Optimal Personalized Filtering Against Spear-Phishing Attacks

Aron Laszka, Yevgeniy Vorobeychik, and Xenofon Koutsoukos

Institute for Software Integrated Systems
Department of Electrical Engineering and Computer Science
VANDERBILT UNIVERSITY

Malicious E-Mails

Spam

- non-targeted
- usually just a nuisance
(but can waste a lot of time
and money in high volumes)

Spear-phishing

- targeted
- potentially very high losses
(even from a single attack)

Spear-Phishing Examples

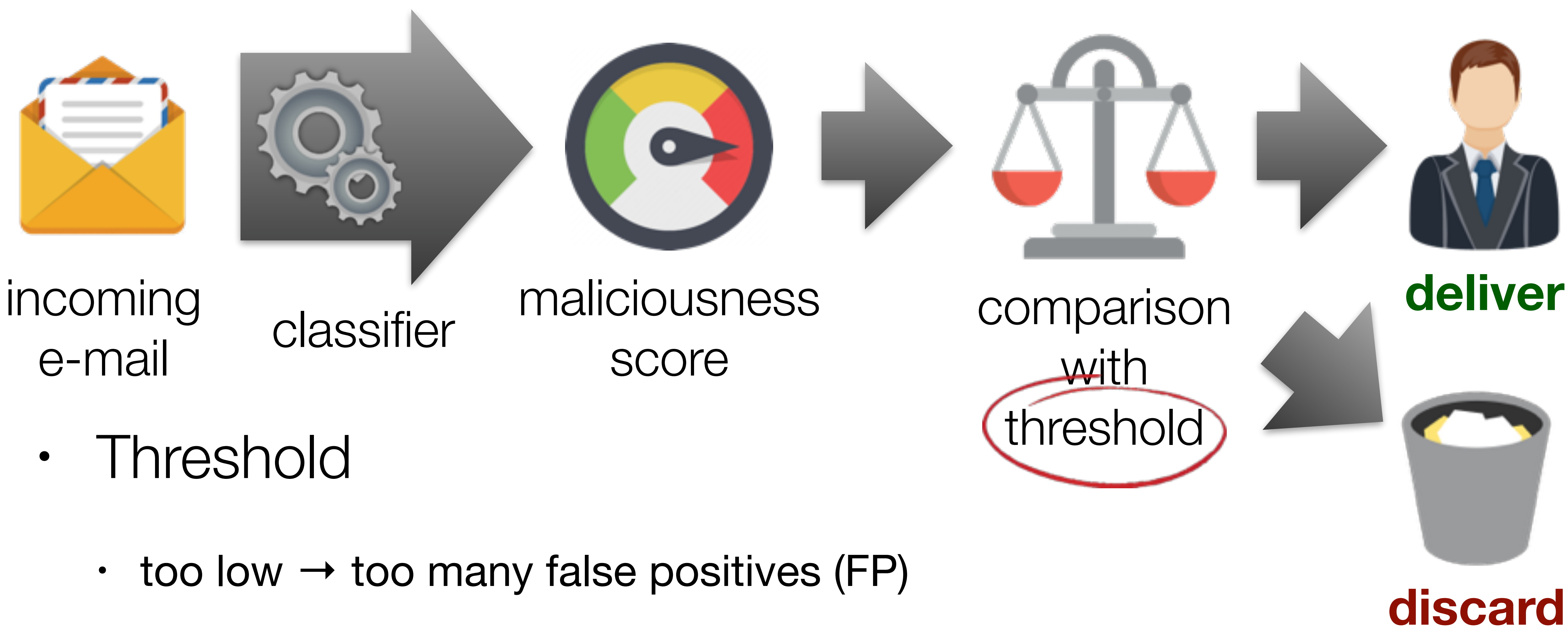
- In 2014, a German steel mill suffered "massive" physical damage due to a cyber-attack
    - first step of the attack was spear-phishing

http://www.wired.com/2015/01/german-steel-mill-hack-destruction/

- In 2013, millions of credit and debit card accounts were compromised due to an attack against Target
    - first step of the attack was spear-phishing

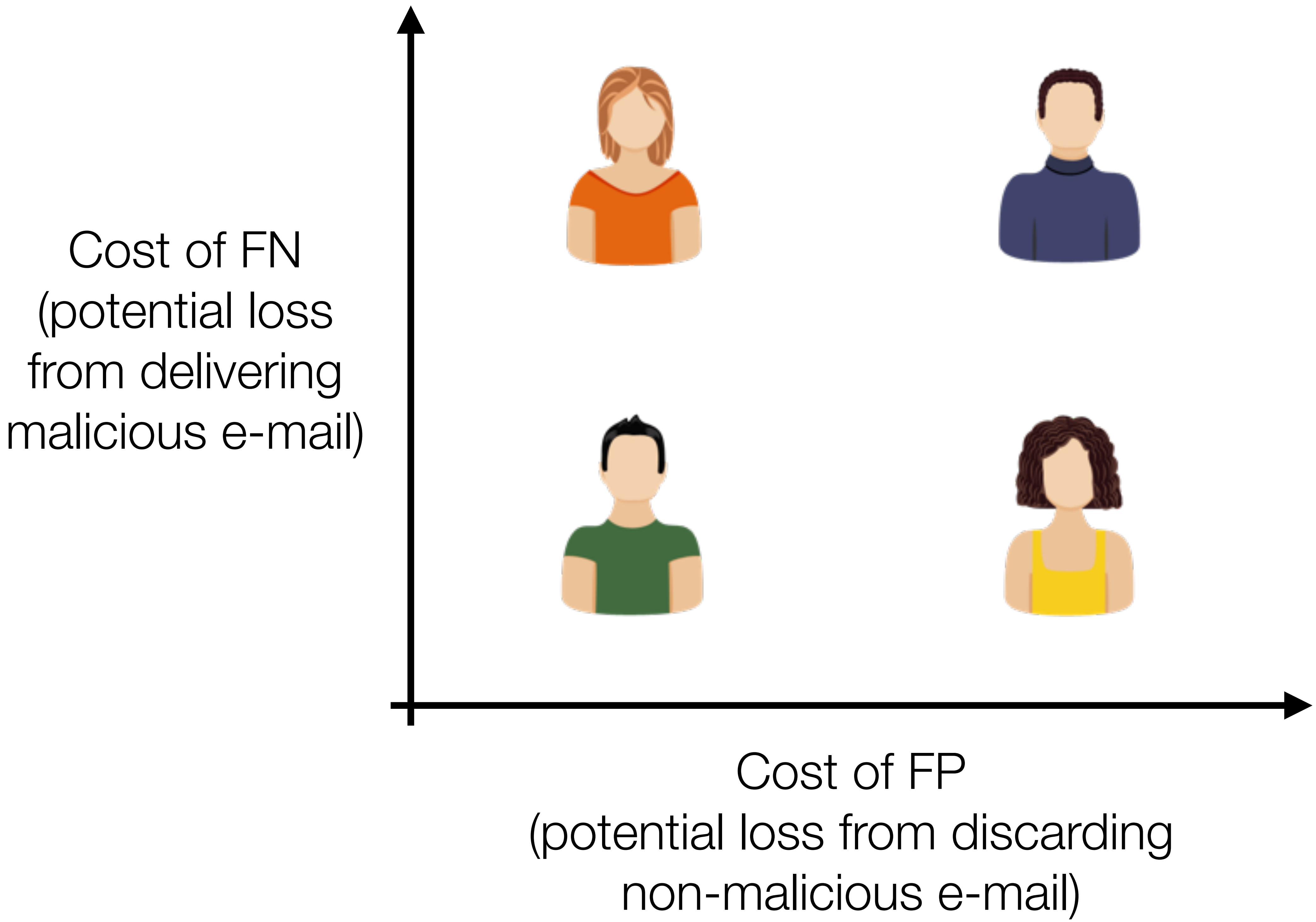
http://www.huffingtonpost.com/2014/02/12/target-hack_n_4775640.html

Filtering Malicious E-Mails

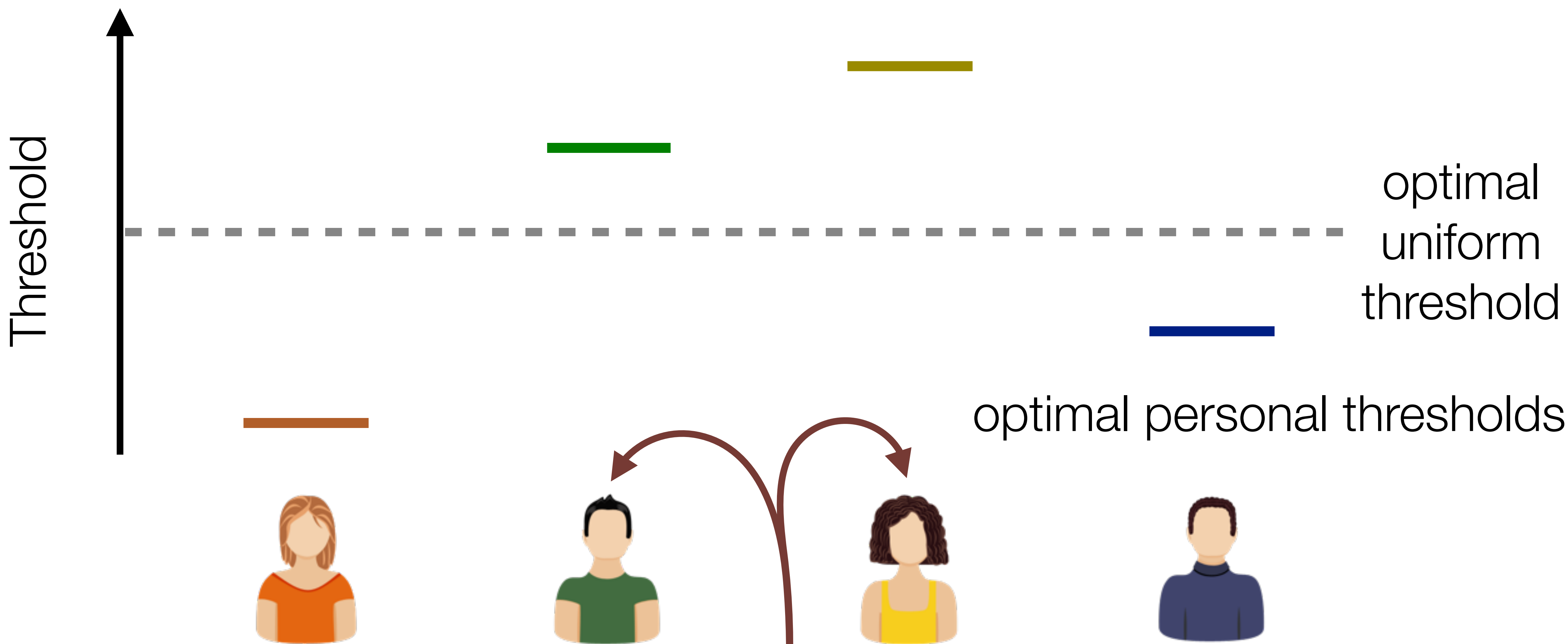


- Threshold
  - too low → too many false positives (FP)
  - too high → too many false negatives (FN)
  - optimal value:
    minimizes **FP rate × cost of FP + FN rate × cost FN**

Multiple Users



Personalized Thresholds



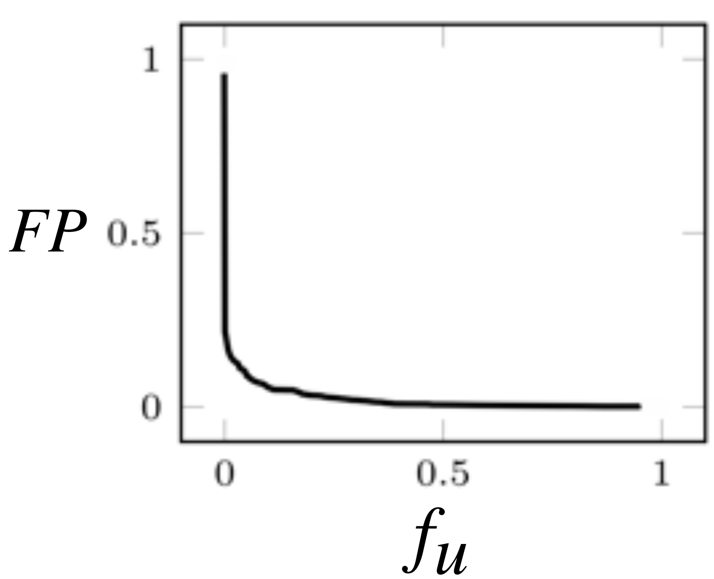
targeting attacker *may exploit* the differences not only between the users but also between the *personalized thresholds*

optimal personal thresholds should also take the attacker's strategy into account
→ **game theory**

Game-Theoretic Model

Defender

- for each user u, selects a false negative rate f_u
- we assume that the feasible FP / FN rate pairs are given by a function $FP(f_u)$



Targeting attacker

- selects a set of users $\mathcal{A}$, and sends them targeted malicious e-mails
- can select at most A users (otherwise the attack is easily detected)

( **Non-targeting attacker(s)**

- non-strategic (not a player))

Game-Theoretic Model (contd.)

Stackelberg (leader-follower) game

1. defender selects a false negative rate f_u for each user u
2. attacker selects a set of users $\mathcal{A}$

expected loss from targeted attacks

Attacker's utility: $$\mathcal{U}_{\text{attacker}} = \sum_{u \in \mathcal{A}} f_u L_u$$

expected loss from non-targeted attacks

Defender's loss: $$\mathcal{L}_{\text{defender}} = \mathcal{U}_{\text{attacker}} + \sum_{u} f_u N_u + FP(f_u) C_u$$

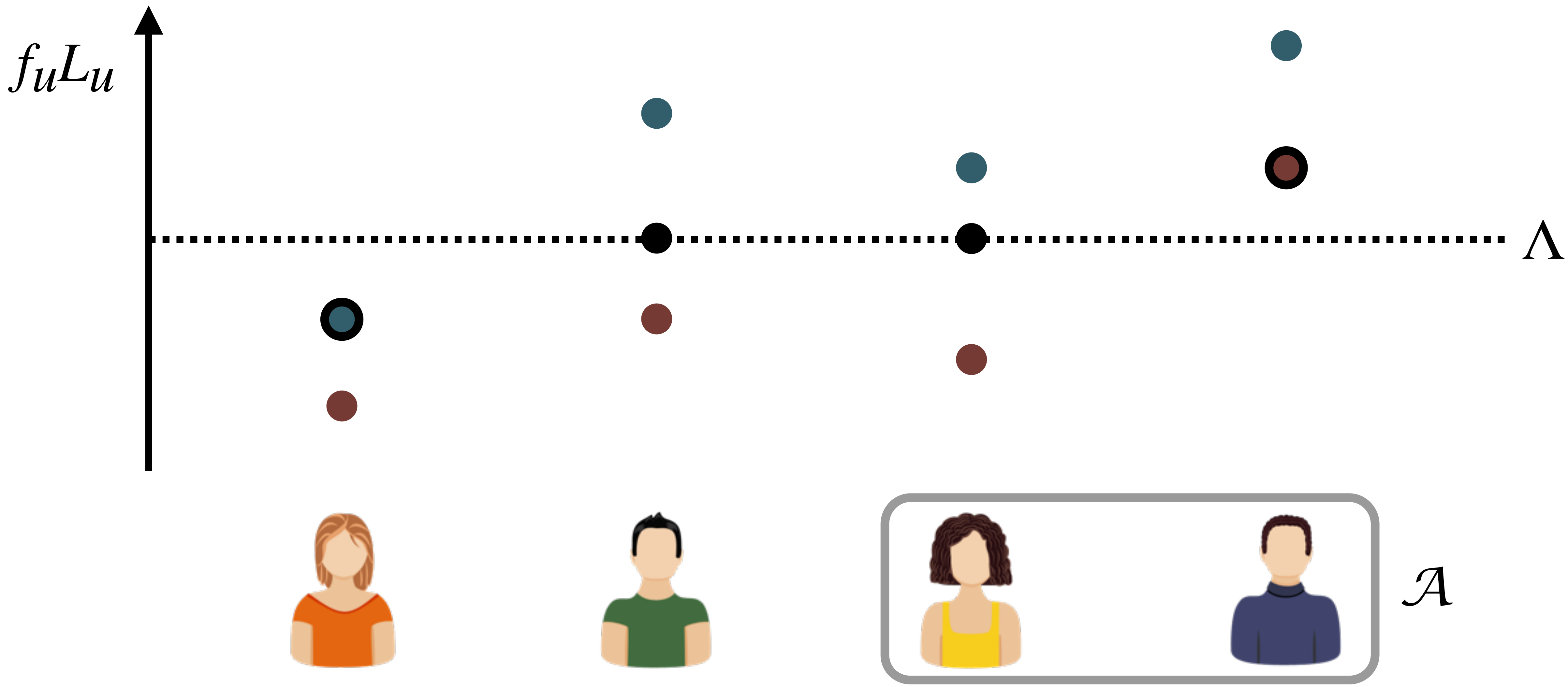
expected loss from from false positives

L_u: potential loss from delivering targeted malicious e-mails

N_u: potential loss from delivering non-targeted malicious e-mails

C_u: potential loss from discarding non-malicious e-mails

Characterizing Optimal Strategies



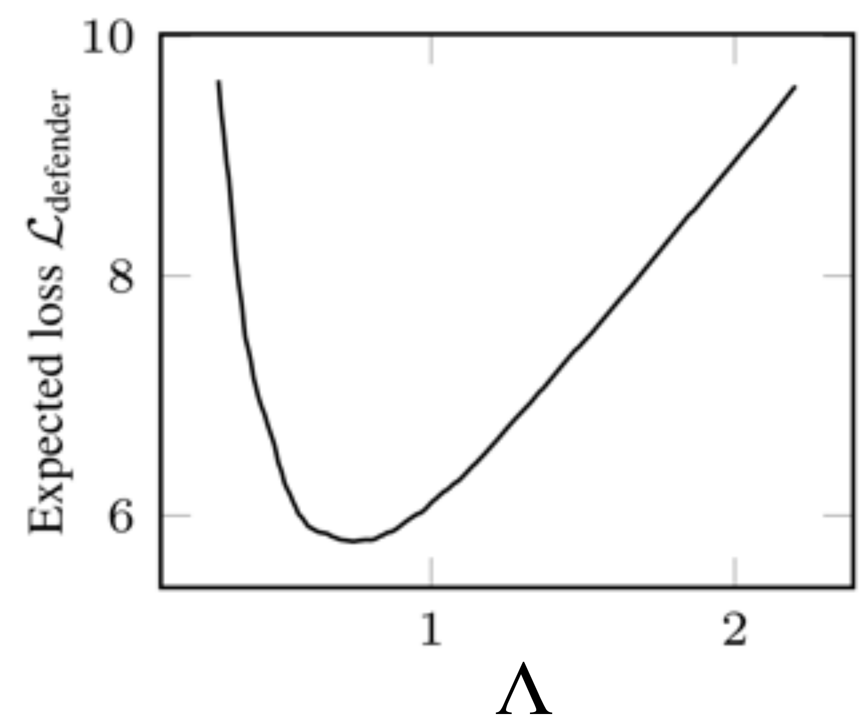
- optimal value for a user given that it is <u>not selected</u> by the attacker
- optimal value for a user given that it is <u>selected</u> by the attacker

Finding an Optimal Strategy

- For a given value of Λ, we can find an optimal strategy using the following polynomial-time algorithm

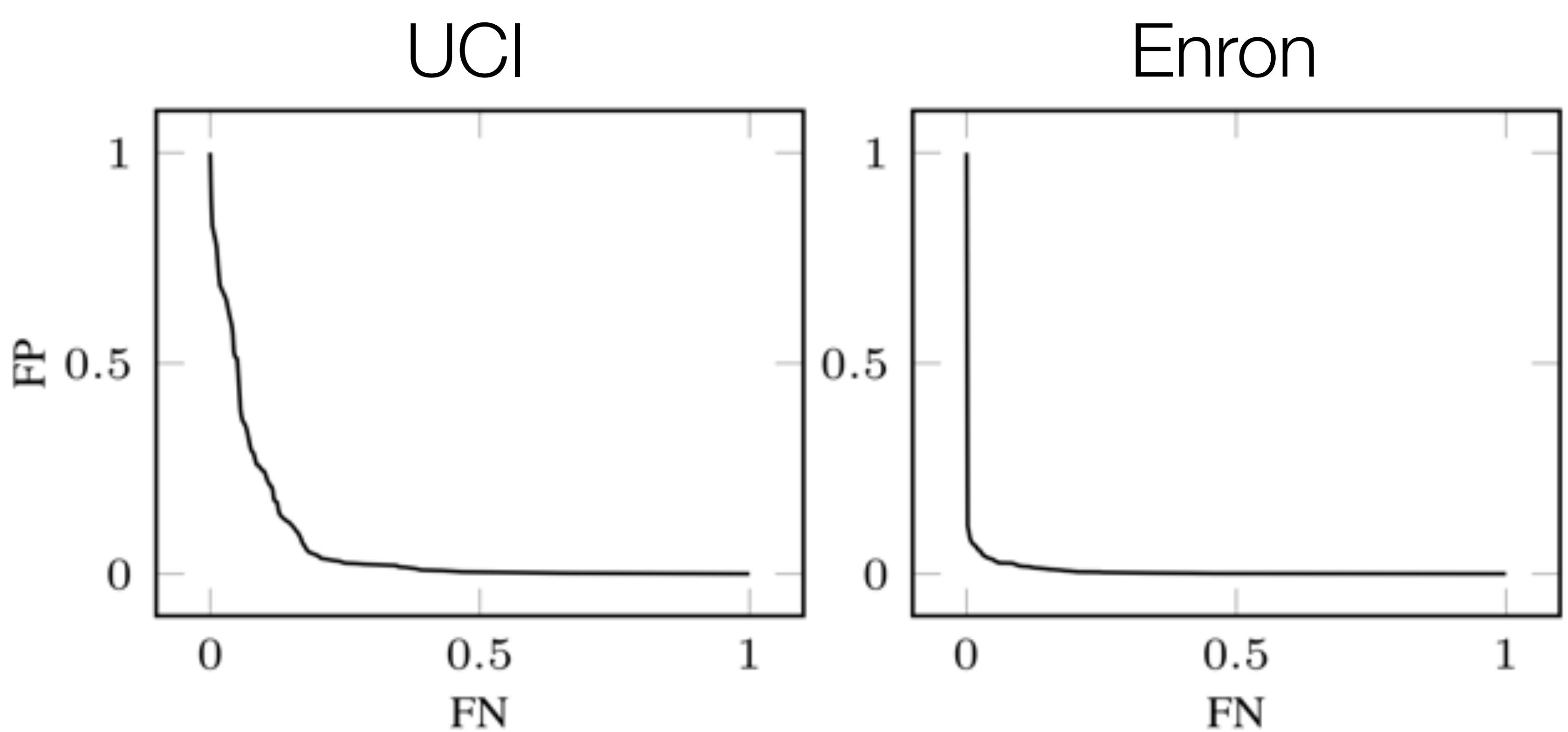
1. *For each user u, compute the loss of user u when it is not targeted as follows: if $f_u^N L_u < \Lambda$, then the loss is $f_u^N N_u + FP(f_u^N)C_u$; otherwise, the loss is $\frac{\Lambda}{L_u} N_u + FP(\frac{\Lambda}{L_u})C_u$.*
2. *For each user u, compute the loss of user u when it is targeted as follows: if $f_u^T L_u > \Lambda$, then the loss is $f_u^T(L_u + N_u) + FP(f_u^T)C_u$; otherwise, the loss is $\frac{\Lambda}{L_u}(L_u + N_u) + FP(\frac{\Lambda}{L_u})C_u$.*
3. *For each user u, let the cost of user u being targeted be the difference between the above computed loss values.*
4. *Select a set $\mathcal{A}$ of A users with the lowest costs of being targeted.*
5. *For every $u \in \mathcal{A}$, let $f_u = f_u^T$ if $f_u^T L_u > \Lambda$, and let $f_u = \frac{\Lambda}{L_u}$ otherwise.*
6. *For every $u \notin \mathcal{A}$, let $f_u = f_u^N$ if $f_u^N L_u < \Lambda$, and let $f_u = \frac{\Lambda}{L_u}$ otherwise.*
7. *Output the strategy $\boldsymbol{f}$.*

- Finally, we can find the optimal value of Λ using a simple binary search



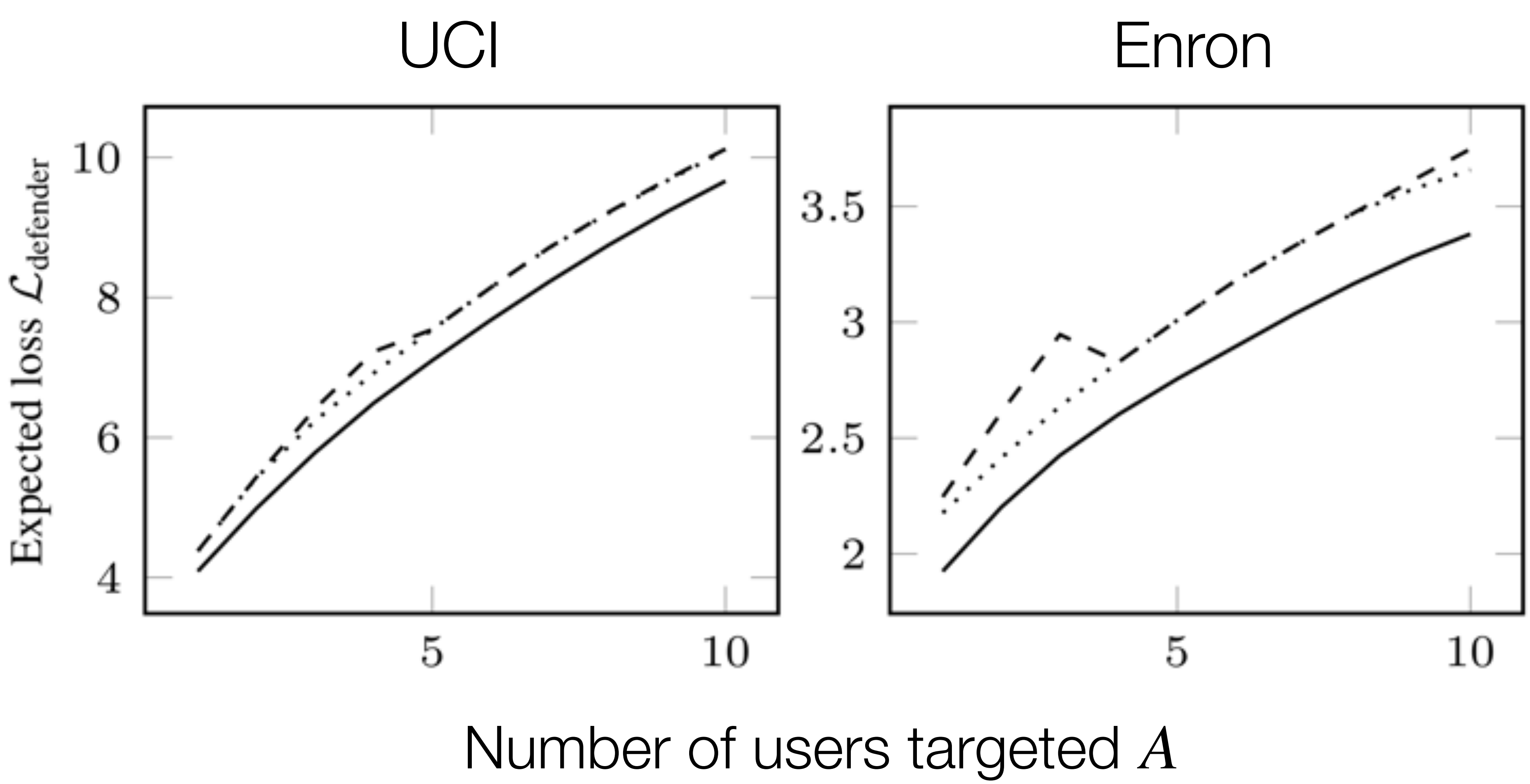
Numerical Examples

- Datasets
  - UCI Machine Learning Repository: 4601 labeled e-mails with 57 features
  - Enron dataset: 13,500 e-mails with 500 features
- Classifier: naive Bayes (note that this is just for the sake of example)
- False positive / false negative rates:



Numerical Examples - Results

- 31 users with parameter values following power-law distributions



— optimal strategy

- - - uniform threshold not expecting strategic attacker

···· uniform threshold expecting strategic attacker

Conclusion & Future Work

- Conclusion
  - filtering thresholds have received less attention in the past
  - we proposed a game-theoretic model for targeted and non-targeted malicious e-mails
  - we showed how to find optimal strategies efficiently
  - numerical results show considerable improvement
- Future work
  - non-linear losses from compromising multiple users

Thank you for your attention!

Questions?