $R_{i,l} = 1$ whenever sub-resource l is affiliated with resource i. Model primitives include also an n-vector b of resource capacities and an m-vector a of mean activity durations; elements b_i and a_j represent the total amount of resource i in the system and the mean duration of activity j, respectively. Define $A := \mathrm{diag}(a)$ for notational convenience. Activities are described further by two nonnegative $k \times m$ matrices: an input (consumption) matrix C and an output (production) matrix P. The element $C_{l,j}$ defines the amount of sub-resources l required to be engaged in activity j. Similarly, $P_{l,j}$ specifies the amount of sub-resources l created upon completion of activity j. Assuming $R1 = b$ (which can be relaxed), we require that $RC = RP$ (conservation of resources at each activity). A plan x is an m-vector of activity levels, which is feasible if

$$RCAx \le b, \quad (C - P)x = 0 \quad \text{and} \quad x \ge 0. \tag{3.4}$$

Components of x are rates at which activities are conducted (thus $x \ge 0$; additional constraints can be imposed, for example to capture routing). The vector $RCAx$ is the amount of resources required under x.

Though formally similar to Harrison [23, (2.1)], our feasibility condition (3.4) is fundamentally different: (3.4) is in terms of resource amounts (counts) while [23] bounds processing rates.

Research agenda for Example 3. First in my agenda is to get CARNets to cover features and systems that are relevant to the present proposal. Some have already been addressed in [4], for example bottleneck identification, open systems, time-varying dynamics and, notably, approximations of resources in conventional heavy-traffic (as in Figure 5: blood-draw and infusion are many-server nodes, but exam is a collection of single-servers in conventional heavy-traffic, as manifested by the exponential waiting time in Figure 9a.)

The CARNet framework will enable AppS of a complete network and all its resources jointly. I plan to take the approach of multi-level AppS, namely iterating between top- and low-level planning (as in [37]). Starting *bottom-up*, and utilizing transaction-level data, one can create a high-resolution intra-day dynamic CARNet for each day in our planning horizon. (Given a daily panel-size, the dynamic closed model is natural over a finite-horizon.) All dynamic models will be aggregated into a static model, over many days (or weeks or months or a year, as in DFCI). The static fluid model, via optimization, will yield an aggregate plan and parameters, that will next serve as long-term constraints for daily planning. (For example, one could determine daily panel size, or the fraction of time that resources are to be engaged in activities, doing the latter by optimizing over P subject to capacity constraint and given a goal-plan x.) Then a *top-down* phase will follow, by redeveloping a dynamic model that will guide AppS at the levels of intra-day and possibly individual-resources. (For example, with panel size given, one can refine the model to account for characteristics such as patients' medical conditions, punctuality (of patients and doctors), familiarity level between doctors and patients (which could affect service-durations) etc.

3.4 Examples 4 & 5: Pilots at DFCI and the Israeli judiciary

The pilots aim at research validation (benefits my research) and technology-transfer (my partners), plus ensuring the continuation of an automatic non-intrusive data-flow into the Technion SEELab.

At **DFCI**, AppS is based on scheduling templates (common practice in healthcare); templates ensure resource accessibility and AppS efficiency+fairness. Interventions of our research will be implemented by modifying the template, which will not disrupt clinical/operational procedures at DFCI. Data sources are DFCI's RTLS and Electronic Health Records (an Epic system).

Re. the Israeli **Judiciary**, Weinshall-Margel's ISF will investigate JW and its measures, causes for it to be high/low, and its effects on judges' wellbeing, work-quality and court-functioning. This will pave the way for my operations research, which will seek data-based ways to better manage and cope with JW. Data sources are mainly Israeli judiciary's computerized case routing and management system (Net Hamishpat [42]: It covers all information on all stages of resolved cases (excluding national security cases, adoption proceedings, etc.), assigned judges, litigants and lawyers, substance, pleadings, motions, preliminary and trial hearings, continuances, temporary injunctions, summations, witnesses, costs, case disposition and outcomes.

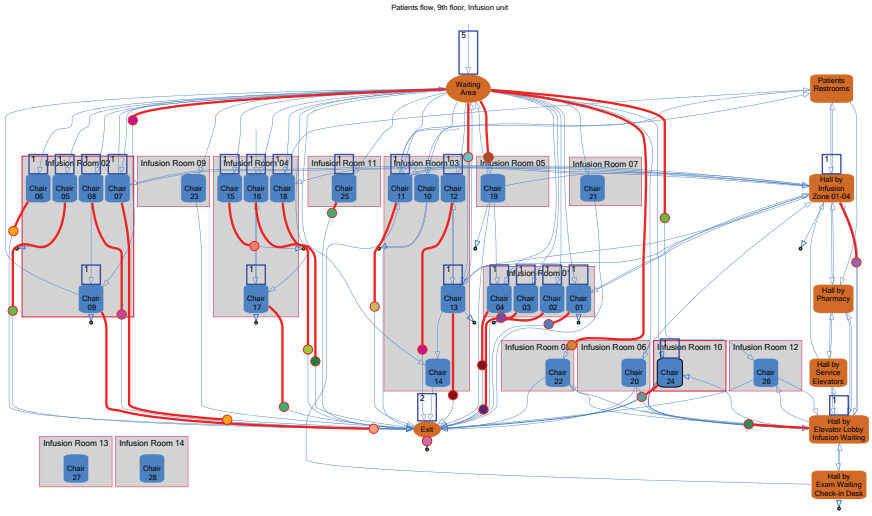


Figure 1: *Patient Flow at DFCI via RTLS data.* Patient location at the infusion unit of the Dana Farber Cancer Institute (DFCI), on Yawkey Center's 9^{th}-floor, at 11:17 on XX-YYY-2014. This snapshot is based on a full-day data-animation, created at the Technion SEELab and accessible here: nodes represent specific locations, *e.g.*, infusion chairs and hallways. Patients, represented by circles traversing edges, are obtaining service at the originating node and reach the destination node upon service completion—for example, the patient/circle traversing the edge emanating from Infusion Chair 25 in Room 11 is currently undergoing infusion treatment in that chair (with slow/fast motion along the edge corresponding to long/short treatment times).

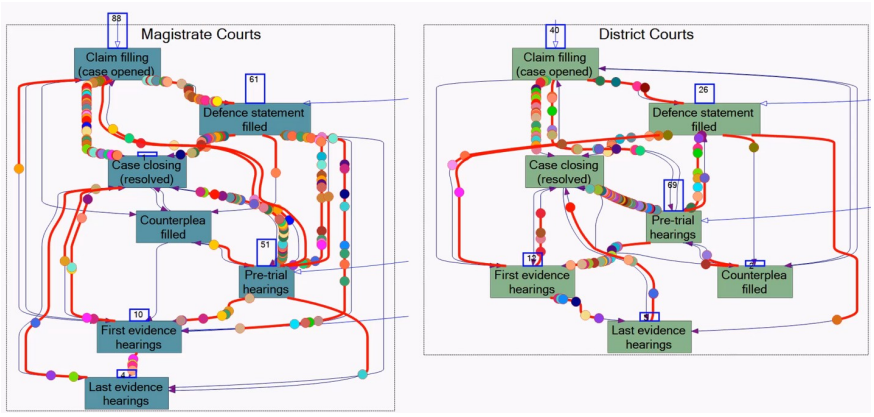


Figure 2: *Flow of civil cases within the Israeli justice system.* Data animation can be viewed here. The two snapshots above describe flows of 2,000 civil cases, filed at the Israeli Magistrate's and District courts (the left and right graphs, respectively), and resolved during 2008–2012. Here circles represent cases in the court system and nodes are states of cases, *e.g.*, "Claim filing (case opened)". With the same interpretation as DFCI above, there are presently 69 cases in District Courts at the "Pre-trial hearings" state.

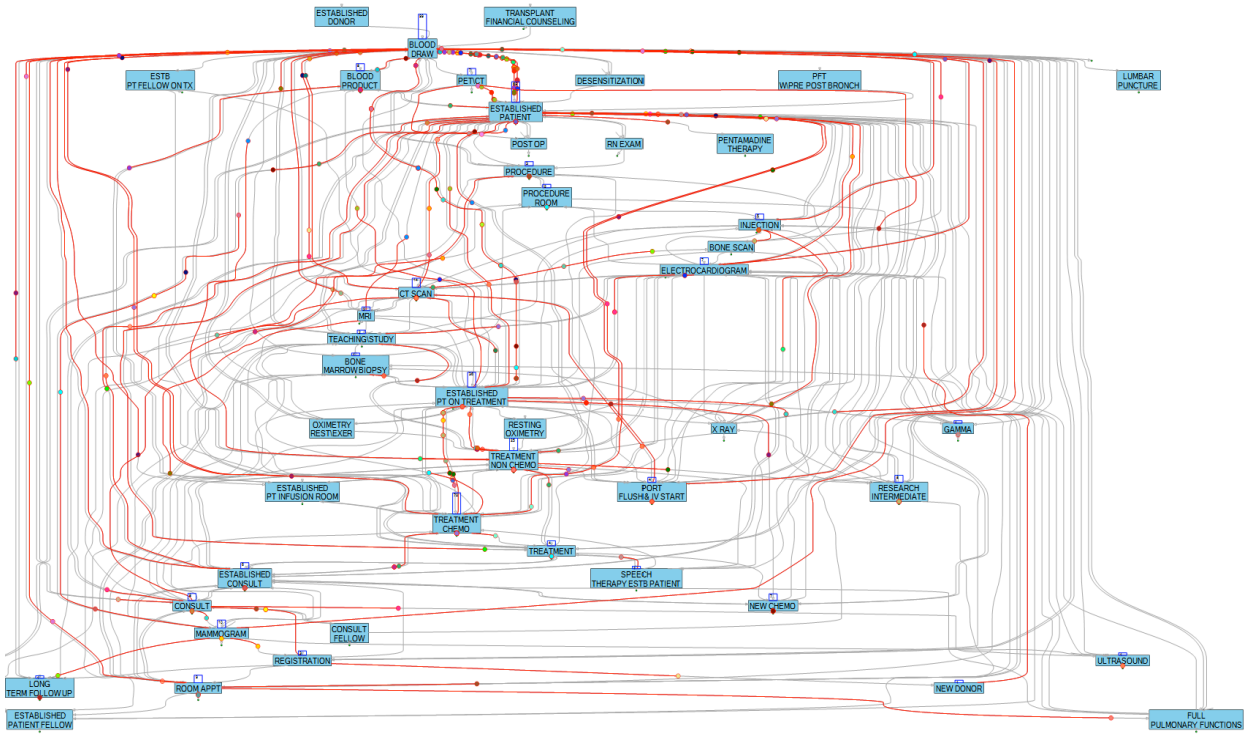


Figure 3: *FULL appointment book/network of DFCI.* Here nodes (rectangles) correspond to *scheduled* activities. Many of the activities require multiple resources.

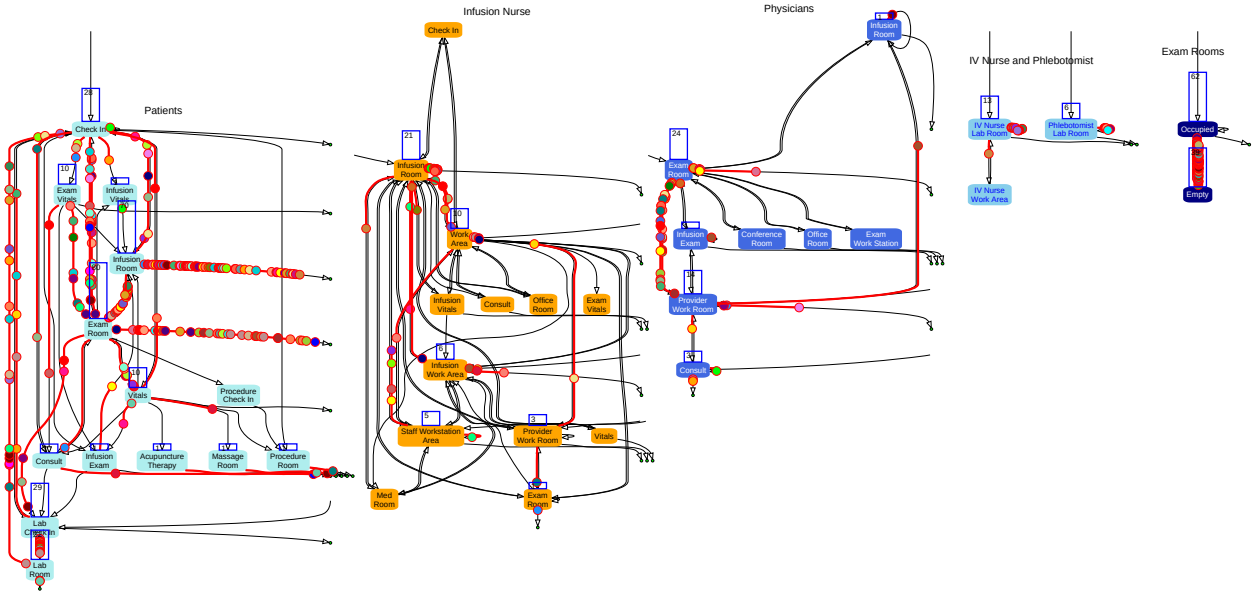


Figure 4: *Network of resources - actual activities.* From left to right: doctors, infusion nurses, doctors, blood-draw nurses, rooms.

Data animation can be viewed here

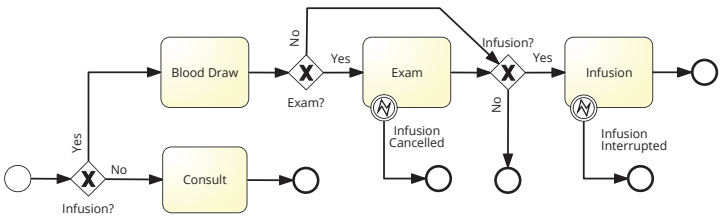


Figure 5: *The main clinical process at DFCI.*

In [28], we focused on infusion (on the right). Scheduling an infusion appointment potentially requires the consideration of multiple resources, in particular infusion chair/bed and infusion nurse (the former being the bottleneck). Beyond infusion, patients may require a blood draw (prior to treatment), an exam, radiation therapy, imaging, or one of auxiliary services provided by translators, nutritionists, social workers, etc. Current scheduling guidelines call for one-hour time gaps between various steps, in order to ensure, for example, that patients are able to make subsequent appointments on time, and that blood draw results (medication) are available to providers (infusion).

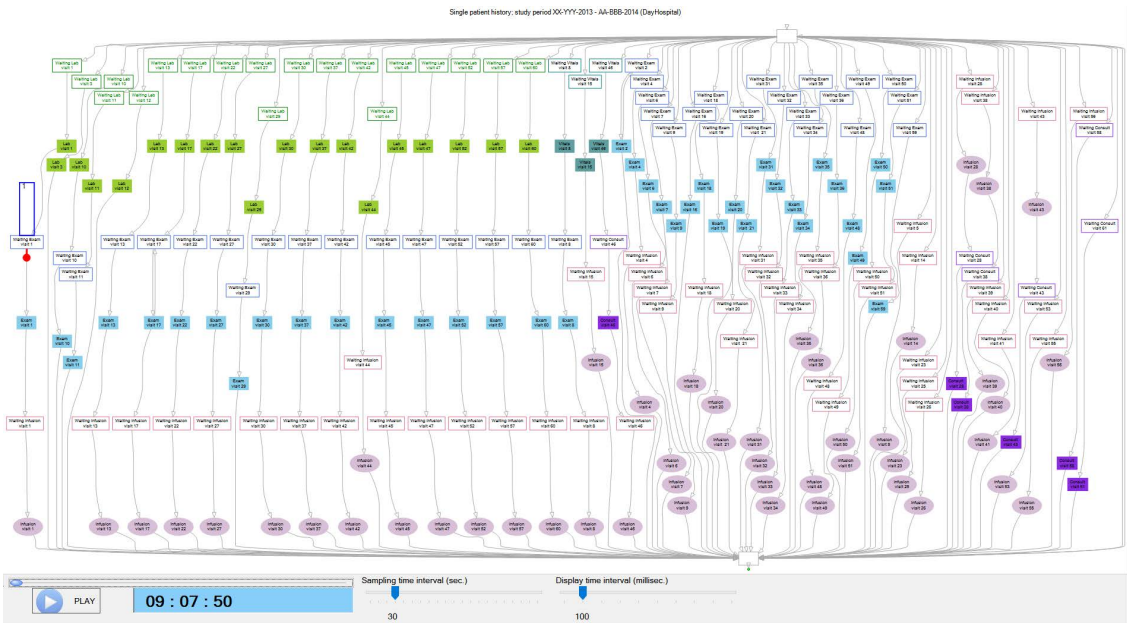


Figure 6: *Complex care-path at DFCI* = 61 visits by a single patient over a 6-month period (sorted by visit-type, as opposed to chronologically).

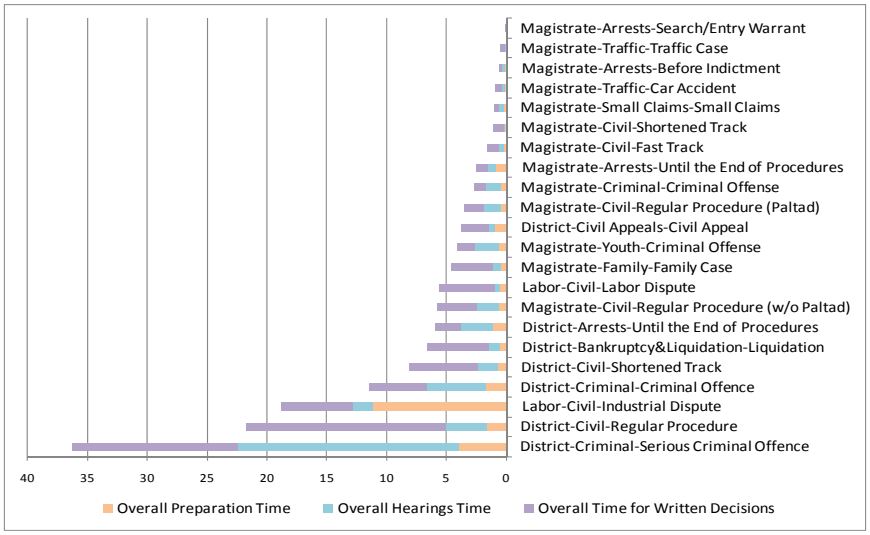


Figure 7: *Average judicial times: variations over 22 leading case types (in hours)* ([40]).

Average times range considerably, from only three minutes in search and entry warrant cases to about 36 hours for serious criminal cases. More refined analysis confirmed that Israeli judges do suffer from high JW, which could have social and legal consequences. For example, average judicial time invested in a district-court's criminal case that goes to full trial is ten fold the time invested in a case resolved in a plea bargain (customers "abandoning" their queues). Judges may be thus incentivized to push towards settlements (encourage abandonment).

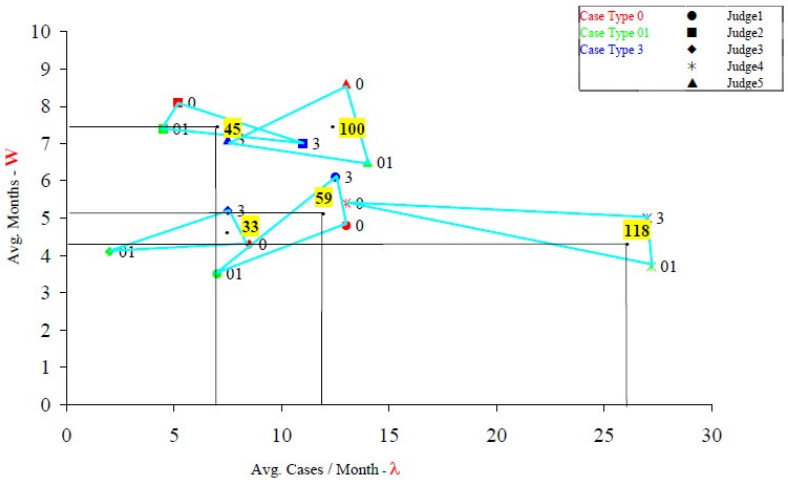
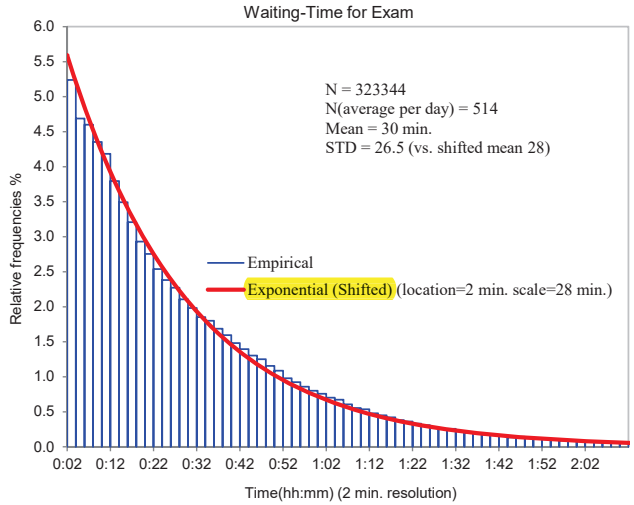


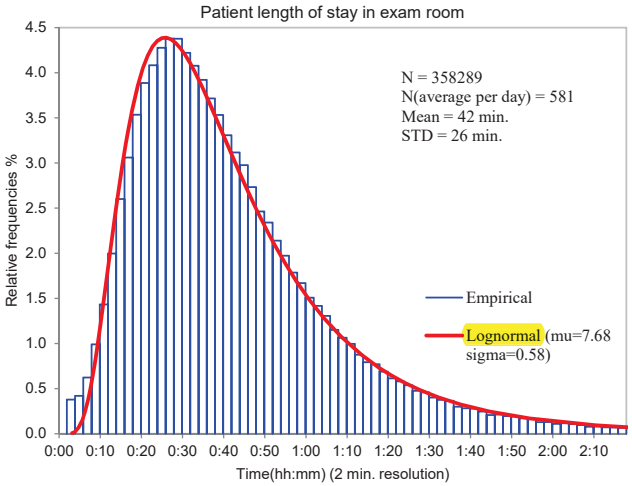
Figure 8: *How to measure operational performance: five judges processing three case types*

The graph shows average handling time W and average throughput rate λ, for 5 judges per 3 case types. Each triangle is a proxy for operational performance, with its center of gravity being average performance. The area of each rectangle, $L = \lambda \times W$, is by Little's Law the average number of pending cases. Commonly, large L is interpreted as low efficiency. However, the graph shows that the judge with the largest L has the largest λ and smallest W, hence it is the best operational performer.

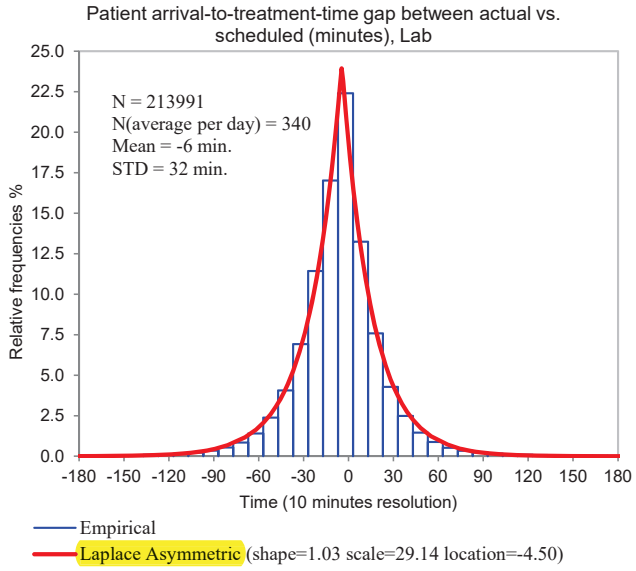
The figure gives rise to a basic AppS challenge: how to match cases with judges (Skills-based Routing)? This question concerns panel size and schedules, but in a judiciary context, it relates also to fairness, ethics (judges' case-mix should be random) and emotional-load that varies across case-types.



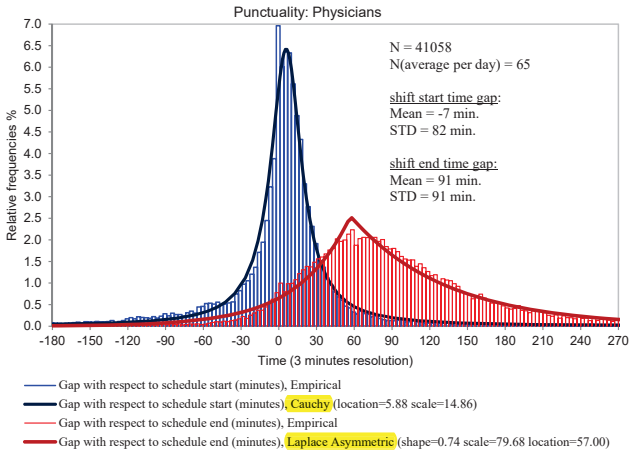
(a) Figure 9a: **Waiting for a doctor (as in $G/G/1$ heavy-traffic)**



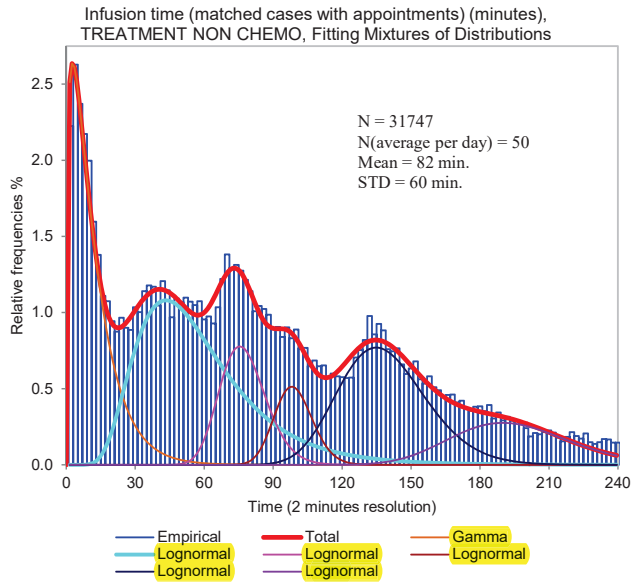
(b) Figure 9b: **LogNormal service durations**



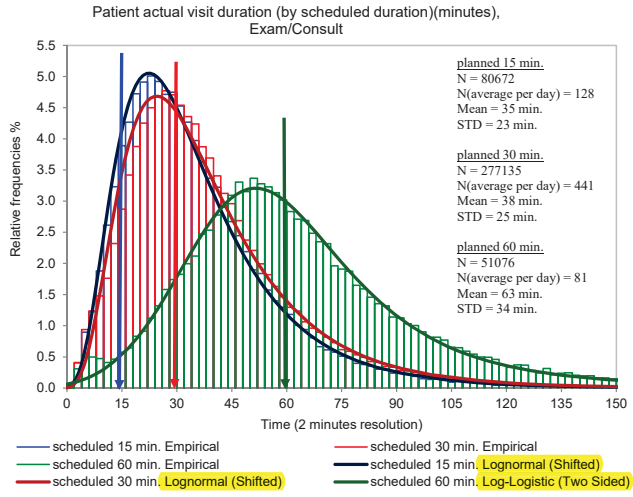
(c) Figure 9c: **Patients punctuality (< 0 = early arrival)**



(d) Figure 9d: **Doctors punctuality (shift-start and -end)**



(e) Figure 9e: **Infusion duration**



(f) Figure 9f: **Actual vs. planned durations (15, 30, 60 min)**

References

[1] Ahmadi-Javid, A., Jalali, Z., and Klassen, K. (2017). Outpatient appointment systems in healthcare: A review of optimization studies. *Eur. J. Oper. Res.*, 258(1):3–34. 2

[2] Armony, M., Atar, R., and Honnappa, H. (2017a). Asymptotically optimal appointment schedules with customer no-shows. *arXiv preprint arXiv:1708.05920*. 2, 4, 8

[3] Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y., Yom-Tov, G. B., et al. (2015). On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems*, 5(1):146–194. 1, 3

[4] Armony, M., Mandelbaum, A., and Momcilovic, P. (2013-2017b). Data-based models of resource-driven activity networks. BSF no. 2014180, Working Paper. http://ie.technion.ac.il/serveng/References/BSF_2014_FINAL.pdf. 4, 9, 10

[5] Atar, R. (2012). A diffusion regime with nondegenerate slowdown. *Operations Research*, 60(2):490–500. 4, 8

[6] Azriel, D., Feigin, P., and Mandelbaum, A. (2014). Erlang-S: A data-based model of servers in queueing networks. Preprint. http://ie.technion.ac.il/serveng/References/Erlang_S_revision2.pdf. 4

[7] Beenstock, M. and Haitovsky, Y. (2004). Does the appointment of judges increase the output of the judiciary? *International Review of Law and Economics*, 24(3):351–369. 2

[8] Berg, B. and Denton, B. (2012). Appointment planning and scheduling in outpatient procedure centers. In Hall, R., editor, *Handbook of Healthcare System Scheduling*, volume 168 of *International Series in Operations Research & Management Science*, chapter 6, pages 131–154. Springer. 2, 5

[9] Best, J. and Tiede, L. B. (2015). Vacancy in justice: Analyzing the impact of overburdened judges on sentencing decisions. Available at SSRN: https://ssrn.com/abstract=2417348. 2

[10] Bray, R. L., Coviello, D., Ichino, A., and Persico, N. (2016). Multitasking, multiarmed bandits, and the Italian judiciary. *Manufacturing & Service Operations Management (MS&OM)*, 18(4):545–558. 2

[11] Cardoen, B., Demeulemeester, E., and Belien, J. (2010). Operating room planning and scheduling: A literature review. *Eur. J. Oper. Res.*, 201(3):921–932. 2

[12] Castro, M. and Guccio, C. (2015). Bottlenecks or inefficiency—an assessment of first instance Italian courts' performance. *Review of Law & Economics*, 11(2):317–354. 2

[13] Cayirli, T. and Veral, E. (2003). Outpatient scheduling in health care: A review of literature. *Prod. and Oper. Management*, 12(4):519–549. 2

[14] Cayirli, T., Veral, E., and Rosen, H. (May–June 2008). Assessment of patient classification in appointment system design. *Production and Operations Management*, 17(3):338–353. 2

[15] Denton, B. and Gupta, D. (2003). A sequential bounding approach for optimal appointment scheduling. *IIE Trans.*, 35(11):1003–1016. 2

[16] Dunn, P., Ghobadi, K., Lennes, I., Levi, R., Marshall, A., Rieb, W., and Zenteno, C. (2017). Real-time outpatient scheduling with patient choice. Preprint. 2

[17] European Commission for the Efficiency of Justice (2014). Report on European judicial systems: Efficiency and quality of justice. http://www.coe.int/t/dghl/cooperation/cepej/evaluation/2014/Rapport_2014_en.pdf. 1

[18] Froehle, C. and Magazine, M. (2013). Improving scheduling and flow in complex outpatient clinics. In Denton, B., editor, *Handbook of Healthcare Operations Management*, volume 184. Springer, New York, NY. 2

[19] Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141. 1

[20] Gocgun, Y. and Puterman, M. (2014). Dynamic scheduling with due dates and time windows: An application to chemotherapy patient appointment booking. *Health Care Manag. Sci.*, 17(1):60–76. 2

[21] Gupta, D. and Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Trans.*, 40(9):800–819. 2

[22] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Oper. Res.*, 29(3):567–588. 2, 8

[23] Harrison, J. M. (2002). Stochastic networks and activity analysis. *Translations of the American Mathematical Society-Series 2*, 207:53–76. 9, 10

[24] Heaton, P. and Helland, E. (2011). Judicial expenditures and litigation access: Evidence from auto injuries. *Journal of Legal Studies*, 40(2):295–332. 2

[25] Heise, M. (2000). Justice delayed? an empirical analysis of civil case disposition time. *Case Western Reserve Law Review*, 50:813–849. 2

[26] Kim, S., Cha, W., and Whitt, W. (forthcoming). A data-driven model of an appointment-generated arrival process at an outpatient clinic. *INFORMS Journal on Computing*. http://www.columbia.edu/~ww2040/Appt_JOC_062617pm.pdf. 2

[27] Liu, N., Finkelstein, S. R., Kruk, M. E., and Rosenthal, D. (2017). When waiting to see a doctor is less irritating: Understanding patient preferences and choice behavior in appointment scheduling. Technical report, Management Science Online. https://doi.org/10.1287/mnsc.2016.2704. 2

[28] Mandelbaum, A., Momčilović, P., Trichakis, N., Kadish, S., Leib, R., and Bunnell, C. (2017). Data-driven appointment-scheduling under uncertainty: The case of an infusion unit in a cancer center. *under revision to Management Science*. http://ie.technion.ac.il/serveng/References/scheduled17.pdf. 2, 4, 5, 6, 8, 13

[29] Mandelbaum, A. and Zeltyn, S. (2013). Data-stories about (im)patient customers in tele-queues. *Queueing Syst. Theory Appl.*, 75(2-4):115–146. 3

[30] Mitsopoulos, M. and Pelagidis, T. (2010). Greek appeals courts' quality analysis and performance. *European Journal of Law and Economics*, 30(1):17–39. 2

[31] Moffett, K., Maltzman, F., Miranda, K., and Shipan, C. (2016). Strategic behavior and variation in the supreme court's caseload over time. *Justice System Journal*, 37(1):20–38. 2

[32] Pinedo, M. (2009). *Planning and Scheduling in Manufacturing and Services*. Springer, New York, 2nd edition. 2

[33] Ponticelli, J. and Alencar, L. S. (2016). Court enforcement, bank loans, and firm investment: Evidence from a bankruptcy reform in Brazil. *The Quarterly Journal of Economics*, 131(3):1365–1413. 2

[34] Qi, J. (2016). Mitigating delays and unfairness in appointment systems. *Management Science*. Published online in Articles in Advance 25 Mar 2016. 2

[35] Rosales-López, V. (2008). Economics of court performance: an empirical analysis. *European Journal of Law and Economics*, 25(3):231–251. 2

[36] Santibáñez, P., Aristizabal, R., Puterman, M., Chow, V., Huang, W., Kollmannsberger, C., Nordin, T., Runzer, N., and Tyldesley, S. (2012). Operations research methods improve chemotherapy patient appointment scheduling. *Jt. Comm. J. Qual. Patient Saf.*, 38(12):541–553. 2

[37] Senderovich, A. (2012). *Multi-Level Workforce Planning in Call Centers*. M.Sc. Thesis, Technion-Israel Institute of Technology. http://ie.technion.ac.il/serveng/References/Arik_Thesis_24_10.pdf. 10

[38] Senderovich, A., Rogge-Solti, A., Gal, A., Mendling, J., Mandelbaum, A., Kadish, S., and Bunnell, C. A. (2015). Data-driven performance analysis of scheduled processes. In *International Conference on Business Process Management*, pages 35–52. Springer. 8

[39] van Leeuwaarden, J. S., Mathijsen, B. W. J., and Zwart, B. (2017). Economies-of-scale in resource sharing systems: tutorial and partial review of the qed heavy-traffic regime. Technical report, preprint arXiv:1706.05397. https://arxiv.org/abs/1706.05397. 1, 2

[40] Weinshall-Margel, K., Galon, I., and Taraboulos, I. (2015). Creating a case weight index for measuring judicial workload. *Hebrew University Law Review*, 44:769–834. 14

[41] Weinshall-Margel, K. and Taraboulos, I. (2013). Hearing deferrals in the court system. Israeli Courts Research Division. http://elyon1.court.gov.il/heb/Research%20Division/doc/Research7.pdf. 2

[42] Weinshall-Margel, K., Yehuda, T., and Shirtz, A. (2011). Reliability of data from the Israeli cases routing and online management software - "net hamishpat". *Jerusalem, Israel: Israeli Courts Research Division*. http://elyon1.court.gov.il/heb/Research%20Division/doc/Research3.pdf. 10

[43] Whitt, W. (2017). Time-varying queues. http://www.columbia.edu/~ww2040/TVQ_082617.pdf. 2

[44] Whitt, W. (Spring 2013). Offered load analysis for staffing. *Manufacturing and Service Operations Management*, 15(2):166–169. 2

[45] Zacharias, C. and Armony, M. (to appear). Joint panel sizing and appointment scheduling in outpatient care. *Management Sci.* 2

[46] Zacharias, C. and Pinedo, M. (2014). Appointment scheduling with no-shows and overbooking. *Prod. and Oper. Management*, 23(5):788–801.

[47] Zacharias, C. and Pinedo, M. (2016). Managing customer arrivals in service systems with multiple servers. Working paper. 2

[48] Zuckerman, A. A. (1999). *Civil Justice in Crisis: Comparative Perspectives of Civil Procedure*. Oxford University Press. 2