# 从恶意代码检测分析体系的下一步进化方向
# 谈恶意代码分类与聚类的研究进展

唐 勇

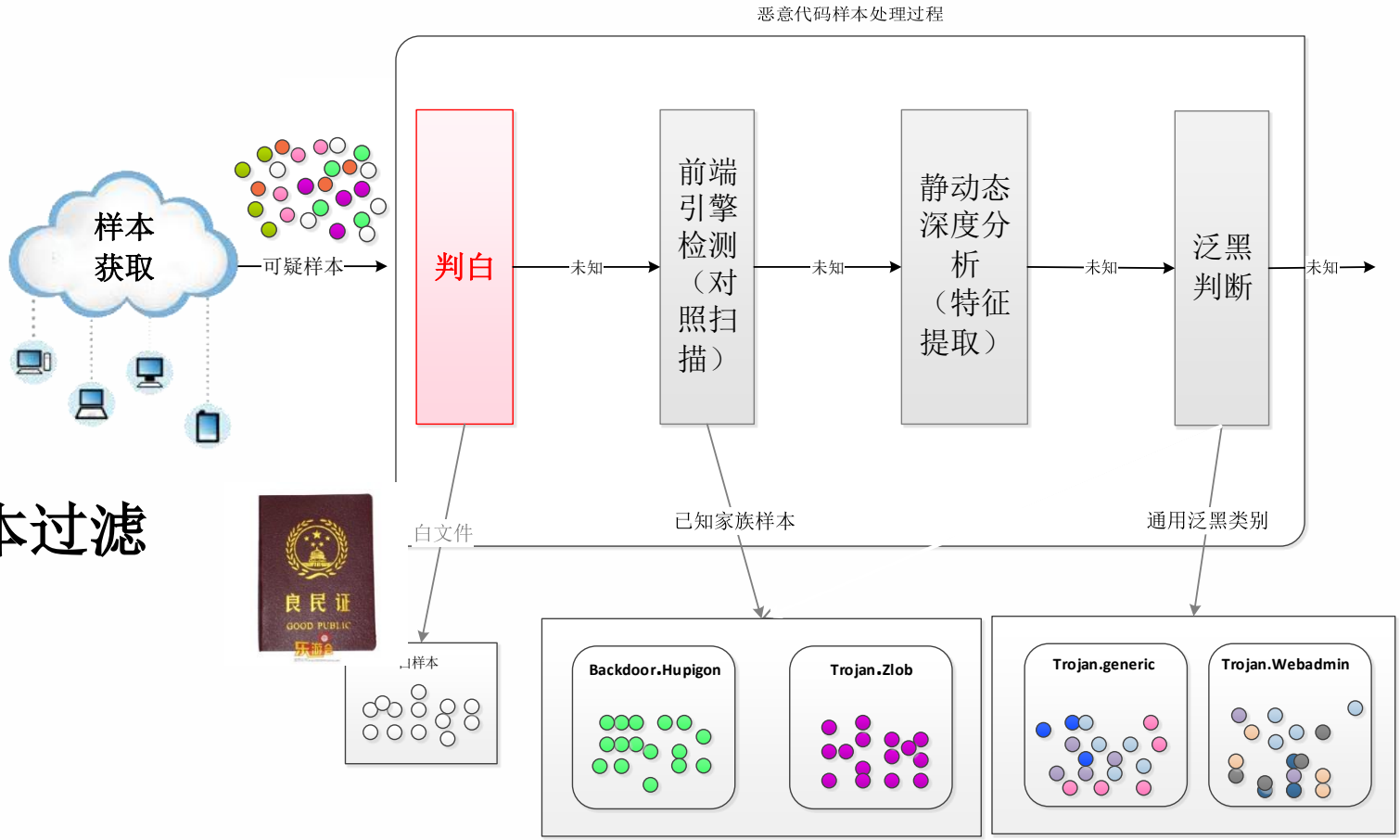
2015年1月21日

国防科学技术大学计算机学院

# 提纲

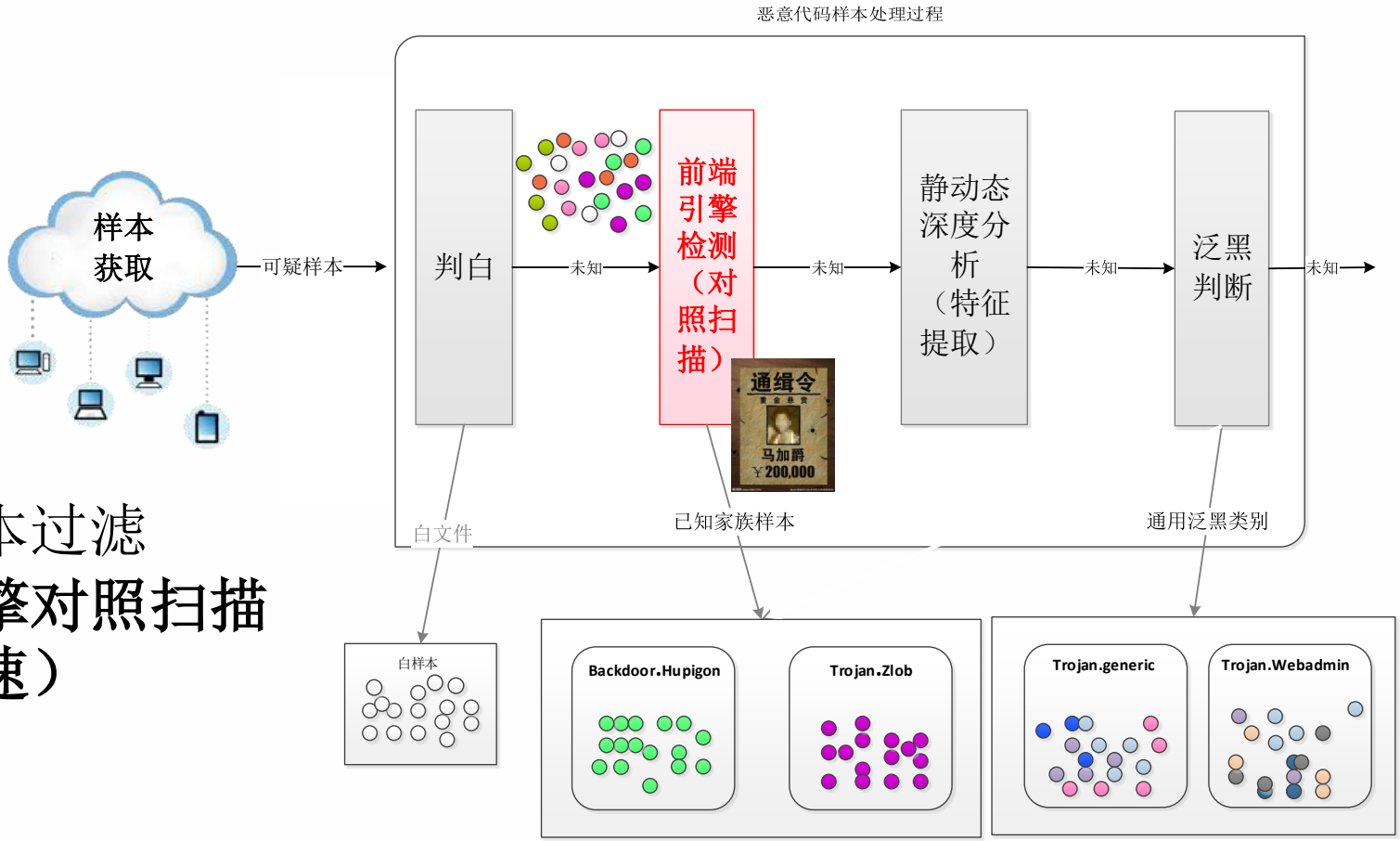
- 现有的恶意代码分析体系
- 一种进化的恶意代码分析体系
- 恶意代码分类聚类学术研究进展
- 数据集

# 安全厂商的样本处理流程（1）



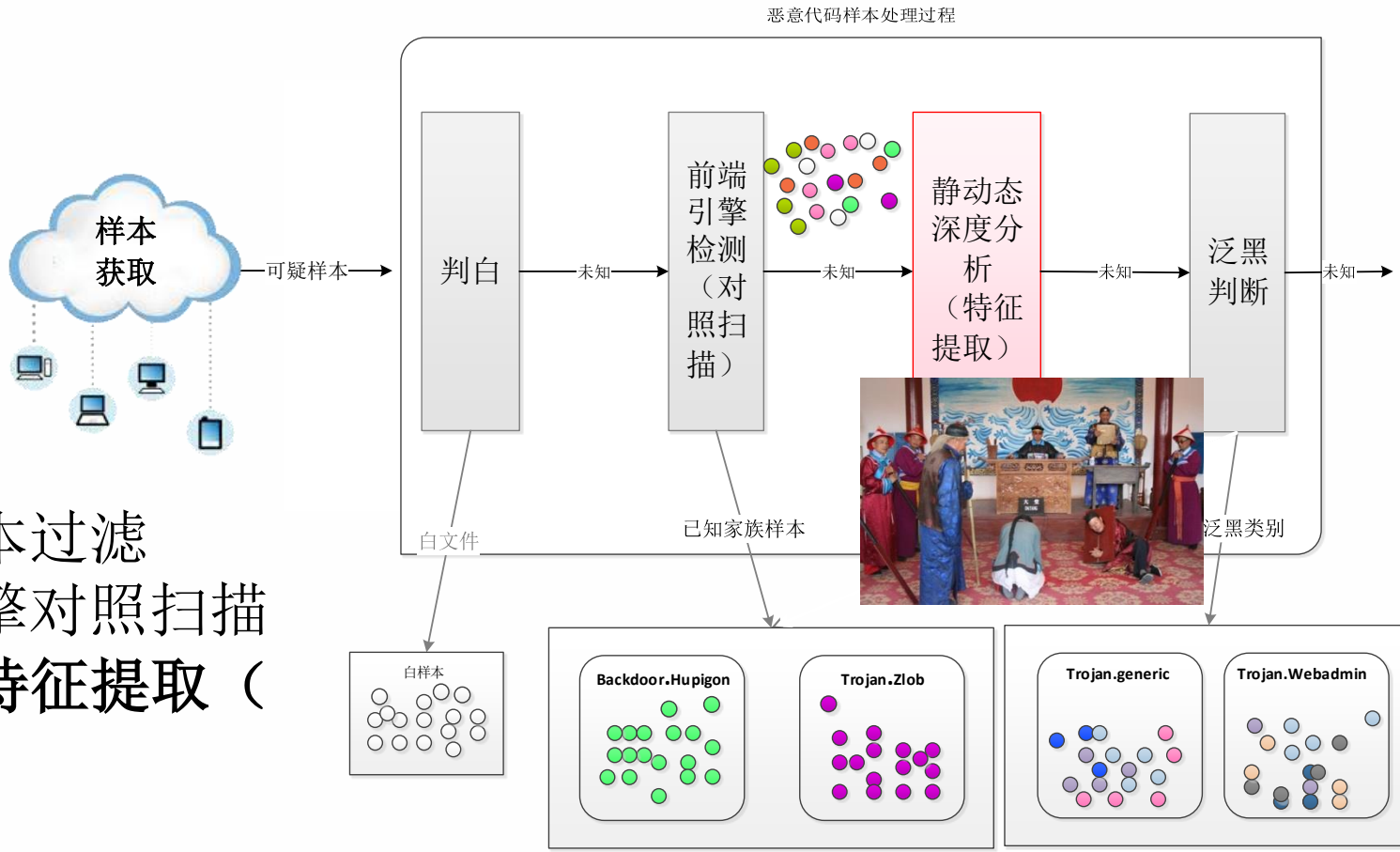

- 白样本过滤

# 安全厂商的样本处理流程（2）



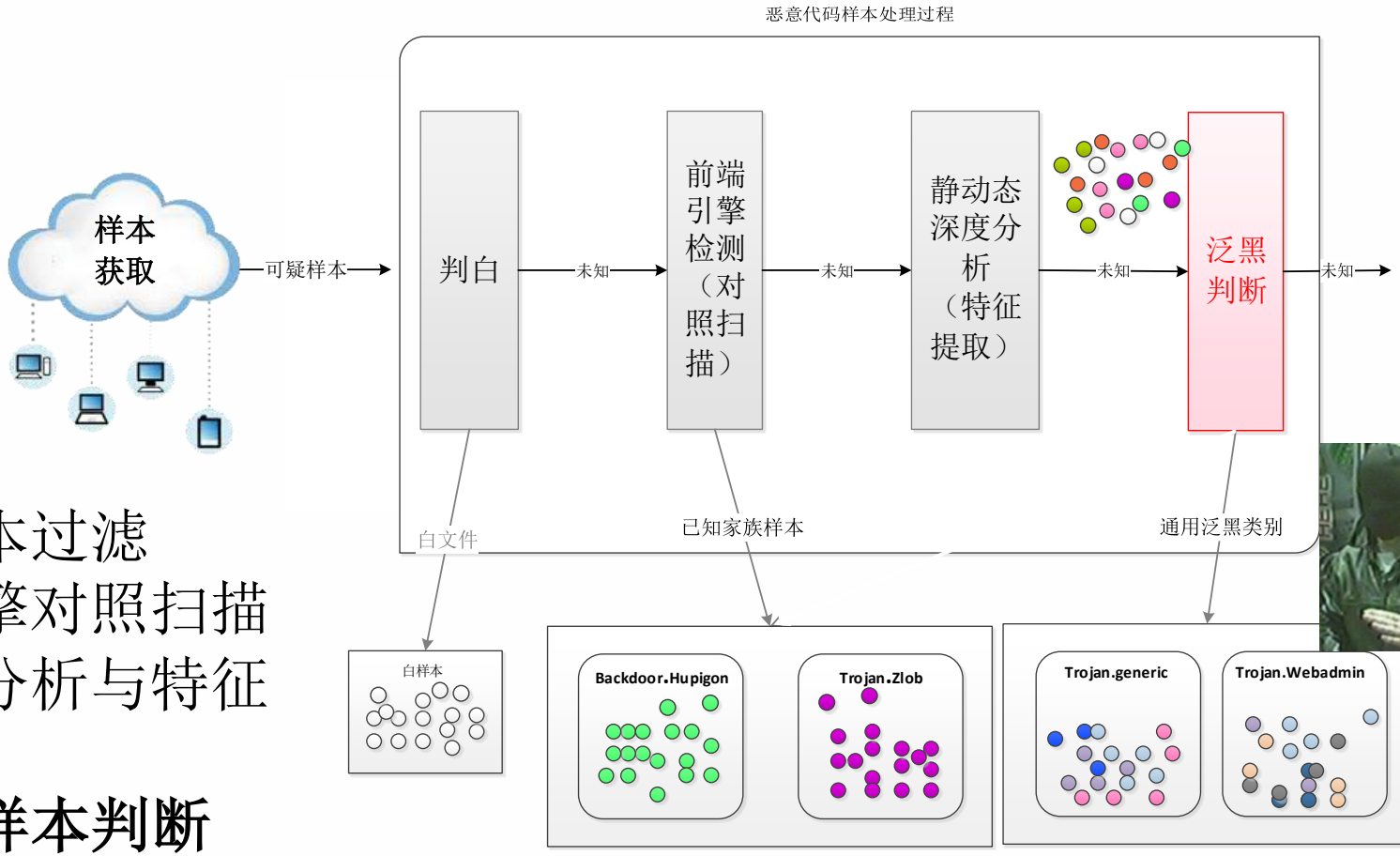

- 白样本过滤
- 多引擎对照扫描（快速）

# 安全厂商的样本处理流程（3）



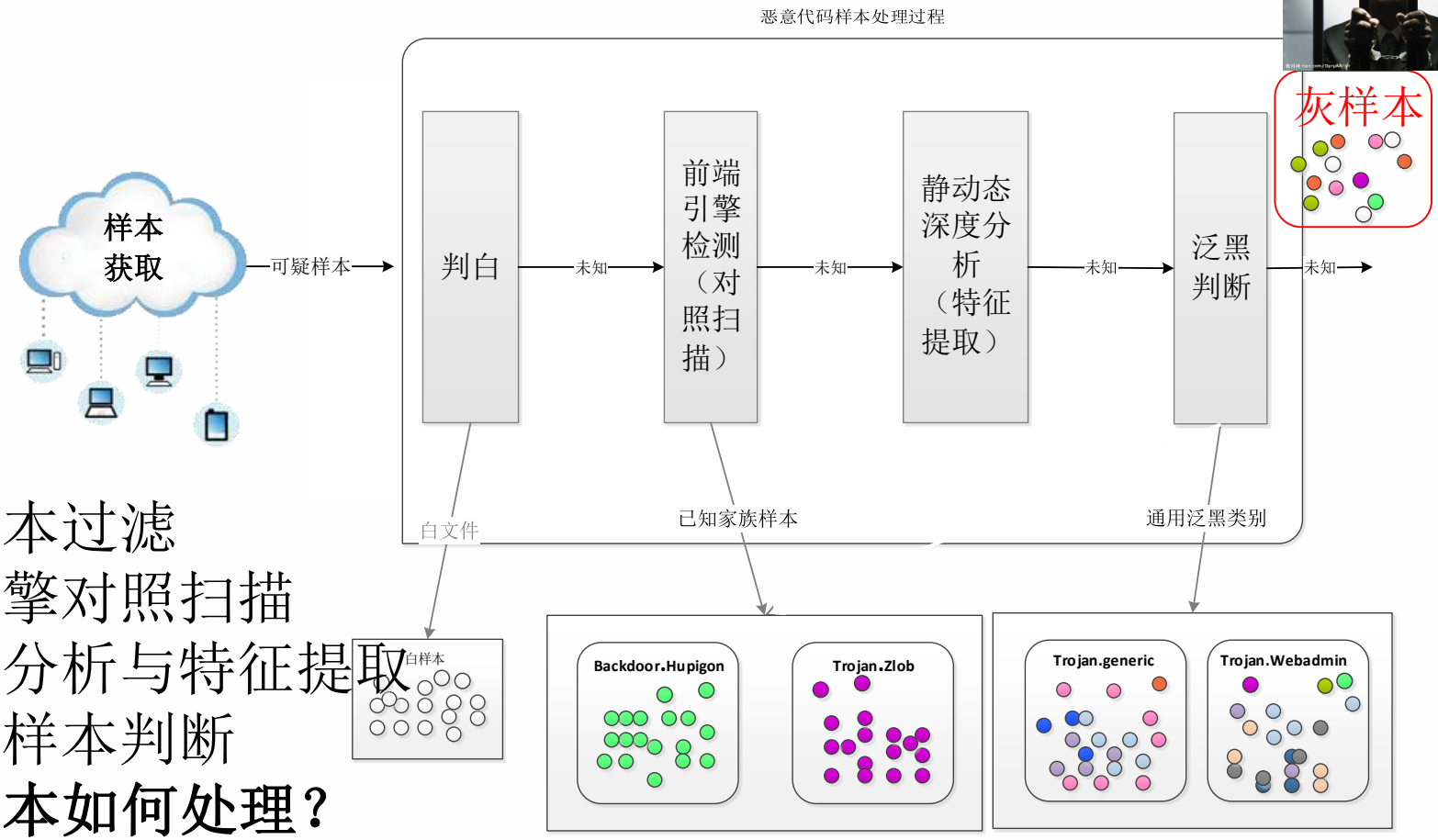

- 白样本过滤
- 多引擎对照扫描
- **深度特征提取（三规）**

# 安全厂商的样本处理流程（4）



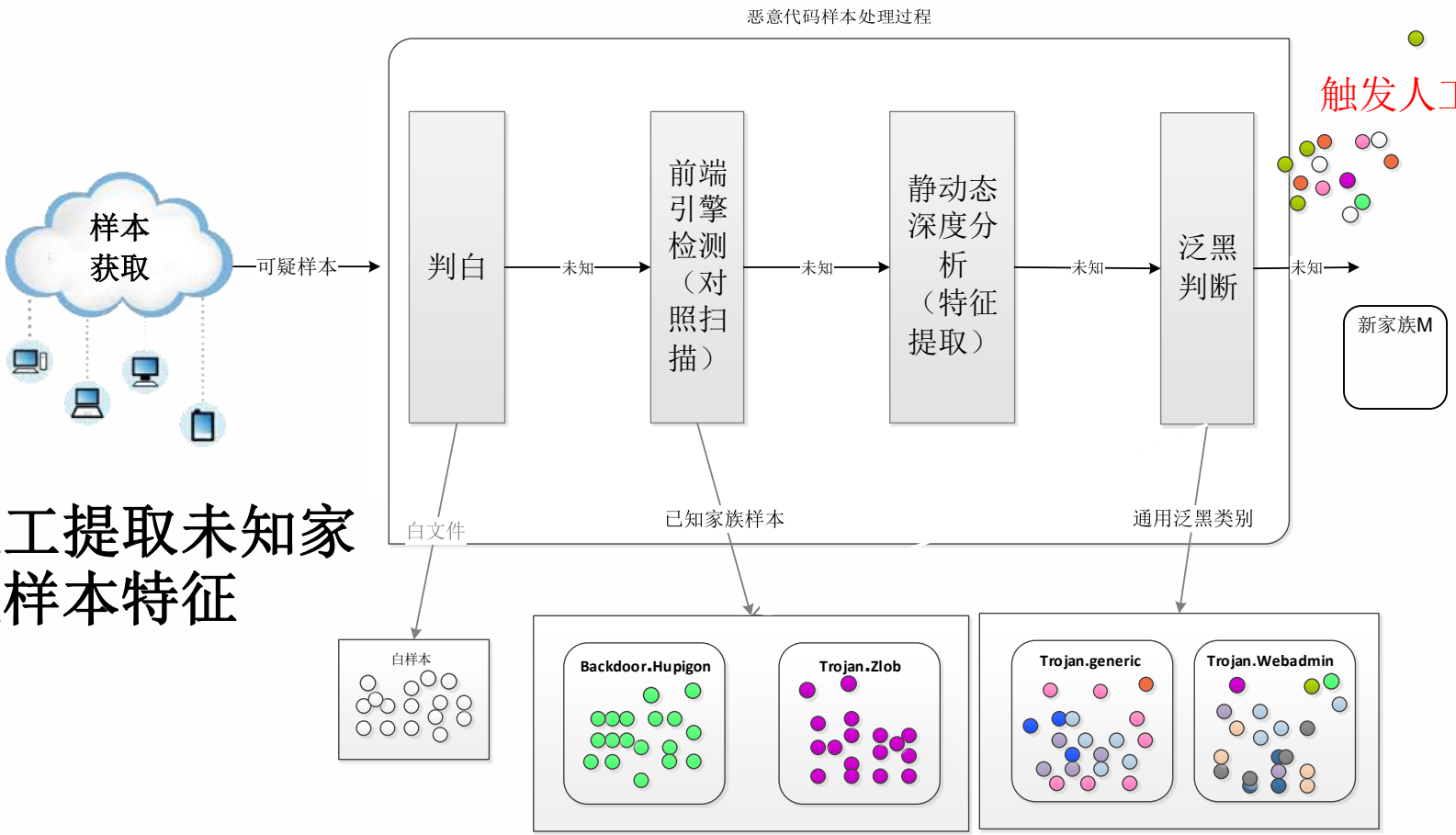

- 白样本过滤
- 多引擎对照扫描
- 深度分析与特征提取
- **泛黑样本判断**

# 安全厂商的样本处理流程（5）



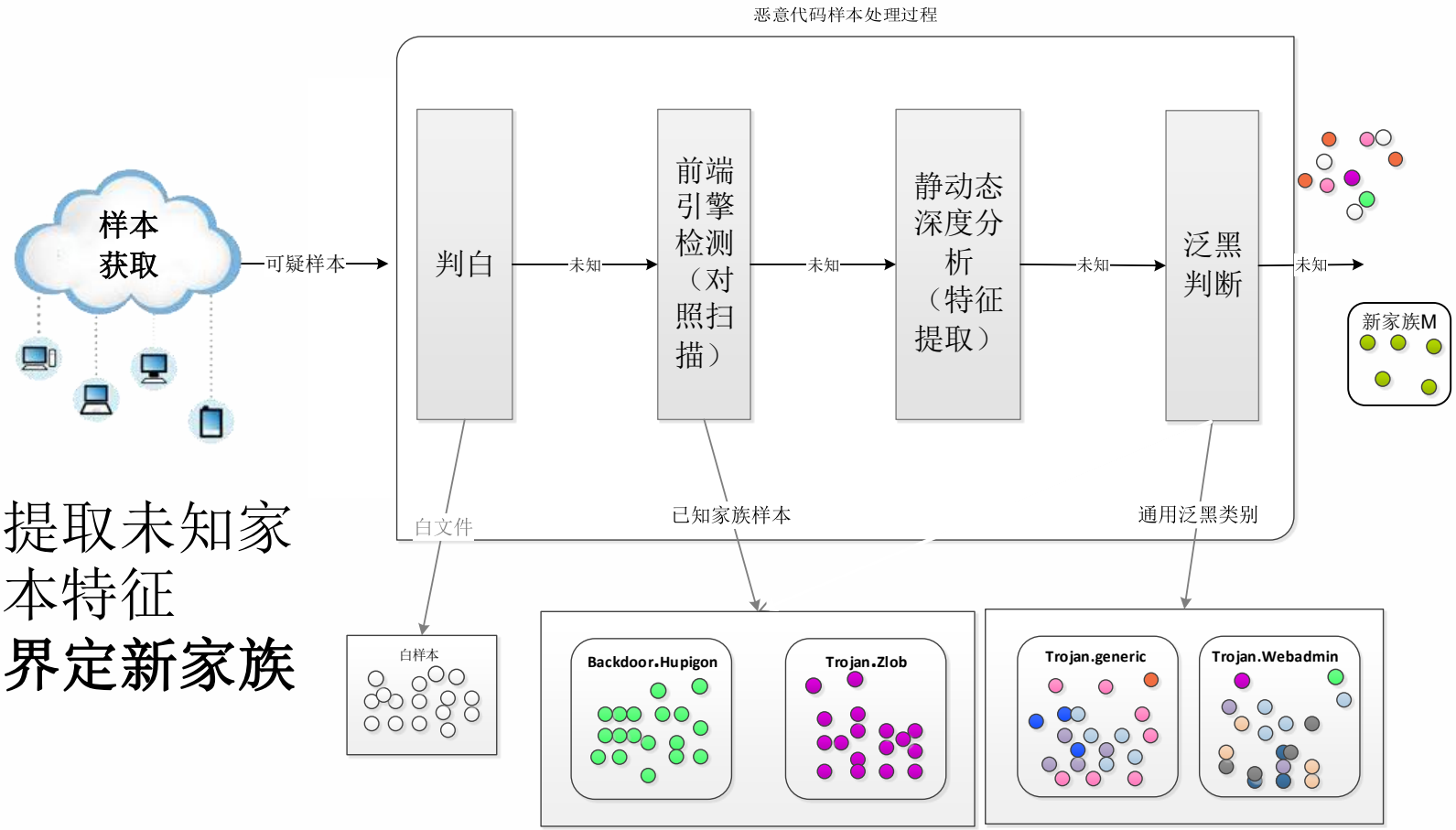

- 白样本过滤
- 多引擎对照扫描
- 深度分析与特征提取
- 泛黑样本判断
- **灰样本如何处理？**

# 未知家族样本触发人工分析（1）



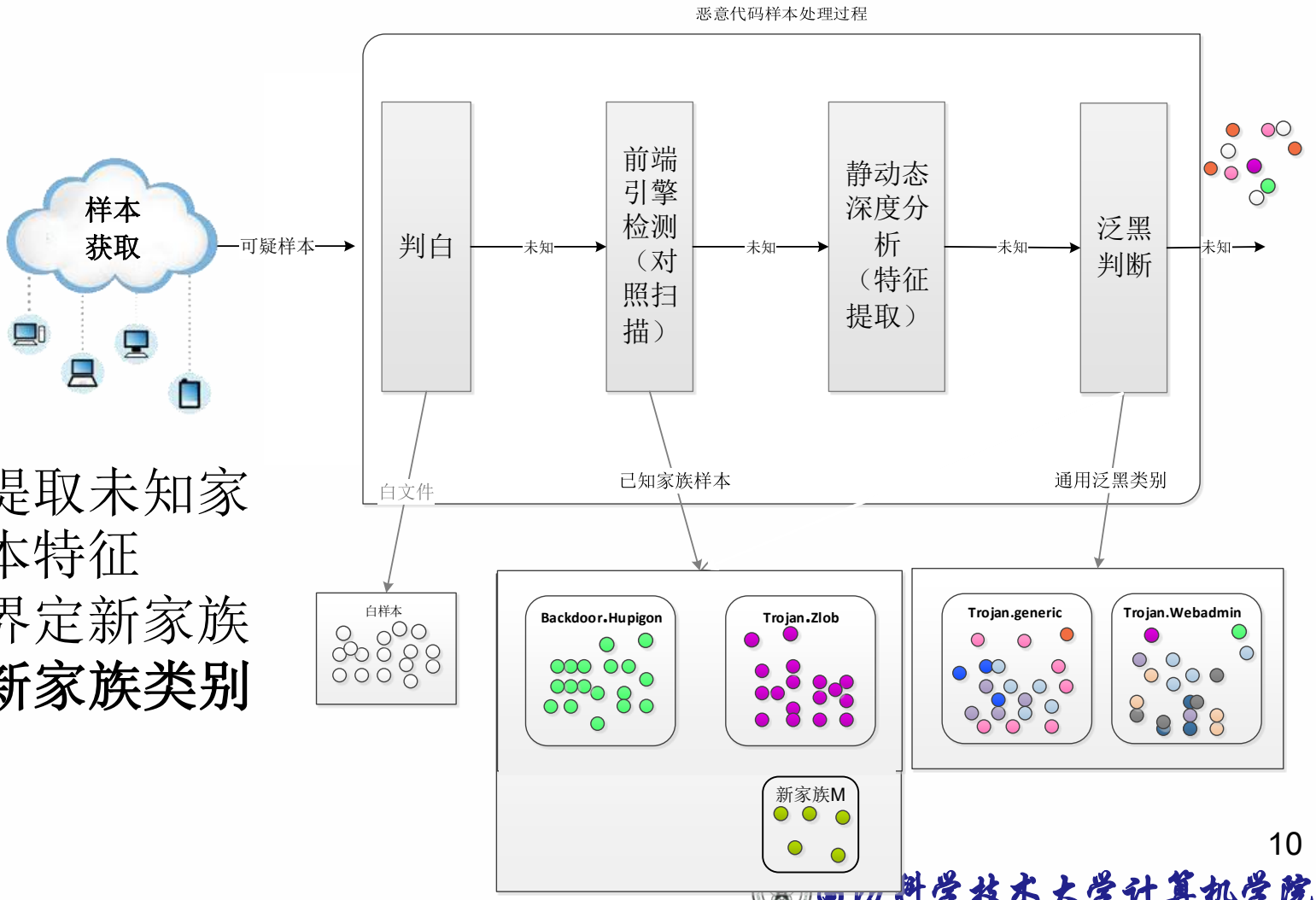

- 人工提取未知家族样本特征

# 未知家族样本触发人工分析（2）



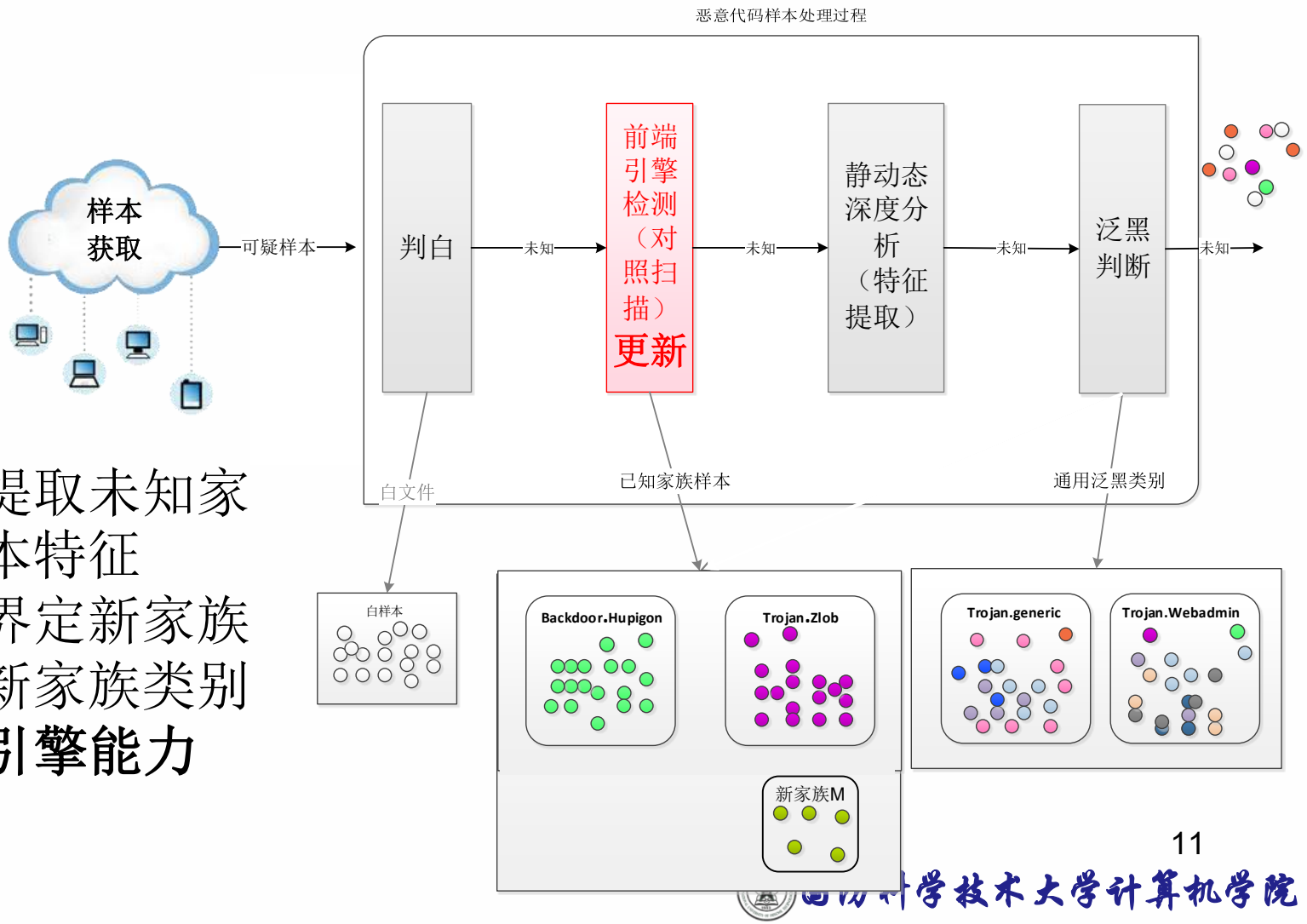

- 人工提取未知家族样本特征
- **人工界定新家族**

# 未知家族样本触发人工分析（3）



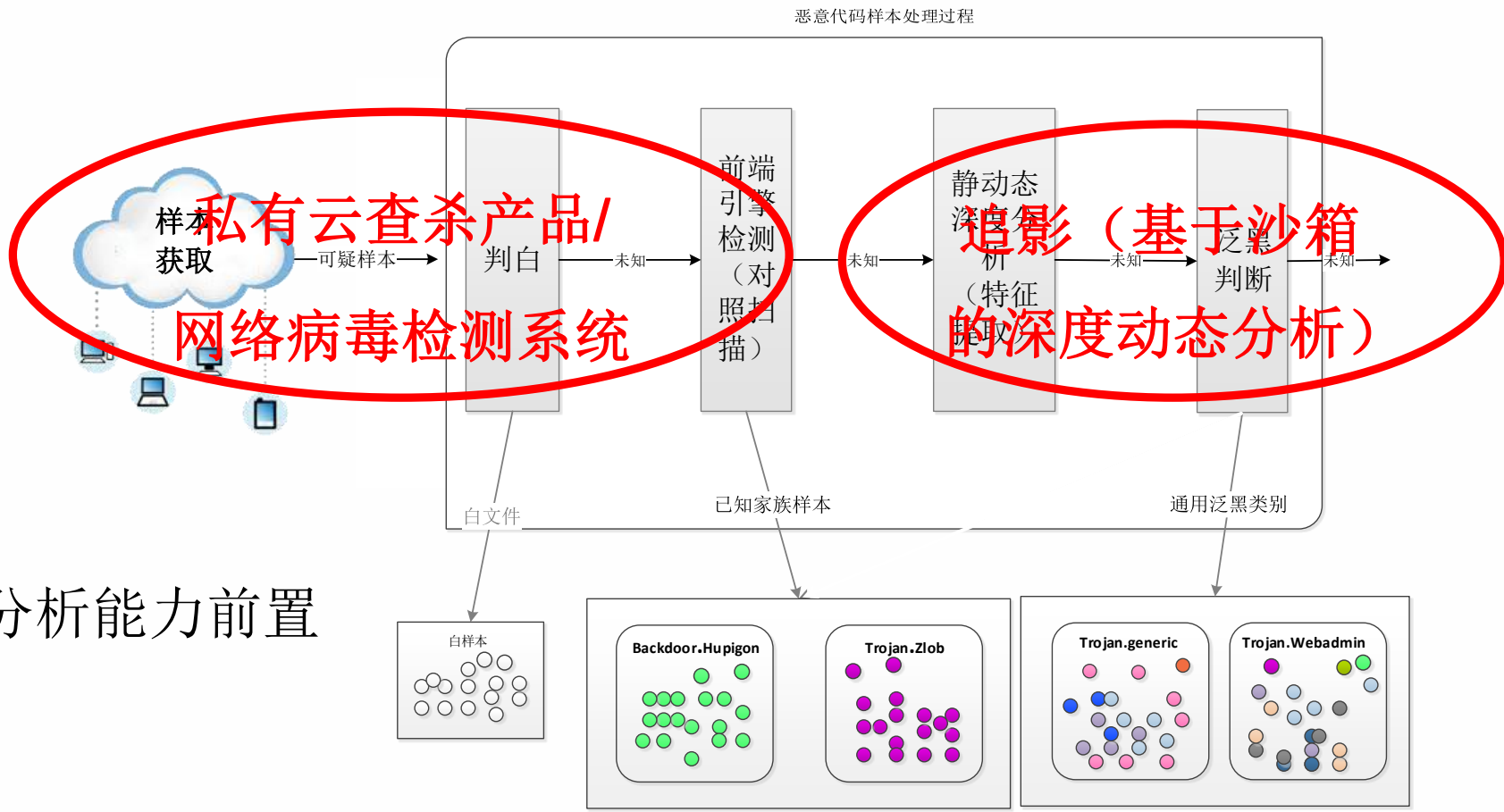

- 人工提取未知家族样本特征
- 人工界定新家族
- **添加新家族类别**

# 未知家族样本触发人工分析（4）



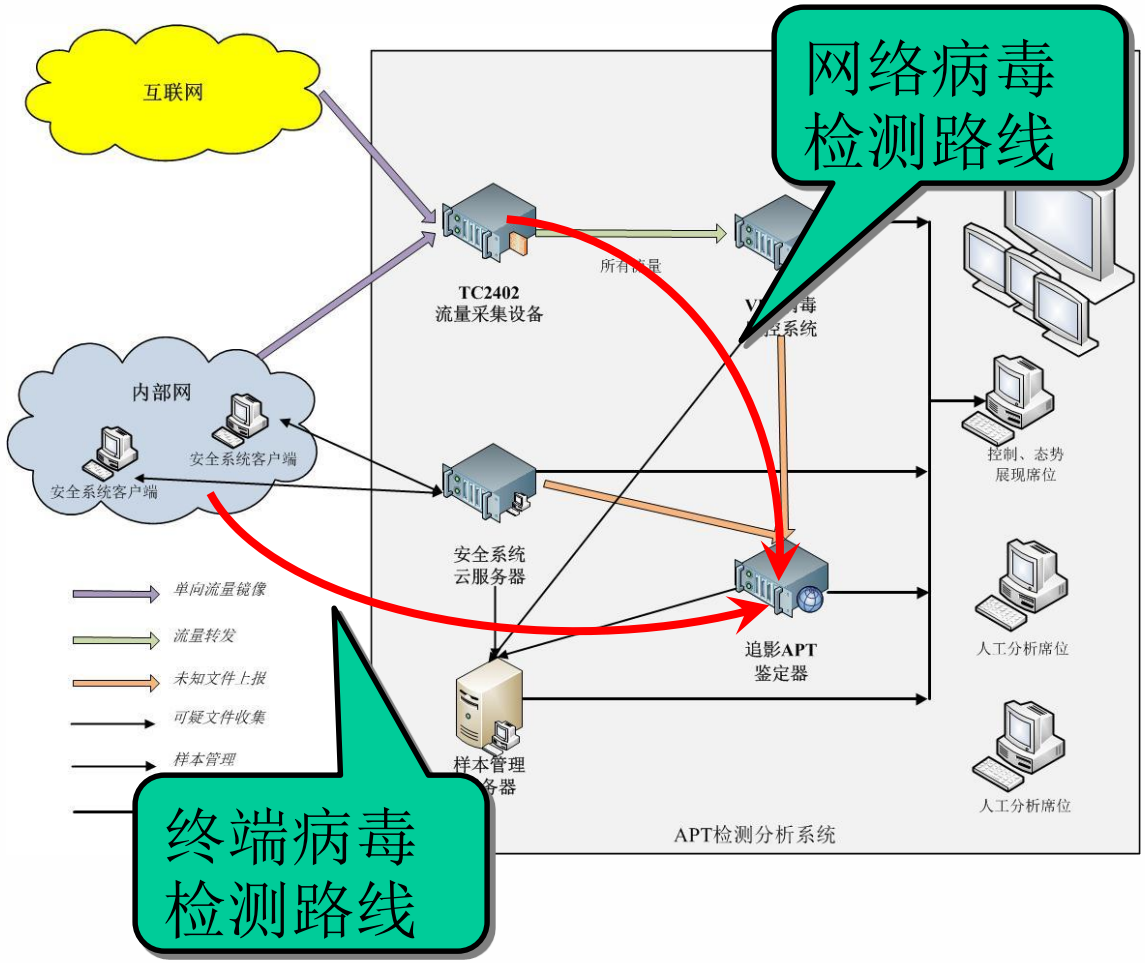

- 人工提取未知家族样本特征
- 人工界定新家族
- 添加新家族类别
- **更新引擎能力**

# 安全厂商的典型恶意代码检测分析设备




病毒分析能力前置

# 我们的APT攻击检测系统结构




- 流量处理
  - 全流量采集与存储
  - 深度会话分析
  - 网络病毒检测
- 样本处理
  - 客户端前置样本捕获
  - 云端黑白名单综合检测与分析
  - 后端虚拟执行动态鉴定

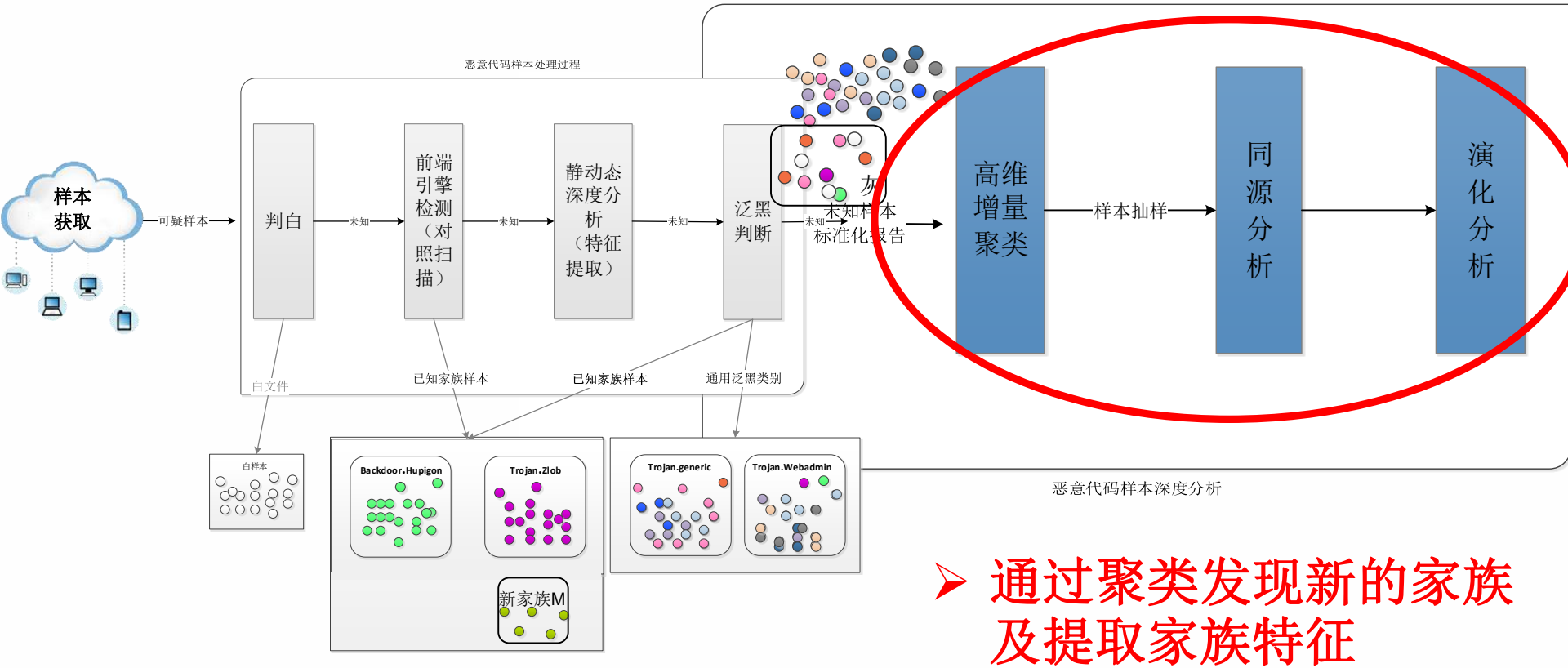
# 发现的灰样本和泛黑样本数量





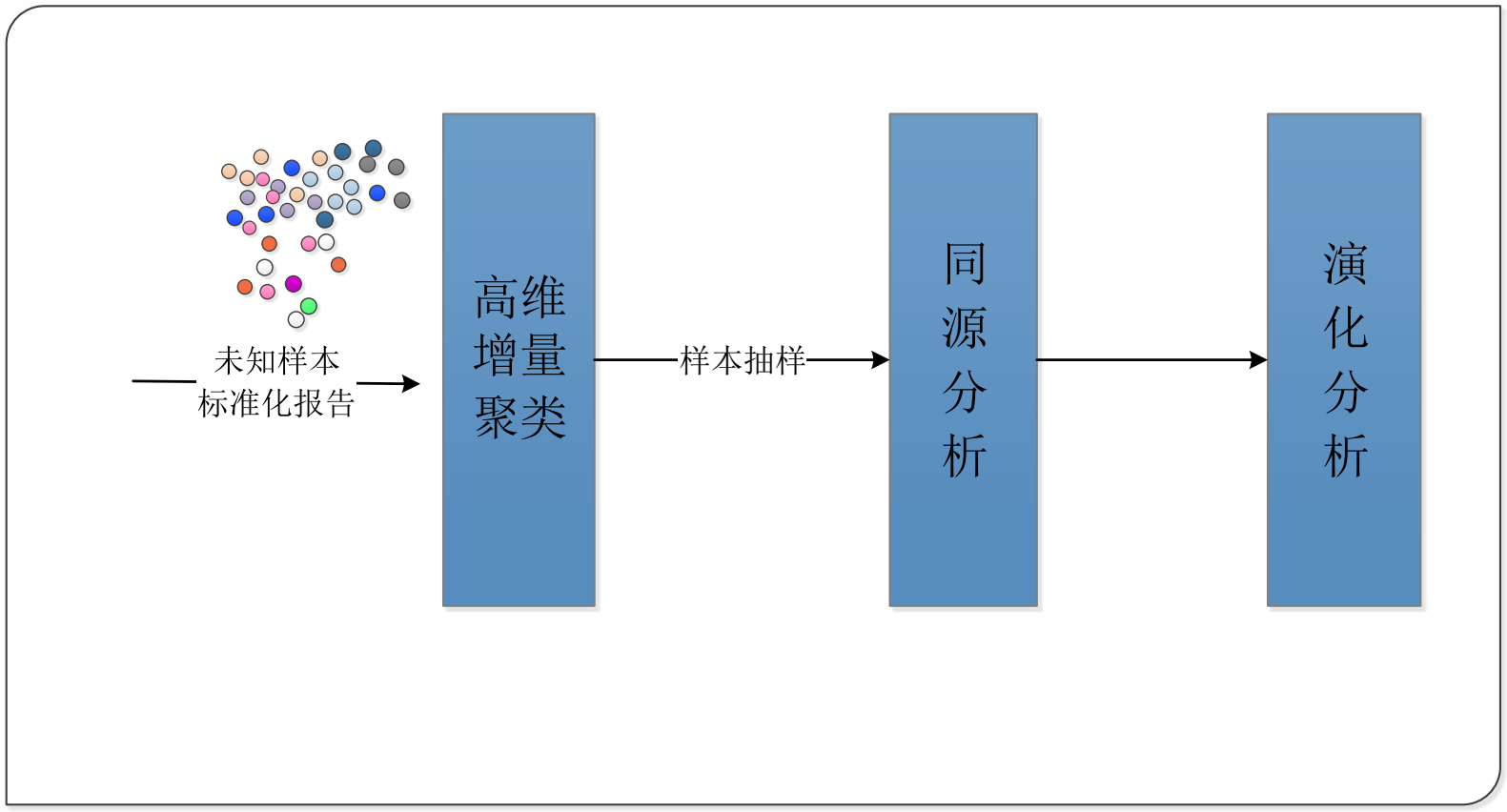
- 泛黑类别淹没了有价值的APT样本

# 恶意代码聚类与同源分析需求(1)



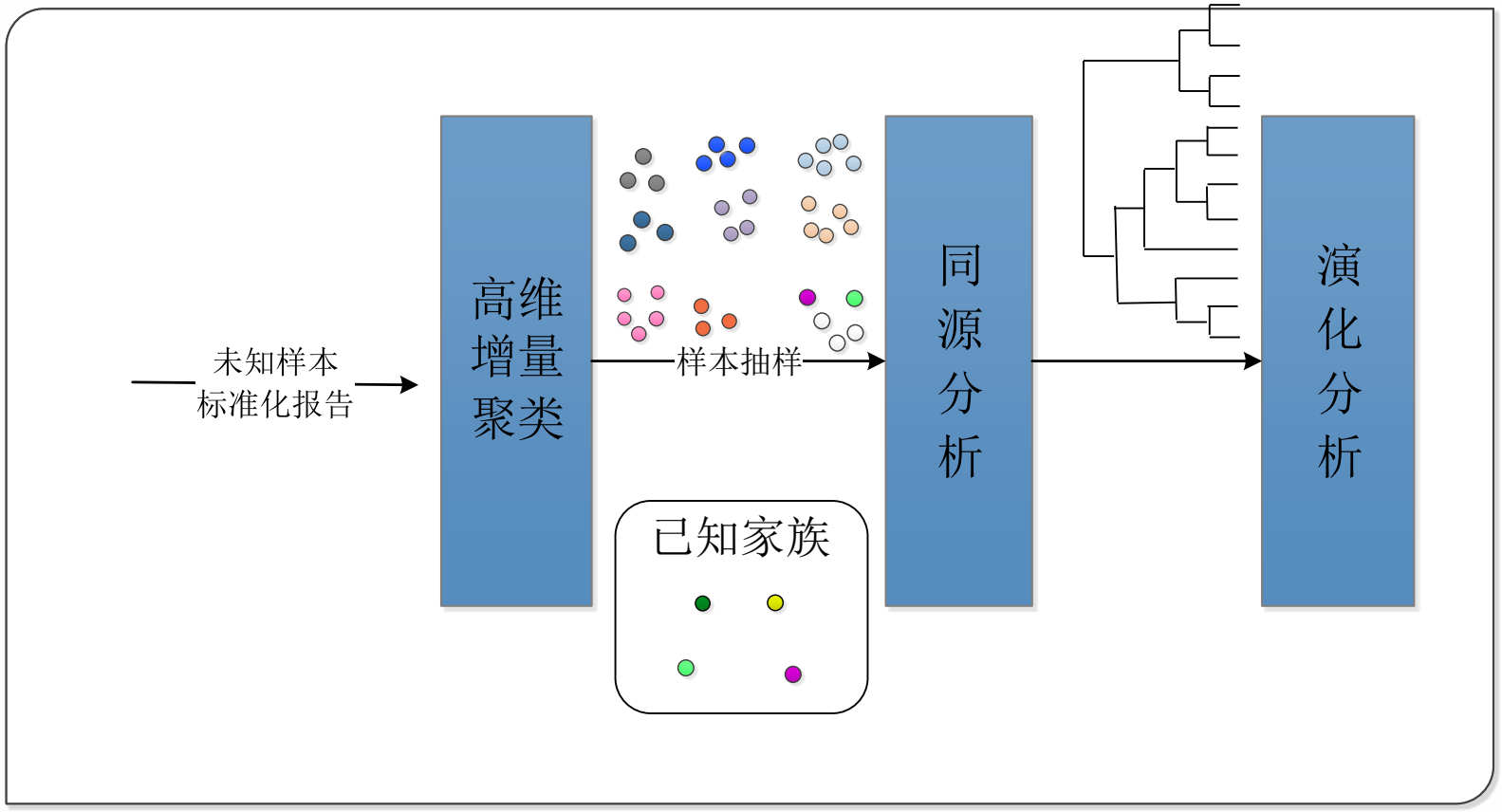

- 通过聚类发现新的家族及提取家族特征
- 自动化分析家族同源和演化关系

# 恶意代码聚类与同源分析需求(2)



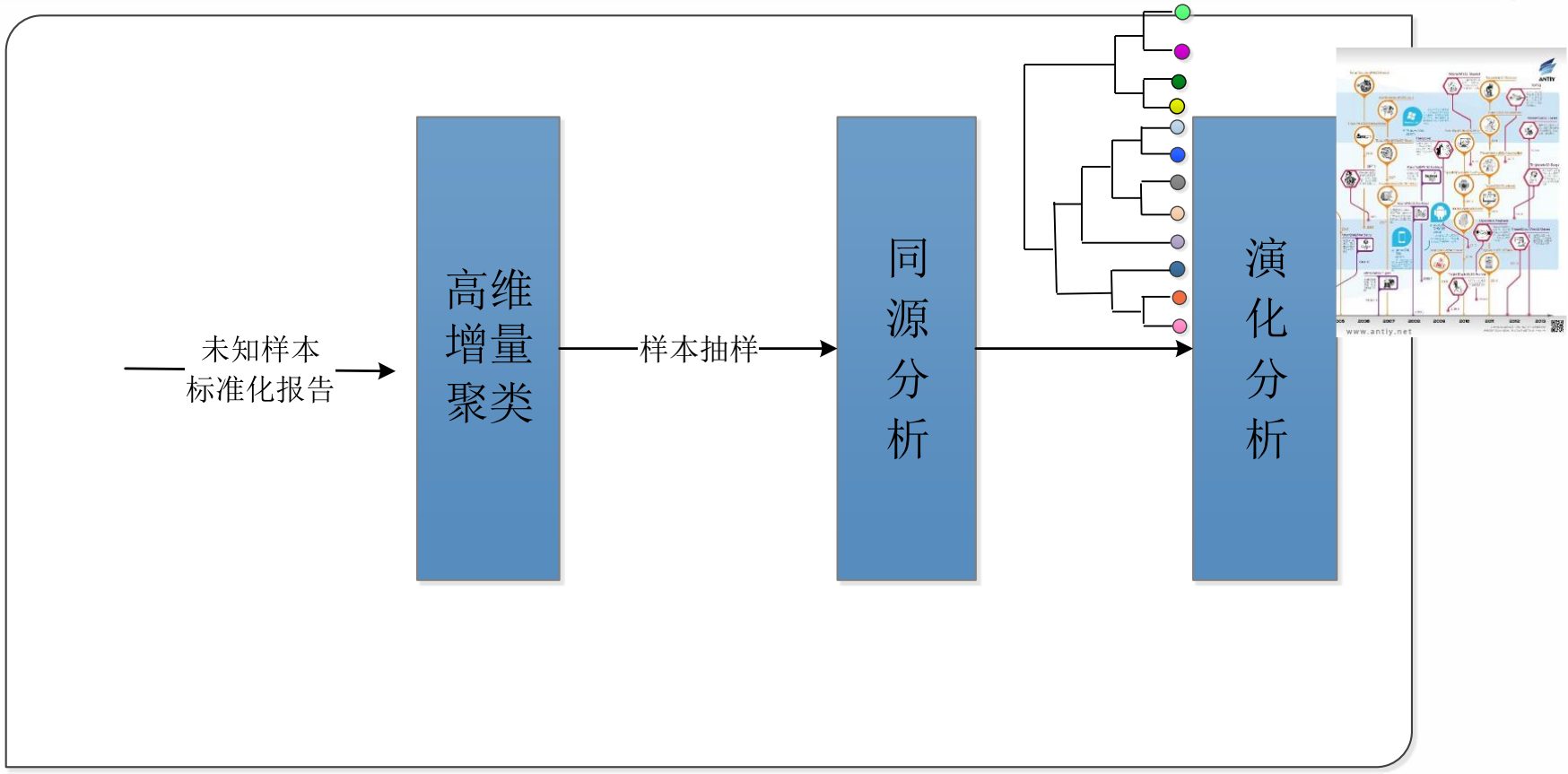

恶意代码样本深度分析

# 恶意代码聚类与同源分析需求(3)




恶意代码样本深度分析

# 恶意代码聚类与同源分析需求(4)




恶意代码样本深度分析

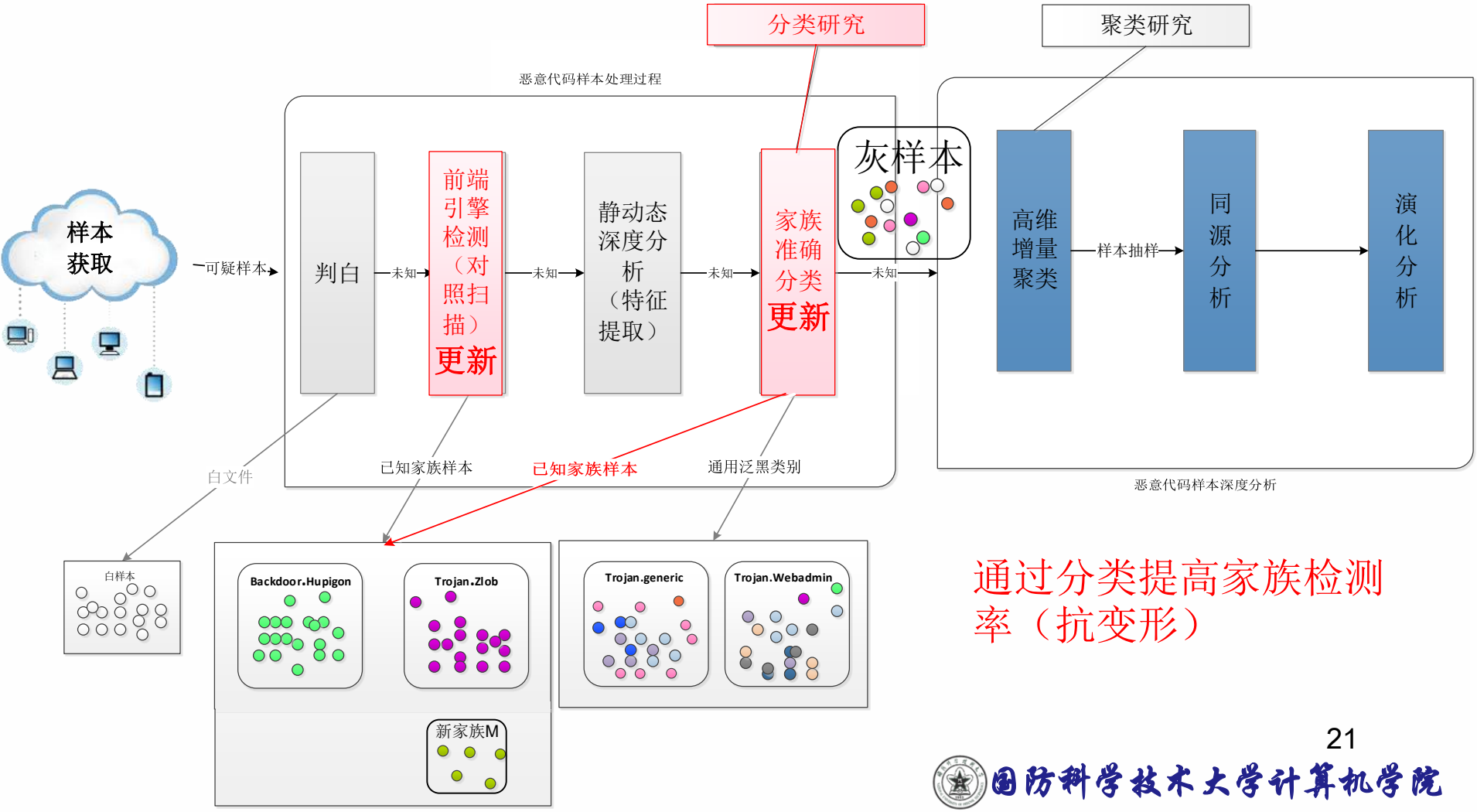
# 改进恶意代码分析流水体系的思考

- 如何识别新的家族？
- 如何用精确的分类算法提升后端家族鉴别的准确性？
- 如何避免泛黑类别划分的“黑洞”效应？
- 能否融合的学术界研究成果？

# 提纲

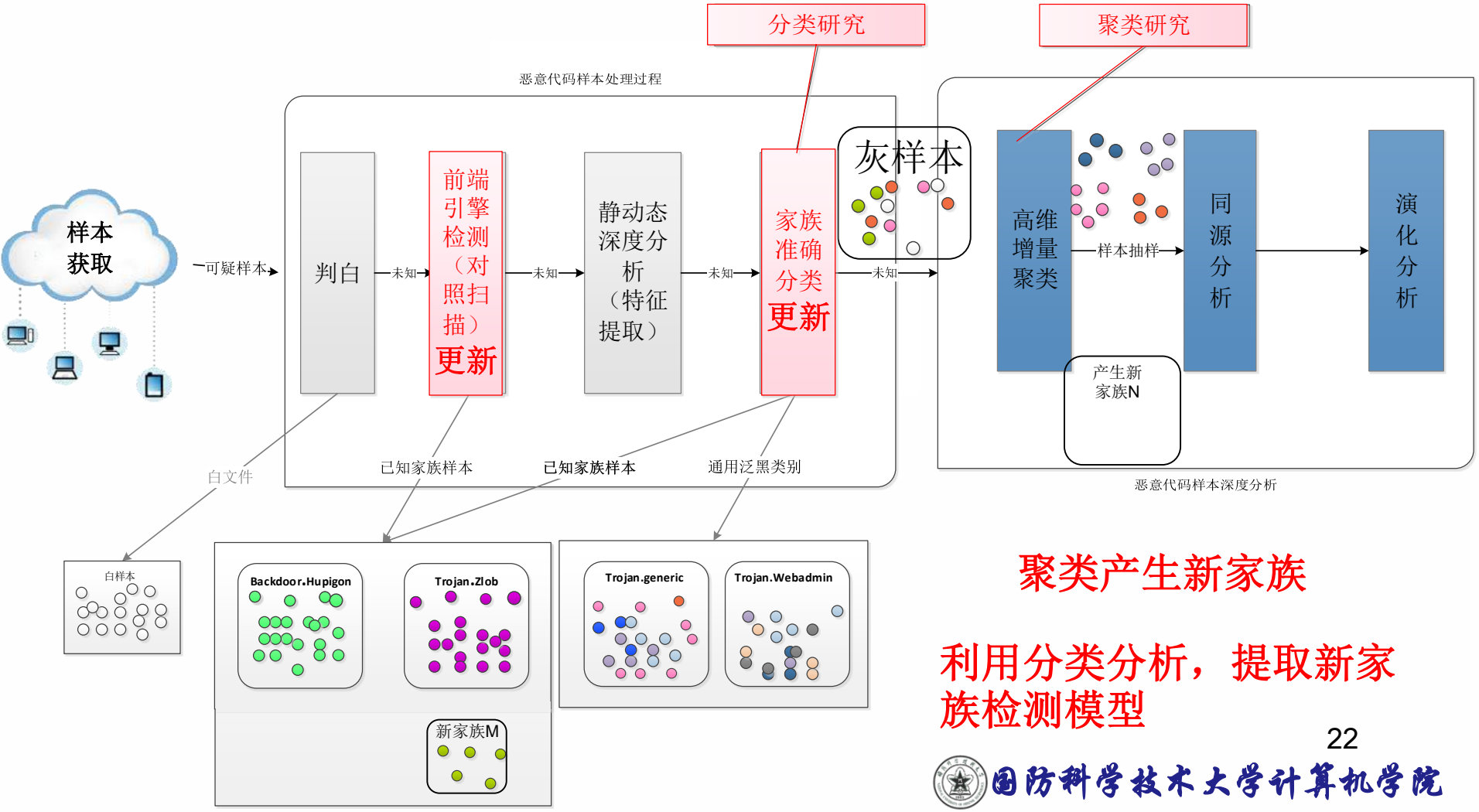
- 现有的恶意代码分析体系
- 一种进化的恶意代码分析体系
- 恶意代码分类聚类学术研究进展
- 数据集

# 一种进化的恶意代码分析体系（1）




通过分类提高家族检测率（抗变形）

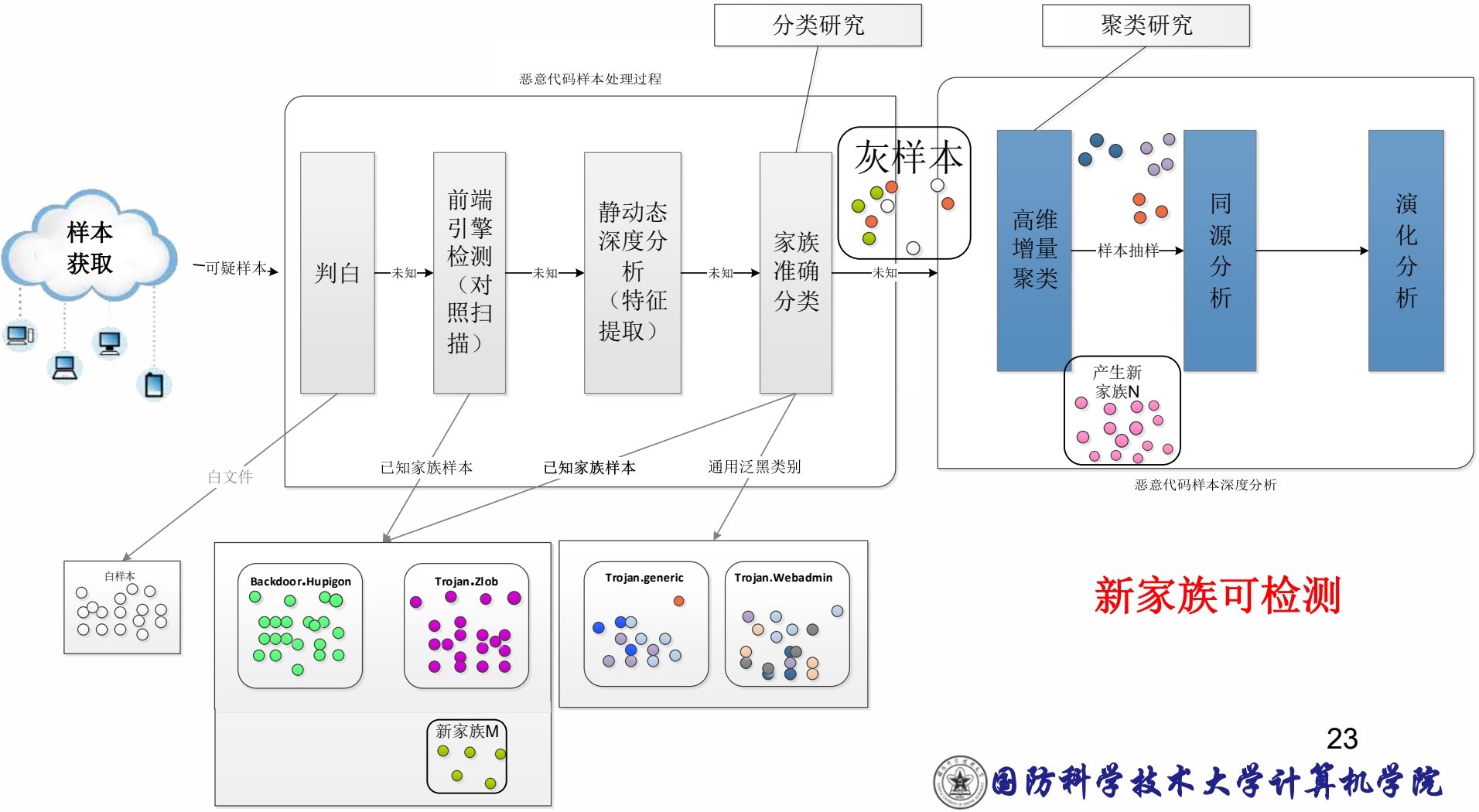
# 一种进化的恶意代码分析体系（2）




聚类产生新家族

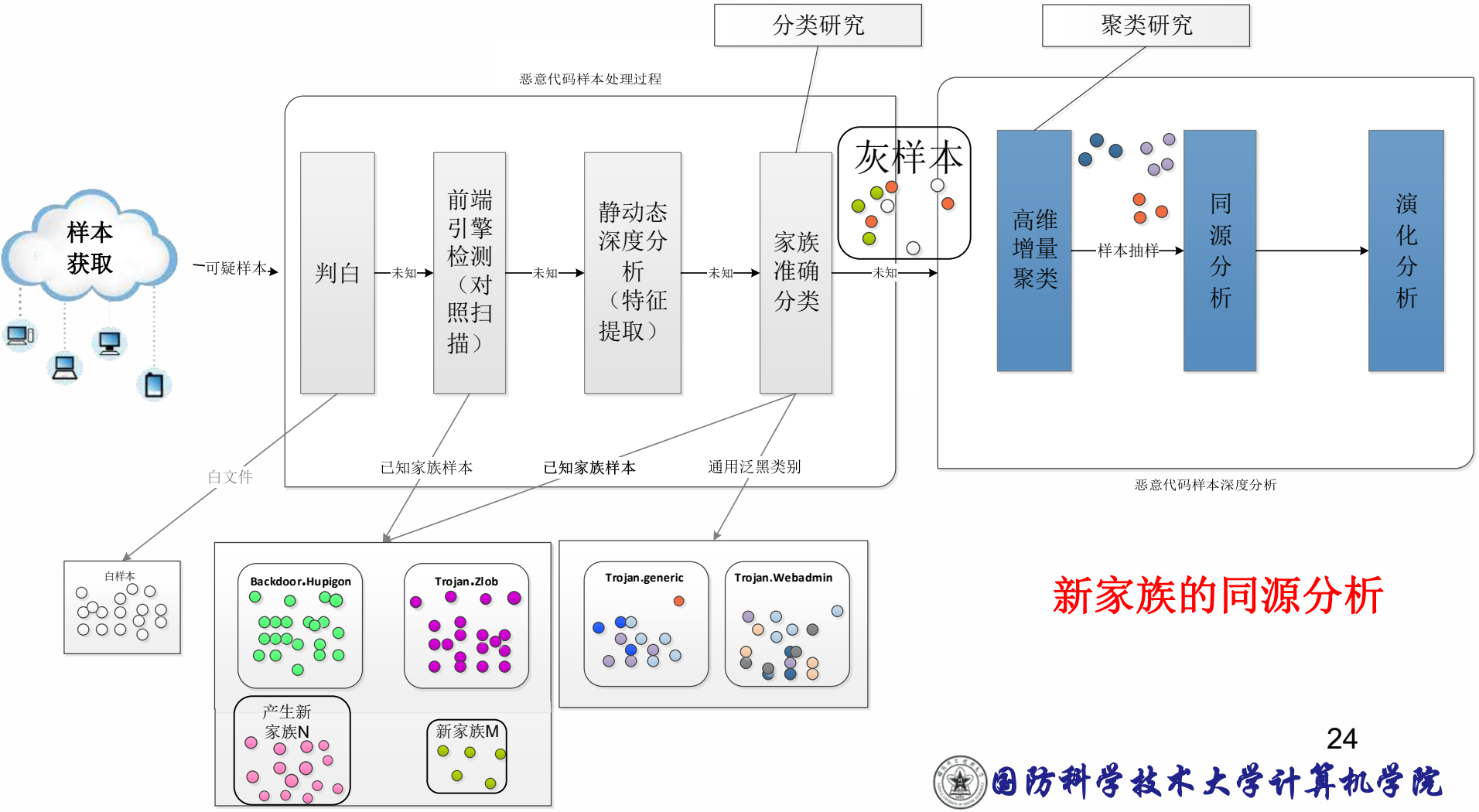
利用分类分析，提取新家族检测模型

# 一种进化的恶意代码分析体系（3）



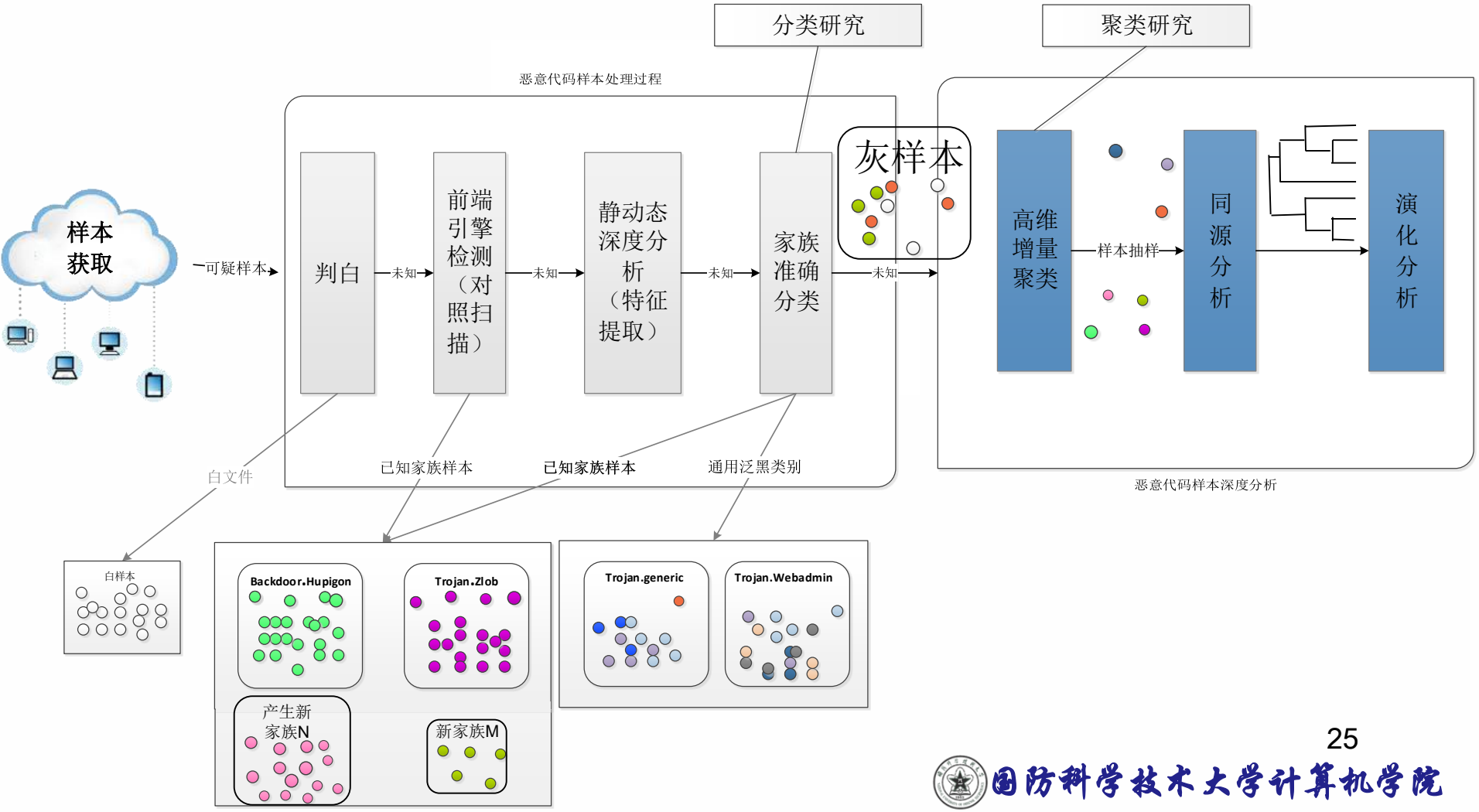

新家族可检测

# 一种进化的恶意代码分析体系（4）




新家族的同源分析

# 一种进化的恶意代码分析体系（5）



# 提纲

- 现有的恶意代码分析流水模型
- 下一代恶意代码分析流水模型
- 恶意代码分类聚类学术研究进展
- 数据集

# “忽如一夜春风来 千树万树梨花开”


一、聚类：

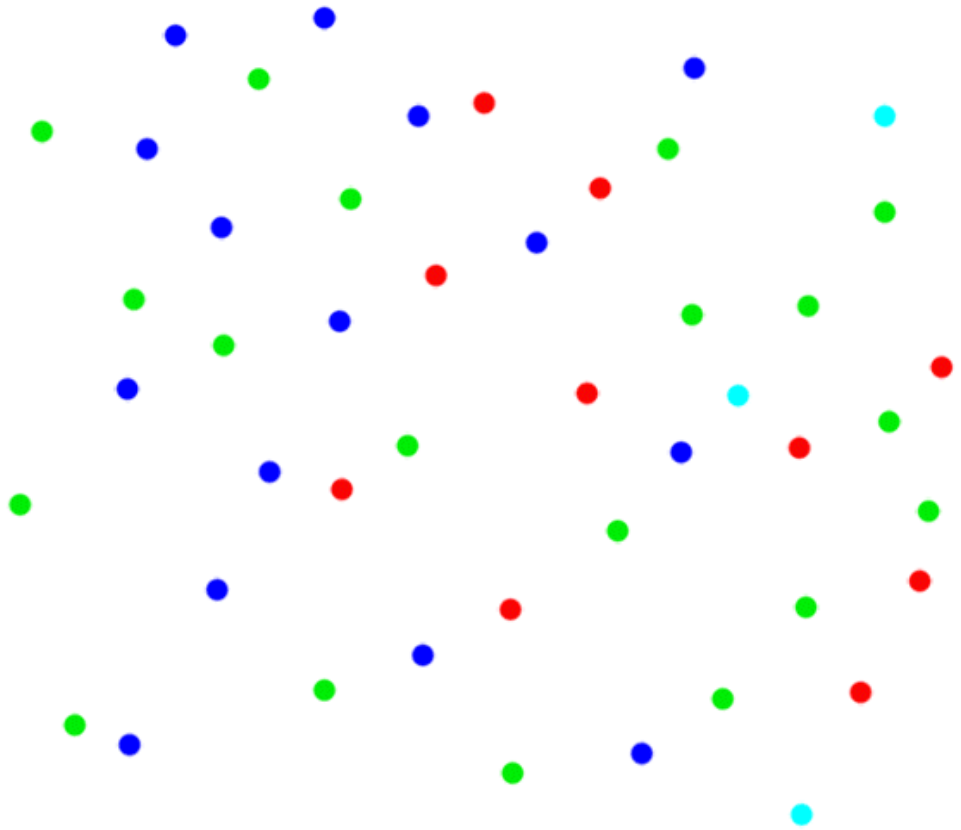
1. Scalable, Behavior-Based Malware Clustering[C]. Bayer U, Comparetti P M, Hlauscheck C, et al. In: 16th Symposium on Network and Distributed System Security (NDSS).2009.
2. Automatic malware cate-gorization using cluster ensemble. Yanfang Ye, Tao Li, Yong Chen, and Qingshan Jiang. InProceedings of the 16th ACM SIGKDD in-ternational conference on Knowledge discovery and data mining, KDD '10, pages 95{104, New York, NY, USA, 2010. ACM.
3. VAMO: Towards a Fully Automated Malware Clustering Validity Analysis. Roberto Perdisci, ManChon U .2012, ACSAC '12: Proceedings of the 28th Annual Computer Security Applications Conference.
4. MutantX-S: scalable malware clustering based on static features. Xin Hu, Sandeep Bhatkar, Kent Griffin, Kang G. Shin. Jun. 2013 Proceedings of the 2013 USENIX conference on Annual Technical Conference.
5. Clustering of Similar Malware Behavior via Structural Host-Sequence Comparison. Horng-Tzer Wang ; Ching-Hao Mao ; Te-En Wei ; Hahn-Ming Lee. Computer Software and Applications Conference (COMPSAC), 2013 IEEE 37th Annual .
6. 基于特征聚类的海量恶意代码在线自动分析模型. 徐小琳，云晓春，周勇林，康学斌 -《通信学报》，2013年 第8期.
7. Poisoning Behavioral Malware Clustering. Battista Biggio, Konrad Rieck, Davide Ariu, Christian Wressnegger, Igino Corona, Giorgio Giacinto, Fabio Roli .2014, AISec '14: Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop.
8. Classy: fast clustering streams of call-graphs. Kostakis, O (Kostakis, Orestis).2014, DATA MINING AND KNOWLEDGE DISCOVERY.

二、分类：

1. Classification of Malware Using Structured Control Flow. Proc. Eighth Australasian Symp. Parallel and Distributed Computing (AusPDC '10), 2010.
2. Fast Malware Classification by Automated Behavioral Graph Matching. Younghee Park, Douglas Reeves, Vikram Mulukutla, Balaji Sundaravel .2010, CSIIRW '10: Proceedings of the Sixth Annual Workshop on Cyber Security and Information Intelligence Research.
3. Malware Classification Based on Call Graph Clustering.J. Kinable and O. Kostakis, J. Computer Virology, vol. 7, pp. 233-245, 2011.
4. A Comparative Assessment of Malware Classification using Binary Texture Analysis and Dynamic Analysis. Lakshmanan Nataraj, Vinod Yegneswaran, Phillip Porras, Jian Zhang .2011, AISec '11: Proceedings of the 4th ACM workshop on Security and artificial intelligence.
5. Malware classification using instruction frequencies.Kyoung Soo Han, Boojoong Kang, Eul Gyu Im.2011, RACS '11: Proceedings of the 2011 ACM Symposium on Research in Applied Computation.
6. Malware classification method via binary content comparison . Boojoong Kang, Taekeun Kim, Heejun Kwon, Yangseo Choi, Eul Gyu Im .2012. RACS '12: Proceedings of the 2012 ACM Research in Applied Computation Symposium.
7. Improving Malware Classfication: Bridging the Static/Dynamic Gap. Blake Anderson, Curtis Storlie, Terran Lane .2012, AISec '12: Proceedings of the 5th ACM workshop on Security and artificial intelligence.
8. Malware Classification based on Extracted API Sequences using Static Analysis.Kazuki Iwamoto, Katsumi Wasaki .2012, AINTEC '12: Proceedings of the Asian Internet Engineeering Conference .
9. Tracking memory writes for malware classification and code reuse identification. André Ricardo Abed Grégio, Paulo Lício de Geus, Christopher Kruegel, Giovanni Vigna. Jul. 2012 Proceedings of the 9th international conference on Detection of Intrusions and Malware, and Vulnerability Assessment.
10. Malwise: an effective and efficient classification system for Packed and Polymorphic Malware. Silvio Cesare, Yang Xiang,Wanlei Zhou. IEEE Transactions on Computers, Volume 62 Issue 6, June 2013, Pages 1193-1206.
11. Discriminant Malware Distance Learning on Structural Information for Automated Malware Classification. Deguang Kong, Guanhua Yan .2013, KDD '13: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.
12. Exploring discriminatory features for automated malware classification. Guanhua Yan, Nathan Brown, Deguang Kong. Jul. 2013 Proceedings of the 10th international conference on Detection of Intrusions and Malware, and Vulnerability Assessment.
13. Unveiling Zeus: automated classification of malware samples. Abedelaziz Mohaisen, Omar Alrawi. May. 2013 Proceedings of the 22nd international conference on World Wide Web companion.
14. Classification of Malware Families Based on N-grams Sequential Pattern Features. Chatchai Liangboonprakong, Ohm Sornil. Industrial Electronics and Applications (ICIEA), 2013 8th IEEE Conference on.
15. Malware Classification based on Social Network Analysis of Call Graph. Jae-wook Jang, Jiyoung Woo, Jaesung Yun, Huy Kang Kim .2014, WWW Companion '14: Proceedings of the companion publication of the 23rd international conference on World wide web companion.
16. Structural Classification and Similarity Measurement of Malware.2014,IEEJ:Institute of Electrical Engineers of Japan.
17. Evolutionary algorithms for classification of malware families through different network behaviors.M. Zubair Rafique, Ping Chen, Christophe Huygens, Wouter Joosen.2014. GECCO '14: Proceedings of the 2014 conference on Genetic and evolutionary computation
18. Blind Separation of Benign and Malicious Events to Enable Accurate Malware Family Classification. Hesham Mekky, Aziz Mohaisen, Zhi-Li Zhang .2014, CCS '14: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security.
19. A Two-Stage Methodology Using K-NN and False-Positive Minimizing ELM for Nominal Data Classification. Akusok, A (Akusok, Anton); Miche, Y (Miche, Yoan); Hegedus, J (Hegedus, Jozsef); Nian, R (Nian, Rui); Lendasse, A (Lendasse, Amaury).2014, COGNITIVE COMPUTATION.
20. Control Flow-Based Malware Variant Detection. Cesare, S (Cesare, Silvio); Xiang, Y (Xiang, Yang); Zhou, WL (Zhou, Wanlei).2014, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING.
21. Malware variant detection using similarity search over content fingerprint. Ban Xiaofang ; Chen Li ; Hu Weihua ; Wu Qu . Control and Decision Conference (2014 CCDC), The 26th Chinese .
22. Structural Classification and Similarity Measurement of Malware. Shi, Hongbo; Hamagami, Tomoki; Yoshioka, Katsunari. IEEJ TRANSACTIONS ON ELECTRICAL AND ELECTRONIC ENGINEERING 卷: 9 期: 6 页: 621-632 出版年:NOV 2014.

# 恶意代码分类过程

- 家族样本标定与特征提取
- 特征模型生成
- 新样本家族分类

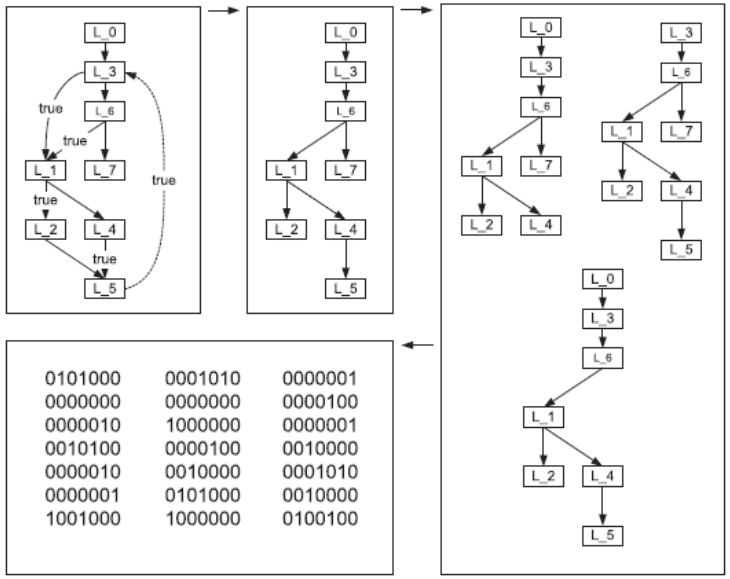


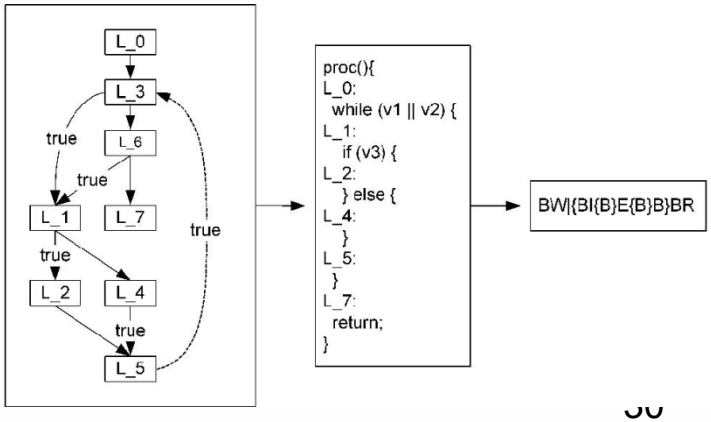
# 分类研究进展


- [1]《Control Flow-Based Malware Variant Detection》
  - Cesare, S (Cesare, Silvio); Xiang, Y (Xiang, Yang); Zhou, WL (Zhou, Wanlei). IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, 2013
- [2]《Exploring Discriminatory Features for Automated Malware Classification》
  - Guanhua Yan, Nathan Brown, Deguang Kong. 2013 Proceedings of the 10th international conference on Detection of Intrusions and Malware, and Vulnerability Assessment.
- [3] Discriminant Malware Distance Learning on Structural Information for Automated Malware Classification.
  - Deguang Kong, Guanhua Yan. KDD '13: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.
- [4]《Malwise—An Effective and Efficient Classification System for Packed and Polymorphic Malware》
  - Silvio Cesare,Yang Xiang,Wanlei Zhou. IEEE Transactions on Computers,2013
- [5]《POSTER: Blind Separation of Benign and Malicious Events to Enable Accurate Malware Family Classification》
  - Hesham Mekky, Aziz Mohaisen, Zhi-Li Zhang. Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security
- [6]《Structural Classification and Similarity Measurement of Malware》
  - Shi, Hongbo; Hamagami, Tomoki; Yoshioka, Katsunari. IEEJ TRANSACTIONS ON ELECTRICAL AND ELECTRONIC ENGINEERING
- [7]《Mal-Netminer: Malware Classification based on Social Network Analysis of Call Graph》
  - Jae-wook Jang, Jiyoung Woo, Jaesung Yun, Huy Kang Kim. 2014, WWW Companion '14: Proceedings of the companion publication of the 23rd international conference on World wide web companion

# 《Control Flow-Based Malware Variant Detection》(1)

- K-Subgraph特征
  - 将函数流程图中的环去除
  - 通过遍历生成访问流程图中k个基本块的所有可能路径
  - 通过Bliss（open-source）将生成的子图转化为标准化字符串
  - 所有生成的字符串共同构成特征向量
- Q-Gram特征
  - 把控制流图转换为“控制流图字符串”
  - 字符串的n-gram



# 《Control Flow-Based Malware Variant Detection》(2)

- 特征选择
  - 可能的特征空间太大
  - 通过学习选取最频繁的500个特征
  - 使用主成分分析方法降维

- 分类（搜索）算法
  - 对字符串特征集合表示样本，字符串特征有三种相似度计算方式
    - 字符串间编辑距离
    - 基因字符串和BLAST
    - NCD距离
  - 向量海量数据的高维索引结构
    - 用Vantage Point Tree索引特征向量
    - 分类时，使用DBM-Tree进行相似度搜索

# 《Control Flow-Based Malware Variant Detection》(3)

- 实验结果

| | ao | b | d | e | g | k | m | q | a |
|---|---|---|---|---|---|---|---|---|---|
| ao | 1.00 | 0.60 | 0.35 | 0.38 | 0.45 | 0.74 | 0.60 | 0.60 | 0.73 |
| b | 0.60 | 1.00 | 0.46 | 0.50 | 0.37 | 0.73 | 0.95 | 0.96 | 0.73 |
| d | 0.35 | 0.46 | 1.00 | 0.64 | 0.59 | 0.36 | 0.46 | 0.46 | 0.35 |
| e | 0.38 | 0.50 | 0.64 | 1.00 | 0.61 | 0.42 | 0.49 | 0.50 | 0.40 |
| g | 0.45 | 0.37 | 0.59 | 0.61 | 1.00 | 0.47 | 0.37 | 0.37 | 0.46 |
| k | 0.74 | 0.73 | 0.36 | 0.42 | 0.47 | 1.00 | 0.73 | 0.72 | 0.86 |
| m | 0.60 | 0.95 | 0.46 | 0.49 | 0.37 | 0.73 | 1.00 | 0.96 | 0.72 |
| q | 0.60 | 0.96 | 0.46 | 0.50 | 0.37 | 0.72 | 0.96 | 1.00 | 0.72 |
| a | 0.73 | 0.73 | 0.35 | 0.40 | 0.46 | 0.86 | 0.72 | 0.72 | 1.00 |

*Levenshtein String Metric on Byte-level Content*

| | ao | b | d | e | g | k | m | q | a |
|---|---|---|---|---|---|---|---|---|---|
| ao | | 0.44 | 0.28 | 0.27 | 0.28 | 0.55 | 0.44 | 0.44 | 0.47 |
| b | 0.44 | | 0.27 | 0.27 | 0.27 | 0.51 | 1.00 | 1.00 | 0.58 |
| d | 0.28 | 0.27 | | 0.48 | 0.56 | 0.27 | 0.27 | 0.27 | 0.27 |
| e | 0.27 | 0.27 | 0.48 | | 0.59 | 0.27 | 0.27 | 0.27 | 0.27 |
| g | 0.28 | 0.27 | 0.56 | 0.59 | | 0.27 | 0.27 | 0.27 | 0.27 |
| k | 0.55 | 0.51 | 0.27 | 0.27 | 0.27 | | 0.51 | 0.51 | 0.75 |
| m | 0.44 | 1.00 | 0.27 | 0.27 | 0.27 | 0.51 | | 1.00 | 0.58 |
| q | 0.44 | 1.00 | 0.27 | 0.27 | 0.27 | 0.51 | 1.00 | | 0.58 |
| a | 0.47 | 0.58 | 0.27 | 0.27 | 0.27 | 0.75 | 0.58 | 0.58 | |

*Exact Matching*

| | ao | b | d | e | g | k | m | q | a |
|---|---|---|---|---|---|---|---|---|---|
| ao | | 0.86 | 0.53 | 0.64 | 0.59 | 0.86 | 0.86 | 0.86 | 0.86 |
| b | 0.88 | | 0.66 | 0.76 | 0.71 | 0.97 | 1.00 | 1.00 | 0.97 |
| d | 0.65 | 0.72 | | 0.88 | 0.93 | 0.73 | 0.72 | 0.72 | 0.73 |
| e | 0.72 | 0.80 | 0.87 | | 0.93 | 0.80 | 0.80 | 0.80 | 0.80 |
| g | 0.69 | 0.77 | 0.93 | 0.93 | | 0.77 | 0.77 | 0.77 | 0.77 |
| k | 0.88 | 0.97 | 0.67 | 0.77 | 0.72 | | 0.97 | 0.97 | 0.99 |
| m | 0.88 | 1.00 | 0.66 | 0.76 | 0.71 | 0.97 | | 1.00 | 0.97 |
| q | 0.88 | 1.00 | 0.66 | 0.76 | 0.71 | 0.97 | 1.00 | | 0.97 |
| a | 0.87 | 0.97 | 0.67 | 0.77 | 0.72 | 0.99 | 0.97 | 0.97 | |

*Q-Grams*

| | ao | b | d | e | g | k | m | q | a |
|---|---|---|---|---|---|---|---|---|---|
| ao | 1.00 | 0.70 | 0.42 | 0.42 | 0.44 | 0.72 | 0.70 | 0.70 | 0.70 |
| b | 0.70 | 1.00 | 0.47 | 0.47 | 0.48 | 0.94 | 1.00 | 1.00 | 0.93 |
| d | 0.42 | 0.47 | 1.00 | 0.71 | 0.80 | 0.48 | 0.47 | 0.47 | 0.48 |
| e | 0.42 | 0.47 | 0.71 | 1.00 | 0.72 | 0.47 | 0.47 | 0.47 | 0.47 |
| g | 0.44 | 0.48 | 0.80 | 0.72 | 1.00 | 0.49 | 0.48 | 0.48 | 0.50 |
| k | 0.72 | 0.94 | 0.48 | 0.47 | 0.49 | 1.00 | 0.94 | 0.94 | 0.96 |
| m | 0.70 | 1.00 | 0.47 | 0.47 | 0.48 | 0.94 | 1.00 | 1.00 | 0.93 |
| q | 0.70 | 1.00 | 0.47 | 0.47 | 0.48 | 0.94 | 1.00 | 1.00 | 0.93 |
| a | 0.70 | 0.93 | 0.48 | 0.47 | 0.50 | 0.96 | 0.93 | 0.93 | 1.00 |

*Levenshtein String Metric*

| | ao | b | d | e | g | k | m | q | a |
|---|---|---|---|---|---|---|---|---|---|
| ao | | 0.70 | 0.28 | 0.28 | 0.27 | 0.75 | 0.70 | 0.70 | 0.75 |
| b | 0.74 | | 0.31 | 0.34 | 0.33 | 0.82 | 1.00 | 1.00 | 0.87 |
| d | 0.28 | 0.29 | | 0.50 | 0.74 | 0.29 | 0.29 | 0.29 | 0.29 |
| e | 0.31 | 0.34 | 0.50 | | 0.64 | 0.32 | 0.34 | 0.34 | 0.33 |
| g | 0.27 | 0.33 | 0.74 | 0.64 | | 0.29 | 0.33 | 0.33 | 0.30 |
| k | 0.75 | 0.82 | 0.29 | 0.30 | 0.29 | | 0.82 | 0.82 | 0.96 |
| m | 0.74 | 1.00 | 0.31 | 0.34 | 0.33 | 0.82 | | 1.00 | 0.87 |
| q | 0.74 | 1.00 | 0.31 | 0.34 | 0.33 | 0.82 | 1.00 | | 0.87 |
| a | 0.75 | 0.87 | 0.30 | 0.31 | 0.30 | 0.96 | 0.87 | 0.87 | |

*Heuristic Approximate Matching*

| | ao | b | d | e | g | k | m | q | a |
|---|---|---|---|---|---|---|---|---|---|
| ao | | 0.86 | 0.49 | 0.54 | 0.50 | 0.87 | 0.86 | 0.86 | 0.86 |
| b | 0.87 | | 0.57 | 0.63 | 0.62 | 0.96 | 1.00 | 1.00 | 0.96 |
| d | 0.61 | 0.64 | | 0.85 | 0.91 | 0.64 | 0.64 | 0.64 | 0.64 |
| e | 0.64 | 0.69 | 0.85 | | 0.90 | 0.68 | 0.69 | 0.69 | 0.68 |
| g | 0.62 | 0.68 | 0.91 | 0.91 | | 0.68 | 0.68 | 0.68 | 0.68 |
| k | 0.88 | 0.96 | 0.58 | 0.62 | 0.61 | | 0.96 | 0.96 | 0.99 |
| m | 0.87 | 1.00 | 0.57 | 0.63 | 0.62 | 0.96 | | 1.00 | 0.96 |
| q | 0.87 | 1.00 | 0.57 | 0.63 | 0.62 | 0.96 | 1.00 | | 0.96 |
| a | 0.87 | 0.96 | 0.58 | 0.62 | 0.61 | 0.99 | 0.96 | 0.96 | |

*Optimal Distance Using Assignment Problem*

False Positives

| Classification Algorithm | Num. False Positives | FP Percentage |
|---|---|---|
| Q-Grams | 10 | 0.62 |
| Q-Grams + Optimal Distance | 7 | 0.43 |

Number of Malware Detected

| Algorithm | Klez | Netsky | Roron | Frethem |
|---|---|---|---|---|
| *Maximum* | *36* | *49* | *81* | *289* |
| Exact | 20 | 29 | 17 | 139 |
| Heuristic Approx. | 20 | 27 | 43 | 144 |
| Q-Grams | 20 | 31 | 79 | 226 |
| Optimal Distance | 22 | 46 | 73 | 220 |
| Q-Grams + Optimal | 20 | 43 | 73 | 217 |

# 分类研究进展


- [1]《Control Flow-Based Malware Variant Detection》
  - Cesare, S (Cesare, Silvio); Xiang, Y (Xiang, Yang); Zhou, WL (Zhou, Wanlei). IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, 2013
- [2]《Exploring Discriminatory Features for Automated Malware Classification》
  - Guanhua Yan, Nathan Brown, Deguang Kong. 2013 Proceedings of the 10th international conference on Detection of Intrusions and Malware, and Vulnerability Assessment.
- [3] Discriminant Malware Distance Learning on Structural Information for Automated Malware Classification.
  - Deguang Kong, Guanhua Yan. KDD '13: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.
- [4]《Malwise—An Effective and Efficient Classification System for Packed and Polymorphic Malware》
  - Silvio Cesare,Yang Xiang,Wanlei Zhou. IEEE Transactions on Computers,2013
- [5]《POSTER: Blind Separation of Benign and Malicious Events to Enable Accurate Malware Family Classification》
  - Hesham Mekky, Aziz Mohaisen, Zhi-Li Zhang. Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security
- [6]《Structural Classification and Similarity Measurement of Malware》
  - Shi, Hongbo; Hamagami, Tomoki; Yoshioka, Katsunari. IEEJ TRANSACTIONS ON ELECTRICAL AND ELECTRONIC ENGINEERING
- [7]《Mal-Netminer: Malware Classification based on Social Network Analysis of Call Graph》
  - Jae-wook Jang, Jiyoung Woo, Jaesung Yun, Huy Kang Kim. 2014, WWW Companion '14: Proceedings of the companion publication of the 23rd international conference on World wide web companion

# 《Discriminant Malware Distance Learning on Structural Information for Automated Malware Classification》 (1)

- 研究目的：怎样的特征作为恶意代码分类的依据效果好？
- 五种特征恶意代码特征
  - 16进制反汇编
  - 操作码和前缀码
  - 操作码
  - PE头
  - 系统调用
- 特征选择是一个经典问题，使用四种方法进行特征选取
  - ReliefF算法
  - Chi-squared算法
  - F-Statistics算法
  - L1-Regularized方法
- 使用四种分类算法对特征选择进行评价
  - Naive Bayes，K-NN，SVM，decision tree (C4.5)
  - 使用准确率，召回率和F1评价分类算法。

# 数据集

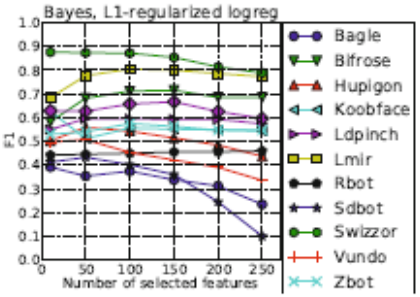
- 数据集
  - 使用Offensive Computing提供的数据库
  - 恶意代码挑选过程
    - 根据杀软描述确定恶意代码所属家族名
    - 挑选可信家族的恶意代码变种
    - 选取在四个杀软厂商中属于同一家族的恶意代码
  - 其中30%的恶意代码加壳
  - 挑选来自图中12个家族的26,848个恶意代码

**Table 1.** Alias resolution and malware selection

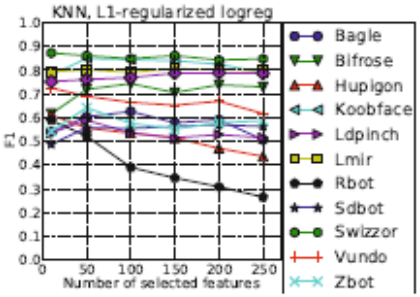
| Family | McAfee | Symantec | Microsoft | Kaspersky | NOD32 | Full | Unpacked |
|---|---|---|---|---|---|---|---|
| Bagle | Bagle | Beagle | Bagle | Bagle | Bagle | 285 | 152 |
| Bifrose | Backdoor-CEP | Bifrose | Bifrose | Bifrose | Bifrose | 2085 | 1677 |
| Hupigon | BackDoor-AWQ | Graybird | Hupigon | Hupigon | Hupigon | 11001 | 4748 |
| Koobface | Koobface | Koobface | Koobface | Koobface | Koobface | 439 | 371 |
| Ldpinch | PWS-Ldpinch | Ldpinch | Ldpinch | Ldpinch | Ldpinch | 310 | 190 |
| Lmir | PWS-Legmir | Lemir | Lemir | Lemir | Lmir | 366 | 181 |
| Rbot | Sdbot | Spybot | Rbot | Rbot | Rbot | 2565 | 923 |
| Sdbot | Sdbot | Sdbot | Sdbot | Sdbot | Sdbot | 629 | 253 |
| Swizzor | Swizzor | Lop | Swizzor | Swizzor | Swizzor | 1826 | 1276 |
| Vundo | Vundo | Vundo | Vundo | Monder | Virtumonde | 3278 | 2853 |
| Zbot | Zbot/PWS-Zbot | Zbot | Zbot | Zbot | Zbot | 1317 | 1233 |
| Zlob | Puper | Zlob | Zlob | Zlob | Zlob | 2747 | 2146 |

# 16进制N-Gram特征

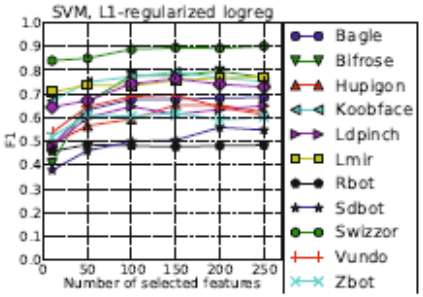
- 使用n-byte滑动窗户提取16进制文件中所有可能的n-byte序列。
- 计算所有可能序列出现的频率。
- 这些频率为16进制N-Gram特征。
- 1-gram特征top10: 00, 40, eb, 24, 10, 89, 8b, cc, 90, ff



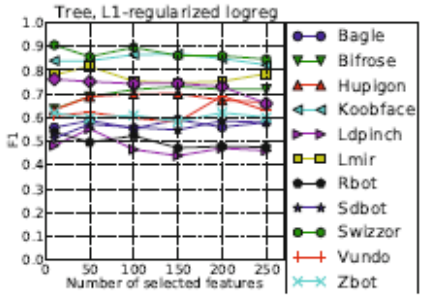

(1) 1-gram, Baves



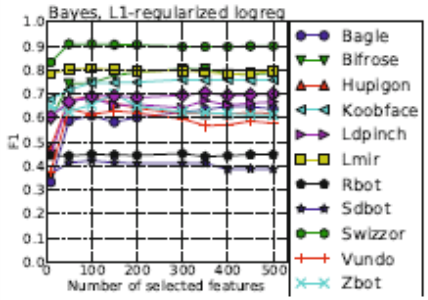

(2) 1-gram, kNN



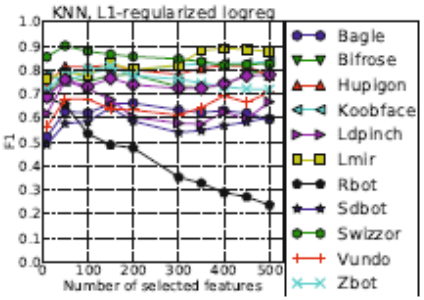

(3) 1-gram, SVM



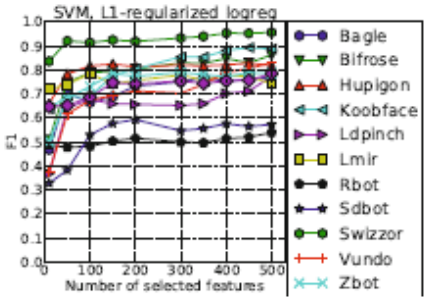

(4) 1-gram, Tree



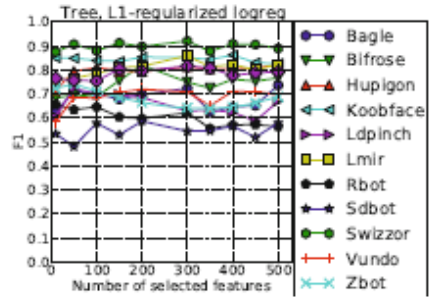

(5) 2-gram, Bayes




(6) 2-gram, kNN



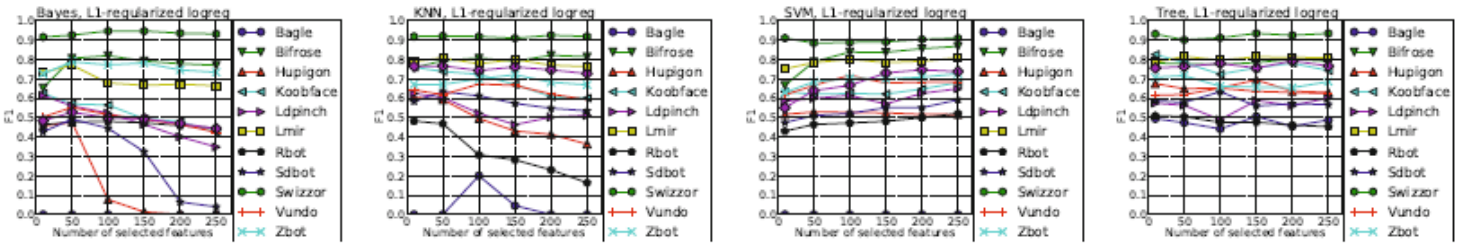

(7) 2-gram, SVM




(8) 2-gram, Tree

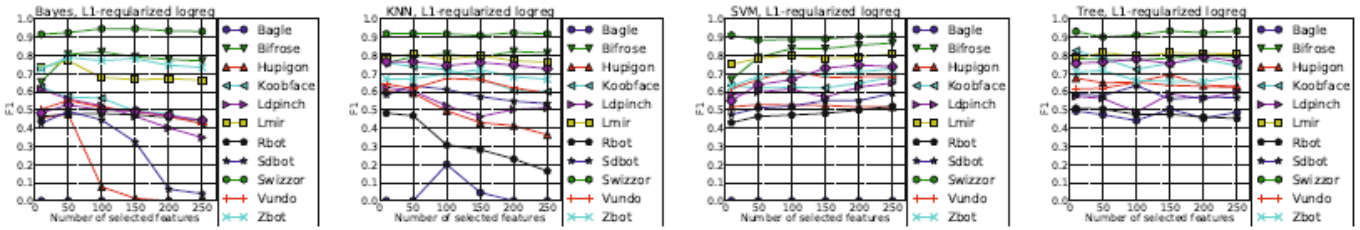
# Objdump

- 连接prefix和opcode作为特征。共有7259个特征
- 线性反汇编的特征top10： lea, jmp,push, add, pushl, cmp, insl, mov, int3, call




(1) Bayes (2) kNN (3) SVM (4) Tree

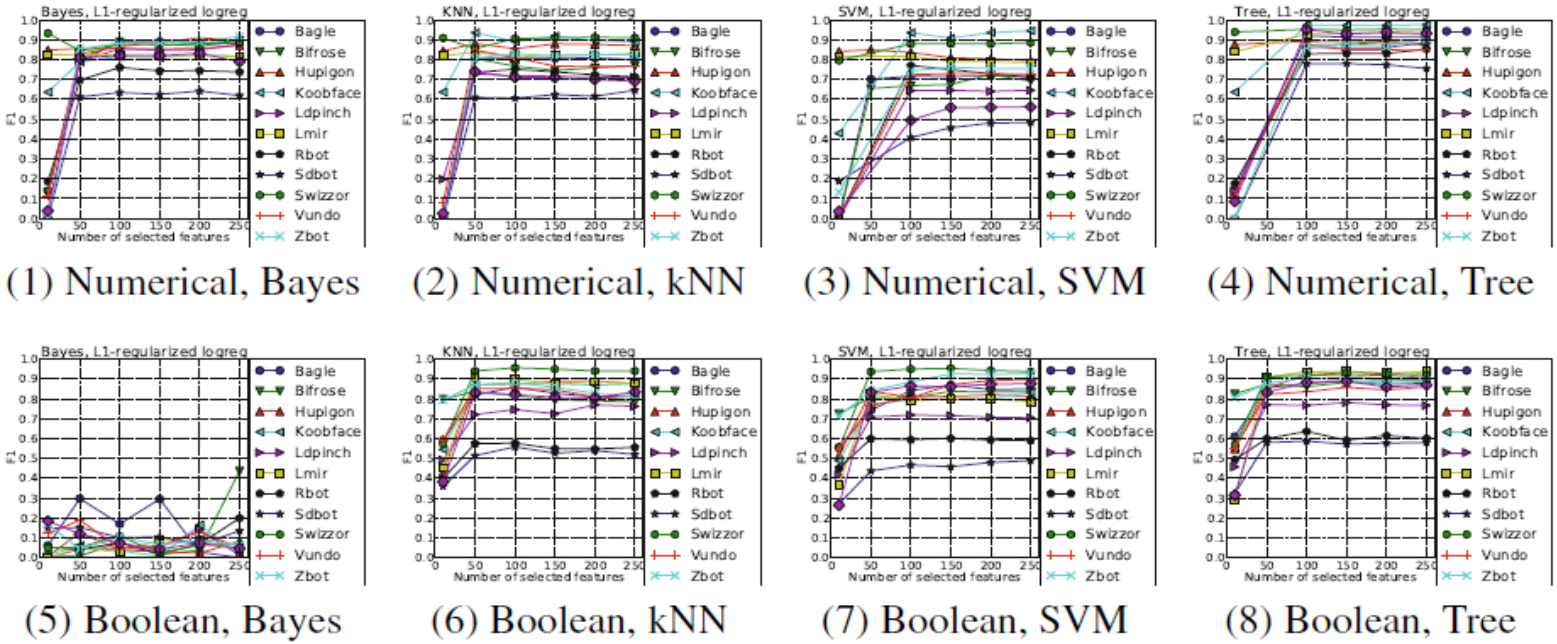
- 递归向下反汇编的特征top10： sub, add, nop, push, jmp, xor, lea, call, mov, dec




(1) Bayes (2) kNN (3) SVM (4) Tree

# PE头

- 数值型特征：基本包含PE头所有域，除去图像NameId字段和特征字段。 共422个特征
- 布尔型特征：特征字段的每个bit，DLL文件是否导入等信息。 共4167个特征。



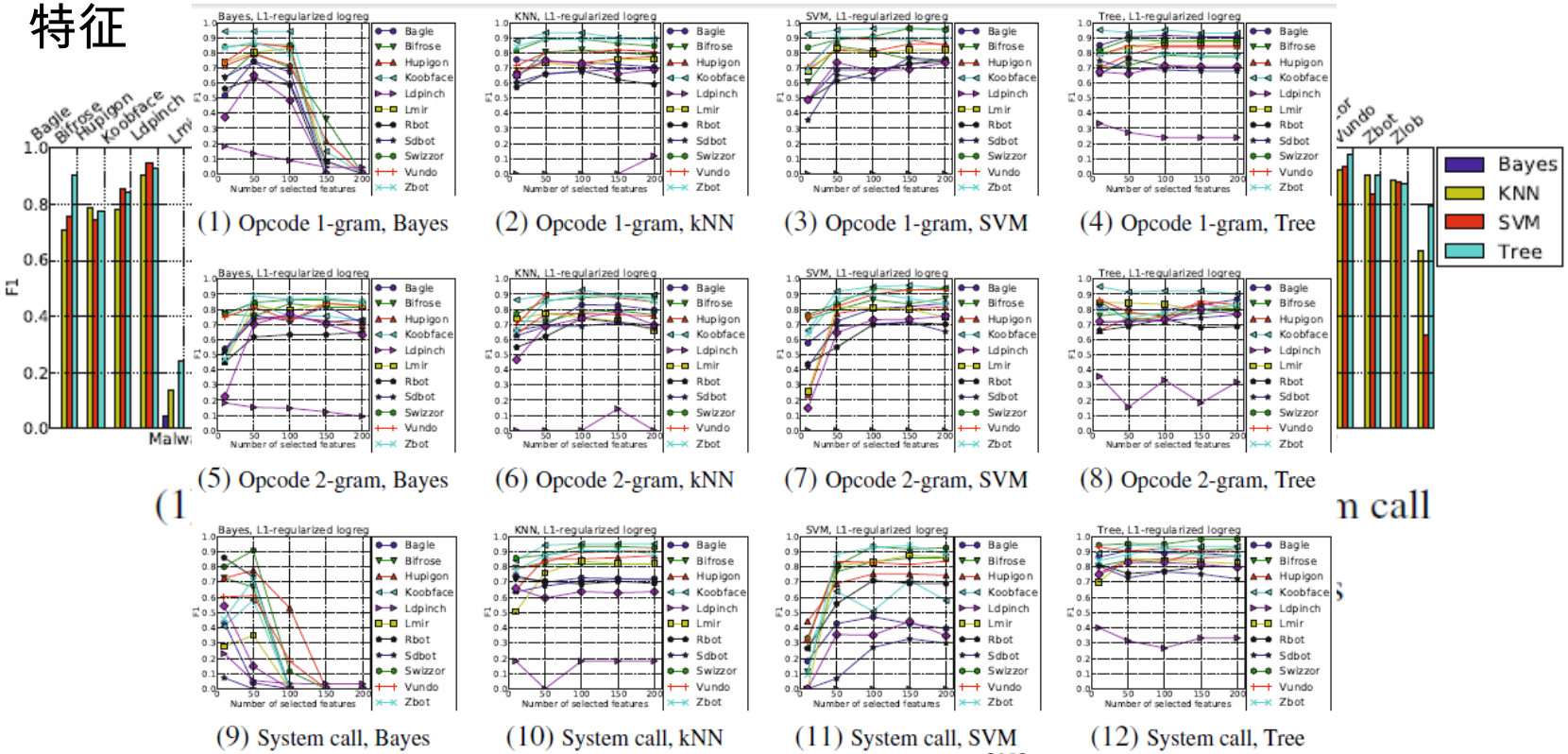

(1) Numerical, Bayes (2) Numerical, kNN (3) Numerical, SVM (4) Numerical, Tree

(5) Boolean, Bayes (6) Boolean, kNN (7) Boolean, SVM (8) Boolean, Tree

**Fig. 11.** Feature selection on PE Header features (L1-regularized logistic regression)

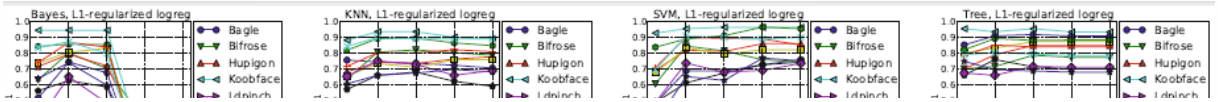
# 动态跟踪

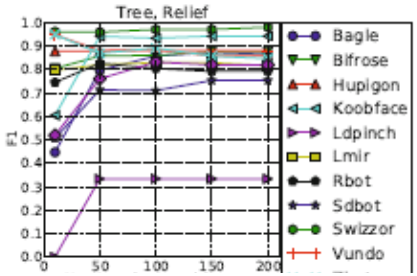
- 使用Intel Pin，只获得46 Lmir, 46 Sdbot and 13 LdPinch 样本的动态跟踪信息
- 从动态跟踪信息得到Opcode1-gram,Opcode 2-gram, 系统调用三种特征

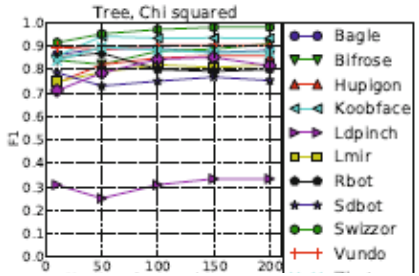


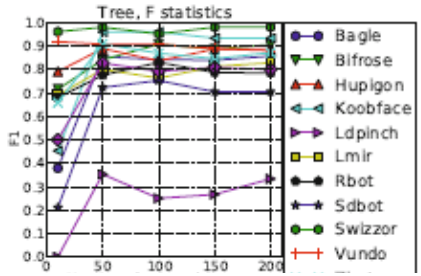
# 动态跟踪

- 使用Intel Pin，只获得46 Lmir, 46 Sdbot and 13 LdPinch 样本的动态跟踪信息
- 从动态跟踪信息得到Opcode1-gram,Opcode 2-gram, 系统调用三种特征



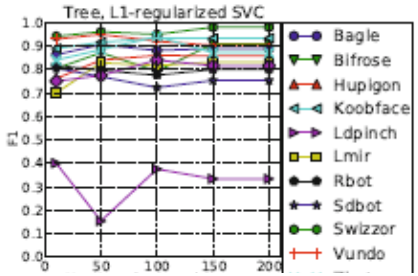




**Table 2.** Top 10 PIN trace features

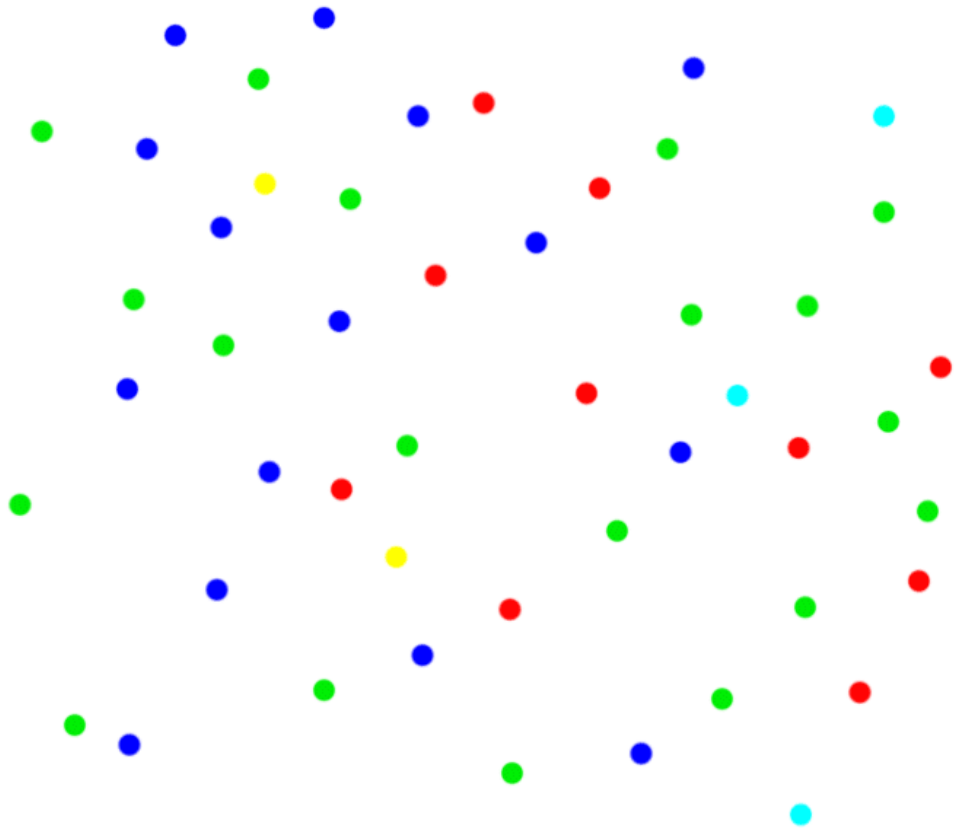
| | |
|---|---|
| 1-gram | add; jmp; mov; cmp; rep movsb; rep movsd; nop; and; xor; push; |
| 2-gram | nop,nop; rep movsb,rep movsb; push,mov; mov,jmp; mov,inc; inc,cmp; jmp,jmp; mov,mov; rep movsd,rep movsd; repne scasb,repne scasb; |
| System | _strcmpi; RtlInitAnsiString; RtlEnterCriticalSection; KiFastSystemCall; RtlAllocateHeap; RtlFreeHeap; NtSetEvent; RtlInitString; RtlNtStatusToDosError; NtPulseEvent |

# 分类研究总结

| | 依赖特征 | 分类算法 | 数据集 | 实验和结果 | 评价 |
|---|---|---|---|---|---|
| [1] | 控制流图（转换为字符串、n-gram向量） | N/A | mwcollect alliance网络下的蜜罐收集17,430个真实的恶意代码 | 误报率小于1%。系统可用于桌面系统和邮件网关。 | 优化了搜算算法，脱壳过程不能对所有加壳恶意代码进行有效、完整的脱壳。 |
| [2] | 16进制反汇编、操作码+前缀、操作码、PE文件头、动态行为 | Naive Bayes, kNN, SVM, and the decision tree (C4.5） | 12个家族的26,848恶意代码 | 对于用不同的权重计算方式的不同特征，使用四种不同的分类方法进行分类。 | 文章提供了五种特征对应四种分类算法的实验，提供了四种计算特征权重的方法<br>具有可靠性，可以用于未来的研究中。 |
| Malwise [3] | 控制流图（转换为字符串） | 字符串相似度比较 | 来自家族Netsky, Klez, Roron, Frethem的样本 | 系统的脱壳和检测效率较好。 | 有效对使用加壳工具的恶意代码进行脱壳，搜索算法不够优化。 |
| [4] | 网络流量（向量） | N/A | Darkness、Shady RAT (SRAT) | F1大于92%。ICA方法可以用于改善分类算法效果。 | ICA可能漏掉恶意代码流量，文章提出的方法只经过了初步验证，面对大量样本的可用性较差。 |
| [5] | DLL调用频率（向量） | 神经网络（GHSOM） | Nepenthes的215个恶意代码样本 | 恶意代码可以进行快速分类，提出的家族相似度计算可以评价不同结构中家族的相似度 | 用DLL调用频率作为输入少欠妥当，挑选合适的特征集可以使得分类算法有更大的实用性。 |
| Mal-Netminer [6] | 系统调用（图） | Naive Bayes、Boosted NB、RIPPER、RBF、C4.5、K-NN | 前期工作收集的2,523个恶意代码和123个良性样本 | 可以达到98%的准确率，第一次尝试使用系统调用图的结构进行分类。 | 使用系统调用图的组织结构进行恶意代码分类思想新颖，文中未提及系统效率，不确定可用性。 |

# 恶意代码聚类分析过程

- 家族样本特征提取
- 基于相似特征聚类样本
- 新样本家族判定



# 聚类研究进展

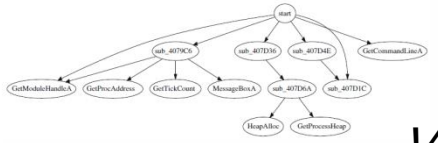
- [1] Classy: fast clustering streams of call-graphs. Kostakis, O (Kostakis, Orestis).
  - 2014, DATA MINING AND KNOWLEDGE DISCOVERY.
- [2] MutantX-S: scalable malware clustering based on static features.
  - Xin Hu, Sandeep Bhatkar, Kent Griffin, Kang G. Shin. Jun. 2013 Proceedings of the 2013 USENIX conference on Annual Technical Conference.
- [3] Clustering of Similar Malware Behavior via Structural Host-Sequence Comparison.
  - Horng-Tzer Wang ; Ching-Hao Mao ; Te-En Wei ; Hahn-Ming Lee. Computer Software and Applications Conference (COMPSAC), 2013 IEEE 37th Annual .
- [4] Poisoning Behavioral Malware Clustering.
  - Battista Biggio, Konrad Rieck, Davide Ariu, Christian Wressnegger, Igino Corona, Giorgio Giacinto, Fabio Roli .2014, AISec '14: Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop.

# Classy: fast clustering streams of call-graphs (1)

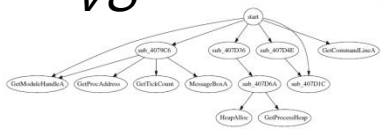
- F-Secure公司的研究员
- 高性能的近似图比对（基于编辑距离）
  - 模拟退火算法

    O(k · |Vmax|2 · dmax)
  - 性能下限算法

    O(n)



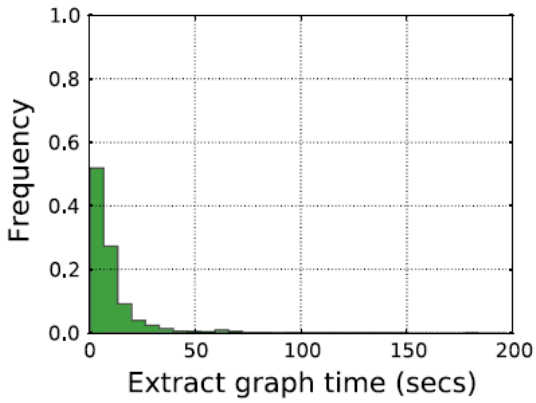
VS



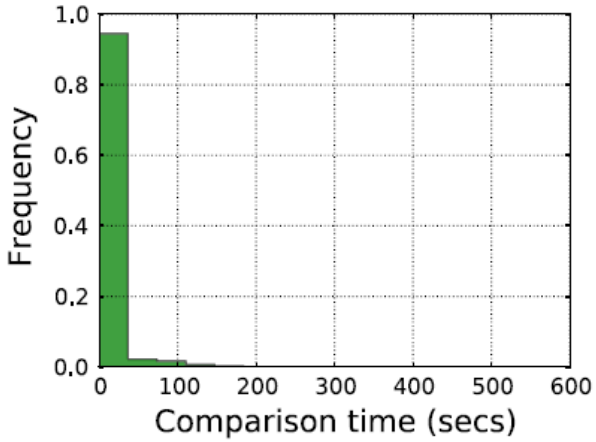
- 聚类算法（对新样本实时聚类）
  - 用快速的“性能下限算法”选择候选类
  - 用“模拟退火算法”在候选类中比较若干样本
  - 没有合适的类别则生成新的类

# Classy: fast clustering streams of call-graphs (2)

- 实验结果




平均提取时间约9秒




平均比较时间约23秒

48核cpu，每天可处理1.1万个函数调用图

| | | Contains clean files | |
|---|---|---|---|
| | | Yes | No |
| Cluster contains malware | Yes | 5.4 (12.42) % | 12.5 (10.2) % |
| | No | 82.1 (77.38) % | – |

# 聚类研究进展

- [1] Classy: fast clustering streams of call-graphs. Kostakis, O (Kostakis, Orestis).
  - 2014, DATA MINING AND KNOWLEDGE DISCOVERY.
- [2] MutantX-S: scalable malware clustering based on static features.
  - Xin Hu, Sandeep Bhatkar, Kent Griffin, Kang G. Shin. Jun. 2013 Proceedings of the 2013 USENIX conference on Annual Technical Conference.
- [3] Clustering of Similar Malware Behavior via Structural Host-Sequence Comparison.
  - Horng-Tzer Wang ; Ching-Hao Mao ; Te-En Wei ; Hahn-Ming Lee. Computer Software and Applications Conference (COMPSAC), 2013 IEEE 37th Annual .
- [4] Poisoning Behavioral Malware Clustering.
  - Battista Biggio, Konrad Rieck, Davide Ariu, Christian Wressnegger, Igino Corona, Giorgio Giacinto, Fabio Roli .2014, AISec '14: Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop.

# MutantX-S系统结构

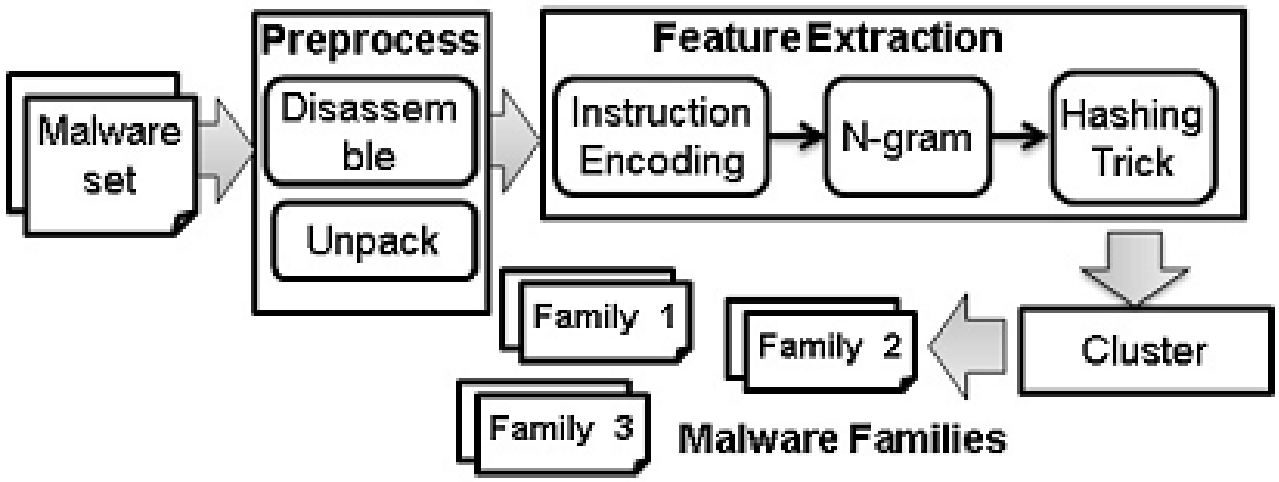
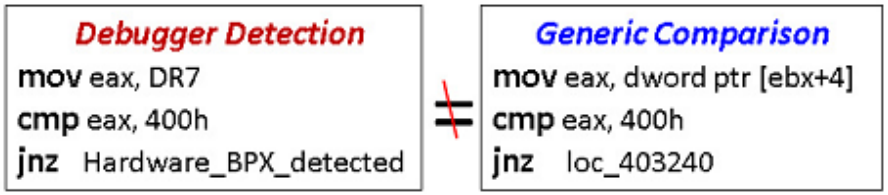



图1 MutantX-S系统概况

# 特征提取和聚类

## (1) Instruction Encoding

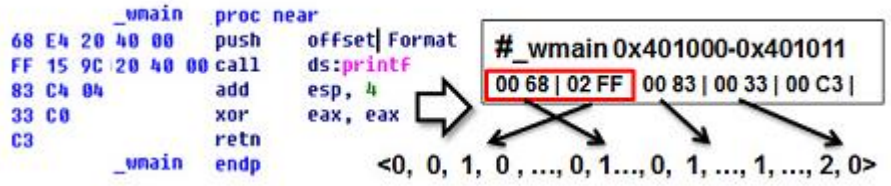
将每条指令转换成操作码系列，捕捉程序更基础的语义




不同的语义共享相同的指令助记符（move, cmp, jnz），用opcode特征标示其不同：“0F 21 3D 75”vs“8B 3D 75”

## (2)N-gram analysis

构建特征向量，计算程序相似性



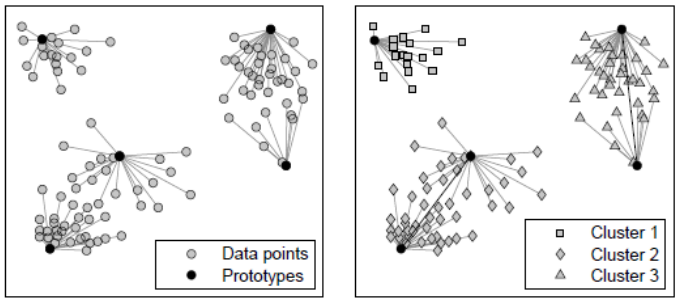

将函数转换成特征向量

## (3) Hashing Trick

对特征向量进行压缩处理，提高相似性计算的速度（精准性损失很少）

## (4) Prototype-Based Clustering

在多个小规模的原始数据子集模型下进行聚类，明显减少计算的时间开销

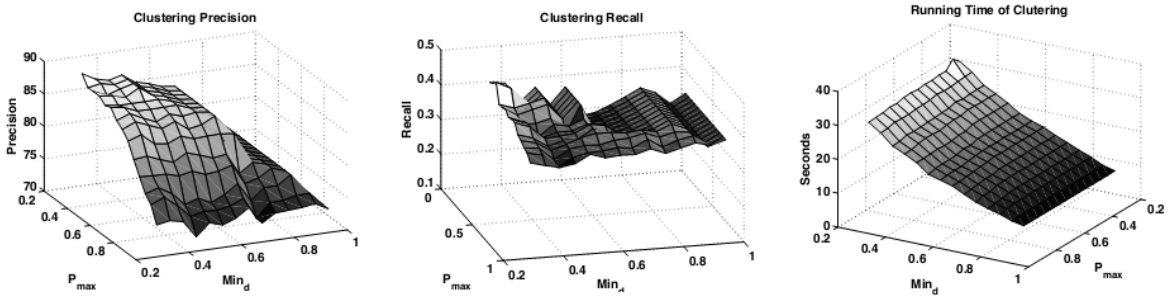



(a) Prototypes (b) Clustering

# 实验

| Family | # | Family | # | Family | # |
|---|---|---|---|---|---|
| Pilleuz | 500 | Bredolab | 301 | Tidserv | 59 |
| Koobface | 496 | Vundo | 249 | Waledac | 34 |
| Silly | 489 | Almanahe | 241 | Ackantta | 32 |
| Fakeav | 489 | Sasfis | 199 | Mebroot | 26 |
| Zbot | 459 | Graybird | 166 | Hotbar | 21 |
| Banker | 449 | Gammima | 126 | Qakbot | 17 |
| Virut | 361 | Mabezat | 107 | | |

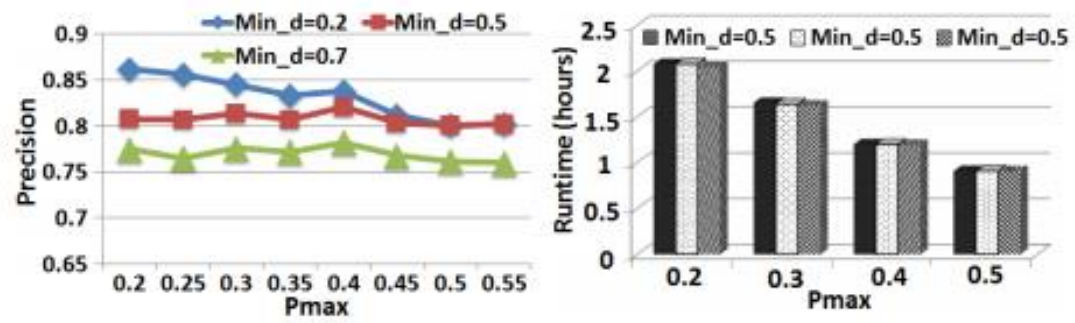
恶意代码家族参考数据集




系统的准确率、召回率和运行时间

| Packer | Diff in IC (%) | NG Dist | Packer | Diff in IC (%) | NG Dist |
|---|---|---|---|---|---|
| PEcompact | 0.88% | 0.068 | ASprotect | 6.70% | 0.133 |
| EXECryptor | 3.20% | 0.176 | UPX | 0.88% | 0.068 |
| EXEStealth | 0.88% | 0.071 | NSPack | 0.87% | 0.069 |
| VMprotect | 2.50% | 0.10 | Armadillo | - | - |

脱壳有效性（前后对比）
（IC：指令数量； NG Dist: N – gram特征向量距离）




对13万个样本进行聚类的准确率和时间：少于1.5小时，准确率接近0.82

# 聚类研究总结

| | 依赖特征 | 聚类算法 | 数据集 | 性能 | 评价 |
|---|---|---|---|---|---|
| Classy[1] | 属性函数调用图 | 改进的BALLS | F-Secure | 平均30+秒/样本 | 最接近实用，可与产业结合 |
| MutantX-S[2] | Opcode的n-gram向量 | Prototype-Based Clustering | 病毒厂商 | 10-30+秒/样本 | 接近实用 |
| [3] | 行为序列的Markov 链 | k-means | CWSandbox | 较差 | 算法探索 |
| [4] | | | | | 研究如何毒害聚类算法 |

# 学术研究的前景是否很美好？

- 研究挑战
  - 海量样本
    - 目前存储样本数量逐步逼近一亿，每天新增样本数平均约为20万
    - 对算法可扩展性的挑战
  - 样本的不平衡
    - 木马程序Zbot的数量达到了35万，而Flame病毒只有57个样本
    - 如何具有大海捞针的能力
  - 可自动化
    - 算法依赖的特征可被自动提取
  - 高性能
    - 并行化计算，大数据挖掘
  - 测试数据集
    - 权威的测试数据集和评测标准

# 总结

- 现有恶意代码检测分析体系的不足
  - 泛黑类别的“黑洞”效应
  - 家族识别困难
  - 限制了AV厂商后端分析前置到企业用户应用部署
- 分类和聚类的需求
  - 提高家族的检测率，抵抗变种
  - 识别新家族，尤其是APT家族
  - 为人工分析减轻负担
- 恶意代码分析体系的进化方向之一，是融合学术界和产业界的工作
- 近两年的学术界研究，似乎正在走出象牙塔，与产业界和结合一下子似乎又看到了希望

谢谢！

欢迎交流与合作研究

ytang@nudt.edu.cn