**A Simple Guide Line for Cyber Security Dataset Related Paper Review**

Muwei Zheng

**Things We Are Looking For…**

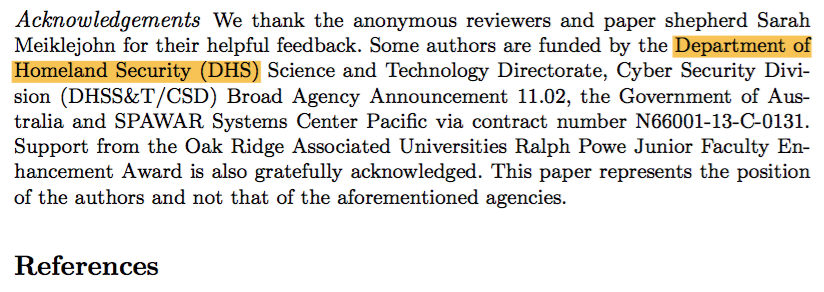
* Whether the research is funded by Department of Homeland Security (DHS) or not.
* Whether there are cyber security related datasets associating with the paper.
* If there are datasets, whether they are public or not.

**Places Where It Might Show up…**

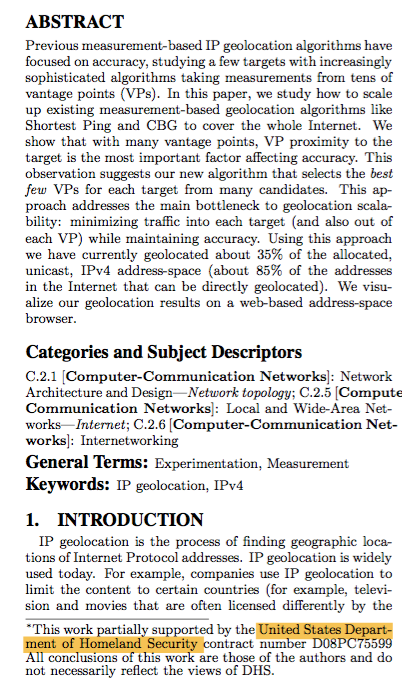
For DHS funding:

* Check **Acknowledgement** section. It generally appears at the end of paper and before References part. If there is a funding associating with this paper, the funding organization will be acknowledged at this part. The Acknowledgement section may also be mentioned at the end of Abstract part or just after Abstract.

Example:



(Before References. Paper: The Bitcoin Brain Drain: A Short Paper on the Use and Abuse of Bitcoin Brain Wallets)

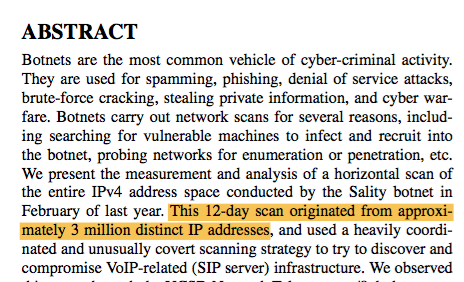
’

(At the first page after Abstract. Paper: Towards Geolocation of Millions of IP Addresses)

For datasets:

* First of all, read the **Abstract** section to get an overall idea what the paper is talking about. Sometimes it will mention whether they are using datasets in the paper. If there is dataset mentioned, GO TO FIND THE DETAIL DESCRIPTION OF THE DATASET WITHIN THE PAPER.

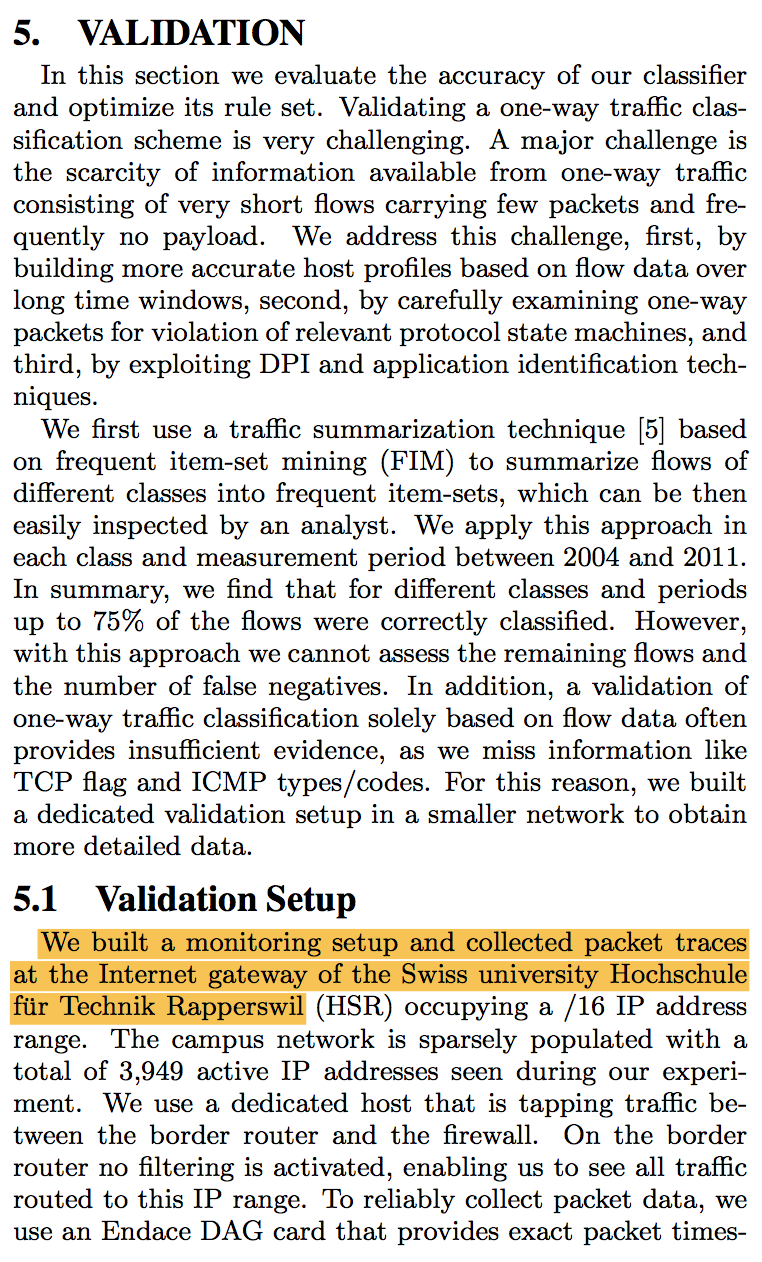
Example:



(Paper: Analysis of a “/0” Stealth Scan from a Botnet)

* If the paper is talking about implementation of a new algorithm, a new app, a new protocol, or any other fancy stuff, probably you can quickly scan through and pass the theory and implementation part. The suspicious section will be **Validation** or **Evaluation.**

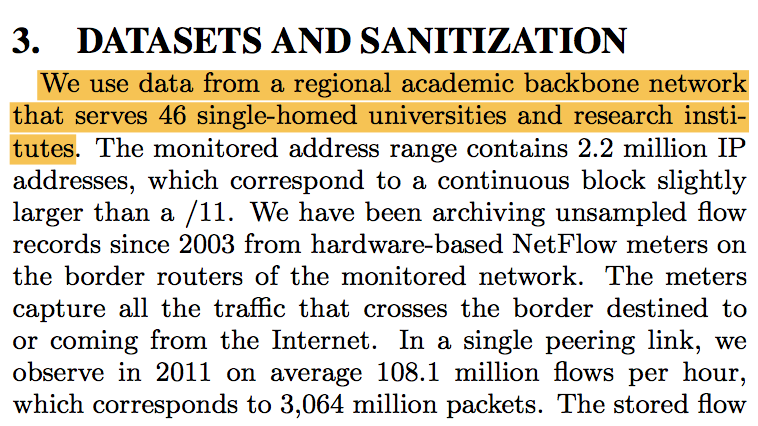
Example:



(Paper: Classifying Internet One-way Traffic)

* There might be a **Dataset** section in the paper, read it carefully. (NOT ALL DATASET RELATED PAPER HAS A DATASET SECTION. UNFORTUNATELY, MANY OF THEM DON’T HAVE ONE.)

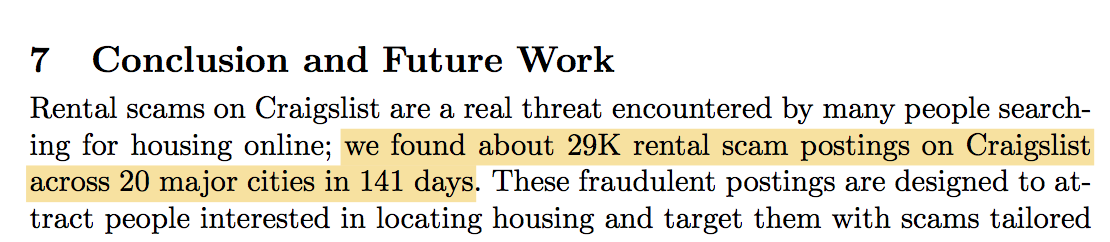
Example:



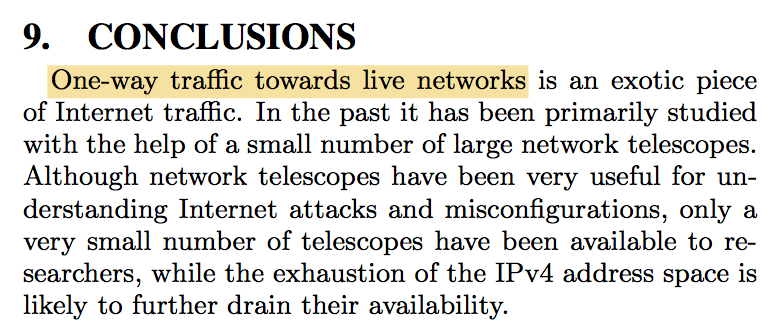
(Paper: Classifying Internet One-way Traffic)

* If the paper is talking about analyzing a large scale dataset, read the **Conclusion** part at the end. *Created Derivative* (type of datasets will be discussed later) type of datasets are more hidden than others, and the **Conclusion** part will give you a hint whether there is one out there.

Example:



(Paper: Understanding Craigslist Rental Scams)



(The paper did not have a direct access to an available one-way traffic dataset, therefore authors must categorize or exacted one-way traffic dataset by themselves. Paper: Classifying Internet One-way Traffic)

* Quickly scan through the **References** section, be alert with reference entries with only a name followed by a URL, they might be an existing type of dataset.

Example:



(These are all datasets. Paper: Understanding Craigslist Rental Scams)

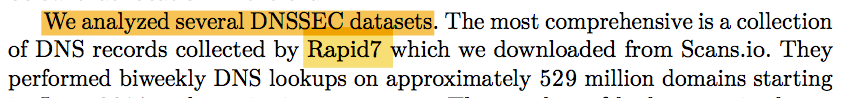
* *ALL OF THESE ARE ONLY THE POTENTIAL PLACES MAY DESCRIBE DATASETS, BUT DATASETS MAY ALSO BE DESCRIBED SOMEWHERE ELSE.*

For public or not:

*We assume that default is non-public.*

* Most of the datasets that have specific names are public available, like Alexa, Rapid7, DB-IP, etc.

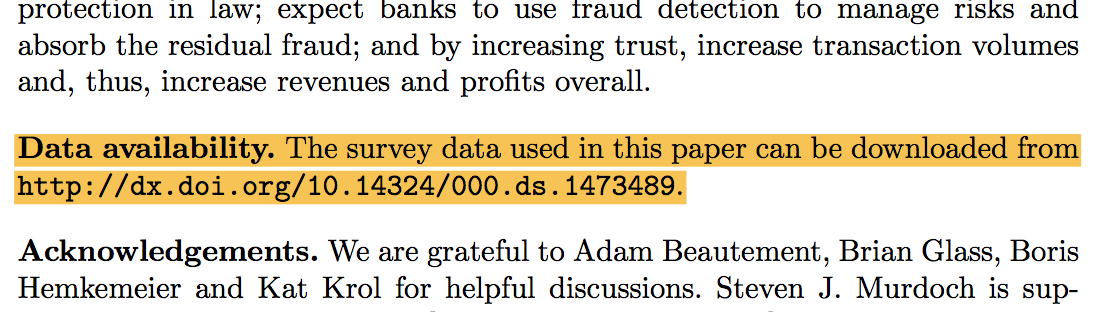
Example:



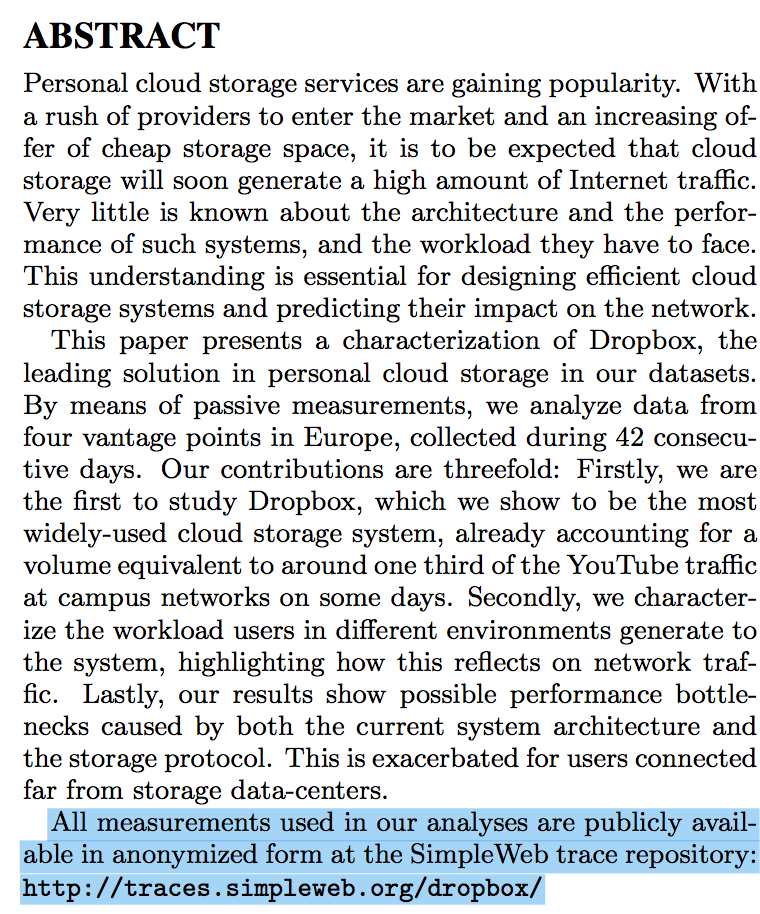
(Paper: Factoring as a Service)

* Some papers will indicate that their datasets are public available. The announcement may appear at different places.

For example:



(Paper: Are Payment Card Contracts Unfair)



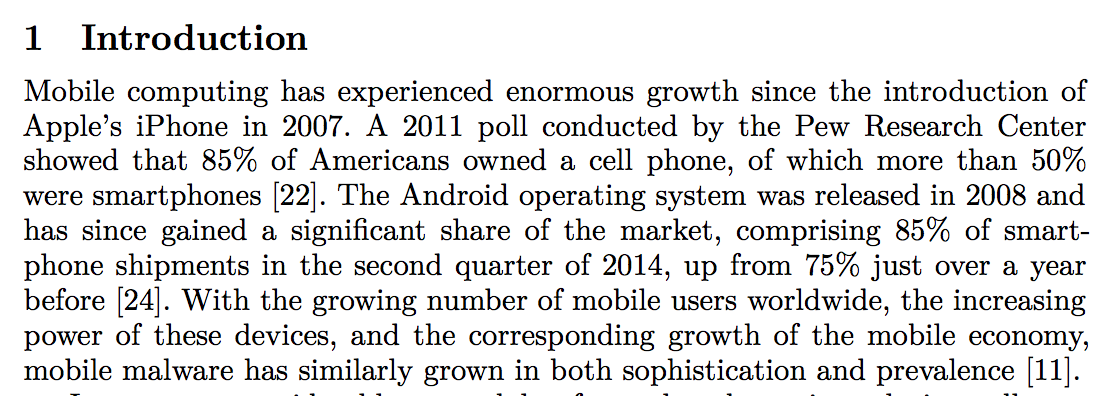
(Paper: Inside Dropbox: Understanding Personal Cloud Storage Services)

* If the dataset is coming from another research paper, probably we need to go back to the original paper and check if the dataset is public available or not.
* At the end of reviewing the paper, search keyword “public” and “availabl” to see if there is any missing capture.

**Things That We Are NOT Looking For…**

* Description of findings of datasets. This happens quite often in Introduction section.

Example:

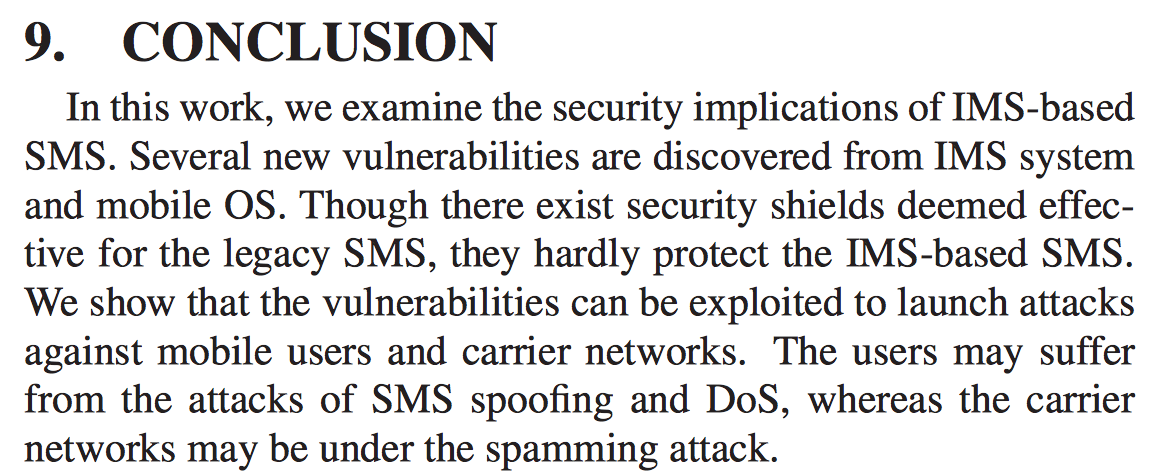


(Paper: CuriousDroid: Automated User Interface Interaction for Android Application Analysis Sandboxes)

In this example, we have information of cell phone market. Of course this information is concluded from some datasets, but authors in this paper only used these conclusions, instead of the datasets themselves, therefore we don’t consider that they are using datasets.

* Datasets should consist of data in a reasonable number of size.

Example:

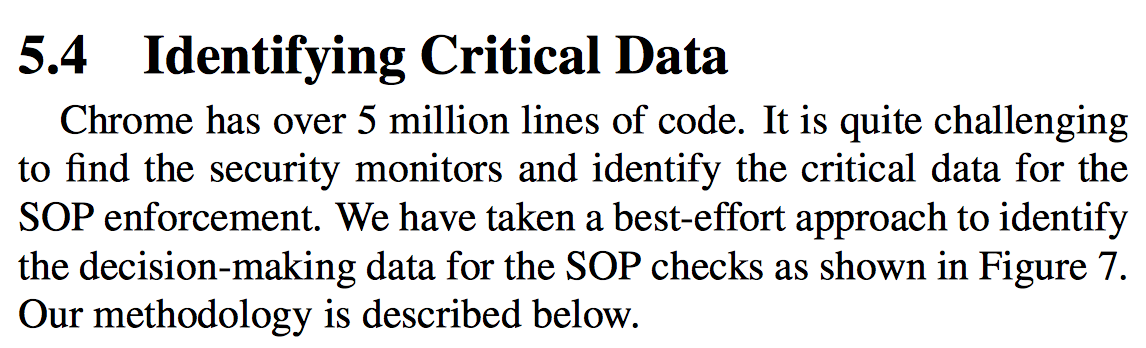


(Paper: New Security Threats Caused by IMS-based SMS Service in 4G LTE Networks)

In this example, there were only 5 vulnerabilities discovered. The number of vulnerabilities is too small to consider it as a dataset.

* Things like XXX lines of code.

Example:



(Paper: The “Web/Local” Boundary Is Fuzzy: A Security Study of Chrome’s Process-based Sandboxing)

This may sound like a dataset, we have keyword “Data”, we have something in large scale “over 5 million lines of code”, but this is not a dataset. Though number of lines of code is huge, but each line of code is not an individual sample, therefore the 5 million lines together is one sample, and one sample doesn’t count as a dataset.

* Experimental datasets.

The experimental datasets can be considered as two types.

* + Type One:

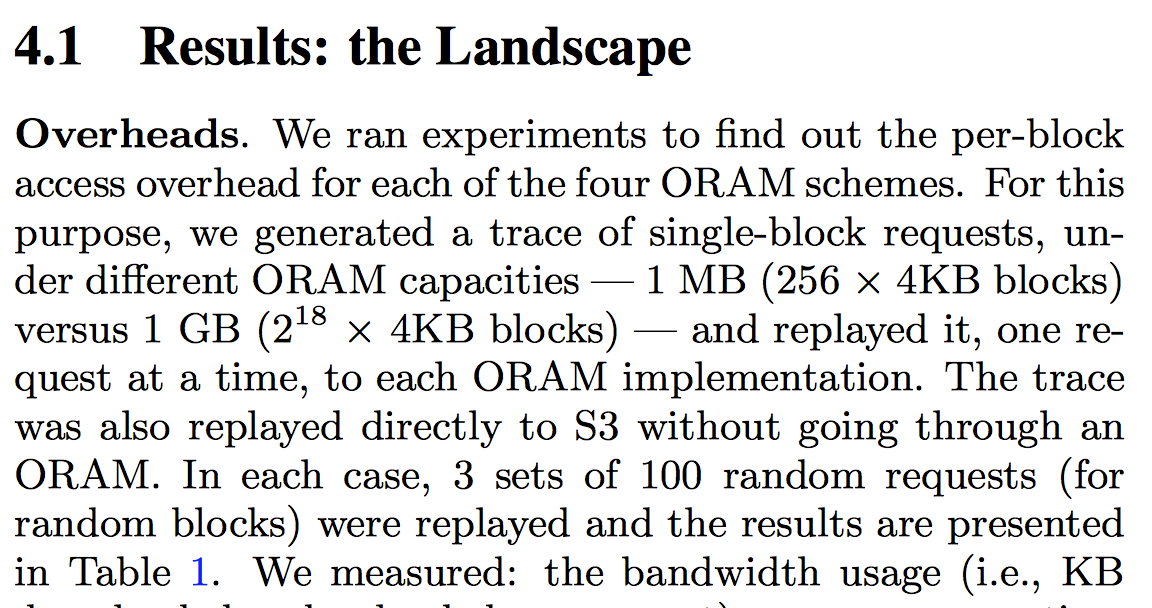
The experiment is set up specially to test their new implementation of models or apps. This kind of dataset are not considered since they are not quite reusable.

(Lack of Sample, since we don’t consider that, I didn’t mark them, I will attach one next time I see it.)

* + Type Two:

The experiment is set up for measuring user’s attitudes, performance of real-world network or applications. These kind of experiment datasets are considered as useful datasets.

Example:



(Paper: Practicing Oblivious Access on Cloud Storage: the Gap, the Fallacy, and the New Way Forward)

**Dataset Origin**

There are three types of dataset origin: Existed, Created Primary, and Created Derivative.

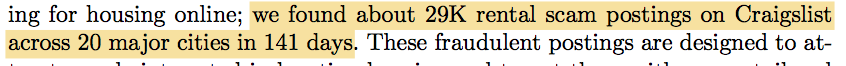
* Existed:

The dataset is already existed BEFORE the research is conducted. A typical example will be any public available dataset downloading online, like Alexa’s Top 1k website, or Google blacklist. Datasets taking from other research papers are also Existed.

* Created Derivative:

The dataset is NOT existed before the research but it is a subset of one or more other existed or not-existed datasets.

Example:



(Paper: Understanding Craigslist Rental Scams)

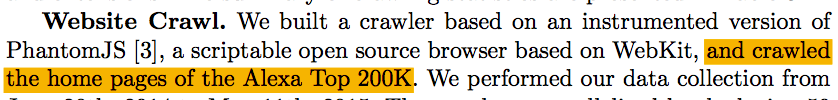
In this example, Craigslist is a dataset contains huge number of ads in different categories. It is a dataset belonging to Population Enumeration under Internet Characteristics. After some study and research, author extracted a dataset consisting of 29K rental scams from this Craigslist dataset. Now the new dataset belongs to Attacks under Attacker Related. In addition, the new dataset is a subset of Craigslist, therefore we consider it as Created Derivative.

(It is not necessary it is a subset of only one dataset, it can consist of subsets from several datasets.)

* Created Primary:

The dataset is NOT existed before the research and it is NOT a subset of other datasets.

Example:



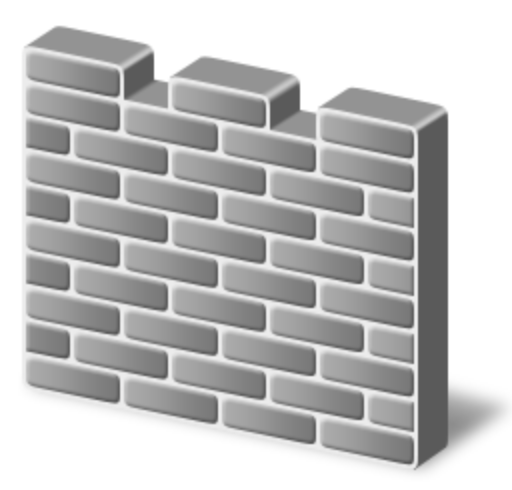
Although we need to use Alexa Top 200K dataset to generate the home pages dataset, the home pages dataset is not a subset of Alexa Top 200K. The dataset is new and it is produced from some work, therefore it is considered as Created Primary.

**Dataset Categorization**

There are four highest level categories, each one associates with one distinct part in cyber security. The four categories are: **Attacker Related**, **Internet Characteristic**, **Defender Artifacts**,and **User & Organization Behaviors**.

Internet Characteristic





Defender Artifacts

User & Organization Behaviors

Attacker Related

* **Attacker Related**
  + Attacks
  + Vulnerabilities
  + Exploits
  + Cybercrime Activities
  + Adverse Events?
* **Internet Characteristic**

(Need more research)

* **Defender Artifacts**

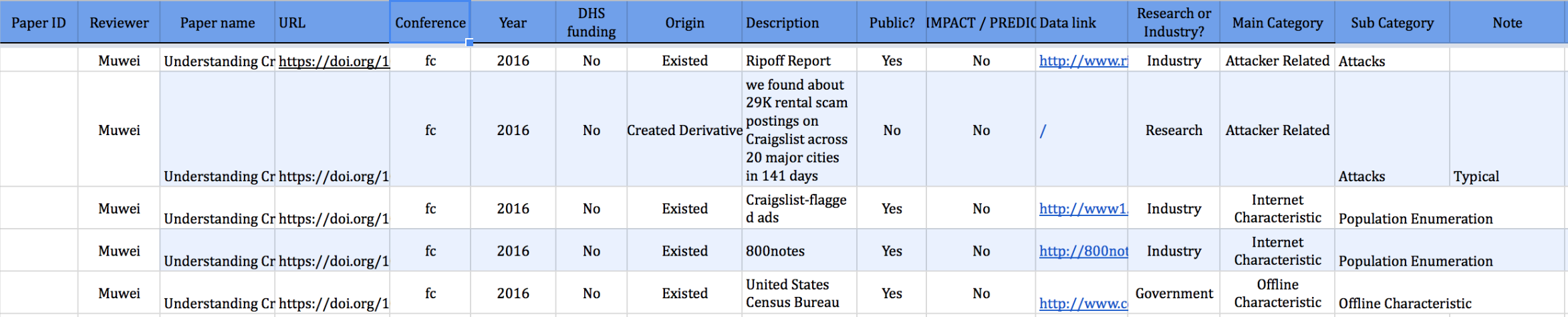
(Need more research)

* **User & Organization Behaviors**

(Need more research)

**Dataset Entry Example**

Every entry is used to describe only one dataset. If there is multiple datasets in one paper, use as many entries as needed.



* Paper ID: (I don’t know what that is)
* Reviewer: Person who review this paper and extract this dataset
* Paper Name: The name of the paper where the dataset is extracted from
* URL: The URL that gives access to the paper
* Conference: The conference in which the paper is presented
* Year: The year when the paper is presented in the conference
* DHS funding: Whether the paper is founding by Department of Homeland Security (DHS)
* Origin: Whether the dataset is Existed, Created Primary, or Created Derivative
* Description: Description of dataset (Should be directly copy from paper if it is possible)
* Public?: Whether the dataset is public available or not
* IMPACT/PREDICT: Whether the dataset is in IMPACT database
* Data link: The URL that gives access to the dataset. If the dataset is from a research paper, name of that paper.
* Research or Industry?:
  + Dataset from a research paper, or from a .org website should be considered as Research
  + Dataset from .com website or from a company should be considered as Industry
  + Dataset from .gov website should be considered as government
* Main Category: Attacker Related, Internet Characteristic, Defender Artifacts, or Users & Organizations Behaviors
* Sub Category: Sub categories under main category
* Note: Anything that is worth notice