463.19 Password Security (Extra)
GoDaddy Data Breach Exposes Over 1 Million WordPress Customers' Data

- Email addresses and customer numbers of up to 1.2 million active and inactive Managed WordPress customers
- Original WordPress Admin password that was set at the time of provisioning was exposed
- sFTP and database usernames and passwords associated with its active customers, and
- SSL private keys for a subset of active customers

According to Wordfence CEO Mark Maunder, "GoDaddy stored sFTP passwords in such a way that the plaintext versions of the passwords could be retrieved, rather than storing salted hashes of these passwords, or providing public key authentication, which are both industry best practices."
Background of Password Guessing
Password Strength Evaluation
Password Reuse
Means of Authentication

• **Something you know**
  – Password or PIN

• **Something you have**
  – Smart card
  – Private key (of a public-private key pair)
  – Phone (running 2FA)

• **Something you are**
  – Biometrics (e.g., iris or fingerprint)
Means of Authentication (Cont.)

• Somewhere you are
  – Location-limited channels

• Someone you know (social authentication)
  – Someone vouches for you
  – You can identify people you should know

• Some system vouches for you
  – Single sign-on
  – PKI Certificate Authorities
Password Advantages

- Familiar to people
- You can have many different ones
- Nothing to carry
- Easy to revoke / replace
- Easy to deploy
- Low cost
- Doesn’t require a trusted third party
- Not linked to an individual*
Disadvantages of Passwords

- Predictability
- Interference between multiple passwords
  - Limits of human memory
  - Password reuse or “trivial” modification
- Requiring a large portfolio of passwords
- Easy to deploy incorrectly / naively
  - System administrators (store in plaintext?)
  - Users
What about Biometrics?

- Fingerprint
- Retina scans
- Face recognition
- Finger/hand geometry
- Voice or speech recognition
- (Many others)
Practical Challenges for Biometrics

• You cannot change them or create a new one (fingerprint?)
• Potentially sensitive data (identifiable info)
• High equipment costs
• Sensitive to changes in the environment
• Biometrics can change over time
• Easy to forge?
Password Guessing: Two Threat Models

• Online guessing
  – Usually has a rate limit
  – Must guess it correctly within a few attempts

• Offline guessing
  – To crack the password hashes
  – Leaked pwd databases where pwds are stored in a hashed format
  – Inefficient if the password is also “salted”
Passwords, Hashes, Salt

- **Password database**
  - Not a good idea to store plaintext directly

- **Login without directly matching plaintext password:**
  - HASH(input password + salt) → password hash
  - Plaintext password is stored in other places
  - Password hash and salt is used to authenticate users
Security of Server-side Password Storing

- Worse way: storing password in plaintext
  - e.g.: username1, password1_plaintext

- Slight better, but still not secure
  - E.g., username1, hashed(password1)

- The right way: adding salt
  - Salt: a fixed length random long string
  - E.g., username1, hashed(password1+salt1), salt1
Background of Password Guessing
Password Strength Evaluation
Password Reuse

Slides credit: Blase Ur
How strong is a particular password?
By looking at them?

- iloveyou
- n(c$JZX!2dfa^dafdIAX^N
- j@mesb0nd007

Leet transformation
How to **Measure** password strength?

- Number of characters, types of characters
- Shannon entropy
- John the Ripper (password cracking software)

- Which one is better?
Old metric: Entropy

• Calculated based on input symbol size (many)
  – Doesn’t account for human patterns
• NIST back-of-envelope estimate (NIST 2006)
  – Vague, not empirical
• Estimated Shannon entropy (Shay 2010)
  – Requires big sample sizes, underestimates

• Average, doesn’t tell you about your weak links
Better Way: Guessability (Offline Guessing)

• How many guesses to reach password?
  – Subject to guessing algorithm, training data
  – Calculate quickly via lookup algorithm
  – Most research focuses on offline guessing model

• Result: guess number or beyond cutoff
  – Model real attacker
  – Per-password estimates

<table>
<thead>
<tr>
<th>Password</th>
<th>Guess number</th>
</tr>
</thead>
<tbody>
<tr>
<td>12345678</td>
<td>4</td>
</tr>
<tr>
<td>Password178</td>
<td>$1.4 \times 10^6$</td>
</tr>
<tr>
<td>jn%fKXsl!8@Df</td>
<td>Beyond cutoff</td>
</tr>
</tbody>
</table>

Example:
Perception vs. Reality
Evaluating Password Pairs

- iloveyou88
- ieatkale88
- 4,000,000,000 x more secure!

- brooklyn16
- brooklynnqy
- 300,000 x more secure!
Ways People Were Wrong

• Overstated security benefits of:
  – Digits
  – Character substitutions (e.g., a→@)
  – Keyboard patterns (e.g., 1qaz2wsx3edc)

• Did not recognize common words/phrases
Many Ways People Were Right

• Capitalize letters other than the first
• Put digits and symbols in middle, not end
• Use symbols rather than digits
• Avoid:
  – Common first names
  – Words related to account
  – Years and sequences
Different Ways to Guess Passwords

• Guessing attacks are data-driven
  – Previously stolen passwords
  – Natural-language corpora

• Array of tools
  – Cracking software
  – Academic algorithms
Markov Models

- Predicts future characters from previous
- Approach requires weighted data:
  - Passwords
  - Dictionaries
Markov Models: Basic Idea
Markov Models: Basic Idea
Deep Learning based Password Guessing

Background of Password Guessing
Password Strength Evaluation
Password Reuse

Data-Driven Analysis: Password Reuse & Modification

• Collect massive password datasets with email addresses
  – Link the same users’ passwords across services

• Data collection method
  – Searched through online forums, data archives, darknet markets
  – Obtained 107 public password datasets leaked during 2008-2016

<table>
<thead>
<tr>
<th>Raw Dataset</th>
<th>Reverse Hashed PWs</th>
<th>Plaintext Passwords</th>
<th>Final Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>107 Services</td>
<td></td>
<td>460 Million Plaintext PWs (93%)</td>
<td>107 Services</td>
</tr>
<tr>
<td>428 Million Users</td>
<td></td>
<td></td>
<td>28 Million Users</td>
</tr>
<tr>
<td>497 Million PWs</td>
<td></td>
<td></td>
<td>61 Million PWs</td>
</tr>
</tbody>
</table>
## Diverse Categories of Online Services

<table>
<thead>
<tr>
<th>Category</th>
<th>#Plain PWs</th>
<th># Datasets</th>
<th>Top 3 Largest Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social</td>
<td>286,000,000</td>
<td>7</td>
<td>Myspace, VK.com, LinkedIn</td>
</tr>
<tr>
<td>Adult</td>
<td>75,200,000</td>
<td>9</td>
<td>Zoosk, Mate1, YouPorn</td>
</tr>
<tr>
<td>Game</td>
<td>40,800,000</td>
<td>13</td>
<td>Neopets, 7k7k, Lbsg</td>
</tr>
<tr>
<td>Entertainment</td>
<td>30,700,000</td>
<td>4</td>
<td>Lastfm, Swingbrasileiro, LATimes</td>
</tr>
<tr>
<td>Internet</td>
<td>16,400,000</td>
<td>18</td>
<td>000webhost, Comcast, Yahoo</td>
</tr>
<tr>
<td>Email</td>
<td>9,600,000</td>
<td>3</td>
<td>Gmail, Mail.ru, Yandex</td>
</tr>
<tr>
<td>Forum</td>
<td>1,100,000</td>
<td>25</td>
<td>CrackingForum, Abusewith.us, Gawker</td>
</tr>
<tr>
<td>Shopping</td>
<td>340,000</td>
<td>12</td>
<td>RedBox, 1394store, Myaribags</td>
</tr>
<tr>
<td>Others</td>
<td>210,000</td>
<td>7</td>
<td>Data1, Data2, Data3</td>
</tr>
<tr>
<td>Business</td>
<td>10,000</td>
<td>9</td>
<td>Movatiathletic, Hrsupporten, 99Fame</td>
</tr>
</tbody>
</table>
Research Questions

• How often do users reuse or modify passwords across services?

• How long does it take for users to update their reused passwords after data breaches?

• How guessable are the modified passwords?
Measuring Password Reuse and Modification

- 37 million password pairs from the same users
- Given a pw pair, determine “reused”, “modified”, or “unknown”
Highlights of Findings

• 53% of the 28.8 million users reused/modified passwords
  – 38% users once reused the same password
  – 21% users once modified an existing password for a new service

• Sensitive services received most reused/modified passwords
  – Ratio = (\# reused+modified pws) / (\# pws of a service category)
  – Shopping services have the highest ratio (85%)
  – Email services are at the second place (62%)
Password Guessing

Password modification patterns have a low variance

• Given a user’s leaked PW \(\rightarrow\) guess modified PW of un-breached services
• Possible for online guessing

Training-based guessing schemes

• Learn the different rules of transform one password to a new one
• Given a password, learn the optimized orders to apply the transform rules
  – Bayesian inference model
Password Guessing Results

0.1% training data, guess the rest of 99.9% modified passwords

30% of modified PWs guessed in 10 attempts
(4.2M out of 14.4M modified PWs)

12.2M reused → 1 guess

Modified passwords highly predicable
Summary

- Offline and online guessing model are very different!
- Password reuse and modification are still common
- Modified passwords are highly predictable
- Password strength meters should consider online guessing models