Searching among Queues under Quality Differentiation

For customers seeking service, selecting a provider is rarely without hassle. Customers are typically not fully aware of the quality of a particular service provider; nor do they know for certain how long they have to wait in real time before a provider is available. While quality of services naturally varies from one provider to another, the availability of services also differs due to short-run fluctuations of demand and supply. As a result, customers expend efforts garnering information about both quality and availability of services before making a choice.

Such search frictions, costs incurred in finding a "good" service provider, are pervasive, for instance, in health care services, and are believed, along with high patient demand, to be contributing to long wait times for elective surgery widely documented in many OECD countries with public health systems, e.g. Canada, the UK, etc. As a result, two policy proposals meant to tackle the long wait times have gained considerable traction. The first policy innovation is reducing search frictions of patients by providing them with more information to facilitate their choice over healthcare providers. For instance, British Columbia Minister of Health Services in Canada launched a surgical wait times website that publishes the length of waiting lists at the surgeon level. Meanwhile, search is also made easier by a growing number of websites that provide quality ratings and reviews of healthcare providers, such as zocdoc.com and ratemds.com, just to name a few. The second policy innovation is allowing privately funded health care to reduce the burden on public health systems. The rationale behind these two policies is that reducing search frictions or the system load reduces the average wait.

This paper develops a stylized model to formally investigate the above rationale when customers' search behavior changes in response to changes in the environment caused by theses policies. We also examine how the two policy interventions balance waiting times with other worthy goals for policy makers, such as the quality of providers with which patients receive treatment and average customer welfare. We consider a population of strategic, delay-sensitive customers who arrive to a queueing system with a collection of vertically
differentiated parallel servers. Customers observe neither quality nor queue length of any server upon arrival. Each customer conducts sequential search by paying a fixed search cost each time she probes a server. On each probe, she observes both the quality and queue length of that server, and decides whether to join, or continue searching, to maximize the expected payoff conditioned on the observed state. We note that while quality distribution is exogenously given, the queue length distribution is endogenously determined by customer search behavior. For analytical tractability, we characterize customers' search equilibrium in a mean field model with an infinite number of servers. The equilibrium search strategies are of threshold type: for each of the servers' quality levels, a customer adopts a search threshold such that she joins a server of that quality level if the queue length is below the threshold, and continues to search otherwise. The higher the quality level, the higher the threshold. We show that there always exist either pure or mixed strategy equilibria, and that in any equilibrium, search improves the average quality obtained by customers. We interpret the first policy proposal above (enhancing information provision) as a fall in the search cost, and the second policy proposal (encouraging private practices) as a fall in the arrival rate.

In the model, reducing the search cost induces a search threshold effect and a quality substitution effect on the waiting time. The search threshold effect prompts customers to adopt lower search thresholds in equilibrium. This implies customers get pickier and are only willing to accept a shorter queue for a given quality level. The search threshold effect puts a downward pressure on the average waiting time. In the mean time, however, customers substitute toward high quality servers for which they are willing to wait longer. This “quality substitution effect” puts an upward pressure on the average waiting time. In general, the overall outcome of the two effects is ambiguous, and there are instances in which the second effect dominates the first effect to the extent that the average waiting time may increase.

Moreover, although customers experience higher quality with a lower search cost, it may not compensate for the longer average waiting time ensued, thus deteriorating customer search outcome (the average customer utility upon joining) without improving average customer welfare.
When the arrival rate falls, the lower system load puts an obvious downward pressure on the waiting time, but the search threshold effect and quality substitution effect are more nuanced. We first investigate cases in which the fall of the arrival rate is small so as not to change the equilibrium search thresholds, i.e., the search threshold effect is off. An incremental decrease in arrival rate is particularly relevant for practice as the shift to privately funded health care is likely to be piecemeal. In this case, the downward pressure from the system load effect makes the conditional waiting time for any given quality level shorter. When the original system load is low, we find that more customers end up joining low quality servers, so the quality substitution effect works in the same direction as the system load effect, giving rise to a shorter average waiting time overall. Nevertheless, when the system load is relatively high originally, the direction of the quality substitution effect is reversed, and its upward pressure on the average waiting time may be strong enough to overshadow the downward pressure from the system load effect, thereby increasing the average waiting time. In cases where the fall in arrival rate changes the search thresholds, we also identify cases where qualitatively similar results arise. On a positive note, the overall search outcome does seem to improve with a lower arrival rate despite the potential exacerbation of either quality or waiting time.

We find that two key drivers for our results are substantive quality differentiation and small enough search costs, both of which are characteristic of an environment where search incentives are strong. Our results imply that the two policy initiatives may be effective in both reducing the waiting time and improving experienced quality under some circumstances, but they may be counter-productive in a system where service providers vary markedly in quality, and where search frictions are already relatively low. Given the growing adoption of information technology, it becomes increasingly relevant that health authorities mitigate severe quality disparity in order for policies of reducing search frictions or the system load to shorten waiting times while improving experienced quality.