Who are you?
A Statistical Approach to Measuring User Authenticity

Sakshi Jain (LinkedIn)
Joint work with David Mandell Freeman (LinkedIn)
Markus Dürmuth (Ruhr Universität Bochum)
Battista Biggio and Giorgio Giacinto (Università di Cagliari)

Motivation

Accounts get attacked all the time!

Motivation

Accounts get attacked all the time!

How?

Motivation

Accounts get attacked all the time!

How?

common password

reuse passwords across sites

get phished

tell someone the password

Motivation

Accounts get attacked all the time!

How?

common password

reuse passwords across sites

get phished

tell someone the password

Why?

Motivation

Accounts get attacked all the time!

How?

common password

reuse passwords across sites

get phished

tell someone the password

Why?

Motivation

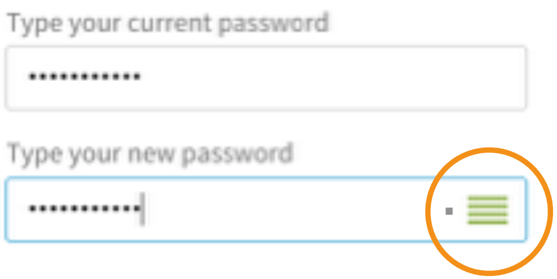
How do we avoid credential leakage?

Effectiveness is limited and attackers get credentials anyway!

Motivation

How do we avoid credential leakage?

Better passwords?

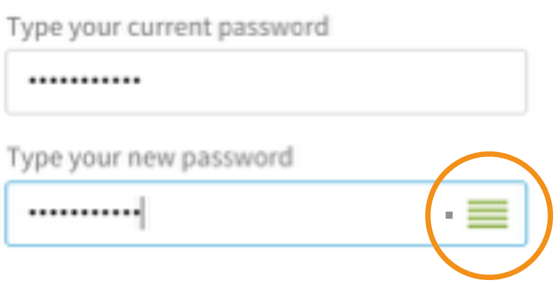


Effectiveness is limited and attackers get credentials anyway!

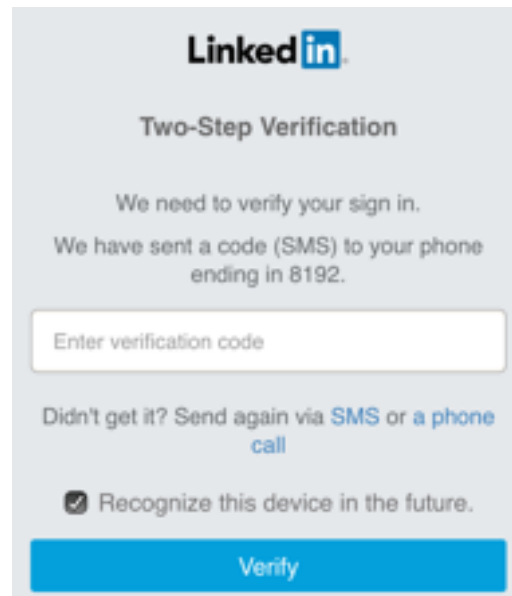
Motivation

How do we avoid credential leakage?

Better passwords?



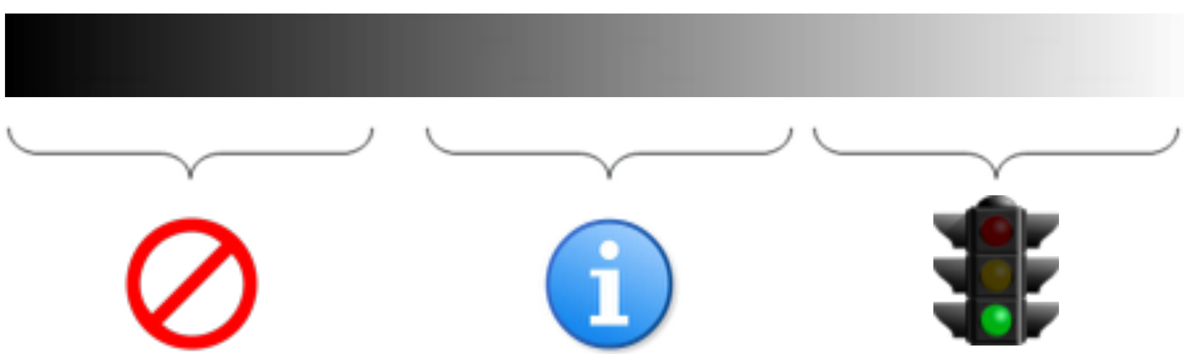
Second Factor?



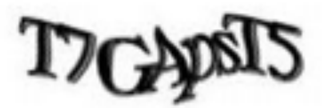
Effectiveness is limited and attackers get credentials anyway!

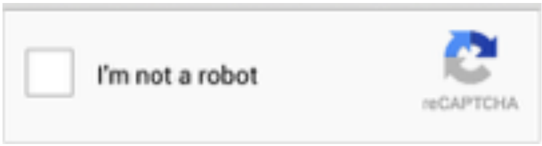
So here’s the problem statement…

For an incoming login request, with **correct credentials**, assess level of suspiciousness **online** and take an action accordingly.



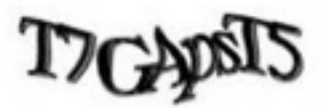
What second factors could we require?

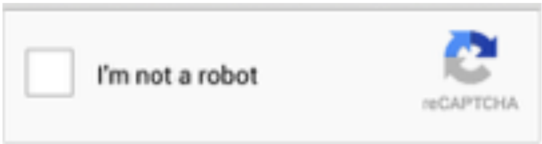




What second factors could we require?

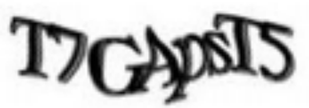
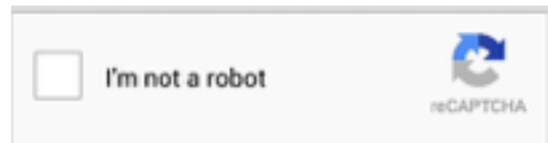
– Prove you’re a human



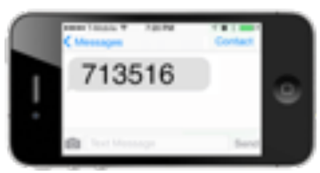


What second factors could we require?

- Prove you're a human

- Establish contact through another channel



What data do we have to score logins?

What data do we have to score logins?

- Request data:
  - IP address (and derived country, ISP, etc.)
  - Browser's user agent (and OS, version, etc.)
  - Timestamp
  - Cookies
  - and more…

What data do we have to score logins?

- Request data:
  - IP address (and derived country, ISP, etc.)
  - Browser’s user agent (and OS, version, etc.)
  - Timestamp
  - Cookies
  - and more…
- Reputation scores

What data do we have to score logins?

- Request data:
  - IP address (and derived country, ISP, etc.)
  - Browser's user agent (and OS, version, etc.)
  - Timestamp
  - Cookies
  - and more…
- Reputation scores
- Global counters

What data do we have to score logins?

- Request data:
  - IP address (and derived country, ISP, etc.)
  - Browser's user agent (and OS, version, etc.)
  - Timestamp
  - Cookies
  - and more…
- Reputation scores
- Global counters
- History of member's previous (successful) logins

Formalizing the problem further…

The scoring model must decide whether

$$\frac{P[\text{attack}|u, X]}{P[\text{legitimate}|u, X]} > 1$$

X = random variable representing vector of user data (timestamp, IP address, user agent, etc.)

u = random variable representing user whose account is being accessed

Computation isn't straightforward…

The scoring model must decide whether

$$\frac{P[\text{attack}|u, X]}{P[\text{legitimate}|u, X]} > 1$$

Hard to estimate likelihood ratio directly from the data:

- Most members are never attacked (numerator is 0)
- Only a few samples per member.
- Members come from previously unseen values of *X* (IP addresses, browsers, etc.)

Computing the likelihood of attack

Computing the likelihood of attack

Assumptions:

- Attack features are independent of the member being attacked
- Features are class conditionally independent

Computing the likelihood of attack

$$\frac{\Pr[\text{attack}|u, X]}{\Pr[\text{legitimate}|u, X]} = \Pr[\text{attack}|X] \cdot \frac{\Pr[X]}{\Pr[X|u]} \cdot \frac{\Pr[u|\text{attack}]}{\Pr[u]}$$

Computing the likelihood of attack

Asset Reputation Score
(interpreted as a probability)

$$\frac{\Pr[\text{attack}|u, X]}{\Pr[\text{legitimate}|u, X]} = \Pr[\text{attack}|X] \cdot \frac{\Pr[X]}{\Pr[X|u]} \cdot \frac{\Pr[u|\text{attack}]}{\Pr[u]}$$

Computing the likelihood of attack

Global likelihood of
seeing data *X*

Asset Reputation Score
(interpreted as a probability)

$$\frac{\Pr[\text{attack}|u, X]}{\Pr[\text{legitimate}|u, X]} = \Pr[\text{attack}|X] \cdot \frac{\Pr[X]}{\Pr[X|u]} \cdot \frac{\Pr[u|\text{attack}]}{\Pr[u]}$$

Computing the likelihood of attack

Global likelihood of
seeing data *X*

Asset Reputation Score
(interpreted as a probability)

$$\frac{\Pr[\text{attack}|u, X]}{\Pr[\text{legitimate}|u, X]} = \Pr[\text{attack}|X] \cdot \frac{\Pr[X]}{\Pr[X|u]} \cdot \frac{\Pr[u|\text{attack}]}{\Pr[u]}$$

Appearance of data *X*
in *u*’s (legitimate) login history

Computing the likelihood of attack

Global likelihood of seeing data *X*

Asset Reputation Score (interpreted as a probability)

Value of account to attacker

$$\frac{\Pr[\text{attack}|u, X]}{\Pr[\text{legitimate}|u, X]} = \Pr[\text{attack}|X] \cdot \frac{\Pr[X]}{\Pr[X|u]} \cdot \frac{\Pr[u|\text{attack}]}{\Pr[u]}$$

Appearance of data *X* in *u*’s (legitimate) login history

Computing the likelihood of attack

Global likelihood of seeing data *X*

Asset Reputation Score (interpreted as a probability)

Value of account to attacker

$$\frac{\Pr[\text{attack}|u, X]}{\Pr[\text{legitimate}|u, X]} = \Pr[\text{attack}|X] \cdot \frac{\Pr[X]}{\Pr[X|u]} \cdot \frac{\Pr[u|\text{attack}]}{\Pr[u]}$$

Appearance of data *X* in *u*'s (legitimate) login history

Likelihood of member *u* logging in

Computing the likelihood of attack

Global likelihood of seeing data *X*

Asset Reputation Score (interpreted as a probability)

Value of account to attacker

$$\frac{\Pr[\text{attack}|u, X]}{\Pr[\text{legitimate}|u, X]} = \Pr[\text{attack}|X] \cdot \frac{\Pr[X]}{\Pr[X|u]} \cdot \frac{\Pr[u|\text{attack}]}{\Pr[u]}$$

Appearance of data *X* in *u*'s (legitimate) login history

Likelihood of member *u* logging in

No per-member attack data required!

Computing the likelihood of attack

Global likelihood of
seeing data *X*

Asset Reputation Score
(interpreted as a probability)

Value of account
to attacker

$$\frac{\Pr[\text{attack}|u, X]}{\Pr[\text{legitimate}|u, X]} = \Pr[\text{attack}|X] \cdot \frac{\Pr[X]}{\Pr[X|u]} \cdot \frac{\Pr[u|\text{attack}]}{\Pr[u]}$$

Appearance of data *X*
in *u*’s (legitimate) login history

Likelihood of member *u*
logging in

Remember we said members come from previously unseen values of x (IP addresses, browsers, etc.) …

Smoothing

Q: How do we estimate $\Pr[X|u]$ when X is an IP address that u has never logged in from?

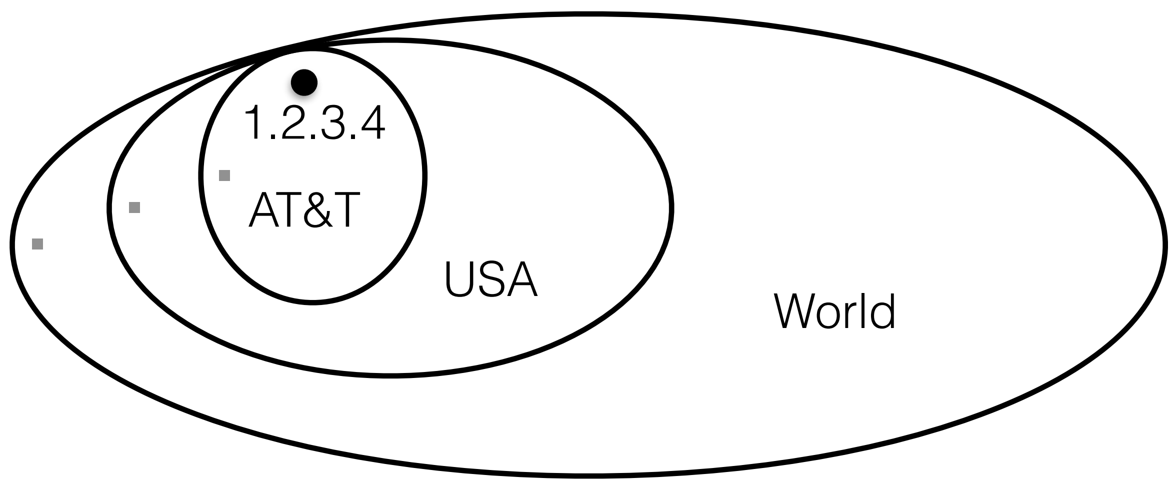
A: We have auxiliary information about unseen IPs:

- Use ISP- or country-level data to estimate probabilities.
- Give higher weight to **unseen** events from a **known** ISP.

Smoothing

Q: How do we estimate $\Pr[X|u]$ when *X* is an IP address that *u* has never logged in from?

A: We have auxiliary information about unseen IPs:



- Use ISP- or country-level data to estimate probabilities.
- Give higher weight to **unseen** events from a **known** ISP.

Smoothing via Backoff

$$P_{\text{backoff}}[X|u] = P_{K=k}[X|u]$$

where K represents level of granularity and k represents the most granular level.

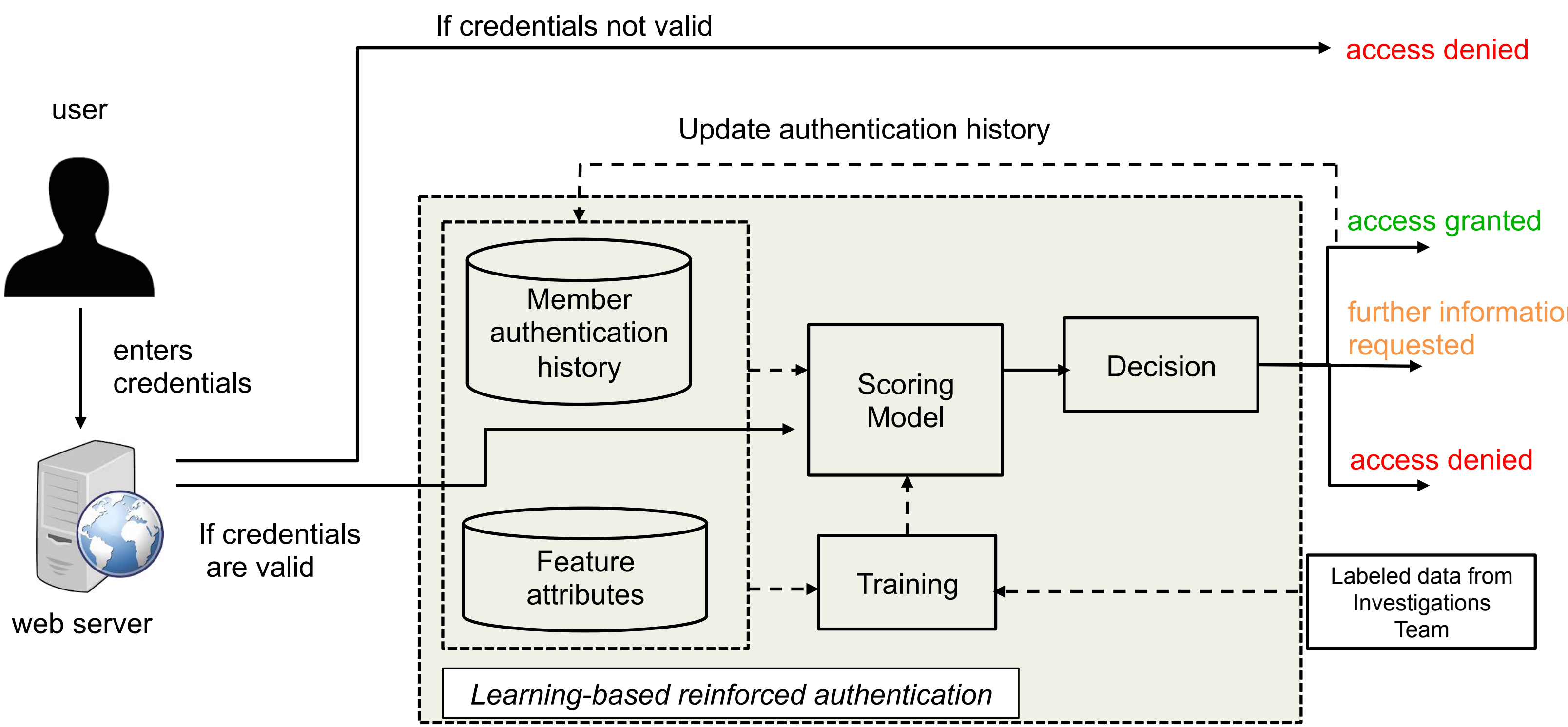
Smoothing via Interpolation

Or, take a linear combination of the estimates $P_K[X|u]$

$P_{\text{interp}}[X|u] = \sum_K \lambda_K P_K[X|u]$

where K represents various levels of granularity.

System architecture



Experiments

Prototype model using two features:

IP hierarchy & user-agent hierarchy

Test data:

- 6 months of successful login attempts (compromised and legitimate)
- unsuccessful login attempts from botnet observed in Jan 2015

Simple Heuristic: Country Mismatch

- 99% of Jan 2015 attack blocked on country mismatch
- 6 Month dataset:
  - Detection rate: 7% , False Positives: 4%

Experiments

Attacker	AUC	TP @ 10% FP
Dumb password-only	1.00	1.00
Simulated botnet	0.99	0.99
Researching	0.99	0.99
Phishing	0.92	0.74
Real Botnet	0.97	0.95
Compromised accounts	0.93	0.77

Simulated four attacks:

- Dumb attack: single IP, scripting useragent
- Botnet attacker: rotates IPs and useragents
- Researching attacker: scrapes target's country info
- Phishing attacker: captures IP and user agent data

Further directions

Can the adversary learn the classification boundary?

- How many queries are necessary?

Use nearline scoring to further classify “gray area.”

- Combine login score with post-login activity.

More features!

Questions?

sjain2@linkedin.com

[p.s. we're hiring!]