Friedrich-Alexander-Universität
Erlangen-Nürnberg

When is a synthetic disk image realistic?

Lena L. Voigt

Joint Work with Felix Freiling and Christopher Hargreaves

Why Do We Need Synthetic Disk Images?

- Skills: education and training, proficiency testing
- Tools: development, testing of tools facilitating forensic tasks

Research Questions:

- How can we define and measure the *realism* of synthetic disk images? Which challenges arise?

Feedback and Collaboration

Please connect with us to:

- give feedback on our general idea,
- propose further realism metrics,
- suggest previous/related work,
- provide sources for synthetic or real-world data.

Defining Realism of Forensic Datasets

- We define realism of a synthetic dataset S based on the set of features F of the data that are statistically indistinguishable from a real-world dataset R, denoted as $S \cong_F R$
- We distinguish different types of realism
  - Strong F = set of all features
  - Possible F = set of all features that can potentially be satisfied within legal/operational restrictions
  - Controlled F = set of (possible) features required for specific use case

Research Assumption

Considering the definitions we propose for realism, the evaluation of realism can only be an approximation given a set of "known" and observable features. Both context and use case are important. We cannot define a standalone metric; instead, a combination of different approaches is desirable.

Measuring "Features" of Disk Images

- Qualitative Evaluation:
  - Candidate features:
    - artifact coherence,
    - narrative consistency, ...
  - Challenges: Time consuming, tedious, necessity of expertise to make a decision
- Quantitative Metrics:
  - Candidate features:
    - overall timespan,
    - number/variety of files,
    - number/variety of events,
    - time between events,
    - distribution of events over time,
    - number of applications installed,
    - number of browser entries, ...
  - Challenges: Limited/unknown expressiveness

First Results

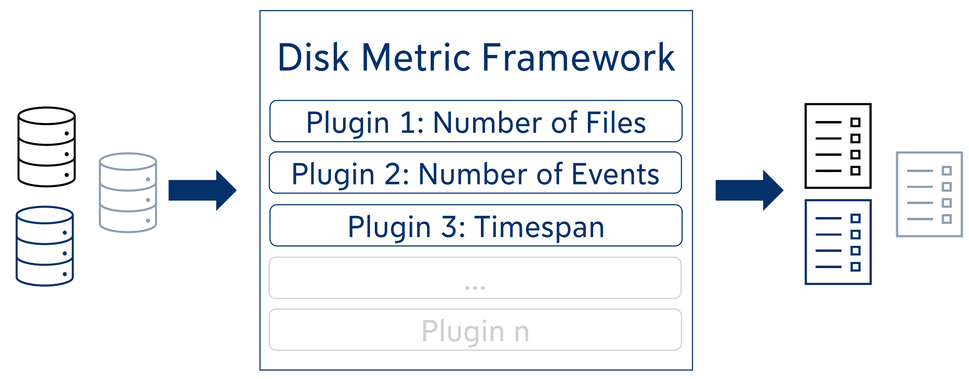

Figure 1: A framework for the extraction of quantitative disk image metrics

ID	Created	Type	Files	Events	Timespan
1	2016	Linux	314104	975221	12
2	2017	Linux	334020	950417	2
3	2020	Windows	639975	2808016	609
4	2020	Linux	521547	1154879	16
5	2021	Windows	552883	2286776	31
6	2022	Windows	84610	1193988	228
7	2023	Linux	271451	951413	9
8	2023	Windows	728713	2787882	11

Table 1: Comparison of some metrics of manually created synthetic disk images for education

Open Questions

- Which features are relevant for determining the realism of a synthetic disk image?
- With which features can we distinguish sets of real-world from sets of synthetic disk images?

Related Work

- Garfinkel, Farrell, Roussev & Dinolt (2009). Bringing science to digital forensics with standardized forensic corpora. Digital Investigation, 6, S2-S11.
- Du, Hargreaves, Sheppard & Scanlon (2021). TraceGen: User activity emulation for digital forensic test image generation. Forensic Science International: Digital Investigation, 38, 301133.
- Göbel, Baier & Breitinger (2023). Data for Digital Forensics: Why a Discussion on "How Realistic is Synthetic Data" is Dispensable. Digital Threats: Research and Practice, 4(3), 1-18.

Cybercrime and Forensic Computing
Research Training Group 2475