Сравнение потоков исполнения программ на основе списка системных вызовов

Магистерская диссертация

Автор: Ханов А.Р. awengar@gmail.com

Руководитель: Баклановский М.В.

Рецензент: Тимофеев А.В.

Работа в рамках проекта CODA

Санкт-Петербург, 2012

Проблемная область

Исследование программ

Динамическое исследование программ

Поведенческий анализ

Сложности:

- Невозможность проверки эквивалентности исходных кодов
- Методы противодействия анализу
  - Обфускация
  - Упаковка
  - Модификация кода

Постановка задачи

Основная задача:
Идентификация процессов по их поведению в операционной системе

Выделение особенностей взаимодействия процесса и операционной системы с целью дальнейшего обнаружения присутствия этого процесса в системе.

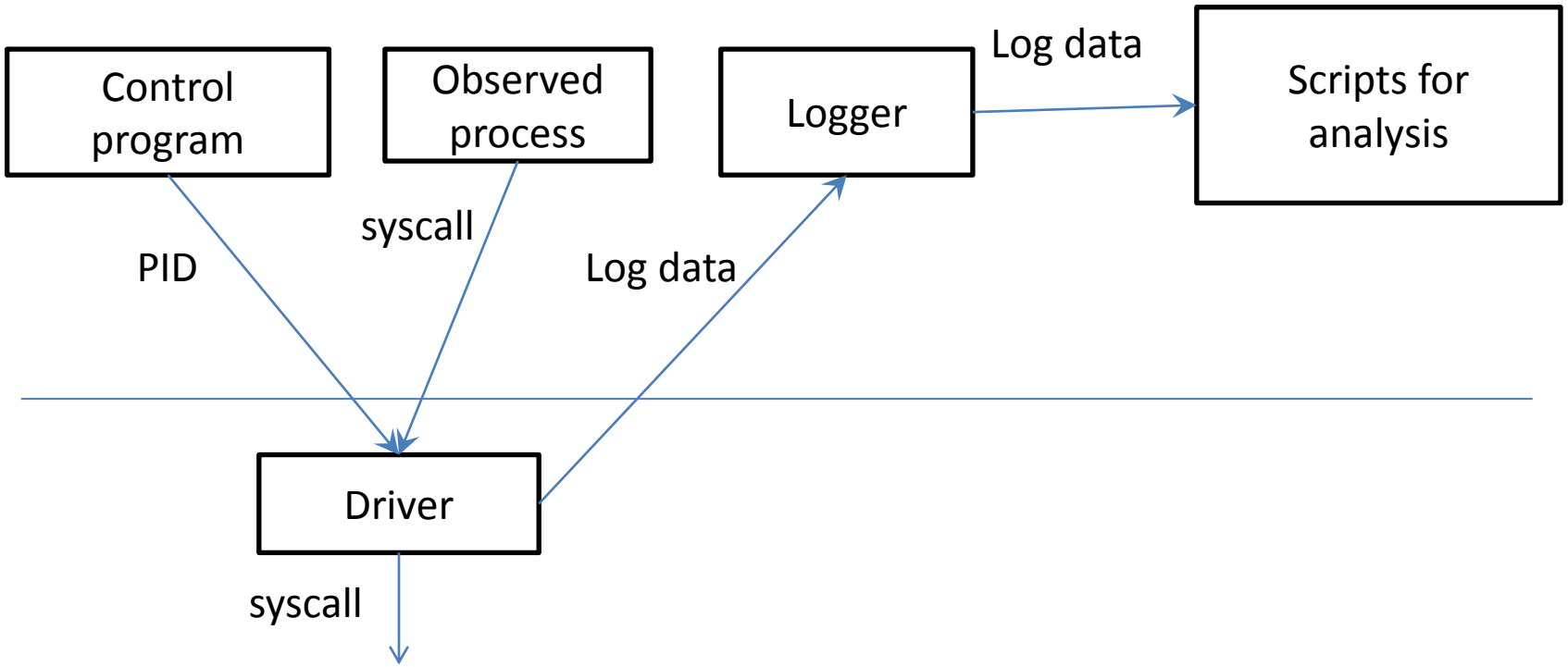
Обзор существующих методик

- Статический анализ
  - Сигнатурные методы
  - Темпоральные модели
  - Символьное исполнение
- Динамический анализ
  - Зависимость API-вызовов по данным
  - Динамические последовательности инструкций
  - Итеративные шаблоны событий

Получение информации о поведении процесса

Информация о системных вызовах

- Быстро извлекается
- Нет вмешательства в исполняемый процесс
- Трудно интерпретировать



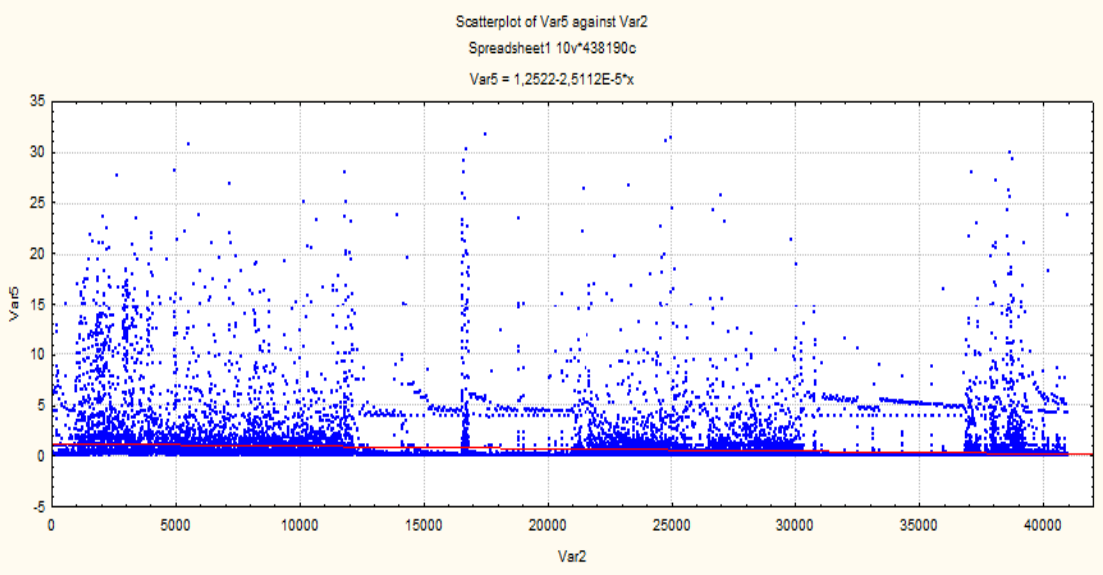
Вычисление энтропии

1) PPM – алгоритм оценки вероятности появления символа в контексте префикса.
2) PPM – работает в потоке, адаптивен.
3) Символ в новом контексте несет больше информации.

$$H(x) = -\sum_{i=1}^{n} p(i) \log_2 p(i).$$ - энтропия события x с n состояниями

$$\log_2 \frac{1}{p(i)}$$ - частная энтропия – степень неопределенности, устраняемой i-м символом

Частная энтропия вызовов



Internet Explorer контекст 9 символов

Процесс	Среднеквадратическое отклонение энтропии
Explorer	2,0018
IE	1,9066
Devenv	1,7847
Notepad	1,766
Opera	1,6362
MF	1,4614
Calc	1,2754
Mines	1,2259
WMPlayer	0,9948
VLC	0,8618

Выводы:

1) Зашумленность вызовов (оценка – среднеквадратическое отклонение)
2) Интерпретация: активность и пассивность в поведении

Корреляция n-грамм

1) Построим гистограммы n-грамм для каждой последовательности вызовов
2) Посчитаем линейный коэффициент корреляции между ними
3) Корреляция близка к 1 – потоки схожи, близка к 0 или меньше 0 – потоки различаются.

$$\mathbf{r}_{XY} = \frac{\mathbf{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}.$$

При изучении более мелких участков можно находить схожие участки потока вызовов

Корреляция n-грамм

Корреляция между различными участками по 50000 вызовов процесса Internet Explorer (серый - более 0,5)

Таблица »	0-50000	50000-100	100000-15	150000-20	200000-25	250000-30	300000-35	350000-40	400000-45	450000-50	500000-55	550000-60	600000-65	650000-70	700000-75	750000-80	800000-85	850000-90
0-50000	1	0,7748	0,8777	0,8979	0,4846	0,8908	0,9275	0,8949	0,3489	0,8524	0,9419	0,8356	0,3745	0,3852	0,3386	0,3995	0,8131	0,847
25000-750	0,7748	1	0,8197	0,3202	0,0595	0,2411	0,4385	0,4748	0,6384	0,244	0,6861	0,7628	0,6016	0,5118	0,6714	0,7446	0,7712	0,801
50000-100	0,8777	0,8197	1	0,7689	0,4547	0,7199	0,718	0,547	0,4188	0,6782	0,7351	0,8043	0,5513	0,5309	0,5832	0,6268	0,7851	0,8188
75000-125	0,8979	0,3202	0,7689	1	0,8166	0,8353	0,9081	0,8074	0,6779	0,7776	0,8916	0,5354	0,2155	0,2555	0,0865	0,1889	0,8567	0,8189
100000-15	0,4846	0,0595	0,4547	0,8166	1	0,8003	0,7881	0,9449	0,7299	0,639	0,4681	0,0059	0,0529	0,0884	-0,0214	0,0985	0,0392	0,0305
125000-17	0,8908	0,2411	0,7199	0,8353	0,8003	1	0,9692	0,9757	0,9524	0,3824	0,9367	0,5465	0,2717	0,313	0,1193	0,2126	0,688	0,6187
150000-20	0,9275	0,4385	0,718	0,9081	0,7881	0,9692	1	0,9521	0,8824	0,5404	0,9573	0,2259	0,1855	0,2344	0,0615	0,1563	0,7112	0,6797
175000-22	0,8949	0,4748	0,547	0,8074	0,9449	0,9757	0,9521	1	0,9597	0,3785	0,7692	0,1878	0,2103	0,2858	0,174	0,2386	0,424	0,4298
200000-25	0,3489	0,6384	0,4188	0,6779	0,7299	0,9524	0,8824	0,9597	1	0,1692	0,1719	0,324	0,3627	0,4283	0,4588	0,4675	0,5146	0,5049
225000-27	0,8524	0,244	0,6782	0,7776	0,639	0,3824	0,5404	0,3785	0,1692	1	0,9086	0,1873	0,2342	0,2909	0,0683	0,2334	0,4517	0,4543
250000-30	0,9419	0,6861	0,7351	0,8916	0,4681	0,9367	0,9573	0,7692	0,1719	0,9086	1	0,6744	0,2781	0,3159	0,1627	0,2517	0,7897	0,8126
275000-32	0,8356	0,7628	0,8043	0,5354	0,0059	0,5465	0,2259	0,1878	0,324	0,1873	0,6744	1	0,9368	0,885	0,9708	0,967	0,9638	0,9689
300000-35	0,3745	0,6016	0,5513	0,2155	0,0529	0,2717	0,1855	0,2103	0,3627	0,2342	0,2781	0,9368	1	0,9662	0,9503	0,9566	0,9661	0,9464
325000-37	0,3852	0,5118	0,5309	0,2555	0,0884	0,313	0,2344	0,2858	0,4283	0,2909	0,3159	0,885	0,9662	1	0,9217	0,9263	0,9431	0,9135
350000-40	0,3386	0,6714	0,5832	0,0865	-0,0214	0,1193	0,0615	0,174	0,4588	0,0683	0,1627	0,9708	0,9503	0,9217	1	0,9552	0,9953	0,995
375000-42	0,3995	0,7446	0,6268	0,1889	0,0985	0,2126	0,1563	0,2386	0,4675	0,2334	0,2517	0,967	0,9566	0,9263	0,9952	1	0,9976	0,9965
400000-45	0,8131	0,7712	0,7851	0,8567	0,0392	0,688	0,7112	0,424	0,5146	0,4517	0,7897	0,9638	0,9661	0,9431	0,9953	0,9976	1	0,995
425000-47	0,847	0,801	0,8188	0,8189	0,0305	0,6187	0,6797	0,4298	0,5049	0,4543	0,8126	0,9689	0,9464	0,9135	0,995	0,9965	0,995	1

Корреляция между различными участками по 100000 процессов Internet Explorer и Mozilla Firefox

Numbers	0-100000	100000-200000	200000-300000	300000-400000	400000-500000	500000-600000	600000-700000	700000-800000	800000-900000	900000-1000000	1000000-1100000	1100000-1200000	1200000-1300000	1300000-1400000	1400000-1500000	1500000-1600000
0-100000	0,3645	0,0564	0,2261	0,4922	0,6549	0,617	0,2488	0	-0,0295	-0,0784	0	0,3174	0,5247	-0,0008	0	0,0769
100000-200000	0,989	0,9967	0,8524	0,8433	0,1454	0,2683	0,1901	0	-0,4797	0,9887	0,261	0,8912	0,6468	0,7258	0	0,9576
200000-300000	0,8475	0,8832	0,9857	0,8848	0,4089	0,3467	0,1225	0	-0,3984	0,9418	0,6888	0,9734	0,8941	0,9407	0	0,954
300000-400000	0,94	0,9377	0,9707	0,9831	0,4316	0,4563	0,127	0	-0,9045	0,9999	0,4992	0,987	0,8484	0,8848	0	0,9998
400000-500000	0,265	0,3658	0,7495	0,4828	0,3041	0,1341	0,0879	0	-0,3727	0,795	0,9952	0,5818	0,7851	0,9323	0	0,7197
500000-600000	0,0221	0,1998	0,2156	0,2461	0,1922	0,0886	0,0791	0	-0,553	-0,4056	1	0,1502	0,1174	0,4822	0	-0,219
600000-700000	0,2657	0,3953	0	0,2919	0,694	0,5793	0,133	0	0,8238	-0,0913	0	0,3156	0,7044	-0,25	0	0,4397
700000-800000	0	0	0	0	0	0	0,8693	0	0	0	0	0	0	0,8998	0	0
800000-900000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Выводы:

- Можно находить общие участки потоков исполнения
- Не подходит для установления соответствия потоков

Термы

Терм – последовательность вызовов, которая повторяется в последовательности более одного раза.

Свойства:

1) Число появлений в последовательности
2) Максимальный и средний периоды
3) Длина терма как последовательности
4) Пересечение, вхождение, содержание других термов, самопересечение

Нахождение всех термов = нахождение всех дубликатов в строке

Термы

Обоснование:

1) Повторение последовательностей - связь вызовов
2) Термы – участки с низкой энтропией (борьба с шумом)

Правила фильтрации:

1) Чем длиннее, тем больше в нем информации
2) Чем больше раз встречается, тем вероятнее его встретить

Алгоритм :

1) Выделение всех дубликатов: построение суффиксного массива и LCP. Множество LCP – дубликаты максимальной вправо длины.

2) Фильтраця термов:
    1) Перебираем термы от самых длинных до самых коротких
    2) Если терм встретился не менее K раз, не пересекаясь с собой и другими термами, то помечаем его вхождения
    3) Продолжаем пока не получим нужное число термов

Идентификация с помощью термов

Обучение:

1) Список вызовов потока программы за большой период работы
2) Выделение термов
3) Фильтрация термов

Идентификация:

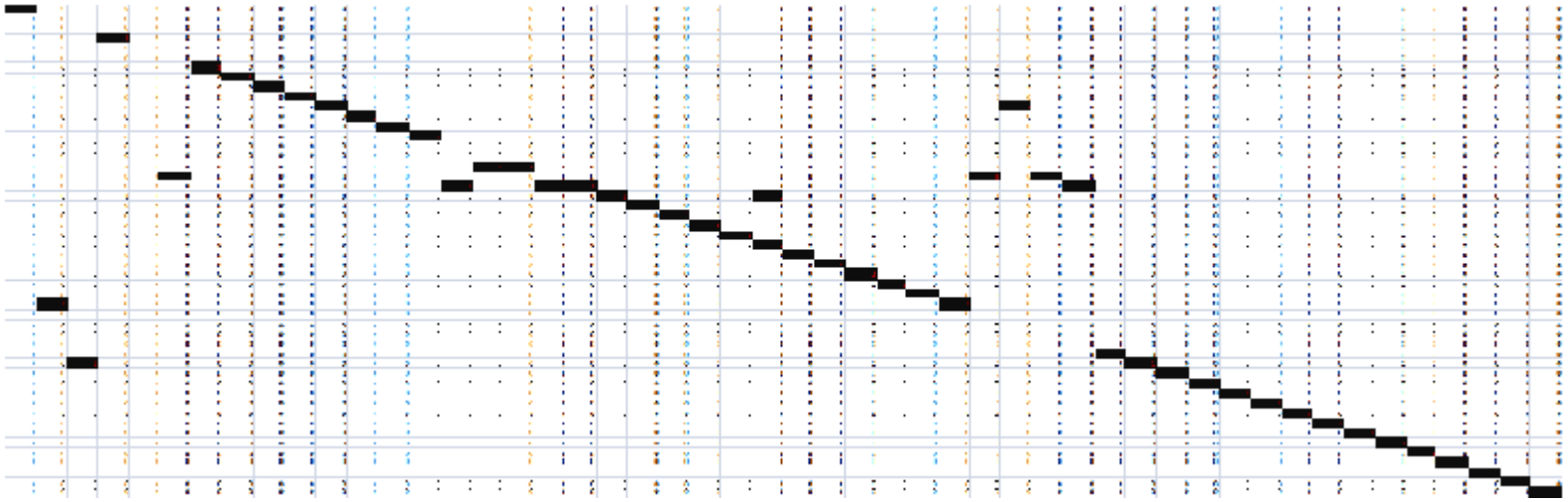
1) Список вызовов потока программы за короткий период
2) Поиск термов в последовательности, вычисление количества уникальных термов и максимальной длины терма.
3) Оценка соответствия потока ранее изученному

Тест

5 потоков из 10 процессов

По вертикали – потоки, из которых извлекались термы

По горизонтали – потоки, в которых искались термы (тестовая выборка)



Тестирование

Проведено 10 тестов: идентификация главных потоков программ. Практически все успешны.

Тест

- Не более 2000 термов, встретившихся не менее 3 раз
- Извлекаем термы из списков вызовов порядка 1000000
- тестируем первые 30000 нового экземпляра процесса

Количество термов в потоке[максимальная длина терма]

Процесс, Термы	Calc	Devenv	Explorer	IE	Mines	Mozilla	Notepad
CALC	964 [360]	103 [93]	37 [11]	77 [32]	61 [134]	32 [11]	86 [103]
DEVENV	3 [62]	145 [585]	2 [45]	3 [47]	6 [71]	0 [0]	23 [98]
EXPLORER	105 [55]	192 [49]	85 [28]	176 [45]	62 [23]	42 [21]	87 [41]
IE	3 [29]	8 [34]	1 [27]	322 [110]	1 [23]	0 [0]	0 [0]
MINES	1 [71]	0 [0]	0 [0]	0 [0]	400 [722]	0 [0]	0 [0]
Mozilla	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	174 [1426]	0 [0]
NOTEPAD	105 [80]	148 [65]	14 [47]	50 [39]	45 [60]	13 [18]	224 [165]

Выделение термов

Достоинства:

- Низкая вычислительная сложность
- Высокое качество распознавания
- Основан на информации о системных вызовов, поведенческий анализ.

Недостатки:

- Неравномерность встречаемости термов
- Применены эвристики, необходимо более глубокое изучение проблемы.

Результаты

1) Разработан алгоритм идентификации процессов
2) Проведено тестирование алгоритма
3) метод оценки активности потоков
4) алгоритм поиска схожих участков в потоке исполнения
5) Опубликовано на конференции СПИСОК-2012

Благодарности

Группа “Профайлер ядра” (Дудин В., Одеров Р., Тенсин Е.)