Managing Supply in the On-Demand Economy: Flexible Workers or Full-Time Employees?

The gig or on-demand economy has gradually become an integral part of the global economy, and it is projected to continue to grow in the coming years. Naturally, not all on-demand services are delivered in the same manner. For example, ride-sharing applications, such as Uber (uber.com) and Lyft (lyft.com), rely solely on self-scheduling independent contractors to fulfill ride requests from customers. In contrast, several on-demand startups, such as Instacart (instacart.com) and Sprig (sprig.com), have recently shifted away from staffing a workforce of independent contractors and currently rely on full-time employees instead. There are also multiple companies, such as Walmart (walmart.com) and Netflix (netflix.com), which rely on a blended workforce i.e., they meld, as a deliberate business strategy, a layer of contingent workers with a core of permanent employees. Given that diverse landscape of alternative workforce models, on-demand service providers face the major problem of deciding on the “right number of right people to staff at the right time”, by appropriately weighing the pertinent tradeoffs. In this paper, we address that problem from an operational point of view.

Modelling framework and overview. We study the staffing problem that an on-demand service provider faces in the context of a stylized queueing model. There are multiple periods, customer demand rates are time-varying, and the agent pool comprises fixed and/or flexible servers. Because part of the agent pool may be flexible, the total number of available servers in our queueing model may be random. We assume that both types of workers, fixed or flexible, provide the same quality of service. However, the two types of workers differ in their unit
operating costs, required working periods, and show-up rates. Specifically, the flexible workforce is not as reliable as the permanent workforce, and may not show up when needed. The service quality is measured by the customers’ waiting experience, and whether or not they abandon. In an initial planning stage, the system manager must decide on the number of fixed servers and the expected number of flexible servers since she cannot enforce a realized number of servers.

Because the optimization problem faced by the system manager is analytically intractable, we rely instead on an asymptotic, many-server, mode of analysis. Specifically, we consider two alternative problem formulations, which correspond to two regimes. The first formulation assumes that uncertainty effects dominate stochastic fluctuations. In this regime, we derive the optimal staffing levels by solving a stochastic-fluid optimization problem which ignores stochastic variability. The second formulation assumes that both uncertainty effects and stochastic fluctuations are negligible. In this regime, we derive the optimal staffing levels by solving a fluid optimization problem instead.

**Main contributions.** We derive optimal staffing policies based on fluid and stochastic-fluid approximations with both stationary and time-varying demand, and rigorously justify their accuracies by quantifying their corresponding errors in large systems. We also show that the optimal staffing policy involves a base capacity and an appropriate uncertainty hedge. We distinguish between three regimes: (i) variability-dominated, where there is no concrete benefit from an uncertainty hedge over the regular square-root staffing hedge; (ii) “moderately” uncertainty-dominated, where the uncertainty hedge embodied in a newsvendor-problem-based solution, which ignores the dependence between uncertainty in supply and
staffing prescription, is extremely accurate; (iii) “strongly” uncertainty-dominated, where there is additional benefit from using an uncertainty hedge which accounts for the dependence between uncertainty in supply and staffing prescription.

Contrary to conventional wisdom in workforce management, we show that a “cheaper resource is not always better”. For example, if a cheap flexible resource entails “high variability”, then it may even be cost-effective to staff a workforce consisting solely of more expensive employees. We also characterize when it would be optimal to use a blended workforce. In particular, if demand rates do not fluctuate over time, then a manager should not blend. If demand rates do fluctuate, then blending may or may not be optimal: A manager should blend only if the fixed resource is cheaper than the flexible resource, but not by much. Then, the manager should rely strictly on the fixed resource in low-demand periods, and rely on a blended workforce in high-demand periods.

We now turn to quantifying the impact of those staffing policies on the quality of service experienced by customers. A well-known maxim in service science is that “variability hurts in queueing”. Thus, we would customers to be disadvantaged by supply-side uncertainty. That is, with other system parameters held fixed, we would expect that customers will be worse off with a more variable pool of servers. In a numerical study, we illustrate that, counter-intuitively, variability in supply need not hurt customer-service levels. In fact, those levels may actually improve with increased variability in supply. The key to understanding this phenomenon lies in quantifying how the service provider hedges against uncertainty in her supply, which in turn depends on system specifics such as customer impatience rates and staffing costs. We characterize that dependence and draw relevant insights.