Обзор задач и методов их решения в области классификации сетевого трафика★

А. И. Гетьман <thorin@ispras.ru>
Ю. В. Маркин <ustas@ispras.ru>
Д. О. Обыденков <obydenkov@ispras.ru>
Е. Ф. Евстропов < john0606@yandex.ru>
ИСП РАН, 109004, Россия, г. Москва, ул. А. Солженицына, дом 25

Аннотация. В статье рассматривается задача классификации сетевого трафика: характеристики, используемые для её решения, существующие подходы и области их применимости. Перечисляются прикладные задачи, требующие привлечения компонента классификации и дополнительные требования, проистекающие из особенности основной задачи. Анализируются свойства сетевого трафика, обусловленные особенностями среды передачи,а также применяемых технологий, так или иначе влияющие на процесс классификации. Рассматриваются актуальные направления в современных подходах к анализу и причины их развития.

Ключевые слова. Анализ сетевого трафика, сетевая безопасность, классификация сетевого трафика, машинное обучение, DPI

DOI: 10.15514/ISPRAS-2017-29(3)-8

Для цитирования: Гетьман А.И., Маркин Ю.В., Евстропов Е.Ф., Обыденков Д.О. Обзор задач и методов их решения в области классификации сетевого трафика. Труды ИСП РАН, том 29, вып. 3, 2017 г., стр. 117-150. DOI: 10.15514/ISPRAS-2017-29(3)-8

1. Введение

В общем виде задача классификации сетевого трафика может быть сформулирована следующим образом: получение на вход некоторых характеристик сетевого трафика с выдачей на выходе класса, к которому данный вид трафика относится. В качестве входных характеристик могут выступать как данные пакетов, так и различные частотные характеристики, а в качестве выходных, как идентификатор конкретного приложения, ответственного за генерацию этого трафика, так и идентификатор вида

★ Работа поддержана грантом РФФИ 14-07-00606 А

трафика, например VoIP-трафик. Данная задача является одной из центральных тем в области организации сетевого взаимодействия. Исторически эта задача была наиболее актуальна в области управления трафиком для повышения эффективности использования существующих каналов связи и качества предоставляемых услуг для конечных пользователей. Однако на данный момент актуальность данной задачи значительно возросла, в связи с расширением её области применения, в которую на данный момент входят как системы применения политик, так и сфера информационной безопасности. Практически любая система анализа трафика в том или ином виде включает в себя компонент классификации.

Задача классификации трафика исследуется достаточно давно: её анализу и поиску эффективных решений в различных условиях и ограничениях посвящено значительное количество исследовательских работ, в том числе и за последние годы. Это связано, в том числе из тем, что сетевой ландшафт быстро меняется и методы, и алгоритмы, ещё недавно показывавшие хороший результат, в новых условиях значительно теряют свою эффективность или становятся вовсе неприменимыми. Среди условий, которые значительно влияют на применимость различных методов, можно выделить быстрый рост количества передаваемого трафика и пропускных способностей каналов связи – это приводит к необходимости поиска алгоритмов с пониженной вычислительной сложностью. Ещё одной тенденцией является значительное увеличение доли шифрованного трафика, что приводит к неприменимости подходов на основе анализа содержимого. Кроме того, в условиях распространения средств анализа и фильтрации многие разработчики сетевых приложений развивают механизмы, противодействующие идентификации используемых протоколов, что также усложняет анализ. В качестве примера такой тенденции можно привести историю развития P2P протоколов, которые начали активно фильтроваться со стороны интернет-провайдеров из-за того, что они слишком сильно нагружали существующие каналы связи, ухудшая качество сервиса для других пользователей этих же каналов. Это в свою очередь привело к ответной реакции от разработчиков P2P-клиентов в виде обфускации используемых протоколов для усложнения их идентификации.

В прикладной области данная задача также широко представлена – существует большое количество как коммерческих, так и свободно распространяемых систем, важнейшие компоненты которых отвечают за её решение.

Спектр предлагаемых решений достаточно широк - известны программные, программно-аппаратные, и полностью аппаратные реализации. Отчасти это связано с тем, что решение данной задачи имеет больше количество практических приложений, среди которых можно выделить:

- системы сбора статистики;
- системы управления трафиком, например, обеспечивающие качество связи (QoS, QoE) и оптимизирующие пропускную способность канала

(Wan Optimization);

- защитные системы: межсетевые экраны (NGFW), системы обнаружения и предотвращения вторжений (IDS/IPS), системы блокировки спама;
- системы применения политик к сетевому трафику (PCEF, PCRF, NAC).

2. Систематизация алгоритмов классификации

Для решения задачи классификации предложено большое количество алгоритмов, которые, в свою очередь можно классифицировать по используемым в них подходам.

С ростом числа подходов возникла потребность в их классификации. Один из вариантов классификации подходов, использующийся в компании Cisco [1], приведён на рис. 1.

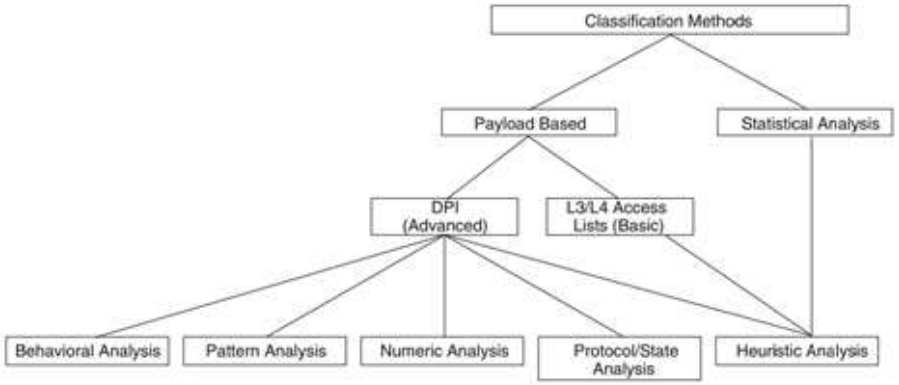

Рис. 1. Подходы к решению задачи классификации
Fig. 1. Approaches to solving the classification problem

Два основных направления – анализ содержимого передаваемых пакетов (payload-based) и статистический анализ характеристик передачи (statistical analysis): последовательность размеров пакетов, временные интервалы между пакетами и т. д.

Два наиболее распространённых метода применяемых для анализа содержимого передаваемых пакетов это поиск сигнатур (Pattern Analysis на рис. 1) и применение разборщиков данных различных протоколов (Protocol/State analysis). Сигнатуры обычно формулируются в терминах регулярных выражений. Хотя эта языковая конструкция является недостаточно выразительной и многие особенности протоколов не могут быть описаны в её терминах, преимуществом данного подхода является его скорость и масштабируемость по количеству сигнатур. В случае тривиальных

строковых сигнатур применяется группа алгоритмов массового поиска строк – типа Ахо-Корасик и турбо Бойер-Мур. В случае регулярных выражений применяются подходы на основе детерминированных и недетерминированных автоматов и гибридные варианты. Подробнее эти подходы рассмотрены в работе [2]. Подход на основе применения разборщиков, с другой стороны, имеет ряд недостатков:

- сложность разработки полноценного разборщика сообщений по сравнению с относительно простыми сигнатурами;
- более низкую скорость работы, которая зависит от применяемых алгоритмов разбора;
- плохую масштабируемость по количеству поддерживаемых протоколов – в худшем случае линейную, так как для гарантированного распознавания может потребоваться полный перебор всех разборщиков на отдельном пакете.

Последний недостаток обычно пытаются компенсировать, улучшая сложность в среднем, за счёт интеллектуального выбора последовательности применения разборщиков. Такие подходы будут подробнее рассмотрены ниже.

Основным преимуществом данного подхода является его высокая точность, так как проверяется полное соответствие структуры сообщения некоторому формату. Это свойство позволяет использовать данный подход для верификации работы других алгоритмов.

Статистические методы, в свою очередь, можно разделить на группы, в зависимости от того, характеристики какого объекта используются для классификации:

- характеристики отдельных пакетов в рамках отдельного потока (packet based);
- характеристики потоков в целом (flow based);
- характеристики нескольких потоков одного сетевого узла (host based);
- характеристики графов потоков (graph-based), применяющиеся в основном для детектирования P2P-протоколов.

Выбор конкретного метода определяется такими факторами как:

- Необходимая пропускная способность – у алгоритмов packet based она минимальна, так как выше объём обрабатываемых данных, у graph-based – максимальна по тем же причинам.
- Необходимая скорость принятия решения о классификации, т.е. количество информации которое надо собрать перед принятием решения. Packet based подходы, как правило позволяют классифицировать поток во время его активности, в то же время для graph-based подходов требуется длительное время наблюдения для сбора статистики.

- Возможные точки в топологии сети для получения информации о трафике, т.е. точки подключения системы классификации или её компонентов ответственных за сбор данных. Так для graph-based требуется возможность получения информации о значительной доли трафика всей сети.

3. Проблемы, возникающие при реализации новых подходов

Однако, несмотря на достаточно активное развитие области классификации сетевого трафика во многих работах отмечается ряд объективных факторов, сдерживающих это развитие [3]. Одним из таких факторов является отсутствие открытого набора данных для тестирования, в качестве которых обычно выступают сохранённые и размеченные сетевые трассы. Вследствие этого затруднено как тестирование качества разрабатываемого алгоритма, так и его сравнение с другими алгоритмами. В частности, это приводит к необходимости решения двух проблем в процессе разработки каждого нового алгоритма.

- Получение собственной сетевой трассы на внутренней сети, от партнёров по исследованию или из публичных источников. Осложняющим фактором является проблема приватности и возникающих рисков информационной безопасности. Для нивелирования этих факторов получаемые трассы, как правило, предварительно подвергаются процедуре анонимизации [4]. Это, в свою очередь, приводит к неприменимости подходов на основе анализа содержимого, так как основным методом анонимизации, помимо прочего, является удаление содержимого пакета уровня приложения.
- Эталонная разметка трассы по протоколам и приложениям, для последующего контроля качества разрабатываемого алгоритма, которая может выполняться несколькими основными способами, в зависимости от того, контролируется ли процесс снятия сетевой трассы или трасса получена из внешнего источника:
- Трасса из внешнего источника:
- Разметка вручную, что является очень ресурсоёмким процессом, с высокой вероятностью ошибок.
- Автоматически с помощью доступных средств классификации трафика. Данный подход приводит к проблеме качества классификации используемого эталонного средства, известной как «ground truth problem» [5].
- Контролируемое получение трассы:
- Ручная разметка по приложениям, которые в момент снятия трассы

выполнялись в системе.

- Использование автоматических средств разметки в процессе снятия трассы, например, [6].

В результате, в большинстве исследовательских работ используются разные трассы, полученные в разных точках разных сетей, на разных сценариях, в разные по длительности промежутки времени.

С другой стороны, требование приватности приводит к более активному развитию статистического направления классификации. Это происходит вследствие того, что для этой группы не требуется доступ к данным пакетов, а достаточно только общих характеристик, таких как размер и метка времени[7,8]. Таким образом, в качестве входных данных подходит большое количество открытых сетевых трасс, прошедших процедуру анонимизации.

4. Системы, использующие классификацию трафика

Помимо вопроса используемого подхода другим важным фактором является прикладная задача, решаемая конкретной системой, в рамках которой реализуется, компонент классификацию. В зависимости от этого, например, может заметно отличаться приемлемый уровень точности результатов классификации. Кроме того, может значительно отличаться и набор групп, на которые разбивается множество классифицируемых объектов. Наиболее грубая классификация используется, как правило, в системах управления трафиком, основной задачей которых является эффективное использование доступной полосы пропускания. Например, провайдер интернета может выделять три основные группы трафика.

- Чувствительный – вид трафика, который чувствителен к задержкам и требует скорейшей доставки. К нему можно отнести VoIP, потоковое видео, трафик онлайн игр.
- Нежелательный – спам и вредоносные типы трафика.
- Остальной – трафик, которому выделяется полоса, оставшаяся после обслуживания потоков чувствительных данных.

Защитные системы и системы применения политик подразумевают, как правило, значительно более точную классификацию – требуется определить конкретное приложение, генерирующее соответствующий трафик. В некоторых случаях необходимо проводить полный разбор трафика с выделением передаваемых команд и высокоуровневых объектов, таких как веб-страницы и другие виды файлов. Это может требоваться, например, для обнаружения потенциально опасного содержимого. Для оценки грубости конкретного подхода используется термин «гранулярность».

Важной характеристикой алгоритма классификации является то, в какой момент времени от начала поступления данных некоторого сетевого потока принимается решение о его принадлежности к тому или иному классу. Для

описания этой характеристики в работе [9] используется термин «ранняя классификация», подразумевающий, что результат классификации появляется вскоре после получения первых пакетов потока (в работе – от 1 до 4 пакетов), что позволяет использовать его, например, в процессе маршрутизации для присвоения приоритета на основе вида трафика. Эта характеристика влияет и на то, в каком классе систем, из приведённых выше, данный алгоритм может использоваться. Например, если алгоритм классификации принимает решение в момент завершения сетевого соединения, то это вполне приемлемо для систем сбора статистики, но неприемлемо для защитных систем.

Фактором, влияющим на оценку подхода, является скорость обработки, т.е. пропускная способность алгоритма. Данная характеристика складывается из двух – количества данных, которые алгоритм должен обработать для получения результата и сложности алгоритма относительно длины входа. Эта характеристика наиболее актуальна для DPI-подходов, которые используют максимальное количеств данных для обработки – всё содержимое отдельных пакетов. Данный вопрос подробно исследуется в большом числе работ, в основном в контексте выбора типа автомата для поиска сигнатур различных протоколов: детерминированный, недетерминированный или некоторый гибридный вариант [10-15].

Важной комплексной характеристикой системы классификации трафика является область ей применимости. В неё входит, как способность системы обрабатывать отдельные виды трафика (шифрованный, p2p и т.д.), так и то, в каких условиях данная система может функционировать и к каким особенностям передачи трафика она устойчива (потери и перестановки пакетов, асимметрия и т.д.). Указанные особенности трафика и их влияние на применимость, точность и скорость работы различных подходов будут рассмотрены ниже.

4.1 Оценка систем классификации

С учётом вышесказанного в работе [16] вводятся параметры конкретных реализации системы классификации, и приводится оценка подходов к классификации по этим параметрам. Эти оценки приведены на рис. 2.

- Точность – общая характеристика, отражающая долю правильно идентифицированного трафика от общего количества проанализированного трафика. Точность результатов определяется в основном тем, насколько хорошо выбраны признаки, по которым осуществляется классификация и качеством применяемой эвристики.
- Время реакции – время от момента получения первого пакета некоторого сетевого потока до момента его классификации. Является критичным для систем, работающих «на потоке», в частности защитных и систем управления трафиком. В это понятие также входит общая производительность алгоритма.

- Надёжность – отражает область применимости системы (например, возможность анализировать зашифрованный трафик) и устойчивость к возникающим в процессе передачи эффектам, таким как потери пакетов, асимметрия и т.д.

Classification methods	Port-based	Payload-based	Statistical classification	Host behavior based
Accuracy	Low	Low	Higher	Higher
Real-time	High	Middle	Higher	Higher
Robustness	Low	Low	Higher	Higher
Advantages	Simple,small computational overhead	No	Robustness, accuracy, fine-grained	Simple,small computational overhead
Disadvantages	Low accuracy, cannot be used alone	Almost useless, privacy risk	Large computational overhead,a lot of training,not stable when traffic changes	Coarseclassification,useless when transport layer encrypted, degradation in case of NAT
Status	Not in use	Not in use	Under test	Under test

Рис. 2. Оценка различных подходов к классификации трафика
Fig. 2. Evaluation of different approaches to traffic classification

Также в работе [16] приведены формулы, использующиеся для оценки точности алгоритма (в приведённых выше терминах). Для этого используется ряд метрик, основанных на доле истинных и ложных результатов классификациитrue/falsepositives/negatives (TP, TN, FN, FP). Наиболее часто используемыми метриками являются следующие [17].

- Правильность: Accuracy = (TP + TN)/(TP + TN + FP + FN).
- Точность: Precision = TP/(TP + FP).
- Полнота: Recall = TP/(TP + FN).
- F-мера. Так как максимальная точность и полнота недостижимы одновременно, то приходится искать некий баланс, который может оцениваться с помощью гармонического среднего между точностью и полнотой:
    - F=2*(Precision * Recall) / (Precision + Recall), в случае одинакового веса точности и полноты.В работе [18] был выбран другой набор параметров для оценки алгоритмов классификации. В него вошли: используемые для классификации данные, гранулярность результата, время принятия решения и вычислительная сложность подхода. Результаты сравнения подходов по этим параметрам

приведены на рис. 3.

Approach	Properties exploited	Granularity	Timeliness	Comput. Cost
Port-based	Transport-layer port	Fine grained	First Packet	Lightweight
Deep Packet Inspection	Signatures in payload	Fine grained	First payload	Moderate, access to packet payload
Stochastic Packet Inspection	Statistical properties of payload	Fine grained	After a few packets	High, eventual access to payload of many packets
Statistical	Flow-level properties	Coarse grained	After flow termination	Lightweight
	Packet-level properties	Fine grained	After few packets	Lightweight
Behavioral	Host-level properties	Coarse grained	After flow termination	Lightweight
	Endpoint rate	Fine grained	After a few seconds	Lightweight

Рис. 3. Сравнение подходов к классификации сетевого трафика
Fig. 3. Comparison of approaches to classification of network traffic

Совокупность оценок подходов, приведённых на рис. 2 и 3 могут быть использованы для определения применимости конкретного подхода в заданных ограничениях. Формулы для количественных оценок могут использоваться для сравнения нескольких реализаций алгоритмов классификации в рамках выбранного подхода.

5. Особенности сетевого трафика, влияющие на скорость и точность классификации

Помимо особенностей выбранного подхода и качества его реализации на скорость, точность и применимость конкретного алгоритма значительное влияние могут оказывать особенности передачи данных по сети. Следует отметить, что большинство описываемых далее особенностей не касаются напрямую задачи классификации, а являются предпосылками, на основе которых могут формулироваться требования, предъявляемые к системе, использующей компонент классификации. По сути, большая часть особенностей формирует набор необходимых видов предобработки, которые необходимо применять к данным, передаваемым посредством сетевых пакетов, прежде чем передавать их на вход алгоритму классификации. Отсутствие соответствующих видов предобработки может приводить либо к сужению области применимости, либо к потенциальному снижению точности некоторых подходов, либо к падению пропускной способности алгоритма.

5.1 Асимметрия

В зависимости от топологии сети и месте размещения компонента классификации может возникать ситуация, при которой не все передаваемые по сети пакеты будут проходить через компонент классификации – возникает асимметрия трафика. Подобная ситуация, является весьма распространённым явлением в корпоративных сетях [19], а также на магистральных каналах [20]. В работе [21] приведена типичная схема сети (рис. 4), приводящая к возникновению асимметрии. Там же приводится классификация видов асимметрии.

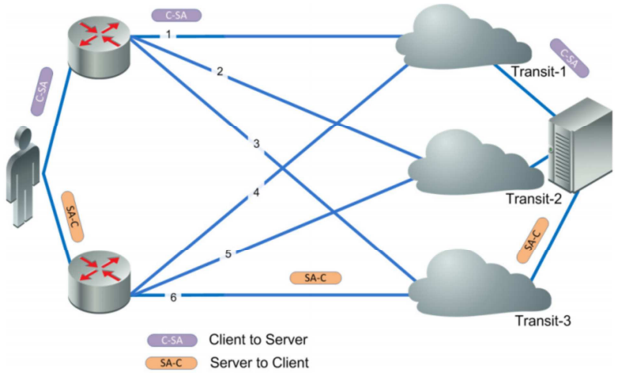

Рис. 4. Пример маршрутизации EQMP, при которой возникает асимметрия
Fig. 4. An example of EQMP routing, in which asymmetry arises

- Потоковая асимметрия.
  - Полная потоковая асимметрия, при которой любой пакет в обоих направлениях может проходить через любой шлюз.
  - Частичная консистентная потоковая асимметрия, при которой все пакеты одного потока заданного направления проходят через один шлюз, а все пакеты обратного направления этого же потока – через другой.
- IP-асимметрия.
  - Полная IP-асимметрия, при которой все потоки с одного IP-адреса могут проходить через любые шлюзы.
  - Частичная IP-асимметрия, при которой все пакеты одного потока для конкретного IP-адреса в обоих направлениях проходят через один шлюз, но разные потоки этого IP-адреса могут проходить через разные шлюзы.
  - Частичная консистентная IP-асимметрия, при которой при которой все пакеты одного IP-адреса заданного направления проходят через один шлюз, а все пакеты обратного

направления этого же IP-адреса – через другой.

При организации спутниковых каналов связи, асимметрия трафика является неотъемлемым свойством передачи данных [22]. Асимметрия может значительно влиять на качество классификации. Её наличие может приводить к:

- усложнению TCP-нормализация в отсутствие ACK-пакетов,
- невозможности IP-дефрагментации из-за отсутствующих пакетов,
- невозможности осуществить декодирование потокового сжатия,
- нарушению временных и размерных характеристик трафика.

В работе [21] рассматриваются различные способы решения проблемы асимметрии – от замыкания всех потоков в одной точки, до реализации распределённой системы классификации.

5.2 Группировка пакетов в потоки и сеансы

В процессе отдельного сеанса связи последовательность пакетов передаётся по сети в рамках некоторого потока, обладающего характеристиками, общими для соответствующей группы пакетов. Существует ряд используемых определений потока, наиболее распространённые из которых приведены на ресурсе [23]. В данной работе, при упоминании*потока* будет подразумеваться «односторонний поток транспортного уровня» – последовательность пакетов передающихся с заданного IP-адреса и TCP/UDP порта на данный IP-адрес и TCP/UDP порт, с указанием протокола транспортного уровня (TCP/UDP). Таким образом, поток задаётся пятёркой <srcIP, srcPort, dstIP, dstPort, protocol>. Так как пакеты, относящиеся к одному потоку, в некоторый промежуток времени, как правило, генерируются одним и тем же приложением, то обычно объектом классификации является не отдельный пакет, а поток целиком. Это позволяет оптимизировать задачу классификации, уменьшив объём обрабатываемых данных. Для этого используется принцип «до первого срабатывания», подразумевающий последовательный анализ пакетов одного потока, до момента его идентификации – последующие пакеты могут игнорироваться. Применение этого принципа может приводить к ошибкам (или неточностям) классификации, вследствие возможной вложенности протоколов. В частности, многие приложения, такие как Skype используют протокол HTTP в качестве транспортного для передачи своих данных.

Важность группировки возрастает в случае использования поточных алгоритмов сжатия – невозможно осуществить декодирование пакета, недекодировав перед этим предыдущие пакеты потока. Примерами таких алгоритмов являются Deflate и Gzip, которые будут подробнее рассмотрены ниже. Обобщением этой особенности является необходимость хранения некоторой информации на протяжении жизни потока – его «состояния». Помимо состояния декодера, примером такой информации является состояние протокола, для протоколов с состоянием (stateful). Причём для одного потока

состояний может быть несколько – по одному на каждый stateful протокол из используемого сетевого стека. Например, для stateful протокола QUIC – состояние будет одно, т.к. транспортный протокол UDP, поверх которого он реализован, состояния не имеет, а для протокола FTP – состояний будет два – одно собственно для FTP, второе – для TCP, поверх которого он реализован. Примером средства, архитектура которого предполагает поддержку stateful анализа на всех уровнях является инструмент GAPA [24], реализация которого, однако недоступна.

Потоки, полученные в результате группировки пакетов, в ряде случаев могут быть сгруппированы в сессии – группы связанных потоков, отвечающих за предоставление некоторого сетевого сервиса. Одним из ярких примеров сессии является пара потоков протоколов SIP и RTP, первый из которых отвечает за передачу команд, а второй – данных, при организации сеанса VoIP связи. Понятие *сессия* используется, например, в описании компонента классификации NBAR от компании Cisco [25]. Выделение сессий в некоторых случаях необходимо для корректной классификации. Примером может служить классификация потока данных FTP для FTP-сервера, работающего в пассивном режиме на основе анализа командного потока: в самом потоке данных отсутствуют какие либо заголовки и сигнатуры, так как данные передаются в «сыром» виде в ответ на соответствующую команду, передающуюся в другом потоке.

Возможна и обратная ситуация, когда несколько сессий (в смысле взаимодействий, соответствующих некоторому сетевому сервису) могут быть упакованы в один поток транспортного уровня. Простым примером такой ситуации является режим постоянного HTTP-соединения (keepalive), появившийся в версии 1.1[26] протокола - использование одного TCP-соединения для отправки и получения многократных HTTP-запросов и ответов вместо открытия нового соединения для каждой пары запрос-ответ. Развитием этой идеи является мультиплексирование, появившееся в версии 2.0[27], которое позволяет одновременные многократные запросы/ответы в одном соединении. Более сложным примером может служить протокол CitrixIndependentComputingArchitecture (ICA) [28], который в одном из режимов работы позволяет нескольким приложениям одновременно использовать одно TCP соединение. Этот пример также демонстрирует, что правило «один поток – одно приложение» не является абсолютным.

Следует отметить, что многие средства классификации, такие как nDPI [29], снимают с себя задачу группировки пакетов в потоки, предполагая, что данные, передаваемые на классификацию, уже сгруппированы в потоки некоторым внешним компонентом.

5.3 Изменчивость

С течением времени характеристики сетевого трафика меняются – в существующие протоколы вносятся изменения, появляются новые протоколы.

Это приводит к необходимости поддержания компонента классификации в актуальном состоянии. В общем случае для этого необходимо решать две задачи – выявление новых классифицирующих признаков или уточнение существующих, а также внедрение новых признаков в компонент классификации, работающий «на потоке». В случае port-based подходов это означает необходимость регулярного обновления таблицы соответствия пар <тип транспортного протокола, номер порта> протоколам прикладного уровня [30]. В случае DPI-подходов это означает необходимость автоматического [31] или ручного определения новых «сигнатур» и их добавления в компонент классификации. Среди алгоритмов автоматического поиска сигнатур можно выделить подходы на основе поиска повторяющихся шаблонов [32] и подходы на основе алгоритмов биоинформатики [33]. Альтернативный вариант получения новых сигнатур – заказ у стороннего поставщика соответствующей услуги [34]. Для их получения в этом случае обычно используется специальный программный инструментарий поддержки, примерами которого могут служить [35, 36]. Для статистических подходов, которые, как правило, основаны на алгоритмах машинного обучения это означает необходимость периодического переобучения [37], иначе их точность значительно деградирует [38]. Дополнительной трудностью для статистических алгоритмов является необходимость дополнительного класса «неизвестного трафика», в который требуется относить трафик, плохо подходящий к другим классам по своим характеристикам [39]. Независимо от подхода, возникает задача аналогичная ситуации с уязвимостью «нулевого дня» в информационной безопасности – минимизация времени от момента обнаружения неизвестного трафика до момента его корректного распознавания на потоке. В случае DPI-подходов это время суммируется из двух компонент – время на создание новой сигнатуры и время на перестроение представления, использующегося для распознавания (например, детерминированного автомата). В случае алгоритмов машинного обучения – это время на переобучение, обеспечивающее качественное распознавание нового протокола. В обоих случаях возникает дополнительная подзадача – получение достаточного количества материала для анализа и его предварительного разбиения на классы, в соответствии с которыми будет выделяться сигнатура или осуществляться переобучение. Это приводит к необходимости присутствия компонента «дополнительной классификации» трафика (возможно включающего ручной анализ), с которым не справился основной компонент классификации. В рамках этого компонента должны сохраняться образцы нераспознанного трафика, уточняться сигнатуры и эвристики, проводится переобучение (в случае алгоритмов машинного обучения) перед обновлением основного компонента классификации.

Общий вид архитектуры системы классификации, с учётом описанных подзадач имеет вид, приведённый на рис. 5, из работы [40].

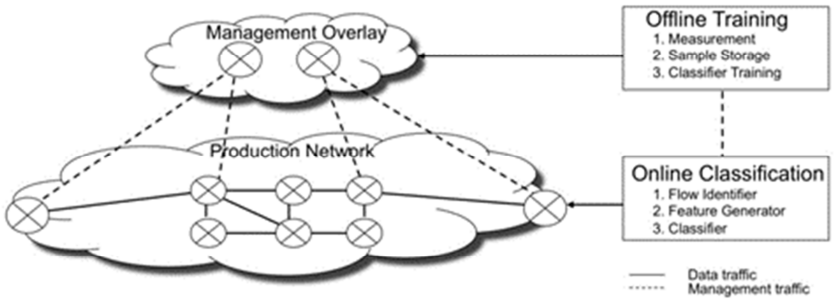

Рис. 5. Общая архитектура системы классификации трафика
Fig. 5. General architecture of the traffic classification system

5.4 Шифрование

Один из наиболее распространённых протоколов, использующихся для шифрования передаваемых данных – SSL, в частности обеспечивающий слой шифрования в HTTPS. Особенностью шифрованных данных является то, что к ним неприменимы алгоритмы классификации, использующие данные пакетов, уровня приложения (DPI на рис. 1). Таким образом, шифрование снижает применимость этого, достаточно обширного, класса алгоритмов. Кроме того, попытка применять данные подходы на шифрованном трафике, не отделяя его предварительно от остального потока, существенно снижает пропускную способность компонента классификации, так как приводит к частому проявлению «худшего» случая – просмотру всех данных пакета при отсутствии решения по его принадлежности к некоторому классу [41]. Для преодоления этого недостатка могут использоваться подходы, определяющие наличие факта сжатия или шифрования на основе измерения энтропии [41, 42]. Для классификации шифрованного трафика разрабатываются специализированные подходы. Некоторые из них основаны на анализе первых нескольких пакетов соединения – т.н. «рукопожатие», при котором стороны договариваются об используемом алгоритме шифрования, его параметрах и т.д. [43]. Однако наиболее перспективным выглядит применение подходов на основе машинного обучения, т.к. именно этому подходу посвящено большинство работ по классификации шифрованного трафика [44, 45]

В настоящее время, общая доля шифрованного трафика по данным Sandvine [44] за 2015 год составляет порядка 30% и имеет выраженную тенденцию к увеличению.Этому, в частности, способствует два основных фактора:

- решение крупного поставщика видеоконтента Netflix (наряду с YouTube) перейти на шифрованную передачу данных по HTTPS;

- инициативы Let'sEncrypt группы InternetSecurityResearchGroup (ISRG) и HTTPSEverywhere от ElectronicFrontierFoundation по автоматическому предоставлению свободных и бесплатных сертификатов всем желающими принудительного использования шифрованного соединения.

Несмотря на то, что напрямую DPI подход к шифрованному трафику не применим, во многих программных продуктах, предназначенных для классификации, в отдельных частных случаях классификация такого трафика всё же осуществляется. Это возможно, если сервис, с которым осуществляется обмен использует расширение Server Name Identification (SNI) протокола TLS для явного указания имени хоста, с которым будет осуществляться обмен. Соответственно классификатор может поддерживать базу имён хостов для наиболее часто используемых сервисов и на основании этого имени идентифицировать трафик, например к хосту google.com. Такая база, по сути, является аналогом базы портов IANA [30] с тем отличием, что идентифицирующим протокол признаком является не номер порта, извлекаемый из заголовка транспортного уровня, а строковое имя хоста, извлекаемое из заголовка TLS. Изначально это расширение было реализовано для того, чтобы была возможность предоставлять различные сертификаты для сайтов, использующих один и тот же IP-адрес и TCP-порт, что в настоящее время достаточно распространено.

5.5 Туннелирование

Туннелирование – это процесс, в ходе которого создается логическое соединение между двумя конечными точками посредством инкапсуляции различных протоколов. Может применяться для создания защищённых соединений (VPN) или для организации взаимодействия сетей, использующих разные протоколы (IPv6 и IPv4). Например, протокол GRE применяется для инкапсуляции пакетов сетевого уровня (OSI) в IP-пакеты и используется, в частности, для доступа в интернет с мобильных устройств.

От обычных многоуровневых сетевых моделей (OSI, TCP/IP) туннелирование отличается тем, что инкапсулируемый протокол относится к тому же или более низкому уровню сетевого стека, чем используемый в качестве туннеля.

В процессе туннелирования принимают участие следующие типы протоколов:

- транспортируемый – протокол объединяемых сетей;
- несущий – протокол транзитной сети;
- инкапсулирующий – помещает пакеты транспортируемого протокола в поле данных пакетов несущего протокола.

Использование туннелирования потенциально может вызывать трудности у всех подходов, но по разным причинам. В случае подходов на основе анализа содержимого (payload-based на рис. 1), требуют дополнительных усилий по разбору дополнительных заголовков, относящихся к организации туннеля,

чтобы получить данные приложения (в случае DPI) или последний заголовок транспортного уровня (в случае port-based подходов, L3/L4 Access Lists на рис.1). В случае применения статистических подходов требуется адаптация к конкретному туннелю, так как его наличие вносит искажение как в размеры пакетов (за счёт добавления заголовков), так и во временные характеристики (за счёт необходимости дополнительной обработки).

5.6 Сжатие

Используется многими протоколами для уменьшения объёма передаваемых данных, примерами могут служить DNS, где вместо повторов строк ставятся ссылки на первое вхождение и HTTP в котором применяются алгоритмы Deflate и GZIP. Влияние данной особенности с одной стороны сходно с влиянием туннелирования – могут меняться временные характеристики и размеры пакетов, а с другой – с влиянием шифрования на подходы DPI: отсутствие предварительного декодирования снижает точность и приводит к замедлению. Основное отличие от зашифрованного трафика – возможность декодирования «на лету» без знания сторонней информации (ключей шифрования). Однако само декодирование может быть ресурсоёмкой задачей, требующей дополнительных вычислительных ресурсов. Близость сжатых и шифрованных данных по их влиянию на DPI подходы подчёркивается общим термином «непрозрачные»(opaque) данные, использующимся для обозначения этих видов данных в работе [41]. Некоторые дополнительные аспекты, связанные с анализом непрозрачного трафика приведены в работе [44].

5.7 Фрагментация данных передаваемых по сети

Необходимость фрагментации связана с физическим ограничением на максимальный размер данных, которые могут быть переданы в одном пакете с использованием конкретной среды передачи (физический уровень модели OSI), а также с особенностями реализации конкретных протоколов (количеством байт в полях, хранящих размеры пакета). В частности, фрагментация имеет место на сетевом уровне (уровень IP TCP/IP стека), а её аналог - сегментация на транспортном уровне (уровень TCP TCP/IP стека). Фрагментация может иметь место и на уровне приложений – примером может служить режим chunkedtransferencoding, появившийся в версии 1.1 протокола HTTP. Фрагментация может влиять как на размеры пакетов, так и на то, что шаблон, используемый для классификации, может быть разделён между несколькими пакетами. Это особенно актуально для DPI-подходов, однако, в случае специально сгенерированных пакетов, это может повлиять даже на port-based подходы. Такие пакеты, в частности, используются для обхода межсетевых экранов при организации сетевых атак [45, 46].

6. Актуальные направления развития

С учетом факторов, затрудняющих классификацию с использованием разработанных ранее подходов, можно выделить несколько направлений для преодоления отдельных групп ограничений. Так для обработки зашифрованного трафика наряду с статистическими методами, точность которых пока относительно невысока применяется подход на основе идентификации сервиса, которые также комплементарны алгоритмам классификации и в ряде случаев позволяют повысить их точность. Для повышения скорости работы алгоритмов DPIпри сохранении относительно высокой точности применяются различные гибридные подходы, позволяющие уменьшить количество обрабатываемых данных и число проверяемых сигнатур. Для адекватного сравнения различных инструментов в условиях отсутствия открытой базы сетевых трасс и проблем с приватностью создаются модульные системы классификации. В рамках таких систем, отдельные инструменты и алгоритмы реализуются в виде независимых модулей, что позволят их оценивать и сравнивать на одних и тех же доступных конкретному исследователю трассах. Эти направления подробно рассматриваются в следующих разделах.

Для анализа P2P приложений, а также ситуаций, когда нет доступа к данным пакетов (наличие шифрования или только данных о потоках) применяются методы анализа поведения отдельных хостов и характеристик групп потоков на уровне графов сетевых взаимодействий.

6.1 Подход на основе идентификации сервиса

Одной из важных проблем для многих алгоритмов классификации является большой объём данных для обработки. Наиболее остро эта проблема стоит для DPIподходов, в которых требуется обрабатывать все данные пакетов. Для преодоления этого недостатка, независимо от подхода, используемого в работе [47], была предложена схема классификации на основе идентификации сервиса (service-basedclassification). Данная схема, по сути, является развитием схемы «до первого срабатывания», когда определив протокол по нескольким первым пакетам автоматически предполагается, что все пакеты данного потока также относятся к определённому протоколу. Идея заключается в том, что во многих сетевых взаимодействиях можно выделить серверную сторону, предоставляющую некоторый сервис по фиксированному IP-адресу и порту в течение длительного времени с использованием фиксированного протокола. Выделив серверную сторону и идентифицировав протокол для некоторого отдельного потока, можно с высокой долей вероятности утверждать, что в других сетевых потоках, одной из сторон которых является данный сервер (пара IP-адрес и порт) будет использовать тот же протокол. Ранее данная идея использовалась для построения базы знаний о доступных сетевых сервисах, с последующей их валидацией [48]; в частности, этот подход использовался для выделения P2Pтрафика [49]. Ранее рассмотренный подход классификации

шифрованного трафика на основе имени сервиса в SNI-расширении протокола TLS является частным случаем данного подхода. Среди проблем данного подхода можно указать:

- Высокую цену ошибки классификации, так как в этом случае ошибка распространится на большое число соединений с некоторым сервисом.
- Ограничение применимости к протоколам с динамическими портами, выделяемыми в рамках управляющего потока (FTP, SIP).
- Ограничение применимости к прокси-серверам (SOCKS), которые хоть и предоставляют фиксированный сервис на фиксированных портах, но этот сервис не привязан ни к какому протоколу прикладного уровня и может использоваться различными приложениями.
- Подход не применим к трафику, шифрование которого осуществляется на IP-уровне (IPSec).
- Применимость только к TCP-трафику, так как для идентификации сервера в контексте каждого отдельного соединения используется пара SYN-ACKфлагов в процессе рукопожатия при установке соединения.

6.2 Комбинации подходов и оптимизация алгоритмов

Основным преимуществом использования DPI-систем классификации является то, что помимо решения задачи классификации эти системы, как правило, применимы для полного разбора сетевых протоколов уровня приложения, извлечения передаваемых данных (сайтов, файлов, медиа-потоков и т.д.) и применение к ним высокоуровневых правил фильтрации и различных политик. Среди коммерческих систем данного класса можно указать Qosmos Intelligence Engine [50], ipoque PACE [51], Windriver Content Inspection Engine [52], Procera PacketLogic Content Intelligence [53]. Так как данные системы относятся к классу корпоративных, проблема шифрованного трафика (к которому DPI-подходы напрямую неприменимы) решается с помощью регистрации на всех машинах сети доверенного корневого сертификата, соответствующего используемому DPI-решению и реализацию схемы аналогичной MITM-атаке с распаковкой, анализом и последующей запаковкой передаваемого трафика. Примером реализации такой схемы является технология DPI-SSLот SonicWALL [54].

Среди других ограничений DPI-подхода можно указать уже упоминавшийся большой объём данных для анализа, а также рост сложности анализа с добавлением поддержки новых протоколов. Для преодоления этих недостатков используются гибридные схемы, некоторые из которых будут рассмотрены далее.

Исследование количества анализируемых байт для классификации

Целью исследования авторов работы [55] была попытка анализа известных инструментов DPI (в работе использовался инструмент с открытым исходным кодом L7[56]) на предмет:

- Количества анализируемых пакетов в одном потоке
- Наиболее частого смещения в пакетах сигнатур, поиск которых осуществляют инструменты DPI.

Для анализа количества анализируемых байт был построен график зависимости доли классифицированных потоков от величины смещения сигнатуры, по которой он был классифицирован – см. рис. 5.

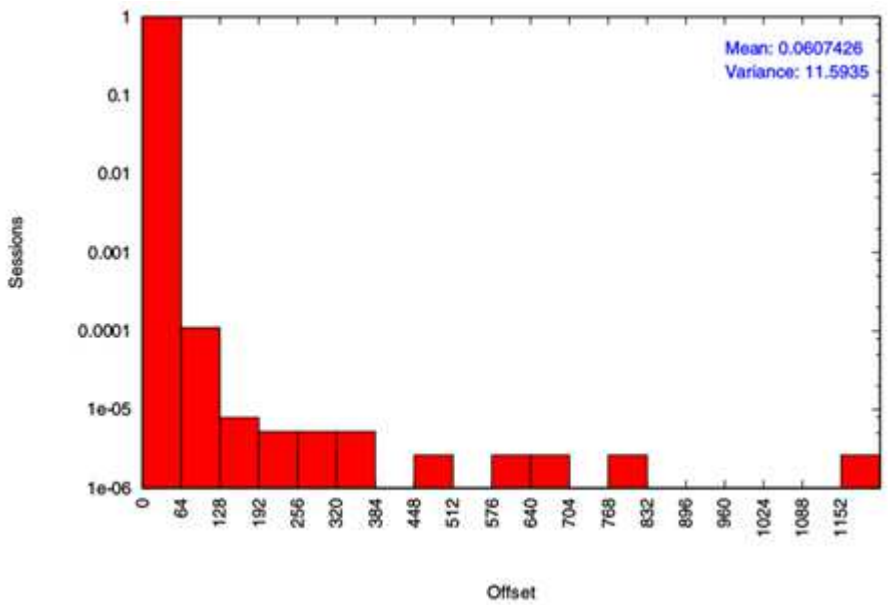

Рис. 5. График зависимости доли классифицированных потоков от смещения классифицирующей сигнатуры для инструмента L7
Fig. 5. Graph of the dependence of the fraction of classified flows on the shift of the classifying signature for the tool L7

График показывает, что в абсолютном большинстве случаев сигнатура расположена в пределах первых 32 байт первого пакета потока. Для проверки точности подхода на основании анализа только первых 32 байт-методом сравнения битовых строк, авторами был разработан инструмент PortLoad, который применял сигнатуры протоколов к первым 32 байтам по аналогии с тем, как применяется сопоставление в подходе на основе портов к 2 байтам номера порта. Было показано, что такой подход позволяет получить точность гораздо более высокую (74%), чем подход на основе анализа портов (24%) и

ненамного более низкую, чем подходна основе полноценного DPI (97%). В то же время скорость анализа была близкой к скорости анализа port-based подхода и значительно превышала скорость DPIанализа, как за счёт уменьшения количества обрабатываемых данных, так и за счёт использования более быстрых алгоритмов сравнения фиксированных строк, а не поиска регулярных выражений. Так, на исследованной трассе время работы реализованного алгоритма всего в 2.8 раза превышало время работы port-basedалгоритма и было в 30 раз меньше времени работы инструмента L7.

Легковесный анализ пакетов

Целью авторов работы [57] являлось создание инструмента классификации сравнимого по точности с известными решениями на основе DPI-подхода, однако использующим для анализа меньше данных, что делало бы его менее ресурсоёмким и более быстрым.

Одной из проблем, возникающих при реализации подхода на основе DPI – ограниченность данных для анализа, вследствие проблем с приватностью: большинство свободно-распространяемых трасс, в частности организацией CAIDA [23] содержит «обрезанные» пакеты. Ограничение длины входных данных, необходимых для анализа позволило бы использовать такие трассы, решая данную проблему.

Свой подход авторы определили, как легковесный анализ пакетов (LightweightPacketInspection, LPI), и реализовали в виде библиотеки с открытым исходным кодом. Идея заключается в использовании совокупности подходов на основе анализа портов, статистического анализа и DPI для нивелирования ограничений при сохранении точности. В качестве сигнатуры используется совокупность значений:

- IP-адресов и портов обоих участников обмена (на основе портов)
- Длины пакетов в обоих направлениях (статистические подходы)
- Первые 4 байта пакетов уровня приложения в обоих направлениях (DPI)

Обоснованием уменьшения размеров анализируемых данных пакетов уровня приложения являются следующие факты:

- в работе [55] было показано, что для классификации большинства приложений используются сигнатуры для префикса пакетов ограниченной длины (в работе использовался префикс длины 32)
- большинство свободно-доступных трасс используют стандартную анонимизацию, которая заключается в удалении содержимого пакетов уровня приложения, за исключением первых 4х байт.

При вызове функции идентификации последовательно вызываются независимые модули распознавания, соответствующие различным протоколам (на данный момент поддерживается около 200 протоколов). Порядок вызова модулей соответствует их приоритетам (задаются вручную), которые

выставляются в соответствии с точностью и простотой функции идентификации и популярностью данного протокола.

Функция идентификации может включать четыре вида правил:

- Правила на содержимое – может включать символ *, соответствующий произвольному символу
- Правила на размер пакета, использующиеся для протоколов не содержащих явных сигнатур в первых 4х байтах
- Правила на номера портов, применяющиеся только для портов с фиксированными протоколами
- Правила на IP-адреса, использующиеся для идентификации специфических сервисов

Оценки качества классификации показывают, что хотя точность анализа и ниже, чем у других открытых инструментов, в ряде случаев она может быть достаточна для практического применения.

Извлечение метаданных

Авторы инструмента с открытым исходным кодом nDP [15] рассматривают решение задачу классификации не как некоторую самоцель, но в комплексе, на примере того какие задачи на её основе решаются в коммерческих инструментах, таких как iPoque [51] и Qosmos [50]. Они приходят к выводу, что помимо определения протокола уровня приложения требуется извлекать метаданные, т.е. значения отдельных полей высокоуровневых протоколов, что в свою очередь требует разработки полноценных разборщиков этих протоколов. Наличие таких разборщиков, в свою очередь означает, что факт их применимости к некоторым данным может служить гарантией того, что данные относятся к протоколу, которому соответствует разборщик. Эта же идея лежит в основе инструмента Wireshark [58]. В этом случае ключевым вопросом производительности является порядок применения разборщиков к данным, чтобы избежать линейного роста сложности по мере увеличения числа поддерживаемых протоколов. Для улучшения скорости анализа в среднем использовались следующие предположения:

- Повышенная вероятность работы приложений с учётом транспортного протокола на закреплённых за ними портах [30].
- Уменьшение количества применяемых разборщиков по мере анализа последовательности пакетов в потоке с отбрасыванием тех, которые должны были сработать на некотором начальном префиксе данных потока.

Также в работе [15] было проведено исследование на предмет минимального количества пакетов, необходимого для классификации потоков различных протоколов. На основании исследования был сделан вывод о том, что эта величина сильно протоколо-зависима и хотя для многих протоколов составляет 1(DNS, NetFlow, SNMP) для специфических протоколов, таких как

BitTorrent она равна 8. Этот порог и был выбран в качестве ограничения для количества анализируемых пакетов в потоке.

Благодаря возможности извлечения метаданных в инструмент был добавлен подход на основе идентификации сервиса по имени хоста (HTTP) или используемым сертификатам в случае шифрованного потока (SNI). Для сопоставления имён с базой известных сервисов использовалась реализация алгоритма Ахо-Корасик. Ещё один побочный положительный эффект от извлечения метаданных – возможность классификации связанных потоков, например, автоматическая классификация потока данных RTP, на основе разбора управляющего потока SIP. Ещё более актуально это для протокола FTP, в случае работы сервера в пассивном режиме – в этом случае поток данных в принципе не содержит специфических сигнатур и заголовков и может быть классифицирован только на основе анализа потока команд.

Сам инструмент nDPI разрабатывался на основе открытой кодовой базы проекта OpenDPI, которая была в значительной мере переработана, в частности, для оптимизации работы в многопоточном режиме.

6.3 Анализ графов взаимодействий

Одной из первых работ, в которой авторы исследовали поведения отдельного сетевого узла с точки зрения шаблонов его взаимодействия с другими узлами является работа [59]. Авторы выделяют несколько уровней шаблонов сетевых взаимодействий отдельного сетевого узла:

- Социальный – количество сетевых узлов (IP-адресов), с которыми данный узел взаимодействует (популярность узла) и их объединение в связанные сообщества.
- Функциональный - набор ролей, в которых выступает данный узел (клиент, сервер или оба варианта). На данном уровне учитываются взаимодействия по разным портам.
- Прикладной – анализ проводится на уровне транспортных потоков (пар IP-адресов и портов) для идентификации отдельных приложений по эвристическим шаблонам, включающим количество пакетов, байт и транспортный протокол передачи.

Авторы показывают, что такой подход позволяет детектировать p2pвзаимодействия, новые неизвестные протоколы, а также некоторые виды атак.

Развитие идея анализа графов для анализа сетей получила в работе [60]. В ней произошла формализация представления массива сетевых взаимодействий в виде графа дисперсии трафика (TrafficDispertionGraph, TDG). Данный граф описывается как пара множеств V-вершин и E-рёбер, где:

- v изV – вершина графа, точка (сетевой узел) в сети, с заданным IP-адресом;

- e из E – ребро, показывающее наличие сетевого потока между сетевыми узлами.

В работе было показано, что клиент-серверные взаимодействия (HTTP)значительно отличаются от взаимодействий типа P2Pпо плотности графа взаимодействий сетевого узла (2|E|/|V|). В случае P2Pэта плотность значительно выше. Всего для анализа использовались три характеристики:

- плотность;
- количество сетевых узлов с двусторонними соединениями (ведут себя и как клиент и как сервер);
- эффективный диаметр (из 90% выборки взять наибольший диаметр графа).

По совокупности этих параметров взаимодействия P2P отличались от HTTP-подобного трафика с практически 100% точностью. На основе этих фактов был предположен метод, основанный на классификации трафика по поведению сетевых узлов, который можно эффективно использовать для распознания P2P трафика. В процессе испытаний метода был выявлен существенный недостаток, который заключается в том, что данный метод плохо отличает P2Pтрафик от других стандартных протоколов типа DNSи SMTP. В процессе развития, для преодоления этого недостатка было найдено два подхода – усложнение анализа графа с учётом динамики его изменения со временем и комбинирование подхода с другими в рамках модульных систем классификации (в частности, подхода на основе анализа портов). Первый из указанных подходов рассматривается в работе [61], в котором показано, что процесс изменения TCG-графа со временем значительно отличается у DNSи P2Pи, таким образом, служит лучшей характеристикой используемого протокола. Для оценки временных изменений в работе были предложен ряд новых метрик, показывающих как изменение в структуре (без учёта конкретных рёбер и вершин), так и изменение в составе вершин и рёбер. Второй подход подробно рассматривается в следующем разделе.

6.4 Модульные системы классификации

В качестве основных аргументов для создания модульных систем классификации, в которых независимые модули реализуют различные подходы и алгоритмы классификации можно указать следующие:

- Необходимость инструмента для качественного сравнения разных подходов с целью выявления их сильных и слабых сторон на одних и тех же входных данных, что в обычных условиях невозможно из-за проблем с приватностью и отсутствия публичных тестовых наборов.
- Комбинация различных подходов с целью взаимного нивелирования их слабых сторон для повышения качества классификации.

Среди известных систем можно указать NetraMark [62]. В качестве базовых принципов разработки указываются:

- Совместимость реализуемых подходов – корректное сопоставление классов/приложений, в рамках которых осуществляется классификация в разных инструментах.
- Воспроизводимость – возможность верификации результатов, полученных разными исследовательскими группами с учётом особенностей входных данных.
- Расширяемость – простота добавления новых подходов в качестве отдельных модулей.
- Синергия – подбор комбинаций классификаторов для взаимного усиления.
- Гибкость настройки – простое переконфигурирование отдельных алгоритмов, быстрый подбор свойств, используемых для классификации.

Примерно эти же принципы были положены в основу разрабатываемой с 2008 года по настоящее время платформы TIE[63,64] (TrafficIdentificationEngine). В данной платформе больший упор сделан на перспективные методы машинного обучения, однако реализовано и большое количество подходов DPI и их оптимизированных версий, рассмотренных в предыдущих разделах. Кроме того, лучше проработаны вопросы архитектуры платформы с выделением отдельных функциональных компонент. Общая схема архитектуры приведена на рис. 6.

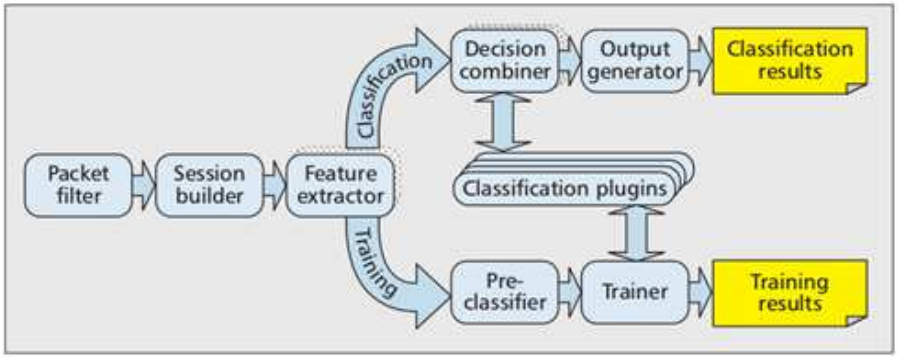

Рис 6. Общая схема архитектура платформы TIE
Fig. 6. General scheme of the architecture of the TIE platform

Исследованию вопросов совместного использования различных походов и алгоритмов классификации с точки зрения как уровней входных данных для анализа, так и осмысленности обмена результатами классификации (с целью

кросс-валидации), посвящена работа [65]. Данная работа содержит перечисление большинства известных подходов с анализом их сильных и слабых сторон, на основании которого строится схема их оптимального применения к входным данным, с учётом взаимного обмена результатами классификации. Полученная схема приведена на рис. 7.

7. Заключение

На основании проведённого обзора подходов к классификации трафика, можно сделать следующие выводы.

- Существует большое количество алгоритмов и подходов с различными достоинствами, недостатками, отличающихся по скорости обработки, области применимости и точности получаемых результатов.
- Сравнение различных алгоритмов значительно затруднено из-за отсутствия общедоступной базы полноценных размеченных сетевых трасс, на которых было бы возможно проводить сравнения. Отсутствие такой базы вызвано объективными причинами, такими как необходимость обеспечения информационной безопасности и приватности пользователей сети. Доступные наборы трасс, например, в базе организации CAIDA являются «анонимизированными», т. е. не содержат данных уровня приложения в пакетах. Это позволяет применять к ним статистические подходы и подходы, использующие заголовки 3 и 4 уровней, но исключает применение подходов на основе DPI.
- Одним из наиболее активно развивающихся на данный момент направлений является применение различных алгоритмов машинного обучения, графового и статистического анализа, по причине их применимости, в том числе к шифрованному трафику (в отличие от DPI-подходов), доля которого быстро растёт. Данное направление, однако, также не избавлено от недостатков. В частности, точность алгоритмов может снижаться, если в анализируемом потоке присутствует трафик приложений, неиспользовавшихся в процессе обучения. Другой проблемой является необходимость периодического переобучения при изменении характеристик известных протоколов и появлении новых.
- Другим развивающимся направлением является разработка комбинированных подходов и систем классификации. Одной из причин для развития является попытка преодоления недостатков отдельных подходов (например, невысокая точность или скорость обработки) и использование их преимуществ. В качестве примеров таких подходов можно привести инструменты Libprotoident и PortLoad. Создание систем классификации обусловлено

необходимостью открытой инфраструктуры, в рамках которой можно реализовать любой новый подход, что позволит сравнить его с другими на одних и тех же собственных данных, не раскрывая их сторонним лицам. Примерами таких систем могут служить NeTraMark, TIE и система из работы [65].

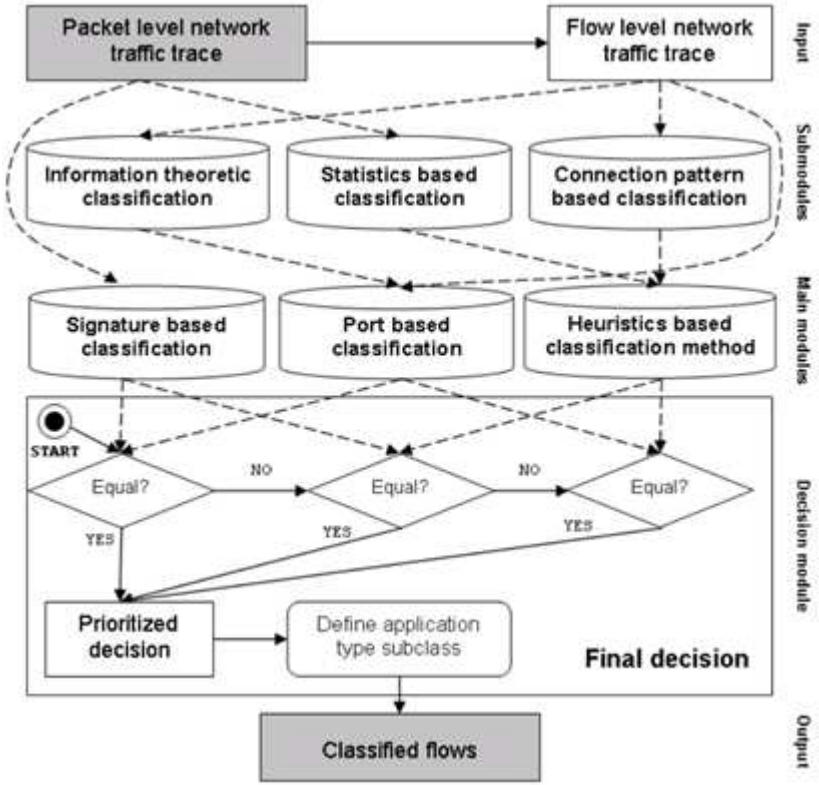

Рис 7. Схема взаимодействия разных компонент классификации трафика
Figure 7. Scheme of interaction between different components of traffic classification

Список литературы

[1]. Cisco WAN and Application Optimization Solution Guide. http://www.cisco.com/c/en/us/td/docs/nsite/enterprise/wan/wan_optimization/wan_opt_sg/chap05.html, дата обращения 01.12.2015

[2]. А.И. Гетьман, Е.Ф. Евстропов, Ю.В. Маркин. Анализ сетевого трафика в режиме реального времени: обзор прикладных задач, подходов и решений. Препринт ИСП РАН, 28, 2015 г., стр. 1-52.

[3]. M.Mellia, A. Pescapè, L. Salgarelli. Traffic classification and its applications to modern networks. Elsevier Computer Networks, Dec. 2008

[4]. T. Farah, L. Trajkovic. Anonym: A tool for anonymization of the Internet traffic. In IEEE 2013 International Conference on Cybernetics (CYBCONF), 2013, pp. 261-266.
[5]. V. Carela-Español, T. Bujlow, P. Barlet-Ros. Is Our Ground-Truth for Traffic Classification Reliable? In Proceedings of the 15th International Conference on Passive and Active Measurement - Vol. 8362. Springer-Verlag New York Inc., New York, NY, USA, 2014, pp. 98-108.
[6]. F. Gringoli, L. Salgarelli, M. Dusi, N. Cascarano, F. Risso, and K. C. Claffy. GT: picking up the truth from the ground for internet traffic //SIGCOMM Computer Communication Review, Volume 39, Issue 5, October 2009, pp. 12-18.
[7]. J. Erman, M. Arlitt, and A. Mahanti. TrafficClassificaton Using Clustering Algorithms. In ACM SIGCOMM MineNet Workshop, September 2006.
[8]. N. Williams, S. Zander, and G. Armitage. Apreliminary performance comparison of five machinelearning algorithms for practical ip traffic flowclassification. In ACM SIGCOMM CCR, Vol. 36, No. 5, pp.7-15, October 2006.
[9]. A. Dainotti, A. Pescapé, C. Sansone. Early classification of network traffic through multi-classification. In Proceedings of the Third international conference on Traffic monitoring and analysis (TMA'11), 2011. Springer-Verlag, Berlin, Heidelberg, pp. 122-135.
[10]. Cascarano N, Ciminiera L, Risso F. Optimizing deep packet inspection for high-speed traffic analysis. Network System Manager. 2011 19(1), pp. 7–31.
[11]. S. Kumar and P. Crowley. Algorithms to Accelerate Multiple Regular Expressions Matching for Deep Packet Inspection. In Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM '06), 2006, New York, USA, pp. 339-350.
[12]. D. Ficara, S. Giordano, G. Procissi, F.Vitucci, G.Antichi, A. Di Pietro. An Improved DFA for Fast Regular Expression Matching. SIGCOMM Comput. Commun. Rev. 38, 5 (September 2008), pp. 29-40.
[13]. F. Yu, Z. Chen, Y. Diao, T. V. Lakshman, and R. H. Katz. Fast and Memory-Efficient Regular Expression Matching for Deep Packet Inspection. In Proceedings of the ACM/IEEE symposium on Architecture for networking and communications systems (ANCS '06). 2006, New York, USA, pp. 93-102.
[14]. S. Kumar, B. Chandrasekaran, J. Turner, and G. Varghese. Curing Regular Expressions Matching Algorithms From Insomnia. In Proceedings of the 3rd ACM/IEEE Symposium on Architecture for networking and communications systems (ANCS '07). 2007,New York, USA, pp. 155-164
[15]. R. Smith, C. Estan, S. Jha, and S. Kong. Deflating the Big Bang: Fast and Scalable Deep Packet Inspection with Extended Finite Automata. In Proceedings of the ACM SIGCOMM conference on Data communication (SIGCOMM '08). 2008, New York, USA, pp. 207-218.
[16]. Cao Z., Cao S., Xiong G., Guo L.Progress in Study of Encrypted Traffic Classification. In Proceedings of International standard conference on trustworthy computing and services, 2012, Beijing, China, pp. 78-86
[17]. M. Sokolova, N. Japkowicz, S. Szpakowicz. Beyond accuracy, f-score and ROC: a family of discriminant measures for performance evaluation //In Proceedings of the 19th Australian joint conference on Artificial Intelligence: advances in Artificial Intelligence (AI'06), Berlin, Heidelberg, 2006, pp. 1015-1021.
[18]. S. Valenti, D. Rossi, A. Dainotti, A. Pescapè, A. Finamore, M. Mellia. Reviewing traffic classification. In DataTraffic Monitoring and Analysis, Springer-Verlag, Berlin, Heidelberg, 2013, pp. 123-147.

[19]. D. Maurizio. Observing routing asymmetry in Internet traffic. https://www.caida.org/research/traffic-analysis/asymmetry, дата обращения 01.12.2015
[20]. K. Fukuda. Difficulties of identifying application type in backbone traffic, 2010 International Conference on Network and Service Management, Niagara Falls, ON, 2010, pp. 358-361
[21]. H. Balakrishnan and V. Padmanabhan. How network asymmetry affects TCP. IEEE Communications Magazine,Vol. 39, pp. 60 -67, April 2001.
[22]. Applying Network Policy Control to Asymmetric Traffic: Considerations and Solutions. https://www.sandvine.com/downloads/general/whitepapers/applying-network-policy-control-to-asymmetric-traffic.pdf, дата обращения 01.12.2015
[23]. CAIDAFlowTypes. https://www.caida.org/research/traffic-analysis/flowtypes/, дата обращения 01.12.2015.
[24]. N. Borisov, D.J. Brumley, H.J. Wang, J. Dunagan, P. Joshi, C. Guo. A Generic Application-Level Protocol Analyzer and Its Language. In Proceedings of 14th Annual Network and Distributed System Security Symposium, 2007.
[25]. CiscoNBAR. http://www.cisco.com/c/en/us/products/ios-nx-os-software/network-based-application-recognition-nbar/index.html, дата обращения 01.12.2015.
[26]. RFC 2616. Hypertext Transfer Protocol -- HTTP/1.1. https://www.ietf.org/rfc/rfc2616.txt, дата обращения 01.12.2015.
[27]. RFC 7540. Hypertext Transfer Protocol Version 2 (HTTP/2). https://tools.ietf.org/html/rfc7540, дата обращения 01.12.2015.
[28]. Administering Cisco QoS in IP Networks. Including CallManager 3.0, QoS, and uOne. 1st Edition, Syngress 2001, eBook ISBN: 9780080481890, pp. 561
[29]. L. Deri, M. Martinelli, T. Bujlow, and A. Cardigliano, "ndpi: Opensource high-speed deep packet inspection," in Wireless Communications and Mobile Computing Conference (IWCMC), 2014 International. IEEE, 2014, pp. 617–622.
[30]. Service Name and Transport Protocol Port Number Registry. http://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml, дата обращения 01.12.2015
[31]. P. Haffner, S. Sen, O. Spatscheck, D. Wang. ACAS: automated construction of application signatures // In Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data (MineNet '05), ACM, New York, NY, USA, 2005, pp. 197-202.
[32]. Y. Wang, Y. Xiang, W. Zhou, S. Yu. Generating regular expression signatures for network traffic classification in trusted network management, Journal of Network and Computer Applications. Volume 35, Issue 3, May 2012, pp. 992-1000
[33]. G. Szabó, Z.Turányi, L. Toka, S. Molnár, A. Santos. 2011. Automatic protocol signature generation framework for deep packet inspection // In Proceedings of the 5th International ICST Conference on Performance Evaluation Methodologies and Tools, Brussels, Belgium, Belgium, 2011, pp. 291-299.
[34]. Перспективный мониторинг. http://amonitoring.ru/service/snort/, дата обращения 01.12.2015
[35]. G. Bossert, F. Guihéry, G. Hiet. Towards automated protocol reverse engineering using semantic information. In Proceedings of the 9th ACM symposium on Information, computer and communications security (ASIA CCS '14). ACM, New York, NY, USA, 2014, pp. 51-62.
[36]. Гетьман А.И., Маркин Ю.В., Обыденков Д.О., Падарян В.А., Тихонов А.Ю. Подходы к представлению результатов анализа сетевого трафика. Труды ИСП РАН, том 28, вып. 6, 2016, стр. 103-110. DOI: 10.15514/ISPRAS-2016-28(6)-7

[37]. O. Mula-Valls. A practical retraining mechanism for network traffic classification in operational environments // Master Thesis Universitat Poliecnica de Catalunya, 2011.
[38]. R. Wang, L. Shi, B. Jennings. Ensemble Classifier for Traffic in Presence of Changing Distributions. In Proceedings of the Symposium on Computers and Communications (ISCC 2013), Split, Croatia, 7-10 July, 2013, pp. 629-635
[39]. J. Zhang, C. Chen, Y. Xiang, .W. Zhou. Robust network traffic identification with unknown applications. In Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security (ASIA CCS '13), 2013, ACM, New York, NY, USA, pp. 405-414.
[40]. R. Wang. Advances in Machine-Learning-Based Traffic Classifiers. https://labs.ripe.net/Members/rwang/advances-in-machine-learning-based-traffic-classifiers, дата обращения 01.12.2015
[41]. A. White, S. Krishnan, M. Bailey, F. Monrose, P. Porras. Clear and Present Data: Opaque Traffic and its Security Implications for the Future. NDSS, 2013.
[42]. J. Olivain, J. Goubault-Larrecq. Detecting subverted cryptographicprotocols by entropy checking. Technical report, Laboratoire Spcificationet Verification, June 2006.
[43]. L.Bernaille, R. Teixeira. Early recognition of encrypted applications. In Proceedings of the 8th international conference on Passive and active network measurement (PAM'07), 2007, Springer-Verlag, Berlin, Heidelberg, 165-175.
[44]. Global Internet Phenomena Spotlight: Encrypted Internet Traffic. https://www.sandvine.com/downloads/general/global-internet-phenomena/2015/encrypted-internet-traffic.pdf, дата обращения 01.12.2015
[45]. IP Fragmentation Attacks on Checkpoint Firewalls. https://www.giac.org/paper/gsec/589/ip-fragmentation-attacks-checkpoint-firewalls/101350, дата обращения 01.12.2015
[46]. M. Handley, V. Paxson, C. Kreibich. Network Intrusion Detection: Evasion, Traffic Normalization, and End-to-End Protocol Semantics. In Proceedings of the 10th conference on USENIX Security Symposium, vol. 10. USENIX Association, Berkeley, CA, USA, 2001, pp. 9-25.
[47]. M. Baldi, A. Baldini, N. Cascarano, F. Risso. Service-based traffic classification: Principles and validation. In Proceedings of the IEEE Sarnoff Symposium (SARNOFF'09), 2009. IEEE Press, Piscataway, NJ, pp. 115–120.
[48]. W. Moore, K. Papagiannaki. Toward the AccurateIdentification of Network Applications. International Workshop on Passive and Active Network Measurement (PAM 2005), 2005, Boston MA, USA, vol. 3431, pp. 41-54.
[49]. T. Karagiannis, A. Broido, M. Faloutsos, Kc. Claffy. Transport layer identification of P2P traffic. In Proceedings of 4th ACM SIGCOMM conference on Internet measurement, 2004, pp. 121 – 134.
[50]. QosmosixEngine. http://www.qosmos.com/products/deep-packet-inspection-engine/, дата обращения 01.12.2015
[51]. Ipoque PACE. https://www.ipoque.com/products/pace, дата обращения 01.12.2015
[52]. Windriver Content Inspection Engine. http://www.windriver.com/products/product-overviews/PO_Wind-River-Content-Inspection-Engine.pdf, дата обращения 01.12.2015
[53]. Procera PacketLogic Content Intelligence. https://www.proceranetworks.com/content-intelligence.html, дата обращения 01.12.2015
[54]. DPI-SSL. https://www.sonicwall.com/ssl-decryption-and-inspection/, дата обращения 01.12.2015

[55]. G. Aceto, A. Dainotti, W. de Donato, A. Pescap. PortLoad: Taking the Best of Two Worlds in Traffic Classification," in IEEE INFOCOM 2010 – WIP Track, 2010.
[56]. L7-filter. http://l7-filter.sourceforge.net/, дата обращения 01.12.2015.
[57]. S. Alcock, R. Nelson, Libprotoident: Traffic Classification Using Lightweight Packet Inspection, Technical report, University of Waikato, 2013. http://www.wand.net.nz/publications/lpireport, дата обращения 01.12.2015
[58]. Wireshark. https://www.wireshark.org/, дата обращения 01.12.2015.
[59]. T.Karagiannis, K.Papagiannaki, M. Faloutsos. BLINC: multilevel traffic classification in the dark. In Proceedings of the SIGCOMM '05. 2005, ACM, New York, NY, USA, pp.229-240.
[60]. M. Iliofotou, H. Kim, M. Faloutsos, M.Mitzenmacher, P. Pappu, G. Varghese. Graph-based P2P traffic classification at the internet backbone. In Proceedings of the INFOCOM'09. 2009, IEEE Press, Piscataway, NJ, USA, pp. 37-42.
[61]. M.Iliofotou, M. Faloutsos, M.Mitzenmacher. Exploiting dynamicity in graph-based traffic analysis: techniques and applications. In Proceedings of the CoNEXT '09. 2009, ACM, New York, NY, USA, pp. 241-252.
[62]. S. Lee, H. Kim, D. Barman, S. Lee, C. Kim, T. Kwon, Y. Choi. NeTraMark: a network traffic classification benchmark. SIGCOMM Comput. Commun. Rev. 41, 1 (January 2011), pp. 22-30
[63]. A. Dainotti, W. Donato, A.Pescapé. TIE: A Community-Oriented Traffic Classification Platform. In Proceedings of the First International Workshop on Traffic Monitoring and Analysis (TMA '09), 2009, Springer-Verlag, Berlin, Heidelberg, pp. 64-74.
[64]. W. Donato, A. Pescape, A. Dainotti. Traffic identification engine: an open platform for traffic classification.In IEEE Network, vol. 28, no. 2, pp. 56-64, March-April 2014.
[65]. G. Szabo, I. Szabo, D. Orincsay. Accurate Traffic Classification. IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, Espoo, Finland, 2007, pp. 1-8.

A survey of problems and solution methods in network traffic classification

A. I. Get'man <thorin@ispras.ru>
Yu. V. Markin <ustas@ispras.ru>
D. O. Obidenkov <obydenkov@ispras.ru>
E. F. Evstropov < john0606@yandex.ru>
Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia

Annotation. The paper discusses the problem of network traffic classification: the characteristics that are used to solve it, existing approaches and their limitations. Applied tasks that require classification are listed, as well as additional requirements that arise from the main problem. Properties of network traffic that root in communication medium specifics are analyzed as well as the technology being used where they influence the classification process. Relevant directions in current approaches to analysis and the reasons for their development are discussed.

Keywords. Network traffic analysis, network security, network traffic classification, machine learning, DPI

DOI: 10.15514/ISPRAS-2017-29(3)-8

For citation: Ge'tman A.I., Markin Yu.V, Evstropov E.F. Obydenkov D.O. A survey of problems and solution methods in network traffic classification classification. *Trudy ISP RAN/Proc. ISP* RAS, vol. 29, issue 3, 2017, pp. 117-150 (in Russian). 10.15514/ISPRAS-2017-29(3)-8

References

[1]. Cisco WAN and Application Optimization Solution Guide. http://www.cisco.com/c/en/us/td/docs/nsite/enterprise/wan/wan_optimization/wan_opt_sg/chap05.html, accessed 01.12.2015

[2]. A.I Get'man, E.F Evstropov, Yu. V. Markin, Wirespeed network traffic analysis: survey of applied problems, approaches and solutions. Preprint ISP RAS, 28, 2015, pp. 1-52 (in Russian)

[3]. M.Mellia, A. Pescapè, L. Salgarelli. Traffic classification and its applications to modern networks. Elsevier Computer Networks, Dec. 2008

[4]. T. Farah, L. Trajkovic. Anonym: A tool for anonymization of the Internet traffic. In IEEE 2013 International Conference on Cybernetics (CYBCONF), 2013, pp. 261-266.

[5]. V. Carela-Español, T. Bujlow, P. Barlet-Ros. Is Our Ground-Truth for Traffic Classification Reliable? In Proceedings of the 15th International Conference on Passive and Active Measurement - Vol. 8362. Springer-Verlag New York Inc., New York, NY, USA, 2014, pp. 98-108.

[6]. F. Gringoli, L. Salgarelli, M. Dusi, N. Cascarano, F. Risso, and K. C. Claffy. GT: picking up the truth from the ground for internet traffic //SIGCOMM Computer Communication Review, Volume 39, Issue 5, October 2009, pp. 12-18.

[7]. J. Erman, M. Arlitt, and A. Mahanti. TrafficClassificaton Using Clustering Algorithms. In ACM SIGCOMM MineNet Workshop, September 2006.

[8]. N. Williams, S. Zander, and G. Armitage. Apreliminary performance comparison of five machinelearning algorithms for practical ip traffic flowclassification. In ACM SIGCOMM CCR, Vol. 36, No. 5, pp.7-15, October 2006.

[9]. A. Dainotti, A. Pescapé, C. Sansone. Early classification of network traffic through multi-classification. In Proceedings of the Third international conference on Traffic monitoring and analysis (TMA'11), 2011. Springer-Verlag, Berlin, Heidelberg, pp. 122-135.

[10]. Cascarano N, Ciminiera L, Risso F. Optimizing deep packet inspection for high-speed traffic analysis. Network System Manager. 2011 19(1), pp. 7–31.

[11]. S. Kumar and P. Crowley. Algorithms to Accelerate Multiple Regular Expressions Matching for Deep Packet Inspection. In Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM '06), 2006, New York, USA, pp. 339-350.

[12]. D. Ficara, S. Giordano, G. Procissi, F.Vitucci, G.Antichi, A. Di Pietro. An Improved DFA for Fast Regular Expression Matching. SIGCOMM Comput. Commun. Rev. 38, 5 (September 2008), pp. 29-40.

[13]. F. Yu, Z. Chen, Y. Diao, T. V. Lakshman, and R. H. Katz. Fast and Memory-Efficient Regular Expression Matching for Deep Packet Inspection. In Proceedings of the

ACM/IEEE symposium on Architecture for networking and communications systems (ANCS '06). 2006, New York, USA, pp. 93-102.

[14]. S. Kumar, B. Chandrasekaran, J. Turner, and G. Varghese. Curing Regular Expressions Matching Algorithms From Insomnia. In Proceedings of the 3rd ACM/IEEE Symposium on Architecture for networking and communications systems (ANCS '07). 2007,New York, USA, pp. 155-164

[15]. R. Smith, C. Estan, S. Jha, and S. Kong. Deflating the Big Bang: Fast and Scalable Deep Packet Inspection with Extended Finite Automata. In Proceedings of the ACM SIGCOMM conference on Data communication (SIGCOMM '08). 2008, New York, USA, pp. 207-218.

[16]. Cao Z., Cao S., Xiong G., Guo L.Progress in Study of Encrypted Traffic Classification. In Proceedings of International standard conference on trustworthy computing and services, 2012, Beijing, China, pp. 78-86

[17]. M. Sokolova, N. Japkowicz, S. Szpakowicz. Beyond accuracy, f-score and ROC: a family of discriminant measures for performance evaluation //In Proceedings of the 19th Australian joint conference on Artificial Intelligence: advances in Artificial Intelligence (AI'06), Berlin, Heidelberg, 2006, pp. 1015-1021.

[18]. S. Valenti, D. Rossi, A. Dainotti, A. Pescapè, A. Finamore, M. Mellia. Reviewing traffic classification. In DataTraffic Monitoring and Analysis, Springer-Verlag, Berlin, Heidelberg, 2013, pp. 123-147.

[19]. D. Maurizio. Observing routing asymmetry in Internet traffic. https://www.caida.org/research/traffic-analysis/asymmetry, accessed 01.12.2015

[20]. K. Fukuda. Difficulties of identifying application type in backbone traffic, 2010 International Conference on Network and Service Management, Niagara Falls, ON, 2010, pp. 358-361

[21]. H. Balakrishnan and V. Padmanabhan. How network asymmetry affects TCP // IEEE Communications Magazine,Vol. 39, pp. 60 -67, April 2001.

[22]. Applying Network Policy Control to Asymmetric Traffic: Considerations and Solutions. https://www.sandvine.com/downloads/general/whitepapers/applying-network-policy-control-to-asymmetric-traffic.pdf, accessed 01.12.2015

[23]. CAIDAFlowTypes. https://www.caida.org/research/traffic-analysis/flowtypes/, accessed 01.12.2015.

[24]. N. Borisov, D.J. Brumley, H.J. Wang, J. Dunagan, P. Joshi, C. Guo. A Generic Application-Level Protocol Analyzer and Its Language. In Proceedings of 14th Annual Network and Distributed System Security Symposium, 2007.

[25]. CiscoNBAR. http://www.cisco.com/c/en/us/products/ios-nx-os-software/network-based-application-recognition-nbar/index.html, accessed 01.12.2015.

[26]. RFC 2616. Hypertext Transfer Protocol -- HTTP/1.1. https://www.ietf.org/rfc/rfc2616.txt, accessed 01.12.2015.

[27]. RFC 7540. Hypertext Transfer Protocol Version 2 (HTTP/2). https://tools.ietf.org/html/rfc7540, accessed 01.12.2015.

[28]. Administering Cisco QoS in IP Networks. Including CallManager 3.0, QoS, and uOne. 1st Edition, Syngress 2001, eBook ISBN: 9780080481890, pp. 561

[29]. L. Deri, M. Martinelli, T. Bujlow, and A. Cardigliano, "ndpi: Opensource high-speed deep packet inspection," in Wireless Communications and Mobile Computing Conference (IWCMC), 2014 International. IEEE, 2014, pp. 617–622.

[30]. Service Name and Transport Protocol Port Number Registry. http://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml, accessed 01.12.2015

[31]. P. Haffner, S. Sen, O. Spatscheck, D. Wang. ACAS: automated construction of application signatures. In Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data (MineNet '05), ACM, New York, NY, USA, 2005, pp. 197-202.
[32]. Y. Wang, Y. Xiang, W. Zhou, S. Yu. Generating regular expression signatures for network traffic classification in trusted network management, Journal of Network and Computer Applications. Volume 35, Issue 3, May 2012, pp. 992-1000
[33]. G. Szabó, Z.Turányi, L. Toka, S. Molnár, A. Santos. 2011. Automatic protocol signature generation framework for deep packet inspection. In Proceedings of the 5th International ICST Conference on Performance Evaluation Methodologies and Tools, Brussels, Belgium, Belgium, 2011, pp. 291-299.
[34]. Perspective monitoring. http://amonitoring.ru/service/snort/, accessed 01.12.2015.
[35]. G. Bossert, F. Guihéry, G. Hiet. Towards automated protocol reverse engineering using semantic information. In Proceedings of the 9th ACM symposium on Information, computer and communications security (ASIA CCS '14). ACM, New York, NY, USA, 2014, pp. 51-62.
[36]. Get'man A. I., Markin Yu. V., Obydenkov D. O., Padaryan V. A., Tikhonov A. Yu. Methods of presenting the results of network traffic analysis. Trudy ISP RAN/Proc. ISP RAS, vol. 28, issue 6, 2016, pp. 103-110 (in Russian). DOI: 10.15514/ISPRAS-2016-28(6)-7
[37]. O. Mula-Valls. A practical retraining mechanism for network traffic classification in operational environments // Master Thesis Universitat Poliecnica de Catalunya, 2011.
[38]. R. Wang, L. Shi, B. Jennings. Ensemble Classifier for Traffic in Presence of Changing Distributions // In Proceedings of the Symposium on Computers and Communications (ISCC 2013), Split, Croatia, 7-10 July, 2013, pp. 629-635
[39]. J. Zhang, C. Chen, Y. Xiang, .W. Zhou. Robust network traffic identification with unknown applications. In Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security (ASIA CCS '13), 2013, ACM, New York, NY, USA, pp. 405-414.
[40]. R. Wang. Advances in Machine-Learning-Based Traffic Classifiers. https://labs.ripe.net/Members/rwang/advances-in-machine-learning-based-traffic-classifiers, accessed 01.12.2015
[41]. A. White, S. Krishnan, M. Bailey, F. Monrose, P. Porras. Clear and Present Data: Opaque Traffic and its Security Implications for the Future. NDSS, 2013.
[42]. J. Olivain, J. Goubault-Larrecq. Detecting subverted cryptographicprotocols by entropy checking. Technical report, Laboratoire Spcificationet Verification, June 2006.
[43]. L.Bernaille, R. Teixeira. Early recognition of encrypted applications. In Proceedings of the 8th international conference on Passive and active network measurement (PAM'07), 2007, Springer-Verlag, Berlin, Heidelberg, 165-175.
[44]. Global Internet Phenomena Spotlight: Encrypted Internet Traffic. https://www.sandvine.com/downloads/general/global-internet-phenomena/2015/encrypted-internet-traffic.pdf, accessed 01.12.2015
[45]. IP Fragmentation Attacks on Checkpoint Firewalls. https://www.giac.org/paper/gsec/589/ip-fragmentation-attacks-checkpoint-firewalls/101350, accessed 01.12.2015
[46]. M. Handley, V. Paxson, C. Kreibich. Network Intrusion Detection: Evasion, Traffic Normalization, and End-to-End Protocol Semantics. In Proceedings of the 10th conference on USENIX Security Symposium, Vol. 10. USENIX Association, Berkeley, CA, USA, 2001, pp. 9-25.

[47]. M. Baldi, A. Baldini, N. Cascarano, F. Risso. Service-based traffic classification: Principles and validation. In Proceedings of the IEEE Sarnoff Symposium (SARNOFF'09), 2009. IEEE Press, Piscataway, NJ, pp. 115–120.
[48]. W. Moore, K. Papagiannaki. Toward the AccurateIdentification of Network Applications. InternationalWorkshop on Passive and Active Network Measurement (PAM 2005), 2005, Boston MA, USA, vol. 3431, pp. 41-54.
[49]. T. Karagiannis, A. Broido, M. Faloutsos, Kc. Claffy. Transport layer identification of P2P traffic. In Proceedings of 4th ACM SIGCOMM conference on Internet measurement, 2004, pp. 121 – 134.
[50]. QosmosixEngine. http://www.qosmos.com/products/deep-packet-inspection-engine/, accessed 01.12.2015
[51]. Ipoque PACE. https://www.ipoque.com/products/pace, accessed 01.12.2015
[52]. Windriver Content Inspection Engine. http://www.windriver.com/products/product-overviews/PO_Wind-River-Content-Inspection-Engine.pdf, accessed 01.12.2015
[53]. Procera PacketLogic Content Intelligence. https://www.proceranetworks.com/content-intelligence.html, accessed 01.12.2015
[54]. DPI-SSL. https://www.sonicwall.com/ssl-decryption-and-inspection/, accessed 01.12.2015
[55]. G. Aceto, A. Dainotti, W. de Donato, A. Pescap. PortLoad: Taking the Best of Two Worlds in Traffic Classification," in IEEE INFOCOM 2010 – WIP Track, 2010.
[56]. L7-filter. http://l7-filter.sourceforge.net/, accessed 01.12.2015.
[57]. S. Alcock, R. Nelson, Libprotoident: Traffic Classification Using Lightweight Packet Inspection, Technical report, University of Waikato, 2013. http://www.wand.net.nz/publications/lpireport, accessed 01.12.2015
[58]. Wireshark. https://www.wireshark.org/, accessed 01.12.2015.
[59]. T.Karagiannis, K.Papagiannaki, M. Faloutsos. BLINC: multilevel traffic classification in the dark. In Proceedings of the SIGCOMM '05. 2005, ACM, New York, NY, USA, pp.229-240.
[60]. M. Iliofotou, H. Kim, M. Faloutsos, M.Mitzenmacher, P. Pappu, G. Varghese. Graph-based P2P traffic classification at the internet backbone. In Proceedings of the INFOCOM'09. 2009, IEEE Press, Piscataway, NJ, USA, pp. 37-42.
[61]. M.Iliofotou, M. Faloutsos, M.Mitzenmacher. Exploiting dynamicity in graph-based traffic analysis: techniques and applications. In Proceedings of the CoNEXT '09. 2009, ACM, New York, NY, USA, pp. 241-252.
[62]. S. Lee, H. Kim, D. Barman, S. Lee, C. Kim, T. Kwon, Y. Choi. NeTraMark: a network traffic classification benchmark. SIGCOMM Comput. Commun. Rev. 41, 1 (January 2011), pp. 22-30
[63]. A. Dainotti, W. Donato, A.Pescapé. TIE: A Community-Oriented Traffic Classification Platform. In Proceedings of the First International Workshop on Traffic Monitoring and Analysis (TMA '09), 2009, Springer-Verlag, Berlin, Heidelberg, pp. 64-74.
[64]. W. Donato, A. Pescape, A. Dainotti. Traffic identification engine: an open platform for traffic classification. In IEEE Network, vol. 28, no. 2, pp. 56-64, March-April 2014.
[65]. G. Szabo, I. Szabo, D. Orincsay. Accurate Traffic Classification. IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, Espoo, Finland, 2007, pp. 1-8.