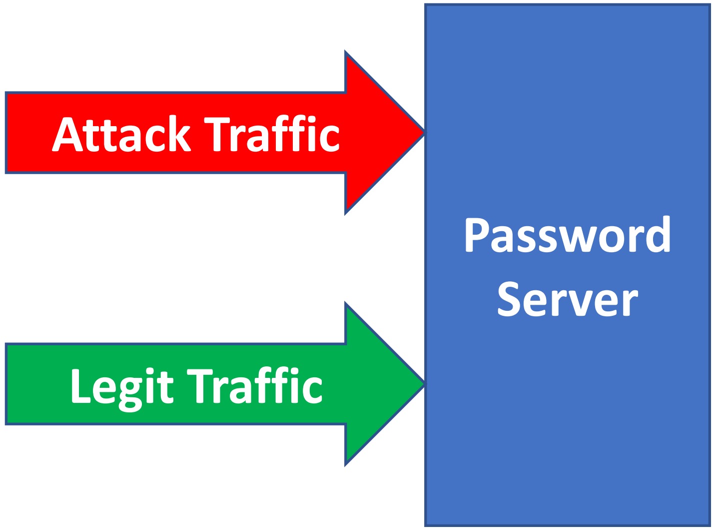
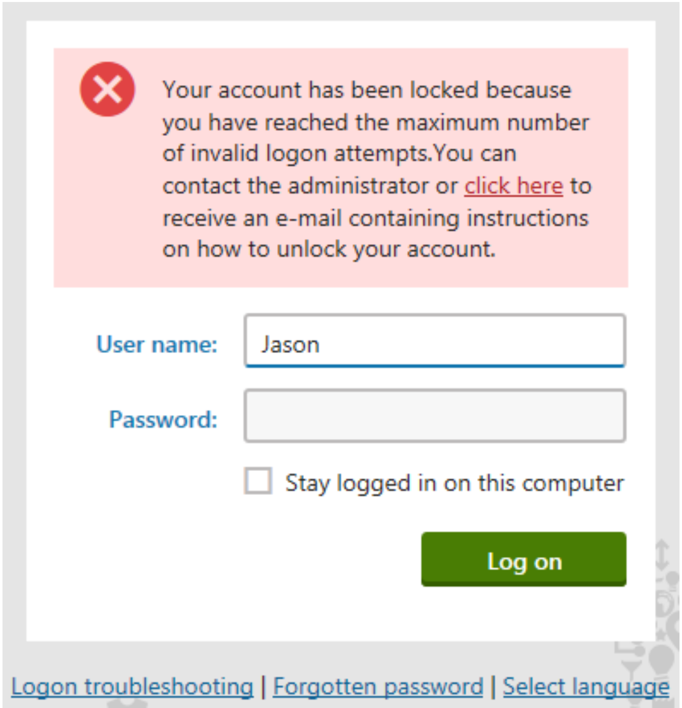
Distinguishing Attacks from Legitimate Authentication Traffic at Scale

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?

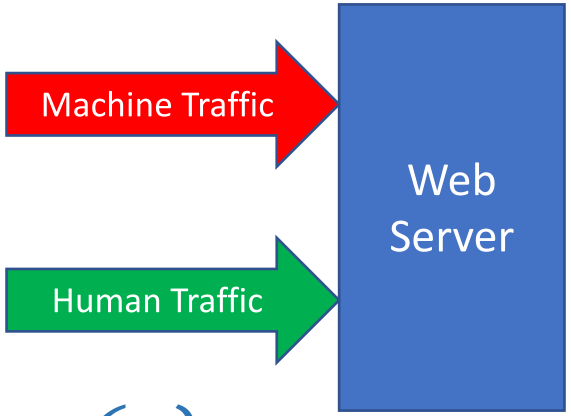
Want P(abuse|x)

X = [username, password, time, IP address, UserAgent, ……]

Goals:

- Minimal assumptions about attack traffic
- Scalability/Maintainability

Back to the drawing board



- Suppose *x is* categorical feature:

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

- If we know *Clean(), α then* odds of being malicious:

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

$$= \frac{Observed(x) - \alpha\, Clean(x)}{(1-\alpha)\, Clean(x)}\, \frac{1-\alpha}{\alpha}$$

Three Observations:

1. Clean(x) is stationary
    - Aggregate behavior of millions of users is ***very*** stable
2. If we can estimate α we can estimate Clean(x)

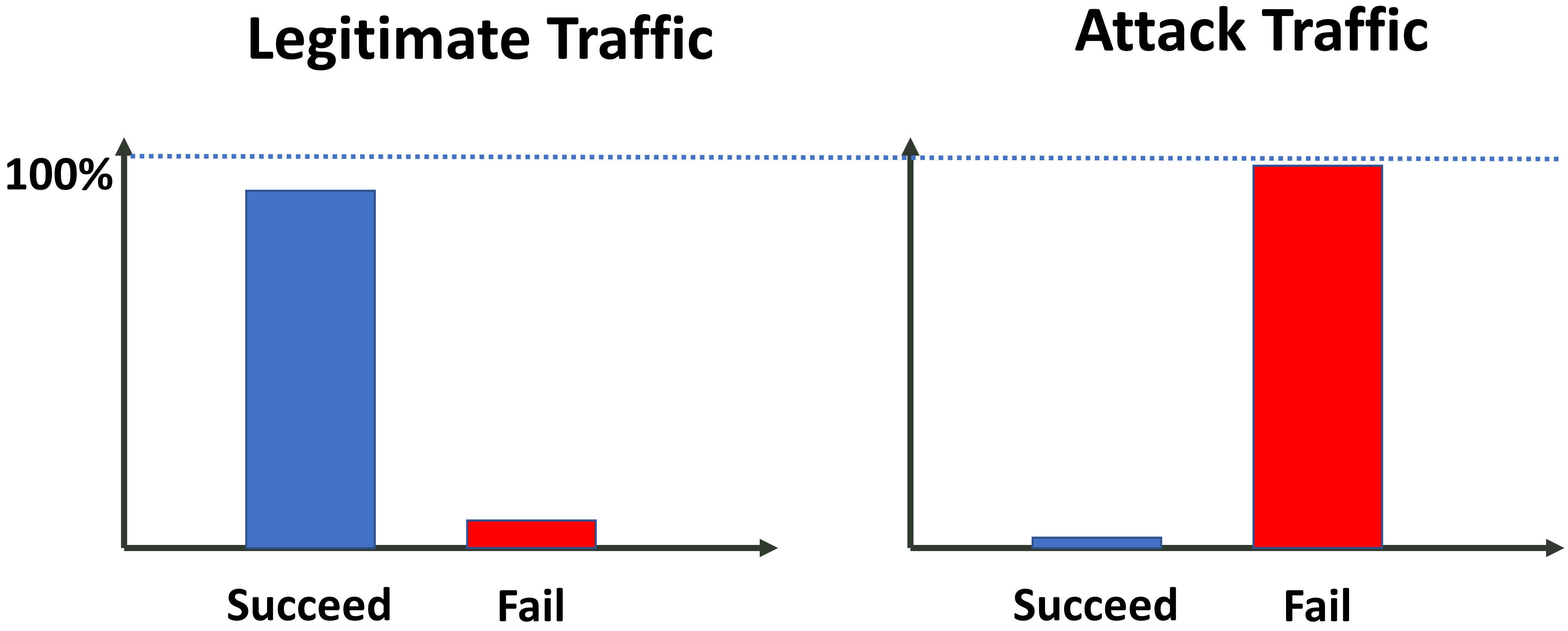
$$Observed(x) = \alpha\, Clean(x) + (1\text{-}\alpha)\, Abuse(x)$$

- That is, $\alpha \approx 1$ =>

$$Observed(x) \approx Clean(x)$$

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



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 = const.

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

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

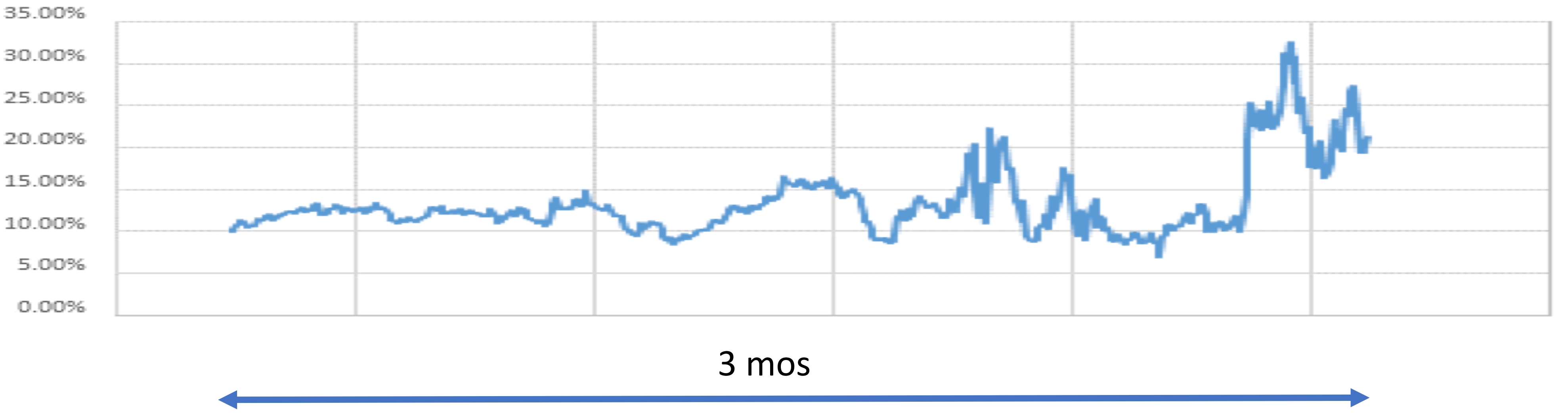
- *Ratio of fails/logins:*

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

Assumptions:

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

We can estimate abuse/legit ratio!!!

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

If we know c, we now know how to calculate (1-α)/α

If we can find a subset where (1-α)/α ≈ 0

$$Observed(x) \approx Clean(x)$$

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

Thought-experiment: attackers' day off

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

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

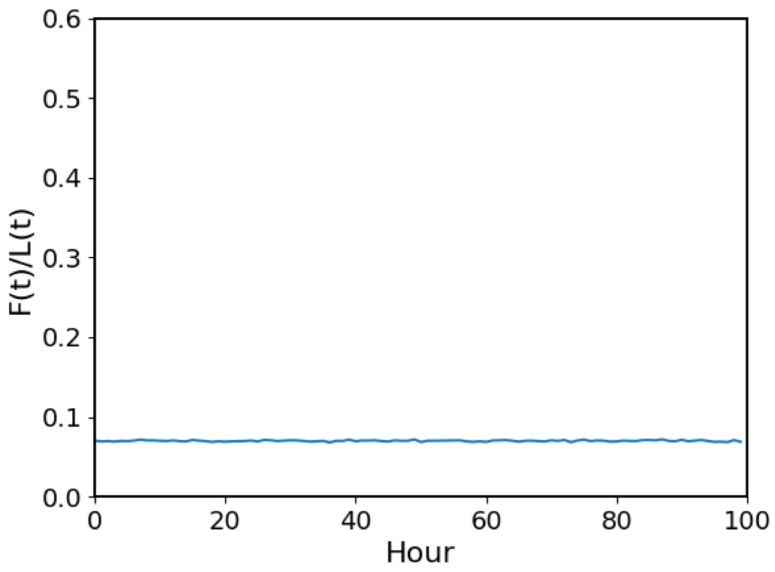
$$Observed(x) \approx Clean(x)$$

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

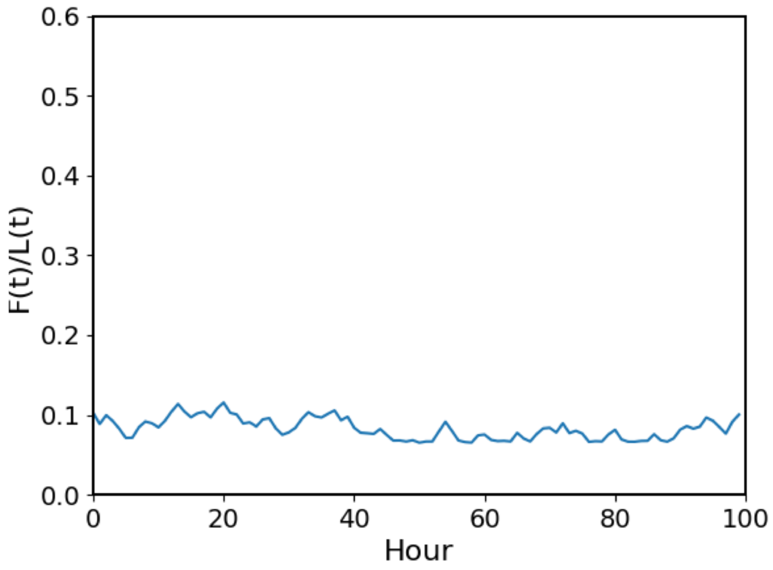
$$\frac{F(t)}{L(t)} = c + \frac{F_m(t)}{L_b(t)} \approx const.$$

Finding an unattacked subset

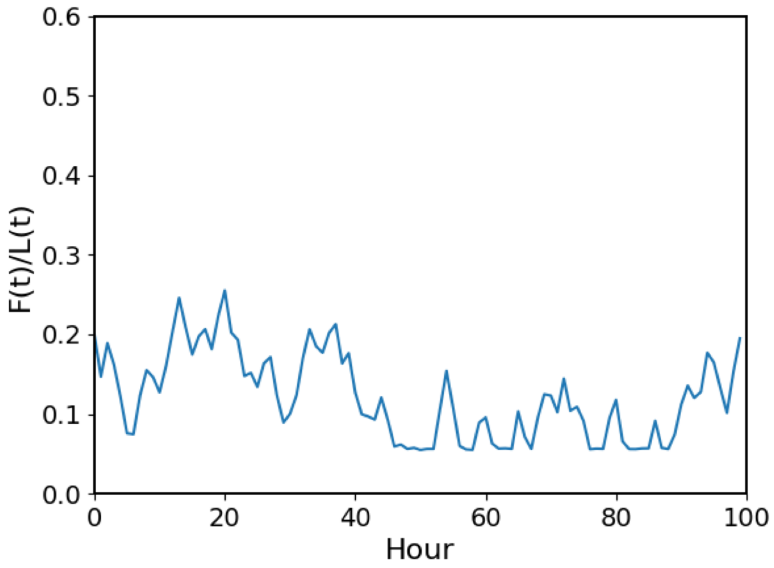
$$\frac{F(t)}{L(t)} = c + \frac{F_m(t)}{L_b(t)}$$



α=1.0



α=0.95



α=0.8

Overall Algorithm

Break into k subsets:

1. $\hat{C} = \min_{k} \frac{F(k)}{L(k)}$
2. $Clean(x) \approx Observed_{k\_min}(x)$

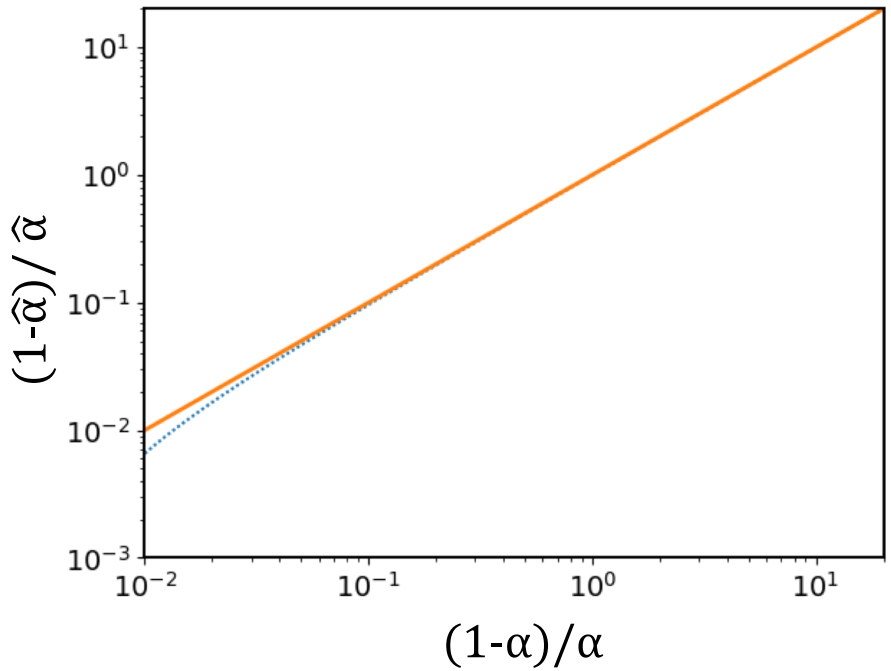
For each subset k = 0,1,2, …, K-1:

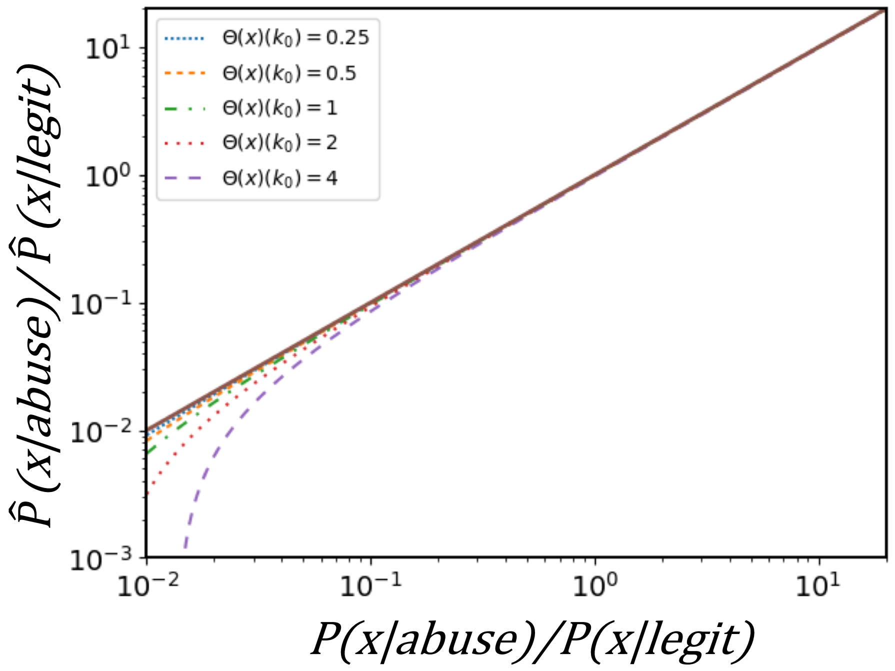
3. $\frac{1-\alpha(k)}{\alpha(k)} \approx \frac{F(k) - c \cdot L(k)}{L(k) \cdot (1+c)}$

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

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

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 distribtuions slowly varying