XIII JORNADAS DE JÓVENES INVESTIGADORES/AS DEL I3A

APOTHEOSIS: Bringing Approximate K-Nearest Neighbor Search to Similarity Digests

Daniel Huici, Ricardo J. Rodríguez, Eduardo Mena

Dept. of Computer Science and Systems Engineering, Universidad de Zaragoza, Spain

Instituto Universitario de Investigación de Ingeniería de Aragón
Universidad Zaragoza

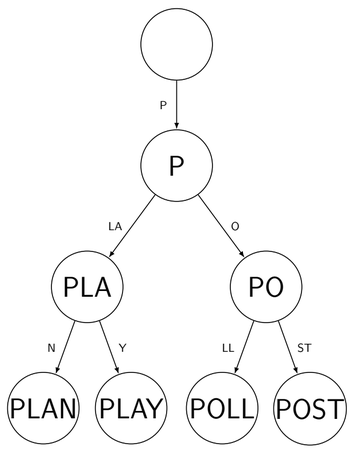
Universidad Zaragoza 1542

Detecting and Analyzing Malicious Artifacts for Timely Protection

- Specialized data structures ∪ Similarity Digest algorithms = APOTHEOSIS, an *approximate SD nearest neighbors* system [1]
- Extensible architecture design for multipurpose usage
- Two query searches: K-NN or similarity threshold
- Allow-list dataset of Windows system processes
- REST API interface (*Software-as-a-Service* model)

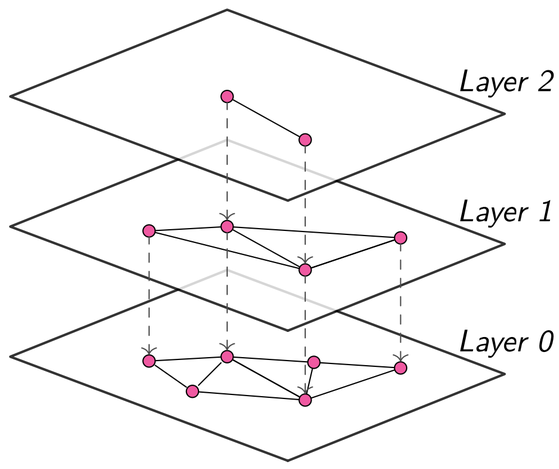
APOTHEOSIS Data Structures

Radix Tree [2]



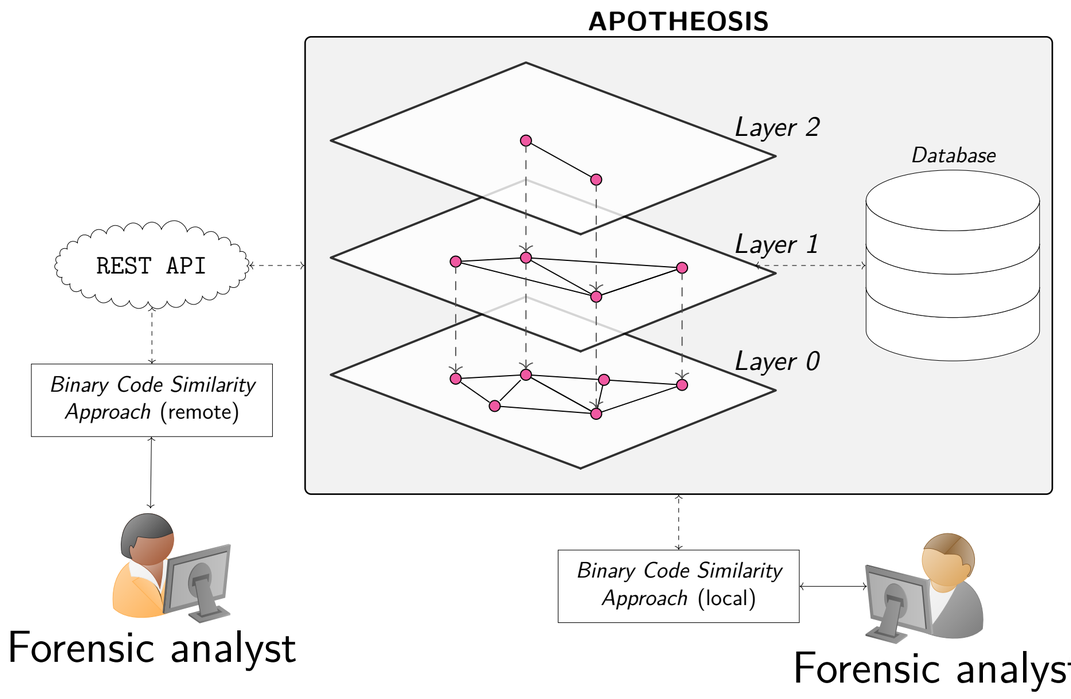
- Insert complexity: $\mathcal{O}(m \cdot \mathcal{A})$
- Search complexity: $\mathcal{O}(m)$ (m: length of the string and $\mathcal{A}$: size of the alphabet)

HNSW [3]



- Insertion and search complexity: $\mathcal{O}(N \cdot \log N)$

Use Cases

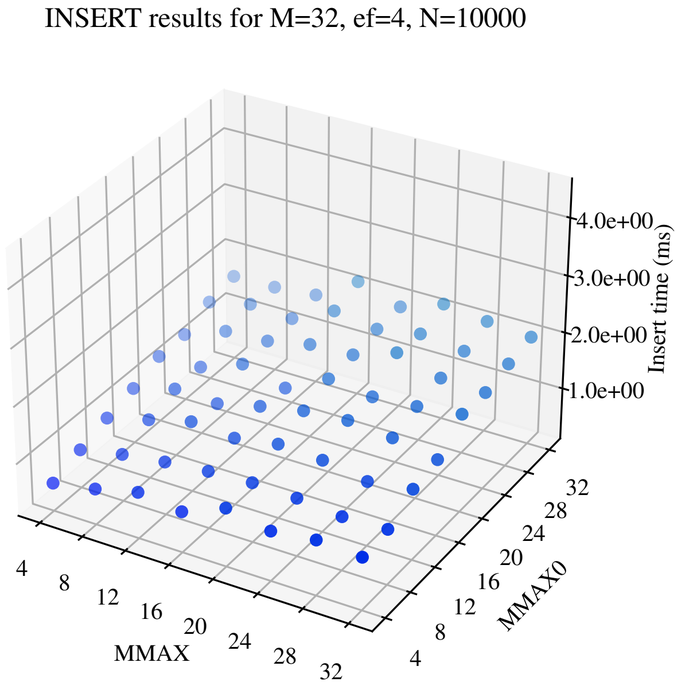


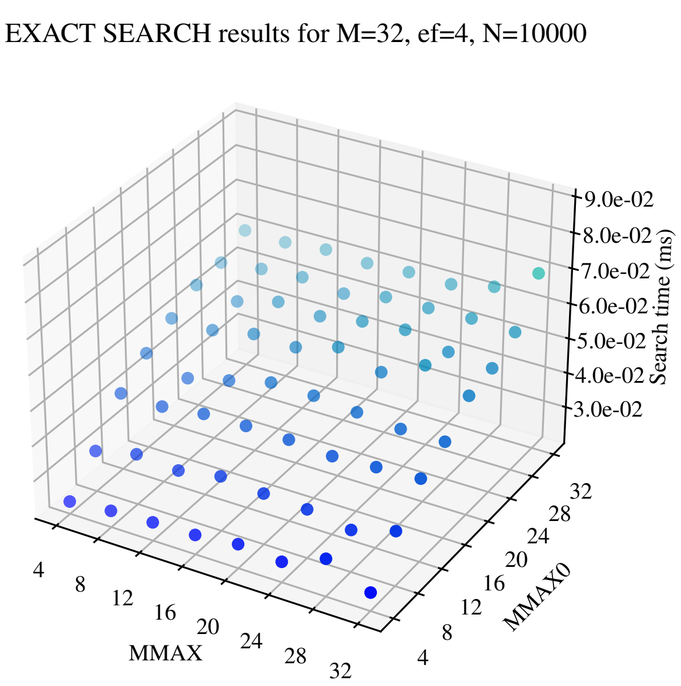
Allow-List Dataset

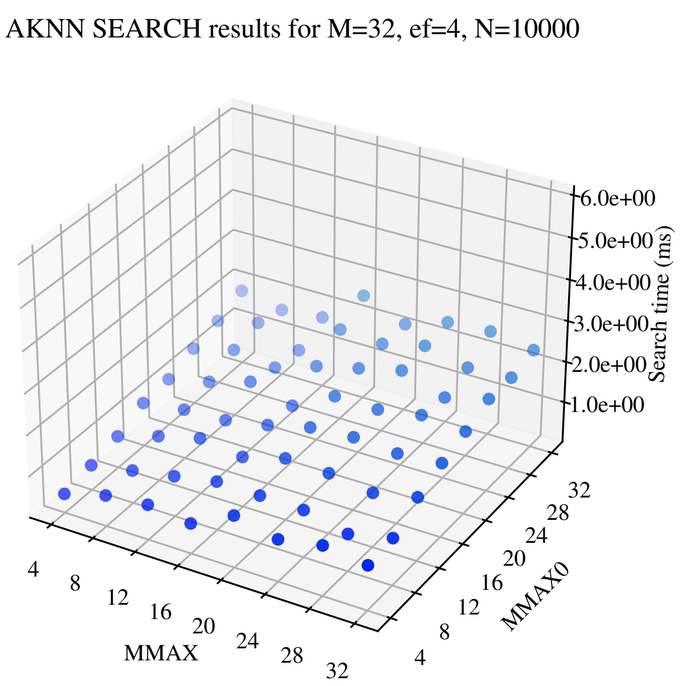
- Windows system module similarity digests
- > 13K Windows system processes dumped
- +2.7M of similarity digests (at page-size granularity)
- Different versions of Microsoft Windows operating system
- Similarity digest algorithms: `TLSH`, `SSDEEP`, `SDHASH`

Scan the QR to see the code on GitHub! (under GNU/GPLv3 license)

Evaluation







Conclusions

- APOTHEOSIS is an extensible, versatile system that leverages approximate search methods for similarity digests
- Database of Windows OS modules built as an allow-list
- REST API interface available for APOTHEOSIS with our database *(free access allowed for CERTs and research centers, contact us)*

ACKNOWLEDGEMENTS

- Spanish Ministry of Science and Innovation under grant TED2021-131115A-I00 (MIMFA); Spanish National Cybersecurity Institute (INCIBE) under the Recovery, Transformation and Resilience Plan funds, financed by the European Union (Next Generation); and University, Industry and Innovation Department of the Aragonese Government under *Programa de Proyectos Estratégicos de Grupos de Investigación* (DisCo research group, ref. T21-23R; SID research group, ref. T42-23R).

References

[1] D. Huici, R. J. Rodríguez, and E. Mena, "APOTHEOSIS: Bringing Approximate K-Nearest Neighbor Search to Similarity Digests," techreport, Dept. of Computer Science and Systems Engineering, Universidad de Zaragoza (Spain), Mar. 2024.

[2] D. R. Morrison, "PATRICIA—Practical Algorithm To Retrieve Information Coded in Alphanumeric," *J. ACM*, vol. 15, p. 514–534, oct 1968.

[3] Y. A. Malkov and D. A. Yashunin, "Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs," *IEEE TPAMI*, vol. 42, no. 4, pp. 824–836, 2020.

Contact data: reverseame@unizar.es