

基于攻击算法的海量真实用户口令数据分析



谢志杰 张 旻 李振汉 王红军

国防科技大学电子对抗学院 合肥 230037

网络空间安全态势感知与评估安徽省重点实验室 合肥 230037

(xzj9510@nudt.edu.cn)

摘要 口令认证是现今主要的身份认证方式,已广泛应用于金融、军事和网络等领域。文中从攻击者的角度对口令的安全性展开研究,利用海量真实的用户数据对口令的一般特征进行统计分析,基于概率上下文无关文法(Probabilistic Context-Free Grammars,PCFG)口令猜测算法、TarGuess-I定向口令猜测模型的口令脆弱性分析,发现了用户在选择生成口令时存在易被攻击者发现并被利用的脆弱行为,如偏好使用简单结构口令、基于模式设计口令、基于语义生成口令以及偏好使用姓名和用户名等个人信息生成口令等,总结了这些脆弱行为的特征,为避免用户设置脆弱口令以及设计口令强度评估方法提供了依据。

关键词: 口令安全;口令猜测;脆弱行为;用户信息

中图法分类号 TP309

Analysis of Large-scale Real User Password Data Based on Cracking Algorithms

XIE Zhi-jie,ZHANG Min,LI Zhen-han and WANG Hong-jun

College of Electronic Engineering,National University of Defense Technology,Hefei 230037,China

Anhui Province Key Laboratory of Cyberspace Security Situation Awareness and Evaluation,Hefei 230037,China

Abstract Password authentication is the main authentication method nowadays. It is widely used in various fields, such as finance, military and internet. In this paper, password security is studied from the perspective of an attacker. Large-scale real user data is used for statistical analyses of password general characteristics, and for password vulnerability analyses based on Probabilistic Context-Free Grammars (PCFG) password guessing algorithm and TarGuess-I targeted password guessing model. Through the above analyses, it is found in users' passwords that there are vulnerable behaviors that can be easily discovered and exploited by attackers, such as choosing simple structure passwords, generating passwords based on patterns, password containing semantics and passwords containing personal information (i. e., name and user name). These vulnerable behavior characteristics are summarized to provide a basis for reminding users to avoid setting weak passwords and studying the method of password strength meter.

Keywords Password security, Password guessing, Vulnerable behaviors, User information

1 引言

随着互联网逐渐进入人们的日常生活中,网络安全已成为影响国家安全的“第五疆域”。身份认证是保障网络安全的主要防线,而口令认证是现今最主要的身份认证手段,因其便于部署的特性,在可预见的未来仍将占据主要地位^[1]。因此,口令认证的安全性一直是人们迫切关注的问题。

影响口令认证安全性的主要因素是用户自身的脆弱性^[2-5]。由于用户的大脑能力有限^[6-7],且需要同时管理几十个甚至上百个不同服务的账号^[8],因此在设置口令时无法选

择完全随机且互不相同的字符串,这就给攻击者猜测用户口令提供了可乘之机。一旦口令被攻击者命中,将对用户的资产以及管理的系统造成不可估量的损失。

随着对口令安全的深入研究,口令猜测逐渐摆脱依靠“奇思妙想”的启发式方法,进入了依靠科学化概率模型算法的新阶段^[1],出现了诸如基于PCFG的方法^[9-10]、基于马尔可夫链(Markov-chain)的方法^[11-12]、基于递归神经网络(Recurrent Neural Network)的方法^[13]、基于生成对抗网络(Generative Adversarial Networks)的方法^[14]等。其中,PCFG算法^[9]和Markov算法^[11]是当前主流的口令猜测算法,是其他概率算

到稿日期:2020-09-09 返修日期:2020-10-08 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:国家自然科学基金(61971473);安徽省自然科学基金项目(1908085QF291)

This work was supported by the National Natural Science Foundation of China(61971473) and Anhui Provincial Natural Science Foundation(1908085QF291).

通信作者:张旻(zhangmindy@nudt.edu.cn)

法的基础,且 PCFG 算法对真实口令数据的拟合效果被证明优于 Markov 算法^[15]。近年来出现了多次大规模个人信息泄露事件,引起了口令安全界的重视^[16],进而出现了利用用户个人信息(Personal Information,PI)进行定向口令猜测的方法,如包含口令重用的方法^[17-18]、包含用户 PI 语义的方法^[18-19]等。相比以往的口令猜测方法,这些定向口令猜测方法在有限次数的命中率上有了明显的提高。其中,TarGuess-I^[18]模型在 PCFG 算法的基础上进行了改进,可以识别包含用户 PI 语义的口令,并经过实验验证了其猜测效果明显优于同类模型^[18]。

本文从攻击者的角度对海量真实的用户口令数据进行研究,通过对口令数据进行统计分析、基于 PCFG 算法分析和基于 TarGuess-I 模型分析,发现用户普遍存在的口令脆弱行为,并总结了易受攻击者猜测命中的口令特征,进而提醒用户避免设置脆弱的口令,提高口令本身抗猜测的强度,并为后续设计口令强度评估方法提供了依据。

2 口令猜测方法

为了研究攻击者如何猜测目标用户的口令,学术界出现了各种各样的口令猜测算法和模型,本节介绍了 PCFG 口令猜测算法以及基于 PCFG 算法设计的 TarGuess-I 定向口令猜测模型。

2.1 PCFG 口令猜测算法

Weir 等于 2009 年首次提出了基于 PCFG 的口令猜测算法^[9]。该算法被证明对口令数据有良好的拟合效果^[15],其核心假设是口令中各字符类型的字段是相互独立的,其定义了一个上下文无关文法模型 $G=(V,\Sigma,S,R)$,其中: V 是一个有限非终结符集合; Σ 是一个有限终结符集合; S 是起始符集合且 $S \in V$; R 是形如 $\alpha \rightarrow \beta$ 规则的有限集合,其中 $\alpha \in V$ 并且 $\beta \in V \cup \Sigma$ 。在 V 集合中,除了 S 是以口令结构为起始符的集合,模型还包含了字母集合 L_n 、数字集合 D_n 和字符集合 S_n ,而 n 表示字段的字符长度。

PCFG 算法包含两个阶段,即训练阶段和猜测阶段,运行流程如图 1 所示。

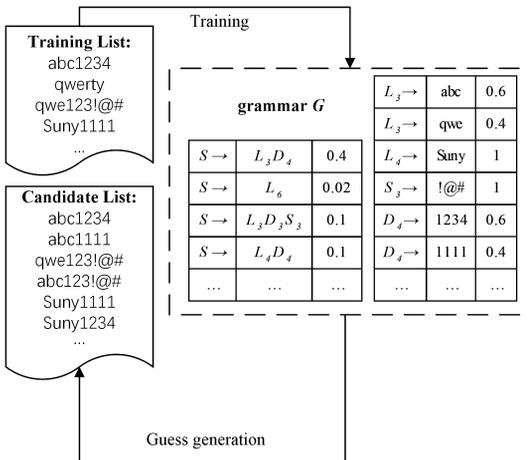


图 1 基于 PCFG 的口令猜测算法的运行流程示例

Fig. 1 Illustration of PCFG-based password guessing algorithm

在训练阶段,该算法将每条口令数据按照字符类别切分成各类字段元素,例如“asd123!”被切分成 $L_3 \rightarrow asd, D_3 \rightarrow 123$ 和 $S_1 \rightarrow !$,进而产生语法结构 $S \rightarrow L_3 D_3 S_1$ 。然后,该算法统计并计算各个集合中元素的概率(即出现频率),每个列表的所有元素概率总和为 1,最终生成上下文无关文法模型 G 。而在猜测阶段,首先依据模型生成候选口令集并计算口令的概率,例如“asd123!”的概率表示为:

$$P(asd123!) = P(S \rightarrow L_3 D_3 S_1) \times P(L_3 \rightarrow asd) \times P(D_3 \rightarrow 123) \times P(S_1 \rightarrow !) \quad (1)$$

然后,将候选口令依照概率由大到小排序,最终生成候选口令列表。

2.2 TarGuess-I 口令猜测模型

TarGuess-I 是由 Wang 等于 2016 年提出的对用户 PI 敏感的 PCFG 口令猜测模型,并被证明其猜测效果优于同类的口令猜测模型^[18]。Wang 等首次提出基于规则的 PI 识别方法,在 PCFG 算法的模型中,除了 L, D, S 这 3 个字符类型标签之外还加入了 6 类 PI 标签,包括 N_n 名字、 U_n 用户名、 B_n 生日、 T_n 电话、 I_n 身份证、 E_n 邮箱。对于每一个 PI 标签,其子集分类与 LDS 标签的分类方式不同, n 代表这个类型 PI 的一种生成规则,详细的 PI 标签生成规则如图 2 所示。

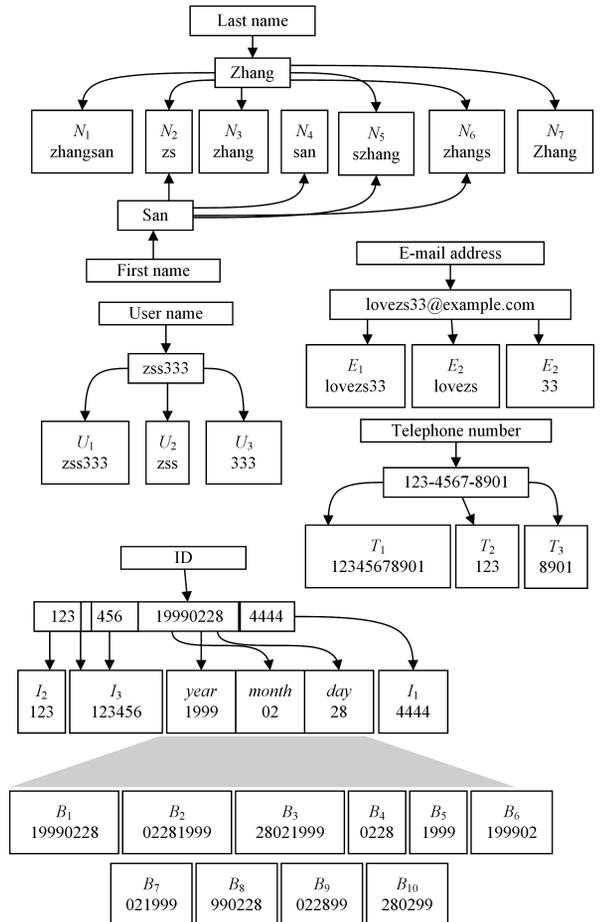


图 2 TarGuess-I 的 PI 标签生成规则

Fig. 2 Rules of PI tag generations of TarGuess-I

TarGuess-I 模型的运行流程示例如图 3 所示。其总体流

程与 PCFG 口令猜测算法相似,但在训练阶段和猜测阶段都多了相应的生成用户 PI 标签元素以及匹配与替换的步骤。

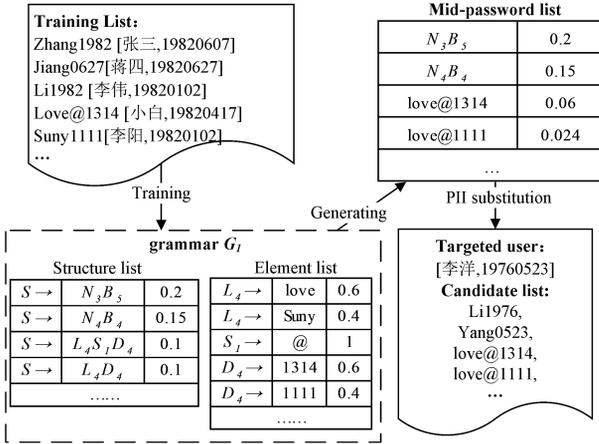


图3 TarGuess-I 模型的运行流程

Fig. 3 Illustration of TarGuess-I model

3 真实数据的口令脆弱性分析

本节使用网络公开泄露的海量真实的用户数据对口令的脆弱性进行分析。本文主要针对国内中文语种的用户,获得

了包含 12306 等超过 50 个国内网站公开泄露的用户数据集,总数据量达亿级。本文依照网站类型对数据源进行大致分类,并列出了 20 个较为典型的口令数据源,详情如表 1 所列。这些数据集一部分由攻击者破解网站后公开,另一部分由网站管理者泄露。尽管这些数据是网络公开的,其中有些数据集甚至被业内广泛用来进行口令安全的研究^[18-21],但它们是敏感的数据,为了解决泄露数据源特征带来的安全问题,本文将后续分析特征的数据来源信息抹去,以编号代替。

表1 典型的网站数据来源

Table 1 Typical website data sources

| 类型 | 来源 | 数据量 |
|------|-------------------------------|-----------|
| 生活购物 | 12306、淘宝、京东、嘟嘟牛 | 27565970 |
| 社交论坛 | YY 语音、微博、珍爱网、人人网、猫扑、天涯论坛、CSDN | 87903530 |
| 通信资讯 | 网易、126 邮箱、163 邮箱、新浪 | 81088662 |
| 视频娱乐 | 盛大、7K7K、暴雪、优酷、爱拍 | 150108490 |

由于缺少包含用户 PI 的数据,我们利用包含用户 PI 的 12306 数据集中用户个人独特的信息字符串(如邮箱),对其他数据集进行匹配,获得了包含用户个人信息和口令的数据共 5070175 条,匹配数据量大小排名前十的来源详细信息如表 2 所列。

表2 数据量排名前十的详细信息

Table 2 Details of top-10 data

| 排名 | 来源 | 类型 | 年份 | 数据量 | 包含 | 不重复口令 | PI 关联 |
|----|-------|----|------|----------|--------------|---------|---------|
| 1 | 12306 | 服务 | 2014 | 1512899 | 邮箱、用户名、PI、口令 | 1001839 | 1145169 |
| 2 | 优酷 | 视频 | 2016 | 92547261 | 邮箱、口令 | 453948 | 552631 |
| 3 | YY | 社交 | 2011 | 32755800 | 邮箱、用户名、口令 | 260571 | 291384 |
| 4 | 盛大 | 游戏 | 2011 | 15313334 | 邮箱、用户名、口令 | 191092 | 223843 |
| 5 | 淘宝 | 购物 | 2016 | 7340364 | 邮箱、口令 | 183724 | 193750 |
| 6 | 太平洋电脑 | 论坛 | 2013 | 5443694 | 邮箱、用户名、口令 | 143889 | 165049 |
| 7 | 嘟嘟牛 | 商务 | 2011 | 16114381 | 邮箱、用户名、口令 | 134429 | 161496 |
| 8 | 天涯论坛 | 论坛 | 2011 | 31006590 | 邮箱、用户名、口令 | 136438 | 158224 |
| 9 | 爱拍 | 视频 | 2016 | 7682232 | 邮箱、用户名、口令 | 129155 | 134327 |
| 10 | 17173 | 游戏 | 2011 | 9480949 | 邮箱、用户名、口令 | 104931 | 122431 |

3.1 口令的特征统计分析

本节对口令的一般特征进行了统计分析,并得到了用户的口令偏好性选择,然后对这些偏好性选择导致的口令脆弱性进行了分析。

3.1.1 口令的长度

口令的长度是口令数据分布最明显的特征。当数据来源的网站具有口令设置策略时,口令的长度分布特征主要受网站策略影响;而当来源的网站没有特定的口令设置策略时,该特征主要受网站的服务类型及重要程度的影响。

本文对 10 个典型数据集中口令的长度进行了分析,结果如图 4 所示。由图 4 可知,相比其他数据集,论坛 5 的口令长度分布较为特殊,长度小于 6 的口令占 3.67%,长度大于 14 的口令占 2.79%,其中长度为 7 的口令占比最高;而游戏 6 中长度为 10 的口令占比最高,其长度为 6,7,8 的口令则相对较少。总体来看,数据分布中长度小于 6 以及大于 14 的口令占很小的一部分,而长度在 8~10 之间的口令普遍占比较高(超过 20%),部分网站数据的口令长度分布的中心有所偏

移,但范围大多集中在 8~10 之间。当攻击者获取目标网站的口令长度特征时,将猜测口令的长度限制在指定范围内可以大大缩小命中所需的猜测空间。

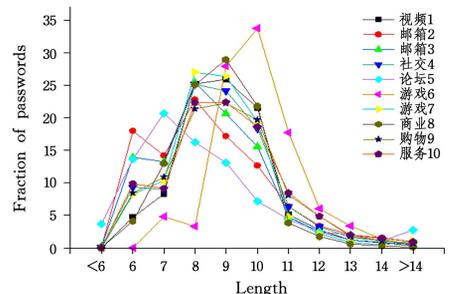


图4 10 个典型数据的口令长度分布

Fig. 4 Distribution of password length of 10 typical data

3.1.2 口令的字符组成

口令的字符组成是口令分布的另一明显特征。本文对 10 个典型的口令数据进行口令字符组成分析,结果如图 5 所示。图 5 显示了 10 个典型的数据源的口令字符组成分布,本

文将口令包含的字符分为数字 D 、大写字母 U 、小写字母 L 和符号 S 等 4 类字符类型,并将字符组成类型依照在数据中的占比从大到小排列, others 包含了占比几乎可以忽略不计的字符组成类型。从图中可以看到,国内用户的口令字符组成类型主要集中在“ DL ”组合,其次则是纯数字口令和纯字母口令,而其余的字符组成类型则普遍占比很低(低于 2%)。商务 7 的口令分布与其他来源数据较为不同,口令中“ DL ”组合以及纯数字口令仍为主要类型,同时其余的字符组成类型也占有一定的比例(大于 2%)且分布较为平均。

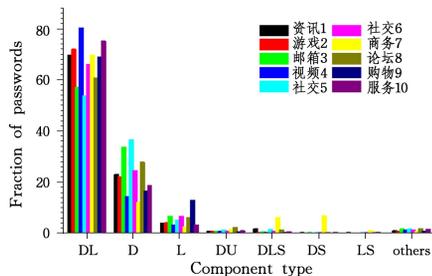


图 5 10 个典型数据的口令字符组成分布

Fig. 5 Distribution of password character component types in 10 typical data

图 5 中的口令字符组成分布体现了中国用户设置口令

的偏好,即习惯使用“ DL ”片段的组合,或者单纯使用数字来设置口令。从数据分布可以推测出,设置纯字母口令在某种情况下可能比设置“ DL ”组合的口令更安全。如果攻击者利用数据驱动的概率猜测算法(如 PCFG 算法),由于数据分布中“ DL ”组合占比最高,则首先输出的应是“ DL ”组合的口令,因此纯字母口令反而以较后的排序位置被输出。

3.1.3 流行口令

本文按口令在数据中出现的次数由大到小排序,获得了流行口令 top-10 列表,结果如表 3 所列,表中最后一行统计了 top-10 的口令在数据集中的占比。表 3 中,有 0.47%~4.11% 的用户账户简单利用 top-10 的口令就可猜测成功。同时,可以发现国内用户偏好于顺序数字(如“123456”和“123321”)、简单的片段组合(如“qq123456”和“a123456”)、重复的字符串(如“123123”和“111111”)以及键盘模式(如“1qaz2wsx”和“1q2w3e4r”)等口令模式。特别是,流行口令中还存在与爱情相关的特殊语义口令(如“5201314”和“woaini1314”)。另外,表中还出现了一些找不到意义的固定字符串(如“7758521”和“aptx4869”),猜测可能是虚假账户,或者可能是由网站文化或网站名称衍生的特殊语义口令。由于这类口令只在几个数据集中出现,无法进一步统计分析,因此不做赘述。

表 3 10 个典型数据的流行口令 top-10 排名

Table 3 Top-10 popular passwords of 10 typical data

| 排名 | 服务 1 | 游戏 2 | 视频 3 | 商务 4 | 论坛 5 | 游戏 6 | 购物 7 | 论坛 8 | 视频 9 | 社交 10 |
|----------------------|------------|------------|------------|------------|-----------|------------|------------|-----------|------------|------------|
| 1 | 123456 | 123456 | 1qaz2wsx | 123456 | 123456 | 123456 | 123456 | 123456 | 123456 | 123456 |
| 2 | a123456 | 111111 | qq123456 | 123456 | 111111 | 111111 | 111111 | 111111 | a123456 | 111111 |
| 3 | 123456a | 000000 | 1q2w3e4r | a123456 | 000000 | 123456789 | 000000 | 000000 | 123456789 | 123456789 |
| 4 | qq123456 | a123456 | 5201314 | woaini1314 | 123123 | 123123 | 123123 | 123456789 | 111111 | 123123 |
| 5 | woaini1314 | 123456789 | woaini1314 | 5201314 | 123456789 | 5201314 | 123456789 | 5201314 | 123456789a | 5201314 |
| 6 | 111111 | 123123 | 123456aa | 111111 | a123456 | a123456 | a123456 | 123123 | 5201314 | 000000 |
| 7 | 5201314 | 5201314 | aptx4869 | 123456789 | 5201314 | 000000 | 5201314 | 7758521 | 000000 | a123456 |
| 8 | 1qaz2wsx | 123456a | 123456789a | 123456a | 7758521 | qq123456 | qq123456 | 123321 | 123123 | 1qaz2wsx |
| 9 | 123123 | woaini1314 | 7758521 | 111111 | 1qaz2wsx | 1qaz2wsx | woaini1314 | 12345678 | qq123456 | 12345678 |
| 10 | 1q2w3e4r | qq123456 | aa123456 | 000000 | 1314520 | woaini1314 | 1qaz2wsx | 1qaz2wsx | woaini1314 | woaini1314 |
| top-10 流行口令在数据中的占比/% | 1.06 | 2.49 | 0.47 | 1.29 | 2.59 | 1.98 | 1.44 | 3.26 | 4.11% | 2.20% |

3.2 基于 PCFG 口令猜测算法的分析

本文利用 PCFG 口令猜测算法分析口令数据中的语法结构,结果如表 4 所列。表 4 所列为 10 个典型数据的口令语法结构 top-10 排名,以及这些语法结构排名在总口令数据中的占比。可以发现,排名前十的口令语法结构在总体数据中的占比近乎达到了 50%。

为了进一步挖掘口令数据中用户对语法结构的偏好性,本文定义了 3 类语法结构类型:纯字符类型、简易结构类型以及复杂结构类型。其中,纯字符类型的语法结构表示为单纯由一类字符类型组成的口令(如 D_6 和 D_7 由纯数字组成),可以发现,表中排名第一的口令语法结构均为纯字符类型。而简易结构类型表示口令类型为 $L_m D_n$ 的语法结构,即由 m 位 L 字母字段与 n 位 D 数字字段前后组合而成,语法结构中每

个字符类型的字段仅出现一次。

从表 4 中可观察到,除去部分纯字符类型的语法结构,其余的语法结构均为简易结构类型。然而,口令数据中还存在另一类语法结构,与简易口令的特征相反,即复杂结构类型。在该语法结构中,有一个或多个字符类型的字段出现了两次及两次以上,且各字符类型的字段交叉出现。例如“1qaz2wsx”被分析为 $D_1 L_3 D_1 L_3$,其中 D_1 字段和 L_3 字段交叉出现了两次。

我们统计了纯字符类型、简易结构以及复杂结构的口令语法结构类型的分布,如表 5 所列。由表 5 可以看出,口令数据中大部分为纯字符类型口令和简易结构口令,复杂结构口令也占了 4.75%~10.54%,尤其是视频 3 中复杂结构口令占了 10.54%。口令在 PCFG 算法中依照字符类型被切分成

多个字段,且每个字段都包含在上下文无关文法模型 G 中, 并依照在训练数据中出现的频率进行了排序。

表 4 10 个典型数据基于 PCFG 口令猜测算法分析的口令语法结构 top-10 排名

Table 4 Top-10 password semantic structures of 10 typical data analyzed by PCFG-based password guessing algorithm

| 排名 | 服务 1 | 游戏 2 | 视频 3 | 商业 4 | 论坛 5 | 游戏 6 | 购物 7 | 论坛 8 | 视频 9 | 社交 10 |
|------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 1 | D_6 | D_7 | D_7 | D_7 | D_6 | D_7 | D_7 | D_6 | D_6 | D_6 |
| 2 | $L_3 D_6$ | D_6 | $L_3 D_6$ | $L_3 D_6$ | D_7 | D_6 | D_6 | D_7 | D_7 | D_8 |
| 3 | D_7 | D_8 | $L_2 D_8$ | $L_2 D_7$ | D_8 | D_8 | D_8 | D_8 | D_8 | D_7 |
| 4 | $L_2 D_6$ | $L_2 D_7$ | $L_2 D_7$ | $L_1 D_7$ | $L_3 D_6$ | $L_3 D_6$ | $L_2 D_7$ | $L_3 D_6$ | $L_3 D_6$ | $L_3 D_6$ |
| 5 | D_8 | $L_3 D_6$ | D_8 | D_6 | D_9 | $L_2 D_6$ | $L_3 D_6$ | D_9 | $L_2 D_6$ | $L_2 D_6$ |
| 6 | $L_2 D_7$ | $L_1 D_7$ | $L_3 D_7$ | $L_2 D_6$ | $L_2 D_6$ | $L_2 D_7$ | $L_1 D_7$ | $L_2 D_6$ | $L_2 D_7$ | D_9 |
| 7 | $L_3 D_7$ | $L_2 D_6$ | $L_2 D_8$ | $L_3 D_7$ | $L_3 D_7$ | $L_1 D_7$ | $L_1 D_9$ | D_{10} | D_9 | $L_2 D_7$ |
| 8 | $L_4 D_4$ | $L_3 D_7$ | $L_4 D_4$ | D_8 | $L_4 D_4$ | $L_3 D_7$ | $L_3 D_7$ | $L_4 D_4$ | $L_4 D_4$ | $L_3 D_7$ |
| 9 | $L_2 D_8$ | $L_2 D_8$ | $L_1 D_7$ | $L_2 D_8$ | $L_2 D_7$ | $L_4 D_4$ | D_6 | $L_3 D_7$ | $L_1 D_7$ | $L_1 D_7$ |
| 10 | $L_1 D_7$ | D_9 | D_6 | $L_1 D_8$ | L_8 | $L_2 D_8$ | $L_2 D_6$ | $L_2 D_7$ | $L_3 D_7$ | $L_4 D_4$ |
| top-10 口令语法结构在数据中的占比/% | 38.06 | 45.11 | 37.74 | 42.52 | 46.51 | 43.36 | 32.67 | 49.57 | 43.55 | 41.20 |

表 5 口令数据的语法结构类型分布

Table 5 Distribution of password syntax structures

(单位: %)

| 来源 | 包含 K | 简易结构 | 纯字符 | 复杂结构 |
|-------|--------|-------|-------|-------|
| 服务 1 | 1.86 | 71.42 | 22.44 | 6.14 |
| 游戏 2 | 1.76 | 65.09 | 30.16 | 4.75 |
| 视频 3 | 1.94 | 71.51 | 17.95 | 10.54 |
| 商务 4 | 1.59 | 72.97 | 22.72 | 4.31 |
| 论坛 5 | 1.49 | 50.56 | 43.02 | 6.42 |
| 游戏 6 | 1.85 | 67.79 | 26.89 | 5.33 |
| 购物 7 | 1.55 | 64.83 | 29.94 | 5.24 |
| 论坛 8 | 1.26 | 47.13 | 46.52 | 6.36 |
| 视频 9 | 1.75 | 59.21 | 34.44 | 6.36 |
| 社交 10 | 1.61 | 62.53 | 31.65 | 5.82 |

对于简单结构的口令,由于其每一个字符类型的字段只出现一次,因此依照上述流程,PCFG 算法能够很好地将该类口令的真实分布体现出来。然而,对于复杂口令,由于口令中同一个字符类型的字段出现两次以上,因此 PCFG 算法会产生出不同于训练数据的新口令。例如,对于服务 1,训练生成的模型中包含:

$$\begin{cases} P(S \rightarrow D_1 L_3 D_1 L_3) = 0.002 \\ P(D_1 \rightarrow 1) = 0.11 \\ P(D_1 \rightarrow 2) = 0.08 \\ P(L_3 \rightarrow \text{qaz}) = 0.12 \\ P(L_3 \rightarrow \text{wsx}) = 0.11 \end{cases} \quad (2)$$

表 6 10 个典型数据基于 TarGuess-I 口令猜测模型分析的口令语法结构排名 top-10

Table 6 Top-10 password syntax structures of 10 typical data analyzed by TarGuess-I model

| 排名 | 服务 1 | 游戏 2 | 视频 3 | 商业 4 | 论坛 5 | 游戏 6 | 购物 7 | 论坛 8 | 视频 9 | 社交 10 |
|----------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 1 | D_6 | D_7 | D_7 | E_1 | D_6 | D_7 | U_1 | D_6 | D_6 | D_6 |
| 2 | D_7 | D_6 | $N_2 D_6$ | D_7 | D_7 | D_6 | E_1 | D_7 | D_7 | E_1 |
| 3 | $N_2 D_6$ | D_8 | $N_2 D_7$ | D_6 | D_8 | D_8 | D_7 | D_8 | D_8 | D_7 |
| 4 | U_1 | U_1 | D_8 | $L_1 D_7$ | U_1 | $N_2 D_6$ | D_6 | D_{10} | U_1 | D_8 |
| 5 | D_8 | E_1 | $L_1 D_7$ | $N_2 D_7$ | E_1 | $L_1 D_7$ | D_8 | B_1 | $N_2 D_6$ | U_1 |
| 6 | $N_2 D_7$ | $L_1 D_7$ | D_6 | $N_2 D_6$ | U_3 | $N_2 D_7$ | $N_2 D_7$ | E_1 | U_3 | $N_2 D_6$ |
| 7 | E_1 | $N_2 D_7$ | $L_2 D_6$ | D_8 | D_9 | E_1 | N_1 | D_9 | E_1 | $N_2 D_7$ |
| 8 | $L_2 D_6$ | $N_2 D_6$ | $L_2 D_7$ | $L_1 D_8$ | B_1 | U_1 | $L_1 D_7$ | B_8 | D_9 | D_9 |
| 9 | $N_1 D_3$ | U_3 | E_1 | $L_2 D_7$ | B_8 | $L_2 D_7$ | $N_2 D_6$ | $N_2 D_6$ | B_1 | $L_1 D_7$ |
| 10 | U_3 | D_9 | $N_1 D_3$ | $U_2 D_7$ | $N_2 D_6$ | $L_2 D_6$ | $N_1 D_3$ | $N_2 D_7$ | $L_2 D_6$ | B_1 |
| TOP-10 口令语法结构 在数据中的占比/% | 21.27 | 26.84 | 20.14 | 24.55 | 32.11 | 24.43 | 36.03 | 33.98 | 26.94 | 27.27 |

如表 6 所列,在 PCFG 文法中加入 PI 标签后,虽然纯数字的口令语法结构仍出现在排名中,但是排名前十的其他类

则对应生成的口令中有:

$$\begin{cases} P(1\text{qaz}1\text{qaz}) = 3.4848 \times 10^{-7} \\ P(1\text{qaz}2\text{wsx}) = 1.9008 \times 10^{-7} \end{cases} \quad (3)$$

可以发现,口令“1qaz1qaz”并没有出现在服务 1 的 top-10 流行口令中,但却比排名第八的流行口令“1qaz2wsx”更先被模型输出。从这一角度来看,使用复杂结构口令能降低被 PCFG 口令猜测算法中的可能性。

然而,正如“1qaz2wsx”可以明显看出是键盘模式口令,在复杂结构口令中仍有相当一部分口令可以被特定的口令模式识别。本文通过在 PCFG 算法的模型中加入键盘模式识别标签 K 来分析口令数据,结果如表 5 第二列所列。键盘模式口令在整体口令数据中占 1.49%~1.94%,尤其在视频 3 中键盘模式口令占据了 1.94%。由此可见,基于模式生成复杂结构的口令可能无法帮助用户降低被攻击者猜测命中的风险,尤其是基于流行的口令生成模式。

3.3 基于 TarGuess-I 口令猜测模型的分析

本文利用 TarGuess-I 口令猜测模型对 10 个典型数据进行分析,获取了包含 PI 标签的语法结构排名 top-10 以及这些语法结构在整体口令数据中的占比,结果如表 6 所列。表 6 中,用户 PI 标签为姓名 N 、电话 T 、邮箱 E 、用户名 U 、身份证号 I 和生日 B ,其尾部的数字为生成规则的类型,具体的生成规则如图 2 所示。

型口令语法结构都包含了用户 PI 标签。通过进一步分析发现,原本被识别为纯数字语法结构的口令中部分被识别为包含用户生日 B 、身份证号 I 以及电话 T ,而包含对应 PII 标签的口令结构在总体口令数据中的占比较小,并没有在 top-10 中出现。

本文在分析口令数据中包含 PI 标签的语法结构分布的基础上,进一步统计了各类用户 PI 在口令数据中的占比,结果如图 6 所示。从图 6 中可以看出,10 个典型数据中超过 60% 的用户口令包含 PI 标签。另外,从分布中可知国人选择 PII 构造口令的偏好性,即:姓名>用户名>生日>邮箱地址>身份证号~电话号码。如表 7 所列, N_2 标签(姓名简写)在口令数据中出现的频次最高, N_3 标签(姓的全拼)和 N_1 标签(名的全拼)也排在前列,生日 B 的各类生成规则也被频繁使用。值得注意的是,邮箱 E 和用户名 U 的各类生成规则也出现在排名中。邮箱和用户名等信息是由用户自主生成的字符串,上述现象说明用户也偏好于利用自主生成的字符串设置口令。然而,由于这些字符串出现在公开的个人信息公开中,可被攻击者发现并利用,因此包含这些字符串的口令也是不安全的。

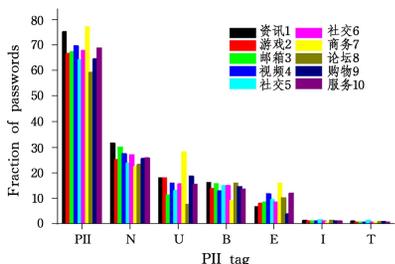


图 6 用户 PII 在口令数据中的占比

Fig. 6 Proportion of user PII in password data

表 7 10 个典型数据中口令包含的 PII 生成规则 top-10

Table 7 Top-10 PII generation rules in passwords of 10 typical data

| 排名 | 服务 1 | 游戏 2 | 视频 3 | 商业 4 | 论坛 5 | 游戏 6 | 购物 7 | 论坛 8 | 视频 9 | 社交 10 |
|----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 1 | N_2 | N_2 | N_2 | N_2 | N_2 | N_2 | U_1 | N_2 | N_2 | N_2 |
| 2 | U_3 | U_3 | N_3 | U_3 | U_3 | U_3 | E_1 | E_3 | U_3 | U_3 |
| 3 | N_3 | U_2 | U_3 | U_2 | E_3 | N_3 | N_2 | B_5 | U_2 | E_1 |
| 4 | N_1 | N_3 | B_5 | N_3 | N_1 | U_2 | U_3 | U_3 | N_3 | N_3 |
| 5 | U_2 | B_5 | N_1 | E_1 | N_3 | B_5 | U_2 | N_3 | N_1 | N_1 |
| 6 | B_5 | N_1 | E_3 | B_5 | B_5 | N_1 | N_1 | N_1 | B_5 | U_2 |
| 7 | B_1 | B_1 | U_2 | N_1 | B_1 | B_1 | N_3 | B_1 | B_1 | B_5 |
| 8 | B_8 | E_3 | B_1 | E_3 | B_8 | E_3 | B_5 | B_8 | B_8 | E_3 |
| 9 | E_3 | E_1 | B_8 | B_1 | E_1 | E_1 | E_3 | E_1 | U_1 | B_1 |
| 10 | E_1 | U_1 | E_1 | B_8 | U_1 | B_8 | B_1 | B_4 | E_1 | U_1 |

4 口令脆弱行为的特征

本节对第 3 节分析发现的用户口令数据中存在的脆弱行为进行总结。

4.1 口令的偏好性设置

通过对口令的统计分析发现,若网站没有对口令进行规范,用户设置的口令则普遍集中在 8~10 位的字符长度,利用数字和小写字母组合设置口令。这类口令偏好设置可被攻击者用于优化其口令猜测模型,从而大大缩小命中口令所需的猜测范围。攻击者可利用 PCFG 口令猜测算法学习数据中用户的口令设置偏好,包括字符组成和字符长度等口令的一般特征,以及口令的语法结构。若 PCFG 算法的模型中只

包含字符类型的标签,则对于具有复杂结构的口令没有良好的猜测效果,从这一角度出发,用户设置复杂结构的口令更能避免被攻击者攻破。

然而,除上述一般特征之外,国内用户还存在诸如顺序字符串、重复字符串、键盘模式等口令模式的设置偏好。若这些口令属于简单的语法结构类型,则 PCFG 算法能够很好地将这些口令的真实分布体现出来。然而,对于具有复杂的语法结构类型的口令,如键盘模式口令“1q2w3e4r”,使用只包含字符类型标签的 PCFG 模型进行分析训练则会生成无效的猜测口令。尽管如此,攻击者仍可以通过加入对应的模式识别标签对猜测模型进行优化。因此,也应避免设置常用的模式口令。

同时,包含爱情等特殊语义的口令也被国内用户广泛使用,语义中相关的词句如“woaini”“520”以及“1314”等在用户口令中频繁出现。这类口令也容易被包含相应语义标签的 PCFG 口令猜测模型猜测成功。

4.2 口令包含个人信息

基于对 PI 语义敏感的 TarGuess-I 口令猜测模型对口令进行语义分析,发现国内有超过 70% 的用户存在利用自己的 PI 构造口令的行为。这些行为是可被理解的,因为使用个人信息可帮助用户记忆和管理口令。然而,普遍使用相同类型的 PI,依照相同的规则将 PI 转换成口令的组成元素,容易被 PI 语义敏感的 TarGuess-I 口令猜测模型攻破。口令中出现的个人信息频率最高的类型为姓名,国内用户习惯将名字简写作为口令的组成元素。意外的是,用户名中包含的数字在口令中出现的频率高于其他类型的个人信息生成规则。用户可能对某些含有特殊意义的数字有偏爱,但这类数字同时出现在用户名与邮箱地址等公开的信息和口令中也是不安全的行为。

结束语 本文通过对海量用户口令数据进行统计分析、基于 PCFG 口令猜测算法的语法结构分析和基于 TarGuess-I 定向口令猜测模型的 PI 语义分析,发现了国内用户设置口令时存在的脆弱行为,诸如简单结构偏好、模式口令偏好、语义偏好以及口令包含的个人信息偏好等,并对这些口令脆弱行为进行了总结,为避免设置易受攻击者猜测命中的弱口令和提高口令安全性提供了依据。然而这仅仅是对用户口令脆弱行为的分析,我们将以此为基础,进一步开展针对口令脆弱性评估的分析研究。

参考文献

[1] WANG P, WANG D, HUANG X. Advances in Password Security [J]. Journal of Computer Research and Development, 2016, 53(10): 2173-2188.

[2] ADAMS A, SASSE M A. Users are not the enemy [J]. Communications of the ACM, 1999, 42(12): 40-46.

[3] YAMPOLSKIY R V. Analyzing user password selection behavior for reduction of password space [C] // Proceedings 40th Annual 2006 International Carnahan Conference on Security Technology, 2006: 109-115.

[4] WANG D, WANG P, HE D, et al. Birthday, name and bifacial security: understanding passwords of chinese web users [C] //

- 28th USENIX Security Symposium (USENIX Security 19). 2019;1537-1555.
- [5] LIU G, QIU W, MENG K, et al. Password Vulnerability assessment and recovery based on rules mined from large-scale real data[J]. Chinese Journal of Computers, 2016, 39(3): 454-467.
- [6] BEAUTEMENT A, SASSE M A, WONHAM M. The compliance budget: managing security behaviour in organisations[C]// Proceedings of the 2008 New Security Paradigms Workshop. 2008;47-58.
- [7] NITHYANAND R, JOHNSON R. The password allocation problem: Strategies for reusing passwords effectively[C]// Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society. 2013;255-260.
- [8] FLORENCIO D, HERLEY C. A large-scale study of web password habits[C]// Proceedings of the 16th international conference on World Wide Web. 2007;657-666.
- [9] WEIR M, AGGARWAL S, DE MEDEIROS B, et al. Password cracking using probabilistic context-free grammars[C]// 2009 30th IEEE Symposium on Security and Privacy. 2009;391-405.
- [10] VERAS R, COLLINS C, THORPE J. On Semantic Patterns of Passwords and their Security Impact[C]// NDSS. 2014.
- [11] MA J, YANG W, LUO M, et al. A study of probabilistic password models[C]// 2014 IEEE Symposium on Security and Privacy. 2014;689-704.
- [12] NARAYANAN A, SHMATIKOV V. Fast dictionary attacks on passwords using time-space tradeoff[C]// Proceedings of the 12th ACM Conference on Computer and Communications Security. 2005;364-372.
- [13] MELICHER W, UR B, SEGRETI S M, et al. Fast, lean, and accurate: Modeling password guessability using neural networks[C]// 25th USENIX Security Symposium (USENIX Security 16). 2016;175-191.
- [14] HITAJ B, GASTI P, ATENIESE G, et al. Passgan: A deep learning approach for password guessing[C]// International Conference on Applied Cryptography and Network Security. 2019;217-237.
- [15] WANG D, HE D, CHENG H, et al. fuzzyPSM: A new password strength meter using fuzzy probabilistic context-free grammars[C]// 2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). 2016;595-606.
- [16] FLORÊNCIO D, HERLEY C, VAN OORSCHOT P C. An administrator's guide to internet password research[C]// 28th Large Installation System Administration Conference (LISA14). 2014;44-61.
- [17] DAS A, BONNEAU J, CAESAR M, et al. The tangled web of password reuse[C]// NDSS. 2014;23-26.
- [18] WANG D, ZHANG Z, WANG P, et al. Targeted online password guessing: An underestimated threat[C]// Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. 2016;1242-1254.
- [19] LI Y, WANG H, SUN K. A study of personal information in human-chosen passwords and its security implications[C]// IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications. 2016;1-9.
- [20] LI Z, HAN W, XU W. A large-scale empirical analysis of chinese web passwords[C]// 23rd USENIX Security Symposium (USENIX Security 14). 2014;559-574.
- [21] WANG D, CHENG H, WANG P, et al. Zipf's law in passwords [J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2776-2791.



XIE Zhi-jie, born in 1995, postgraduate, is a member of China Computer Federation. His main research interests include password security.



ZHANG Min, born in 1966, Ph.D, professor, Ph.D supervisor. His main research interests include communication network security and intelligent computing.