TIFS'17 - Zipf's Law in Passwords

概述：用户口令的实际分布一直是一个开放问题，本文深入探讨这个问题，基于自然语言处理和计算统计技术，通过14个大型的真实数据集提出了两个Zipf-like model用于描述口令的分布，其中PDF-Zipf可以描述流行口令，而CDF-Zipf可适用于整个口令数据集。接下来根据口令分布，提出评估口令数据集强度的方法，并且简述了Zipf model的应用场景及将会影响的领域。最后对未来工作进行了展望.

优点：①解决了安全领域的一个基本问题：用户生成的口令是如何分布的。

② 提出的观点（及参数选择）都进行了详尽的实验论证和分析.

③ 对该领域的背景、现状和未来工作都进行了讨论（很详实的Conclusion）

问题：① 本文可进一步分析不同数据集的特征对最终分布的影响，如对于CDF-Zipf Model，C'和S'与数据集的特征的关系.

② 第IV节的B部分，文章设计了从完美Zipf分布中取样的实验，共考虑了3组可能影响拟合的实验，包括Zipf分布、样本大小和LF，应简要分析这样选取实验参数的原因.

③ 相对于PDF-Zipf和Clauset'09的方案，在拟合数据集时，CDF-Zipf的时间复杂度较大，在实际测试中时间（14.67 hours）远超另外两种（32.40 seconds和69.39 seconds），应简要讨论原因。由于CDF-Zipf在KS统计量D和口令集覆盖范方面表现较优，应简要讨论提升拟合速度的方法.

▲口令策略会影响用户口令的选择，若设计较好，口令策略可以在保证usability的情况下改善口令安全. ▲→ 口令策略设计者应如何评估策略强度？管理员应如何选择口令策略？（可用性）

→由于口令分布（一个网站内）可能随时间产生变化，安全管理员应评估口令分布的强度，并根据此强度调整自己的策略.

→如何精确评测基于某策略的口令分布的强度？引出另一基础问题：用户生成的性服从什么分布函数？

！关于口令分布的不合实际的假设通常会引起严重的安全性和可用性问题.

研究口令分布的重大意义（其中的两个）.

① Schechter等提出了基于流行度的口令生成策略，当一个口令超过给定的流行度阈值T则被拒绝，可有效抵抗漫步猜测攻击. 但T若无法合适选择，可能影响安全性和可用性。而当得知用户选择的口令的分布，可更合理地选取阈值.

② 在衡量基于口令的认证方案抗离线字典空间的能力时，主要考虑口令搜索空间大小，并根据单次猜测成本来衡量敌手的成本-收益。 ↳取决于用户选择口令方式

▲ 口令分布的强度评测可以依据单一攻击方法.

▲口令数据集：14个大规模真实口令集，来自不同服务类型和规模的网站，用户的地理位置、语言信仰和文化背景也有不同。另外，对于不同数据集中的口令长度、字符构成等进行了统计，并分析了产生这些特征的可能原因.

▲两个统计学技术

①线性回归：拟合可能存在线性关系的观测数据，求解两个变量之间的关系（常用最小二乘法）

决定系数 $R^2 \in [0,1]$ 衡量回归线与实际数据点拟合优劣的一个评价指标：R^2 越接近1越好.

② Kolmogorov-Smirnov检验：针对离散数据的最流行的非参数检验方法，通过假设检验测量样本和理论分布模型之间的"距离"：→实际数据的累积分布函数CDF $F_n(x)$ 与理论分布CDF $F(x)$ 的距离：$D = \sup|F_n(x) - F(x)|$. $D \in [0,1]$ 表示两条CDF的最大距离，D越小越好.

→由于口令分布不是固定不变，采用Monte Carlo方法：先使用理论模型拟合给定口令集得到口令分布函数并计算D，接下来使用从实际口令集拟合的分布参数生成合成数据集，使用理论模型拟合合成数据集，计算D'，使用p-value代表D'大于D的比例。若p-value较大，可以表明实际口令数据分布与理论模型无显著差异. (it is safer)

用户口令中的Zipf's law.（Zipf定律最早用来刻画自然语言中单词"排名 vs. 出现频率"的关系，$f_r = \frac{C}{r^s}$）

▲ 去除每个口令集中最不流行的口令（次数少于3次或5次），进行线性回归拟合，⇒ 口令分布服从相似的规律，对于口令集DS一个口令的排名r及频率f_r满足 $f_r = \frac{C}{r^S}$，C和S是依赖DS的常数.

取对数：$\log f_r = \log C - S \cdot \log r$； R^2 很接近于1（与数据集大小及数据泄漏方式有关）

Zipf理论模型模拟人口的相应离散分布函数PDF，故称之为PDF-Zipf模型（注：$S \in [0,1]$）.

在KS检验中，p-value比较小，分析得到假设不一致的结论的原因

① 非流行部分的口令（$f_Y < 1$）占整个口令集的比重为80%~92%，构成长尾部分。

根据大数定律，其在口令集出现的频率不能反应在口令分布中的真实概率。

通过将给定口令 pw_i 的每次观测看作一个服从Bernoulli分布的随机变量，得到相对标准误差：

→ RSE：$\frac{\sigma_{\hat{p}}}{p_{pw_i}} \approx \sqrt{\frac{1}{f_{pw_i}}}$，则当 f_{pw_i} 较大时，真实概率 p_{pw_i} 才可用经验概率 $\frac{f_{pw_i}}{|DS|}$ 拟合。

⇒ 不流行的口令较大程度上会干扰拟合

② 收集到的口令集相对于根据策略产生的口令集仍然太小，由于Zipf分布具有口令概率的多次式递减特性，小样本中低概率事件将压倒高概率事件。

▲进一步证明用产生成的口令样本服从Zipf定律，从完美Zipf理论分布中随机抽取一些样本，观测拟合特性。⇒考虑3组可能影响拟合的变量，Zipf分布3种，样本大小8种，最小频率阈值4F 5种

→共120个实验，表明给定Zipf分布（确定N和S），刚开始更大的F会导致更好的回归，而F进一步增加会导致拟合恶化。

→找到最优拟合，表明样本量越大，更大比例的流行事件将用于拟合。

收集到的口令集远小于样本空间，故应先去除非流行部分的口令。

▲ CDF-Zipf Model（直接对口令的CDF建模）。

$F_Y = C' \cdot Y^{S'}$，F_Y 为排名至 Y 的口令的累积频率，C' 和 S' 是取决于口令集的常数（采用最小割法）

由 F_Y 为阶梯函数：$f_Y = F_Y - F_{Y-1} = C' \cdot Y^{S'} - C' \cdot (Y-1)^{S'}$（若视为连续函数 $f_Y = \frac{d(F_Y)}{dY} = C' \cdot S' \cdot Y^{S'-1}$

KS检验统计量D为0.000670~0.045874（均值0.018457），比PDF-Zipf小。

当给定一个足够大的样本，小且非显著的差异会在统计学中认为是显著的，故p-value较小，而当样本容量较小（10^5）时，多数CDF-Zipf模型拟合可通过KS检验（p-value > 0.01）

→KS统计量D的最大改进量被限制在区间[0, 0.018457]内，CDF-Zipf model可改进空间有限。

通过与Clauset'09的方案对比，在KS统计量D方面：CDF-Zipf和Clauset'09优于PDF-Zipf，对于百万规模以上的口令集，CDF-Zipf通常更好，百万级以下的，Clauset'09表现更好。

③从口令集覆盖范围方面，CDF-Zipf表现最佳，PDF-Zipf次之，Clauset'09较差

另外，给定中等的计算资源，三种模型均可在可接受的时间内完成拟合。

▲数据集涵盖了不同的服务类型、规模、泄漏原因等，表明了Zipf model的广泛适用性

Zipf' law带来的三项启示

1> 基于口令的密码协议：在CDF-Zipf model中，敌手优势刻画为：$Adv_{A,P}(k) = C' \cdot Q(k)^{S'} + \varepsilon(k)$。

2> 口令创建策略：对基于流行度的口令生成策略的阈值做更合理的选择。

$P(rank \le Y) = C' \cdot Y^{S'}$，$P(rank \le N) = C' \cdot N^{S'} = 1 \Rightarrow N = (\frac{1}{C'})^{\frac{1}{S'}}$，$N$ 为独立口令的个数。

前 Y 个独立口令的累积频率：$P(rank \le Y \cdot N) = C' \cdot (Y \cdot N)^{S'} = C' \cdot (Y \cdot (\frac{1}{C'})^{\frac{1}{S'}})^{S'} = C' \cdot Y^{S'} \cdot (\frac{1}{C'}) = Y^{S'}$

若口令频率 f_Y 为连续型变量，$f_Y = \frac{d(F_Y)}{dY} = \frac{d(C' \cdot Y^{S'})}{dY} = C' \cdot S' \cdot Y^{S'-1} \cdot \Rightarrow T = C' \cdot S' \cdot Y^{S'-1}$

则流行度高于T的比例为：$\eta = (\frac{T}{C' \cdot S'})^{\frac{1}{S'-1}} \cdot (C')^{\frac{1}{S'}}$

·潜在受影响的用户比例：$W_p(\eta) = \eta^{S'}$，而真实受影响的用户比例为：$W_a(\eta) = (1-S') \cdot \eta^{S'}$

3> 口令分布强度指标α-guesswork：在两种情形下实际上是非参数化的，在Zipf model下，α-guesswork的应用可大为简化

▲口令数据集的强度指标

对于一个聪明的攻击者来说，给定数据集 χ，使用破解结果 $\lambda_\chi^S(n)$ 作为强度指标：

$\lambda_\chi^S(n) = \sum_{i=1}^{n} p_i(\chi) = \frac{1}{|DS|} \sum_{i=1}^{n} f_i(\chi)$，$|DS|$ 为数据集大小，n 为猜测数。

由 $p_i(\chi) = \frac{f_i(\chi)}{|DS|} = F_i(\chi) - F_{i-1}(\chi) \approx C' \cdot i^{S'} - C'(i-1)^{S'}$。

⇒ $\lambda_\chi^S(n) = \sum_{i=1}^{n} p_i(\chi) = F_n(\chi) \approx C' \cdot n^{S'} = \lambda_\chi(n)$

通过真实数据集作图，$\lambda_\chi^S(n)$ 和 $\lambda_\chi(n)$ 的图像大多重叠，通过与其它方案对比，基于CDF-Zipf的强度指标 $\lambda_\chi(n)$ 最好。