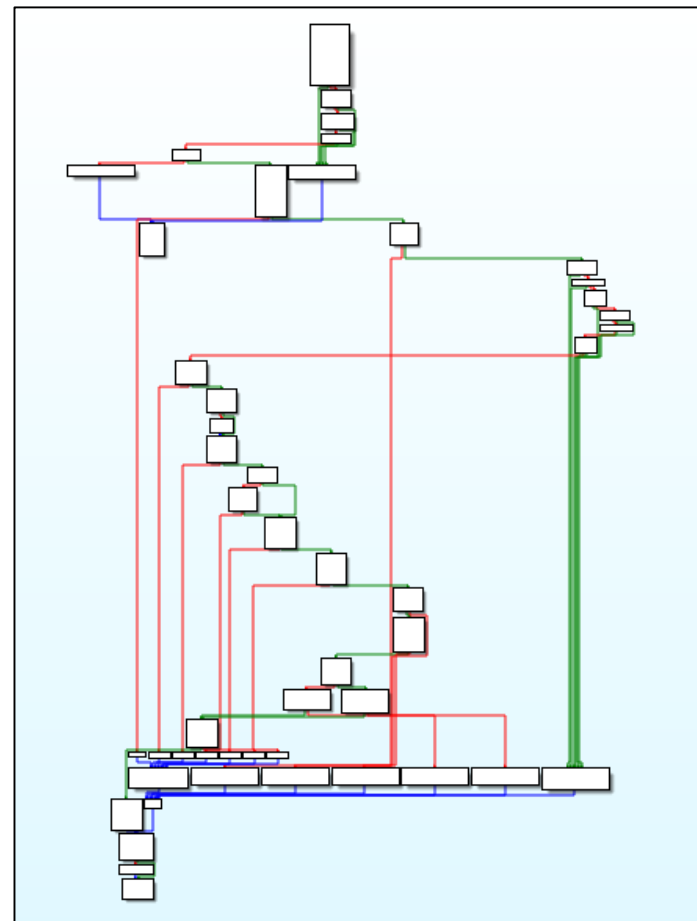
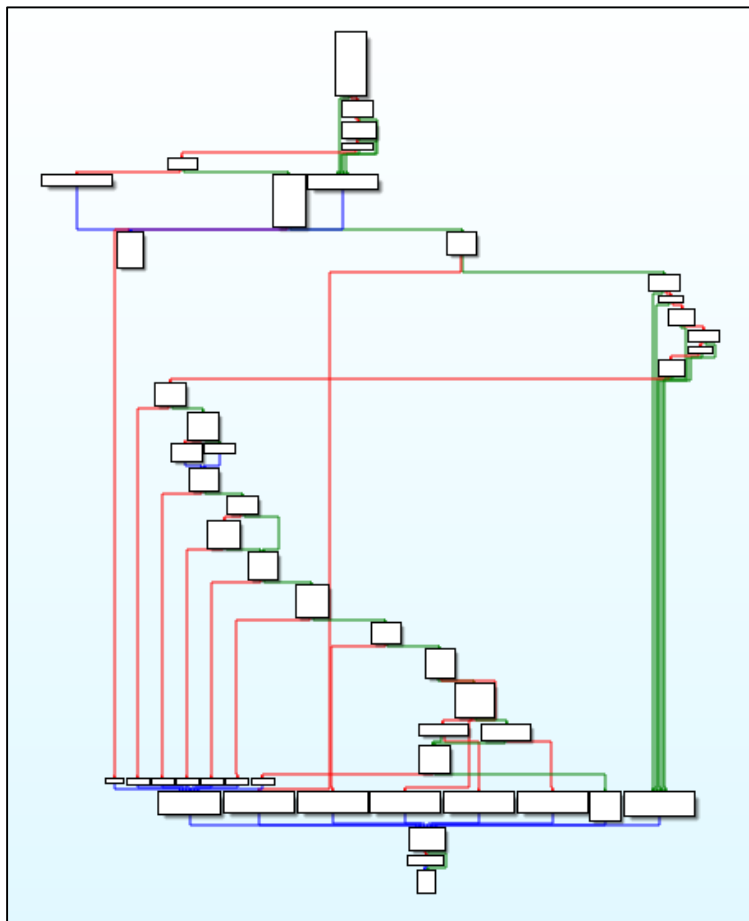


함수 유사도 측정

김영철

2016. 7. 20.

함수 비교



유사도 측정 방법

- IDA pro를 이용하여 함수 정보 추출
 - basic block의 수
 - edge의 수
 - call 명령어의 수
 - cmp 명령어의 수

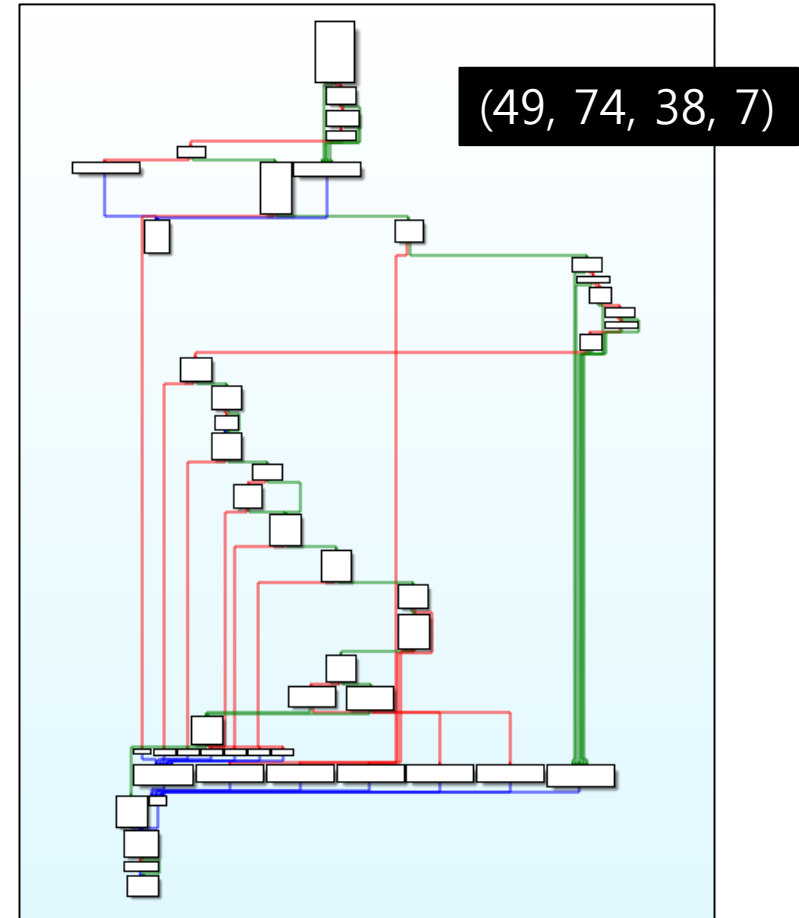
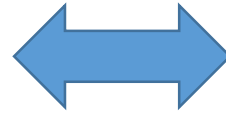
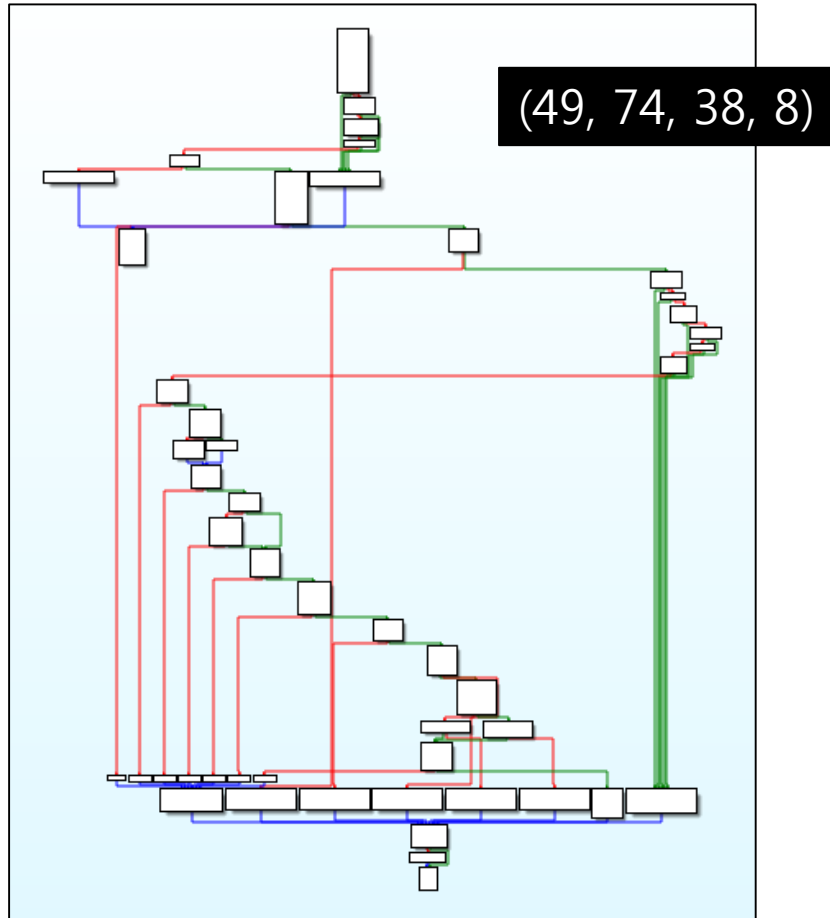
함수 수준 특징정보 기반의 오픈소스 소프트웨어 모듈 탐지, 김동진
프로그램 구조와 상수 값을 이용하는 바이너리 실행 파일의 차이점 분석, 박희완

유사도 측정 방법

- IDA pro를 이용하여 함수 정보 추출
 - 수집 제외 함수
 - 명령어의 수가 2보다 작은 경우
 - 블록의 수가 2보다 작으면서 edge의 수가 2보다 작은 경우
➔ 블록 하나에 명령어가 많은 경우도 제외됨

함수 수준 특징정보 기반의 오픈소스 소프트웨어 모듈 탐지, 김동진
프로그램 구조와 상수 값을 이용하는 바이너리 실행 파일의 차이점 분석, 박희완

유사도 측정 방법



유사도 측정 방법

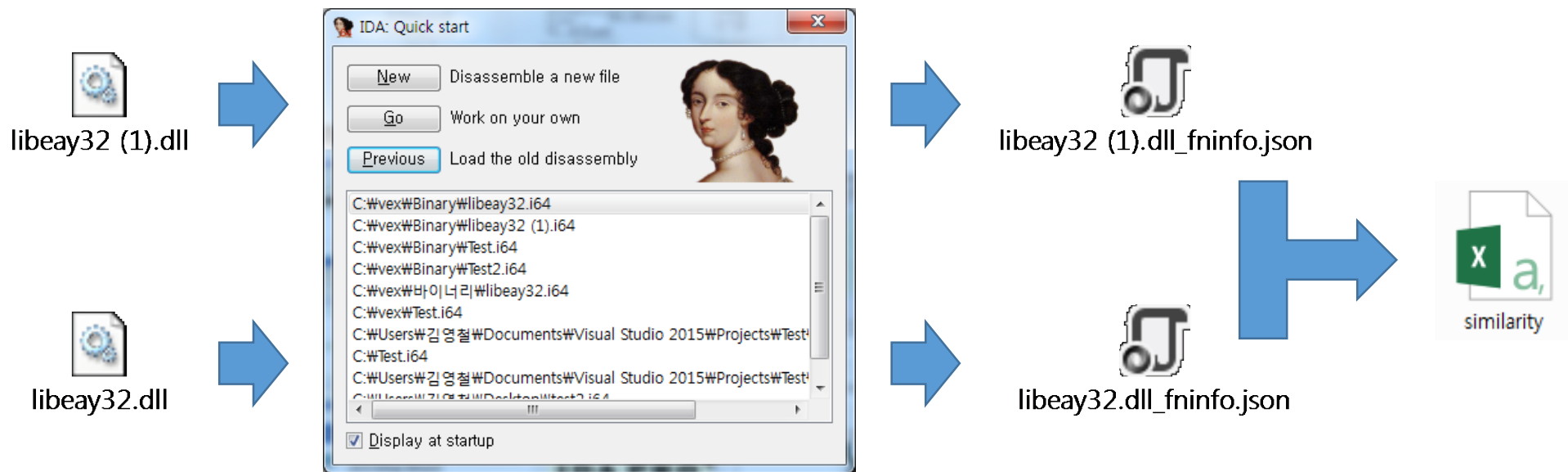
- 추출한 정보들을 코사인 유사도 공식에 대입하여 계산
 - (3, 1, 0, 0), (6, 2, 0, 0) 이러한 경우에도 코사인 유사도 결과는 1
 - 이러한 경우를 제외하기 위해서 벡터의 크기 비교
 - 유사도가 0.975 이상인 함수 요약 정보만 저장

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

<코사인 유사도 공식>

실험

- 다른 버전의 crypto++ 동적 라이브러리를 가지고 실험

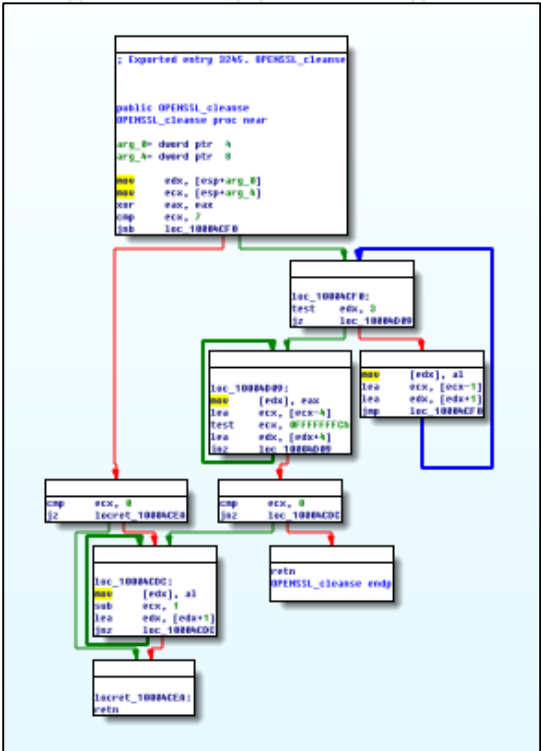
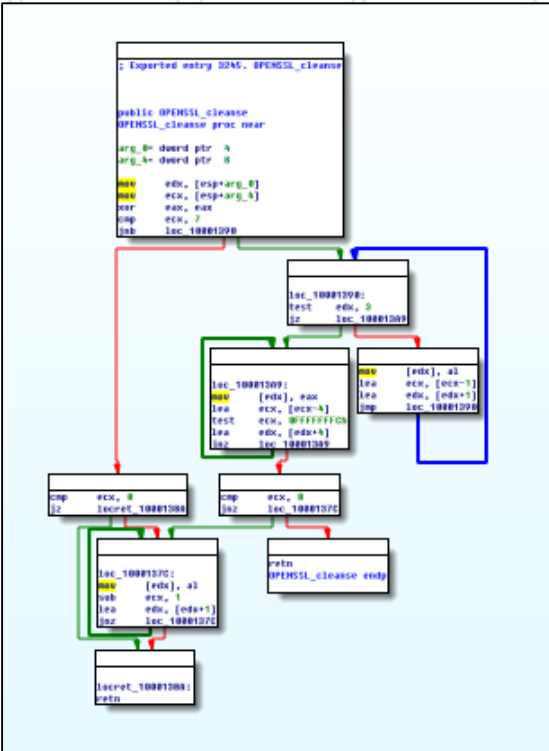
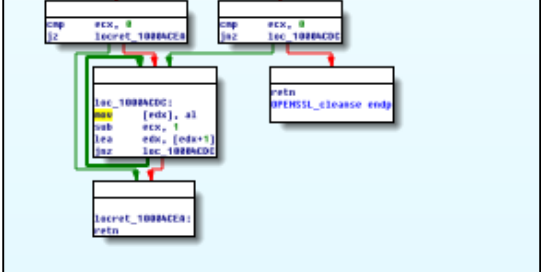
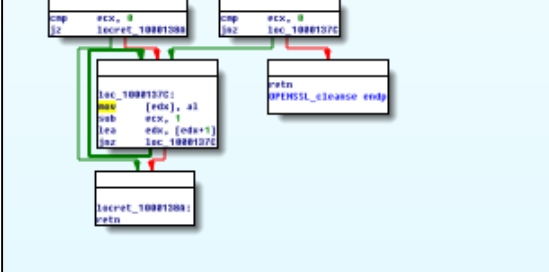


실험 결과

target	blocks1	edges1	calls1	cmps1	source	blocks2	edges2	calls2	cmps2	sim
0x1000L	26	42	0	17	0xbdfe0L	26	44	2	9	0.980945
0x1360L	9	13	0	3	0x4cc0L	9	13	0	3	1
0x1360L	9	13	0	3	0x49fd0L	9	13	0	2	0.988417
0x1360L	9	13	0	3	0x59930L	9	13	1	4	0.981273
0x13d0L	4	5	0	1	0x4d30L	4	5	0	1	1
0x13d0L	4	5	0	1	0x26980L	4	5	0	0	0.97619
0x13d0L	4	5	0	1	0x2d310L	5	4	0	1	0.97619
0x13d0L	4	5	0	1	0x5ad90L	4	5	1	0	0.97619
0x13d0L	4	5	0	1	0x5ae50L	4	5	1	1	0.976744
0x13d0L	4	5	0	1	0x73f40L	4	5	1	1	0.976744
0x13d0L	4	5	0	1	0x73f70L	4	5	1	0	0.97619
0x13d0L	4	5	0	1	0xb67b0L	4	5	0	0	0.97619
0x1440L	3	3	0	1	0x4950L	3	3	0	1	1
0x1440L	3	3	0	1	0x59c0L	3	3	0	1	1
0x1440L	3	3	0	1	0x1b480L	3	3	0	1	1

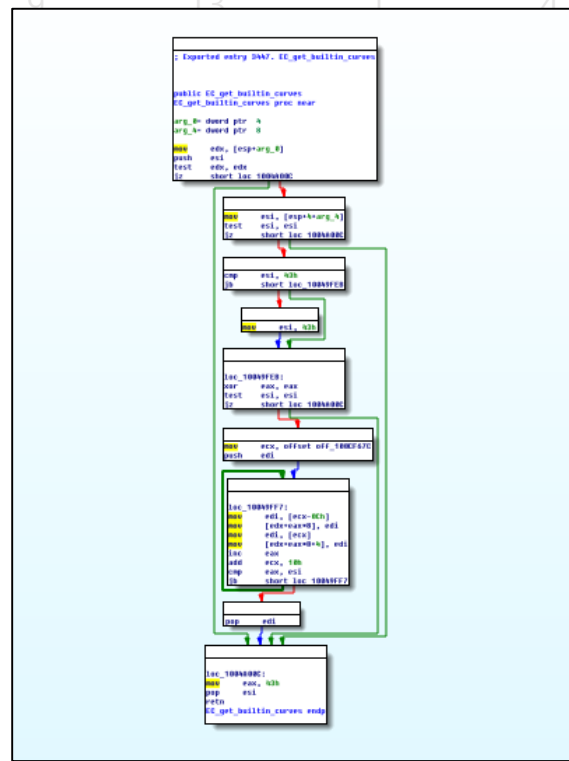
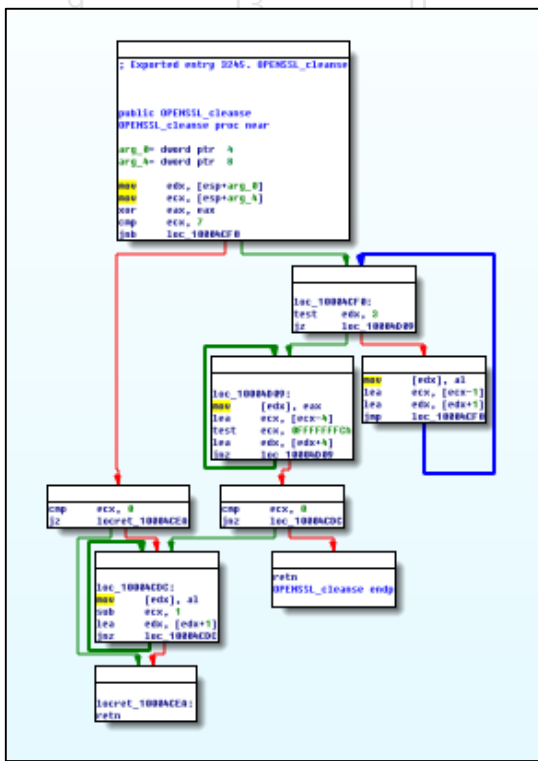
실험 결과

target	blocks1	edges1	calls1	cmps1	source	blocks2	edges2	calls2	cmps2	sim
0x1000L	26	42	0	17	0xbdfe0L	26	44	2	9	0.980945
0x1360L	9	13	0	3	0x4cc0L	9	13	0	3	1

0x1360L		3	0x49fd0L		0.988417
0x1360L		3	0x59930L		0.981273
0x13d0L		1	0x4d30L		1
0x13d0L		1	0x26980L		0.97619
0x13d0L		1	0x2d310L		0.97619
0x13d0L		1	0x5ad90L		0.97619
0x13d0L		1	0x5ae50L		0.976744
0x13d0L		1	0x73f40L		0.976744
0x13d0L		1	0x73f70L		0.97619
0x13d0L		1	0xb67b0L		0.97619
0x1440L		1	0xb67b0L		1
0x1440L		1	0x4950L		1
0x1440L		1	0x59c0L		1
0x1440L		1	0x1b480L		1

실험 결과

target	blocks1	edges1	calls1	cmps1	source	blocks2	edges2	calls2	cmps2	sim
0x1000L	26	42	0	17	0xbdfe0L	26	44	2	9	0.980945
0x1360L	9	13	0	3	0x4cc0L	9	13	0	3	1
0x1360L	9	13	0	3	0x49fd0L	9	13	0	2	0.988417
0x1360L	9	13	0	3	0x59930L	9	13	1	4	0.981273
0x13d0L					0x4d30L					1
0x13d0L					0x26980L					0.97619
0x13d0L					0x2d310L					0.97619
0x13d0L					0x5ad90L					0.97619
0x13d0L					0x5ae50L					0.976744
0x13d0L					0x73f40L					0.976744
0x13d0L					0x73f70L					0.97619
0x13d0L					0xb67b0L					0.97619
0x1440L					0x4950L					1
0x1440L					0x59c0L					1
0x1440L					0x1b480L					1



실험 결과

- 요약 정보의 수치가 작은 경우, 같다고 측정되는 함수가 너무 많이 추출됨.
- 버전이 다른 경우, 같은 이름의 함수이지만 요약 정보의 차이가 많이 나는 경우 발생.

결론

- 함수 요약 정보만으로는 부족
- 그래프 비교 필요
- 비교 대상을 줄일 수 있는 방법 필요