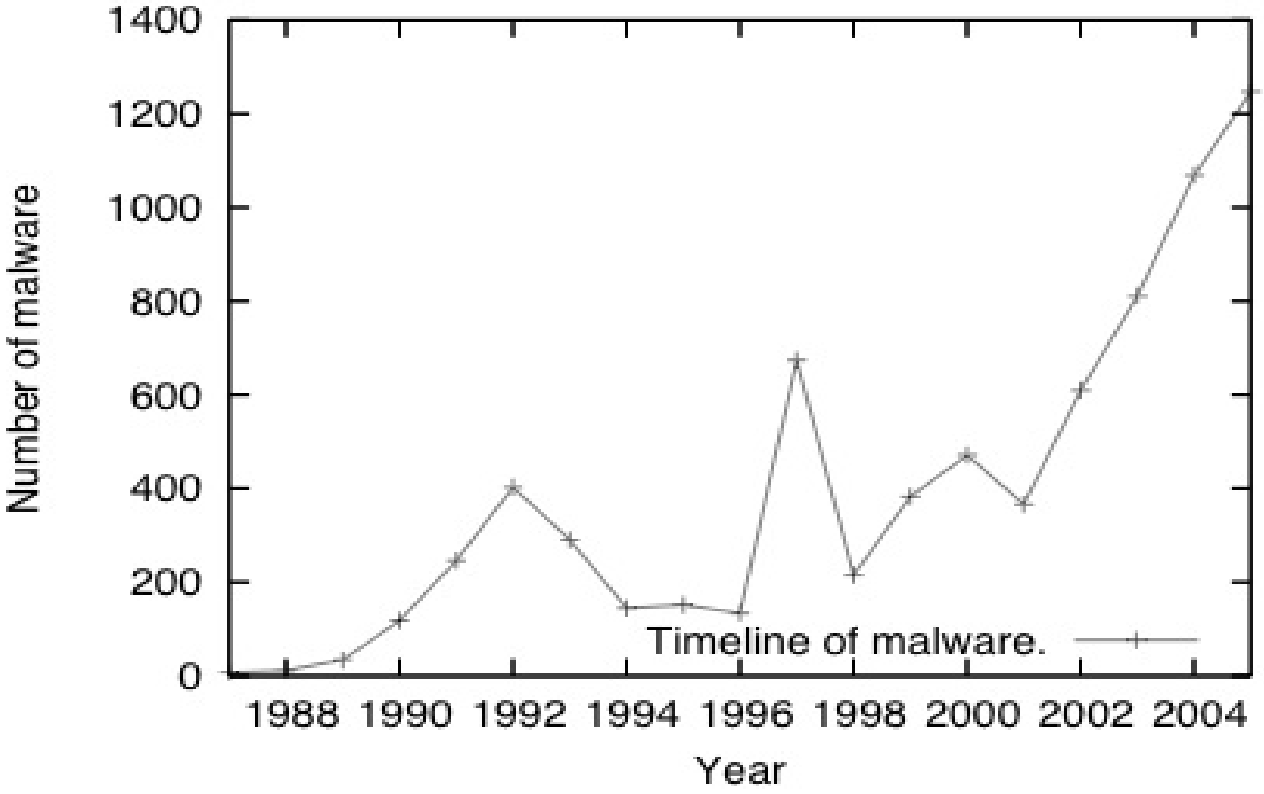
An Empirical Study of Malware Evolution

January, 2009

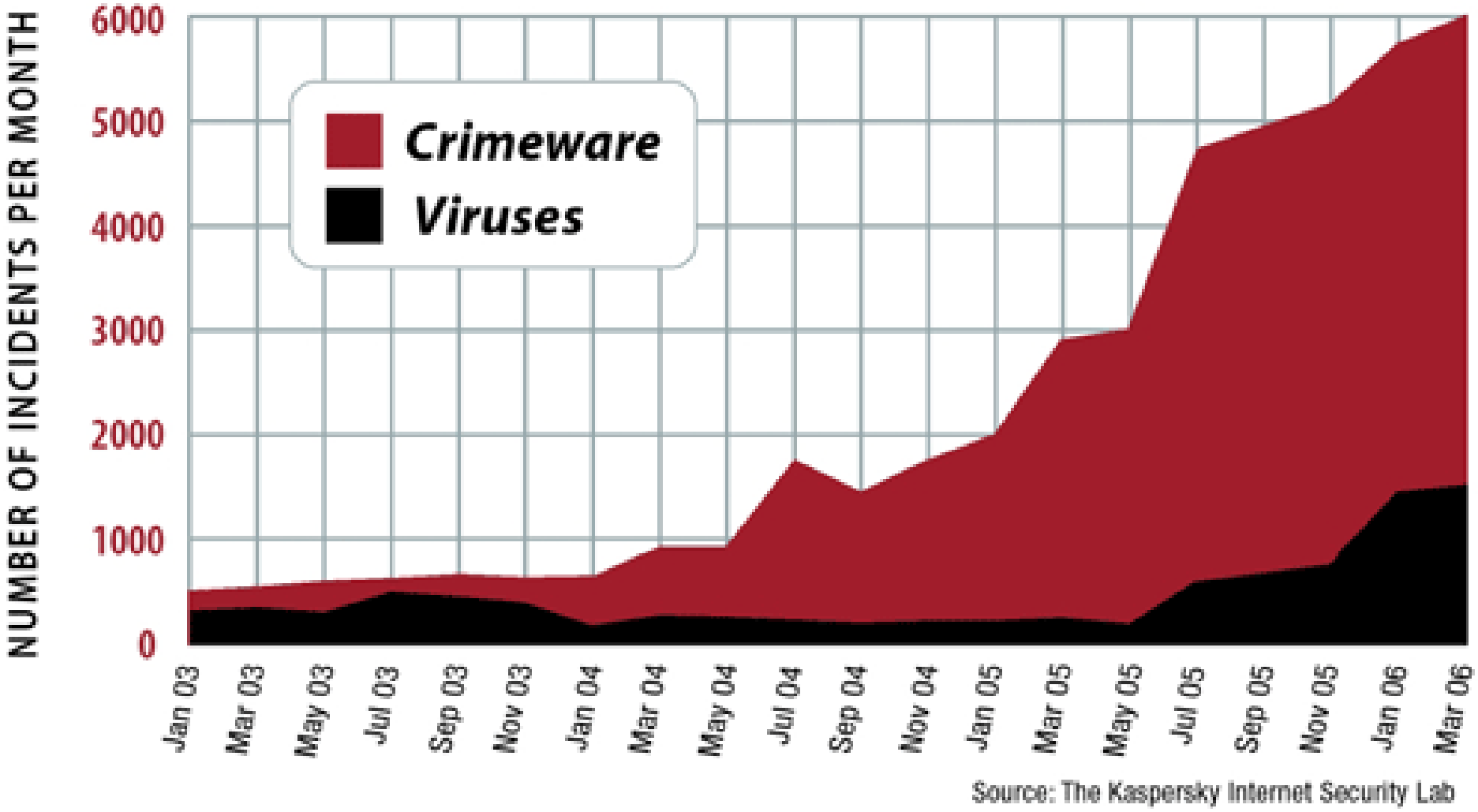
Archit Gupta, Pavan Kuppili
Aditya Akella, Paul Barford
Sanjay Bharadwaj
Computer Science
University of Wisconsin

Threat of malware



Monthly average of new malware doubled from 06 to 07

Continuing criminalization of Internet



Why evolutionary analysis

Many new malware are variants

Evolutionary history of malware can speed up response

Evolutionary trends can enable proactive development of defenses

Our approach

Focus on metadata

Text mining for malware properties

Creation of graph for related malware

Malware related through evolutionary links and commonality in properties

Output : A forest of Malware Family Trees

McAfee Avert Labs threat library

Database of 44,504 malware instances collected between 1987 - 2006

Malware characteristics

Methods of infection

Indications of infection

3182 malware instances used in study

Example entry

/32/Bagle.q@MM

ype	Virus
ubType	E-mail worm
iscovery Date	03/17/2004
ength	25,600Bytes

haracteristics:

his W32/Bagle variant bears the following characteristics:

- contains its own SMTP engine to construct outgoing messages

ymptoms:

resence of the following files in the %SysDir% folder :

- directs.exe (25,600 bytes)
- directs.exeopen (26,807 bytes)

ethod of Infection:

- Mail Propagation

Malware metadata mining

Map malware instances with properties

Automation is key

Metadata is unstructured

Use *Information Retrieval* to mine properties

IR for properties

Extract *maximal frequent phrases*

Not all phrases are propertiesof malware

Use *Google search engine* **for filtering**

4500 *malware properties* **found**

Establishing malware families

Derive a directed graph *G*

Vertices are malware instances

Edges imply strongly related instances

Edges indicate direction of evolution

Graph pruning

Start with*completely connected***graph**

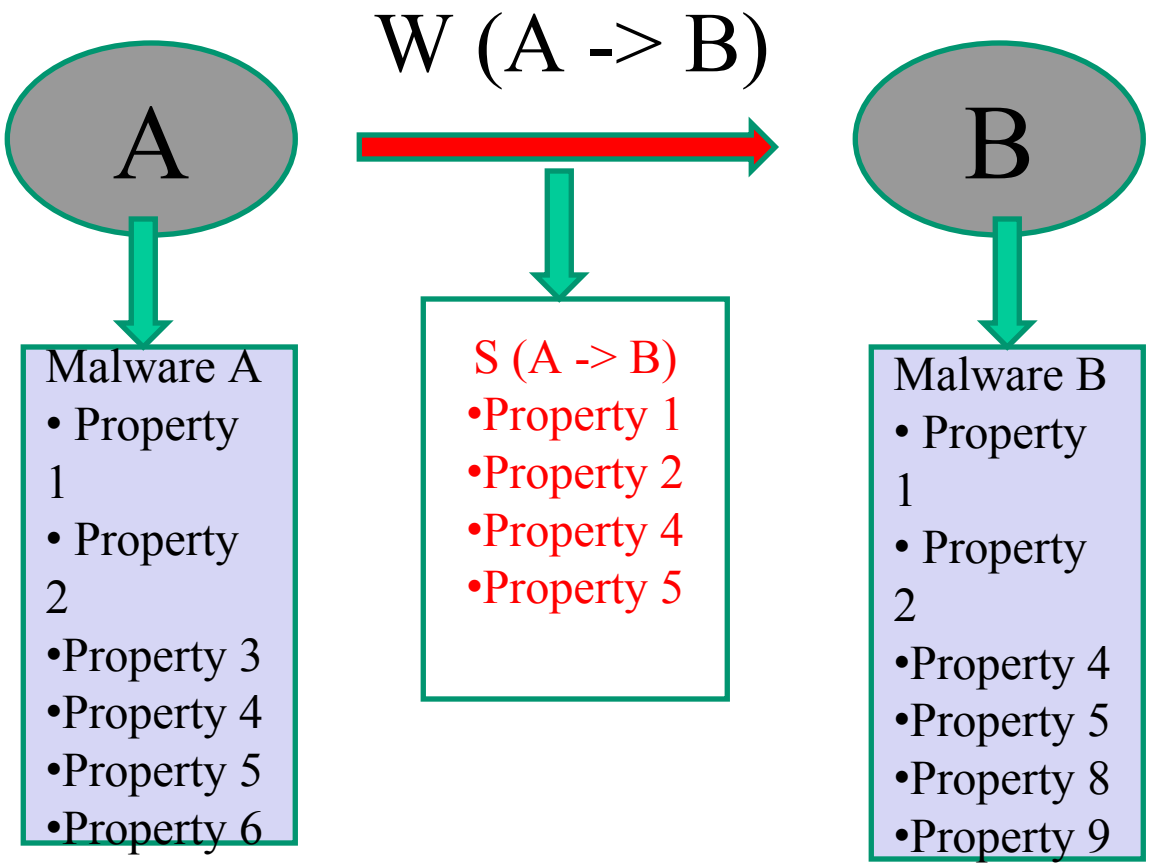
Edge weights reflect commonality in properties

A threshold δ_1 signifies independent development of malware

A threshold δ_2 used to prune spurious edge

Result – ***forest*** **of malware families**

Getting the malware families



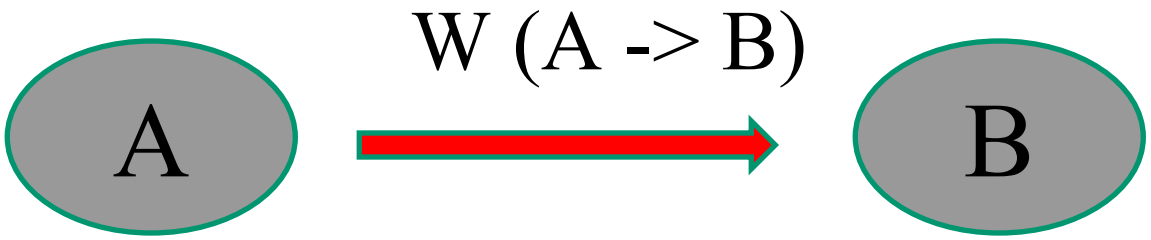
$|B| = 6$

$|S(A \rightarrow B)| = 4$

$W(A \rightarrow B) = |S(A \rightarrow B)|/$

$= 4/6$

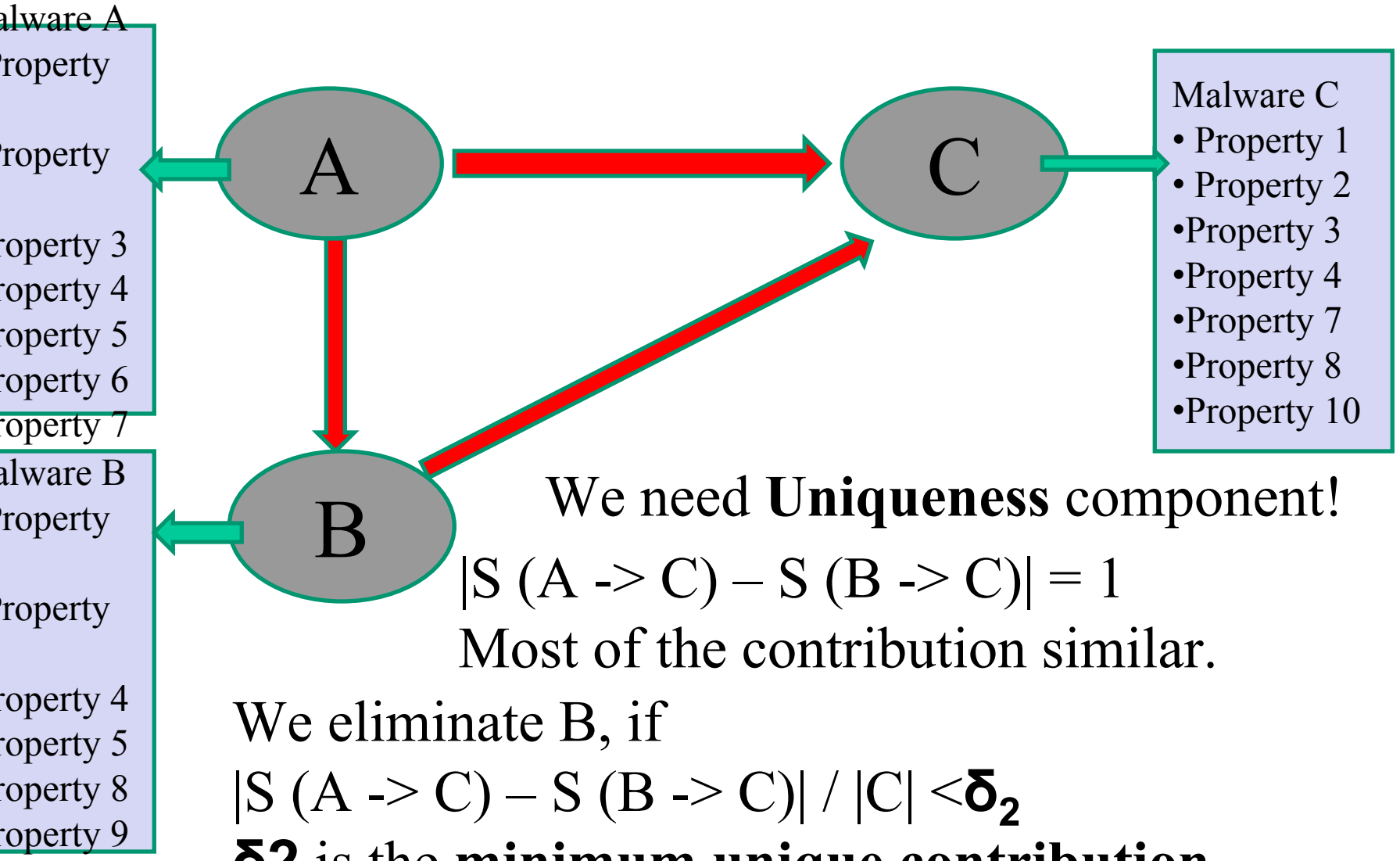
Is B a root of a family?



If W(A->B) > δ_1

δ_1 is the **minimum inheritance threshold**. Ensures significant fraction of properties of B are inherited.

Who spawned C?



We need **Uniqueness** component!

$|S(A \rightarrow C) - S(B \rightarrow C)| = 1$

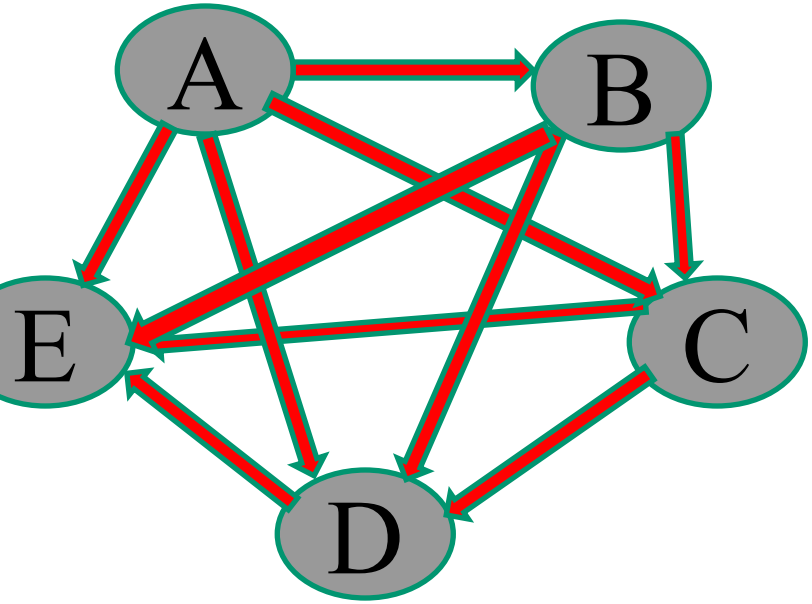
Most of the contribution similar.

We eliminate B, if

$|S(A \rightarrow C) - S(B \rightarrow C)| / |C| < \delta_2$

δ2 is the **minimum unique contribution**

Algorithm in motion



Prune edges based on δ_2 threshold. Remove edges which do not uniquely contribute enough.

W(C->D) $<\delta_1$
D does not inherit enough properties from C.

Final Malware family derived!

We start with a completely connected graph.
Directions are based on date of origin of malware.

Results

General results

Evolutionary dynamics

External factors

Mytob family

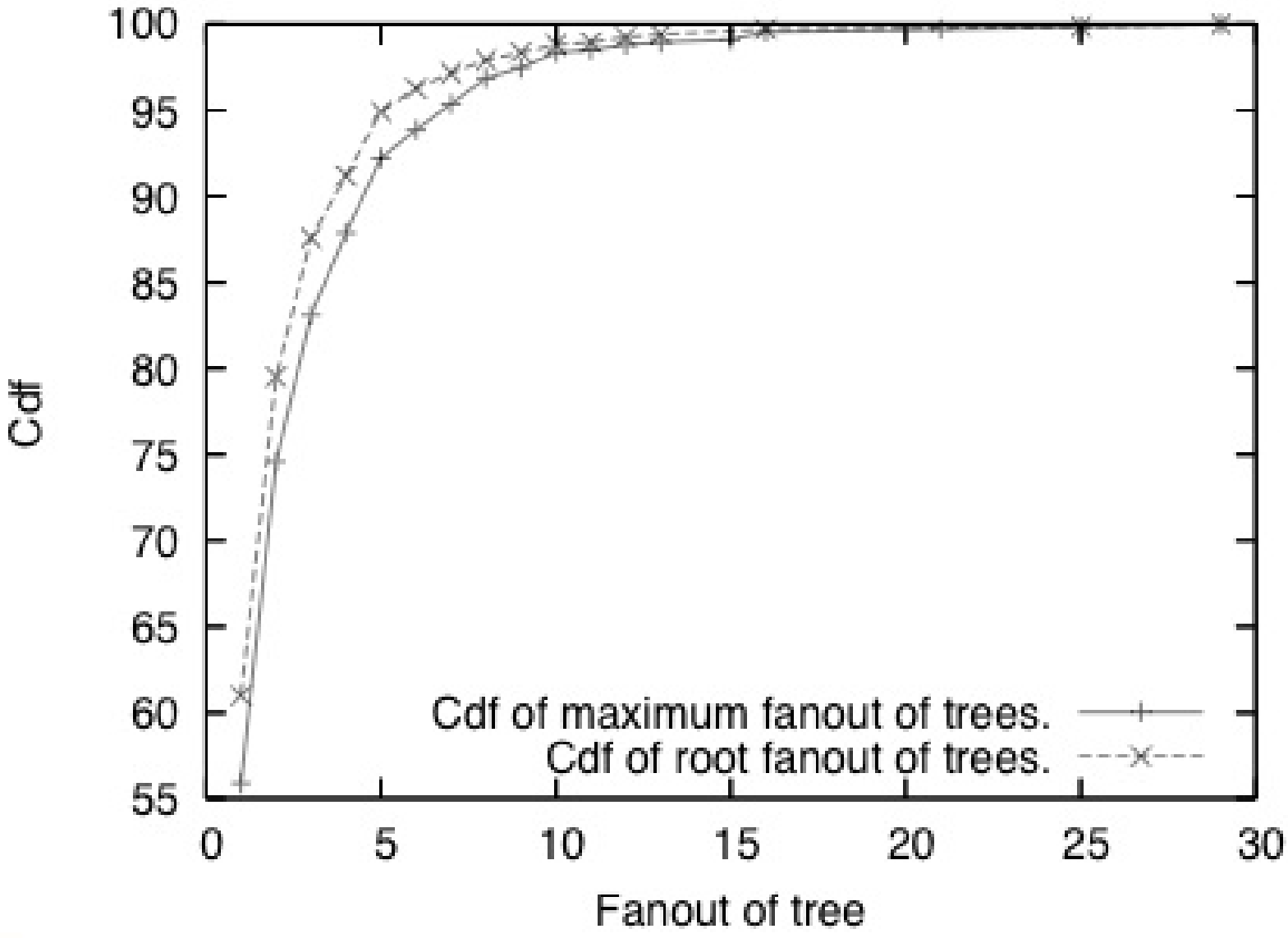
General Results

After experimentation, $\delta_1 = 0.7$ and $\delta_2 = 0.3$

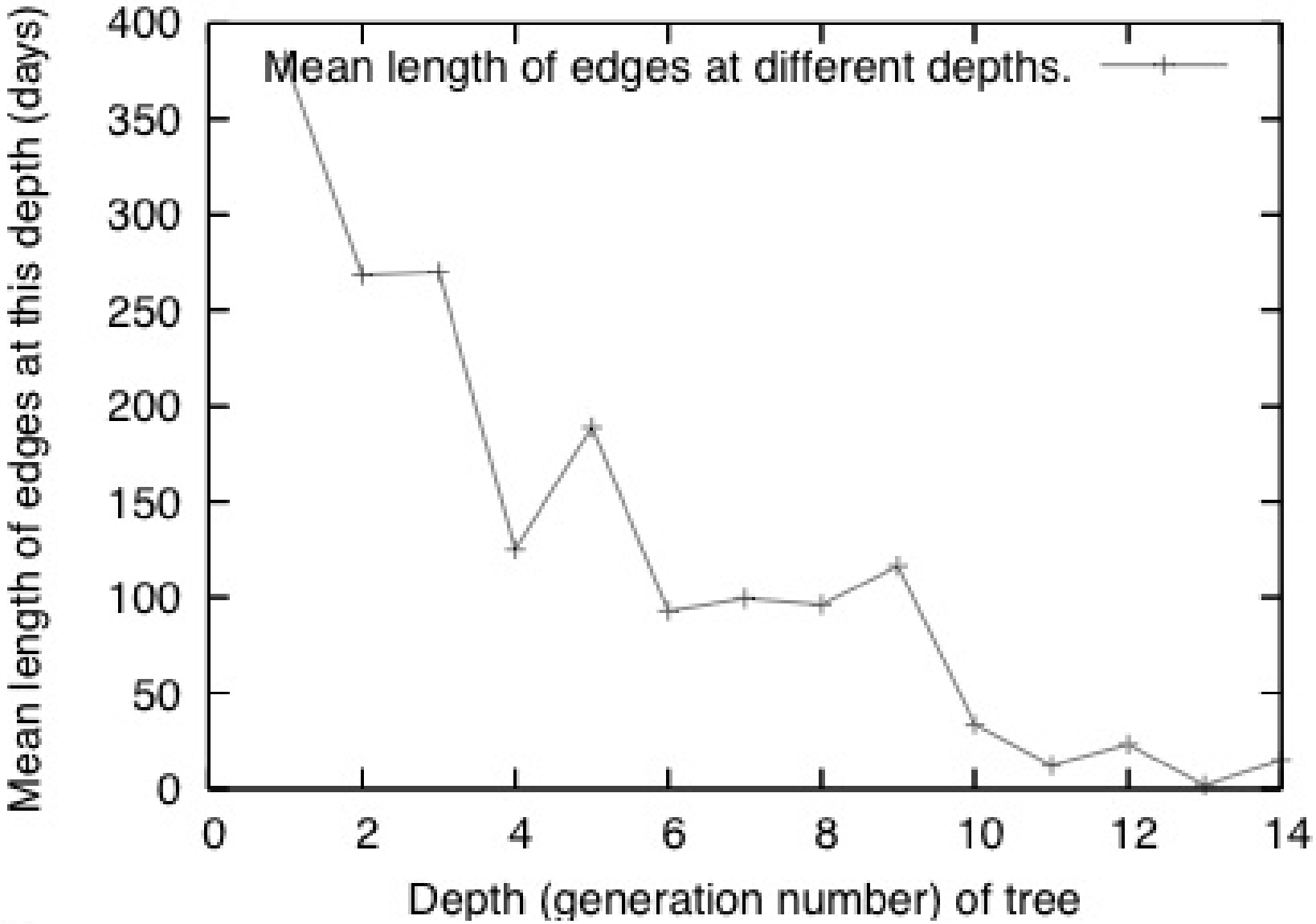
We have 669 families

Entropy metric = 1.19 with a maximum of 9

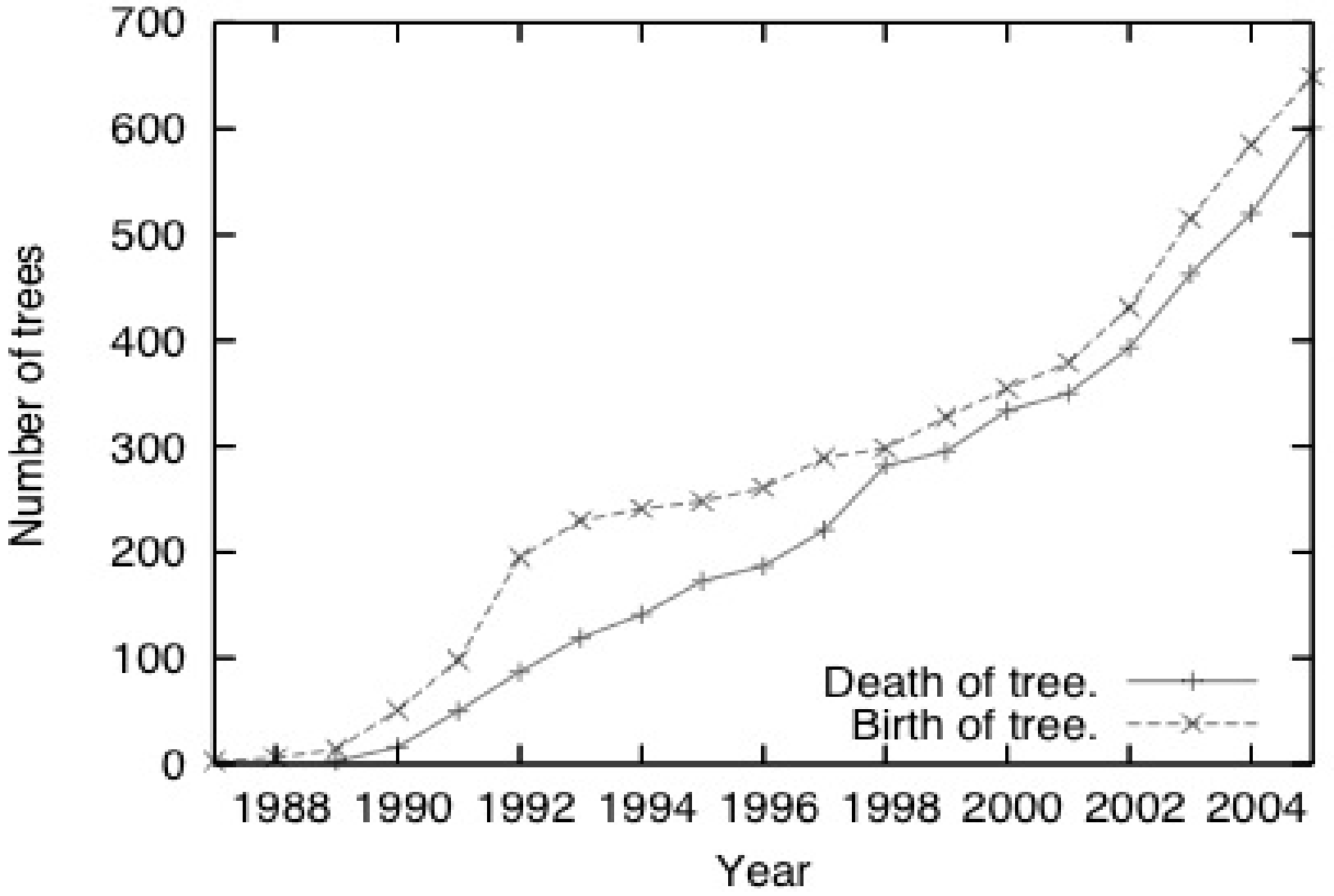
Fan-out


100
95
90
85
80
75
70
65
60
55
Cdf
0
5
10
15
20
25
30
Fanout of tree
Cdf of maximum fanout of trees.
Cdf of root fanout of trees.

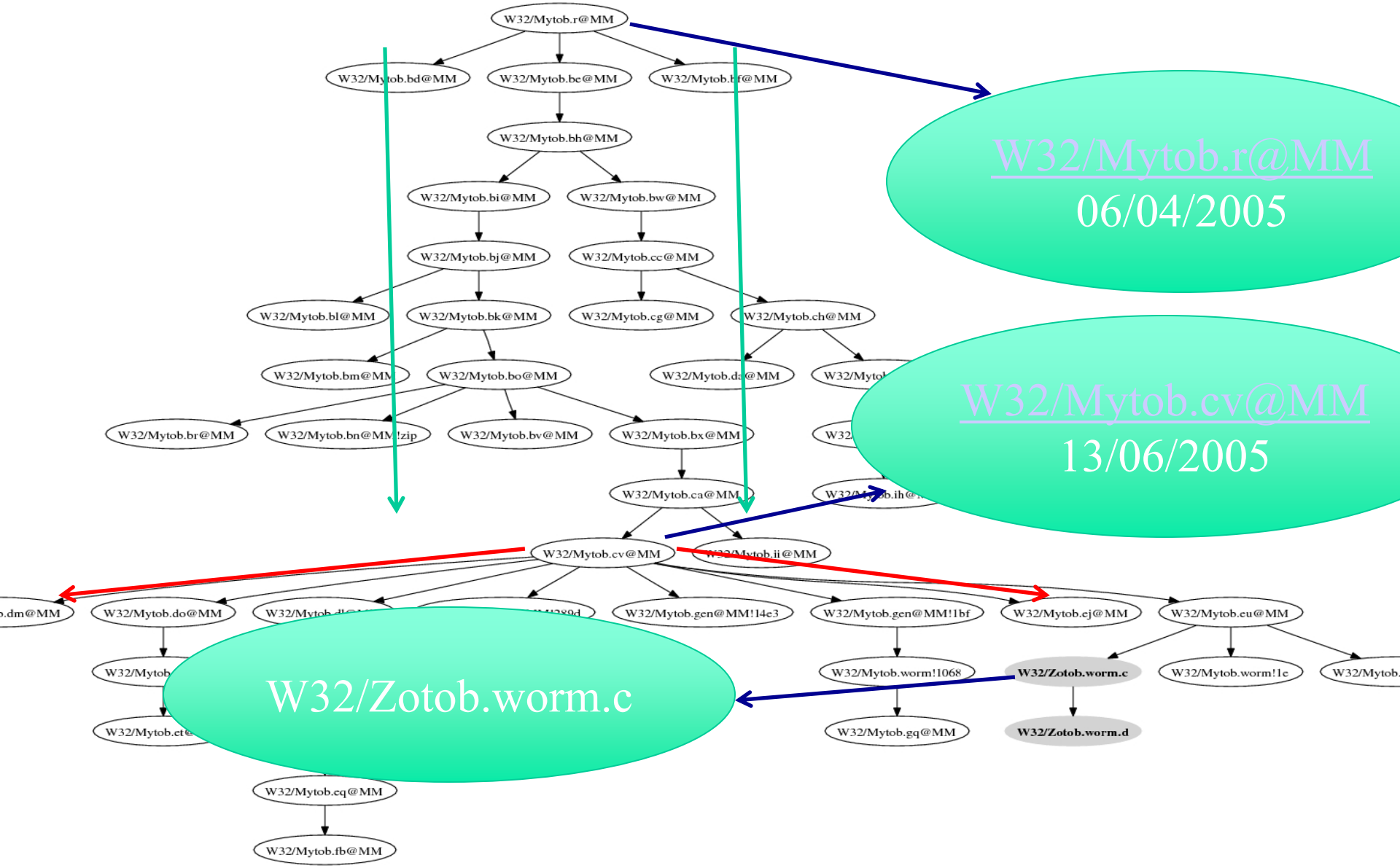
Evolution dynamics



AV vsBlackhats?



Mytob family



Validation

Use an *Entropy Metric*

McAfee named families used as reference

Entropy of a McAfee family

$$e = \sum_{1}^{k} (-f.\log(f))$$

Entropy of our output is the weighted mean ove all McAfee families

Related work

Empirical studies ofmalcode

Malware evolution

Text mining

Conclusion

Usage of metadata allows long term view

A novel method of analysis

Malware families have unique characteristics

www.cs.wisc.edu/~archit/projects/malware/

Future work

Expand analysis with other data sets

Link to other analyses e.g., call graphs

Relate to defenses e.g., selecting signatures

Thank you