A Static Birthmark of Binary Executables Based on API Call Structure

2007.12.9

Seokwoo Choi, Heewan Park, Hyun-il Lim, Taisook Han
Programming Languages Lab, EECS,
Korea Advanced Institute of Science and Technology

Contents

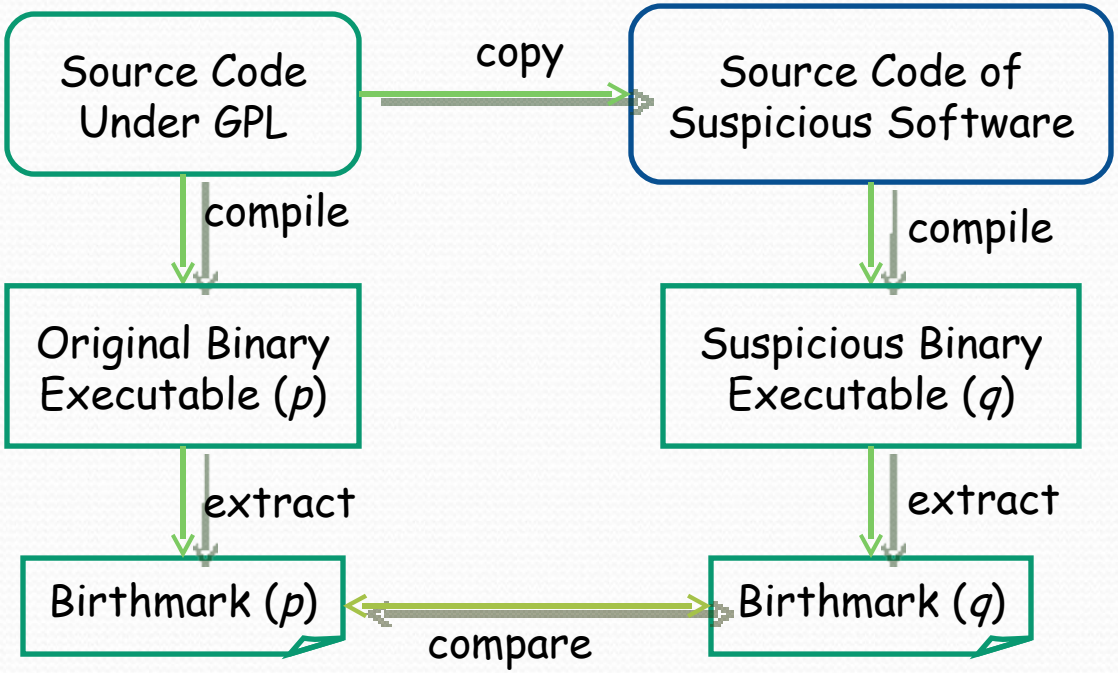
- Introduction
- Static API Birthmark
  - Extraction
  - Similarity Calculation
- Evaluation
  - Implementation
  - Experiment
  - Result
- Conclusion

Software Birthmark

- Inherent characteristics of a program extracted from the program itself for program identification

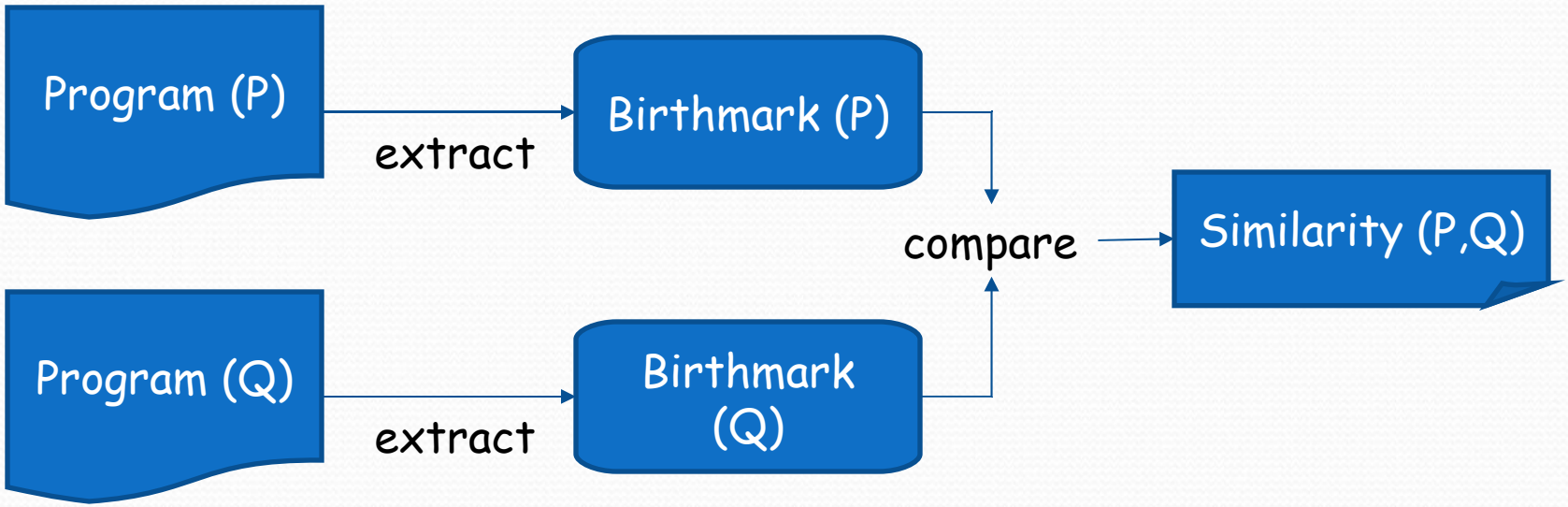
- Software birthmark can be used as an evidence that a program is a copy of another

Software Birthmark (cont)



Software Birthmark (cont)

- Comparing two programs with birthmark



Static API Birthmark

- "A Static API Birthmark based on API call Structure"

- Features
  - API call
  - Static
  - call Structure

Static API Birthmark
- Windows API

- Windows API function calls as Birthmark
  - API function calls are essential to build Windows applications
  - Windows API functions are very frequently called

Static API Birthmark
- Static Birthmark

- Dynamic - API calls at runtime
  - Depends on inputs and OS environments
  - Appropriate for non-Interactive programs
    - Image readers, XML parsers [Schuler 2007]
    - MP3 tag editors with no inputs [Tamada 2007]

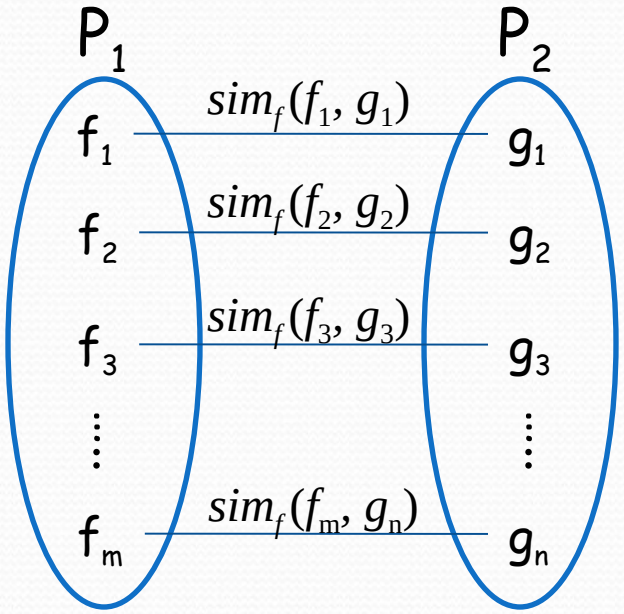
- Static - API calls in program itself
  - Applicable for Interactive programs
    - Sequence of method calls [Tamada 2005]

Static API Birthmark

- Comparing two programs

- Program = set of functions
- Match functions between two programs
  - Calculate function similarity by comparing function birthmarks
- Calculate program similarity with function similarities

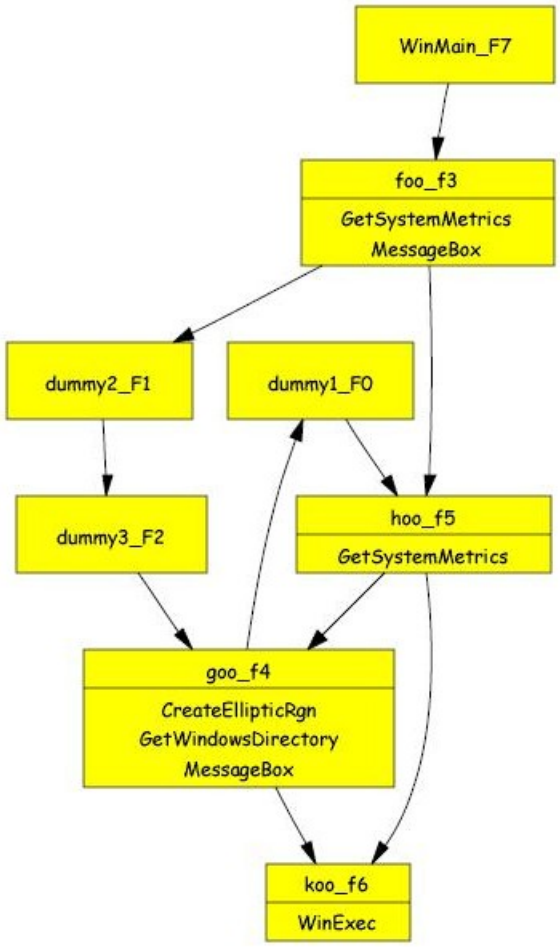
Program → Partition
Function → Node
Edge → Matching
Weight → Similarity



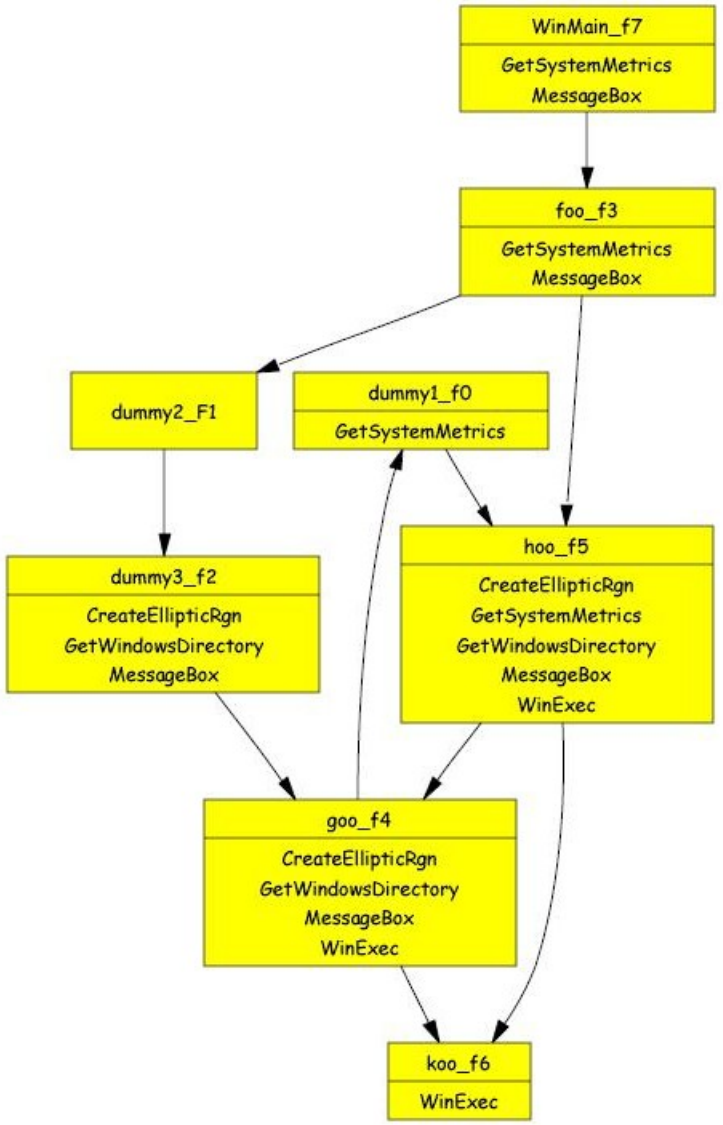
Static API Birthmark
- Function Birthmark

- Birthmark Structure (for a function)
  - Sequence of API calls
  - Set of API calls
  - Control flow graphs with API calls

- Function Birthmark (with call depth)
  - Set of API function calls found in a function f and its sub-functions of which the call depths are within k from f



function birthmarks with k=0



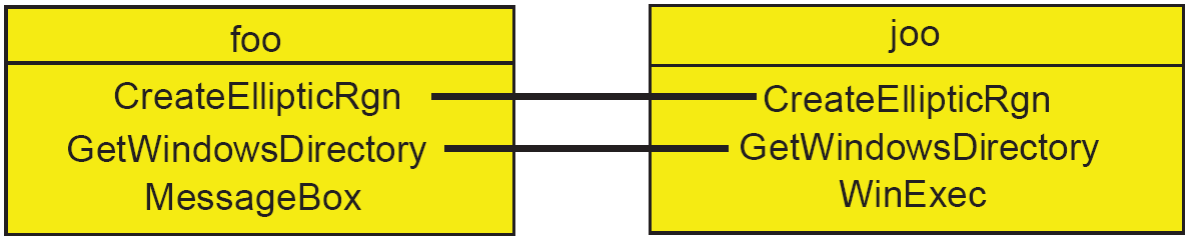
function birthmarks with k=1

Static API Birthmark

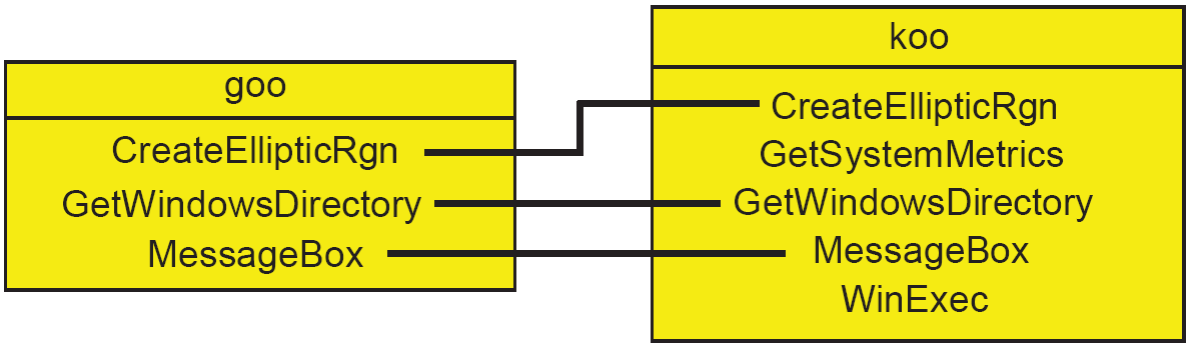
- Function Similarity

- Let f_1, f_2 be functions, function similarity $sim_f(f_1, f_2)$ is defined as

$$sim_f(f_1, f_2) = \frac{2 \times \left| Birthmark_f(f_1) \cap Birthmark_f(f_2) \right|}{\left| Birthmark_f(f_1) \right| + \left| Birthmark_f(f_2) \right|}$$



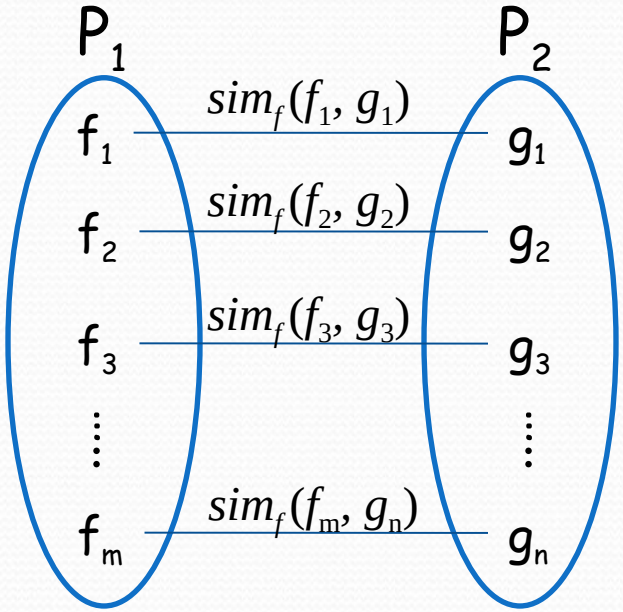
$sim_f(foo, joo) = 2 * 2 / (3+3) = 2/3$



$sim_f(goo, koo) = 2 * 3 / (3+5) = 3/4$

Static API Birthmark
- Function Matching

- Function Matching between Two Programs
  ≈ Weighted Bipartite Matching
- Maximum Weighted Bipartite Matching



Static API Birthmark

- Program Similarity

- Let P_1, P_2 be programs and N_1, N_2 be the numbers of functions of P_1, P_2. Program similarity $sim_p(P_1, P_2)$ is defined as

$$sim_p(P_1, P_2) = \frac{2 \times \sum_{(f_1, f_2) \in match(P_1, P_2)} sim_f(f_1, f_2)}{N_1 + N_2}$$

  where $match(P_1, P_2)$ is a set of matched functions between P1 and P2 which maximize $sim_p(P_1, P_2)$

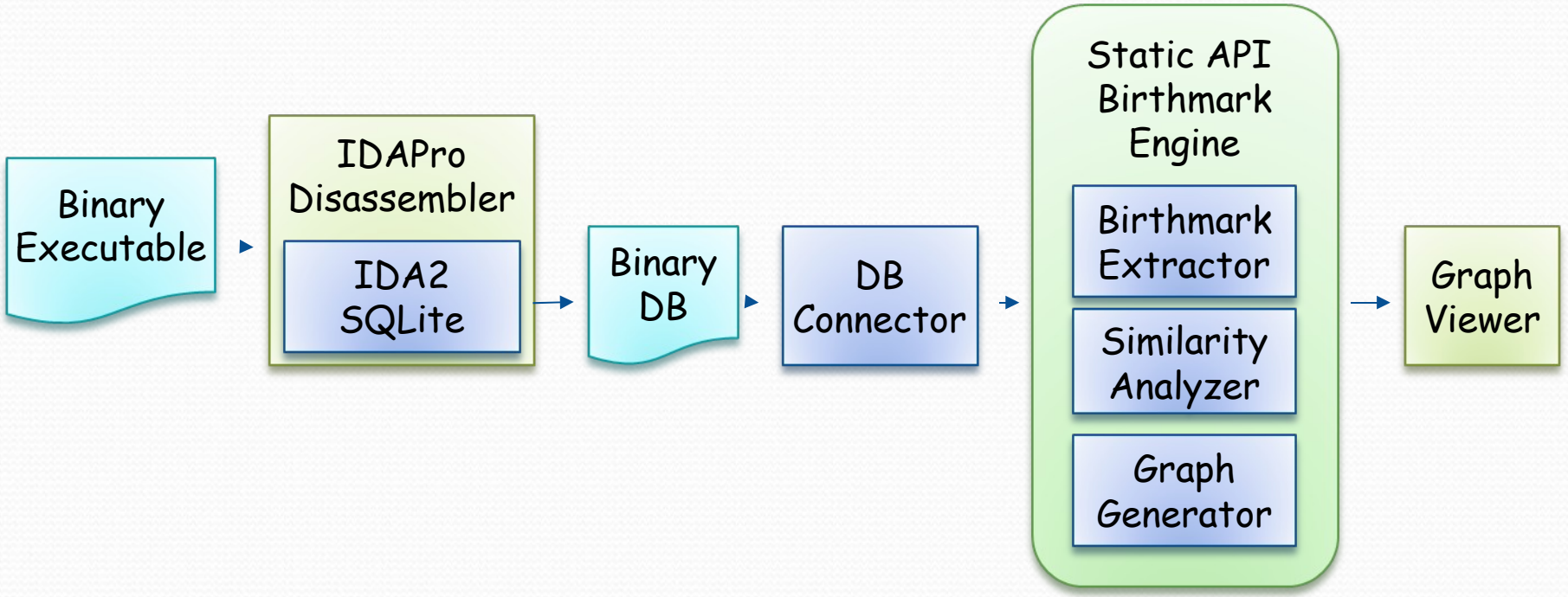
- Applied Hungarian ($O(n^3)$) algorithm to maximize similarity.

Static API Birthmark

- Judgment of Copy

Similarity	Decision
$0.00 \leq sim_{\mathrm{P}}(P_1,P_2) < 0.35$	Independent (Not Copy)
$0.35 \leq sim_{\mathrm{P}}(P_1,P_2) < 0.70$	Inconclusive
$0.70 \leq sim_{\mathrm{P}}(P_1,P_2) \leq 1.00$	Similar (Copy)

Static API Birthmark System



Evaluation

- Evaluation Criteria for Birthmarks

- Credibility
  - Ability to say that different programs are different and similar programs are similar
- Resilience
  - Ability to resist program transformation such as code obfuscation and compiler optimizations

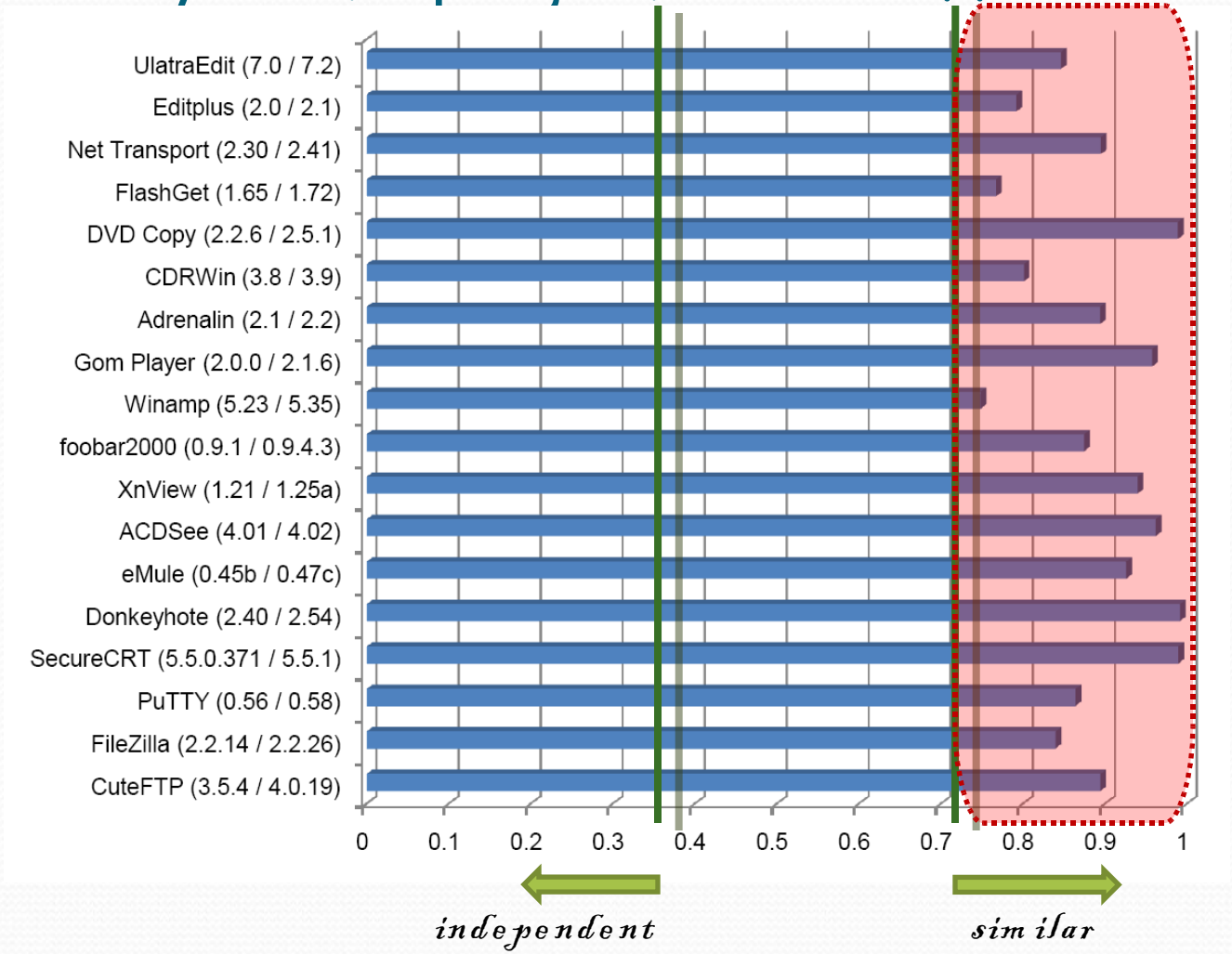
Experiment

- Sample Programs (10 categories, 39 executables)

Category	Program 1	Program 2
Text Editor	UltraEdit 7.0 / 7.2	EditPlus 2.0 / 2.1
FTP Client	FileZilla 2.2.14 / 2.2.26	Cute FTP 3.5.4 / 4.0.19
Terminal	Putty 0.56 / 0.58	Secure CRT 5.5.0 / 5.5.1
P2P Client	Donkeyhote 2.40 / 2.54	eMule 0.45b / 0.47c
Image Viewer	ACDSee 4.01 / 4.02	XnView 1.21 / 1.25a
Audio Player	Winamp 5.23 / 5.35	foobar2000 0.9.1 / 0.9.4
Video Player	GOMPlayer 2.0.0 / 2.1.6	Adrenalin 2.1 / 2.2
CD Burner	CDRWin 3.8 / 3.9	DVDCopy 2.2.6 / 2.5.1
Download Manager	Flashget 1.6.5 / 1.7.2	NetTransport 2.3.0 / 2.4.1
Disk Image Emulator	Daemon Tools 4.3.0 / 4.9.0	CDSpace 5.0

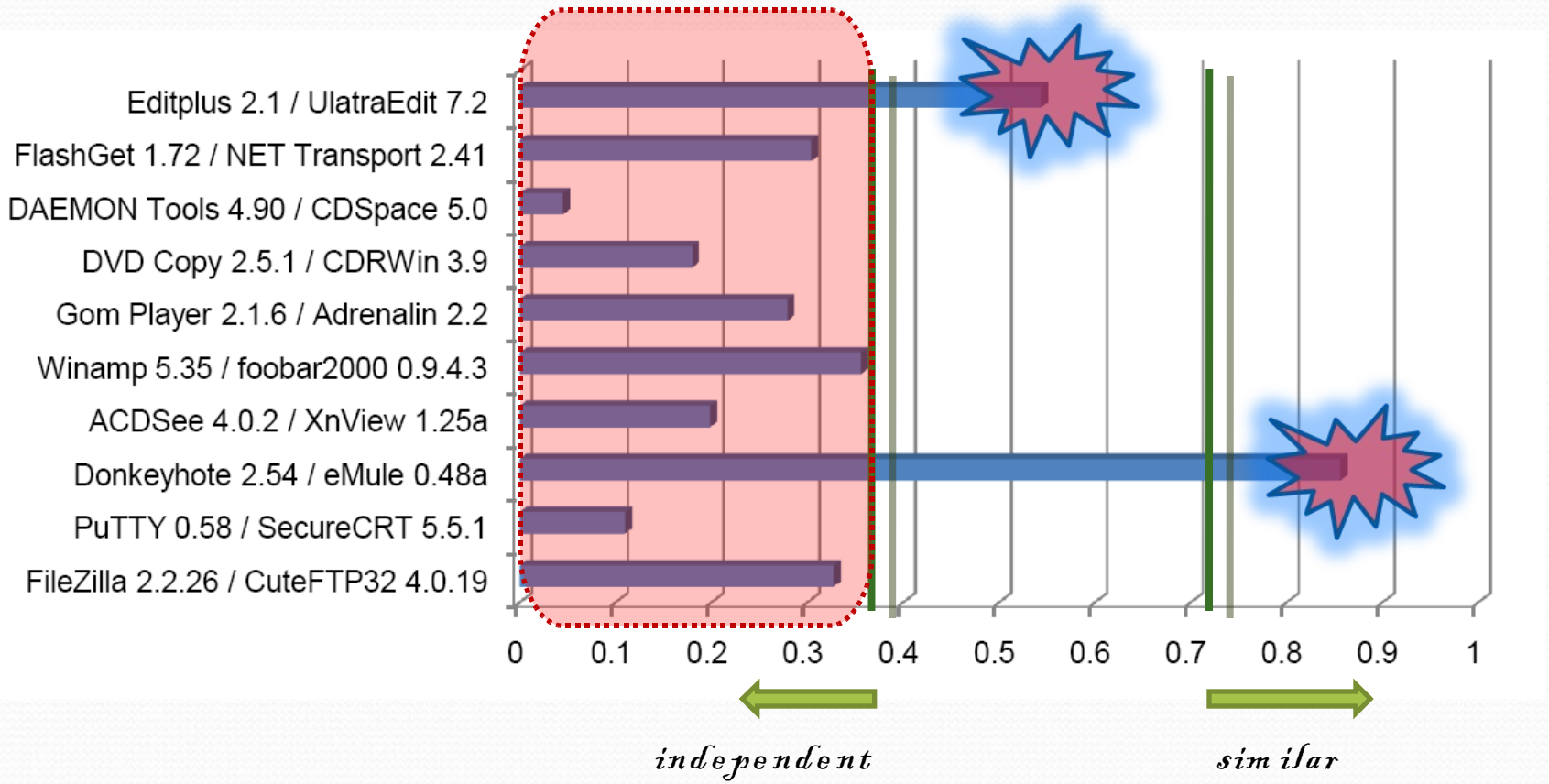
Result

- Credibility : same programs with different versions



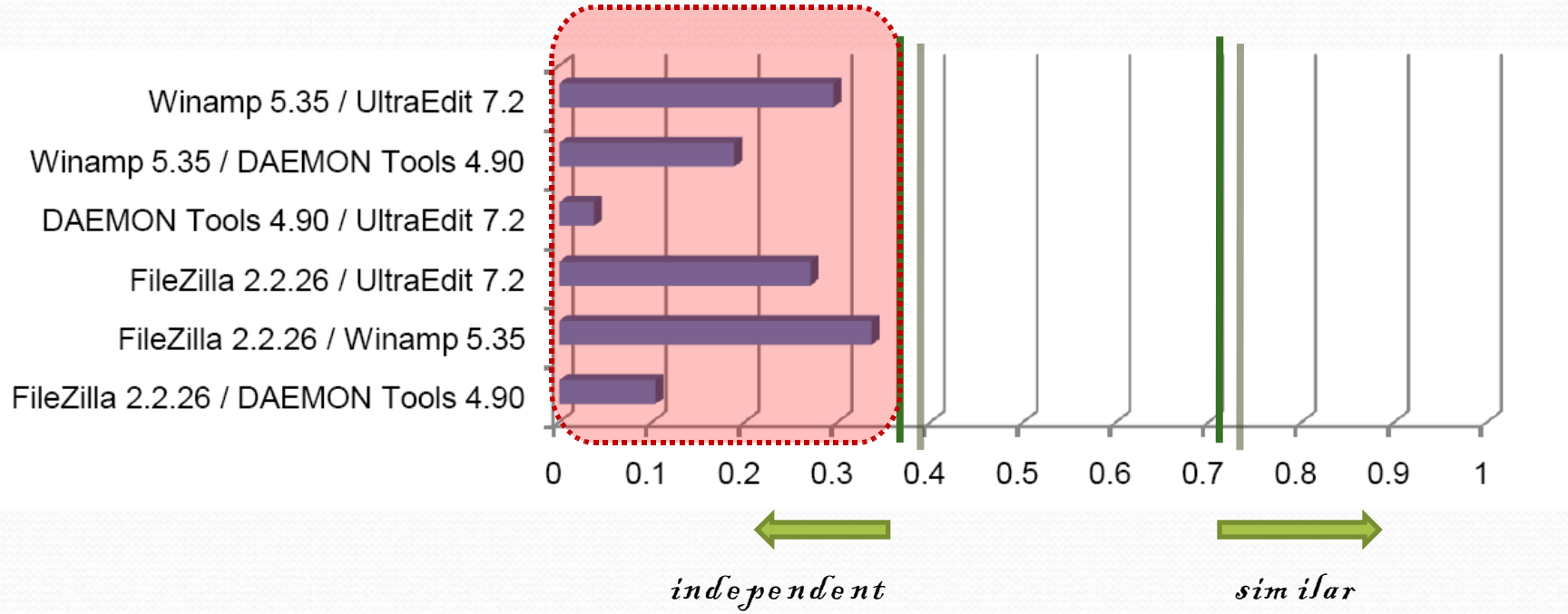
Result

- Credibility : programs with same category



Result

- Credibility : programs with different category



Result

- Resilience : different compilers

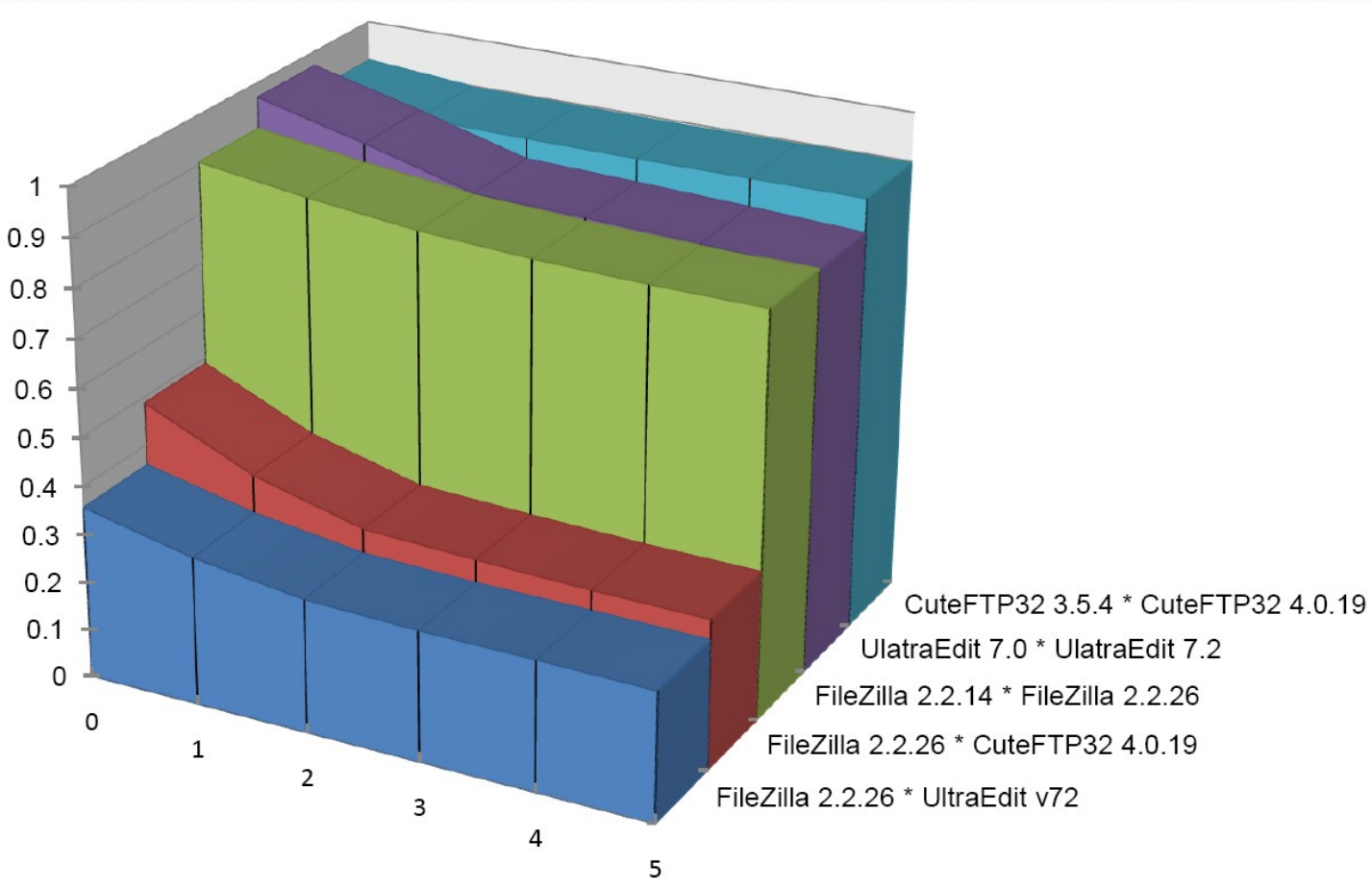
- Sample program : fr-hed hex editor
- Compilers
  - Microsoft Visual Studio 6.0 (VC6.0)
  - Microsoft Visual Studio 2003 (VC7.0)
  - Microsoft Visual Studio 2005 (VC8.0)

		VC6		VC7		VC8	
		Debug	Release	Debug	Release	Debug	Release
VC6	Debug	1.0000	0.9823	0.9809	0.9767	0.9694	0.9797
	Release		1.0000	0.9751	0.9755	0.9590	0.9780
VC7	Debug			1.0000	0.9900	0.9793	0.9924
	Release				1.0000	0.9857	0.9977
VC8	Debug					1.0000	0.9881
	Release						1.0000

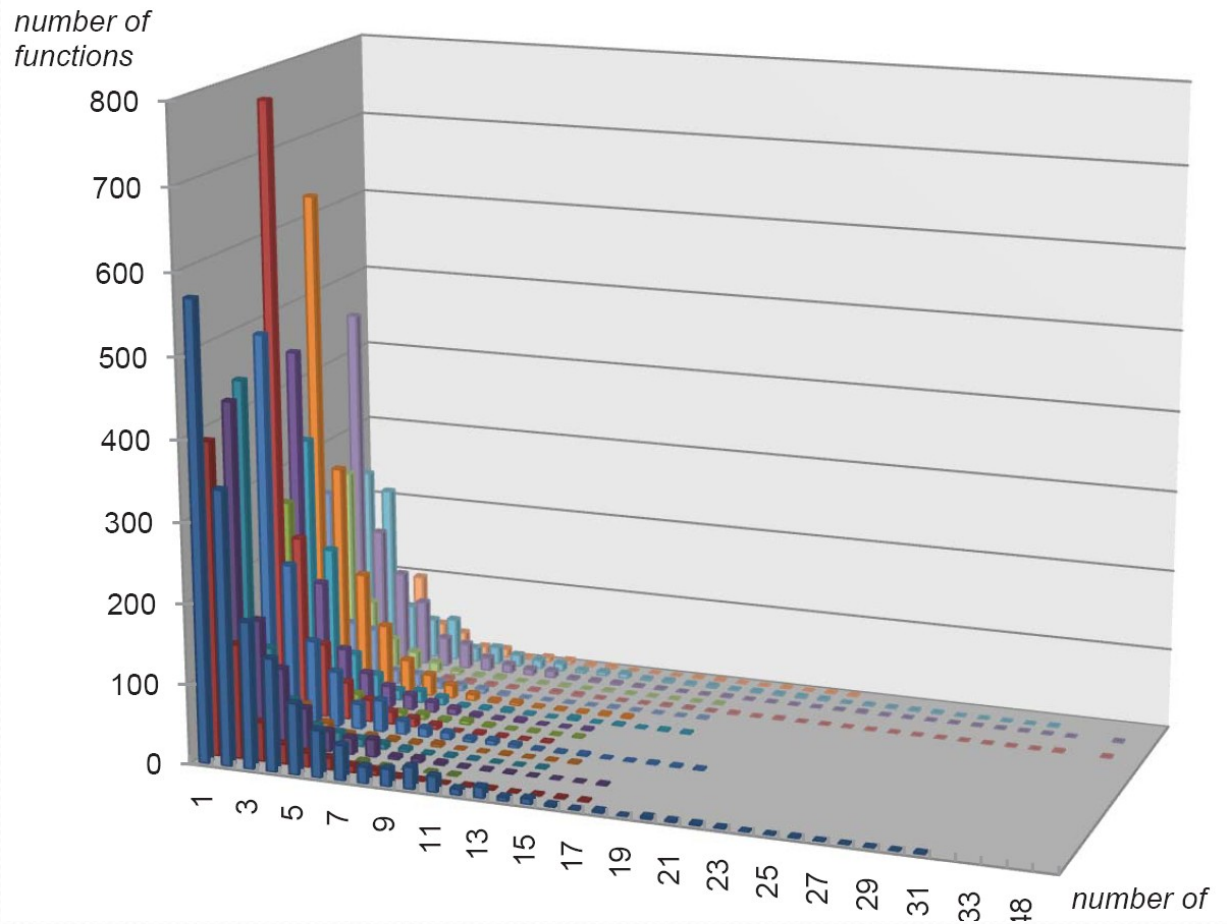
Conclusion

- We presented a static API birthmark technique for Windows programs
- Our birthmark can be applied to real Windows applications with appropriate credibility and resilience
- Limitation
  - Handle API using programs
  - Dependent on IDAPro for API function calls

Similarity Change according to call depth



API distribution of sample programs



Visualization