SOFIA: Finding and Profiling Malware Source Code in Public Archives at Scale

PI: Michalis Faloutsos, University of California, Riverside
Graduate Researchers: B. Treves, M.R. Masud, O.F. Rokon
https://maverics.cs.ucr.edu/sofia/
NSF CISE SATC Award ID#: 2132642

The Missed Opportunity

There is a significant amount of publicly-accessible malware source code on GitHub. Security research could greatly benefit from accessing such malware source code!

The Problem

How can we distinguish malicious from benign repositories at scale?

Note: these repositories are the malware created by the author of the repo (think of the hacker's source code).

Challenges

- Selecting among 32 million public repositories.
- Data is unstructured and heterogeneous.
- Finding the right algorithmic solutions:
  - Identifying the most discerning features
  - Using appropriate representations per feature.
- There is no existing ground truth.

Scientific Impact

- Develop effective methods to identify malware repositories.
- Develop methods to model the malware ecosystem.

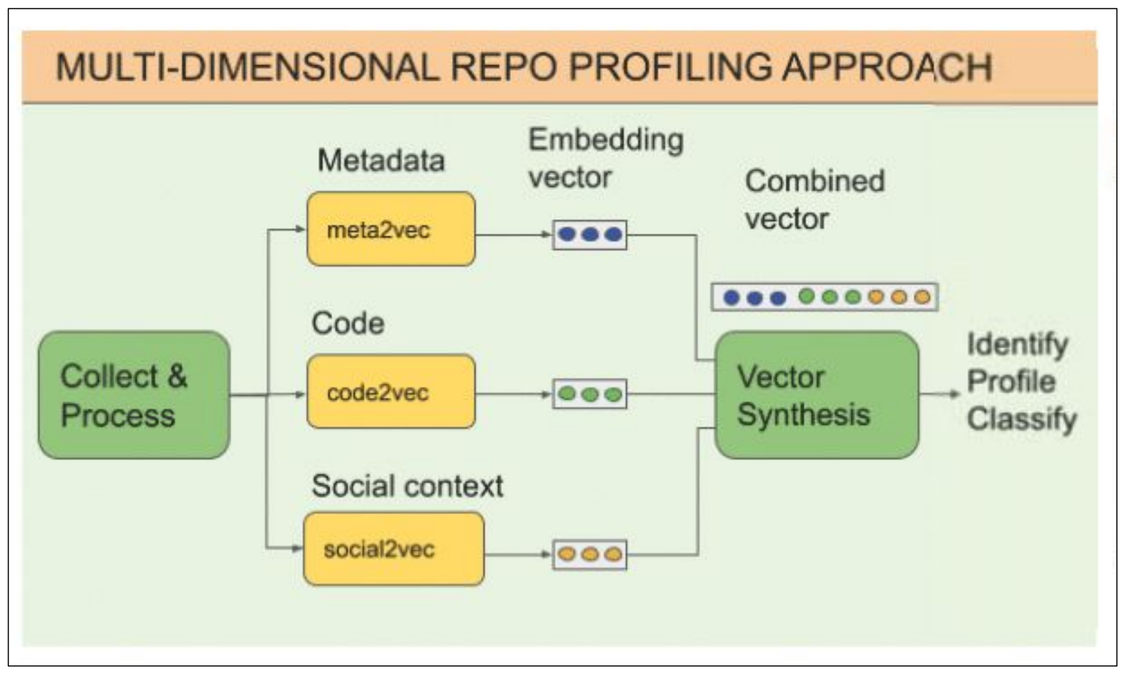
Initial Results

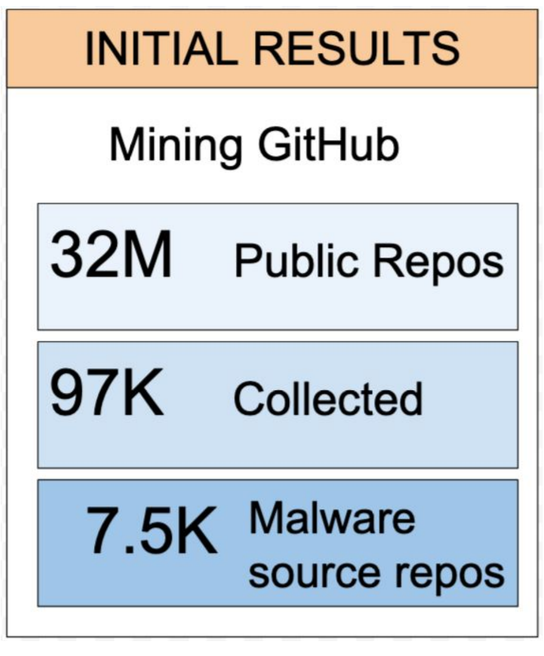
We identify 7.5K malware source code repos on GitHub with 89% precision using only metadata information.

Solution

A novel systematic approach for identifying malware repositories that leverages:

1. **Repo Metadata:** Repo title, description, topics, readme file, and relevant dates.
2. **Source Code:** Often contains structure or wording that reveals malicious nature of the repo.
3. **Social Context:** Behavioral patterns, roles, and interactions among repos and their authors.





Broader Impact (on society)

A critical piece towards reducing the $10.5T cost of cybercrime by introducing proactive measures
Key novelty:
Proactively find and detect hackers behaviors and activities
Reduce hacker's first mover advantage

Broader Impact (Education and outreach)

We develop a new course material.
Enhance the BPC activities of the PI, who has a record of an inclusive lab, with many current and past minority students including u-grads.
Target: Engage with 3-4 new minority students.

Broader Impact and Broader Participation (Quantify potential impact)

Accelerate malware research by providing our DB of sourcecode towards reducing the cost of cyber attack estimated at $1.1M per incident.