

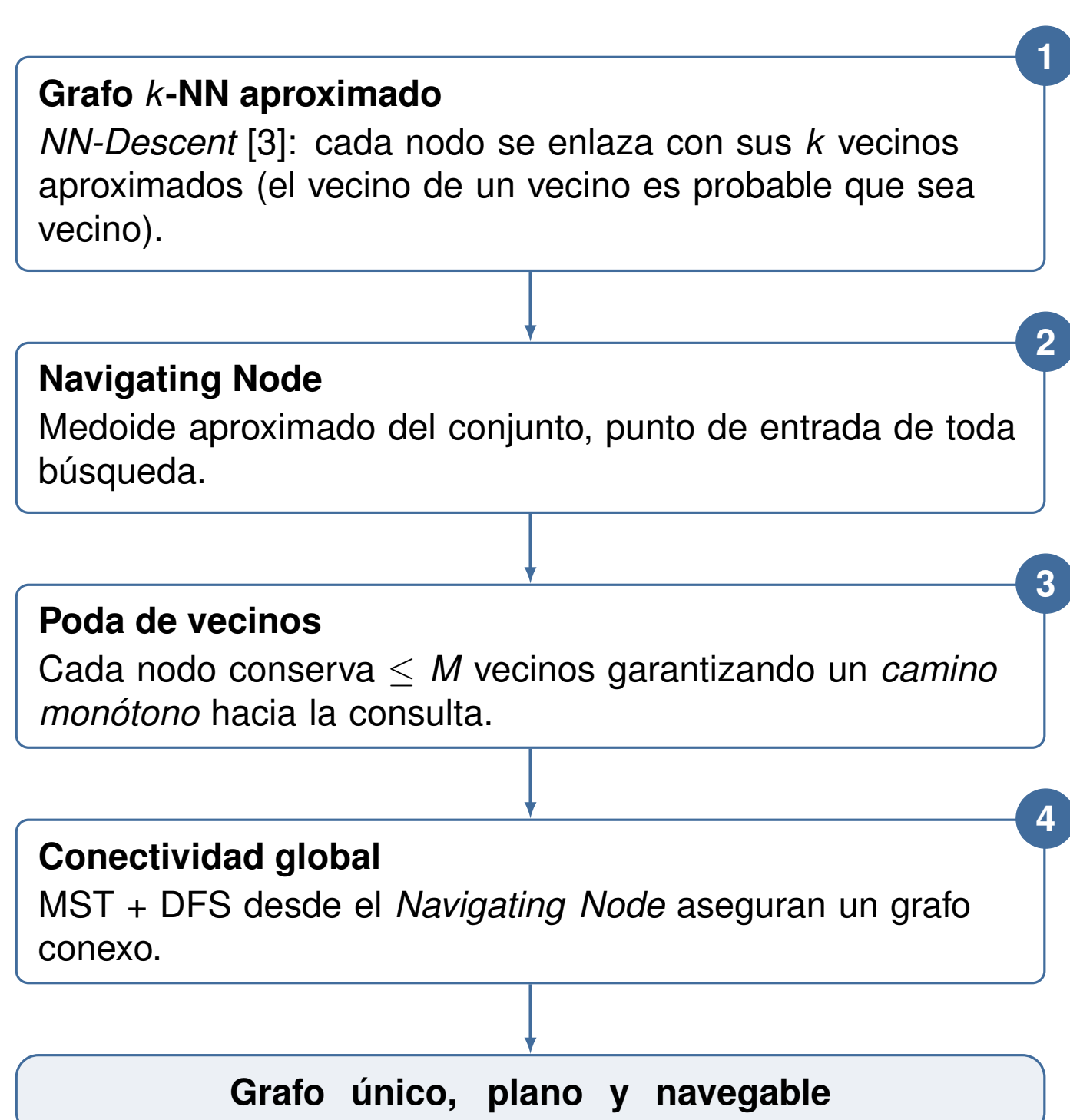
Extendiendo APOTHEOSIS, un sistema eficiente de búsqueda aproximada de similitudes

Miguel Ángel Royo, Daniel Huici, Ricardo J. Rodríguez
Dpto. de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, España

Motivación

- El **crecimiento exponencial** de los datos digitales ha incrementado significativamente la complejidad de las investigaciones forenses.
- Los **algoritmos de resumen de similitud** (TLSH, ssdeep) condensan cada artefacto en una huella digital o resumen comparable.
- La **búsqueda aproximada de vecinos** (ANN) acelera la identificación de coincidencias relajando el requisito de exactitud.

Construcción de NSG



Conjunto de datos

- Windows** [4]: funciones extraídas mediante análisis estático de binarios de sistema de varias versiones de Windows. 100k muestras para el índice y 1k muestras como consultas.

Conclusiones

- NSG permite realizar **búsquedas hasta 3,7 veces más rápido** que HNSW manteniendo un *recall* comparable. Además, su integración en APOTHEOSIS no supone cambios en la interfaz actual.
- Su punto débil actual es el **mayor tiempo de construcción** del índice, atribuible a la construcción del grafo k -NN que posteriormente es refinado por NSG.
- Como **trabajo futuro**, se plantea la exploración de mejores inicializaciones del grafo k -NN que permitan acelerar su construcción, así como la evaluación del comportamiento de APOTHEOSIS + NSG en conjuntos de datos de mayor escala.

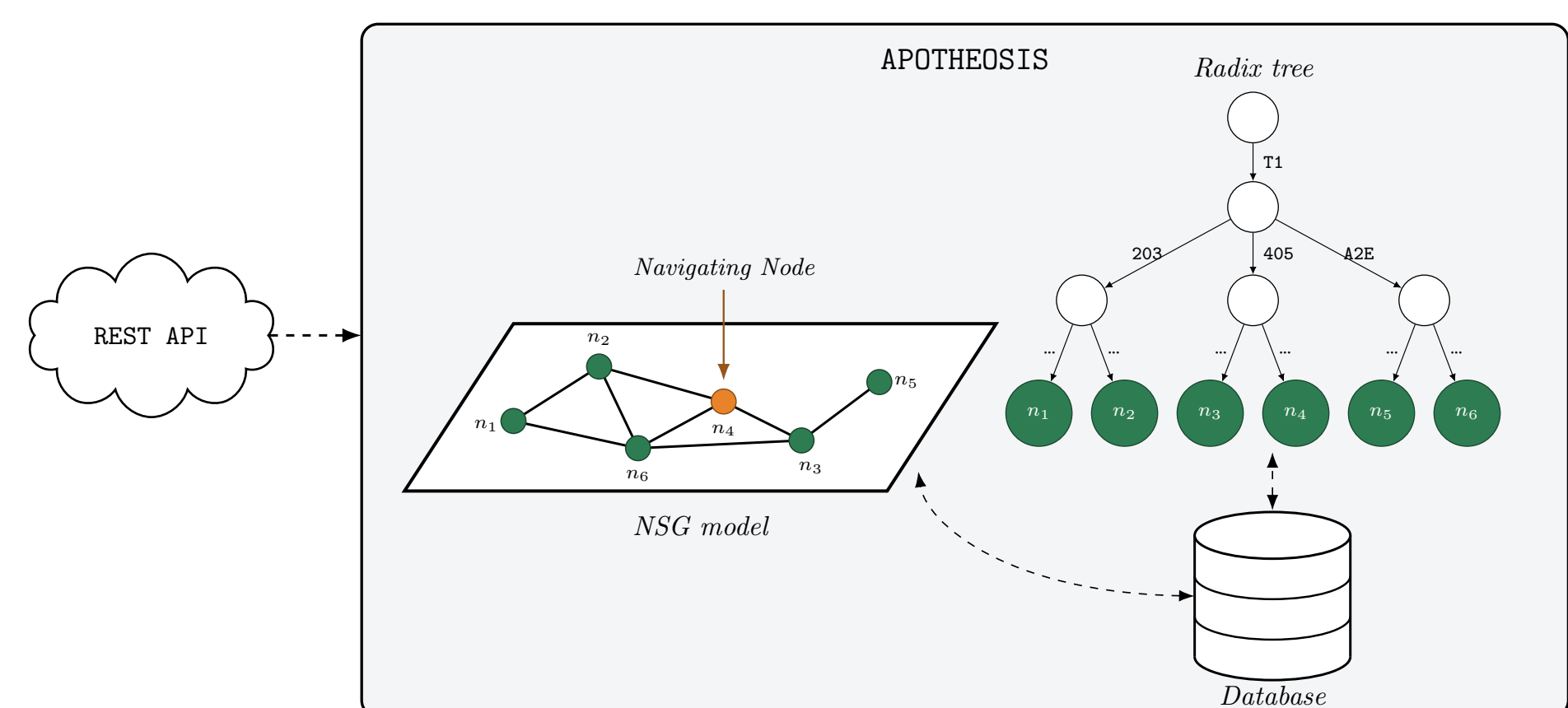
APOTHEOSIS

APOTHEOSIS (*APrOximaTe search systEm Of Similarity digests*) [1] combina dos estructuras complementarias:

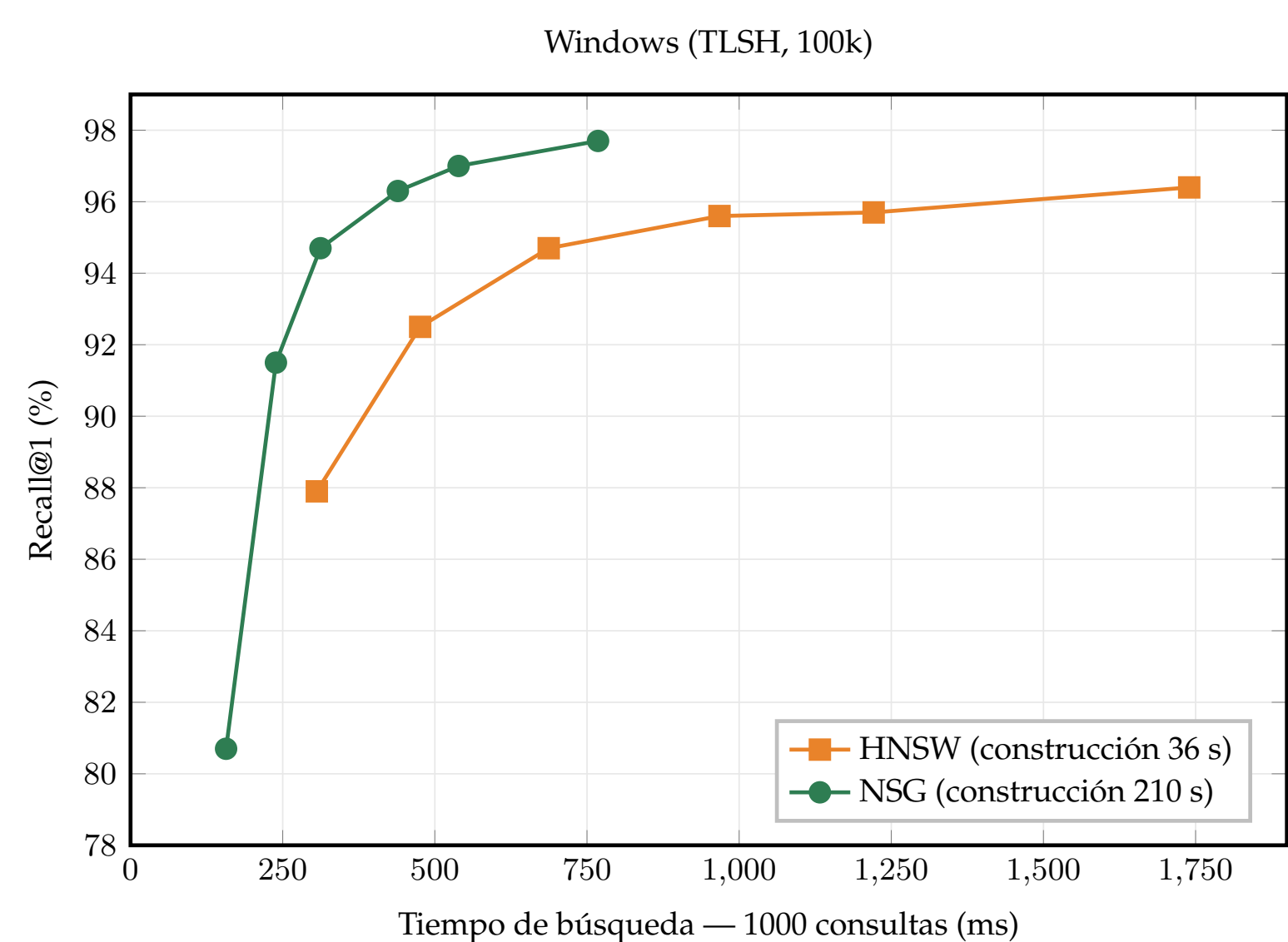
- Un **índice ANN** basado en grafos para localizar rápidamente los artefactos más similares a una consulta.
- Un **radix tree** que organiza y recupera los resúmenes de forma eficiente mediante sus prefijos.

La versión original emplea **HNSW** como índice. Este trabajo integra **NSG** [2] como alternativa, y evalúa su impacto en rendimiento, escalabilidad y precisión.

Arquitectura



Resultados



Referencias

- Huici, D., Rodríguez, R.J., Mena, E.: APOTHEOSIS: An efficient approximate similarity search system. *SoftwareX* **29**, 102016 (2025)
- Fu, C., Xiang, C., Wang, C., Cai, D.: Fast approximate nearest neighbor search with the navigating spreading-out graph. *Proc. VLDB Endow.* **12**(5), 461–474 (2019)
- Dong, W., Moses, C., Li, K.: Efficient k-nearest neighbor graph construction for generic similarity measures. In: *Proceedings of the 20th International Conference on World Wide Web*, pp. 577–586. WWW '11, Association for Computing Machinery, Hyderabad, India (2011)
- Huici, D., Rodríguez, R.J.: A dataset of Windows system binaries and similarity digests for enhanced forensic analysis. *Data in Brief* **62**, 111993 (2025)

Agradecimientos

Este trabajo ha sido financiado en parte por el proyecto PID2023-151467OA-I00 (CRAPER), financiado por MICIU/AEI/10.13039/501100011033 y por FEDER/UE, y por la Cátedra Inetum en Ciberseguridad y Cloud de la Universidad de Zaragoza.