ESORICS'15 - The Emperor's New Password Creation Policies

优点：①从系统性、基于证据的方法评估口令策略，比之前定义分析好
②对选取网站，评估方法及结果的描述好（如何论证）

问题：①口令策略可以根据实际存储的用户信息来制定防止预测猜测的方法。
②应论述口令策略是为在线猜测攻击而制定的这一说法的合理性。
③可跟踪享主流网站探索口令策略变化的内在原因

▲研究背景

1）研究范围：口令构造策略（Password creation policy）。
①由于用户选择的口令不够安全，相应的帐户容易被破坏，经过良好的设计，口令构造策略可以帮助用户选择可记忆但安全的口令，几乎所有web服务都采用了一定的口令策略。
②口令构造策略包括：password composition rules：要求口令符合一定的复杂度，促使用户使用更强的口令；password strength meter：在用户注册时给口令强度做visual或verbal反馈

2）历史研究的不足或未探索的部分
①广泛部署的口令构造策略在多大程度上可以被依赖。
→之前的结论（i）各个网站的rule和meter不同且它们没有执行较好的。
（ii）更高的安全需求不意味着更严格的rule，反而对可用性无要求的网站会执行严格的rule
（iii）随时间转变，口令策略未做出明显的改善。
②之前的研究是5年前做的，不仅用户数量增多，口令策略可能会变，而且可以使用泄漏的数据集来评估口令强度。
③部分研究关注口令策略的一方面，许多网站使用强制的口令规则但建议的meter，故许为weak的口令仍可能被接受，网站拒绝弱口令抗在线猜测攻击的能力未知。
④很少有工作研究非英文网站，特别是网站世界中的中文网站。

▲研究问题及思路

1）提出systematic, evidence-grounded 方法评估网站策略。调查50个领先服务的策略，包括10类服务（IT, web, 游戏，金融和技术等），每一类3个中文网站，2个英文网站

2）探索中英文网站口令策略选择的不同

3）以更可信赖的方式采用3个口令破解技术评估316个测试口令的强度（4种，漫步十定向）

▲采用的方法

1）选择代表性的网站：10个种类，英文选流量前5，中文选流量前10，15个网站中选5个，共50个，30中文，20英文。
→代表3主流的用户常用的，口令为主要认证方式的网络服务，会影响绝大部分终端用户且可以成为其它网站的model。

2）评估口令策略强度：通常要通过评估在此策略下生成口令数据集的强度完成
由于guessing entropy和α-guesswork需依赖真实数据集，故使用Florencio和Herley的简单方法 $N_{min}\log_2 C_{min}$，N_{min}指最小长度，C_{min}是最小字符集的基数（例数字为10）可以评估口令策略强度的下界

3）利用真实的口令数据集（7个数据集，5个收，2个中文，用于训练破解算法并学习用户口令实践行为的统计结果）

4）评估口令强度：抵抗猜测攻击的能力，本文考虑漫步攻击者和利用用户姓名的定向攻击者
→只考虑不低于4个字符的姓名片段，构建基于name的字典模型。

5）选择测试口令：针对两种攻击采用不同的口令各8个。
其中对于漫步攻击采用3部分中文风格的常见口令如520与1314，对于定向攻击假设用户为中文用户

6）从网站中收集数据：创建真实帐户来评估口令构造规则和强度评测。
未采用自动化方法的原因：①3%网站不允许自动注册，用户需解决CAPTCHA ②18%网站需输入用户手机验证码完成注册 ③8%网站在输入口令前需输入邮箱验证码 ④每个站点的信息是高度异构的，没有网站共享口令策略，批处理不可行。

▲结果

1）实际的口令构造规则：检查的内容包括①长度限制②字符集需求③是否接受特殊符号，④是否使用黑名单，⑤是否检查用户信息⑥是否正式给出口令规则
①增加口令长度通常比扩展字符集对口令安全性的增强效果好。
②没有协商好的口令实践，口令在实现层面未被标准化。
③来自权威机构的建议互不相同且不可行；安全知识、资本、工程资源与策略强度关系不大
④英文网站通常比中文网站执行更严格的策略。

2）混乱的口令强度评估
①及时的、易于理解的、准确反馈的评估可显著提高口令安全性。
②26个网站（16中，10英）提供强度评估，5个是口头显示（3英2中），9个强制要求满足口令规则（5英4中）。注：部分网站在用户口令满足构造策略后才提供强度反馈。
③从用户的角度，部分IT、安全和学术网站未提供口令强度评估，但使用更严格的构造策略，可能令用户费解；从服务器角度，口令组成规则的工程成本更低

3）在线猜测攻击者的攻击
① 50×2×8=800个测试用例，541个被网站接受，其中257个未给出任何强度信息，83个在weak时接受，每个网站至少接受7个实例，许多网站口令策略的目的尚未达到，抵抗在线猜测
②口令强度评测应在过滤不好的口令方面占更重要地位，且前不应用少，且给出高度不统一
③3.3个用于在线猜测的口令被接受，许多评估方法高估了定向猜测中弱口令的强度。