Distinguishing Attacks from Legitimate Authentication Traffic at Scale

Cormac Herley and Stuart Schechter*
Microsoft Research, Redmond

* Work done while at MSR
Online password guessing

- Account lockout (3 strikes, etc)?
- IP blocking?
- Machine Learning?
Want $P(\text{abuse}|x)$

$X = \{\text{username, password, time, IP address, UserAgent, ......}\}$

Goals:

• Minimal assumptions about attack traffic
• Scalability/Maintainability
Back to the drawing board

• Suppose $x$ is categorical feature:

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

• If we know $\text{Clean}()$, $\alpha$ then odds of being malicious:

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

$$= \frac{Observed(x)-\alpha \text{Clean}(x)}{(1-\alpha) \text{Clean}(x)} \frac{1-\alpha}{\alpha}$$
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)**
   - $Observed(x) = \alpha Clean(x) + (1-\alpha) Abuse(x)$
   - That is, $\alpha \approx 1 \Rightarrow$
     
     \[ Observed(x) \approx Clean(x) \]

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

100%

Succeed Fail

Attack Traffic

Succeed Fail
Ratio of fails/logins

Failures: $F = F_b + F_m$
Logins: $L = L_b + L_m$

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

\[
\approx \frac{F_b}{L_b} + \frac{F_m}{L_b} = C + \frac{F_m}{L_b}
\]

Assumptions:
• $L_m/L_b \approx 0$
• $F_b/L_b = \text{const.}$
• **Ratio of fails/logins:**
  \[
  \frac{F}{L} \approx C + \frac{F_m}{L_b}
  \]

• Abuse increases F/L, never decreases

Assumptions:
• \( L_m/L_b \approx 0 \)
• \( F_b/L_b = \text{const.} \)
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!!!
If we know $c$, we now know how to calculate $(1-\alpha)/\alpha$

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

\[ \text{Observed}(x) \approx \text{Clean}(x) \]

OK, so how do we find $c = F_b / L_b$?
Thought-experiment: attackers’ day off

\[ \text{Observed}(x) = \alpha \text{Clean}(x) + (1-\alpha) \text{Abuse}(x) \]

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

\[ \text{Observed}(x) \approx \text{Clean}(x) \]

2. We’ll know it when we see it:

\[ \frac{F(t)}{L(t)} = c + \frac{F_m(t)}{L_b(t)} \approx \text{const.} \]
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 \( k = 0, 1, 2, ..., K-1 \): 

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

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

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

$X = \text{Failure from Top-1000 passwords}$

- $P(X|\text{abuse}) = 0.97$, $P(X|\text{legit}) = 0.005$

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

For accounts on attackers list:

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

Accounts not on list

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