The Public Impact of Queueing Theory: From Queen Elizabeth to Internet to Emergency Rooms

Soroush Saghafian

Oct. 2022

Blog Series: PUBLIC IMPACT ANALYTICS SCIENCE (PIAS)

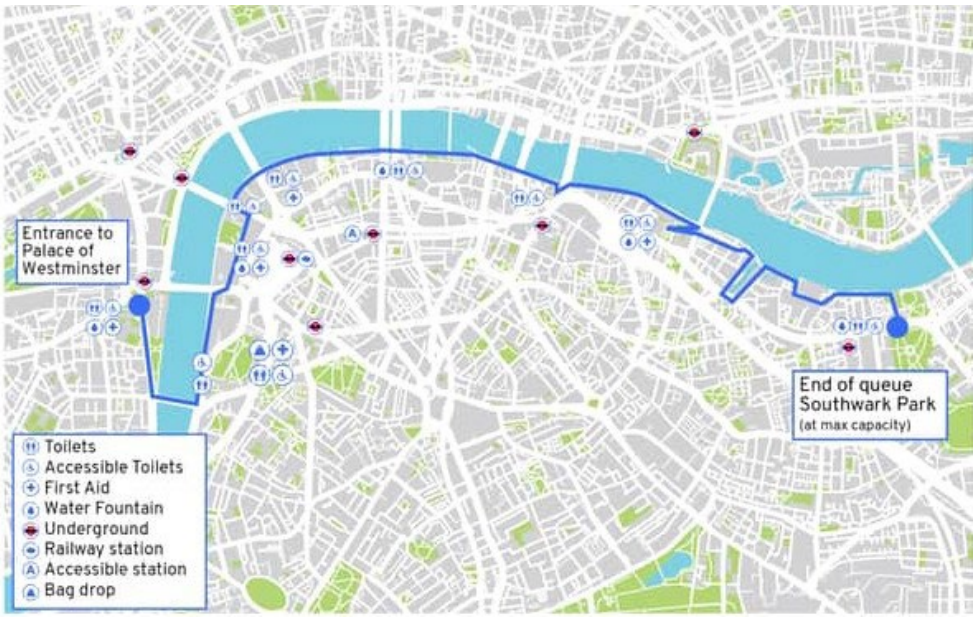

Figure: A map of the long queue formed for Queen Elizabeth II in September 2022 [Credit: Kim Mogg / NationalWorld]

On September 8, 2022 Britain's Queen Elizabeth II— the longest reigning monarch in modern history—died at the age of 96. Six days later, hundreds of thousands of people who wanted to view the late monarch's coffin tested Britain's famous queuing skills to their limit as they queued up opposite the Palace of Westminster in London [1]. Authorities were quite prepared; they had consulted queueing experts to design a gigantic queueing structure with a length of 10 miles for people to line up (see the figure above).

BBC reported that the queue's maximum length ended up being around 10 miles (6.9 miles from Westminster to Southwark, and a three-mile zigzag queue in Southwark Park), and the wait time for some was more than 24 hours [2].

It is hard to imagine what these numbers—maximum queue length and waiting times—would be, if the queueing experts were not involved. Equally, however, one can question whether queueing experts created an "optimal queue." Couldn't they do better? After all, a queue in which some people end up waiting over 24 hours does not seem to be anywhere near "optimal."

But optimizing, or at least improving, a large-scale queue for people interested to take part in a historical event is not the only time that queueing experts are needed. Interested to know why? Well, it is because waiting occurs way more often than we think.

Two facts about waiting lines are just amazing: (a) we spend a good portion of our lives waiting in lines, and (b) waiting lines are everywhere. So it is not surprising that analytics scientists have been trying to find ways to optimize them. The science of optimizing waiting lines is called Queueing Theory, a subject that is very near and dear to my heart. But Queuing Theory is not just about optimizing waiting lines to make them shorter. It is also a method of *capacity management*. By quantifying the relationship between the level of capacity and the amount of waiting, Queueing Theory allows answering questions such as: how much capacity is needed to meet the demand? Given that allocating capacity is often not cheap, what is an appropriate level of resources needed? And how should resources be allocated to achieve the best performance?

Queuing Theory, and the experts using it, have affected your life in many ways that you might not have even noticed. In early 1960's, Leonard Kleinrock, a brilliant analytics scientist who was at the time a doctoral student of Claude Shannon—father of information theory and mathematical theories of communication—at MIT developed a mathematical theory for efficient routing of data in data networks, which was instrumental in designing what we now know as the Internet. He found that he could make use of queuing theory to optimize how data is transferred in such networks.

At the time, the main idea behind how data could be transferred was based on what is known as *circuit switching*. In circuit switching—a useful method of transferring data in phone calls—the bandwidth through which data are transferred between the sender and receiver is constant. As long as a call is going, the bandwidth exists and has the same capacity. Kleinrock realized that it would be a bad idea to use circuit switching for allowing computers to communicate to each other. For starters, computers do not send data at a constant rate. In the words of Kleinrock:

> "They go blast! and they are quiet for a while. A little while later, they suddenly come up and blast again" [3].

Is this pattern weird? Well, not if you think about it. Restaurants, roads, and even emergency rooms see sort of similar patterns in their demand. Around lunch time, your favorite restaurant sees a bunch of customers arriving. But almost no one enters in the afternoon. And then it comes dinner time, when the restaurant again sees quite a few people trying to get a table at the same time.

Similarly, roads see a spike in demand during rush hours, but then in between rush hours, it seems that there are almost no cars on them. And in emergency rooms (ERs)? If you look at the patterns of arrivals to emergency rooms, like the one in the figure below, you realize that around noon many people rush to them. But then after midnight, when not many people seem to need emergency care. What is more, the figure shows that similar patterns, though shifted to the right, can be seen for when hospital beds are requested for ER patients that need to be hospitalized after their ER visit (and also when such patients leave the ER).

So how should we match demand and capacity when demand comes in blast? This is where queueing theory can help a lot. In the context of ERs, for example, various studies, including many of my own, have shown how principles of queening theory can be used to save lives in ERs by cutting long lines

(see, also, [4,5,6,7,8,9,10]). You would not want to waste capacity by providing a constant bandwidth capable of handling maximum demand at all time, a solution that circuit switching would recommend.

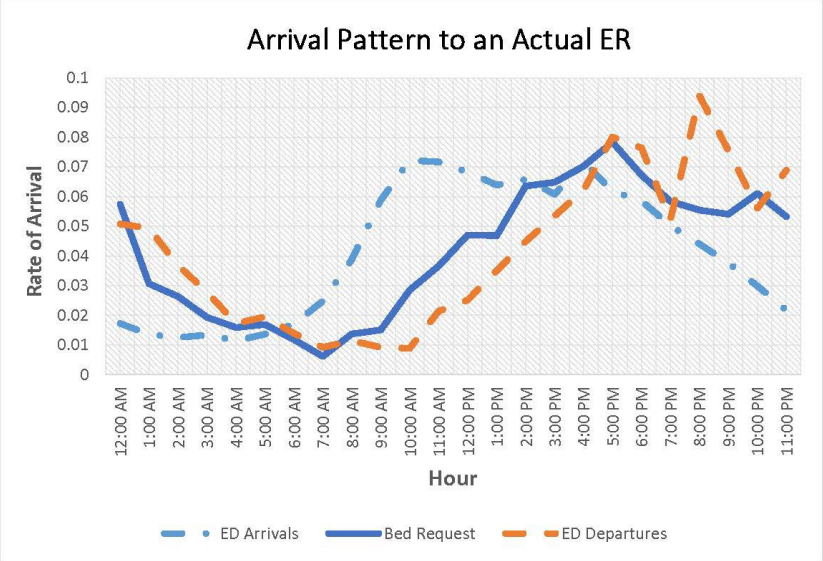

Figure: Pattern of arrivals to an actual emegrncy room (ER) [Source: [4]]

Addressing Societal Problems Using Queueing Theory

In 1976, Kleinrock published an influential two-volume book entitled "Queueing Systems." In it, he not only described and extended the mathematical models underlying waiting lines, but also brought attention to the importance of Queueing Theory in understanding the world we live in:

> "Recently, I made the mistake of flying across the country in a Boeing 747. As a queueing systems analyst, I should have known better! As soon as I arrived at the airport, I immediately realized my error, for there was a mob of passengers waiting to be checked in […]. I, of course, had two extremely heavy suitcases filled with notes on queueing systems (what else?) and so I could not morally avoid this queue. The situation was a multiple-server multiple queue system with clearly unequal rates of service; however, once in invested in a (particularly slow) queue, I could not afford to risk giving up my position. After clearing the check-in procedure, I then found my way to the departure lounge where an enormous queue had been formed in a snakelike fashion awaiting seat assignment and boarding passes."

He continued by further highlighting the importance of studying real-life systems that involve queues, underscoring the public impact that it might have:

> "That travel adventure is just one of many similar situations that all of us have encountered. As systems analysts, we have a moral and personal obligation to study

> these real-life systems and provide more relief from their aggregations even if they do not land themselves easily to analysis [….] The class of systems that generally lands itself to queueing analysis is the (huge) one in which customers compete for access to limited (i.e., finite-capacity) resources. In fact, many of today's significant problems can be reduced to the problem of resource allocation and resource sharing."

Although Kleinrock mainly studied resource allocation in computer networks, he was spot on in stating that many of the problems we face in the society can be reduced to resource allocation. Similar to how in the Internet we need to use available resources (computers, communications channels, etc.) wisely for good performance, we need to be smart in how we allocate our limited societal resources to different tasks that it needs to carry. And this is where the main insights from Queueing Theory are vital.

In addressing societal problems, however, we need to be mindful that understanding what "good performance" means is often not that easy. Interestingly, even in allocating resources on the Internet, what "good performance" involves is not completely agreed upon. For example, while Kleinrock focused on designing data networks so that information can be transferred *efficiently*, other researchers worked on other aspects of performance such as *reliability*.

Similarly, if we think about the question raised earlier—whether queueing experts had designed an "optimal queue" for those who wanted to pay respect to Queen Elizabeth II in September of 2022—we realize that what constitutes good performance might not be easily understood. Understanding better ways of measuring performance—going beyond traditional queueing measures such as average waiting time, sojourn time, or average length of the queue—becomes important when we notice that policymakers have long used queueing as a *rationing mechanism*: when resources are limited, they consciously or not ration them by introducing queues. And it is interesting to note that long queues formed because of this rationing mechanism are often welcomed, because they are interpreted as a *signal* of the importance or high utility of whatever is being rationed.

In a Russian novel published first in 1983 about queueing entitled "The Queue," the author Vladimir Sorokin points to this phenomenon [11]. The main character joins an extremely large queue in Moscow during the time of the Soviet Union. Similar to the queue that we saw was formed in 2022 in London to give respect to the longest reigning monarch in modern history, this queue in Moscow was so long that the main characters could not see what is happening at the end. Yet, joining the queue is so attractive that the main character could not resist. Why should anyone join such a long queue, you may wonder? Well, whatever is being rationed at the end of the queue must be of high utility. Otherwise, what would be the reason for people to form such a long a queue? As the queueing guru, Richard Larson of MIT—who I have known for a long time and recently chatted with when writing these—describes it: "long lines are like magnets of attraction" [12].

Finally, in addressing societal problems using queueing theory, we need to be mindful about the fact that in queueing analysis, one often resorts to "stylized" models. For example, for tractability one might need to assume that there is only one server, or that the interarrival times have specific properties such as being memoryless, time-stationary, or independent of the speed of service. Many such assumptions do not hold in most real-world systems. Stylized models, however, can still provide important insights into how the system can be improved. But it is important to be aware of "model ambiguity," and include in the analyses the uncertainly that stems from not knowing which model or set of assumptions best represent the reality. New advancements in including model ambiguity in queueing analyses, allowing

data-driven approaches that can reduce the dependency of analyses to a specific queueing model can go a long way (see, e.g., [13]).

References

[1] Lawless, J. (2020). Huge line to view monarch's coffin is queue fit for queen. Washington Post.

[2] BBC (2022). Queen's lying-in-state: How long was the queue? https://www.bbc.com/news/uk-62872323

[3] Christian B, and Griffiths T (2016). *Algorithms to live by: The computer science of human decisions*. Macmillan.

[4] Saghafian S, Kilinc D, and Traub SJ (2022). Dynamic Assignment of Patients to Primary and Secondary Inpatient Units: Is Patience a Virtue? *HKS Working Paper No. RWP17-010, Harvard University, Cambridge, MA.*

[5] Saghafian S, Hopp W, Iravani S, Cheng Y, Diermeier D (2018). Workload management in telemedical physician triage and other knowledge-based service systems. *Management Science* 64(11):5180–5197.

[6] Saghafian S, Hopp W, Van Oyen M, Desmond J, Kronick S (2014) Workload management in telemedical physician triage and other knowledge-based service systems. *Manufacturing and Service Operations Management* 16(3):329–345.

[7] Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2012) Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* 60(5):1080–1097.

[8] Saghafian S, Austin G, Traub SJ (2015) Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. *IIE Transactions on Healthcare Systems Engineering* 5(2):101–123.

[9] Traub S, Bartley A, Didehban R, Smith V, Lipinski C, Saghafian S (2016a) Physician in triage versus rotational patient assignment. *Journal of Emergency Medicine* 50(5):784–790

[10] Traub S, Saghafian S, Judson K, Russi C, Madsen B, Cha S, Tolson H, Sanchez L, Pines J (2018) Interphysician differences in emergency department length of stay. *Journal of Emergency Medicine* 54(5):702–710.

[11] Sorokin V (2008). *The Queue*. New York Review of Books.

[12] Larson RC (2022). *Model Thinking for Everyday Life*. INFORMS. Chapter 9.

[13] Bren A and Saghafian S (2019). Data-Driven Percentile Optimization for Multi-Class Queueing Systems with Model Ambiguity: Theory and Application. *INFORMS Journal on Optimization*, 1(4), 267-287.