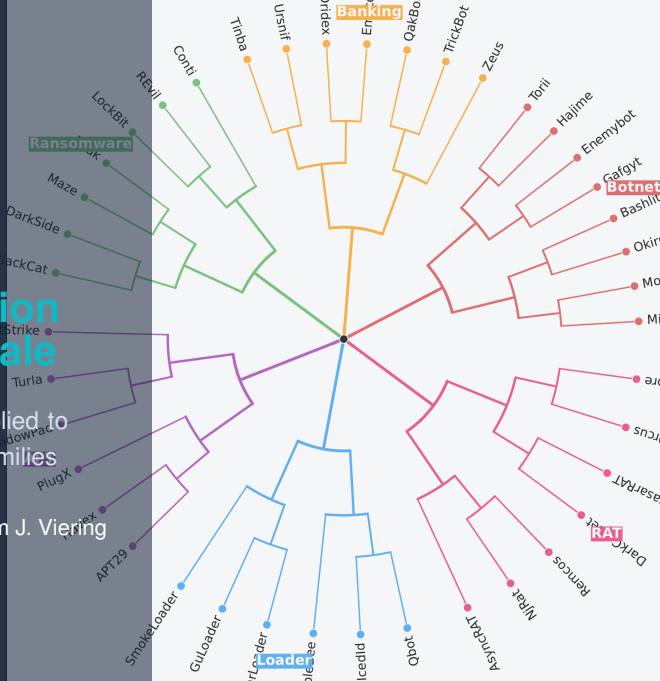
MalTree

Tracing Malware Evolution from Embeddings at Scale

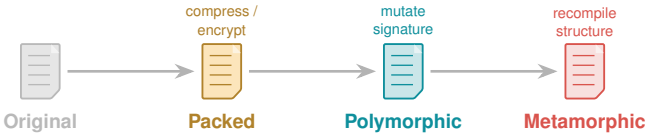
Bioinformatics-inspired phylogenetics applied to 103,883 malware samples across 538 families

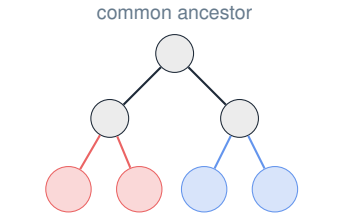
Akash Amalan Georgios Smaragdakis Tom J. Viering

Delft University of Technology • ICML 2026



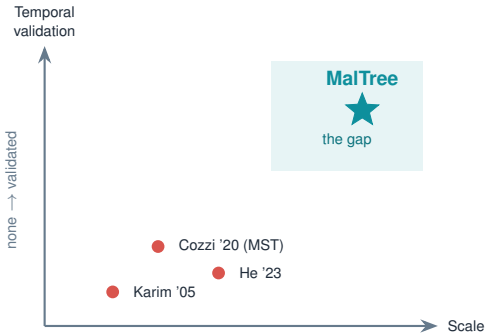
Detection is reactive, malware evolves





Takeaway. Malware evolves into **lineages**; MalTree models them with **phylogenetics**.

The gap in prior work: scale and validation



- **Small-scale**, single-modality
- **No validation**
- **Scale + validation**

What MalTree does

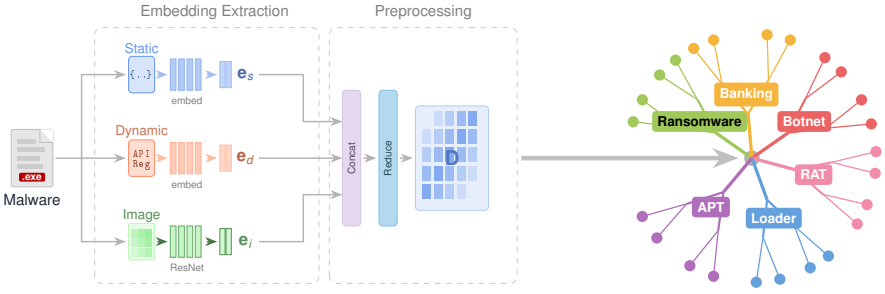
1. **Validated phylogenetic trees**
2. **Temporal validation** (VirusTotal)
3. **Inter-family lineages** (Mirai)

87.1%
temporal consistency

103,883
malware samples

538
families

The MalTree pipeline



pseudo-static $\mathbf{e}_s$: 3512-d | dynamic $\mathbf{e}_d$: 1000-d | image $\mathbf{e}_i$: 2048-d

concatenate 6560-d → reduce to 1000-d → distance matrix **D** → tree

Do the embeddings capture tree-ready distances?

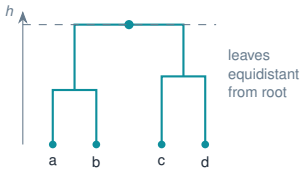
- **Good distances** needed
- **Families well-separated**
- **Linear probe** confirms

Family classification (linear probe)

Embedding	Accuracy (%)
Image ($\mathbf{e}_i$)	94.91 ± 0.16
Combined ($\mathbf{e}$)	$\mathbf{93.69 \pm 0.19}$
Pseudo-static ($\mathbf{e}_s$)	87.51 ± 0.48
Dynamic ($\mathbf{e}_d$)	87.23 ± 0.19
Random guessing	19.36 ± 0.17

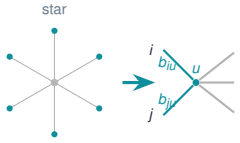
Tree construction: UPGMA and Neighbor-Joining

UPGMA



Assumes a **molecular clock**; tree is **rooted**.

Neighbor-Joining

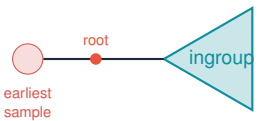


No clock: rates can vary; tree is **unrooted**.

Rooting the tree

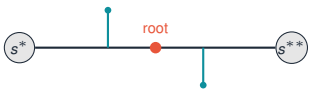
Neighbor-Joining is **unrooted**, so it needs a rooting step.

Outgroup rooting



Root via the **earliest-timestamp** sample. **Wins.**

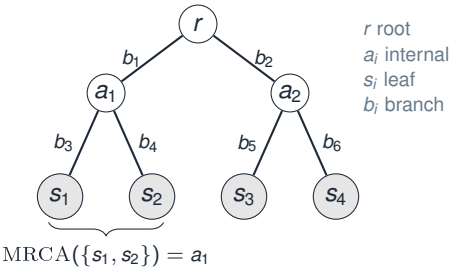
Midpoint rooting



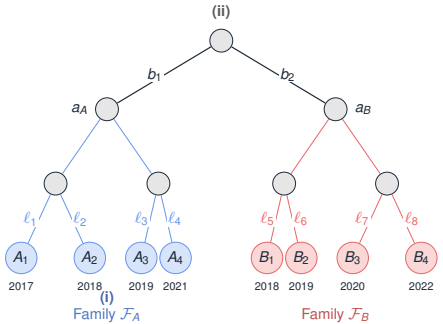
Root at the **midpoint of the longest path**; assumes a constant rate.

How we infer, prerequisite: the MRCA

- Tree $\mathcal{T} = (V, E, w)$: leaves are samples, internal nodes inferred ancestors, w = **branch lengths**.
- **Path distance** $d_{\mathcal{T}}(u, v) = \sum_e w(e)$; **MRCA**$(S)$ = deepest node whose subtree contains S.



How we infer: MRCA path-length, intra and inter-family



Leaves show first-submission year. **(i)** intra-family, **(ii)** inter-family.

- **Path-length (local clock).** For siblings, shorter L_i means earlier emergence; needs only local rate consistency.
- **Intra-family.** Shared MRCA, compare path lengths: $L_i < L_j \Rightarrow t(s_i) < t(s_j)$.
- **Inter-family.** Compare median distance to the deep MRCA; smaller $\tilde{d}$ diverged first.

Validation against VirusTotal timestamps

- **Sibling-order agreement**
- **Label-independent**
- **87.1%** vs 50%

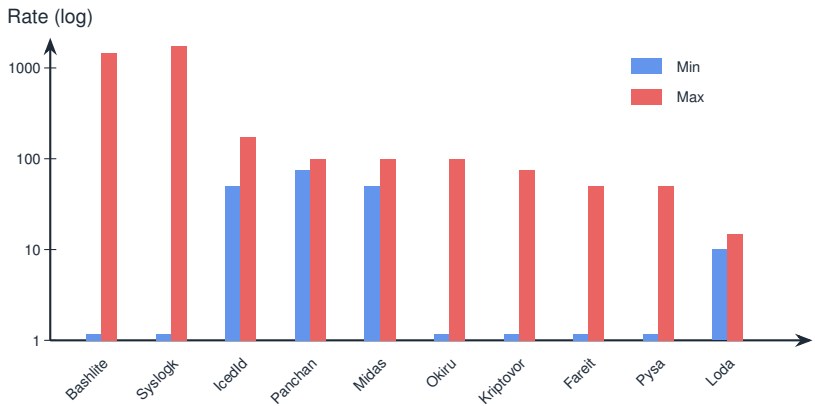
Year-level (full tree)

Method	Default	Outgroup	Midpoint
NJ	0.811	**0.871**	0.631
UPGMA	0.860	0.601	0.562

Month-level

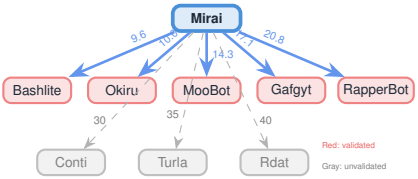
Method	Nominal	Outgroup	Midpoint
NJ	0.793	**0.853**	0.597
UPGMA	0.569	0.553	0.507

Does the molecular-clock assumption hold?



No: embedding drift varies $>10\times$ across families, so the clock fails, which favors Neighbor-Joining.

Case study: the Mirai botnet



Five validated descendants; lower weight = stronger link.

Import/export symbol Jaccard

	Mir	Oki	Bash	Gaf	Moo
Mirai	—	**.82**	.58	.07	.04
Okiru	.82	—	.60	.06	.05
Bashlite	.58	.60	—	.17	.04
Gafgyt	.07	.06	.17	—	.05
MooBot	.04	.05	.04	.05	—

Discussion and future work

What we showed

- Scales to malware
- NJ $>$ UPGMA
- Recovers lineages

Future work

- Phylogenetic networks
- Online updates
- Cleaner labels

Conclusion

MalTree builds and validates large-scale malware phylogenies, **87.1% temporal consistency** on 103,883 samples and 538 families, and recovers documented lineages such as **Mirai**.

A foundation to move beyond classifying samples toward understanding how malware families evolve over time.

github.com/AJ730/MalwareEvolution