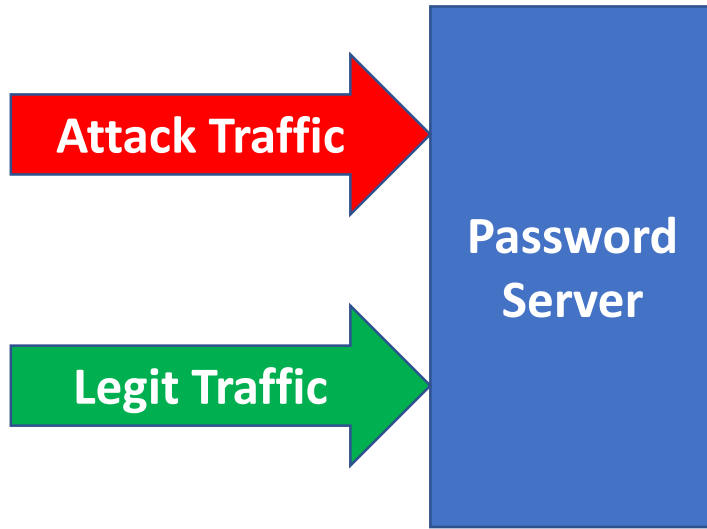


# Distinguishing Attacks from Legitimate Authentication Traffic at Scale

Cormac Herley and Stuart Schechter\*  
Microsoft Research, Redmond

\* Work done while at MSR

# Online password guessing



A screenshot of a login page. At the top, a red error message box with a white 'X' icon contains the text: "Your account has been locked because you have reached the maximum number of invalid logon attempts. You can contact the administrator or [click here](#) to receive an e-mail containing instructions on how to unlock your account." Below the message are two input fields: "User name:" with the text "Jason" and "Password:". There is a checkbox labeled "Stay logged in on this computer" which is unchecked. A green "Log on" button is positioned below the password field. At the bottom of the page, there are three links: "[Logon troubleshooting](#)", "[Forgotten password](#)", and "[Select language](#)".

- Account lockout (3 strikes, etc)?
- IP blocking?
- Machine Learning?

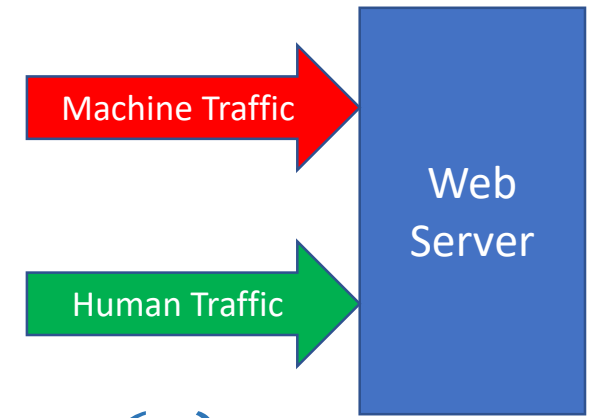
# Want $P(\text{abuse} | x)$

$X = [\text{username}, \text{password}, \text{time}, \text{IP address}, \text{UserAgent}, \dots]$

## Goals:

- Minimal assumptions about attack traffic
- Scalability/Maintainability

# Back to the drawing board



- Suppose  $x$  is categorical feature:

$$Observed(x) = \alpha Clean(x) + (1-\alpha) Abuse(x)$$

- If we know  $Clean()$ ,  $\alpha$  then odds of being malicious:

$$\begin{aligned} \frac{P(abuse|x)}{P(legit|x)} &= \frac{(1 - \alpha) Abuse(x)}{\alpha Clean(x)} \\ &= \frac{Observed(x) - \alpha Clean(x)}{(1-\alpha) Clean(x)} \frac{1-\alpha}{\alpha} \end{aligned}$$

# Three Observations:

## 1. Clean(x) is stationary

- Aggregate behavior of millions of users is **very** stable

## 2. If we can estimate $\alpha$ we can estimate Clean(x)

$$\textit{Observed}(x) = \alpha \textit{Clean}(x) + (1-\alpha) \textit{Abuse}(x)$$

- That is,  $\alpha \approx 1 \Rightarrow$

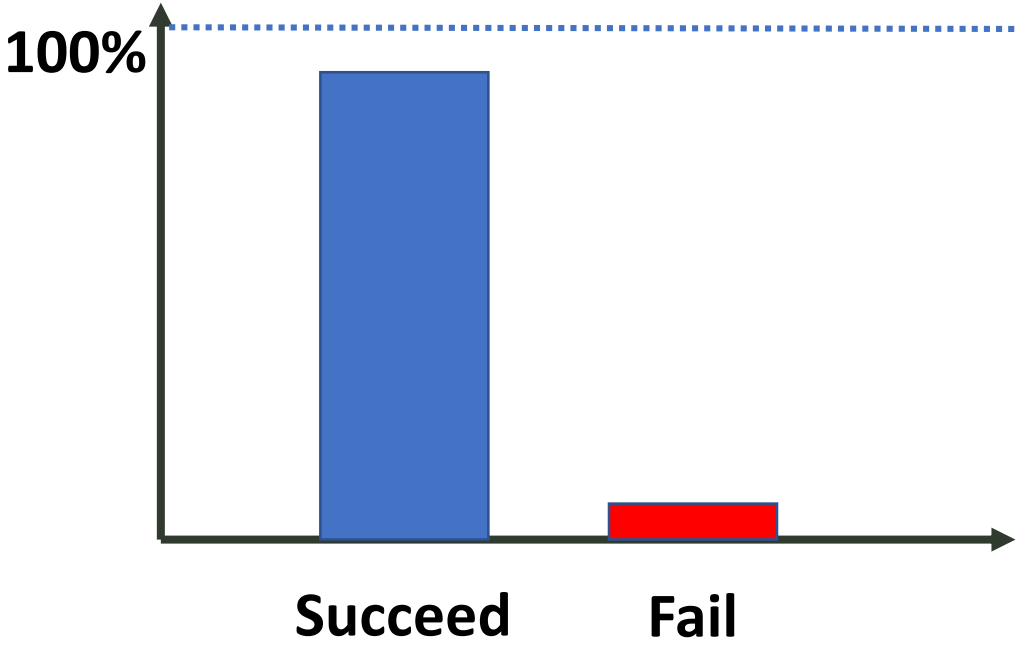
$$\textit{Observed}(x) \approx \textit{Clean}(x)$$

## 3. We have a lot of data:

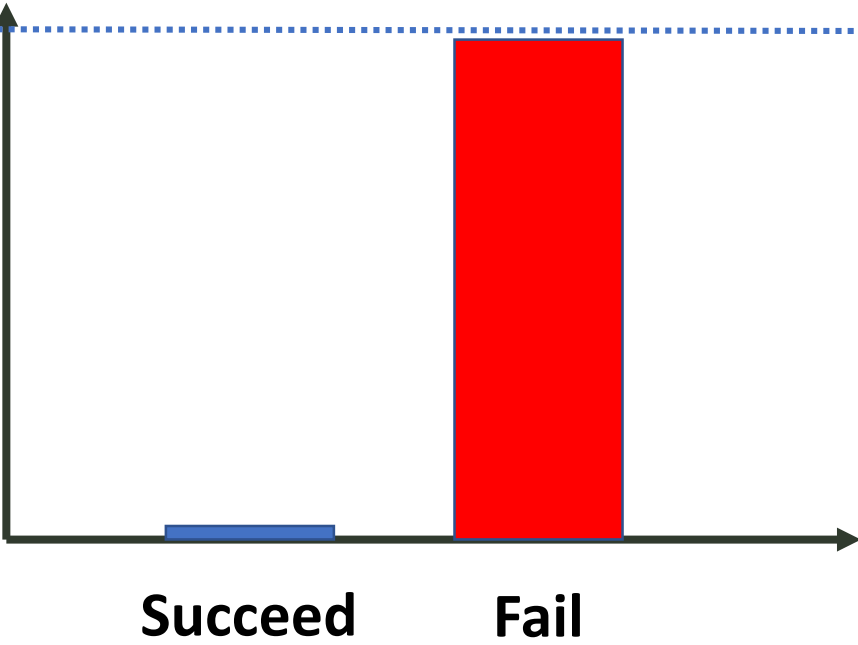
- E.g., subset that's 1% of 1% of 1bn/day

# A feature that separates legit/attack well

**Legitimate Traffic**



**Attack Traffic**



# Ratio of fails/logins

Failures:  $F = F_b + F_m$

Logins:  $L = L_b + L_m$

Assumptions:

- $L_m/L_b \approx 0$
- $F_b/L_b = \text{const.}$

$$\frac{F}{L} = \frac{F_b + F_m}{L_b + L_m} = \frac{F_b/L_b + F_m/L_b}{1 + L_m/L_b}$$

$$\approx \frac{F_b}{L_b} + \frac{F_m}{L_b} = C + \frac{F_m}{L_b}$$

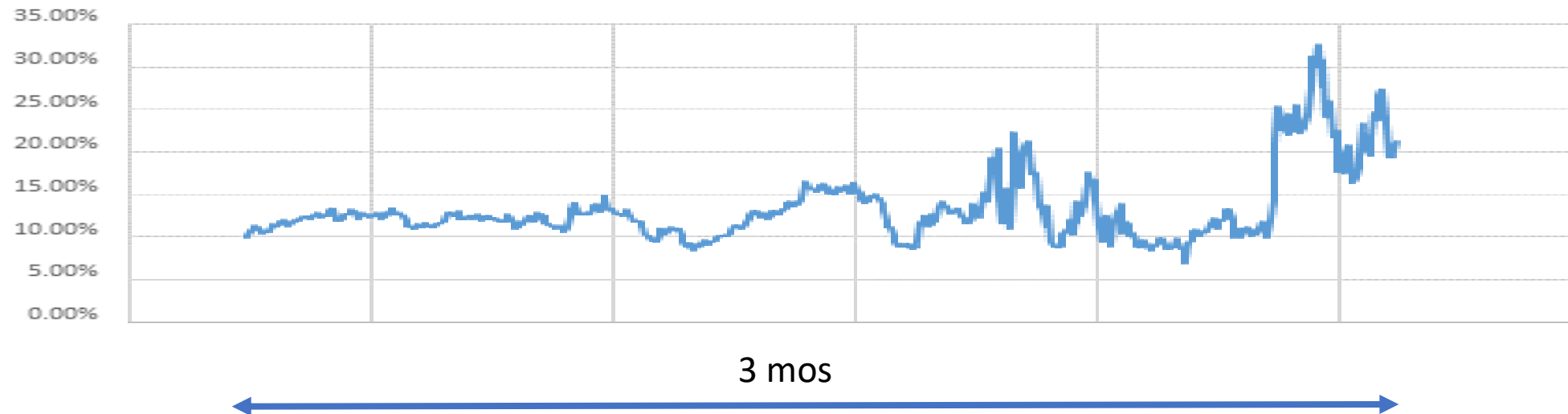
- *Ratio of fails/logins:*

$$\frac{F}{L} \approx C + \frac{F_m}{L_b}$$

Assumptions:

- $L_m/L_b \approx 0$
- $F_b/L_b = \text{const.}$

- Abuse increases  $F/L$ , never decreases





- If we knew  $c$  then:

$$F_m \approx F - c \cdot L$$

$$\frac{1-\alpha}{\alpha} \approx \frac{F_m}{L_b + F_b} = \frac{F - c \cdot L}{L(1+c)}$$

Assumptions:

- $L_m/L_b \approx 0$
- $F_b/L_b = \text{const.}$

***We can estimate abuse/legit ratio!!!***

$$\textit{Observed}(x) = \alpha \textit{Clean}(x) + (1-\alpha) \textit{Abuse}(x)$$

If we know  $c$ , we now know how to calculate  $(1-\alpha)/\alpha$

If we can find a subset where  $(1-\alpha)/\alpha \approx 0$

$$\textit{Observed}(x) \approx \textit{Clean}(x)$$

OK, so how do we find  $c = F_b/L_b$ ?

# Thought-experiment: attackers' day off

$$\textit{Observed}(x) = \alpha \textit{Clean}(x) + (1-\alpha) \textit{Abuse}(x)$$

1. If can identify an un-attacked block of (time, IPs, accounts, uAgent...)

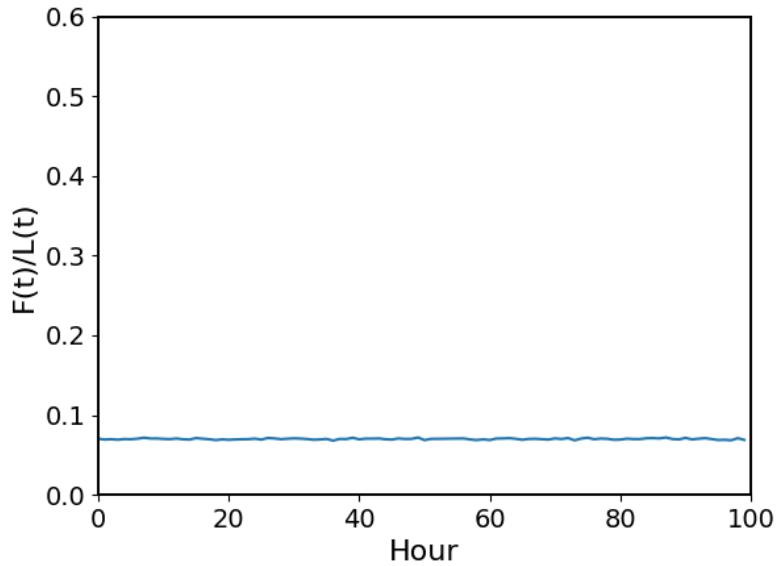
$$\textit{Observed}(x) \approx \textit{Clean}(x)$$

2. We'll know it when we see it:

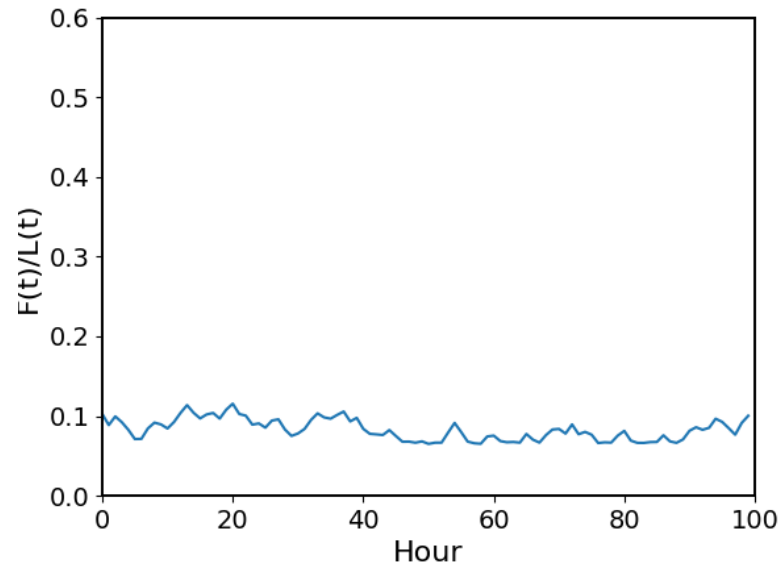
$$\frac{F(t)}{L(t)} = c + \frac{F_m(t)}{L_b(t)} \approx \textit{const.}$$

# Finding an unattacked subset

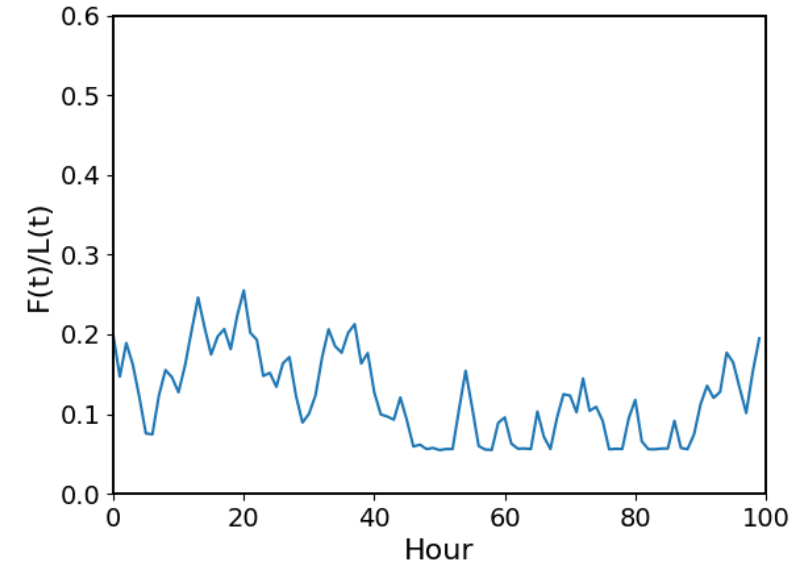
$$\frac{F(t)}{L(t)} = c + \frac{F_m(t)}{L_b(t)}$$



$\alpha=1.0$



$\alpha=0.95$



$\alpha=0.8$

# Overall Algorithm

*Break into  $k$  subsets:*

$$1. \hat{C} = \min_k \frac{F(k)}{L(k)}$$

$$2. \text{Clean}(x) \approx \text{Observed}_{k_{\min}}(x)$$

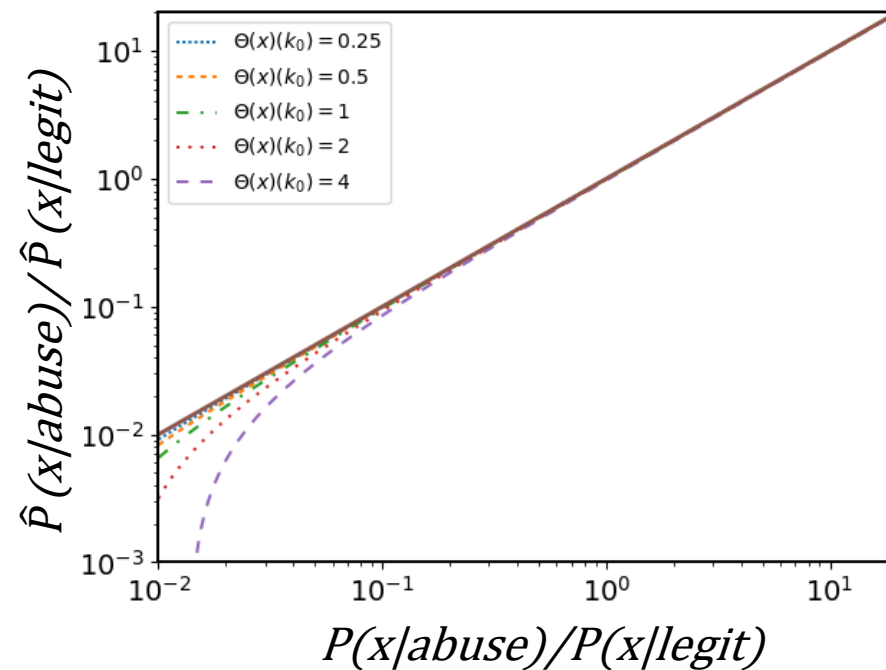
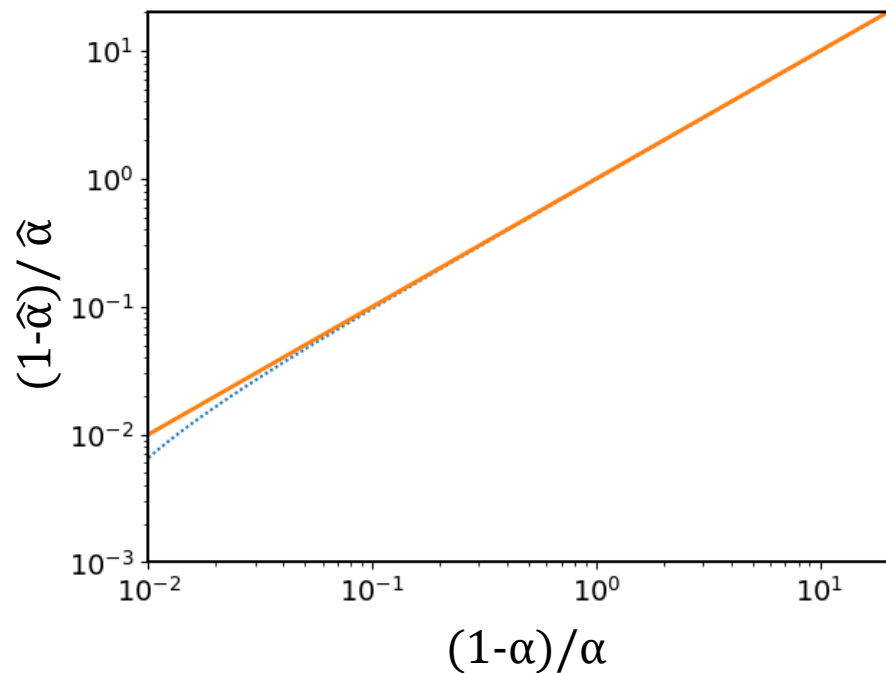
*For each subset  $k = 0, 1, 2, \dots, K-1$ :*

$$3. \frac{1-\alpha(k)}{\alpha(k)} \approx \frac{F(k) - c \cdot L(k)}{L(k) \cdot (1+c)}$$

$$4. \frac{P(x|\text{abuse})(k)}{P(x|\text{legit})(k)} = \frac{\text{Observed}_k(x) - \alpha(k) \text{Clean}(x)}{(1-\alpha(k)) \text{Clean}(x)}$$

$$\text{Odds malicious} = \frac{P(x|\text{abuse})}{P(x|\text{legit})} \cdot \frac{1-\alpha}{\alpha}$$

# Sensitivity Analysis: $c = 0.07, \hat{c} = 0.0732$



$$\frac{P(abuse|x)}{P(legit|x)} = \frac{P(x|abuse)}{P(x|legit)} \cdot \frac{1-\alpha}{\alpha}$$

# Toy example

X = Failure from Top-1000 passwords

- $P(X|abuse) = 0.97$ ,  $P(X|legit) = 0.005$

25% of traffic is abuse, but attacker has list of only 80% accounts.

For accounts on attackers list:

$$\frac{P(abuse|x)}{P(legit|x)} = \frac{P(x|abuse)}{P(x|legit)} \cdot \frac{1-\alpha}{\alpha} = \frac{0.97}{0.005} \cdot \frac{0.25/8}{0.75/10} \approx 80.8$$

Accounts not on list

$$\frac{P(abuse|x)}{P(legit|x)} = \frac{P(x|abuse)}{P(x|legit)} \cdot \frac{1-\alpha}{\alpha} = \frac{0.97}{0.005} \cdot 0 \approx 0$$

# Conclusions

- Simple way to estimate amount of attack traffic
- Simple way to find least-attacked subsets
- Simple way to est. odds that any event is malicious

## Main assumptions:

- Attacker fail rate is high
- Clean distributions slowly varying