

# Автоматическая генерация сигнатур сетевых протоколов

Дурнов Алексей Николаевич

Московский физико-технический институт  
Физтех-школа радиотехники и компьютерных технологий  
Кафедра системного программирования

Научный руководитель к.ф.-м.н. Гетьман А. И.

Москва, 2024 г.

## Методы классификации сетевого трафика:

- основанные на идентификации по номеру порта сервера
- основанные на DPI подходе
  - **сигнатурный**
- основанные на статистических характеристиках

## Сигнатура

Это набор байтов (подстроки), обеспечивающий идентификацию. Пример сгенерированной сигнатуры для HTTP:

- {"GET /", " HTTP/1.", "\r\nHost: "}

## Свойства сигнатур для сетевых протоколов

- короткие общие подстроки
- несколько потоков с разным набором подстрок (пример, FTP)
- необходимость частого обновления

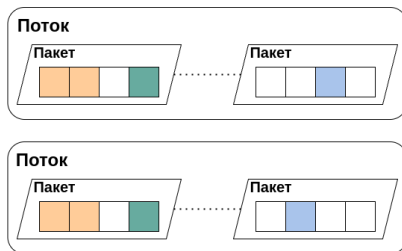
### Цель:

разработка и реализация метода автоматической генерации сигнатур полезной нагрузки сетевого трафика для классификации этого трафика в соответствии с используемым протоколом.

### Необходимо решить следующие задачи:

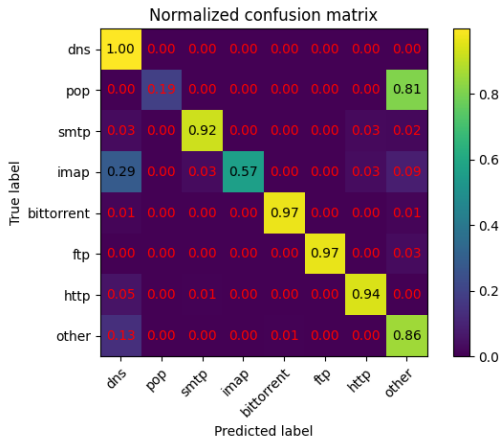
- провести исследование литературы по соответствующей теме
- собрать набор сетевых трасс для последующего тестирования
- выбрать формат сигнатуры сетевых протоколов
- разработать алгоритм генерации сигнатур
- разработать классификатор сетевого трафика для проверки сгенерированных сигнатур
- встроить генератор сигнатур и классификатор как модули в систему анализа сетевого трафика, разрабатываемую в ИСП РАН

- 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



- **LASER** (Application Signature ExtRaction)
  - **LCS** (Long Common Subsequence)
    - сигнатура пакета
    - + не требуется сборка сессии
    - + не хранит все потоки
    - 1 последовательность подстрок
- AutoSig
  - сигнатура потока
  - + набор последовательностей подстрок
  - требуется сборка сессии
  - хранит все потоки
- SigBox
  - сигнатура потока
  - + не требуется сборка сессии
  - хранит все потоки
  - 1 последовательность подстрок

Протокол	Количество потоков: $\geq 5$ pkts	Количество сигнатур	Среднее количество потоков на 1 сигнатуру
bittorent	214	51	4,1
dns	11027	682	16,1
ftp	725	9	80,5
http	1500	145	10,3
imap	65	10	6,5
pop	40	2	20,0
smtp	799	167	4,8



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



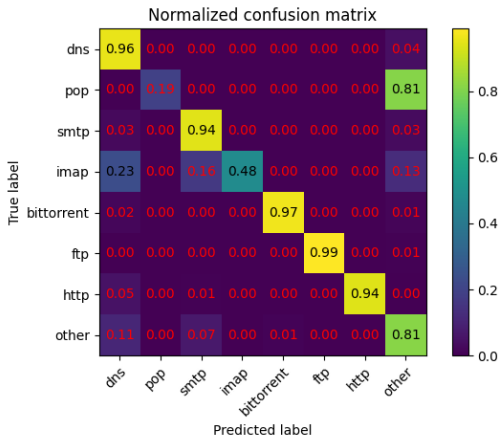
## Избыточность

Это отношение числа потоков, идентифицированных двумя и более сигнатурами, к общему числу потоков, идентифицированных набором сигнатур.

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

- Такая постобработка не влияет на результат классификации.
- Сильно сократилось количество сигнатур, однако показатель избыточности всё ещё высокий.

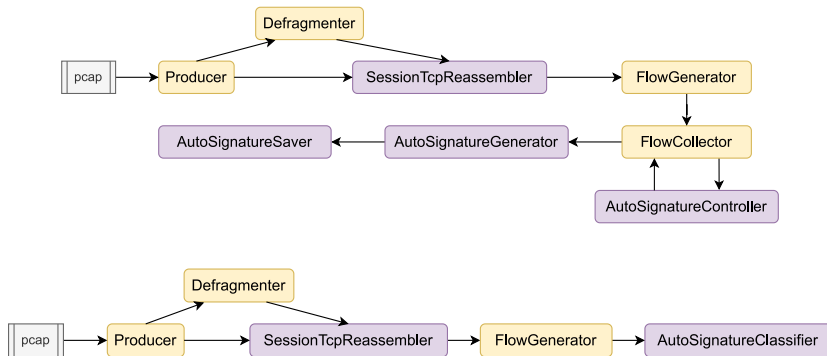
# Результаты работы алгоритма LASER: со сборкой последовательных пакетов TCP-сессии



## 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

- При частичной сборке TCP-сессии ухудшились несильно.
- При полной сборке TCP-сессии, результаты значительно хуже и сильно зависят от ограничения рассматриваемой полезной нагрузки.



- 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.

- 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 для отбора сигнатур при постобработке результатов.

Протокол	Размер, МБ	Количество пакетов	Количество поток	Avg bytes/pkt	Avg pkts/stream	Количество поток: $\geq 5$ 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

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



# Результаты работы алгоритма LASER: с полной сборкой TCP-сессии

