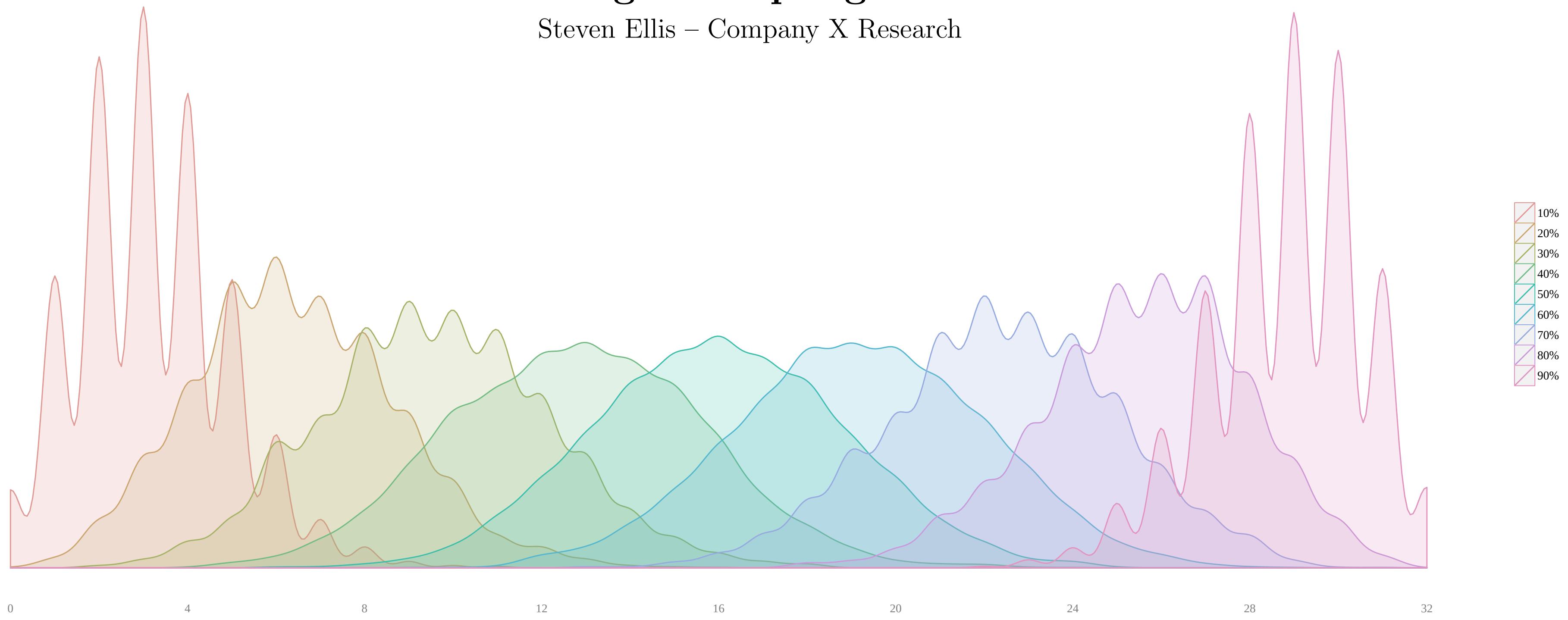
Modeling a Sampling Problem

Steven Ellis – Company X Research



Server Load & Request Allotment

Load balancing is a computer networking method for distributing workloads across multiple computing resources, such as computers, a computer cluster, network links, central processing units or disk drives. Load balancing aims to optimize resource use, maximize throughput, minimize response time, and avoid overload of any one of the resources. Using multiple components with load balancing instead of a single component may increase reliability through redundancy. Load balancing is usually provided by dedicated software or hardware, such as a multilayer switch or a Domain Name System server process.

Operational Situation

The heroku.com stack only supports single threaded requests. Even if your application were to fork and support handling multiple requests at once, the routing mesh will never serve more than a single request to a dyno at a time. A load-balancer can manage 288 dynes, blocked into groups of 32. Given 32 dynos, new requests are randomly assigned to each dyno.

SRSWOR

We iterate along these ranges (1:32, 33:65) to see how common a given persist load is likely to be within each load-balancer. Simple Random Sampling Without Replacement (SRSWOR) models situations where individuals, once picked as part of a sample from a population, are not returned to the population for potential re-sampling.

Predicted Server Load

Persists	Min	1st Q.	Median	Mean	3rd Q.	Max
10%	0	2	3	3.2	4	11
20%	1	5	6	6.4	8	14
30%	1	8	10	9.6	11	20
40%	4	11	13	12.8	15	22
50%	7	14	16	16	18	26
60%	10	17	19	19.2	21	30
70%	13	21	23	22.4	24	30
80%	17	24	26	25.6	27	32
90%	22	28	29	28.8	30	32

Server Options[a]

Persists per server	32	24	16	Pt.-to-Pt.
Model A	x	y	z	q
Model B	x	y	z	q
Model C	x	y	z	q
Model D	x	y	z	q

[a] BW = bandwidth in mbps