Datacon 口令安全WriteUP - 数据收集与整理

StarCrossedLovers战队 复旦大学 陈诗宇

一、确定搜集范围

1、赛题论文指向的数据集

根据题目提供的赛题信息说明，指向以下论文

> 赛题创意来源：*Yunkai Zou, Maoxiang An, and Ding Wang. "Password guessing using large language models." 34th USENIX Security Symposium (USENIX Security 25). 2025.*
>
> *Wang, Ding, et al. "Targeted online password guessing: An underestimated threat." Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. 2016.*
>
> *Ma, Jerry, et al. "A study of probabilistic password models." 2014 IEEE Symposium on Security and Privacy. IEEE, 2014.*

经阅读整理后，总结出论文中做研究预测工作所用的主要数据集及其运用，整理为以下表格：

数据集名称(英文)	简介	数据集大小	权威性	泄露时间	被引用论文	数据集用途

数据集名称(英文)	简介	数据集大小	权威性	泄露时间	被引用论文	数据集用途
MySpace	美国社交网站 MySpace 的密码泄漏样本，在一次重大数据泄露中被公布。研究中使用的数据包含约3.7万条独特明文密码。	约 3.7 万条密码	真实泄漏数据；MySpace 是早期知名社交平台，其泄漏数据常用于密码模型的泛化测试。	2006 年	*PassGPT: Password Modeling and (Guided) Generation with LLMs* (arXiv 2023)	用作模型泛化能力评估的测试集（不参与模型训练）。
PhpBB	开源论坛 PhpBB 的用户密码泄漏。2009年1月，黑客从 phpBB.com 泄露的 MD5 哈希中破解出约25万条用户密码。	约25万条密码	真实泄漏数据（通过破解哈希获得明文）；常被用作小规模密码数据集来对比模型效果。	2009 年1 月	*A Study of Probabilistic Password Models* (IEEE S&P 2014)； *PassGPT: Password Modeling...* (2023)	用于密码模型性能评估和对比。在近期LLM模型中作为小型测试集检验模型泛化能力。
RockYou	社交游戏平台 RockYou 的大型密码泄漏数据集。泄漏包含约3.26亿账户，其中约3260万条独特明文密码已公开。	约 3.26 千万条密码	真实泄漏密码；这是密码研究中最著名的大规模数据集之一，几乎所有密码猜测相关研究都会使用该数据。	2009 年12 月	*A Study of Probabilistic Password Models* (2014)；*Targeted Online Password Guessing...* (CCS 2016)；*Password Guessing Using LLMs* (USENIX 2025)；*PassGPT... with LLMs* (arXiv 2023)	广泛用于密码概率模型和猜测工具的训练与测试。例如，作为基准语料库训练 Markov模型、PCFG模型和最新的大模型，并用于评估新模型的猜测成功率。

数据集名称(英文)	简介	数据集大小	权威性	泄露时间	被引用论文	数据集用途
Hotmail	微软 Hotmail 电子邮件服务的密码泄漏数据。2009年前后，有黑客将数以万计的 Hotmail 帐号凭据发布在网上（例如 Pastebin）。研究数据集中包含约0.89万条独立的账户密码。	约 0.9 万条密码	真实泄漏密码；可能来源于网络钓鱼所得的帐户列表，小规模但典型，常用于验证模型在不同数据集上的表现。	约 2009 年（确切时间不详）	*PassGPT: Password Modeling...* (2023)	用作小规模测试集，评估密码猜测模型对不同数据来源的适应能力。
Rootkit	黑客论坛 Rootkit 的用户密码数据。2011年论坛数据库泄露时密码以 MD5 哈希存储，研究者使用高级猜测技术将其中约97%破解为明文供研究使用。总计约6.94万条密码及部分用户个人信息字段被获取。	约 6.9 万条密码	真实泄漏密码；包含用户邮箱等PII字段，是少数带有用户个人信息的密码数据集之一。TarGuess 等研究中将其破解后用于针对性密码攻击实验。	2011 年2 月	*Targeted Online Password Guessing...* (CCS 2016)；*Password Guessing Using LLMs* (2025)	用于针对特定用户的猜测场景研究。例如结合泄漏的邮箱、用户名等PII评估定向攻击效果。在通用模型训练中由于规模较小，更多用作辅助测试集。

数据集名称(英文)	简介	数据集大小	权威性	泄露时间	被引用论文	数据集用途
CSDN	中国知名开发者社区网站CSDN的用户密码泄漏。泄漏包含约642万条软件开发者账户密码明文。	约642万条密码	真实泄漏密码；这是2011年底中国大规模泄漏事件之一，因覆盖技术社区用户而备受关注。该数据集已被广泛用于密码安全研究，代表中文用户密码习惯。	2011年12月	*A Study of Probabilistic Password Models* (2014)；*Targeted Online Password Guessing...* (2016)；*Password Guessing Using LLMs* (2025)	用于大型密码模型的训练和测试。例如，用来比较中英文用户密码强度分布，训练概率猜测模型，以及在最新大模型中作为训练语料提升对中文密码的泛化能力。
Dodonew（嘟嘟牛）	中国付费网络游戏平台“嘟嘟牛”的用户密码泄漏。由于涉及货币交易，用户密码相对复杂。泄漏公开了约1626万条该站用户密码。	约1626万条密码	真实泄漏密码；2011年底多家中国网站集中泄漏事件之一。因规模大且包含游戏用户群，该数据集被多篇研究采用。密码包含数字和字母混合等更强模式，被认为安全性高于平均水平。	2011年12月	*A Study of Probabilistic Password Models* (2014)；*Targeted Online Password Guessing...* (2016)；*Password Guessing Using LLMs* (2025)	用于大规模密码猜测模型的训练与评估。在定向猜测研究中，该数据集匹配了用户邮箱等PII，用于评估利用游戏站点密码和用户信息进行跨站猜测的效果。

数据集名称(英文)	简介	数据集大小	权威性	泄露时间	被引用论文	数据集用途
126	中国网易 126 免费邮箱的用户密码泄漏。泄漏发生于2011年底，公开了约639万条邮箱账户密码明文。	约639万条密码	真实泄漏密码；作为2011年著名泄漏事件的一部分，该数据集代表了中文邮箱用户常用密码分布，在学术研究中也经常被采用。	2011年12月	*Targeted Online Password Guessing...* (2016)；*Password Guessing Using LLMs* (2025)	用于训练和评估中文密码猜测模型。例如，LLM模型研究中包含该数据以验证模型对中文邮件站点密码的适应性。
178	中国某大型游戏门户网站“178”的用户密码泄漏。该泄漏公布了约1000万条游戏用户账户的明文密码。	约1000万条密码	真实泄漏密码；与CSDN、嘟嘟牛等同为2011年12月的大规模泄漏之一，反映了中文游戏玩家的密码选择模式。虽然未如CSDN般广为人知，但在密码研究中用于对比不同站点间用户行为。	2011年12月	*A Study of Probabilistic Password Models* (2014)	用于研究不同用户群体（如游戏 vs. 技术社区）的密码强度差异。在模型评估中充当中文密码集，用以验证模型对不同类型网站密码的适用性。

数据集名称(英文)	简介	数据集大小	权威性	泄露时间	被引用论文	数据集用途
Yahoo	雅虎 (Yahoo) 网站用户密码泄漏。2012年黑客组织 D33Ds 公布了约45万条雅虎用户密码明文。	约44万条密码	真实泄漏密码；因为雅虎用户群规模大且国际化，该数据常被用作小型测试集来对比不同模型对英文弱密码的覆盖率。雅虎泄漏密码中近70%为独特值，密码复杂度相对更高。	2012年7月	*A Study of Probabilistic Password Models* (2014)；*Targeted Online Password Guessing...* (2016)	用于模型评估中的对比实验。例如，与RockYou等数据相比，雅虎数据帮助衡量模型对不那么常见密码的猜中率，并用于验证猜测算法在不同网站数据上的表现。
LinkedIn	职业社交网站 LinkedIn 的用户密码泄漏。2012年该网站发生密码泄露（约1.65亿哈希密码，2016年明文破解公布）。研究通常使用其中约6050万条已破解明文密码。	约6050万条密码	真实泄漏密码；作为全球大型社交平台泄漏数据，LinkedIn 密码集在近年来被广泛用于训练先进猜测模型，以提升对不同用户群密码的泛化能力。	2012年（完整数据2016年曝光）	*PassGPT: Password Modeling...* (2023)	常用作密码猜测模型的训练集之一。例如，在GAN和Transformer等模型研究中，LinkedIn数据与RockYou一起用于训练，以评估模型在大规模真实泄漏语料上的表现和对新泄漏数据的适应性。

数据集名称(英文)	简介	数据集大小	权威性	泄露时间	被引用论文	数据集用途
Xiaomi	中国科技公司小米的用户账户密码数据。2014年泄漏的信息中，密码以加盐哈希形式保存，未直接公开明文。研究人员将其用作真实用户目标账户，以测试密码猜测算法的有效性。	约828万条账户记录	真实泄漏数据；由于密码未以明文形式泄出，在研究中主要用于模拟真实环境而非训练语料。其存在代表了一种更严峻的攻击场景：攻击者只有哈希需尝试破解。	2014年5月	*Targeted Online Password Guessing...* (2016)	未用于模型训练；在研究中作为“真实目标”账户来验证在线猜测模型的实战效果。例如，TarGuess 将小米泄漏数据的哈希作为攻击目标，以评估模型在真实未破解密码场景下的成功率。
12306	中国铁路订票官网 12306 的用户密码泄漏。该数据集包含约12.93万条明文密码，并附带用户名、身份证号、手机号等详细个人身份信息。	约12.9万条密码	真实泄漏密码；因包含丰富的用户PII，这一数据集被用于研究密码与个人信息的关联，是定向密码猜测研究的经典数据源之一。	2014年12月	*Targeted Online Password Guessing...* (2016)；*Password Guessing Using LLMs* (2025)	用于评估利用用户个人信息进行密码猜测的效果。例如，将12306泄漏中的密码与同一用户在其它站点的密码进行匹配，测试模型在已知部分个人信息条件下的猜测成功率；同时作为训练数据，使模型学习包含个人信息的密码模式。

数据集名称(英文)	简介	数据集大小	权威性	泄露时间	被引用论文	数据集用途
Xato	Xato 密码合集，由安全研究员 Mark Burnett 于2015年公开。该数据集综合整理了多个泄漏来源的常见密码，总计约1000万条，作为通用密码词典使用。	约1000万条密码	合成数据集；非单一站点泄漏，而是从多种来源收集最常见密码汇总而成。由于覆盖面广，这个列表常被当作攻击字典和基线模型，在学术研究中用于与特定模型生成的候选密码对比。	2015年2月（数据集发布）	*Targeted Online Password Guessing...* (2016)	作为密码猜测的词典基线，用于对比模型生成效果。例如，在TarGuess研究中，将Xato词典作为对照，通过比较模型猜中密码数量验证新算法的改进程度。
000Webhost	免费虚拟主机提供商000Webhost的用户密码泄漏。2015年约1300万用户账户数据遭泄露，包含约1525万条明文密码。	约1525万条密码	真实泄漏密码；作为大规模英文密码数据集之一，被多个密码猜测模型论文所采用。此外，该数据还衍生出带有强度标签的公开数据集，用于密码强度分类研究。	2015年10月	*Targeted Online Password Guessing...* (2016)；*Password Guessing Using LLMs* (2025)；*ROBENS: A Robust Ensemble System...* (IEEE Access 2025)	用于训练和评估通用英文密码猜测模型。例如，作为深度学习模型训练语料提高其对英文弱密码的覆盖。此外，衍生的Kaggle 密码强度数据集基于此泄漏，用于训练机器学习模型进行密码强度分类。

数据集名称(英文)	简介	数据集大小	权威性	泄露时间	被引用论文	数据集用途
Taobao	中国电商平台淘宝网的用户密码泄漏。2016年初媒体报道称淘宝有大量用户凭证泄漏，研究数据集中包含了约1507万条淘宝账户密码。	约1507万条密码	真实泄漏密码；反映电子商务网站用户的密码选择行为。在学术研究中，这一数据被用于丰富中文密码语料，考察不同网站类型（社交 vs. 电商）用户密码差异。	2016年2月	*Password Guessing Using LLMs* (2025)	用于训练和评估中文密码猜测模型。例如，在大规模LLM猜测模型中加入淘宝数据，以确保模型学习电商领域用户的密码模式，从而提升对该类密码的猜测性能。
COMB	Compilation of Many Breaches (COMB)，由2021年黑客在论坛发布的超大规模密码合集。该合集整合了此前约252个泄漏事件的明文凭据，汇总了约32.8亿条邮箱-密码对。尽管不包含新的泄漏来源，但因规模空前而备受关注。	约32.8亿条密码	综合数据集；由多次真实泄漏合并而成，是迄今最大规模的单一密码数据集之一。其数据覆盖全球大量账户，具有极强代表性，适合作为密码分布研究的整体样本。但由于数据重复和冗余信息，也需要经过清洗后使用。	2021年2月	*Password Guessing Using LLMs* (2025)	用于提供最新、海量的密码语料以训练密码猜测模型。例如，大模型研究中将COMB用作训练集的一部分，以获取更全面的密码分布特征，并在评估中作为测试基线，检验模型在超大规模数据上的性能。

数据集名称(英文)	简介	数据集大小	权威性	泄露时间	被引用论文	数据集用途
Kaggle Password Strength dataset	Kaggle 平台上的“密码强度分类数据集”，由 000Webhost 泄漏密码经过筛选并结合密码强度评级生成。数据集收集了约67万条独特密码，并基于 Twitter、Microsoft 等三种商业密码强度评估工具的一致结果将每个密码标注为弱、中或强三类。	约67万条密码	衍生数据集；源自真实泄漏但非直接泄漏集，其密码经过清洗和强度标签标注。由于整合了多款主流强度评测标准，一致性高，可用于训练机器学习模型进行密码强度预测。在学术研究中，它提供了一个有标注的基准，用以验证分类算法对弱/强密码的识别能力。	2021年3月(数据集发布)	*ROBENS: A Robust Ensemble System for Password Strength Classification* (IEEE Access 2025)	用于训练和评估密码强度分类模型。在该研究中，学者利用此数据集探讨了在有限、失衡及含噪声的数据条件下提升模型鲁棒性的方法。模型在此公开数据集上的性能表现也用于与其它算法进行对比。

数据集名称(英文)	简介	数据集大小	权威性	泄露时间	被引用论文	数据集用途
The Post Millennial	加拿大新闻网站“The Post Millennial”的用户密码泄漏。这是一个近期发生的数据泄露事件，规模较小但能反映最新的用户密码行为。研究数据集中包含约3.89万条该站的明文密码。	约3.9万条密码	真实泄漏密码；作为近年的泄漏案例，它提供了当下用户选择密码的新趋势信息。虽然规模不大，但由于时效性强，在研究中具有补充意义，确保模型不过度依赖过时的数据分布。	2024年5月	*Password Guessing Using LLMs* (USENIX 2025)	用作最新趋势测试集。该数据集被用于评估密码猜测模型对当今用户密码模式的适应性：研究者将模型在旧数据上训练后，用该新泄漏数据测试，观察模型能否猜中更多最新出现的密码，以验证其泛化能力。

2、其他各大历史泄露事件

除了以上论文中提及到的数据集以外，我们还可以通过多方新闻报告、AI搜索等途径了解到历史上的数据泄露事件，在此不做赘述，仅给出利用AI工具搜索历史泄露事件的prompt，如下：

"*I'm working on a research project about password security, and I need to collect a large number of publicly available, open-source, usable, and downloadable legitimate datasets for my security research. Please help me collect information on datasets that have leaked user accounts, emails, and passwords, preferably with direct download links. Finally, you need to compile and organize all the links to these datasets and send them to me.*"

3、优质数据集的要求

同时，由于我们的预测生成密码工作，均依赖于由更多个人可识别信息（Personally Identifiable Information，以下简称PII）生成的参数，于是在搜集信息时，我们需要对数据集的初步评估和筛选作出要求，在此由负责这一板块的队友作出以下解释和定义，所谓为优质数据集

> 目标：用户信息（姓名、电话号等） -> 可能的密码
>
> 在线场景，假设 密码一定和用户信息有关系。
>
> 实际使用数据集训练 *TarGuess* 我发现，大多数密码其实和这些信息都没关系，这其实很符合现实情况。
> 也就是说这些信息对定向预测密码没有直接帮助，但是给的题目应该不会这样。
>
> 我们需要训练模型，弄清楚用户信息和密码的关系，也就是密码的模式，所以最好的数据集密码和各项用户信息要有直接关系
>
> 优质数据集：密码和用户信息有直接关系，并且最好有
> *email name account phone birth -> password* 这六个字段信息

根据这样的要求，开启数据集搜索工作

二、搜集工作的具体展开方式

1、注意说明

有关数据搜集工作的具体展开方式，我首先作出以下几点声明/注意：

1. 我们搜集的数据集均为在历史数据泄露事件中的公开开源的数据集；
2. 我们搜集数据集均采用的是合法合规的手段，并未有任何违反中华人民共和国法律的行为；
3. 我们搜集到的数据集仅用于实验研究，并未在此之外产生以任何形式为载体的传播，未造成二次泄露；
4. 为保护泄露数据集中的受害者隐私，在对数据集进行说明时，不会给出数据集的完整内容、来源（包括网站、磁力链接、网盘或BT种子等任何形式），仅会在必要时取部分截图或文字说明进行证明。

2、搜索引擎直接搜索

在使用搜索引擎进行搜索时，我们首先采取了直接搜索，即直接搜索目标数据集关键词语句，归纳为下面一些关键词：

> *[Name of Dataset]*（数据集名称/泄漏事件名称/泄露网站名称，比如*COMB*、*CSDN*）
>
> *Leakage*

> *Data Breach*
>
> *Dataset*

其中发现，在使用"Data Breach"作为关键词进行搜索时，往往会得到更多更有价值的结果，包括泄露事件报告、数据集搜集网站、账户安全评估网站、相关论坛甚至直接的下载链接。

同时，我们也采取了不同搜索引擎、不同规则的搜索方式，包括：

(1)使用Google搜索规则：形如：

```
"<Name of Dataset> data leak" "magnet:?"
```

成功搜索到了包含在各类网页形式中的数据集磁力链接，据此方法搜索到的数据集包括CSDN泄露数据集以及Rootkit泄露数据集

(2)使用其他搜索引擎：

搜索国内相关泄露事件数据集（比如12306、7k7k、CSDN、嘟嘟牛）时，直接使用百度搜索引擎也能搜索得到相关报告、论坛、博客等信息，但是通常不含有直接的下载链接；

使用Tor浏览器的专属搜索引擎和网络（DuckDuckGo以及Ahmia.fi），搜索效果奇佳，在搜索数据集集合网站、论坛、博客方面表现出色，成功搜索到了包含大量磁力链接的专业数据集网站，后文将提及的名单和网站大部分来源于此搜索。

3、论坛和网站搜索

在国外各大著名论坛和开源网站中，我们也做了搜索，得到大量有效信息：

在知名论坛Reddit上搜索Data Breach的有关内容，可以得到大量讨论帖子，其中涉及到大量专业网站链接的分享，其中部分已失效，但仍有部分提供了有效线索，指向一些专业搜集泄露数据集的网站。

在Github上同样搜索到大量开源数据集项目，比如Seclist等，但里面的数据集仅包含账号密码，并未提供更多信息，价值不大。

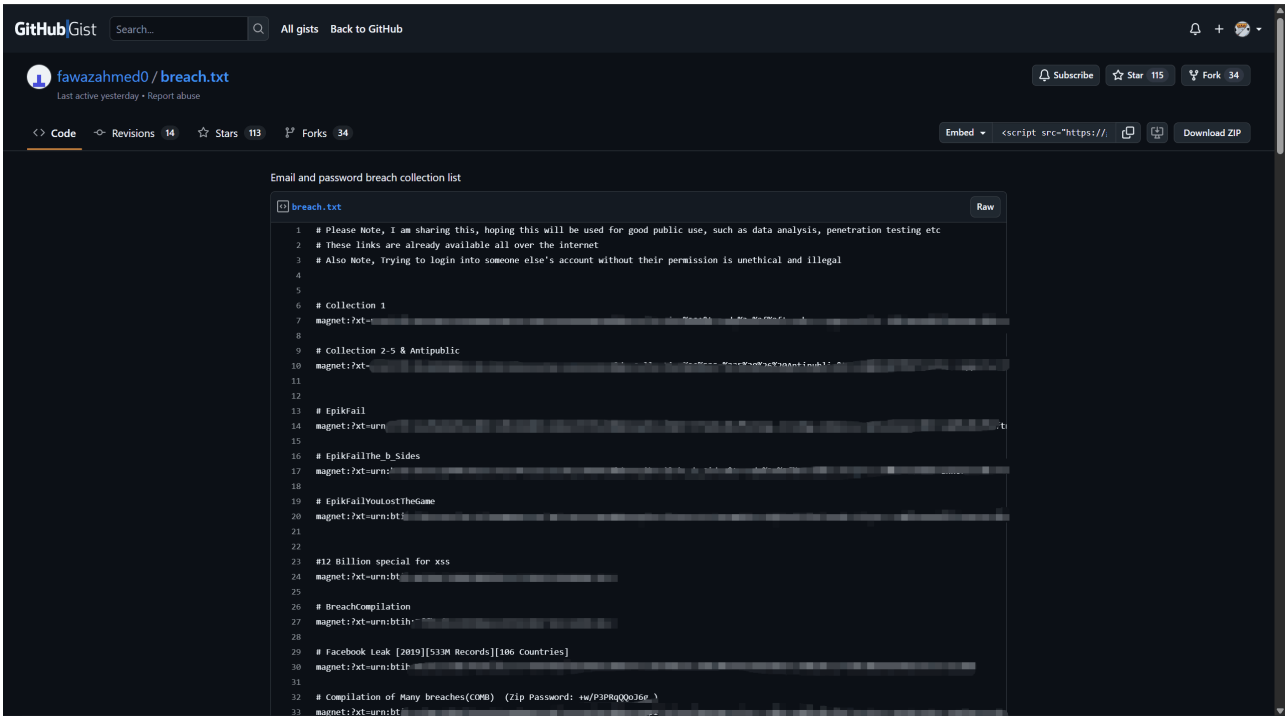
> 实际上，我们并不能轻易将搜索方式简单分为以上几种类型，搜索过程中，通常结果是互相指向而非单独孤立的，尤其是网站和论坛链接。

三、搜索结果

经过以上搜索和初步判断，我们所使用的数据集将从以下几个资源中进行进一步整理和筛选，同时给出资源包含的数据集列表：

1、由Github上某用户整理的磁力链接集合

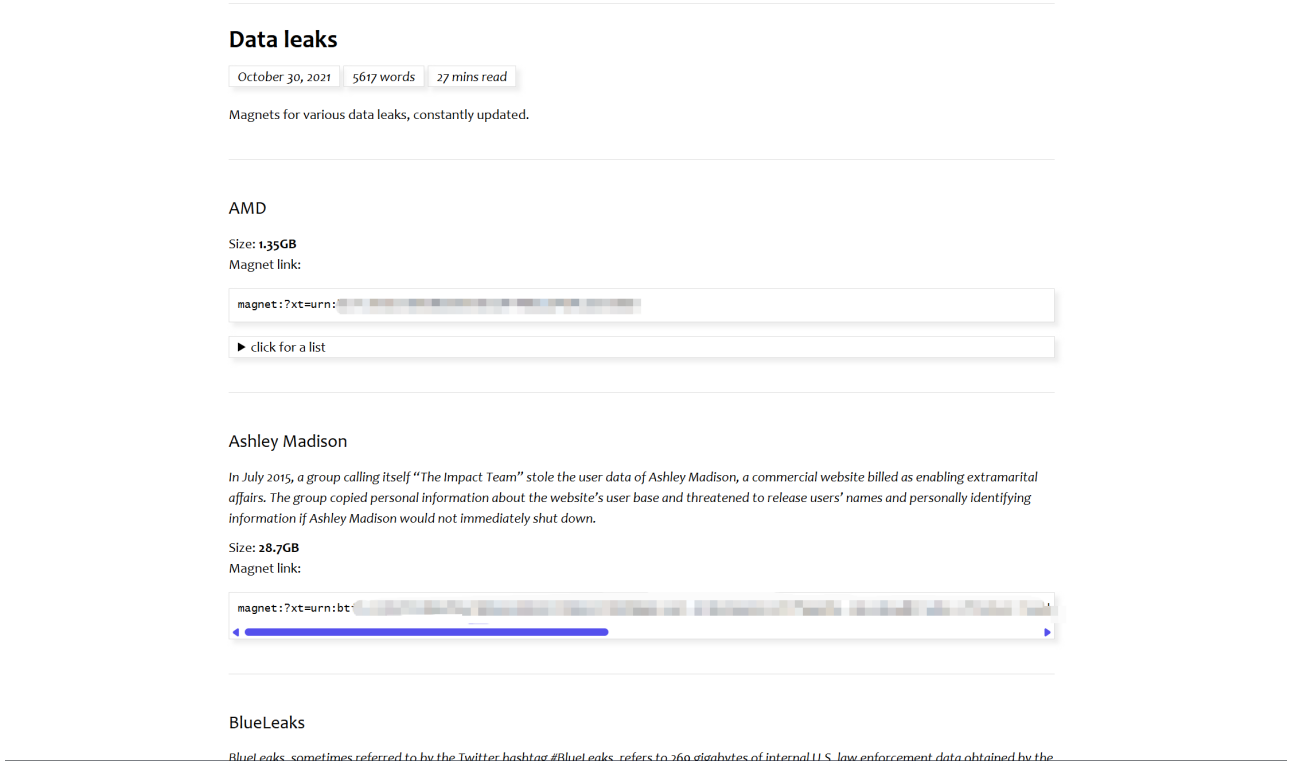
这是一个github上某用户整理的众多数据集链接，挂载到了他的个人Github Gist上，截图如下：



其中包含Collection 1、Collection 2-5 & Antipublic、EpikFail、Compilation of Many breaches(COMB)、Leaked Database Archive.7z等众多泄露数据集整合包，单个整合包中均包含大量数据集。
其中，我们进一步选用了`Leaked Database Archive.7z`进行下载研究，里面包含的数据集名单见`list1.txt`

以及`Collection#3`，该整合包仅包含知名国际交友平台`Fling`的泄露用户数据，在之后用于了我们的研究

2、一个专业搜集泄露数据集的网站



该网站包含了大量由专业黑客群体或个人整理的数据集整合包，从中我们筛选出两份整合包，分别是`list2.txt`和`list3.txt`，详细信息我们不做过多说明

四、数据集整理和筛选

综合上部分我们搜索的结果，我们将从三个list和Collection#3对应的数据整合包中整理和筛选优质数据集，然后用于我们的项目研究。

1、AI初步筛选

通过AI筛选，我们能初步通过list中的文件名搜索对应的网站、企业的相关泄露事件，先初步了解泄露事件的信息，进行第一步的筛选，重点筛选包含明文非加密密码、更多PII的数据集

在此使用的是Claude模型进行初步筛选判断，prompt如下：

> *Carefully examine this list, marking all files that may be related to public leaks (passwords stored in plaintext, leaked user information as comprehensive as possible), along with brief descriptions of the relevant leaks.*

经过AI的初步筛选，我们对这些名单和数据集有了初步认识，同时得到一个初步结论：

有关交友、“约炮”、色情网站的数据泄露集，往往更符合我们的研究需求，属于前文所提及的“优质数据集”需求。

2、人工筛选

在AI进行初步筛选的基础上，囿于硬件设施和技术能力限制，我们将对这些数据整合包进行解压，对里面的文件直接查看、人工筛选出优质数据集。

最终得到有效的优质数据集（文件名或数据集名称）如下：

1. 36k_member.csv
   fling.com_40M_users.sql(由于数据集过大，我们做了切片处理，即pt_00*系列文件)
2. www.naijaloaded.com_Database - INTRAOPS.sql
3. YouPorn.txt
4. mate1.com-plain-november-2015.txt
5. waydate_dump.csv

然后，需要对这些原始文件进行进一步处理，处理为目标可用的数据集格式，比如.csv/.json