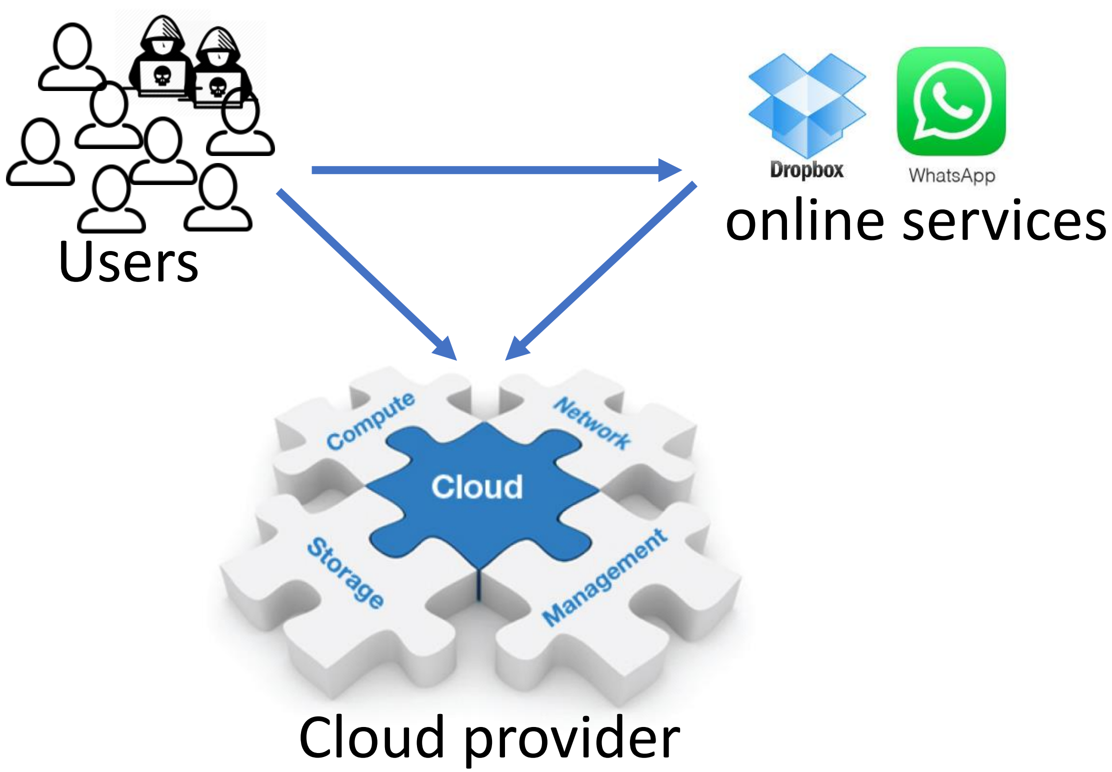
Modeling Approaches to Classification of Cloud Users via Shuffling

Yudong Yang, Vishal Misra, Dan Rubenstein

Columbia University

Problem Description

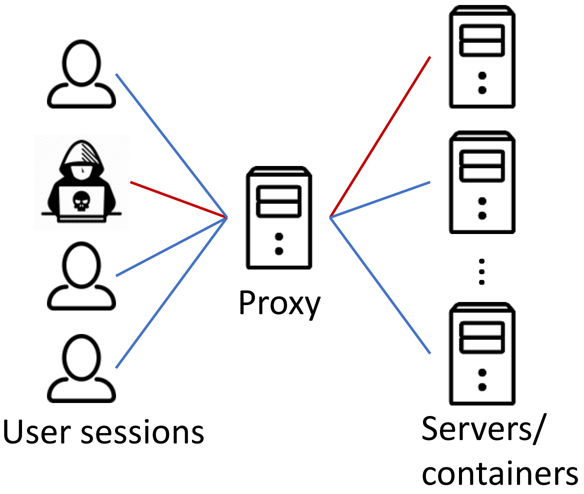


Problem Description

- DDoS attacks

- Attackers abuse the resources of the (back-end) servers

- Servers can be observed to be "attacked" or "not attacked"



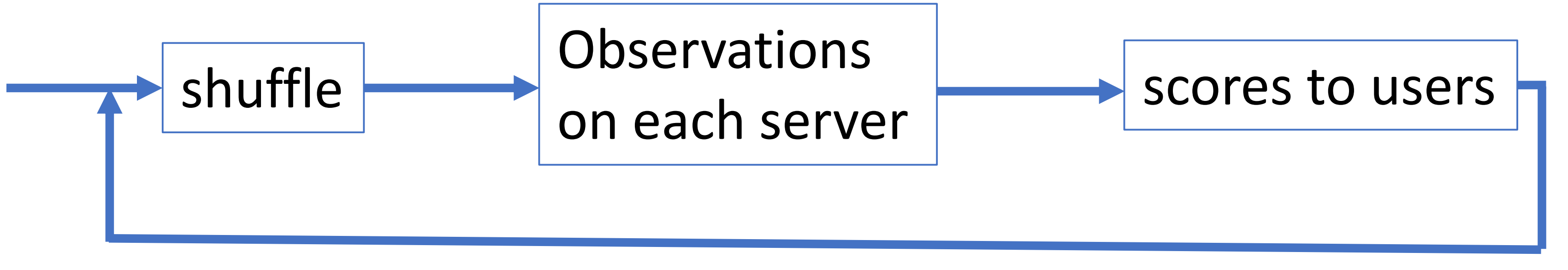
How to detect malicious users?

Periodically shuffle:

1. Shuffle(randomize) the mapping of sessions to servers

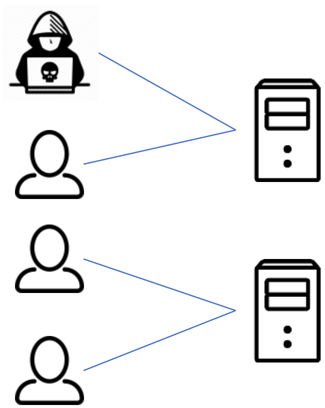
2. Observe the server status

3. Score users based on observation, e.g. +1 or -1

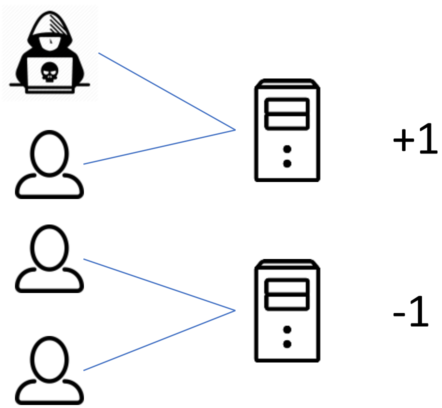


Intuition: The attackers over time have higher scores.

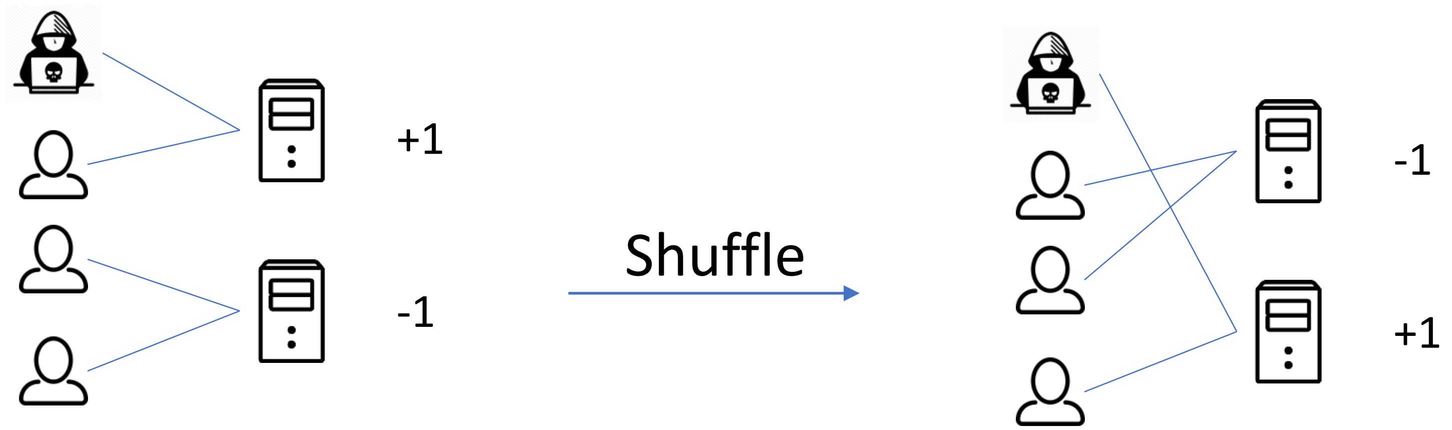
Shuffle Example



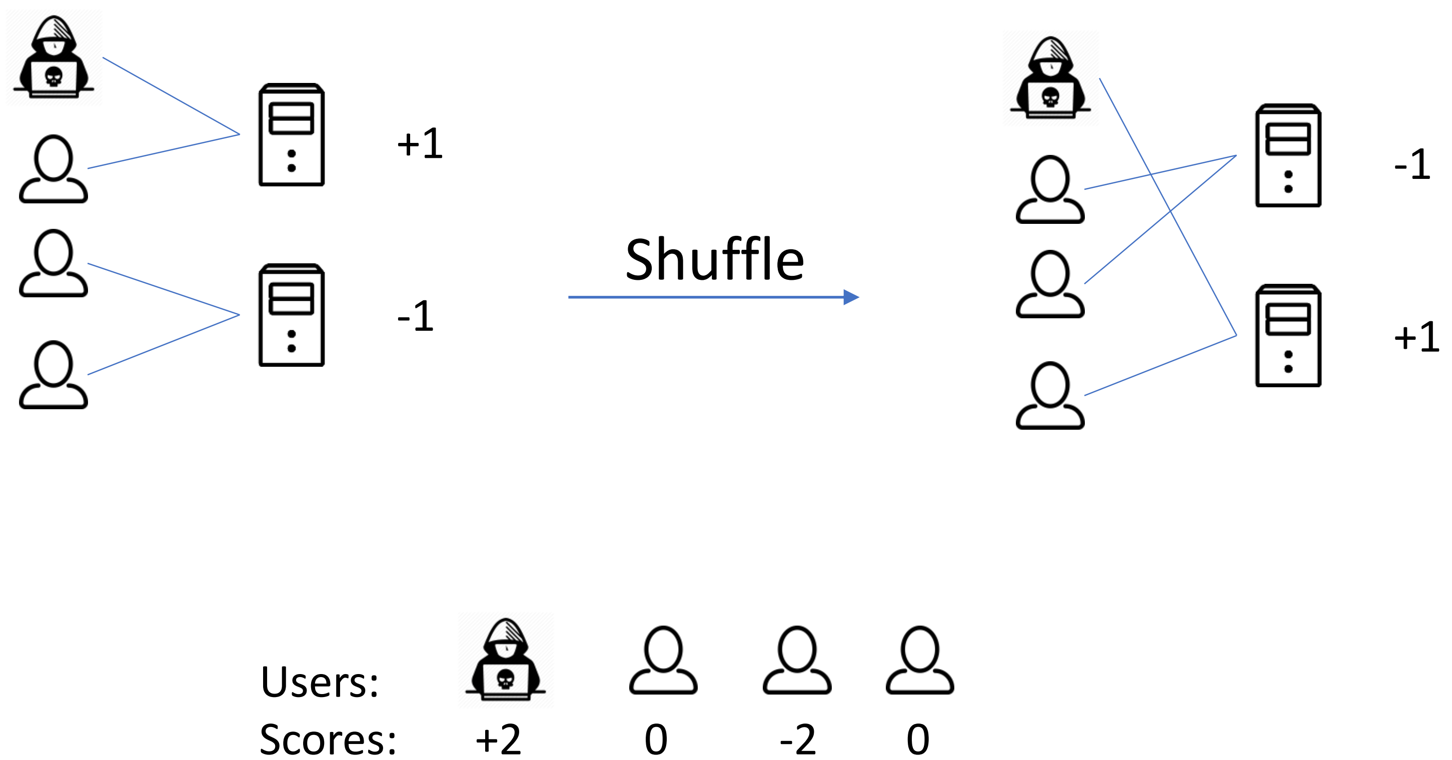
Shuffle Example



Shuffle Example



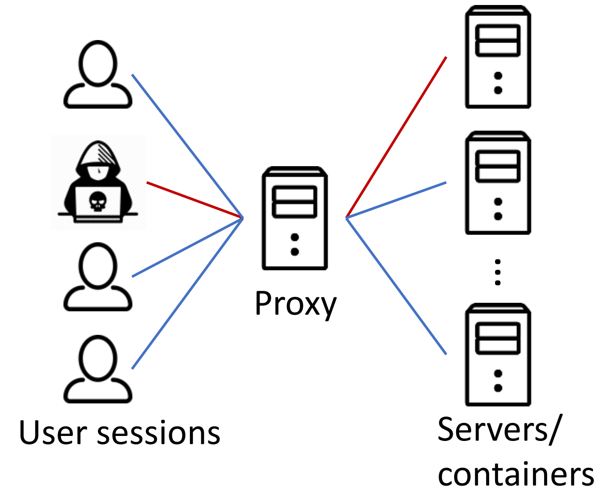
Shuffle Example



Questions

- What is the right scoring function to use?
- How many shuffles are needed?
- What is the optimal group size?

Model



M servers

N sessions

(*K* attackings, *N-K* benign sessions)

Server capacity, *A*. (Server is online if $a \leq A$)

After one shuffle, the server can be:

	Non-attacked($a \leq A$)	Attacked ($a > A$)
Score:	γ_0	γ_1

After s shuffles, the sum of score of session i:

$$\sum_{j=1}^{s} \gamma(x_j^i)$$

Probabilities

There are N sessions (K attacking sessions , U = N - K legitimate sessions)

- Define probability $a(v, k, N, K)$, having k attackers in a random selected subset of v sessions.

$$a(v, k, N, K) = \binom{v}{k}\binom{N-v}{K-k} / \binom{N}{K}$$

Where v=N/M, the average number of sessions per server (group size)

Probabilities

For a given **attacking** session i, the probability i is on:

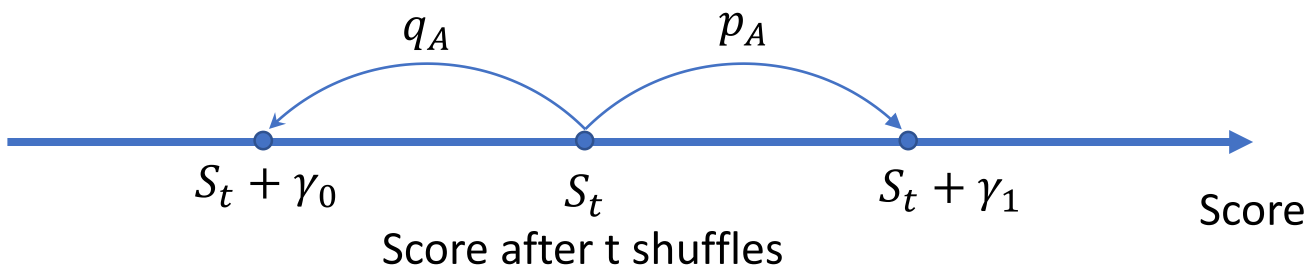
Non-attacked	Attacked
$q_A = \sum_{k=0}^{A-1} a(v-1, k, U, K-1)$	$p_A = 1 - q_A$

For a given **legitimate** session i, the probability i is on:

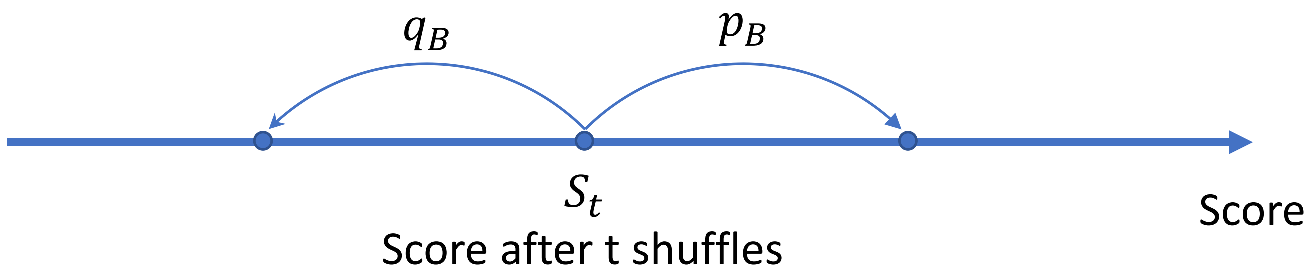
Non-attacked	Attacked
$q_B = \sum_{k=0}^{A} a(v-1, k, U-1, K)$	$p_{\mathrm{B}} = 1 - q_B$

Random walk model

- For attacking sessions



- For legitimate sessions



Mean and Variance

After s shuffles:

For **attacking** sessions,

Mean	Variance
$s(p_A\gamma_1 + q_A\gamma_0)$	$sp_Aq_A(\gamma_0 - \gamma_1)^2$

For **legitimate** sessions,

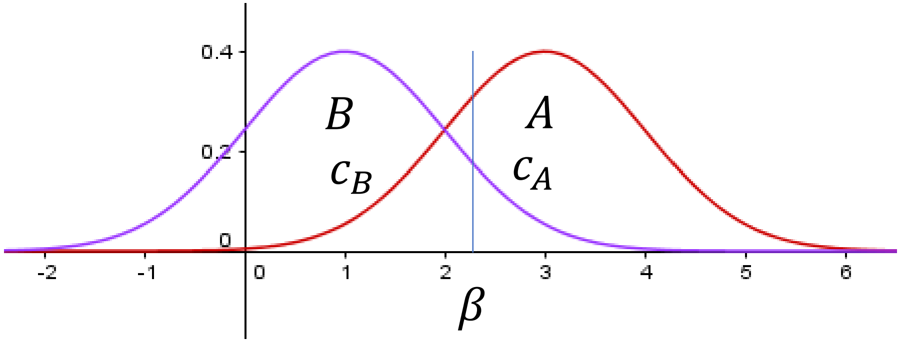
Mean	Variance
$s(p_B\gamma_1 + q_B\gamma_0)$	$sp_Bq_B(\gamma_0 - \gamma_1)^2$

Question

- Question: what is the minimum number of shuffles needed?

Accuracy Level

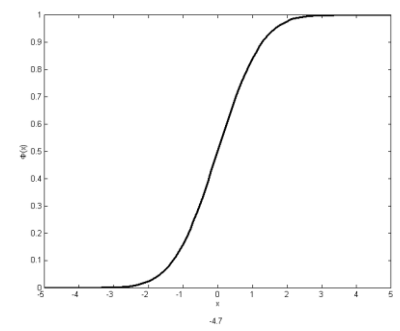
- Decision threshold β
- Accuracy level c_A, c_B



Number of Shuffles

- Question:

Given accuracy level c_A, c_B, what is the minimum number of shuffles needed?



Approximate with normal distribution:

$$c_A = \Phi\left(\frac{\mu_A^s - \beta}{\sigma_A^s}\right), \quad c_B = \Phi\left(\frac{\beta - \mu_B^s}{\sigma_B^s}\right)$$

Number of shuffles:

$$\begin{aligned} \min \quad & s \\ \text{s.t.} \quad & s > 0 \\ & \beta_A = \mu_A^s - \Phi^{-1}(c_A)\sigma_A^s \\ & \beta_B = \mu_B^s + \Phi^{-1}(c_B)\sigma_B^s \\ & \beta_A \geq \beta_B \end{aligned}$$

Solution

$$\begin{aligned} \min \quad & s \\ \text{s.t.} \quad & s > 0 \\ & \beta_A = \mu_A^s - \Phi^{-1}(c_A)\sigma_A^s \\ & \beta_B = \mu_B^s + \Phi^{-1}(c_B)\sigma_B^s \\ & \beta_A \geq \beta_B \end{aligned}$$

s is solved by:

$$s^* = (\gamma_1 - \gamma_0)^2 \left(\frac{\Phi^{-1}(c_A)\sqrt{p_A q_A} + \Phi^{-1}(c_B)\sqrt{p_B q_B}}{\mu_A - \mu_B} \right)^2$$

decision threshold β:

$$\beta^* = \mu_A^s - \Phi^{-1}(c_A)\sigma_A^s = \mu_B^s + \Phi^{-1}(c_B)\sigma_B^s$$

Scoring function

- What is the right scoring function to use? γ_0, γ_1

$$s^* = C_0 \left(\frac{\gamma_1 - \gamma_0}{(p_A - p_B)\gamma_1 + (q_A - q_B)\gamma_0} \right)^2$$

$$\frac{\partial s^*}{\partial \gamma_0} = \frac{(\gamma_0 - 1)(p_A + q_A - p_B - q_B)}{(p_A - p_B) + (q_A - q_B)\gamma_0} = 0$$

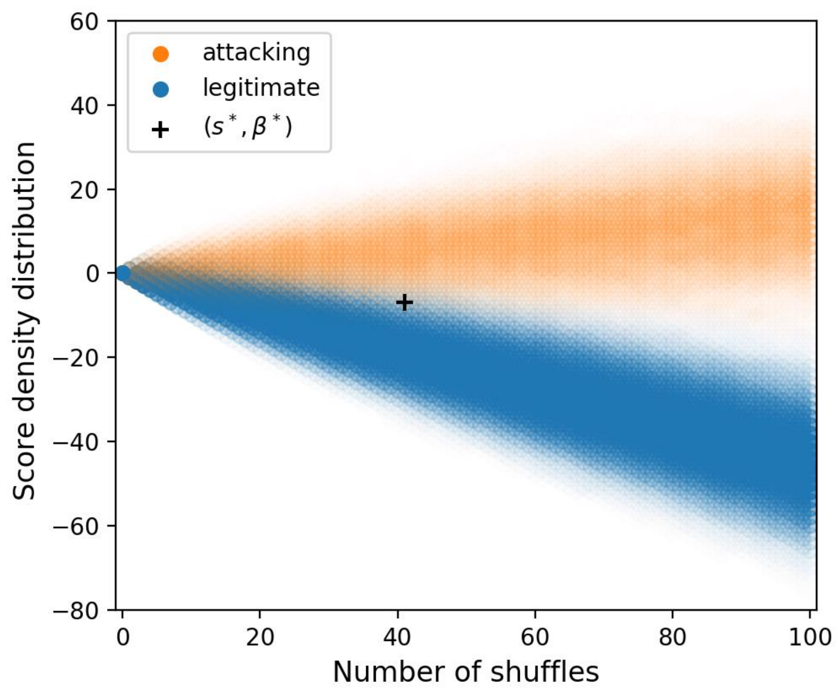
All scoring functions are have the same s^*

Experiment

- N = 12000; K = 2000; M =1000 ; A=2

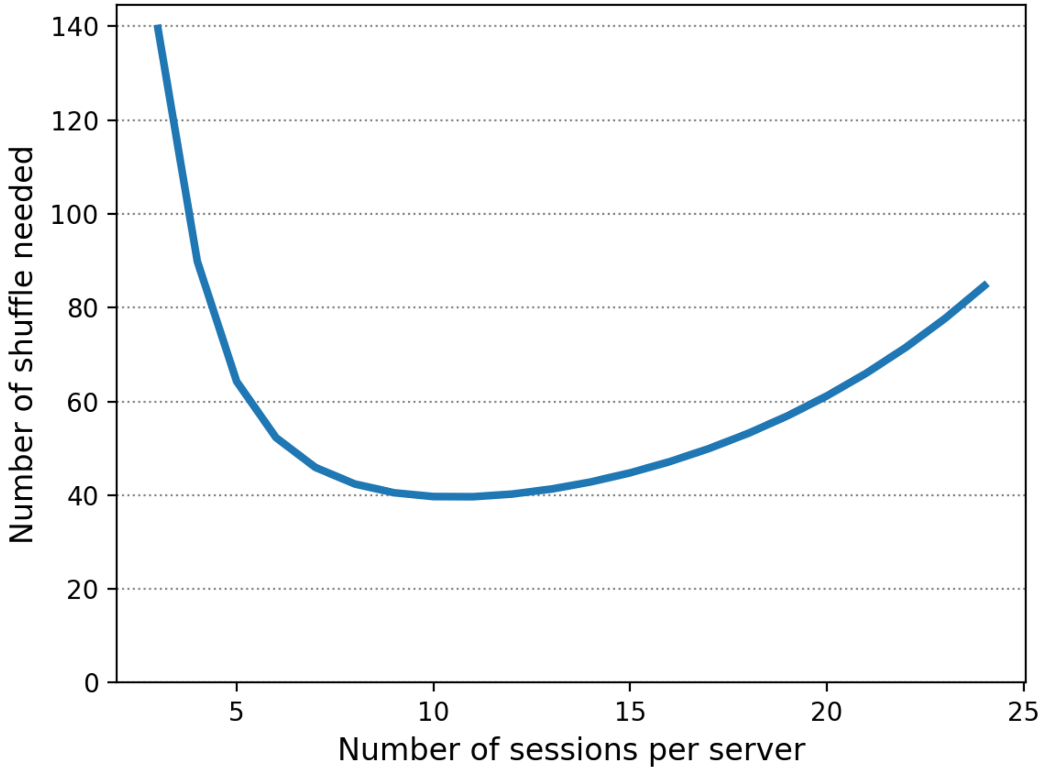
γ_1 = 1; γ_0 = -1;

c_A = c_B = 0.977, thus $\Phi^{-1}(c_A) = \Phi^{-1}(c_A) \approx 2$



Optimal Group Size

- What is the optimal group size?



Thank You!