Availability of Datasets for Digital Forensics & what is missing

Cinthya Grajeda, Dr. Frank Breitinger, & Dr. Ibrahim Baggili

Undergraduate researcher, UNHcFREG member

DFRWS, Austin, Texas, 2017

Cyber Forensics Research & Education Group

Introduction

- Cyber forensics and cybersecurity are ever-growing fields
- Require continuous research to discover and overcome new challenges
- May or may not require datasets. For instance,
  - To perform Android malware analysis, you would need access to malware samples
  - Creating an encryption scheme may not necessarily require datasets
- Focus on the type of research that requires a dataset
- To produce high quality research results, three critical features must be examined:
- The **quality, quantity** and **the availability of the datasets**
- To understand how our community deals with datasets, its limitations in availability, and what types are used:
  - Analyzed **715** cybersecurity and forensics articles from various venues (**2010 – 2015**)

Contributions

- Provide available datasets and where to find them
- Centralized source dataset website
- Insight into datasets that are missing
- Other shortcomings where more research might be needed
- Encourage more researchers to unite in the progress in sharing their datasets

Methodology

- Journal examined: Digital Investigation
- Conferences:
  - IEEE Security & Privacy, Digital Forensic Research Workshop (DFRWS - USA, EU), International Conference on Digital Forensics & Cyber Crime (ICDF2C), and Association of Digital Forensics, Security and Law (ADFSL)
- For each article utilizing a dataset we asked the following questions:
- The availability of the datasets
  - Was the utilized dataset available prior to the research? (re-usage)
  - If the dataset was available prior to the research:
    - Was the origin disclosed and was it freely available?
  - Were there any datasets created through the research?
    - Were they released after publishing? (availability)

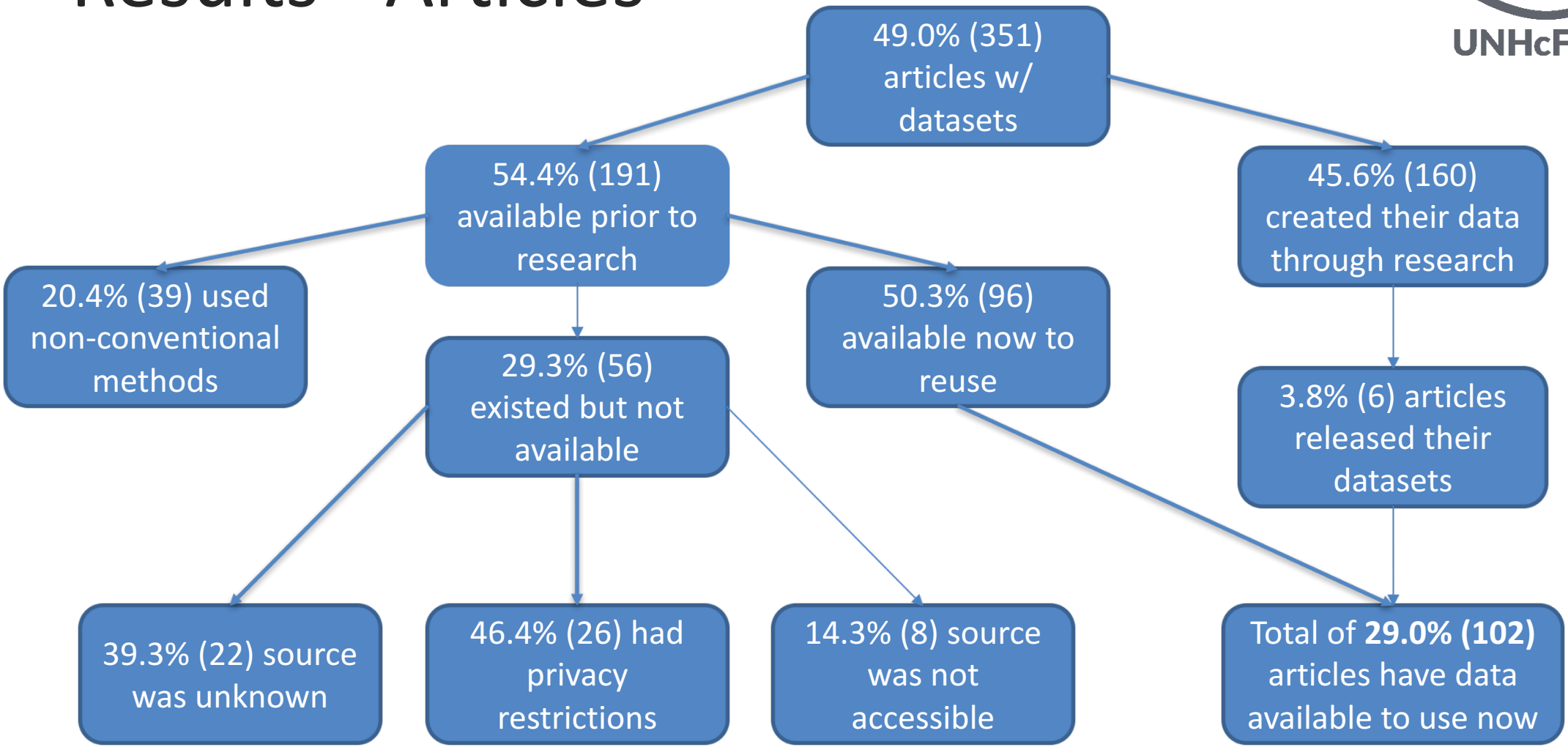
Methodology cont.

- Were there any third party databases, services or online tools used in the creation of datasets?
- Origin of datasets:
  - **Computer generated** (e.g., an algorithm, bot, /dev/urandom)
  - **Experiment generated** (e.g., a user creates specific scenarios)
  - **User generated** (e.g., real world data)
- Kinds of datasets: What datasets exist and can be used by researchers?
- What is missing: What datasets or other things are currently missing?
- Online searches
  - Queried Google for available datasets / repositories not found in the article analysis

Results overview

- Out of the 715 articles, **351** used datasets
- About **49%** articles focused on experiments using datasets
- **51%** focused on standards, techniques, policies and topics on programming, algorithms, etc.
- Interesting fact - the conference with the highest percentage on use of datasets:
  - DFRWS (US, EU) with **~86%** (78 out of 91 articles) within that conference
  - Digital Investigation (Journal) with ~57% (108 out of 190 articles)
  - ICDF2C with 56% (60 out of 107 articles)
  - ADFSL with 33% (29 out of 87)
  - The least was IEEE Security & Privacy with ~ **32%** (76 out of 240 articles)

Results - Articles



Results – Creating vs. re-using datasets

- What do researchers prefer?
  - Re-use datasets than create their own with over **54%**
- Over 45% preferred to create their own datasets
  - Unfortunately, less than **4%** shared those datasets
- Several reasons why they prefer to create their own:
  - Nothing was available
- It also depends on the type of research:
  - E.g., Some train algorithms based on simulated or experiment data so they have to create their own
  - But, for evaluating performance or comparing two algorithms often real world datasets are favored

Results – Origin of datasets

- Experiment generated datasets
- The majority of articles relied on them with over **56%**
- Reasons why:
  - There is a lack of real world datasets available to the community
  - Especially important when conducting experiments on new technologies
    - E.g., Xbox 360 and shared
  - Allows researchers to test and verify the data
  - Common within the area of digital forensics

Results – Origin of datasets cont.

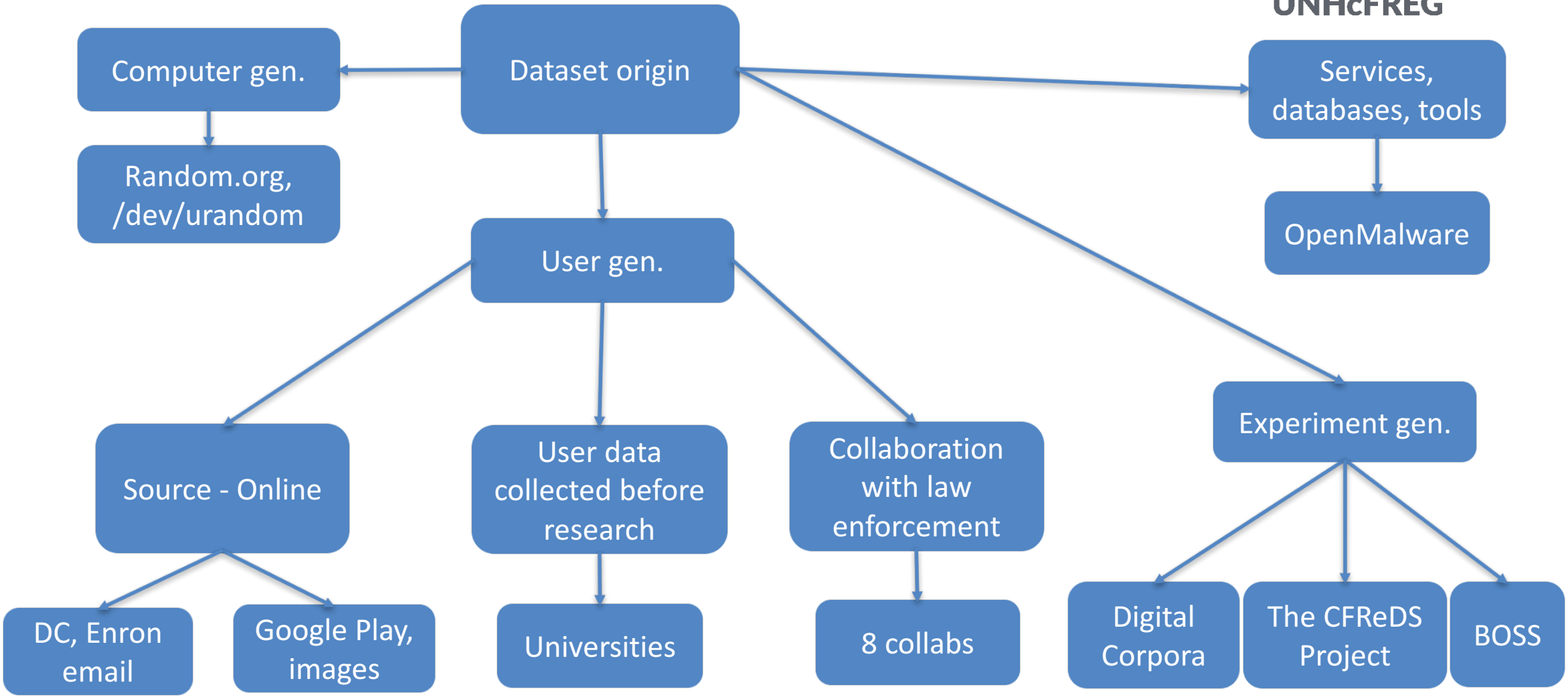
- User generated datasets
- The 2nd most used type with over **36%**
  - Crucial for developing reliable algorithms and tools
  - In the scientific field, scientists do use it more often, share it and recreate experiments but in our field…
  - How can we learn from our past when we don't have real, accessible data to learn from?
  - Two major reasons are:
    - Copyright and privacy laws prohibiting the sharing of datasets with the community

Results – Origin of datasets cont.

- Computer generated datasets
- Less than **5%** of the analyzed articles employ those datasets
- Not a surprise –
- Often researchers in digital forensics want to solve real world problems
- Thus, they cannot use simulated data

Where do they come from?

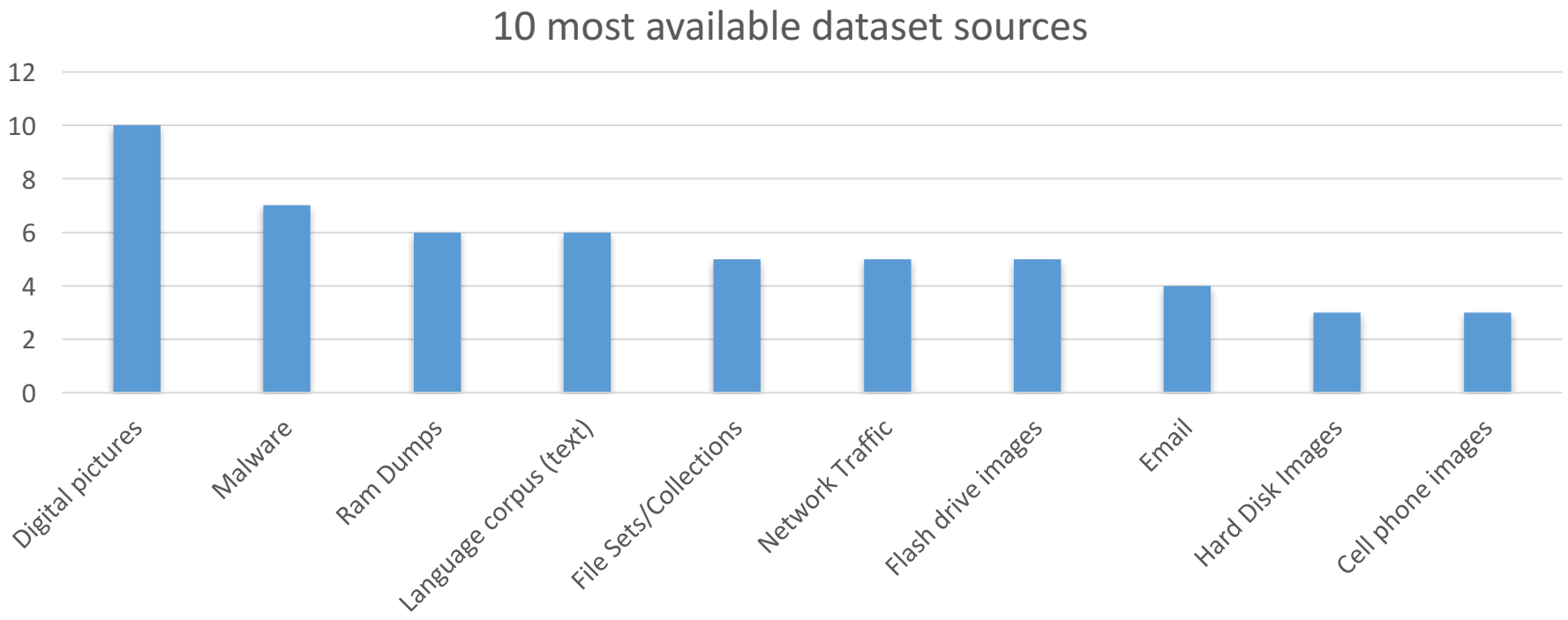
UNHcFREG



Types of Datasets

UNHcFREG

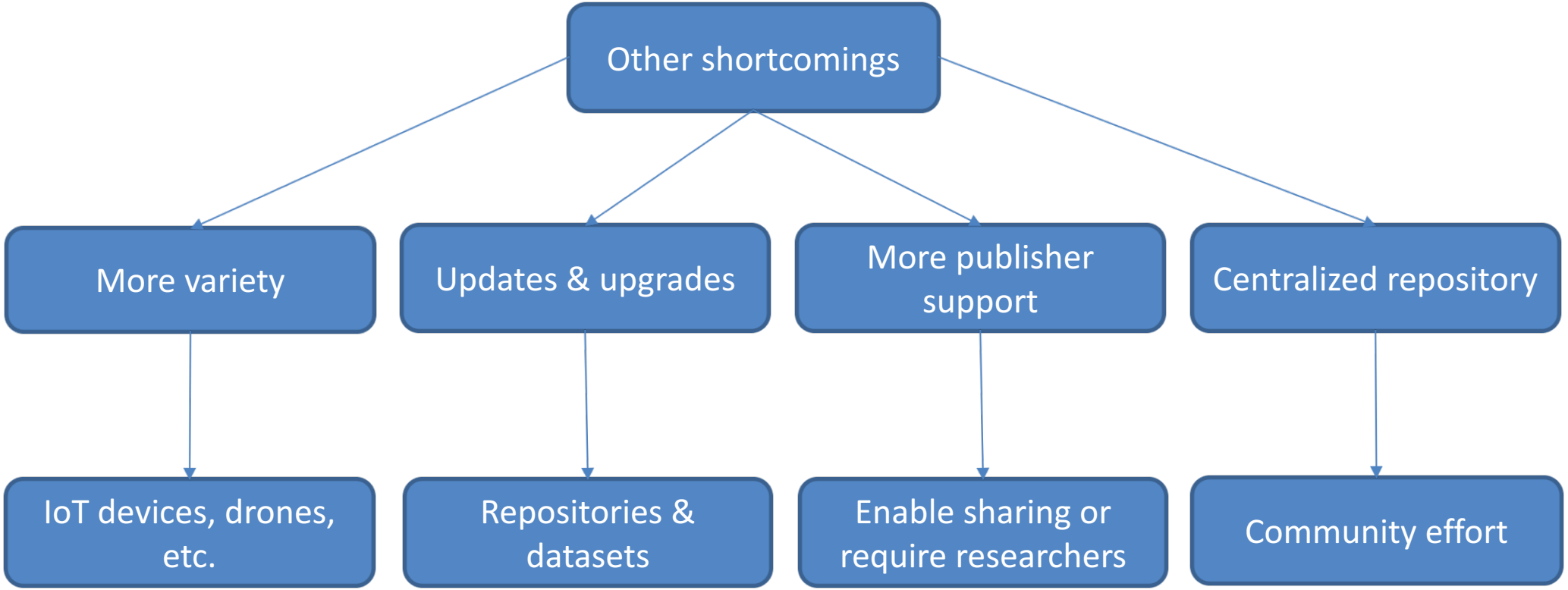
- Found over 70 different types of datasets and organized them in 21 categories



What is missing? Why not share?

- According to researcher feedback
- First, may not have the capability of sharing the set
  - E.g., the dataset is too large or no online resources available
  - R1 - **“At the time of publishing, we did not have a stable platform through which we could provide access to our data”**
- Second, it may be related to privacy concerns
  - R2 - **“I probably wouldn't want to share them (at least not in a publicly accessible manner) because when I picked the content off the Internet, I didn't take into consideration that there might be some privacy or copyright issues that may come up”**
- Third, researchers simply didn’t think of the importance of sharing their data
  - R3 - **“Initially I did not exactly have in mind how important it was to curate and share such data”**

What is missing?



Conclusion

- Researchers **prefer to reuse datasets** if available than creating their own
- They appreciate and utilize them. E.g., Digital Corpora
- Saves time and possibly money
- Especially true of real world datasets to produce high quality and realistic results
- **Lack of shared datasets** among the community still a **problem**
  - Less than 4% shared datasets created through research
- The mindset of researchers needs to change
- Data should be **released** when possible
  - Enable competition and ultimately lead to better results

Future Work

- Continue updating and maintaining our centralized source repository
- Spread the word and encourage others to contribute
- We hope in the future more people will publish their datasets
- **Contribute** to the our website or others:
  - Digital Corpora
  - Impact Cyber Trust

Centralized source repository

- Supplementary datasets and services are located in our central dataset website @
- http://datasets.fbreitinger.de
- Thanks to graduate researcher **Mateusz Topor** for his help in creating the dataset source website

Contact & Questions?

Cgraj1@unh.newhaven.edu
ibaggili@newhaven.edu
http://www.unhcfreg.com

fbreitinger@newhaven.edu
http://datasets.fbreitinger.de

Limitations

- Our analysis was performed by manual inspection
- Human error might have been introduced
- We attempted to alleviate any errors by conducting multiple runs
- The papers come from selected venues and do not include every single paper published in the cyber security and forensics domain
- We believe our results are still applicable and represent the status of datasets in our field

Related work

- Our study was inspired by Abt & Baier (2014) who published an article named "Availability of ground-truth in network security research"
- They analyzed 106 network security papers from 2009 to 2013
  - Many researchers manually produced their datasets (70%)
  - Datasets were often not released after the work was completed (only 10%)
- In our study we focus on all kinds of datasets that might be useful for cyber security and forensics research
  - It expands to a broader number of articles, and results from Google searches
  - Provides an overview of existing datasets and what is available to use for research

Definition of a dataset

- A collection of related, discrete items that has different meanings depending on the scenario and was utilized for some kind of experiment or analysis.
- E.g., files, memory dumps, raw images, pcap files, log files, outputs from /dev/urandom that were analyzed / processed
- Not considered datasets:
- An input that was only used to measure runtime efficiency, results written to log files, or a tool that outputs data which is never used.

Question?