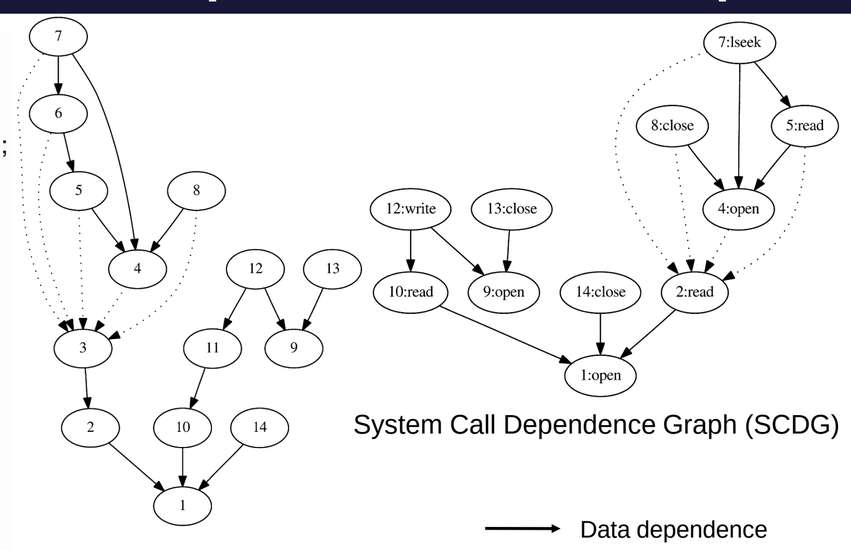
# Behavior Based Software Theft Detection

Xinran Wang, Yoon-Chan Jhi, Sencun Zhu, and Peng Liu

- A plagiarizer or a company might steal code from open source projects and/or its competitors.
- Existing techniques are limited in meeting the following requirements:
  - Resiliency to code obfuscation techniques
  - Capability to detect component theft
  - Scalability to detect large-scale programs
  - Applicability to binary executables
- System call is the only way for a program to talk with kernel (hard to evade or obfuscate)

## System Call Dependence Graph

```
1 :  fd1 = open(path1,"r",...);
2 :  read(fd1, buf, ...);
3 :  if (buf == "1") {
4 :      fd2 = open(path2, "r", ...);
5 :      n = read(fd2,buf,...);
6 :      offset = n + 10;
7 :      lseek(fd2,offset,...);
         ...
8 :      close(fd2);
     }
9 :  fd3 = open(path3,"w",...);
10:  read(fd1, buf, ...);
11:  strcpy(outbuf,buf);
12:  write(fd3,outbuf);
13:  close(fd3);
14:  close(fd1);
```



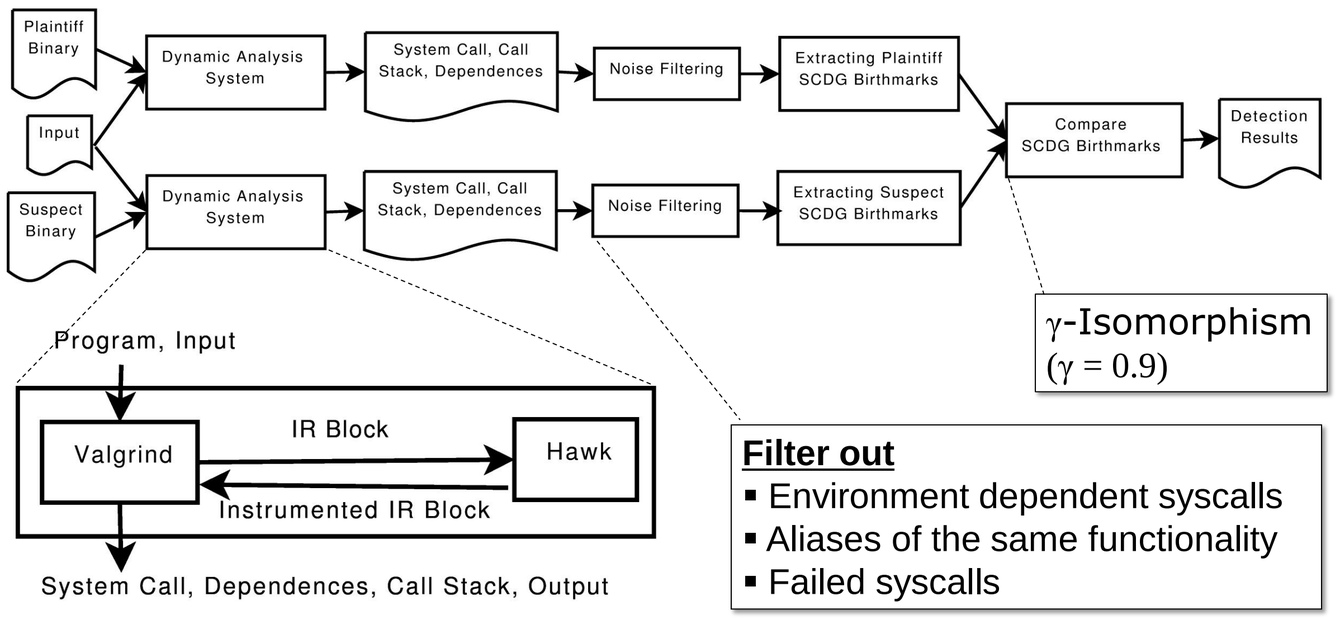

System Call Dependence Graph (SCDG)

Dynamic Dependence Graph (DDG)

Data dependence

Control dependence

## System Design




γ-Isomorphism ($\gamma = 0.9$)

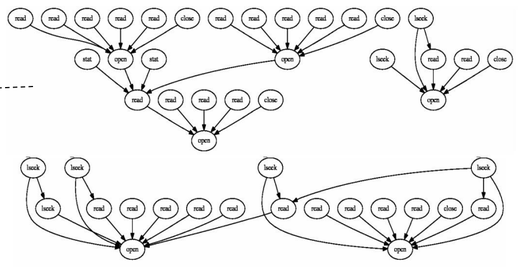
**Filter out**

- Environment dependent syscalls
- Aliases of the same functionality
- Failed syscalls

## Birthmark Extraction

- *For the plaintiff program*, partition SCDG based on dynamic call graph → The subgraphs form *System Call Dependence Graph Birthmark* (SCDGB)
- *For the suspect program*, SCDGB is computed from the whole SCDG
- Remove SCDGBs commonly observed in other programs (Training set is required)

SCDGB of the Gecko HTML rendering engine



## Evaluation

### I. Component Detection

- Test program set

| Program | Version | Type | SCDG Node # | SCDG Edge # |
|---|---|---|---|---|
| Flock | 2.0.3 | Web Browser | 21337 | 9343 |
| Epiphany | 2.22.2 | Web Browser | 16864 | 9011 |
| Konqueror | 3.5.10 | Web Browser | 11850 | 5589 |
| Amaya | 10 | Web Browser | 42701 | 23958 |
| Opera | 9.52 | Web Browser | 58485 | 21361 |
| Songbird | 1.1.2 | Web Browser | 37103 | 25547 |
| Galeon | 2.0.7 | Web Browser | 19825 | 7450 |
| AbiWord | 2.4.6 | Word Processor | 12975 | 5642 |
| KWord | 1.6.3 | Word Processor | 15408 | 6630 |
| LyX | 1.5.3 | Latex Editor | 21977 | 18656 |
| Texmaker | 1.6 | Latex Editor | 6897 | 3223 |
| Kile | 2.0.0 | Latex Editor | 50937 | 24615 |
| Gedit | 2.22.3 | Text Editor | 25113 | 5867 |
| Bluefish | 1.0.7 | Text Editor | 10952 | 3502 |
| GNU Emacs | 22.2.1 | Text Editor | 14807 | 4734 |
| Vim | 7.1.138 | Text Editor | 2582 | 1979 |
| Pidgin | 2.5.2 | Messenger | 10816 | 8014 |
| Kopete | 0.12.7 | Messenger | 16319 | 7144 |
| Kmess | 1.5 | Messenger | 10830 | 6247 |
| GnoCHM | 0.9.9 | CHM Viewer | 21191 | 8354 |
| Evince | 2.22.2 | Doc. Viewer | 16179 | 7095 |
| GV | 3.6.3 | Doc. Viewer | 6508 | 3267 |
| Quod Libet | 1.0 | Media Player | 15839 | 10725 |
| Evolution | 2.22.3 | Email Client | 13798 | 6787 |

- GNU Aspell
  - Opera, Kword, Lyx, Bluefish, Pidgin
  - *0 false positives/negatives*
- The Gecko HTML rendering engine
  - Flock, Epiphany, SongBird and Galeon
  - *0 false positives/negatives*

### II. Code Transformation Resiliency

- Test program set
  - gzip, oggenc, bzip2
- Impact of Compiler Optimization Levels
  - Tested -O0, -O1, -O2, -O3, and -Os of GCC
  - *One less "write" in oggenc with –O3 and -Os*
- Impact of Different Compilers
  - Tested GCC, TinyCC, Watcom C
  - *No change in SCDGB*
- Impact of Obfuscation Techniques
  - Tested Semantic Design Inc.'s C obfuscator, Control Flow Flattening in Loco/Diablo
  - *No change in SCDGB*

## Publications


- X. Wang, Y-C Jhi, S. Zhu, P. Liu. *Behavior Based Software Theft Detection* In Proc. of the 16th ACM Conference on Computer and Communications Security (CCS '09) (To appear)
- X. Wang, Y-C Jhi, S. Zhu, P. Liu. *Detecting Software Theft via System Call Based Birthmarks* In Proc. of the 25th Annual Computer Security Applications Conference (ACSAC '09) (To appear)