Analysis of Malware Dataset

Brad Miller

Outline

- Context & Data Overview
- Infrastructure
- Analysis

CONTEXT & DATA OVERVIEW

In an Ideal World…

- An evaluation dataset would include
  - Full analysis of every file that ever appears
    - Past, Present & Future!
  - The prevalence or importance of each file
  - Gold standard label given by magical oracle

Life in an Imperfect World

- Access to data in academia is limited
- The challenge is to produce solid results despite imperfections in the data
- Our Goals: Examine the viability of a malware dataset for evaluating malware detection

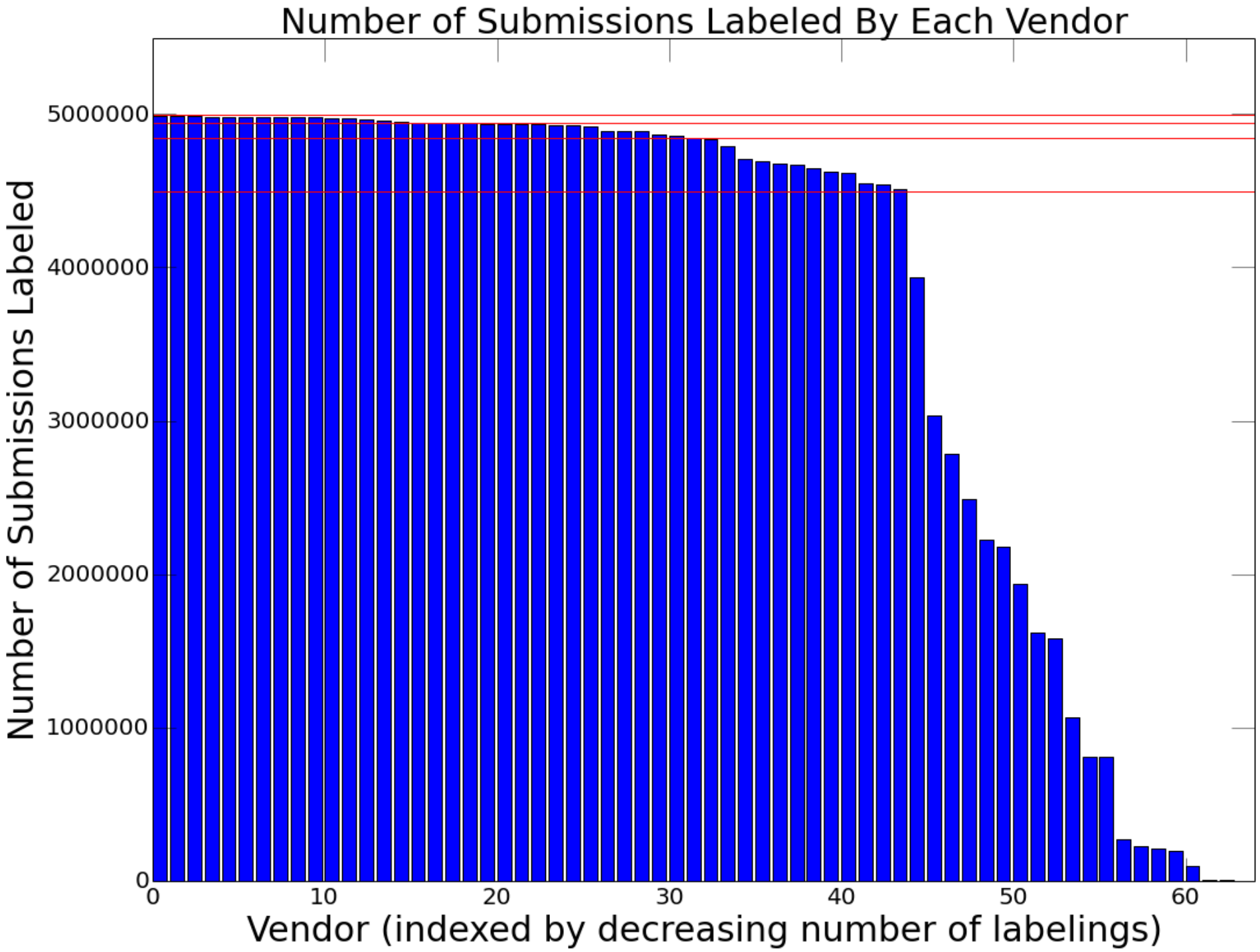
Data Origins

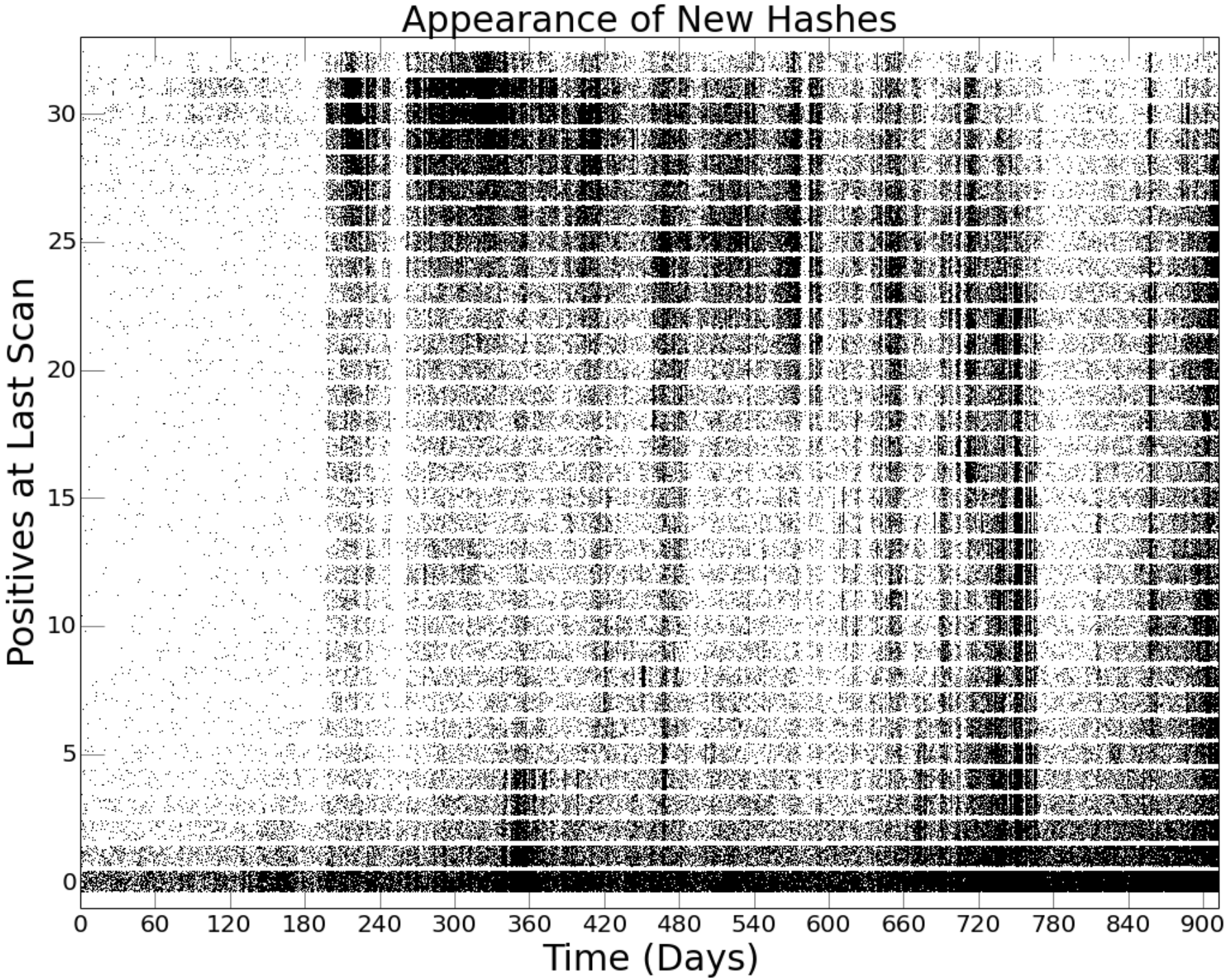
- Hashes provided by VirusTotal & AV Provider
- VirusTotal provides
  - Labeling from many AV vendors
  - Times when each hash is submitted to VirusTotal

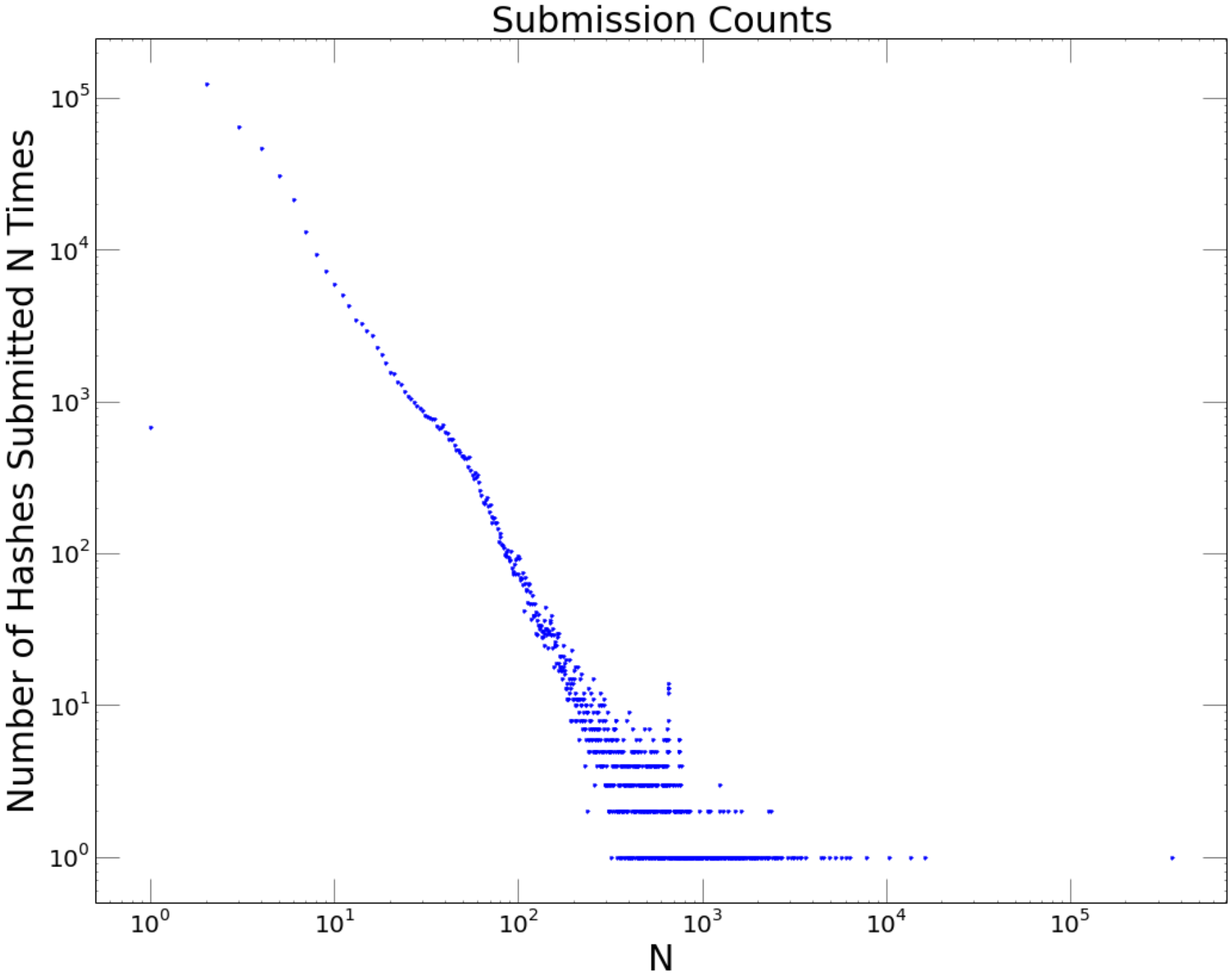
Data Overview

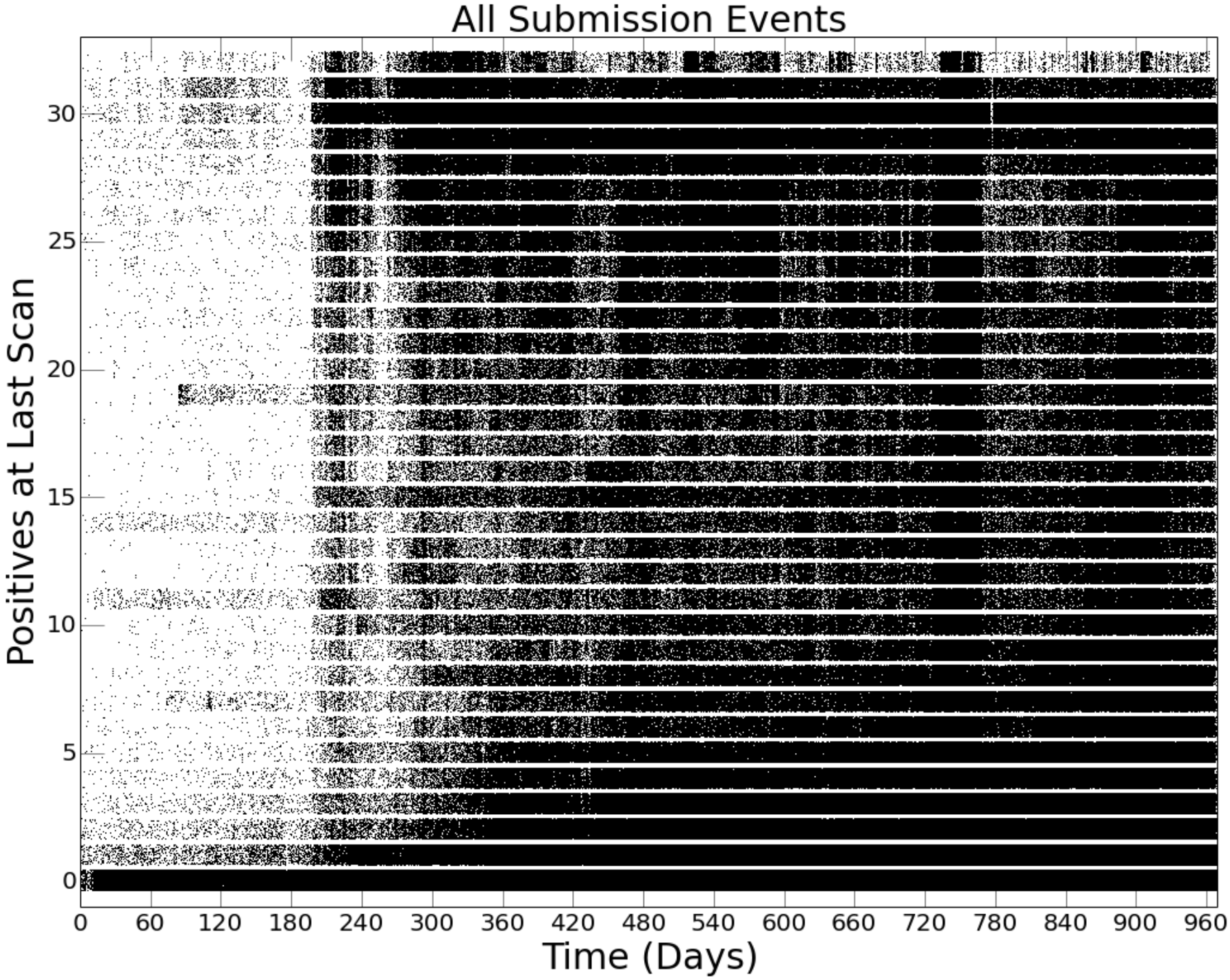
- 1.1M hashes with dynamic data available

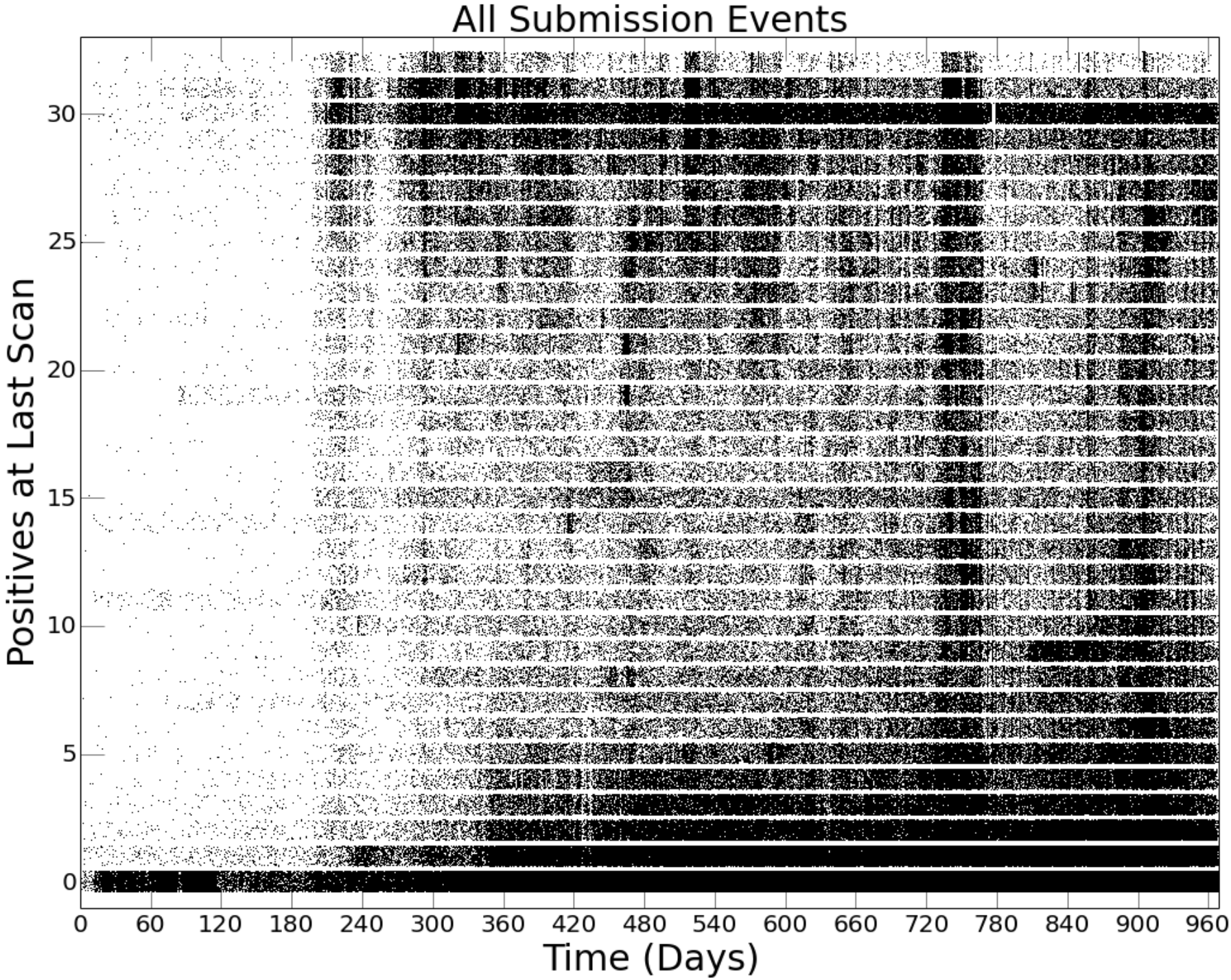
- 400K hashes with multiple submissions
  - We will focus on these today
  - Multiple submissions allows us to observe AV vendor behavior over time


Number of Submissions Labeled By Each Vendor
Number of Submissions Labeled
5000000
4000000
3000000
2000000
1000000
0
0
10
20
30
40
50
60
Vendor (indexed by decreasing number of labelings)


Appearance of New Hashes
Positives at Last Scan
0
5
10
15
20
25
30
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
Time (Days)


Submission Counts
Number of Hashes Submitted N Times
N
10^0
10^1
10^2
10^3
10^4
10^5


All Submission Events
Positives at Last Scan
0
5
10
15
20
25
30
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960
Time (Days)


All Submission Events
Positives at Last Scan
0
5
10
15
20
25
30
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960
Time (Days)

INFRASTRUCTURE

Analysis Stack

matplotlib: Data Visualization

iPython Notebook:
Interactive Computation

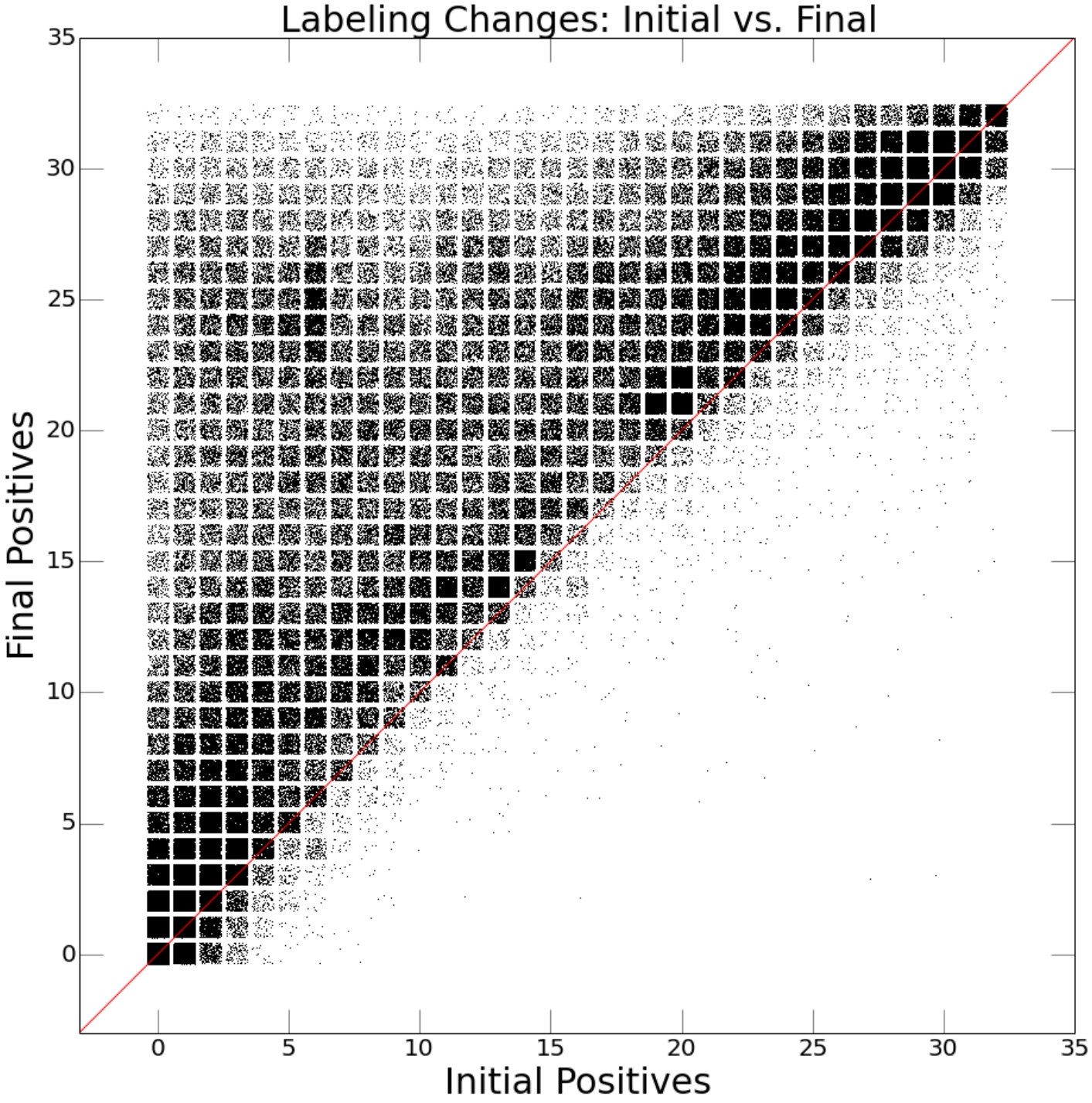
Spark: Distributed (fast) Computation

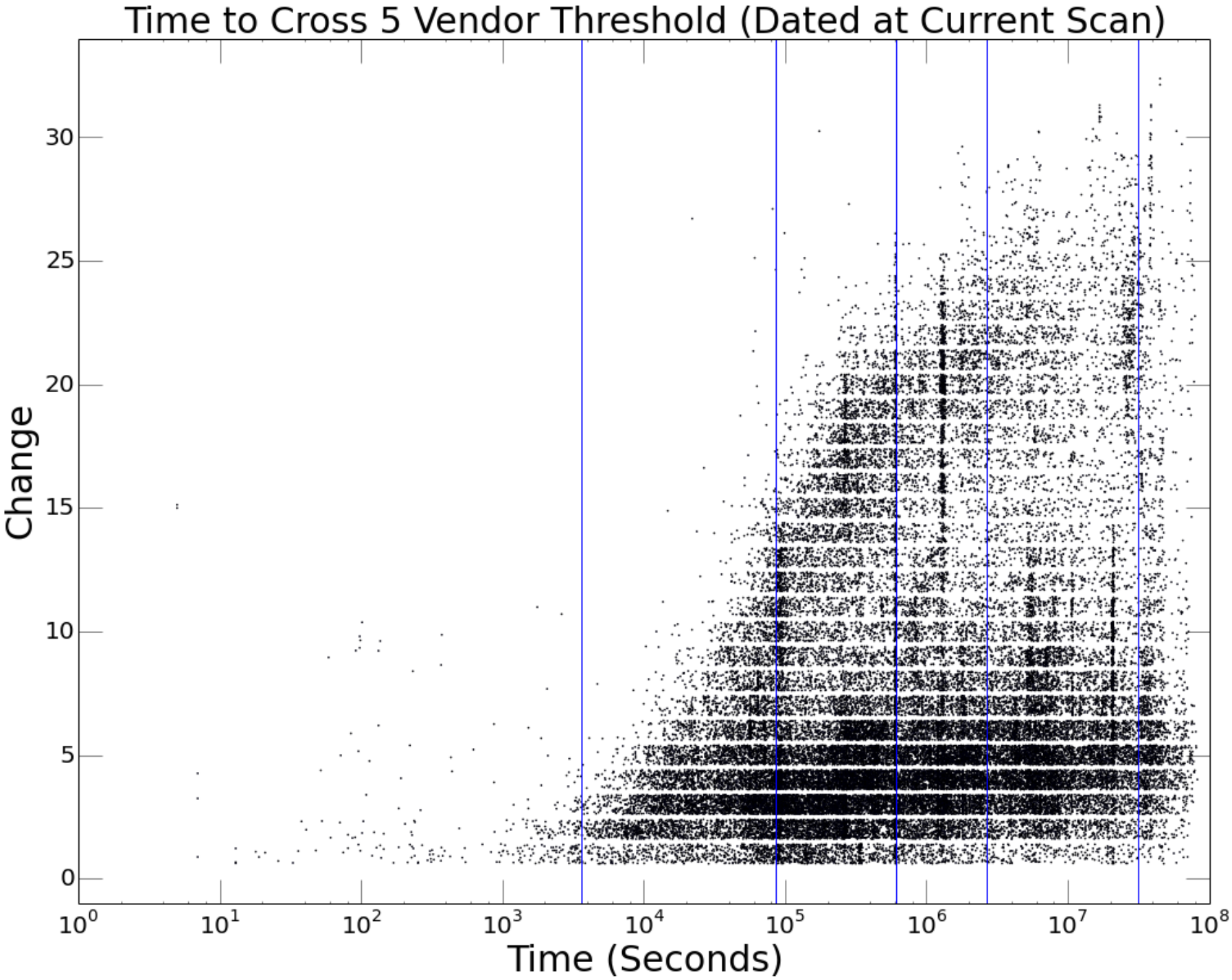
HDFS: Distributed Storage

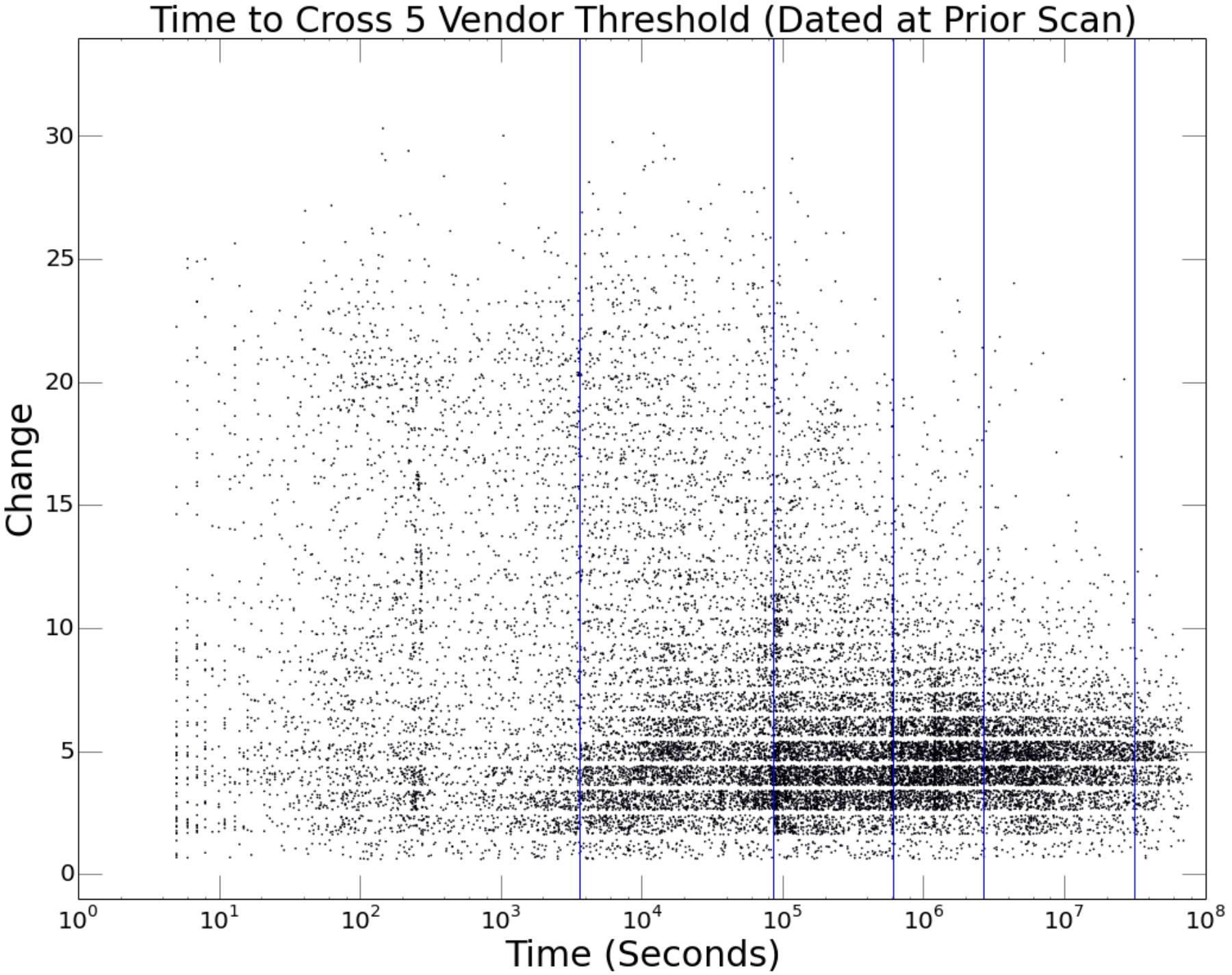
ANALYSIS

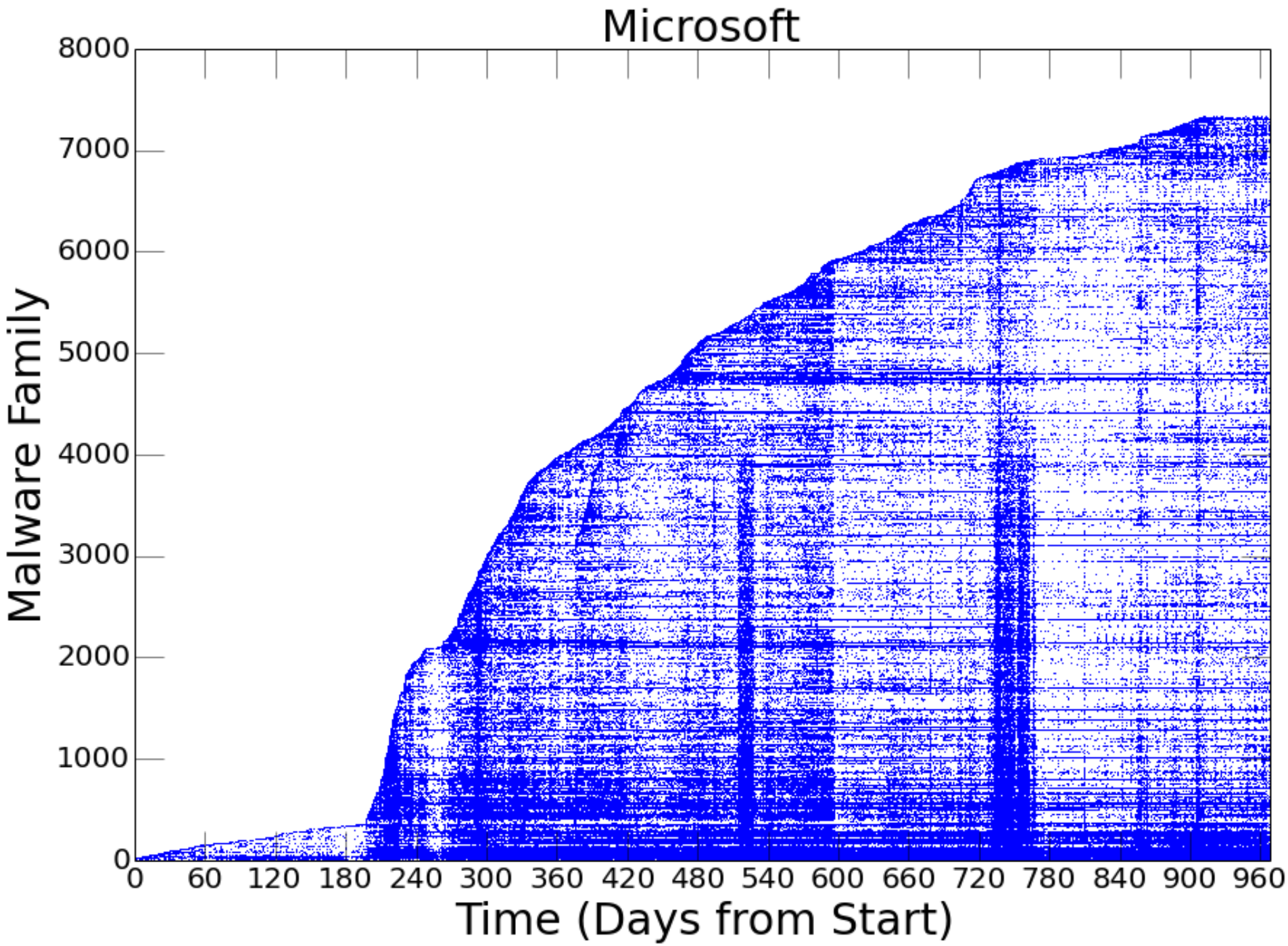
Investigative Targets

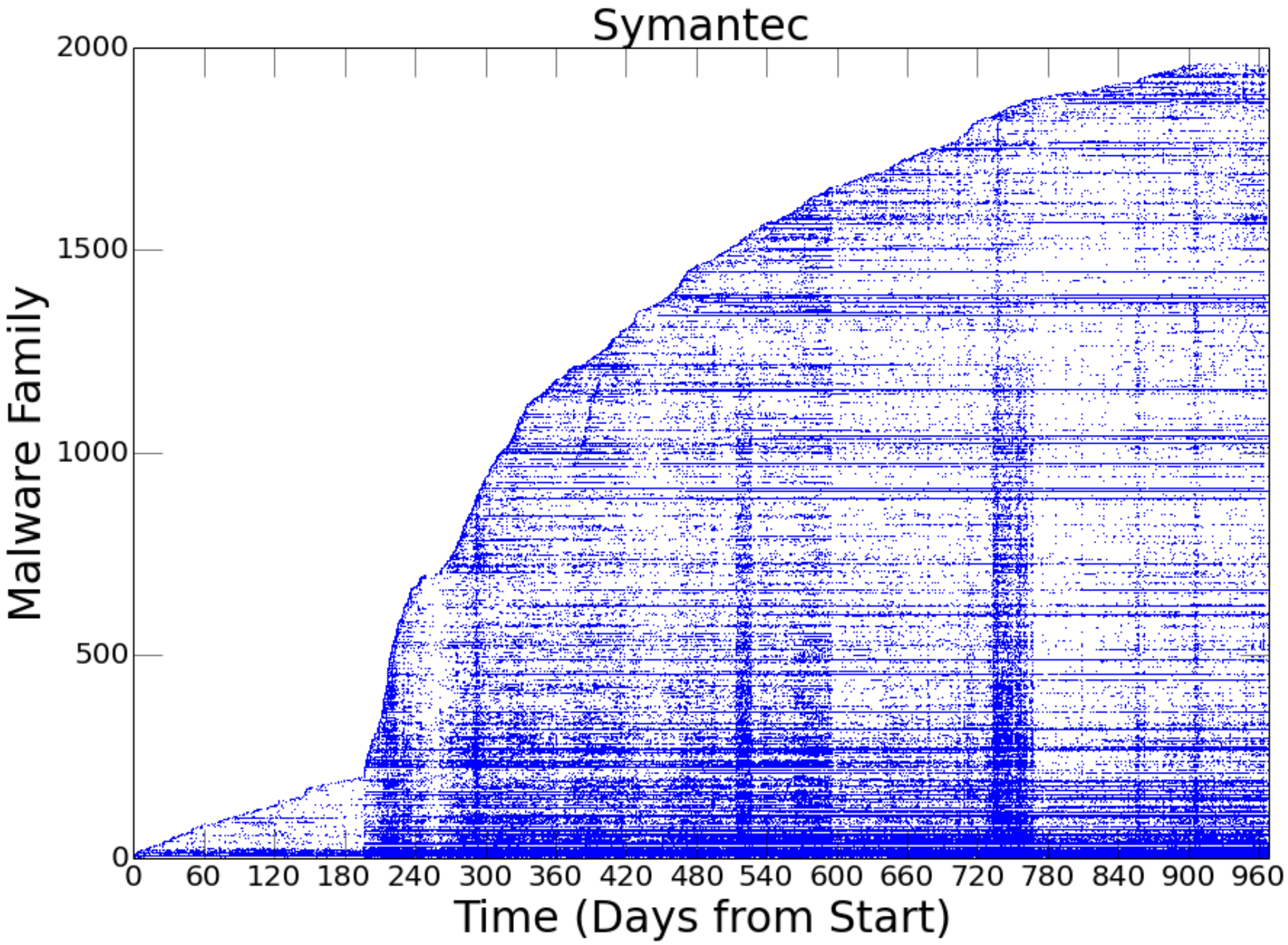
- Is this a good dataset to use for evaluating a malware detection system?
  - Keep in mind the properties of the ideal dataset
- How should we label samples?
  - Vendors may change their minds over time
  - Some vendors may be better than others

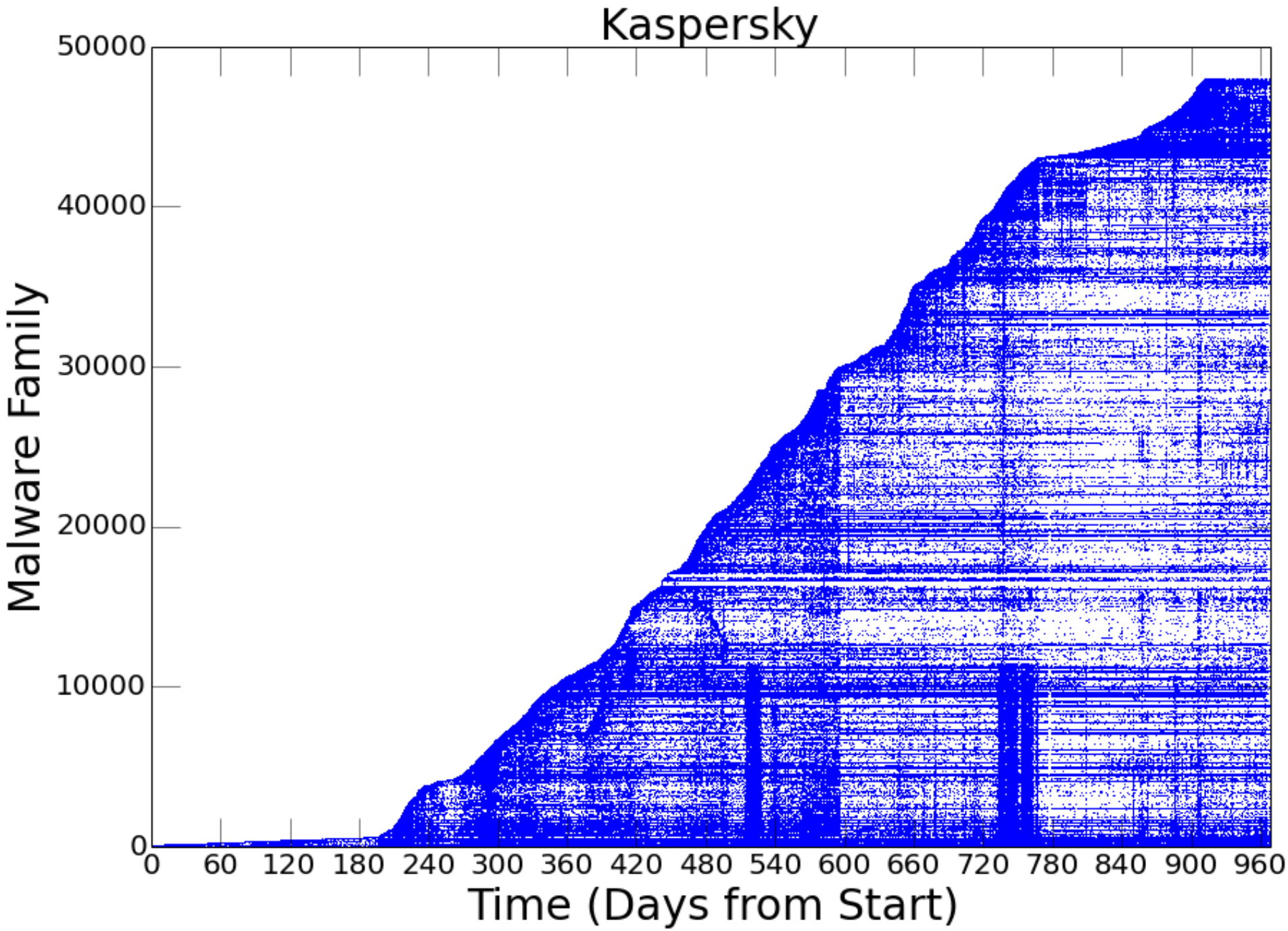

Labeling Changes: Initial vs. Final
Final Positives
Initial Positives
0
5
10
15
20
25
30
35

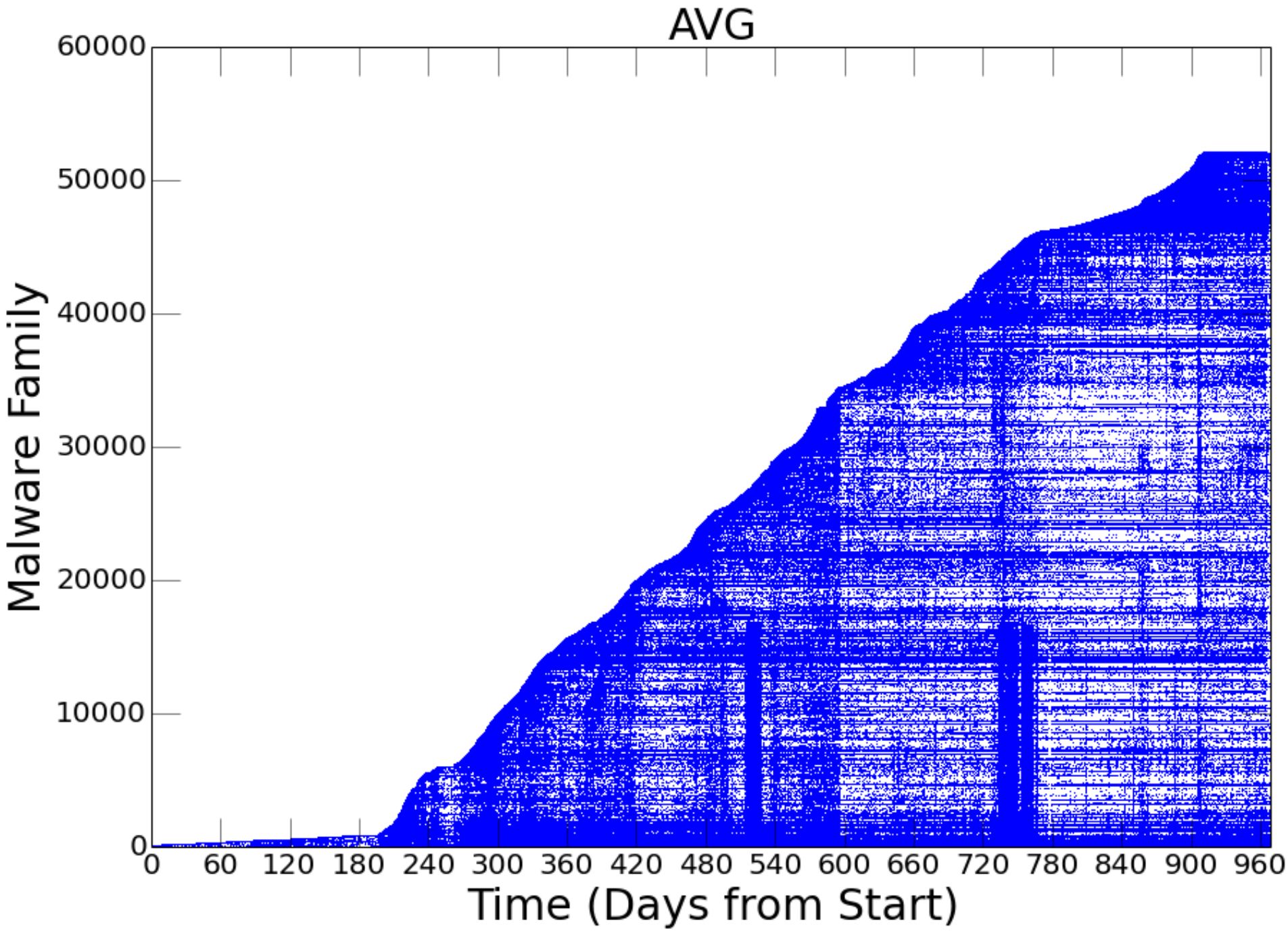

Time to Cross 5 Vendor Threshold (Dated at Current Scan)
Change
0
5
10
15
20
25
30
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7
10^8
Time (Seconds)

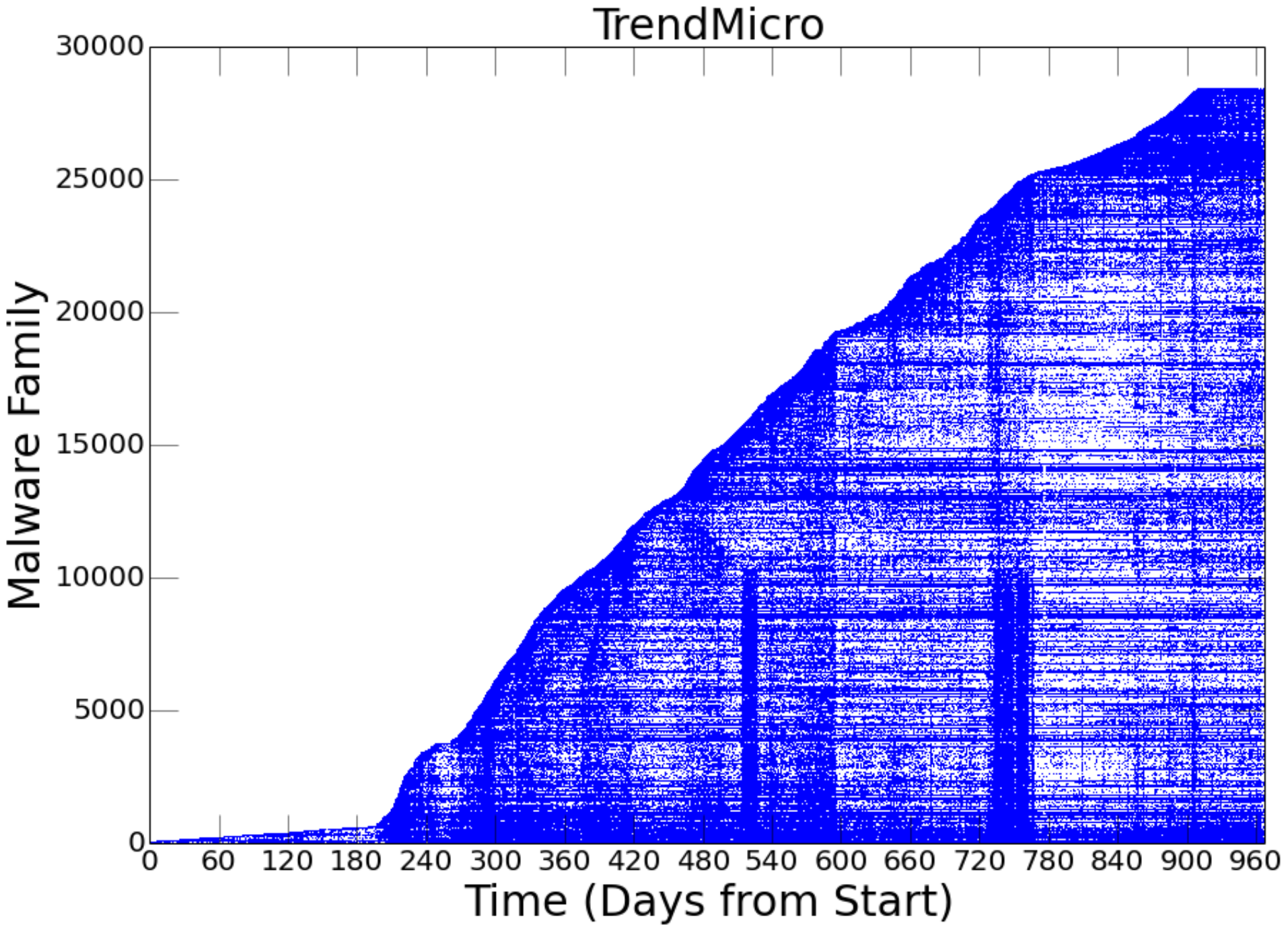

Time to Cross 5 Vendor Threshold (Dated at Prior Scan)
Change
30
25
20
15
10
5
0
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7
10^8
Time (Seconds)

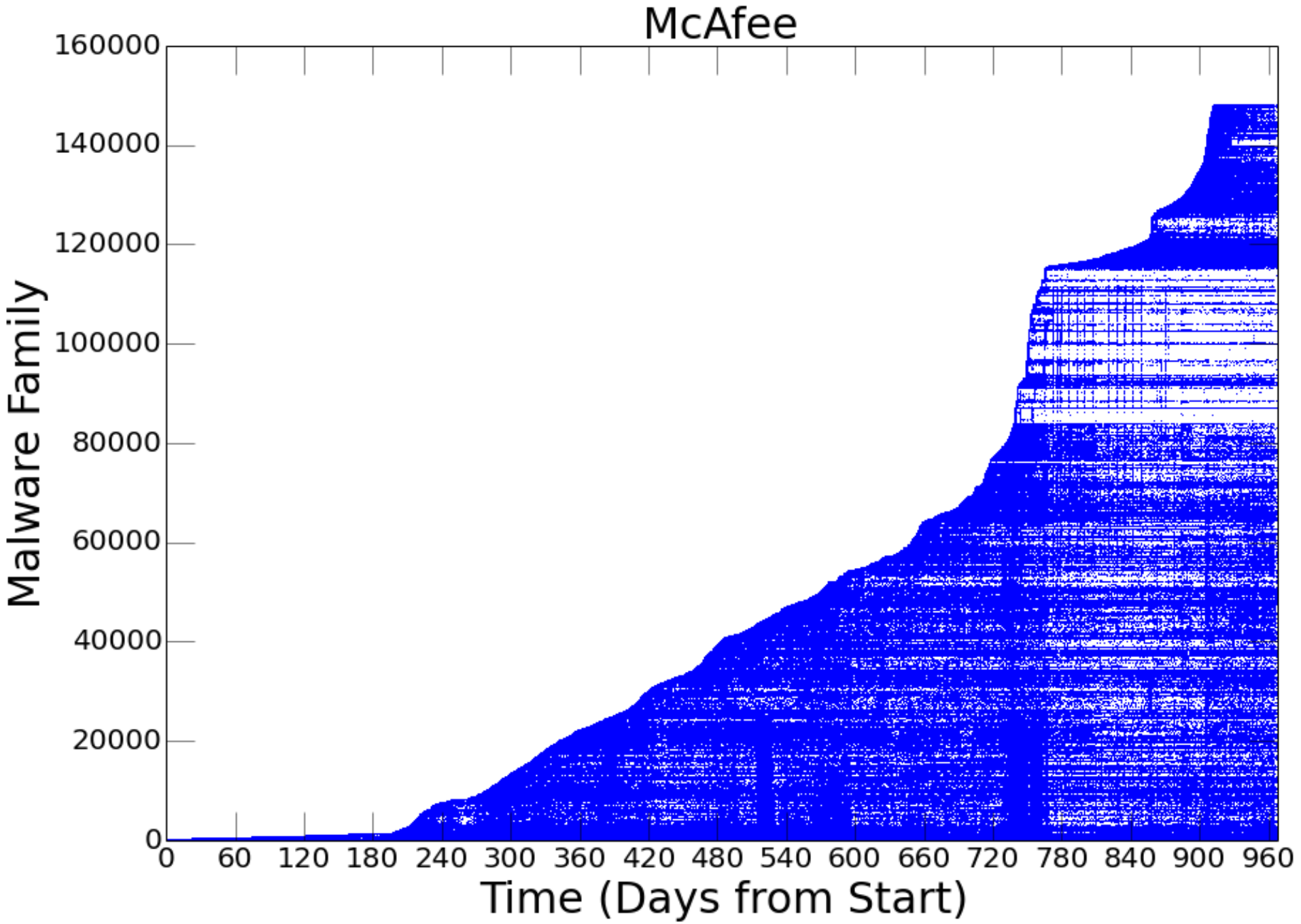

Microsoft
Malware Family
Time (Days from Start)
0
1000
2000
3000
4000
5000
6000
7000
8000
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960

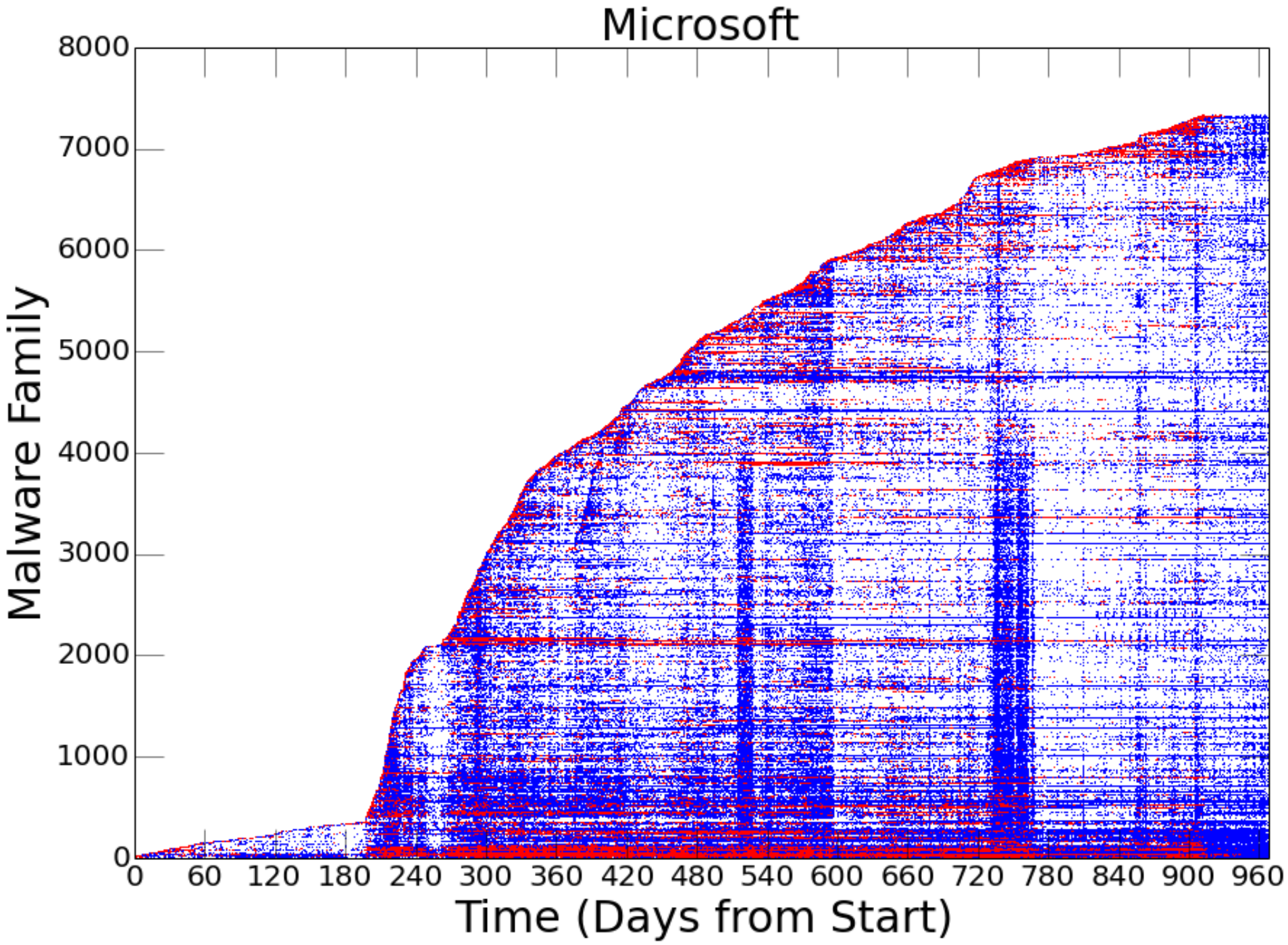

Symantec
Malware Family
2000
1500
1000
500
0
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960
Time (Days from Start)

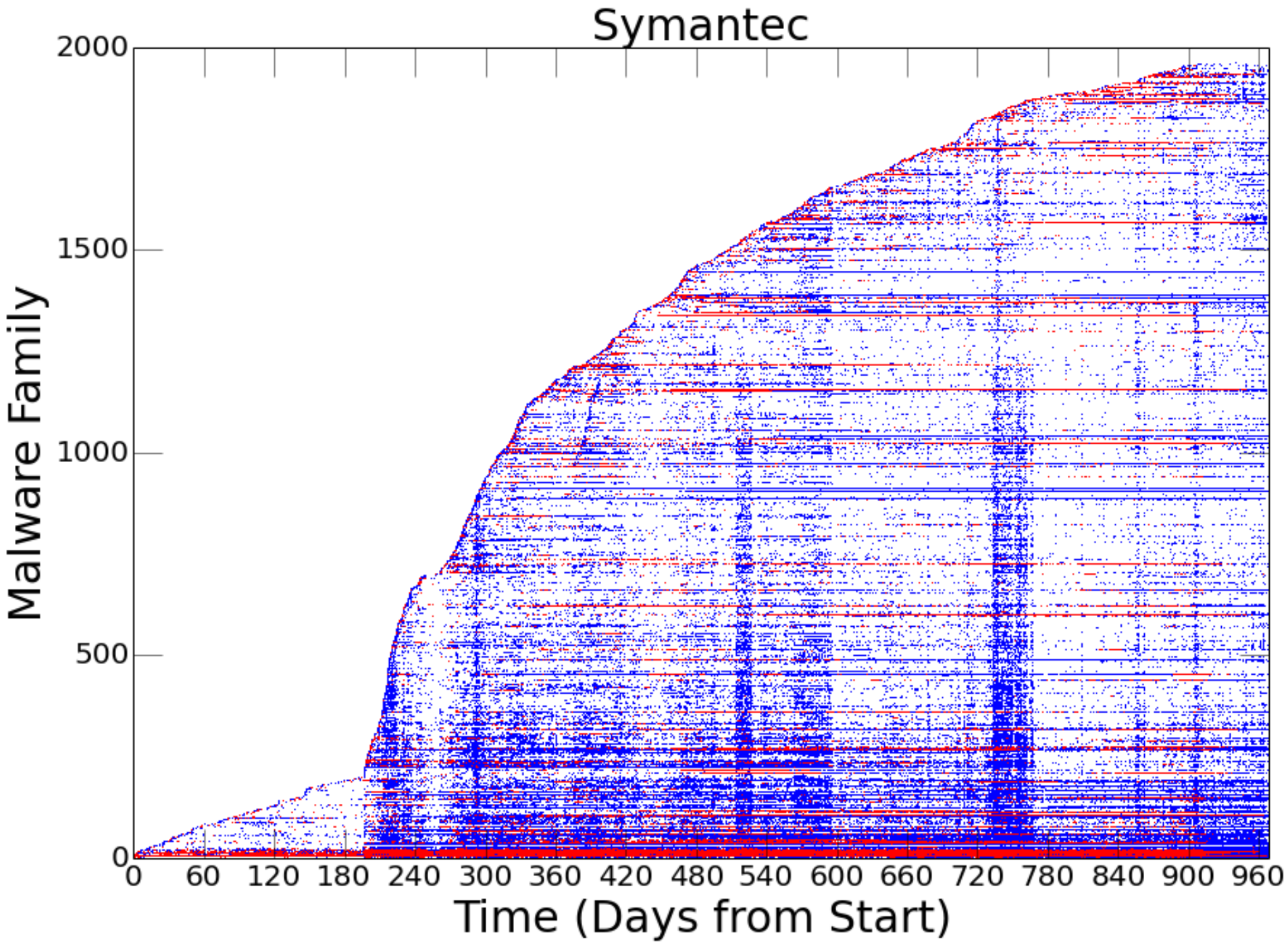

Kaspersky
50000
40000
30000
20000
10000
0
Malware Family
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960
Time (Days from Start)

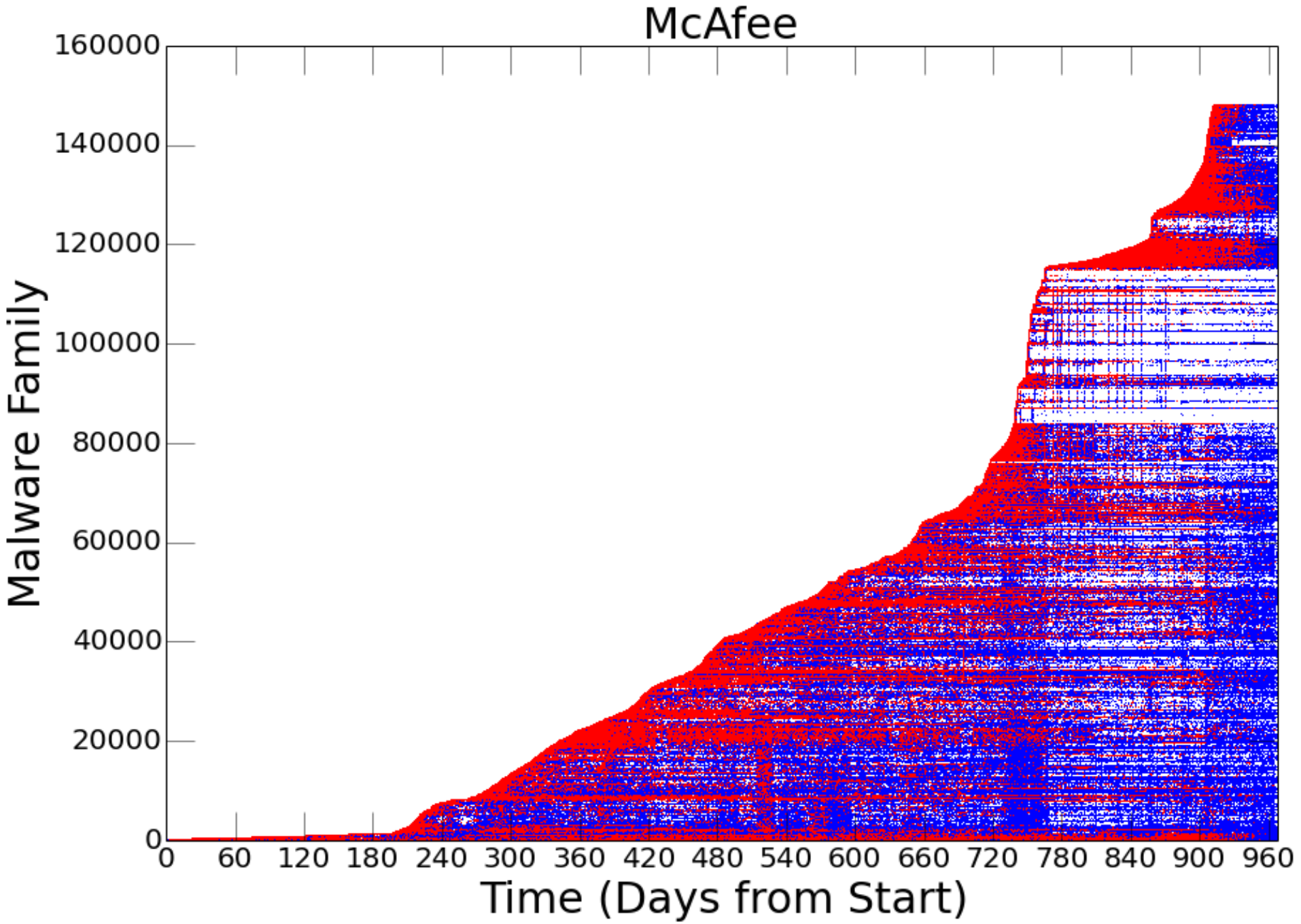

AVG
60000
50000
40000
30000
20000
10000
0
Malware Family
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960
Time (Days from Start)

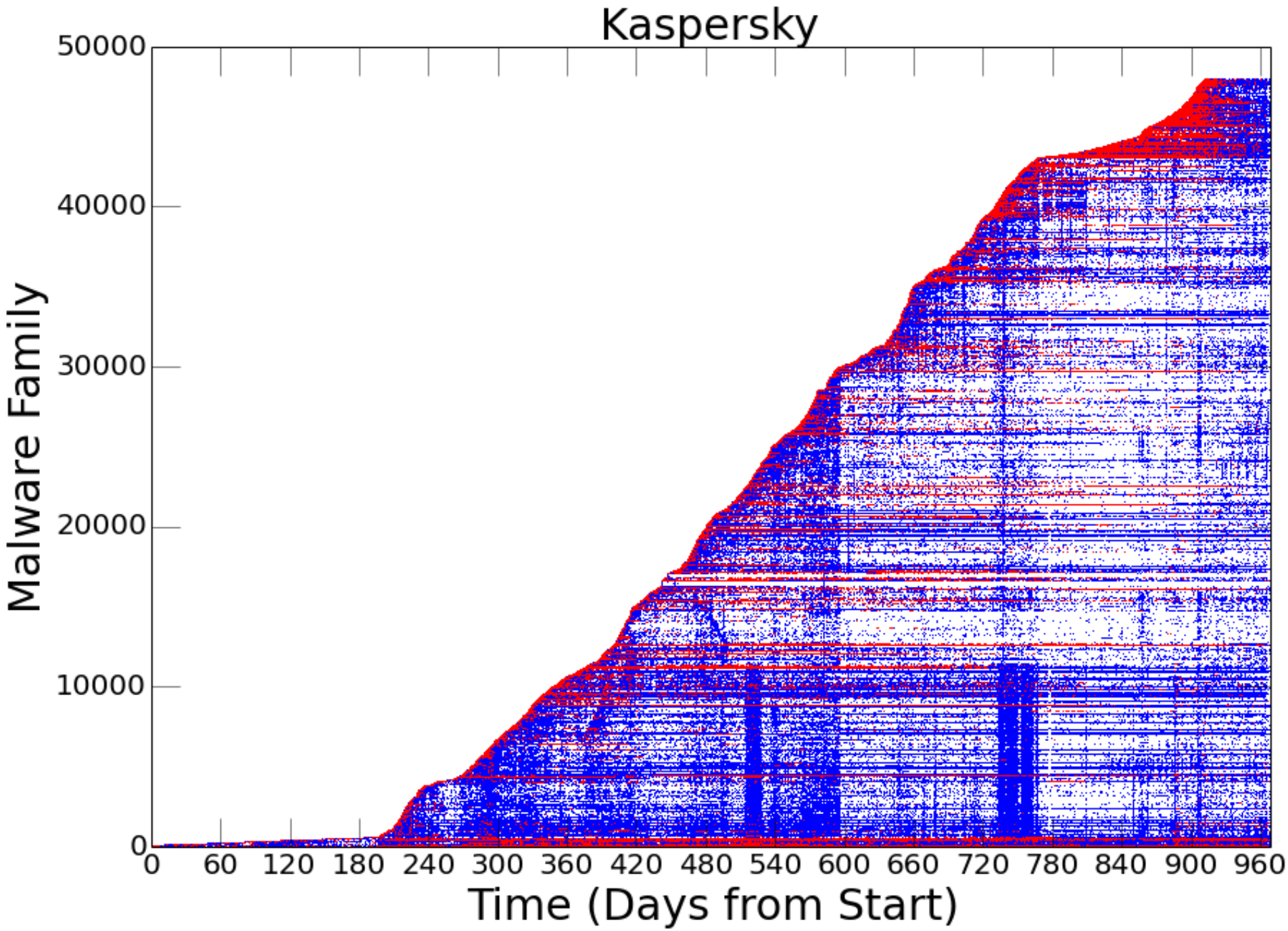

TrendMicro
Malware Family
Time (Days from Start)
0
5000
10000
15000
20000
25000
30000
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960

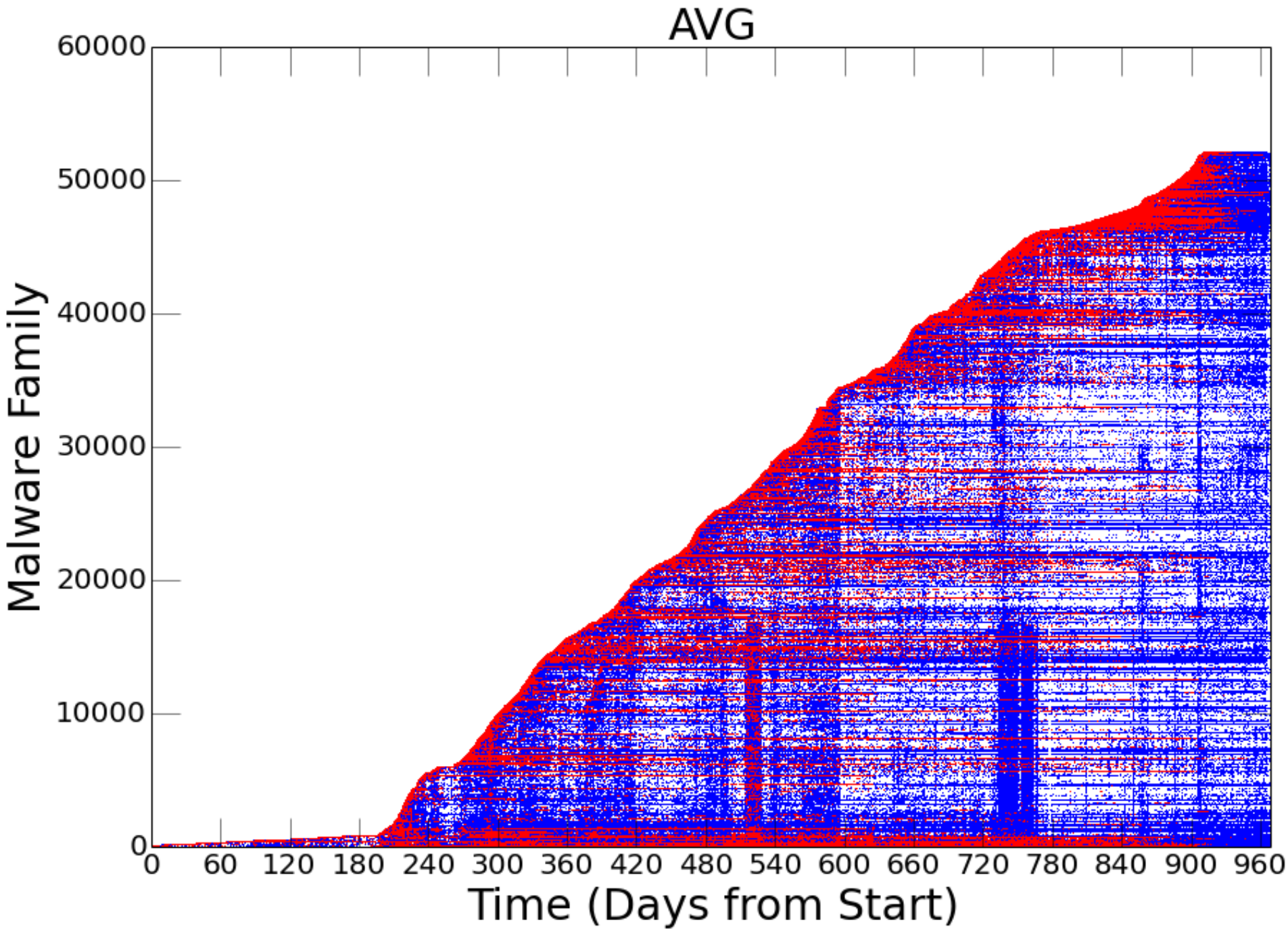

McAfee
160000
140000
120000
100000
80000
60000
40000
20000
0
Malware Family
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960
Time (Days from Start)

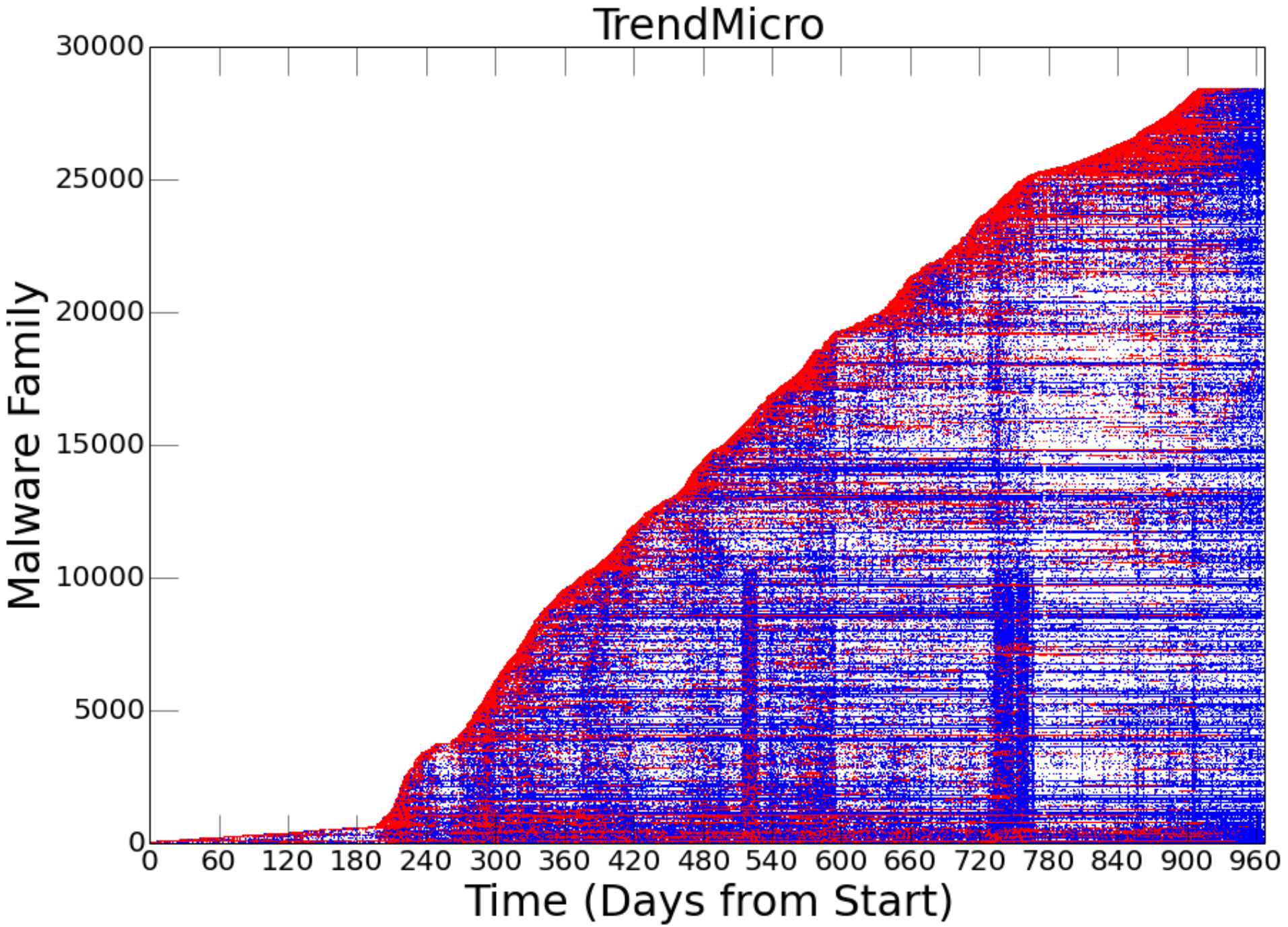

Microsoft
Malware Family
Time (Days from Start)
0
1000
2000
3000
4000
5000
6000
7000
8000
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960

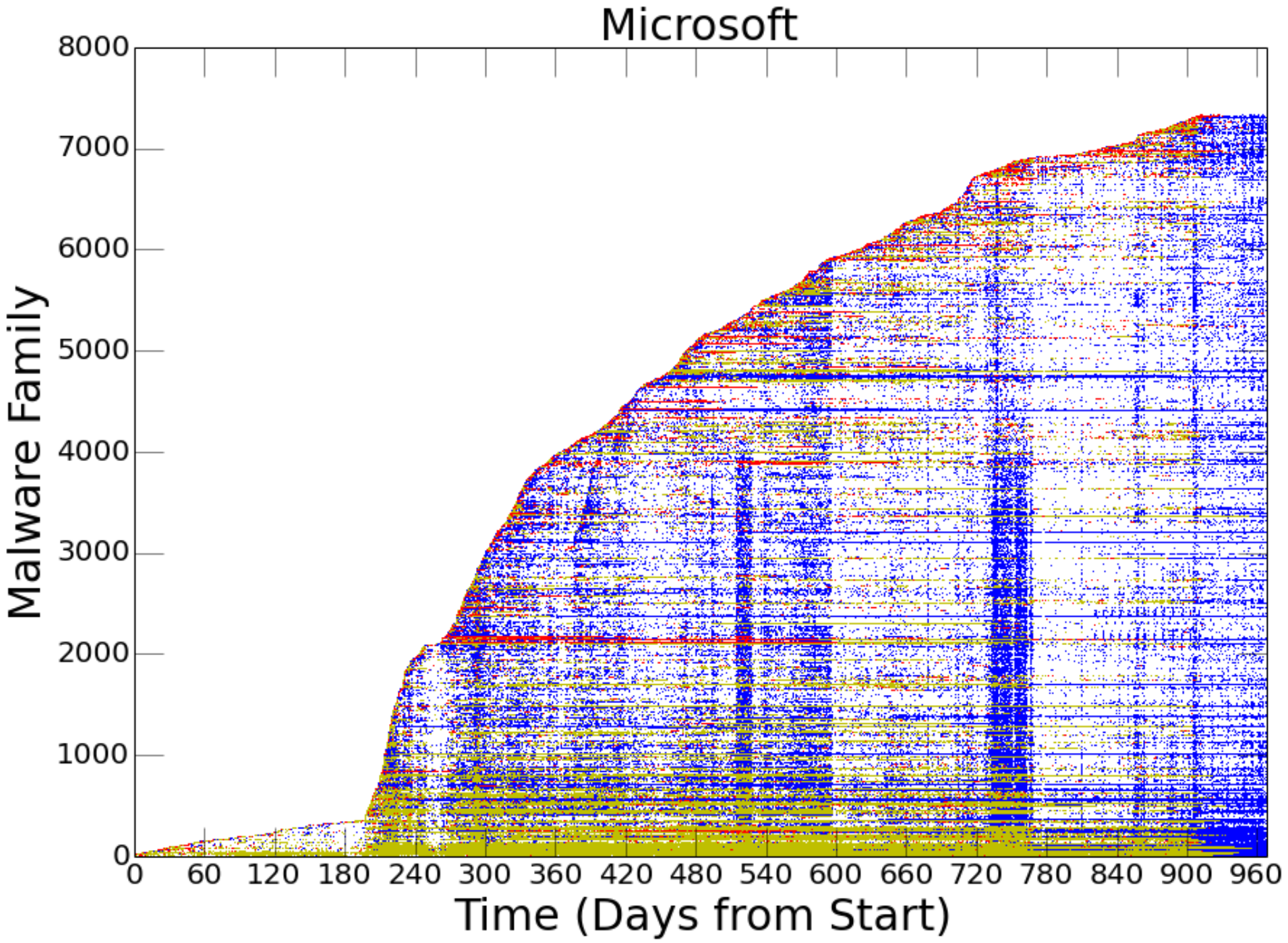

Symantec
2000
1500
1000
500
0
Malware Family
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960
Time (Days from Start)

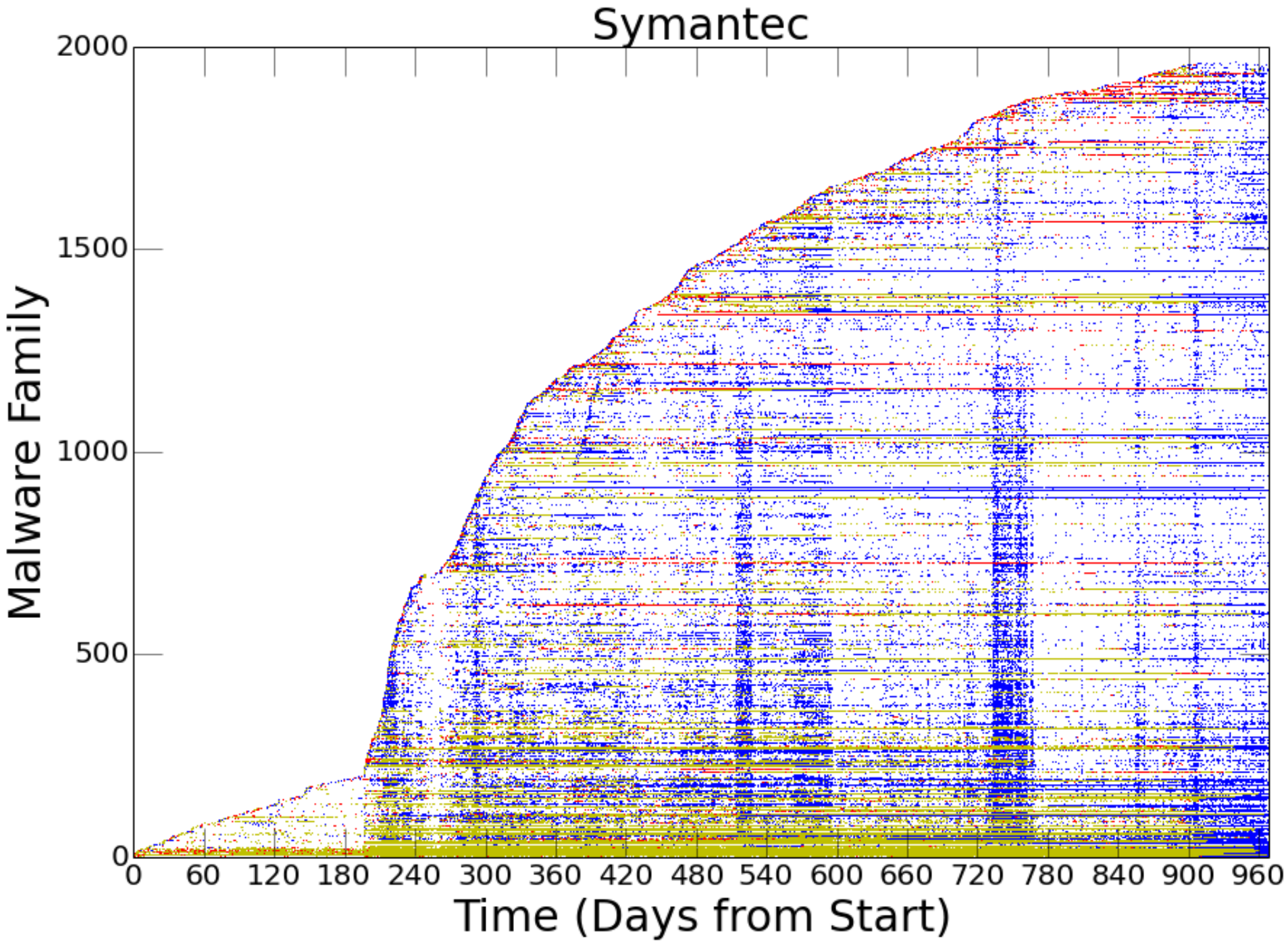

McAfee
160000
140000
120000
100000
80000
60000
40000
20000
0
Malware Family
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960
Time (Days from Start)

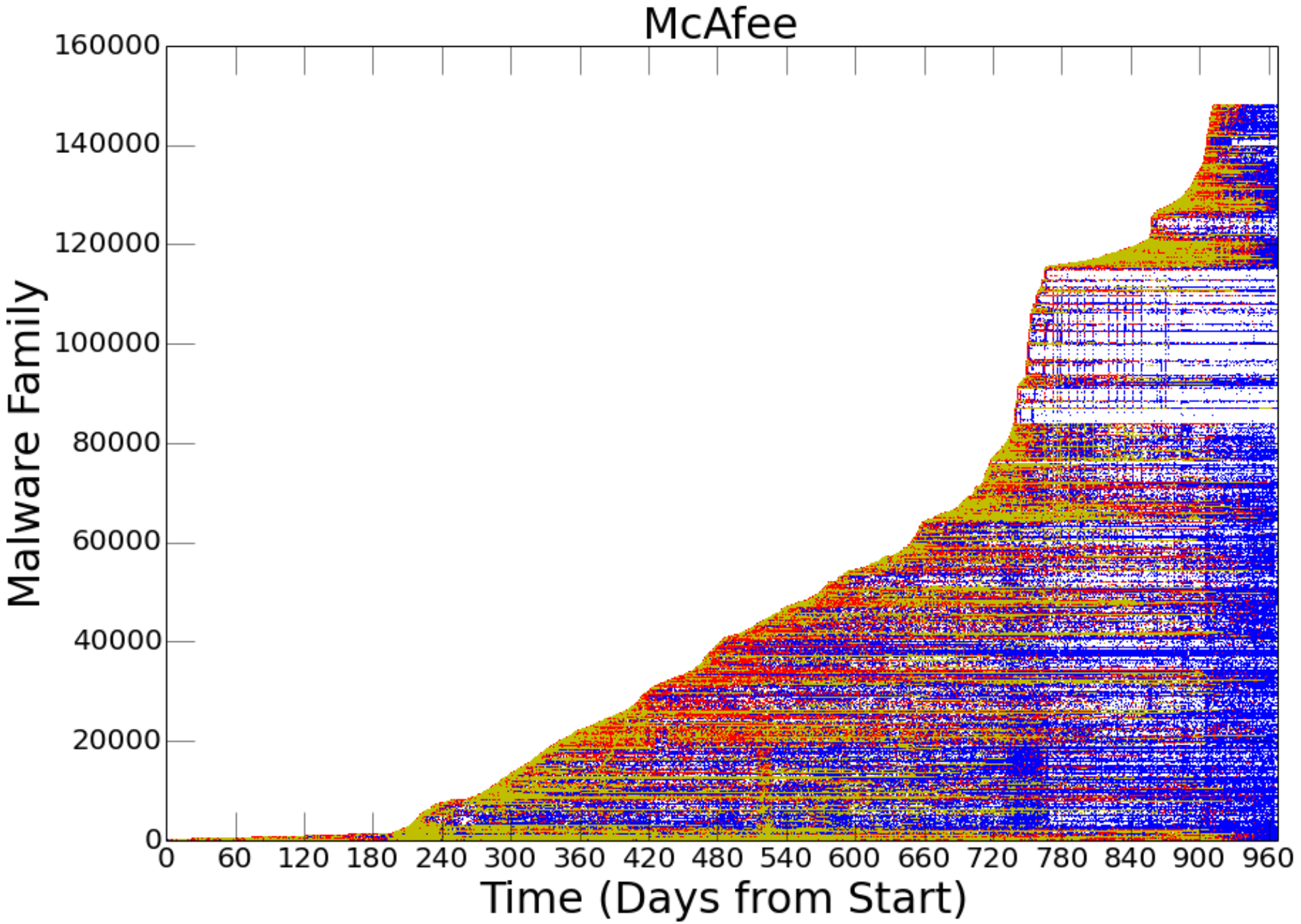

Kaspersky
Malware Family
Time (Days from Start)
0
10000
20000
30000
40000
50000
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960

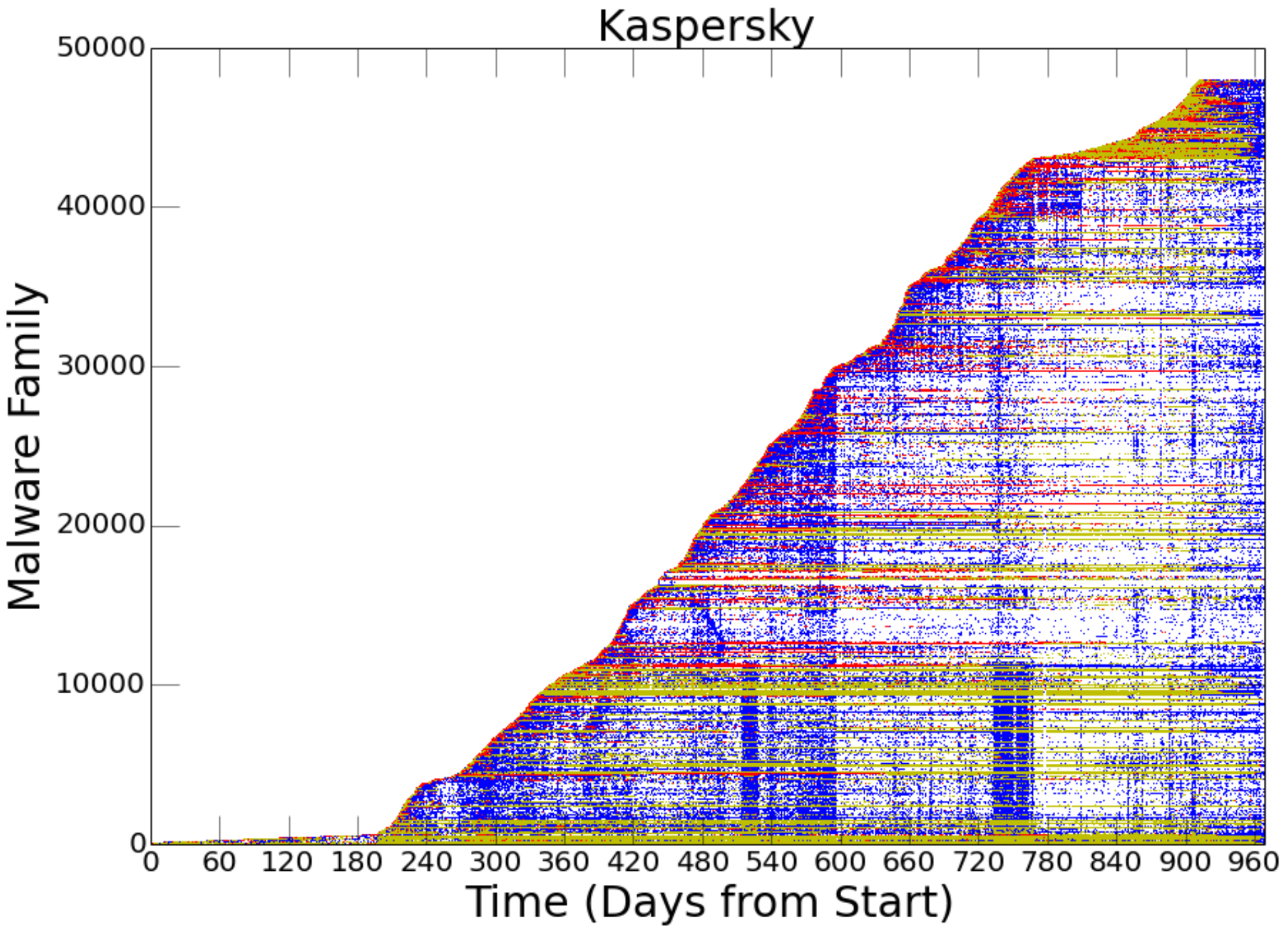

AVG
60000
50000
40000
30000
20000
10000
0
Malware Family
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960
Time (Days from Start)

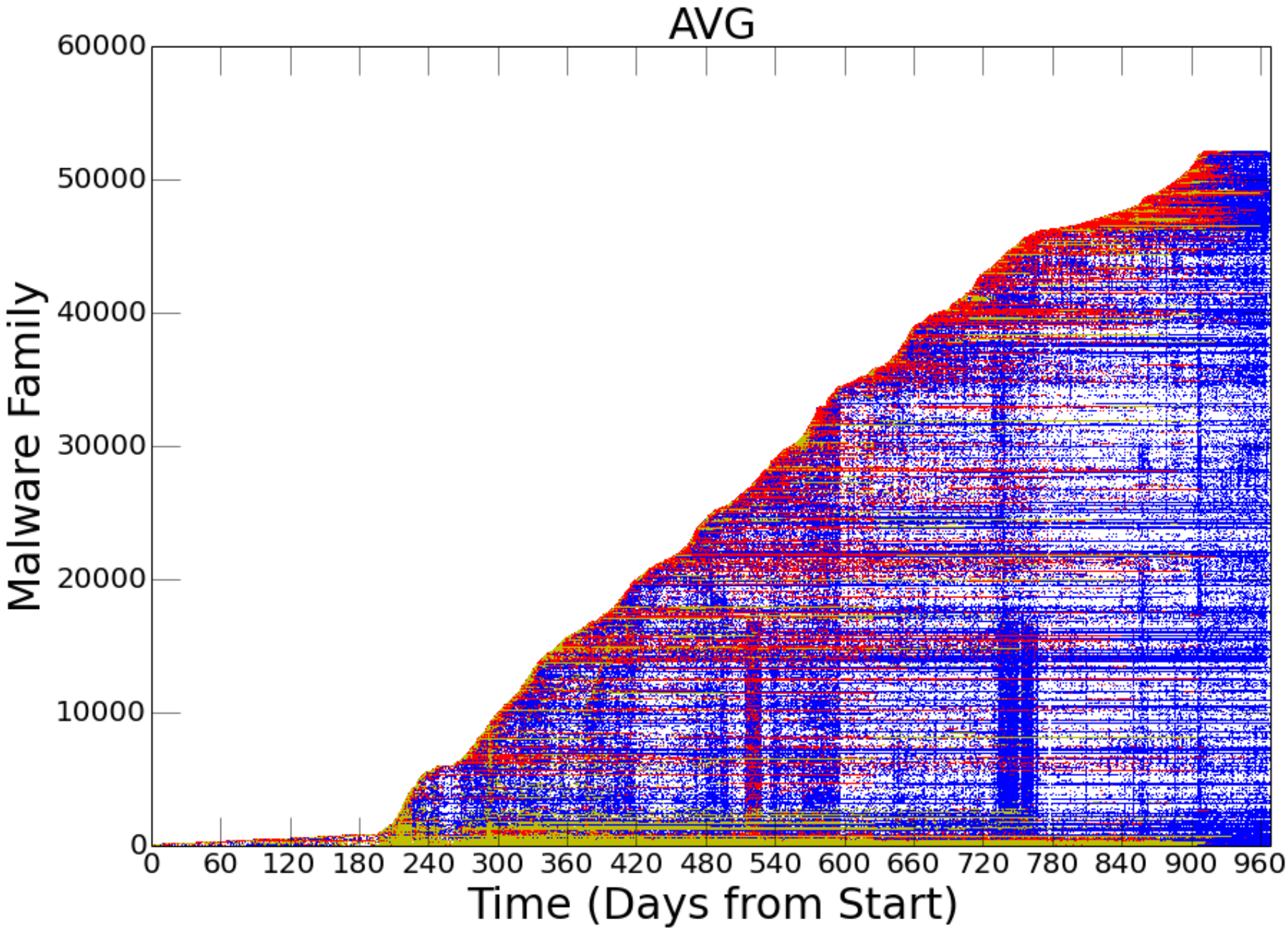

TrendMicro
Malware Family
Time (Days from Start)
0
5000
10000
15000
20000
25000
30000
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960

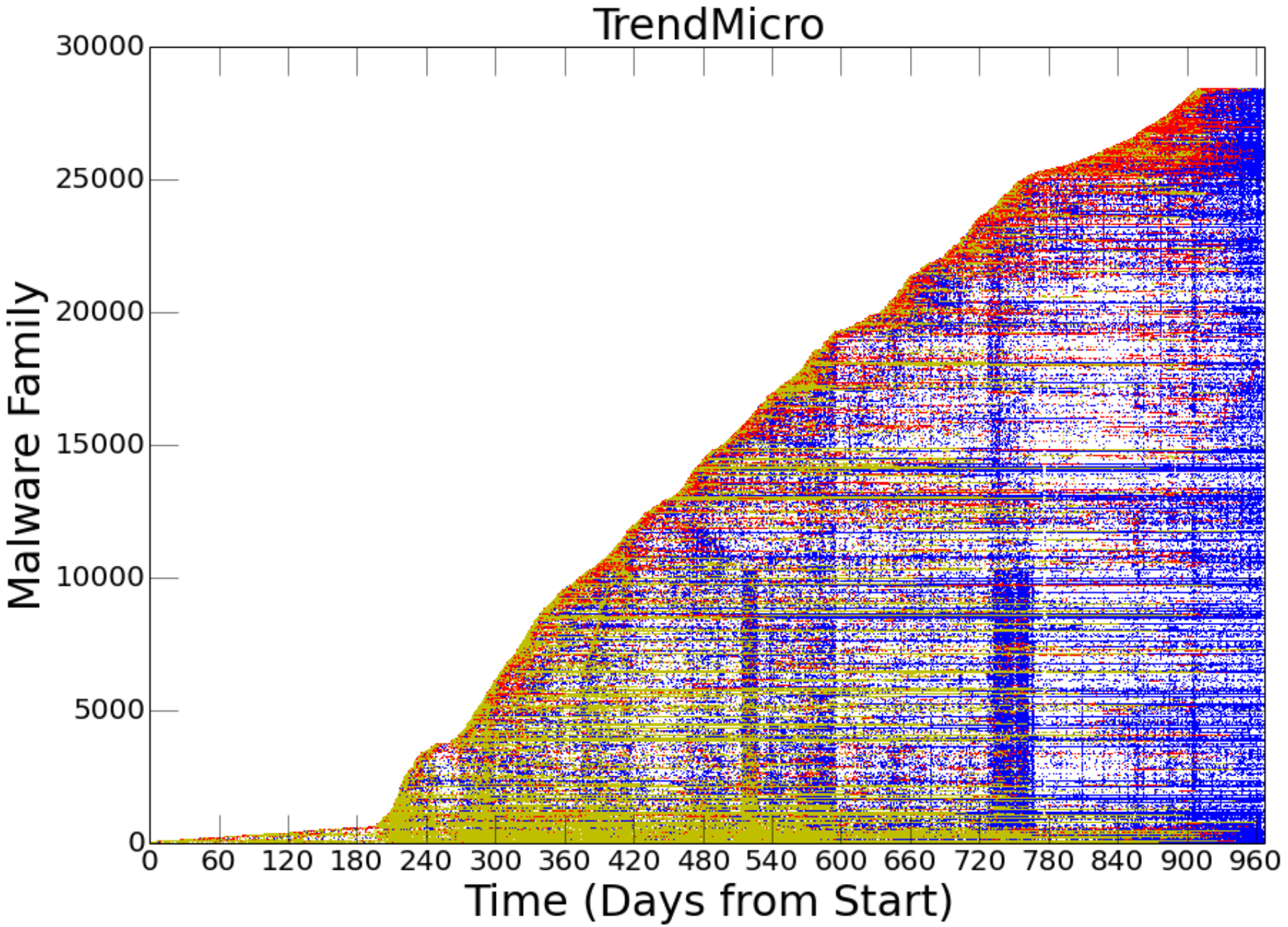

Microsoft
Malware Family
8000
7000
6000
5000
4000
3000
2000
1000
0
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960
Time (Days from Start)


Symantec
Malware Family
0
500
1000
1500
2000
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960
Time (Days from Start)


McAfee
160000
140000
120000
100000
80000
60000
40000
20000
0
Malware Family
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960
Time (Days from Start)


Kaspersky
Malware Family
50000
40000
30000
20000
10000
0
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960
Time (Days from Start)


AVG
60000
50000
40000
30000
20000
10000
0
Malware Family
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960
Time (Days from Start)


TrendMicro
Malware Family
Time (Days from Start)
0
5000
10000
15000
20000
25000
30000
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960