ИСП РАН

Автоматическая генерация сигнатур сетевых протоколов

Дурнов Алексей Николаевич
Научный руководитель: к. ф.-м. н. Гетьман А. И.

18 апреля 2024 г.

МФТИ, кафедра системного программирования, ИСП РАН

Введение

- Методы классификации сетевого трафика:
  1. основанные на идентификации по номеру порта сервера
  2. основанные на DPI подходе
     - сигнатурный
  3. основанные на статистических характеристиках
- Свойства сигнатур для сетевых протоколов и приложений:
  1. короткие общие подстроки
  2. несколько потоков с разным набором подстрок
  3. необходимость частого обновления

Цель

Целью данной работы является разработка и реализация метода автоматической генерации сигнатур полезной нагрузки сетевого трафика для классификации этого трафика в соответствии с использующимся протоколом.

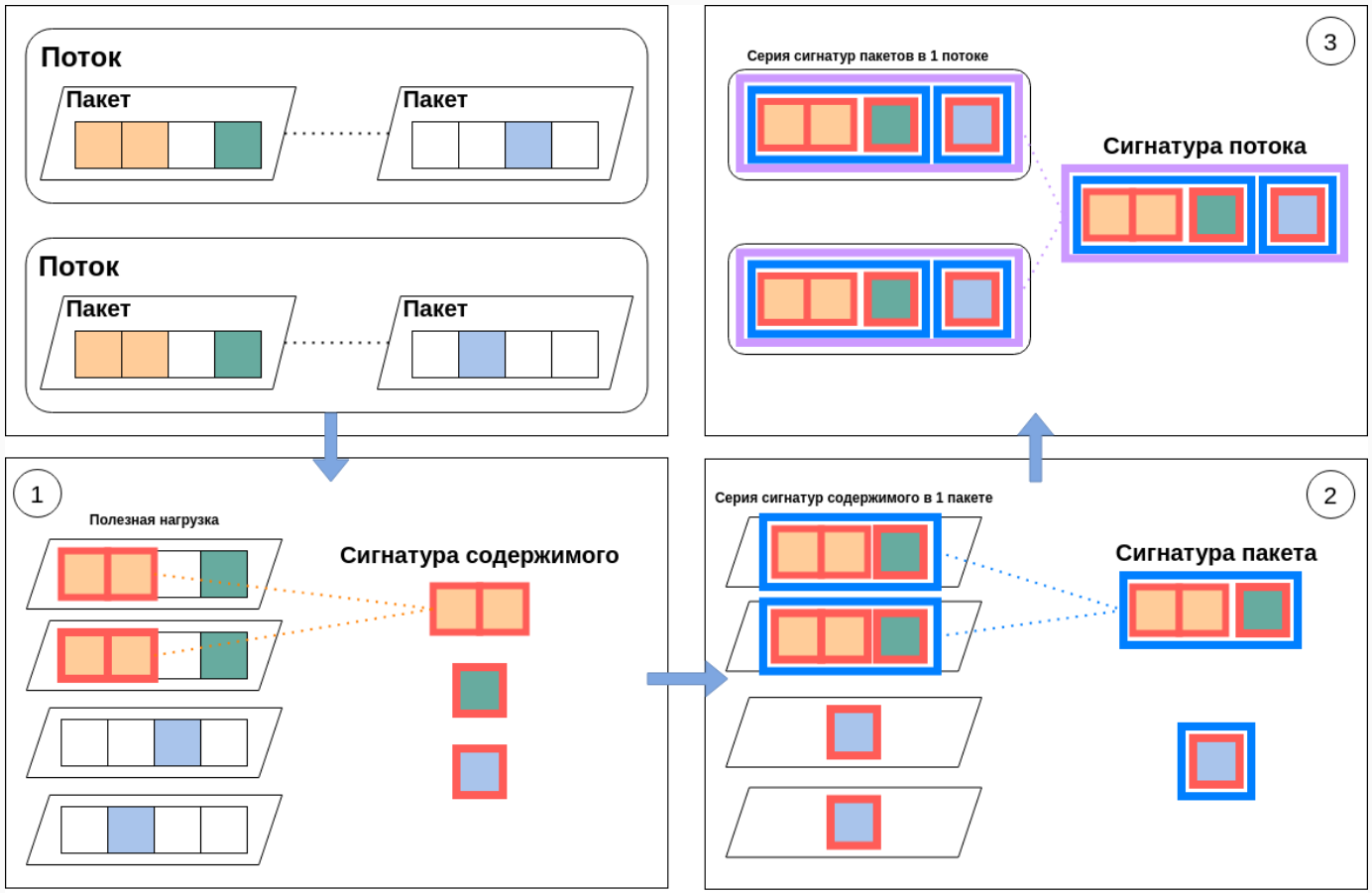
Задачи

- Провести исследование литературы по соответствующей теме.
- Собрать набор сетевых трасс для последующего тестирования.
- Выбрать формат сигнатуры сетевых протоколов.
- Разработать алгоритм генерации сигнатур.
  1. Выбрать методы для автоматической генерации сигнатур.
  2. Рассмотреть ограничения данных методов.
- Разработать классификатор сетевого трафика для проверки сгенерированных сигнатур.
  1. Выбрать метрики, по котором можно оценить качество классификации сигнатур.
  2. Реализовать сопоставление сигнатур.
- Встроить генератор сигнатур и классификатор как модули в систему анализа сетевого трафика, разрабатываемую в ИСП РАН.

Формат сигнатур

- Сигнатура - набор последовательностей подстрок (байт).
- Полезная нагрузка соответствует сигнатуре, если она содержит какую-то последовательность подстрок из этого набора.
- Иерархия сигнатур:
  1. сигнатура содержимого
  2. сигнатура пакета
  3. сигнатура потока

Формат сигнатур: иерархия



Методы автоматической генерации сигнатур

- LASER (Application Signature ExtRaction)
  - LCS (Long Common Subsequence)
  - сигнатура пакета
  + наиболее подходящий для модификаций
  — неустойчив к шумам
  — не утилизирует все потоки
- AutoSig
  - сигнатура потока
  + набор последовательностей подстрок
  + учитывает частотное распределение подстрок
  — избыточное дерево подстрок
  — долгая сходимость алгоритма
- SigBox
  - сигнатура потока
  + поддерживает иерархию
  — долгая сходимость алгоритма

Схема генерации сигнатуры

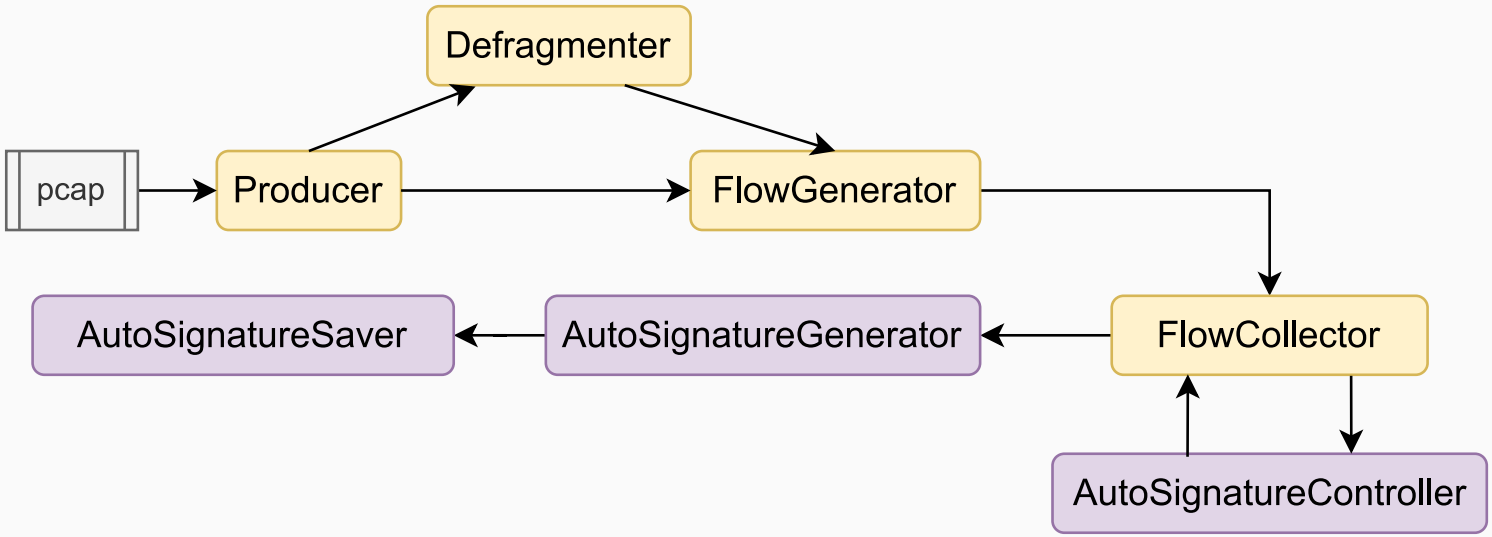
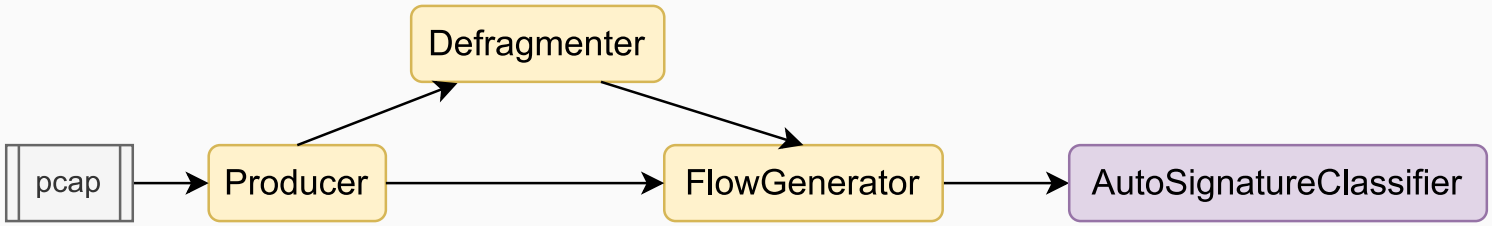


Схема классификации трафика



Данные для генерации сигнатур

Протокол	Размер, МБ	Количество пакетов	Количество потоков	Avg bytes/pkt	Avg pkts/stream	Количество потоков: $\geq$ 5 pkts
bittorrent	272,4	240159	708	1134	339	214
dns	130,7	1283082	204150	102	6	11027
ftp	0,86	16959	735	51	23	725
http	1811,5	799062	13710	2268	58	1500
imap	22,0	27702	66	793	419	65
pop	0,06	919	59	64	16	40
smtp	13,5	59121	1120	229	53	799

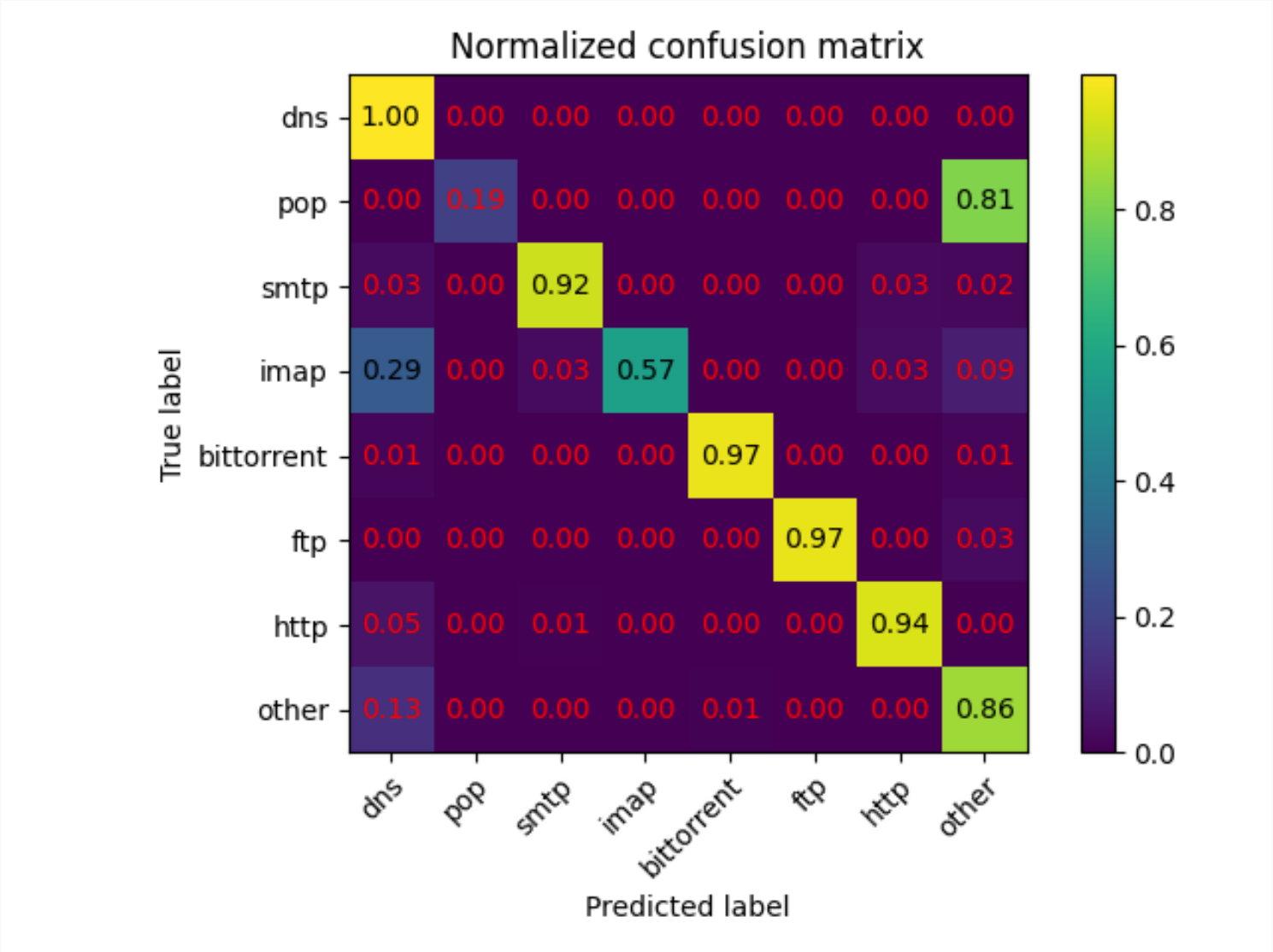
Данные для классификации

Протокол	Размер, МБ	Количество пакетов	Количество потоков	Avg bytes/pkt	Avg pkts/stream
bittorrent	1,26	9409	876	133	11
dns	53,3	664809	27800	80	24
ftp	0,19	4000	294	48	14
http	1494	406030	4607	3681	88
imap	3,38	10587	143	320	74
other	1119	1235122	18846	906	66
pop	0,02	344	21	58	16
smtp	5,8	34428	1018	170	34

Генерация сигнатур: алгоритм LASER

Протокол	Количество сигнатур	Среднее количество потоков на 1 сигнатуру
bittorent	51	4,1
dns	682	16,1
ftp	9	80,5
http	145	10,3
imap	10	6,5
pop	2	20,0
smtp	167	4,8

Результаты работы алгоритма LASER



Результаты работы алгоритма LASER

protocol	precision	recall	f1-score	support
dns	0.93	1.00	0.97	27801
pop	0.80	0.19	0.31	21
smtp	0.94	0.92	0.93	1076
imap	1.00	0.57	0.72	175
bittorrent	0.92	0.97	0.95	886
ftp	0.99	0.97	0.98	295
http	0.99	0.94	0.96	4888
other	0.99	0.86	0.92	12094
accuracy			0.95	47236
macro avg	0.95	0.80	0.84	47236
weighted avg	0.95	0.95	0.95	47236

Постобработка: удаление дубликатов и простых надмножеств

Протокол	Количество сигнатур		Избыточность	
	Было	Стало	Было	Стало
bittorrent	51	7	0.53	0.53
dns	682	64	0.99	0.99
ftp	9	8	0.98	0.98
http	145	47	1.00	1.00
imap	10	4	1.00	0.79
pop	2	1	1.00	0.00
smtp	167	26	0.99	0.99

- Избыточность - отношение числа потоков, идентифицированных двумя и более сигнатурами к общему числу потоков, идентифицированных набором сигнатур.
- Результаты классификации никак не поменялись, при этом сильно упало количество сигнатур.
- Слегка уменьшилась избыточность.

Схема генерации сигнатуры со сборкой TCP-сессий

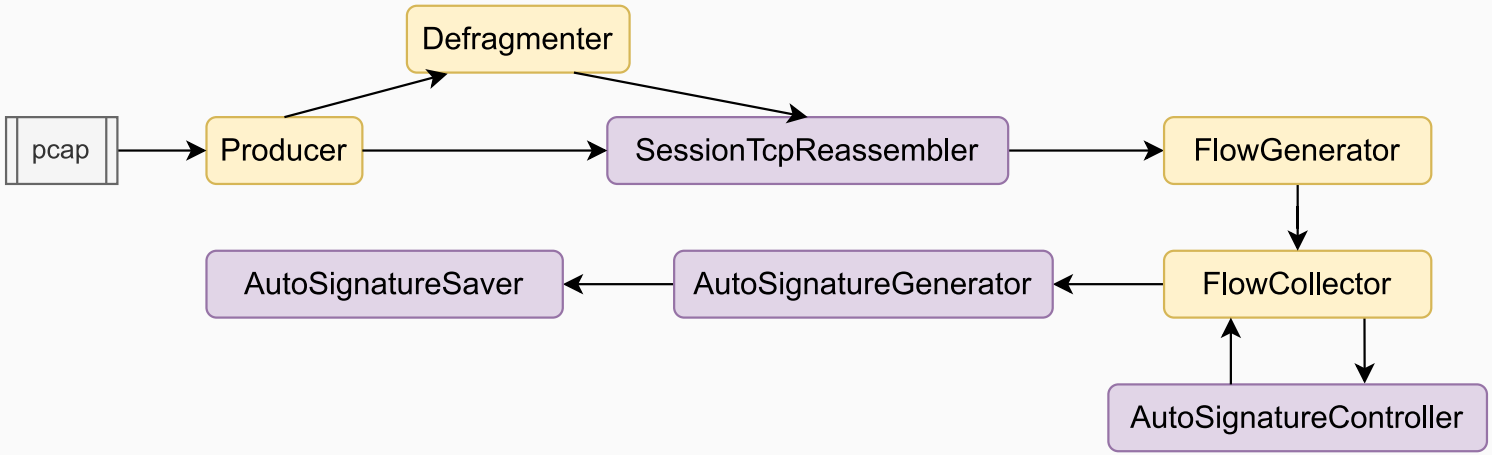
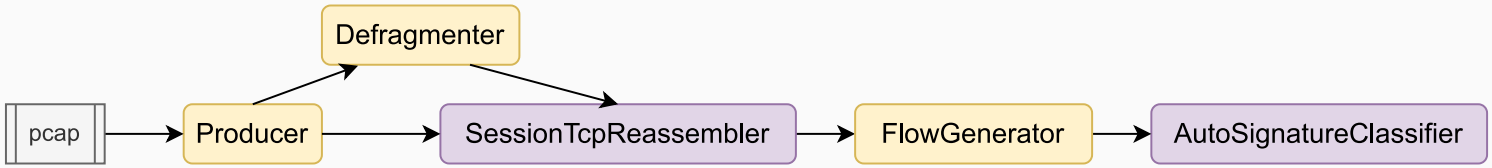
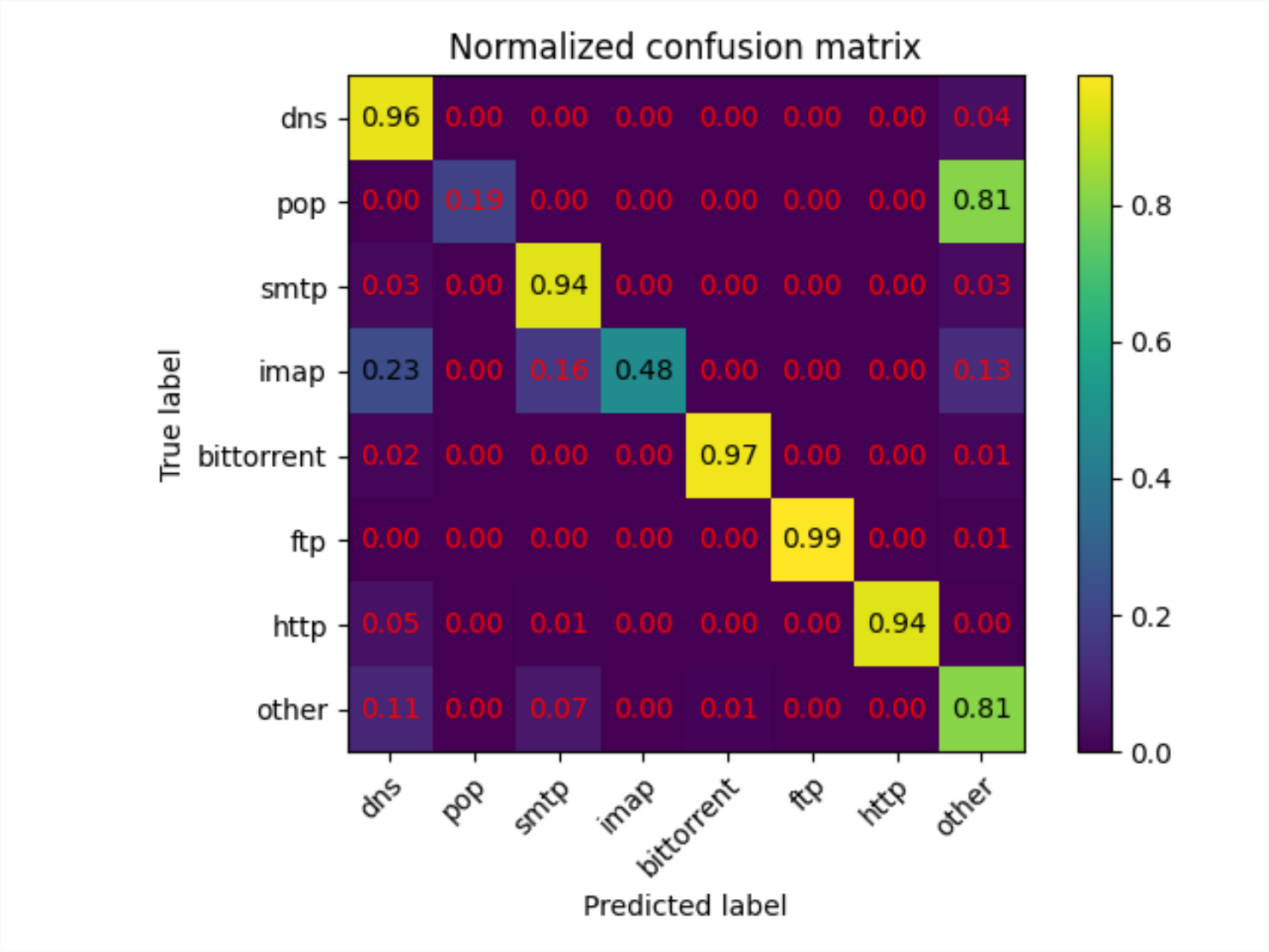


Схема классификации трафика со сборкой TCP-сессий



Результаты работы алгоритма LASER: со сборкой последовательных пакетов TCP-сессии



Результаты работы алгоритма LASER: со сборкой последовательных пакетов TCP-сессии

protocol	precision	recall	f1-score	support
dns	0.92	0.96	0.94	10565
pop	0.67	0.19	0.30	21
smtp	0.78	0.94	0.85	1464
imap	0.91	0.48	0.62	189
bittorrent	0.95	0.97	0.96	853
ftp	0.99	0.99	0.99	288
http	1.00	0.94	0.97	4890
other	0.88	0.81	0.85	4804
accuracy			0.92	23074
macro avg	0.89	0.79	0.81	23074
weighted avg	0.92	0.92	0.92	23074

Результаты

- При частичной сборке TCP-сессии с использованием LASER, полученные результаты получились немного хуже.
- При полной сборке TCP-сессии с использованием LCS, результаты сильно зависят от параметров алгоритма и постобработки. Параметры необходимо подбирать для каждого протокола вручную.
- Могут генерироваться неспецифичные сигнатуры.

Заключение

- Проведено исследование литературы по соответствующей теме.
- Собран набор сетевых трасс для генерации и классификации.
- Выбран формат сигнатуры сетевых протоколов.
- Рассмотрены ограничения выбранных методов и реализован один из них.
- Разработан классификатор сетевого трафика для проверки сгенерированных сигнатур.
- Рассмотрено влияние сборки TCP-сессии и постобработки на результат классификации
- Встроены генератор сигнатур и классификатор как модули в систему анализа высокоскоростного сетевого трафика, разрабатываемую в ИСП РАН.

Будущие исследования

- Реализация и сравнение других методов генерации сигнатур: AutoSig и SigBox.
- Реализация модуля уточнения положения сигнатуры и его влияние на точность.
- Рассмотрение возможности применения машинного обучения (метода SVM) для выбора набора сигнатур при постобработке результатов.

Спасибо за внимание