Incorporating Capacity Constraints (June 18, 2019)

Let

λ = arrival rate of cases,

J = number of judges,

f_h = Pr(judge is of type, i.e., harshness level, h), where $\int f_h \, dh = 1$,

μ_p = service rate of a plea,

μ_t = service rate of a trial,

$P_p(h)$ = Pr(plea bargain | judge type h),

$P_t(h)$ = Pr(trial | judge type h), and

P_h^u = Pr(defendant chooses judge type h | no capacity constraints), which I think can be derived from Can's $p_{i,v,r}(h)$ in her thesis.

Then pleas can be processed by type h judges at rate $Jf_h\mu_p$, and the rate of pleas desired from type h judges is $\lambda P_h^u P_p(h)$.

If $\frac{\lambda P_h^u P_p(h)}{Jf_h\mu_p} > 1$, then judge type h is overloaded. Let harshness level h_1 be such that $\frac{\lambda P_h^u P_p(h)}{Jf_h\mu_p} = 1$, so that judges of type $h \in [0, h_1]$ are overloaded. It needs to be checked whether the function $\frac{\lambda P_h^u P_p(h)}{Jf_h\mu_p}$ is indeed monotone in h, as presumed here. Then define $P_h^c =$ Pr(defendant gets judge type h | capacity constraints), and let $P_h^c = \frac{Jf_h\mu_p}{\lambda P_p(h)}$. So all judges of type $h \in [0, h_1]$ spend all their time on pleas.

Hence, the rate of unsatisfied pleas from these judges is $\lambda \int_0^{h_1} (P_h^u - P_h^c) P_p(h) \, dh$; these pleas need to be transferred to less lenient judges. The slack rate of less lenient judges of type $h > h_1$ is $[Jf_h\mu_p - \lambda P_h^u P_p(h)]^+$. Define harshness level h_2 such that $\lambda \int_0^{h_1} (P_h^u - P_h^c) P_p(h) \, dh = \int_{h_1}^{h_2} [Jf_k\mu_p - P_h^u P_p(h)] \, dh$. Then set $P_h^c = \frac{Jf_h\mu_p}{\lambda P_p(h)}$ for $h \in (h_1, h_2]$, so that these judges also spend all of their time on pleas (due to overflow from the judges with $h \in [0, h_1]$).

Hence, the rate of unsatisfied trials from judges of type $h \in [0, h_2]$ is $\lambda \int_0^{h_2} \lambda P_h^c P_t(h) \, dh$. Since the outcomes of trials do not depend on the judge, we don't have to worry which

judges handle which trials. But we do need an aggregate capacity constraint requiring that judges who are of types $h \in (h_2, \infty)$ can process all their own pleas and all of the trials. This constraint is

$$\frac{J \int_{h_2}^{\infty} f_h \, dh}{\lambda} \left(\frac{\int_{h_2}^{\infty} P_h^u P_p(h) \, dh}{\mu_p} + \frac{\int_0^{h_2} P_h^c P_t(h) \, dh + \int_{h_2}^{\infty} P_h^u P_t(h) \, dh}{\mu_t} \right) \leq 1.$$