Revisiting Membership Inference Under Realistic Assumptions

Bargav Jayaraman†, Lingxiao Wang§, Katherine Knipmeyer†,
Quanquan Gu§ and David Evans†

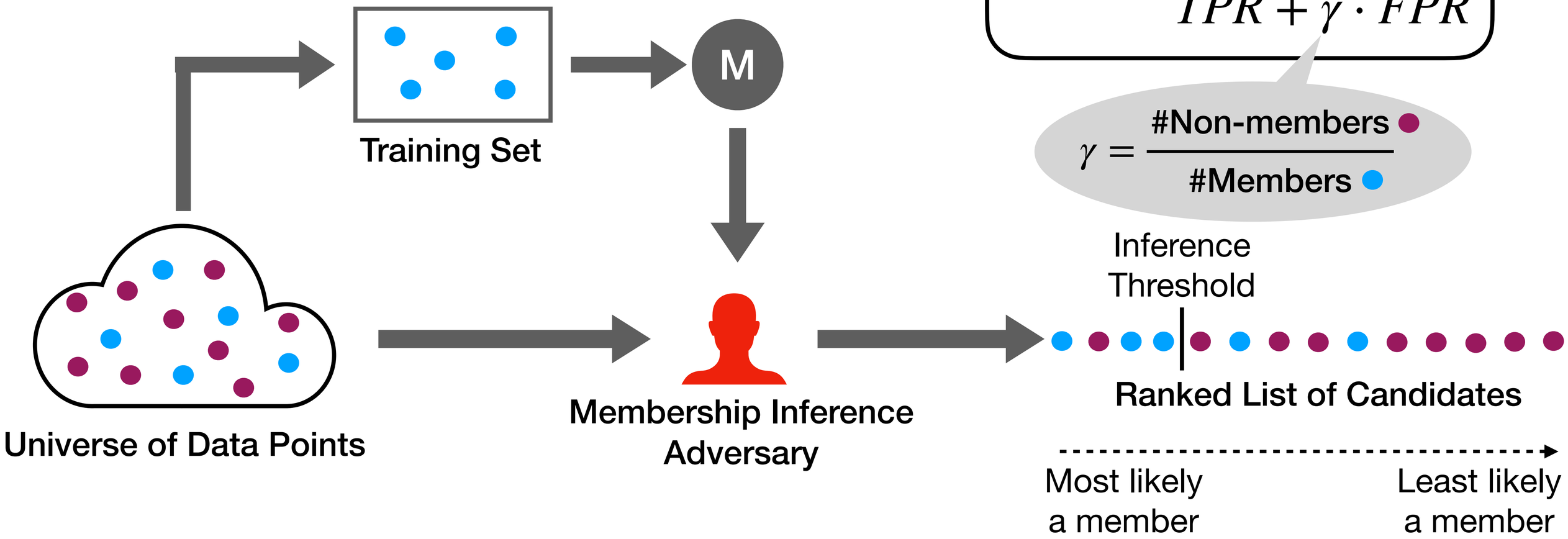
†University of Virginia
§University of California Los Angeles

Membership Inference (MI) Attack

$$Adv = TPR - FPR$$

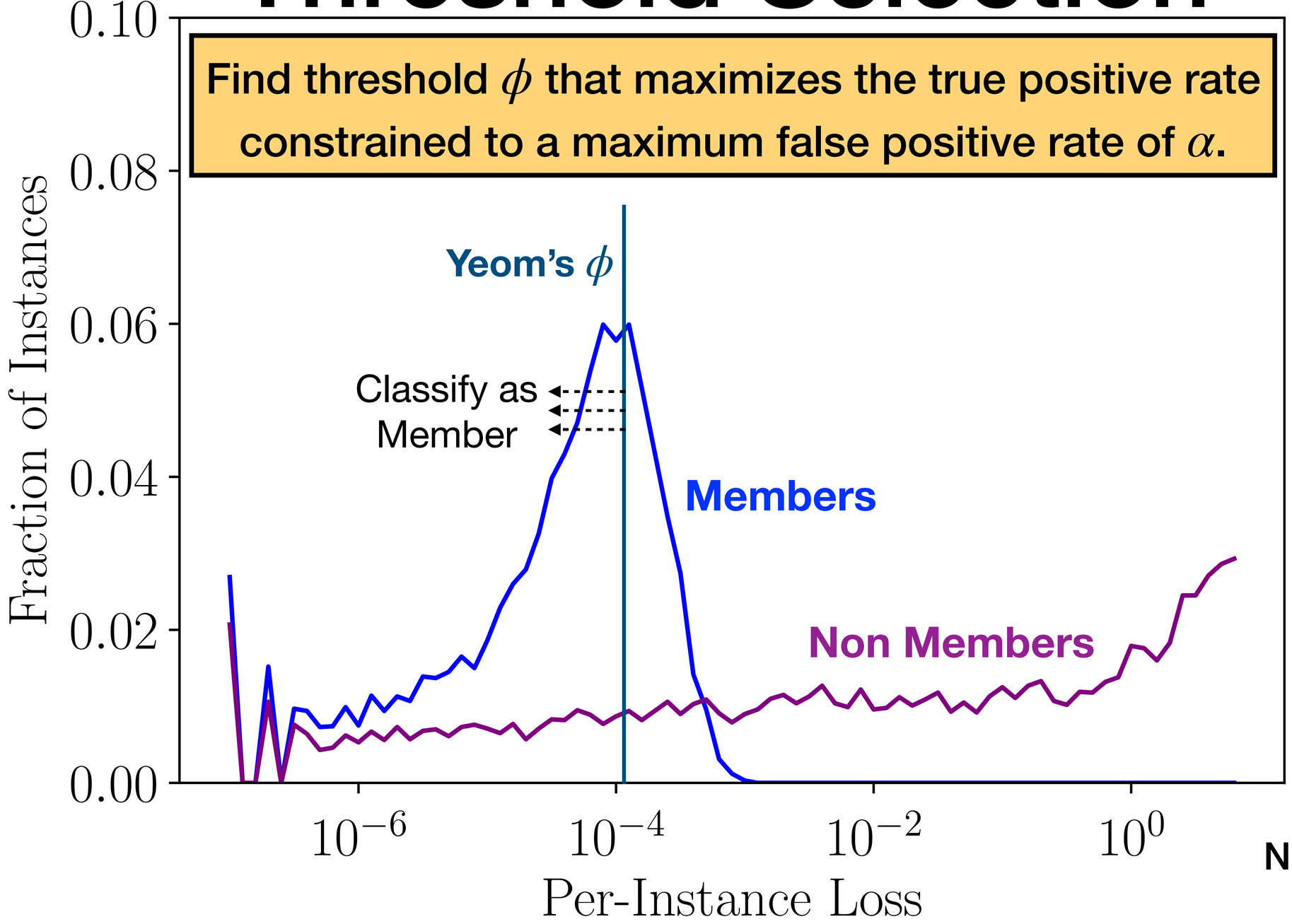
$$PPV = \frac{TPR}{TPR + \gamma \cdot FPR}$$

$$\gamma = \frac{\#\text{Non-members}}{\#\text{Members}}$$



Threshold Selection

Find threshold ϕ that maximizes the true positive rate constrained to a maximum false positive rate of α.

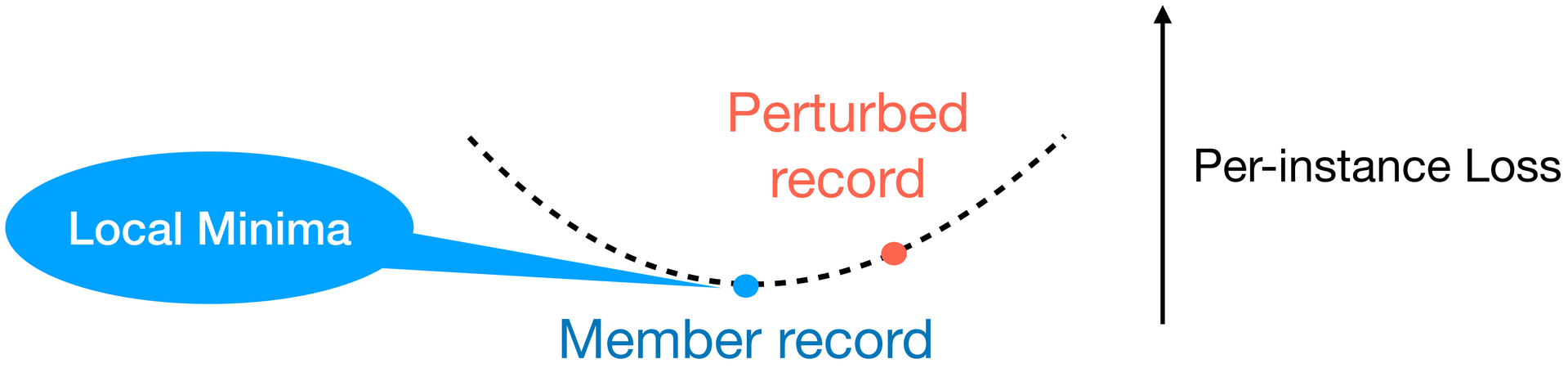


Neural Network on Purchase-100X

New Attack: Merlin

(MEasuring Relative Loss In Neighbourhood)

Key Intuition:
Per-instance loss of members tend to increase when perturbed.



Merlin Algorithm

Input: query record z, model $\mathscr{M}_S$ trained on data set S, number of repeats T, standard deviation σ, threshold ϕ

Output: membership prediction of z (0 or 1)

for T runs do:

  $\zeta \sim \mathcal{N}(0, \sigma^2 I)$ // sample Gaussian noise

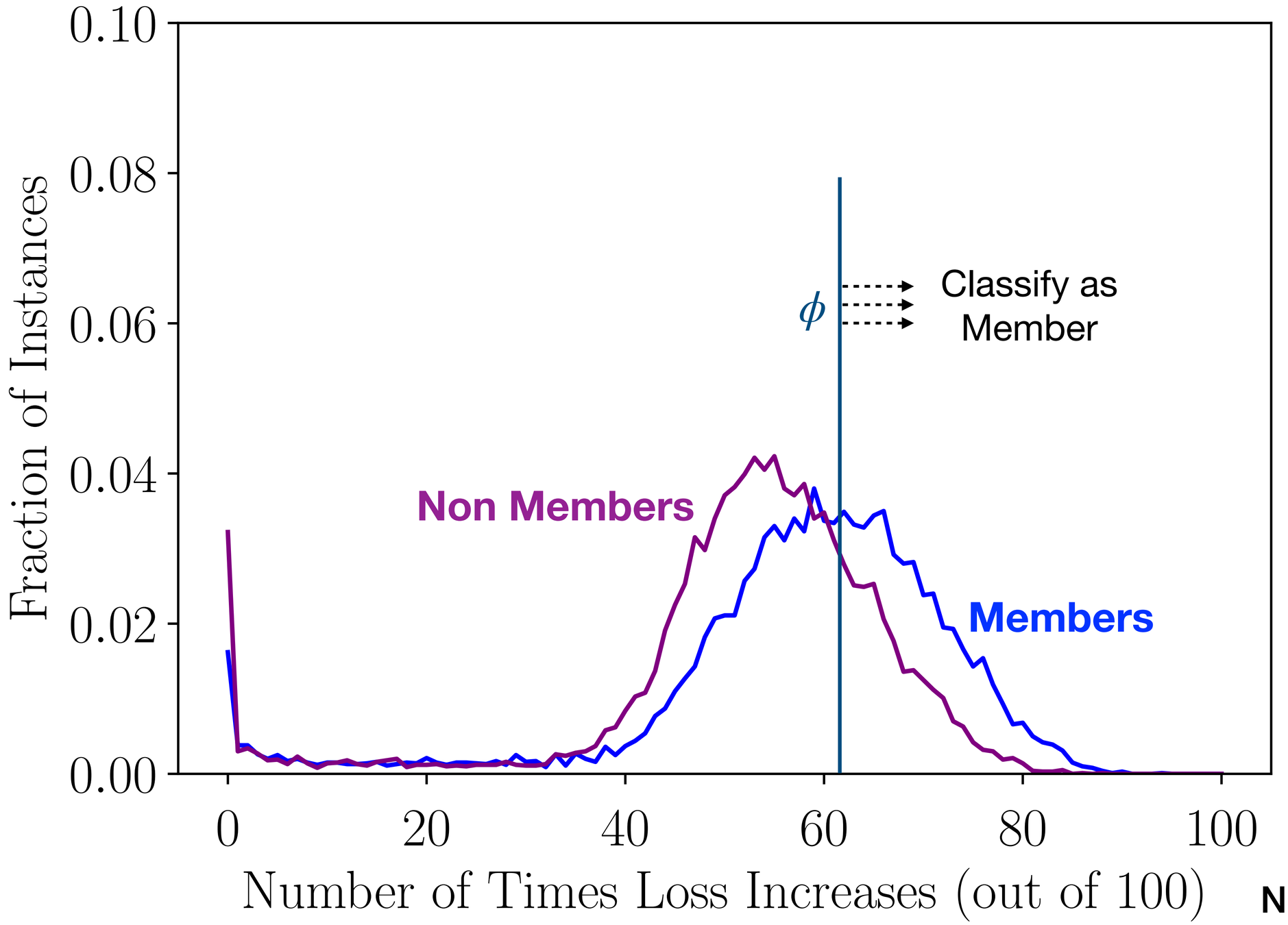
  if $\ell(z + \zeta, \mathscr{M}_S) > \ell(z, \mathscr{M}_S)$ then:

    increment $count$

  end

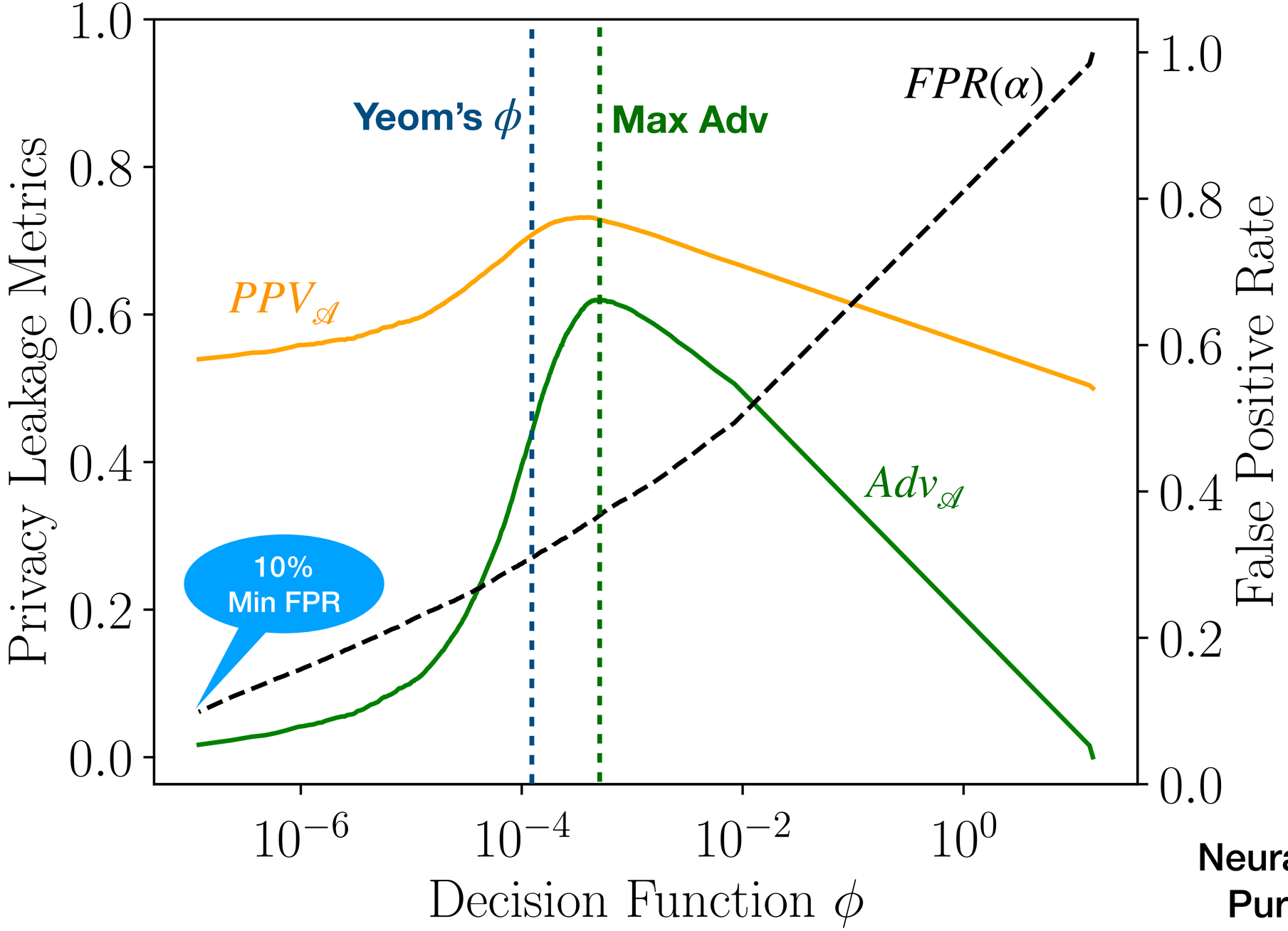
end

return $count/T \geq \phi$ // '1' if member



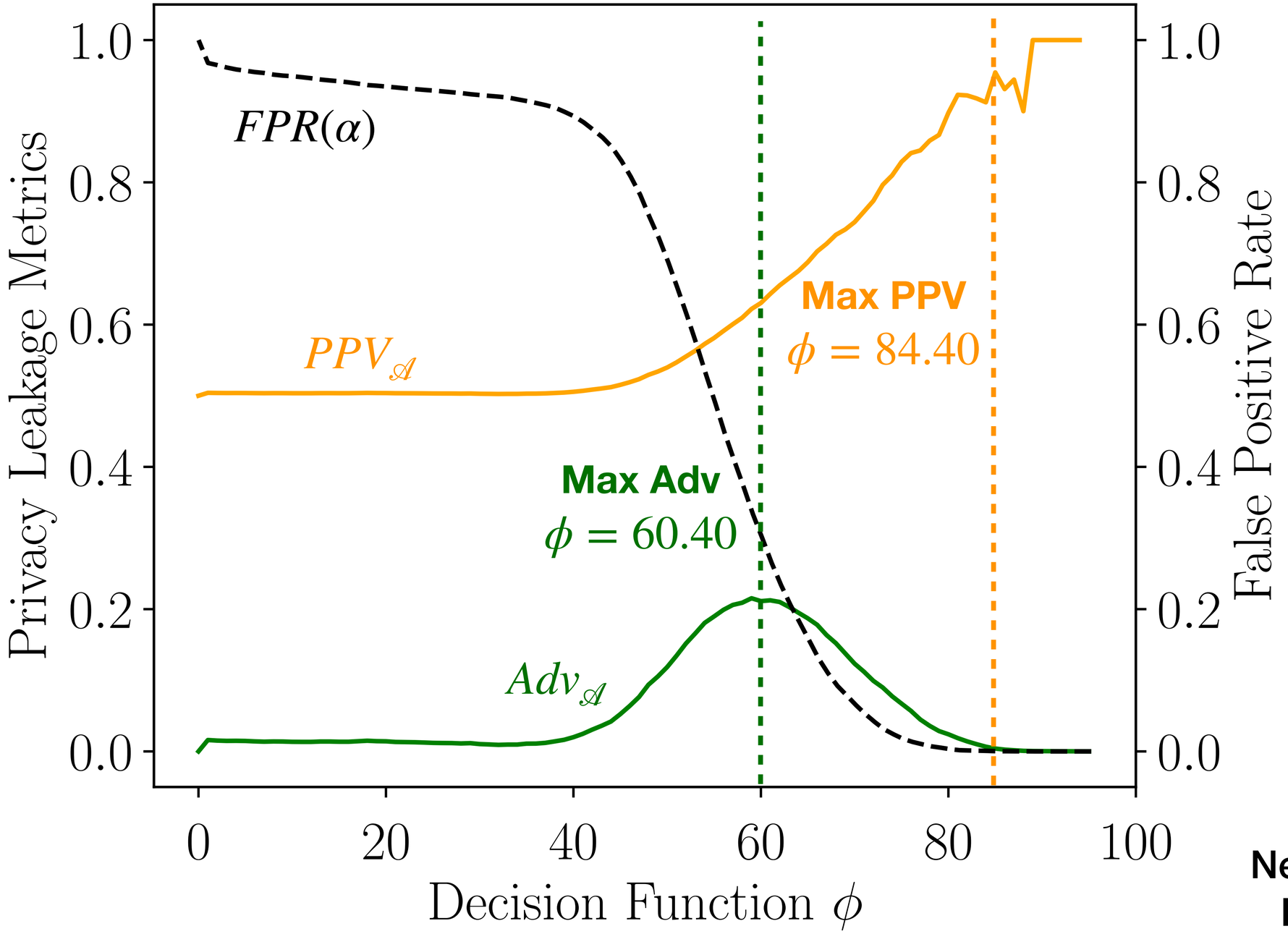
Neural Network on Purchase-100X

Effectiveness of Yeom at different thresholds



Neural Network on Purchase-100X

Effectiveness of Merlin at different thresholds

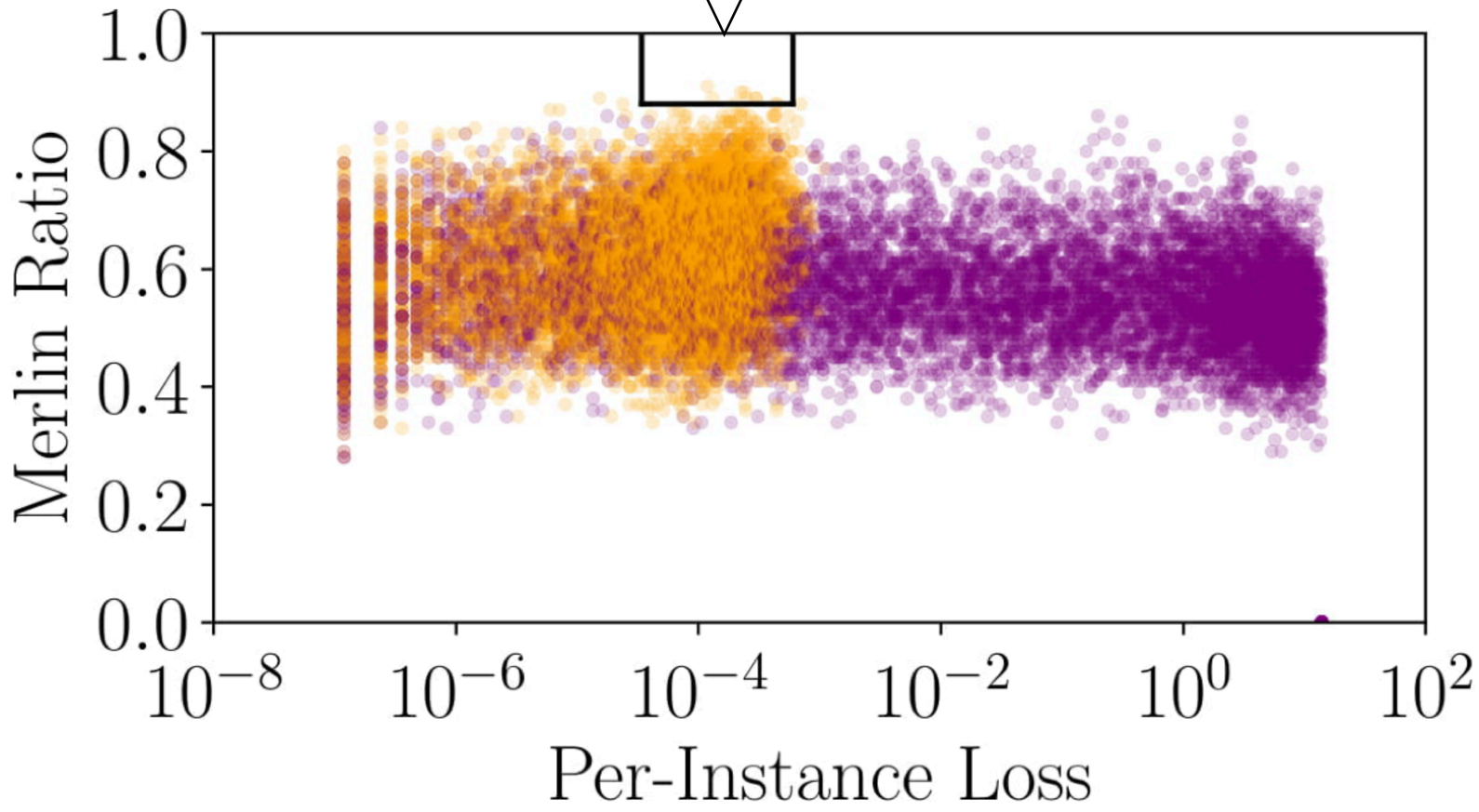


Neural Network on Purchase-100X

New Attack: Morgan

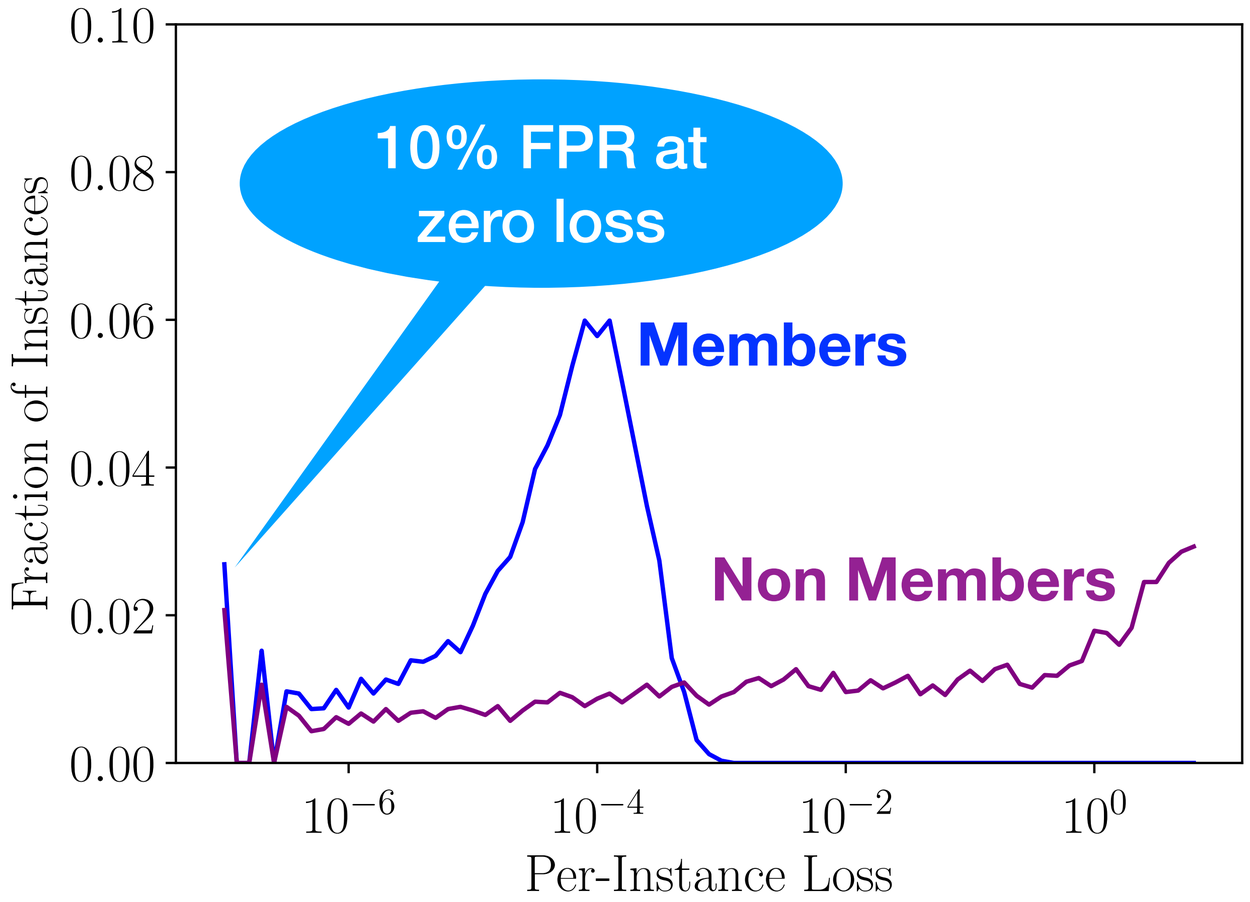
(Measuring lOss Relative Greater Around Neighbourhood)

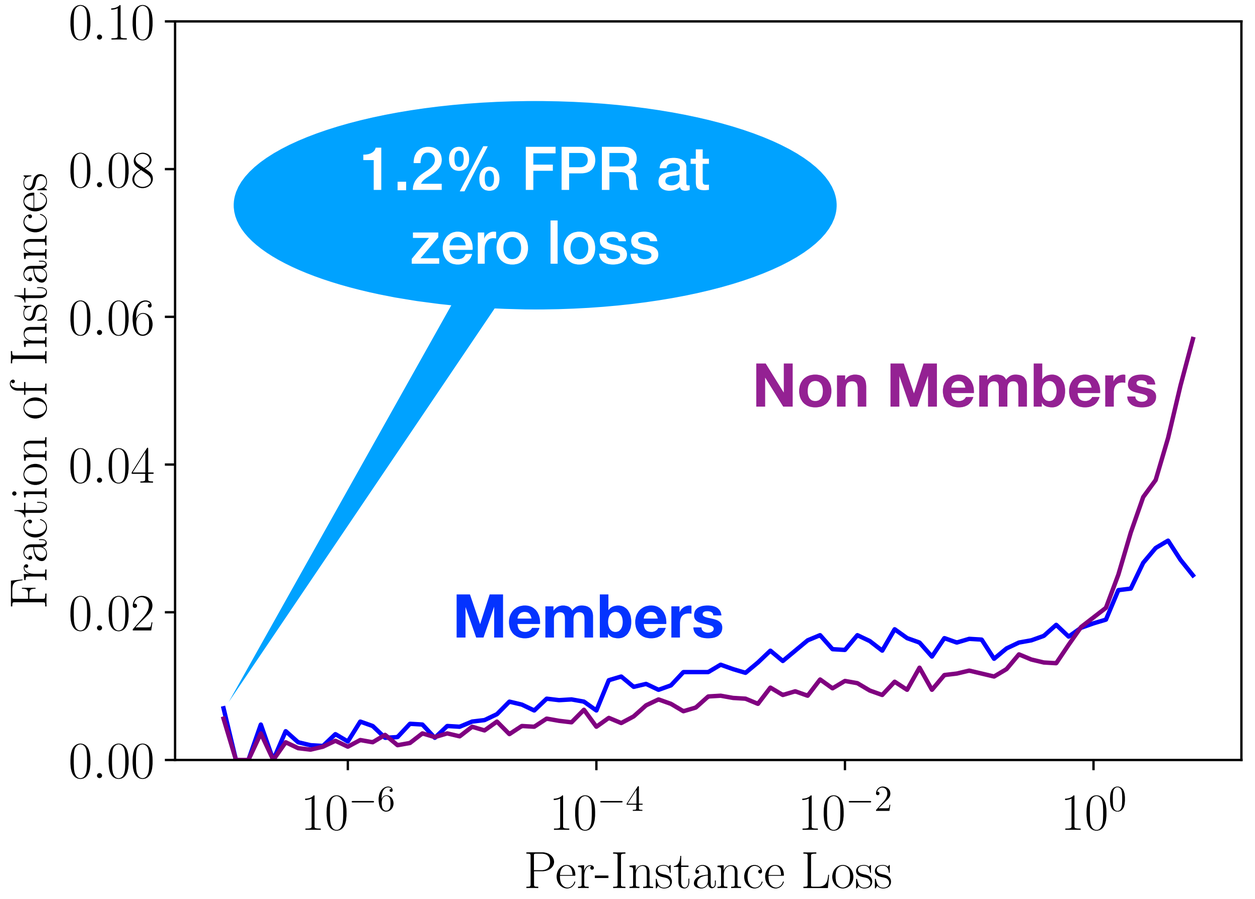
Combining Yeom and Merlin thresholds leads to a stronger attack, Morgan, that can identify the most vulnerable members with very high (~100%) confidence.



Effect of Privacy Noise on Yeom

No Privacy → **GDP with $\epsilon = 100$**

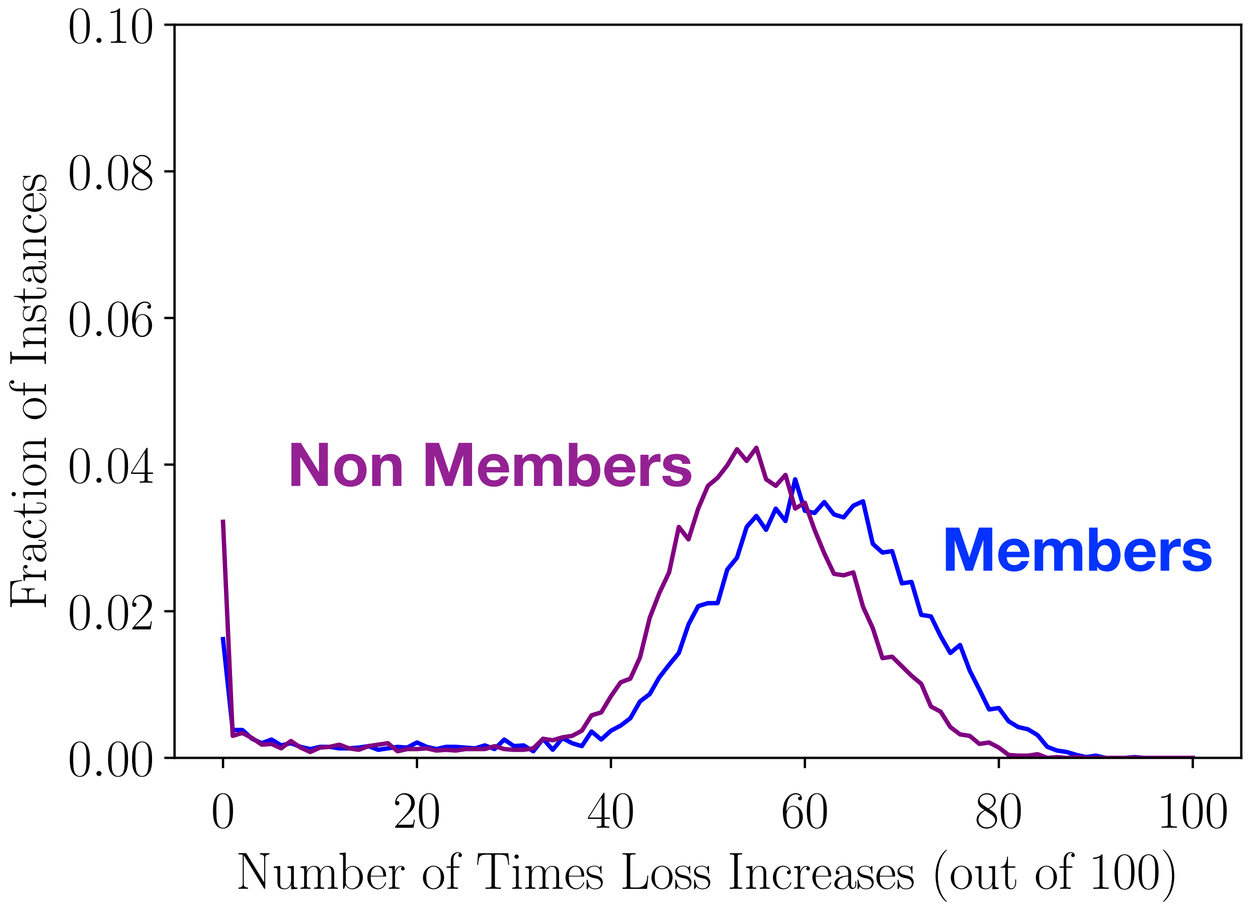


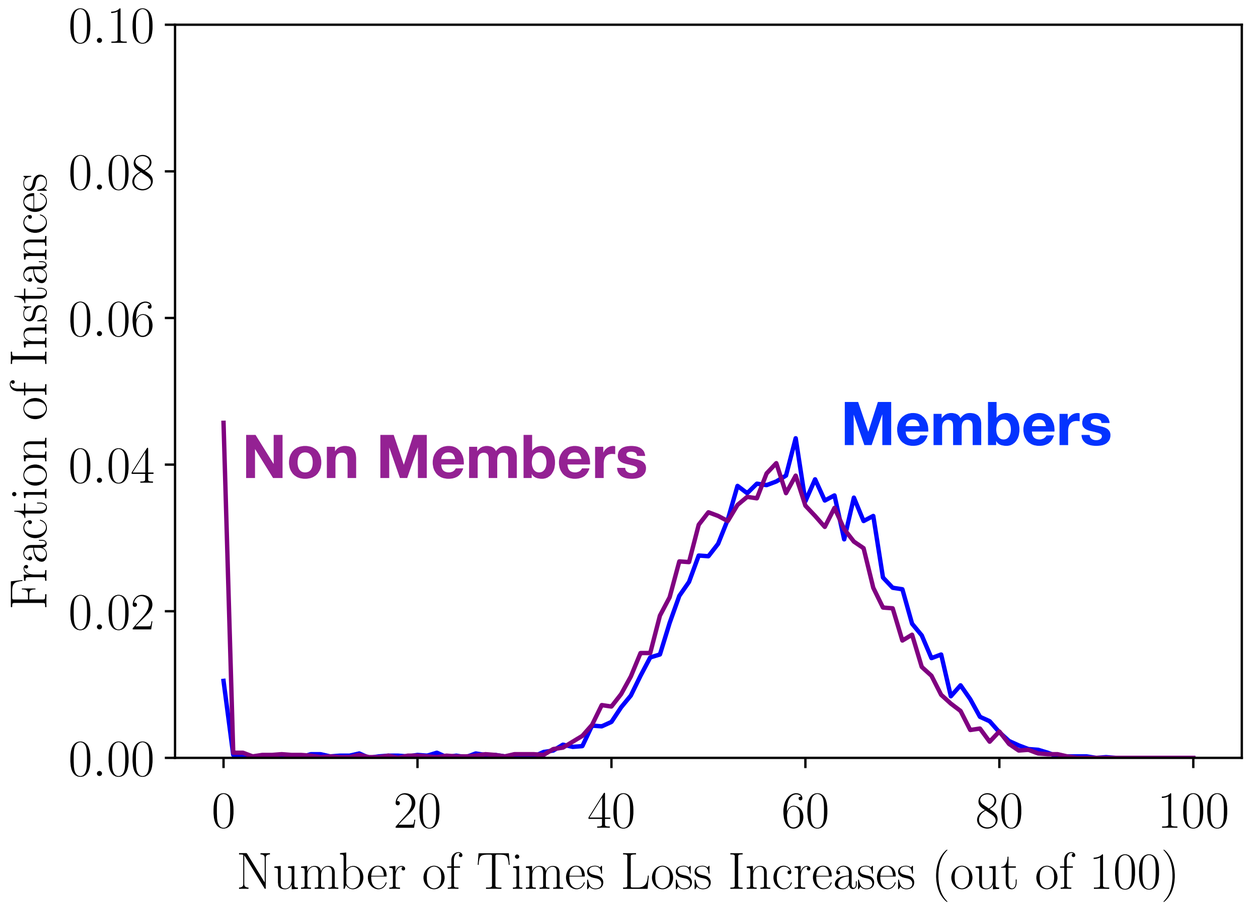


Neural Network on Purchase-100X

Effect of Privacy Noise on Merlin

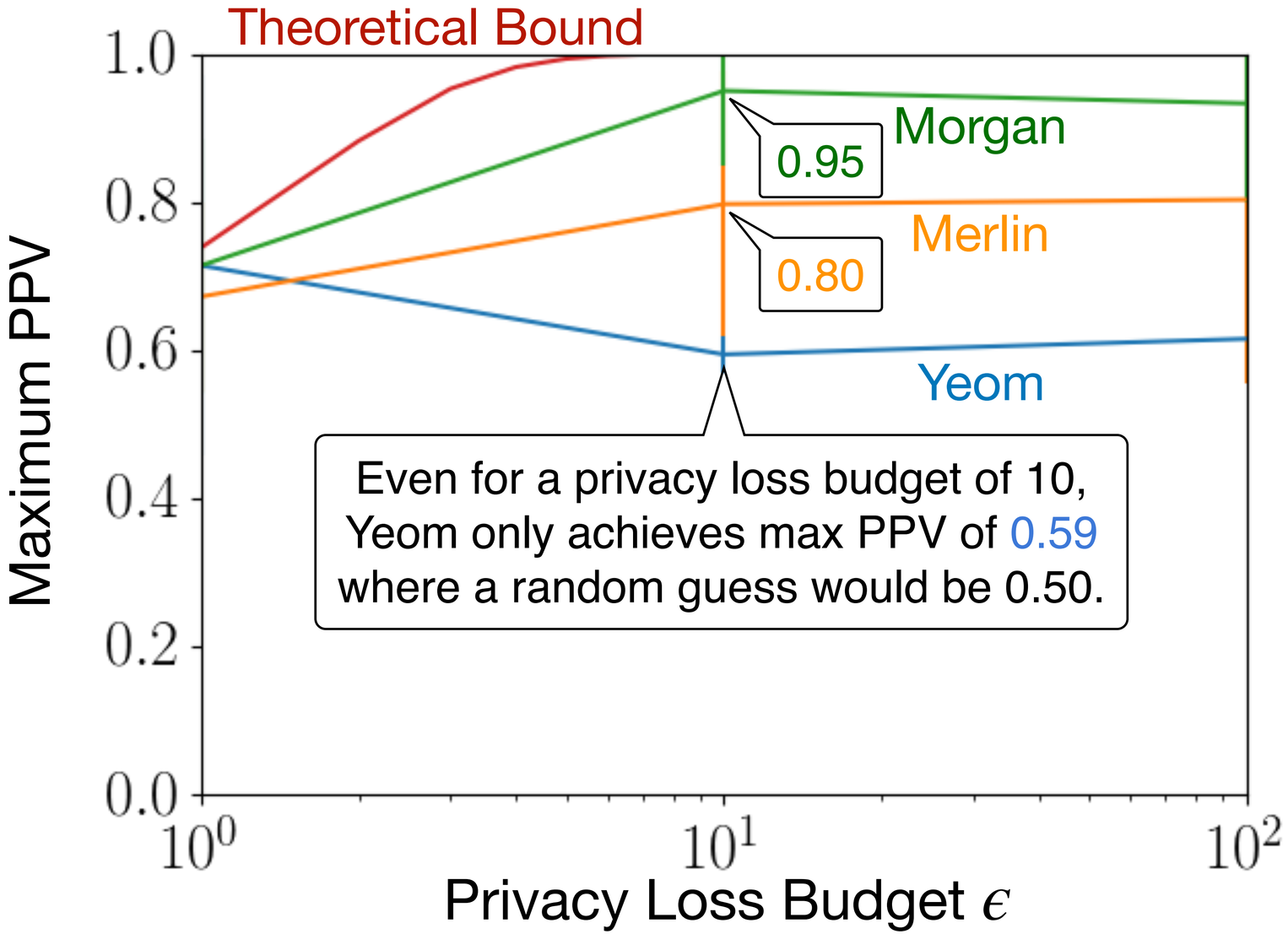
No Privacy → **GDP with $\epsilon = 100$**





Neural Network on Purchase-100X

Max PPV comparison on Purchase-100X data set (Varying Privacy Loss Budget)



What about Imbalanced Priors?

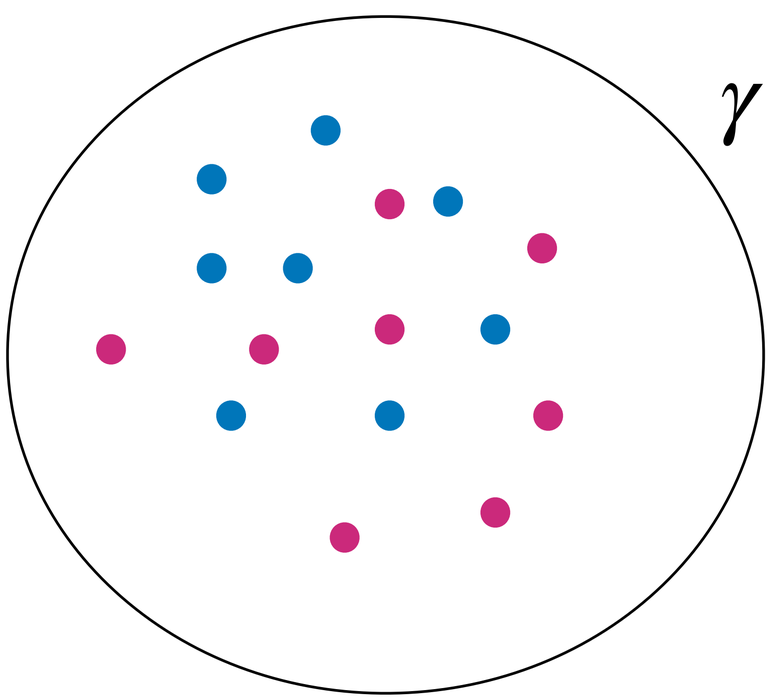
Balanced Prior

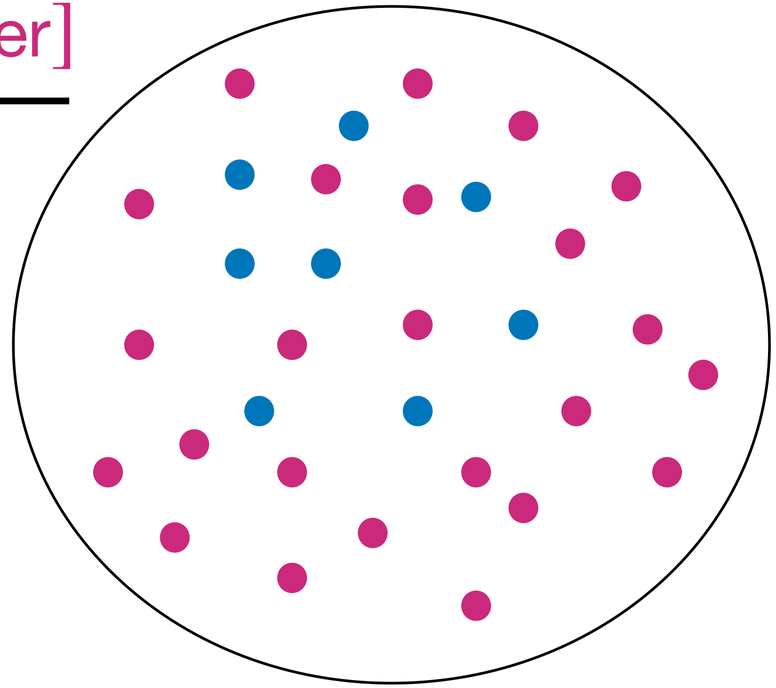
$Pr[\text{Member}] = Pr[\text{Non Member}]$

Imbalanced Prior

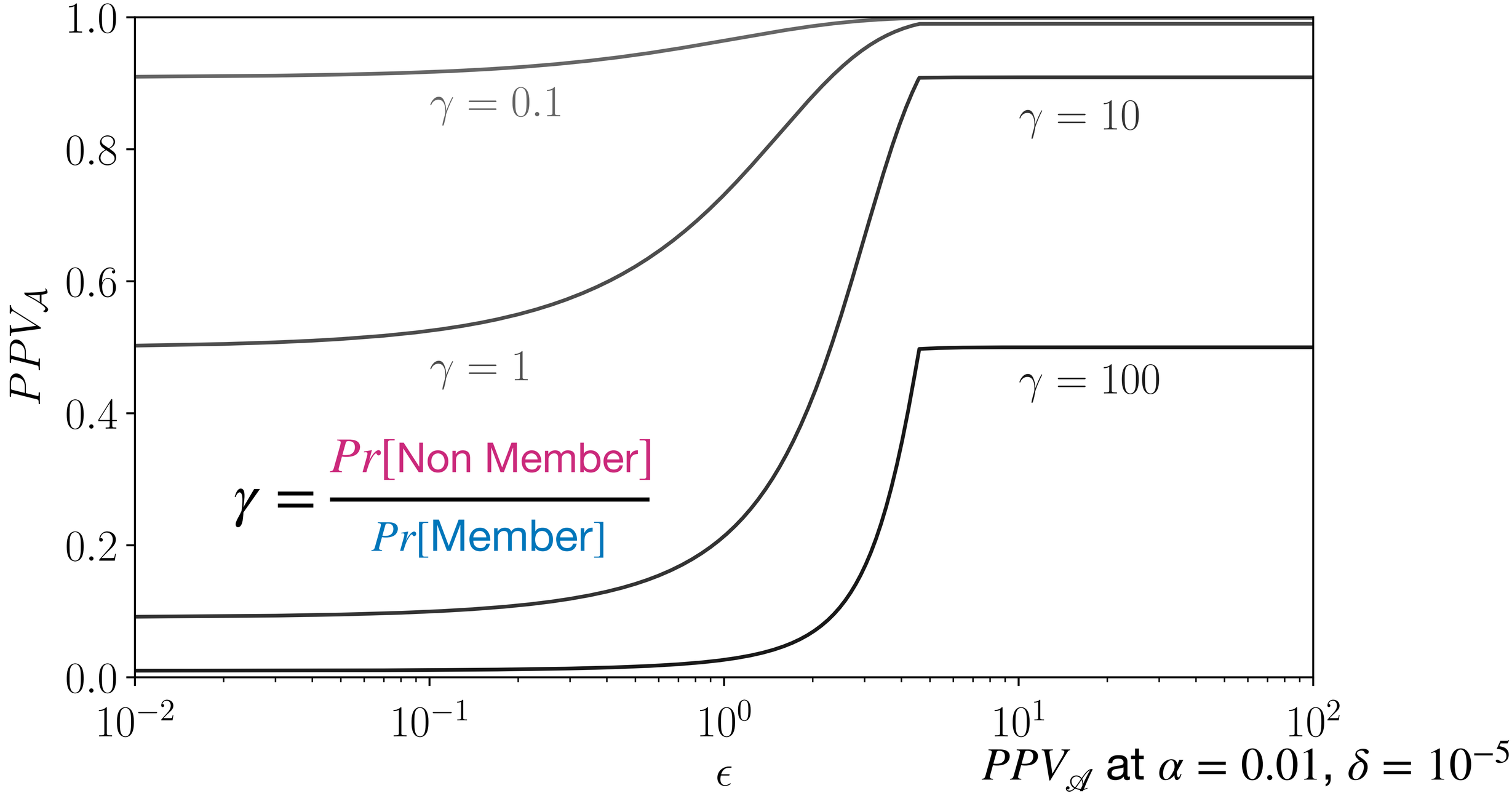
$Pr[\text{Member}] < Pr[\text{Non Member}]$

$$\gamma = \frac{Pr[\text{Non Member}]}{Pr[\text{Member}]}$$

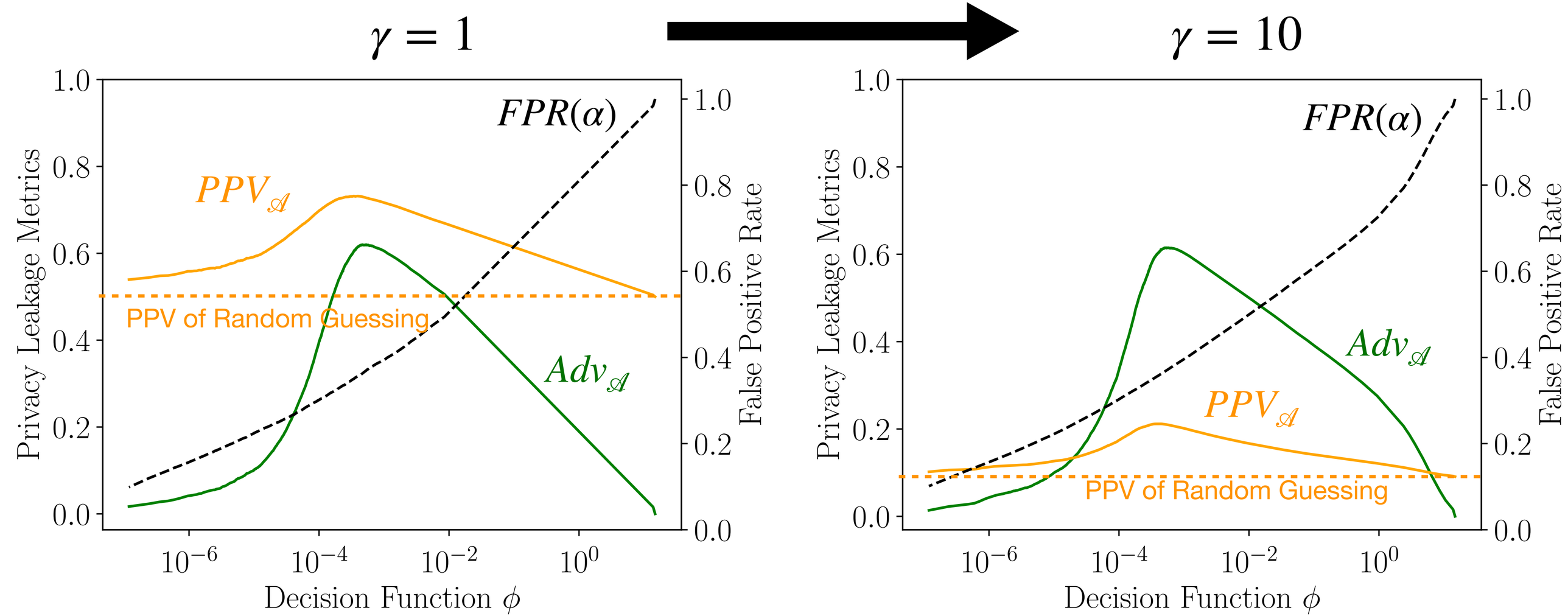




Effect of Prior on the PPV Metric

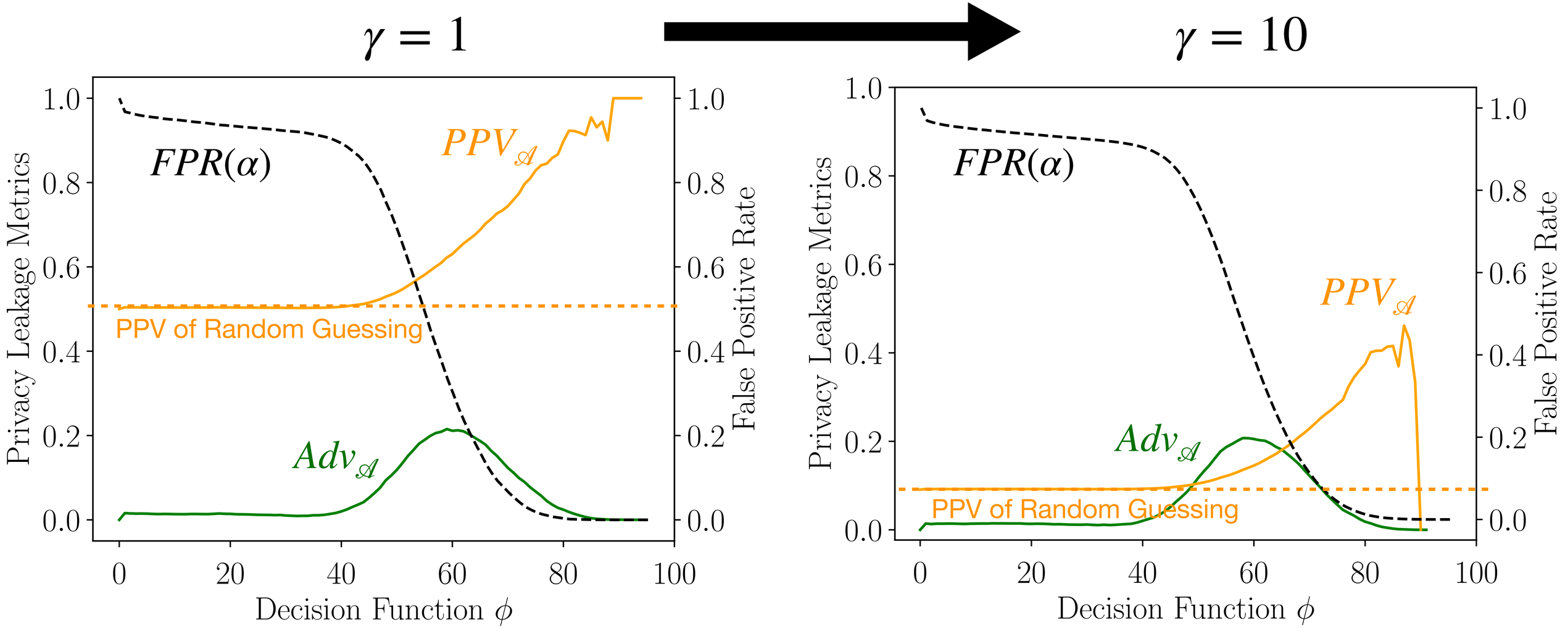


Effect of Prior on Yeom



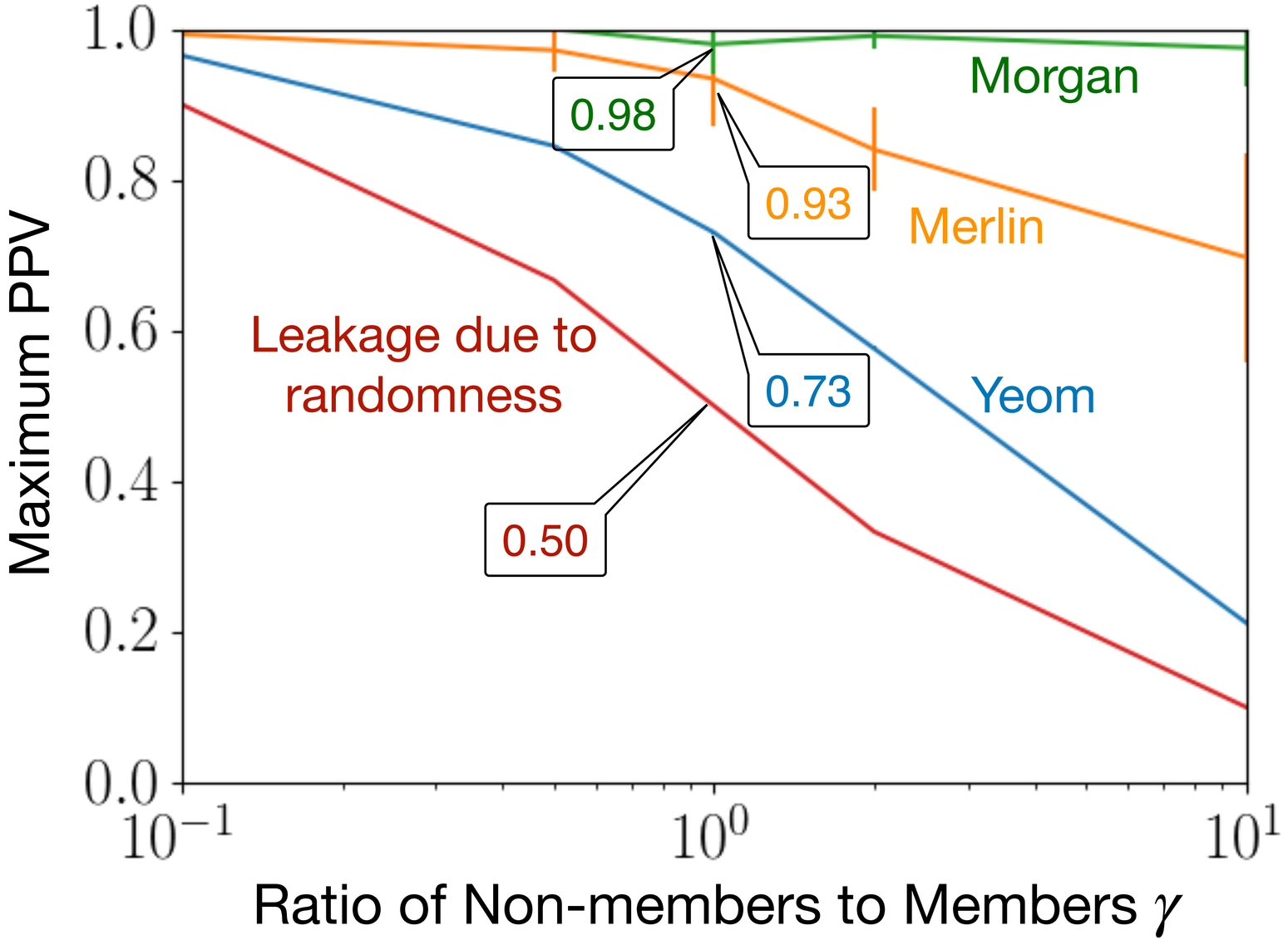
Neural Network on Purchase-100X

Effect of Prior on Merlin



Neural Network on Purchase-100X

Max PPV comparison on Purchase-100X data set (Varying Prior)



Full Paper:

https://arxiv.org/abs/2005.10881

Code:

https://www.github.com/bargavj/EvaluatingDPML

Corresponding Author:

Bargav Jayaraman, bj4nq@virginia.edu