Surveys, interviews, and diary studies

Michelle Mazurek

(some slides adapted from Blase Ur, Lorrie Cranor, and Rich Shay)
Today

• Finish up field studies / crowdsourcing
• Surveys
• Interviews
• Diary studies
Case study: Measuring password strength

ECOLOGICAL VALIDITY, CROWDSOURCING, FIELD STUDIES
Passwords research is everywhere

**CCS 2005** (Narayanan and Shmatikov)

**CCS 2010** (Weir et al.)

**CHI 2011** (Komanduri et al.)

**WWW 2007** (Florencio and Herley)

**IEEE S&P 2012** (Bonneau)

**NDSS 2012** (Castelluccia et al.)

(b) Ten most frequent passwords for different sites. Passwords underlined are shared by at least two services. The wide difference likely depend on background (e.g., Faithwriters) or password rules (e.g., MySpace).

Frequency of occurrence of symbols in passwords created in aggressive condition.
... but good data is hard to find

- Small data sets
- Experimental rather than field data
- Self-reported surveys
- Leaked data of questionable validity
- Minimal-value accounts
- No access to plaintext passwords

**Are the results generalizable?**
Fahl et al.: Password study validity

• Goal: Compare lab study, online study, real passwords

• Methods:
  – Several thousand passwords (plaintext, anonymized)
  – Invite same pool to online or lab study
  – Security priming, or not
  – Manual analysis for similarity

• 583 online, 63 lab participants
Results: Validity

<table>
<thead>
<tr>
<th>%</th>
<th>Online</th>
<th>Lab</th>
<th>Priming</th>
<th>Non</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly valid</td>
<td>46</td>
<td>49</td>
<td>47</td>
<td>44</td>
<td>46</td>
</tr>
<tr>
<td>Somewhat valid</td>
<td>23</td>
<td>32</td>
<td>24</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>Invalid</td>
<td>31</td>
<td>18</td>
<td>29</td>
<td>32</td>
<td>30</td>
</tr>
</tbody>
</table>

• Overall, experimental data can be useful
  – Self-reporting of realistic behavior can help
• No significant difference due to priming
• Lab slightly but significantly better than online
Critique the study design

• What was measured?
  – Expert analysis, composition, self-reported behavior

• Priming vs. non-priming
  – Would you have given different instructions?
  – Are there other conditions you would test?

• Self-selected participants
Implications of the results

• Do these findings apply to other studies in the security/privacy area? How?
Passwords for an entire university

- 25,000 **real, high-value** passwords from CMU
- Contextual data – logs, demographics, survey
- What factors correlate with password strength?
  - New (to passwords) statistical methods
  - Find new results, confirm prior results
- What to do when you don’t have field data?
  - Comparison with leaked and study data
What are CMU passwords?

• 25,459 accounts for faculty, staff, and students
  – Plus 17,104 deactivated accounts

• Single-sign-on for email, financial, grades, registration, health, etc.

• Password requirements:
  – Minimum 8 characters
  – Upper, lower, digit, symbol
  – Dictionary check (241,497 words)
Strength metric: Guessability

• How many guesses to reach each password?
  – Subject to guessing algorithm and training data
• Result: guess number or beyond the cutoff
  – Cutoff = 380 trillion guesses (runs in about 1 day)

Example:

<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%fKXs1!8@Df</td>
<td>Beyond cutoff</td>
</tr>
</tbody>
</table>
Comparing password sets

• Examining CMU password policy
  – Use conforming subset for all leaked data

• Online studies
  – MTsim: Closest match to real CMU experience
  – MTcomp8: Similar password requirements

• Leaked: plaintext
  – RockYou, Yahoo!, CSDN

• Leaked: hashed and cracked
  – Gawker, StratFor
Comparing sets – Guessability

Leaked hashed/cracked: Very easy to guess
Comparing sets – Guessability

Leaked plaintext: RockYou close to CMU, others much tougher
Comparing sets – Guessability

Online studies: Both close, MTcomp8 closer
Other metrics for comparison

• Composition: length, character classes
• Structures
• Entropy (Shay et al., SOUPS 2010)
• Frequency distribution
Comparing sets – Length

Overall: **Online studies closest** across metrics

(Full results in the paper)
Discussion

• Critique the study design
  – Challenges of field studies

• Are there lessons for other HFSP studies?
Quick note on ethics

• Several studies we discussed Thursday / today have significant ethical implications

• We’ll revisit this in a couple of weeks
  – Any comments/questions in the meantime?
Takeaways

• Crowdsourcing / online data can be valuable
  – When used carefully
  – With important caveats

• Field studies require a lot of effort to get right
  – But can give otherwise impossible data

• Ecological validity is a careful balance
SURVEYS
Why a survey?

• A little bit of data (each) from a lot of people
• Quantitative results
  – Generalizable if done correctly
• Quick, easy, unobtrusive, relatively cheap
• Shallow data
  – Multiple choice, short free-response
• Biases: self-reported, question/answer order, etc.
Survey best practices

• Pilot, pilot, pilot!
  – Ensure questions are neutral, are not ambiguous
  – Test different question wordings

• Consider your sample

• Include attention checks

• Don’t make it too long
  – No shortcuts (branch questions equally)

• Offer option not to answer (avoid lying)
Try it!

In groups of 2-3, write a 5-question survey about privacy for student records.