The science of guessing
analyzing an anonymized corpus of 70 million passwords

Joseph Bonneau
jcb82@cl.cam.ac.uk

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

IEEE Symposium on Security & Privacy
≈ Oakland, CA, USA
May 23, 2012
Why do password research in 2012?

Compatible Time-Sharing System, MIT 1961
Research goal

Precisely compute the guessing difficulty of a given population’s password distribution
Research goal

Compare the **guessing difficulty** of password distributions chosen by different populations
Research goal

Compare the **guessing difficulty** of password distributions chosen by different populations
Research goal

Compare the guessing difficulty of password distributions chosen by different populations

Joseph Bonneau (University of Cambridge)
Research goal

Compare the **guessing difficulty** of password distributions chosen by different populations

Password: 

Retype Password: 

VS.

Password: 

Strong: 

Capitalization matters. Use 6 to 32 characters, and don't use your name or Yahoo! ID.

Re-type Password: 

For a more secure password:
- Use both letters and numbers
- Add special characters (such as @, ?, %)
- Mix capital and lowercase letters
Research goal

Compare the **guessing difficulty** of password distributions chosen by different populations
Approach #1: Semantic password evaluation

- How long are the passwords?
- Do they look like English words?
- What kind of characters do they contain?
## Approach #1: Semantic password evaluation

<table>
<thead>
<tr>
<th>Length Char.</th>
<th>No Checks</th>
<th>Dictionary Rule</th>
<th>Dict. &amp; Comp. Rule</th>
<th>10 char. alphabet</th>
<th>94 char alphabet</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>3.3</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>-</td>
<td>-</td>
<td>5</td>
<td>6.7</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>-</td>
<td>-</td>
<td>7</td>
<td>10.0</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>14</td>
<td>16</td>
<td>9</td>
<td>13.3</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>17</td>
<td>20</td>
<td>10</td>
<td>16.7</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>20</td>
<td>23</td>
<td>11</td>
<td>20.0</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>22</td>
<td>27</td>
<td>12</td>
<td>23.3</td>
</tr>
<tr>
<td>8</td>
<td>18</td>
<td>24</td>
<td>30</td>
<td>13</td>
<td>26.6</td>
</tr>
<tr>
<td>10</td>
<td>21</td>
<td>26</td>
<td>32</td>
<td>15</td>
<td>33.3</td>
</tr>
<tr>
<td>12</td>
<td>24</td>
<td>28</td>
<td>34</td>
<td>17</td>
<td>40.0</td>
</tr>
<tr>
<td>14</td>
<td>27</td>
<td>30</td>
<td>36</td>
<td>19</td>
<td>46.6</td>
</tr>
<tr>
<td>16</td>
<td>30</td>
<td>32</td>
<td>38</td>
<td>21</td>
<td>53.3</td>
</tr>
<tr>
<td>18</td>
<td>33</td>
<td>34</td>
<td>40</td>
<td>23</td>
<td>59.9</td>
</tr>
<tr>
<td>20</td>
<td>36</td>
<td>36</td>
<td>42</td>
<td>25</td>
<td>66.6</td>
</tr>
<tr>
<td>22</td>
<td>38</td>
<td>38</td>
<td>44</td>
<td>27</td>
<td>73.3</td>
</tr>
<tr>
<td>24</td>
<td>40</td>
<td>40</td>
<td>46</td>
<td>29</td>
<td>79.9</td>
</tr>
<tr>
<td>30</td>
<td>46</td>
<td>46</td>
<td>52</td>
<td>35</td>
<td>99.9</td>
</tr>
<tr>
<td>40</td>
<td>56</td>
<td>56</td>
<td>62</td>
<td>45</td>
<td>133.2</td>
</tr>
</tbody>
</table>

**NIST “entropy” formula**
Approach #2: Cracking experiments
Approach #2: Cracking experiments

\[ \alpha = \text{proportion of passwords guessed} \]

\[ \mu = \log(\text{dictionary size}) \]

- Morris and Thompson [1979]
- Klein [1990]
- Spafford [1992]
- Wu [1999]
- Kuo [2006]
- Schneier [2006]
- Dell’Amico (it) [2010]
- Dell’Amico (fi) [2010]
- Dell’Amico (en) [2010]
<table>
<thead>
<tr>
<th></th>
<th>semantic</th>
<th>cracking</th>
</tr>
</thead>
<tbody>
<tr>
<td>external validity</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>no operator bias</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>no demographic bias</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>repeatable</td>
<td>✓</td>
<td>?</td>
</tr>
<tr>
<td>easy</td>
<td>✓</td>
<td>?</td>
</tr>
</tbody>
</table>
My approach

1. Collect password data on a huge scale
2. Compare populations as probability distributions
3. Test hypotheses using different populations
My approach

1. Collect password data on a huge scale
2. Compare populations as probability distributions
3. Test hypotheses using different populations
My approach

1. Collect password data on a huge scale
2. Compare populations as probability distributions
3. Test hypotheses using different populations
Goal #1: collect a massive data set

- with cooperation from Yahoo!
- privacy-preserving collection 😊
  - histograms only
- demographic splits collected
Collecting large-scale data at Yahoo!

Internet

Collection Proxy

Login Server

user: joe
pass: 12345

12345

Joseph Bonneau (University of Cambridge)
Collecting large-scale data at Yahoo!

Internet

Collection Proxy

Login Server

user: joe
pass: 12345

\( H(12345) \)
Collecting large-scale data at Yahoo!

Internet

user: joe
pass: 12345

Collection Proxy

$H(K || 12345)$

Login Server
Collecting large-scale data at Yahoo!

SELECT gender, lang, age
FROM users
WHERE user = joe

m, en, 21-34

\[ H(K||12345) \]
m, en, 21-34

user: joe
pass: 12345

Joseph Bonneau (University of Cambridge)
The science of guessing
May 23, 2012 10 / 33
Collecting large-scale data at Yahoo!

- Internet
  - User database
    - SELECT gender, lang, age
      - FROM users
      - WHERE user = joe
    - m, en, 21-34

- Collection Proxy
  - H(K||12345)

- Login Server
  - user: joe
    - pass: 12345
  - user: joe
    - pass: 123456

- Seen users
  - gender=m
  - lang=en
  - age=21-34

Joseph Bonneau (University of Cambridge)
Collecting large-scale data at Yahoo!

Internet

gender=m

lang=en

age=21-34

Login Server
Collecting large-scale data at Yahoo!

- Experiment run May 23–25, 2011
- 69,301,337 unique users
- 42.5% unique
- 328 different predicate functions
Goal #2: model guessing as a probability problem

- Assume **perfect knowledge** of the distribution $\mathcal{X}$
- $\mathcal{X}$ has $N$ events (passwords) $x_1, x_2, \ldots$
- Events have probability $p_1 \geq p_2 \geq \ldots \geq p_N \geq 0$
- Each user chooses at random $X \xleftarrow{\text{R}} \mathcal{X}$

**Question:** How hard is it to guess $X$?
Shannon entropy

\[ H_1(X) = - \sum_{i=1}^{N} p_i \log p_i \]

**Interpretation:** Expected number of queries “Is \( X \in S \)” for arbitrary subsets \( S \subseteq X \) needed to guess \( X \). *(Source-Coding Theorem)*
\[ G_1(\mathcal{X}) = E \left[ \text{\#guesses} \right] = \sum_{i=1}^{N} p_i \cdot i \]

**Interpretation:** Expected number of queries “Is \( X = x_i \)?” for \( i = 1, 2, \ldots, N \) (optimal sequential guessing)
$G_1$ fails badly for real password distributions

Random 128-bit passwords in the wild at RockYou ($\sim 2^{-20}$)

\[
\begin{align*}
ed65e09b98bdc70576d6c5f5e2ee38a9 \\
e54d409c55499851aeb25713c1358484 \\
dee489981220f2646eb8b3f412c456d9 \\
c4df8d8e225232227c84d0ed8439428a \\
bd9059497b4af2bb913a8522747af2de \\
b25d6118ffcc44b12b014feb81ea68e49 \\
aac71eb7307f4c54b12c92d9bd45575f \\
9475d62e1f8b13676deab3824492367a \\
92965710534a9ec4b30f27b1e7f6062a \\
80f5a0267920942a73693596fe181fb7 \\
76882fb85a1a8c6a83486aba03c031c9 \\
6a60e0e51a3eb2e9fed6a546705de1bf \ldots
\end{align*}
\]

$\Rightarrow \quad G_1(\text{RockYou}) > 2^{107}$
Attackers might be happy ignoring the hard values
\[ \mu_\alpha(X) = \min \left\{ \mu \in [1, N] \mid \sum_{i=1}^{\mu} p_i \geq \alpha \right\} \]

**Interpretation:** Minimal dictionary size to succeed with probability \( \alpha \)
\[ G_\alpha(\mathcal{X}) = (1 - \lceil \alpha \rceil) \cdot \mu_\alpha(\mathcal{X}) + \sum_{i=1}^{\mu_\alpha(\mathcal{X})} p_i \cdot i \]

**Interpretation:** Mean number of guesses to succeed with probability \( \alpha \)
Guessing curves visualise all possible attacks

\[
\mu_\alpha(U_{10^4})
\]

\[
\mu_\alpha(U_{10^3})
\]

\[
\mu_\alpha(\text{PIN})
\]

\[
G_\alpha(\text{PIN})
\]

success rate $\alpha$

dictionary size/number of guesses
More intuitive after converting to bits

The science of guessing

May 23, 2012 19 / 33
More intuitive after converting to bits
More intuitive after converting to bits
More intuitive after converting to bits

The science of guessing

Joseph Bonneau (University of Cambridge)
More intuitive after converting to bits
More intuitive after converting to bits

\[ H_\infty \xrightarrow{\sim} G_1 \xrightarrow{\sim} H_0 \xrightarrow{\sim} H_1 \rightarrow H_2 \rightarrow \tilde{\mu}_\alpha(U_{10^4}) / \tilde{G}_\alpha(U_{10^4}) \]

\[ \tilde{\mu}_\alpha(U_{10^3}) / \tilde{G}_\alpha(U_{10^3}) \]

\[ \tilde{\mu}_\alpha(\text{PIN}) \]

\[ \tilde{G}_\alpha(\text{PIN}) \]
Sample size is a major problem for passwords...
Predict our confidence range by bootstrapping

\[ \alpha \text{-work-factor } \tilde{\mu}_\alpha \text{ (bits)} \]

- \( M = 69,301,337 \) (full)
- \( M = 10,000,000 \) (sampled)
- \( M = 1,000,000 \) (sampled)
- \( M = 500,000 \) (sampled)
Extrapolation w/ truncated Sichel-Poisson distribution

\[ \alpha \text{-work-factor } \tilde{\mu}_\alpha \text{ (bits)} \]

\[ M = 69,301,337 \text{ (full)} \]
\[ M = 10,000,000 \text{ (sampled)} \]
\[ M = 1,000,000 \text{ (sampled)} \]
\[ M = 500,000 \text{ (sampled)} \]
Goal #3: Analyze Yahoo! passwords
Goal #3: Analyze Yahoo! passwords

The science of guessing

Joseph Bonneau (University of Cambridge)
Goal #3: Analyze Yahoo! passwords

The science of guessing

May 23, 2012 25 / 33
Demographic trends: nationality

The science of guessing

May 23, 2012 26 / 33
Demographic trends: age

The science of guessing

May 23, 2012 27 / 33
Credit card details make little difference

The science of guessing

Joseph Bonneau (University of Cambridge)
Password strength meter makes little difference
Demographic summary

- there is no “good group” of users
- differences small but statistically significant
- online attack 6–9 bits ($\tilde{\lambda}_{10}$)
- offline attack 15–25 bits ($\tilde{G}_{0.5}$)
Surprisingly little language variation

<table>
<thead>
<tr>
<th>target</th>
<th>de</th>
<th>en</th>
<th>es</th>
<th>fr</th>
<th>id</th>
<th>it</th>
<th>ko</th>
<th>pt</th>
<th>zh</th>
<th>vi</th>
<th>global</th>
</tr>
</thead>
<tbody>
<tr>
<td>de</td>
<td>6.5%</td>
<td>3.3%</td>
<td>2.6%</td>
<td>2.9%</td>
<td>2.2%</td>
<td>2.8%</td>
<td>1.6%</td>
<td>2.1%</td>
<td>2.0%</td>
<td>1.6%</td>
<td>3.5%</td>
</tr>
<tr>
<td>en</td>
<td>4.6%</td>
<td>8.0%</td>
<td>4.2%</td>
<td>4.3%</td>
<td>4.5%</td>
<td>4.3%</td>
<td>3.4%</td>
<td>3.5%</td>
<td>4.4%</td>
<td>3.5%</td>
<td>7.9%</td>
</tr>
<tr>
<td>es</td>
<td>5.0%</td>
<td>5.6%</td>
<td>12.1%</td>
<td>4.6%</td>
<td>4.1%</td>
<td>6.1%</td>
<td>3.1%</td>
<td>6.3%</td>
<td>3.6%</td>
<td>2.9%</td>
<td>6.9%</td>
</tr>
<tr>
<td>fr</td>
<td>4.0%</td>
<td>4.2%</td>
<td>3.4%</td>
<td>10.0%</td>
<td>2.9%</td>
<td>3.2%</td>
<td>2.2%</td>
<td>3.1%</td>
<td>2.7%</td>
<td>2.1%</td>
<td>5.0%</td>
</tr>
<tr>
<td>id</td>
<td>6.3%</td>
<td>8.7%</td>
<td>6.2%</td>
<td>6.3%</td>
<td>14.9%</td>
<td>6.2%</td>
<td>5.8%</td>
<td>6.0%</td>
<td>6.7%</td>
<td>5.9%</td>
<td>9.3%</td>
</tr>
<tr>
<td>it</td>
<td>6.0%</td>
<td>6.3%</td>
<td>6.8%</td>
<td>5.3%</td>
<td>4.6%</td>
<td>14.6%</td>
<td>3.3%</td>
<td>5.7%</td>
<td>4.0%</td>
<td>3.2%</td>
<td>7.2%</td>
</tr>
<tr>
<td>ko</td>
<td>2.0%</td>
<td>2.6%</td>
<td>1.9%</td>
<td>1.8%</td>
<td>2.3%</td>
<td>2.0%</td>
<td>5.8%</td>
<td>2.4%</td>
<td>3.7%</td>
<td>2.2%</td>
<td>2.8%</td>
</tr>
<tr>
<td>pt</td>
<td>3.9%</td>
<td>4.3%</td>
<td>5.8%</td>
<td>3.8%</td>
<td>3.9%</td>
<td>4.4%</td>
<td>3.5%</td>
<td>11.1%</td>
<td>3.9%</td>
<td>2.9%</td>
<td>5.1%</td>
</tr>
<tr>
<td>zh</td>
<td>1.9%</td>
<td>2.4%</td>
<td>1.7%</td>
<td>1.7%</td>
<td>2.0%</td>
<td>2.0%</td>
<td>2.9%</td>
<td>1.8%</td>
<td>4.4%</td>
<td>2.0%</td>
<td>2.9%</td>
</tr>
<tr>
<td>vi</td>
<td>5.7%</td>
<td>7.7%</td>
<td>5.5%</td>
<td>5.8%</td>
<td>6.3%</td>
<td>5.7%</td>
<td>6.0%</td>
<td>5.8%</td>
<td>7.0%</td>
<td>14.3%</td>
<td>7.8%</td>
</tr>
</tbody>
</table>

With 1000 guesses, greatest efficiency loss is only 4.8 (fr/vi)

Joseph Bonneau and Rubin Xu.

## Comparing password analysis methods

<table>
<thead>
<tr>
<th></th>
<th>Semantic</th>
<th>Cracking</th>
<th>Statistical</th>
</tr>
</thead>
<tbody>
<tr>
<td>External validity</td>
<td>✓</td>
<td>✓</td>
<td>?</td>
</tr>
<tr>
<td>No operator bias</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>No demographic bias</td>
<td>?</td>
<td>?</td>
<td>✓</td>
</tr>
<tr>
<td>Repeatable</td>
<td>✓</td>
<td>?</td>
<td>✓</td>
</tr>
<tr>
<td>Easy</td>
<td>✓</td>
<td>?</td>
<td>✓</td>
</tr>
</tbody>
</table>
Comparing password analysis methods

<table>
<thead>
<tr>
<th></th>
<th>semantic</th>
<th>cracking</th>
<th>statistical</th>
</tr>
</thead>
<tbody>
<tr>
<td>external validity</td>
<td></td>
<td>✓</td>
<td>?</td>
</tr>
<tr>
<td>no operator bias</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>no demographic bias</td>
<td>?</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>repeatable</td>
<td>✓</td>
<td>?</td>
<td>✓</td>
</tr>
<tr>
<td>easy</td>
<td>✓</td>
<td>?</td>
<td>✓</td>
</tr>
<tr>
<td>works w/small data</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
The picture so far

![Graph showing the relationship between success rate $\alpha$ and $\tilde{\mu}_\alpha$ (bits).]

- Password (YAHOO)
- Password (RockYou)
- Surname (Facebook)
- Forename (Facebook)
- PIN (iPhone)
- Password [Morris]
- Password [Klein]
- Password [Spafford]
- Mnemonic [Kuo]
- Faces [Davis]
- PassPoints [Thorpe]

Joseph Bonneau (University of Cambridge)

The science of guessing

May 23, 2012
For more information

my email
jcb82@cl.cam.ac.uk

my dissertation
Guessing human-chosen secrets
## Acknowledgements

<table>
<thead>
<tr>
<th>Company/Institution</th>
<th>Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo!</td>
<td>Elizabeth Zwicky, Henry Watts, Ram Marti, Clarence Chung, Christopher Harris</td>
</tr>
<tr>
<td>University of Cambridge Computer Laboratory</td>
<td>Ross Anderson, Richard Clayton, Frank Stajano, Markus Kuhn, Saar Drimer, Andrew Lewis</td>
</tr>
<tr>
<td></td>
<td>Paul van Oorschot, Cormac Herley, Arvind Narayanan</td>
</tr>
</tbody>
</table>
Converting metrics to bits

- Find the size of a uniform distribution $\mathcal{U}_N$ with equivalent security

**Easy case:**

$$\tilde{\mu}_\alpha(\mathcal{X}) = \lg \left( \frac{\mu_\alpha(\mathcal{X})}{\lceil \alpha \rceil} \right)$$

**More complicated:**

$$\tilde{\mathcal{G}}_\alpha(\mathcal{X}) = \lg \left[ \frac{2 \cdot \mathcal{G}_\alpha(\mathcal{X})}{\lceil \alpha \rceil} - 1 \right] - \lg(2 - \lfloor \alpha \rfloor)$$

**Sanity check:**

$$\tilde{\lambda}_\beta(\mathcal{U}_N) = \tilde{\mu}_\alpha(\mathcal{U}_N) = \tilde{\mathcal{G}}_\alpha(\mathcal{U}_N) = \lg N$$
Sample size is a major problem for passwords...

\[
\begin{align*}
\hat{H}_0 & \\
\hat{G}_1 & \\
\hat{H}_1 & \\
\hat{\mu}_{0.25} & \\
\hat{G}_{0.25} & \\
\hat{\lambda}_{10} & \\
\hat{\lambda}_1 & 
\end{align*}
\]
Poor password implementations

Results from a study of password authentication in the wild:

- 29–40% of websites don’t hash passwords during storage
- 41% of websites don’t use any encryption for password submission
  - 22% do so incompletely
- 84% of websites don’t rate-limit against guessing attacks
- 97% of websites leak usernames to simple

Joseph Bonneau and Sören Preibusch.

The password thicket: technical and market failures in human authentication on the web.

*Workshop on the Economics of Information Security, 2010.*