На пути к доказуемому измерению и анализу сетевого трафика с помощью полумаркированных наборов данных трассировки.

Исследования по измерению и анализу сетевого трафика критически зависят от реалистичных, актуальных и хорошо документированных данных. Создание таких данных является сложной задачей, так как реальные данные часто недостаточно аннотированы, а искусственно сгенерированные не отражают специфику реальных сетей.

Как итог — ограничение исследовательских и промышленных возможностей. В настоящей работе авторы предлагают использовать наборы данных с полумаркировкой, состоящие из аннотированных трассировок интересующих пакетов. Этот подход всё ещё не обеспечивает универсального решения проблемы, однако такой подход даст больше преимуществ для исследования сетевого трафика.

Наилучший метод создания набора данных зависит от назначения этих данных. Результатом всегда является компромисс между подлинностью и точностью аннотаций. Применяемые наборы данных должны решать следующие четыре задачи, связанные с их созданием и использованием: *анонимизация, аннотация трафика, параметры захвата, актуальность набора данных.* Для смягчения этих проблем авторы предлагают создать набор аннотированных единиц, которые можно объединить с фоновым трафиком реального мира в наборы данных с полумаркировкой. Создание модуля с аннотациями происходит в заранее подготовленной среде.

После создания аннотированные единицы должны быть нормализованы и введены в фоновый трафик. Цель нормализации — гарантировать, что аннотированные единицы могут обрабатываться единообразно в процессе внедрения. Авторы статьи предлагают следующие шаги: *измените MAC-адреса и IP-адреса на значения, которые обычно не встречаются в трафике, сбросьте временные метки пакетов, чтобы первый из них начинался с нулевого времени.*

- *Анонимность* достигается тривиально, поскольку блок содержит только контролируемый неконфиденциальный трафик.
- *Аннотация* точна, поскольку содержимое блоков полностью известно.
- *Параметры захвата* известны и легко изменяются в силу заранее подготовленной среды.
- *Актуальность данных* достигается либо путём объединения с более новым фоновым трафиком для имитации старых атак в новых условиях трафика, либо замены более новыми версиями атак.

При анализе реальной части набора данных с полумаркировкой может произойти ложноположительное обнаружение (FP). Все обнаруженные события из реальной части следует проверять вручную. Если проверка показывает истинно положительный результат (TP), из связанного трафика может быть создан новый блок с аннотациями. В случае обнаружения FP алгоритм обнаружения улучшается для получения правильного результата. Каждый раз, когда метод совершенствуется, он повторно оценивается с использованием исходного набора данных с полумаркировкой. Повторная оценка продолжается до тех пор, пока показатель FP не достигнет желаемого уровня. Таким образом можно протестировать несколько частично помеченных наборов данных до тех пор, пока не будет достигнута желаемая надёжность метода обнаружения.

Обмен информацией и сотрудничество являются ключевым компонентом эффективных исследований. Это всё ещё остаётся проблемой для исследовательского сообщества в данной сфере. Хотя сообщество сотрудничает на конференциях, возможности для улучшения обмена данными всё же существуют. Пользовательские наборы данных становятся общедоступными только в 4% всех случаев. Однако, если общедоступный набор данных уже существует, он используется повторно в 50% всех случаев. Спрос на общедоступные наборы данных явно заметен.

Проблемы сетевых платформ обмена данными можно разделить на две группы. Первая — это проблемы преобразования данных (*анонимизация и нормализация*), необходимого для

обеспечения общего доступа. Вторая группа — это проблемы, связанные с самой платформой обмена данными, такие как соответствие требованиям исследователей, удобство использования и управляемость в течение длительного периода времени. Некоторые из этих проблем уже привели к закрытию нескольких платформ обмена данными.

Авторы статьи предложили решения для проблем, связанных с анонимизацией, неоднородностью данных и устойчивостью платформы, включая разработку веб-платформы централизованного обмена наборами данных с функциями загрузки, поиска, скачивания и управления учётными записями пользователей. Подчёркивается важность унификации для нормализации данных и укрепления доверия к платформе, развития сообщества и предоставления актуального контента для успешного функционирования платформы обмена данными. Авторы также изложили базовую методологию использования полумаркированных наборов данных для оценки результатов исследований и обеспечения возможности взаимного сравнения различных методов анализа. Были охвачены некоторые области, связанные с проблемой вероятности исследования.

Подход, описанный выше, предоставляет комплексное решение проблем доказуемости исследований и обмена данными. Он всё ещё не обеспечивает универсального решения проблемы, однако это положит начало в обсуждении вышеописанной проблемы и даст больше преимуществ для исследования сетевого трафика.