УДК 004.056.53

АЛГОРИТМ ВЕРОЯТНОСТНОЙ ИДЕНТИФИКАЦИИ ПОЛЬЗОВАТЕЛЕЙ СЕТИ

К. А. Тюрин, Н. В. Болдырихин,

Донской государственный технический университет, Ростов-на-Дону, Российская Федерация
kayvflu@gmail.com
boldyrikhin@mail.ru

Рассмотрен новый подход к решению задачи деанонимизации пользователей сети Интернет. Предложен алгоритм сопоставления некоторого субъекта и его анонимного профиля. Алгоритм основан на определении взаимной корреляции трафика анонимного пользователя и идентифицируемого субъекта. Проведены экспериментальные исследования, подтверждающие эффективность алгоритма. Определены факторы, влияющие на качество работы алгоритма.

Ключевые слова: анонимизация, деанонимизация, статистический анализ

UDC 004.056.53

PROBABALISTIC ALGORITHM OF NETWORK USERS IDENTIFICATION

K. A. Tyurin, N. V. Boldyrikhin

Don State Technical University, Rostov-on-Don, Russian Federation
kayvflu@gmail.com
boldyrikhin@mail.ru

The article describes a new approach of the Internet users' deanonymization. It offers the algorithm for comparing a subject and his anonymous profile. The algorithm is based on the definition of anonymous user traffic correlation and identifiable entity. Experimental studies confirming the effectiveness of the algorithm are conducted. Factors affecting the quality of the algorithm are defined.

Keywords: anonymization, deanonymization, statistical analysis

Введение. Современное развитие интернет-технологий и рост числа пользователей влекут за собой возникновение необходимости идентификации субъектов сети при обращении к различным ресурсам [1–4]. В настоящее время существует множество разработок как в направлении установления личности пользователей сети, так и в направлении анонимизации пользователей [1,3,4]. Основу методов идентификации пользователя составляет поиск уникальных для данного пользователя свойств. Первой задачей в этой области является определение IP-адреса пользователя. Тем не менее, существует множество техник сокрытия реального IP-адреса при помощи различных средств анонимизации [3,4]. В таких случаях для идентификации пользователя (или установления связи между двумя профилями пользователей) используется анализ косвенных признаков, таких, например, как «отпечаток браузера», наличие некоторых временных объектов на стороне клиента («Evercookie») и др. [2]. Однако данные методы зачастую работают с недостаточной точностью. Также имеется ряд технических и административных мер, позволяющих снизить точность идентификации до статистически незначимого уровня. Таким образом, существует необходимость в разработке дополнительных алгоритмов идентификации пользователя, работающих с высокой точностью.

Основные термины. Для описания понятия деанонимизации и её основных задач необходимо привести общую модель сетевого взаимодействия и дать определение понятию анонимности.

Профиль субъекта — набор свойств, известных ресурсу, к которому идет обращение. Зачастую целью деанонимизации является не идентификация субъекта, а подтверждение гипотезы о том, что два профиля принадлежат одному субъекту.

Анонимность — состояние неидентифицируемости субъекта во множестве всех возможных субъектов, обладающих схожими свойствами. Будем называть множество таких субъектов множеством анонимности. Исходя из вышеизложенного определения следует, что устойчивость анонимности находится в прямой зависимости от мощности множества анонимности, в котором находится субъект. Таким образом, анонимность можно определить как степень распространенности совокупности свойств, которыми обладает субъект (например, распространенный «отпечаток браузера»), а также степень их несвязности с объектом (например, не принадлежащий субъекту IP-адрес).

Анонимизация — процесс обеспечения анонимности. Анонимизация чаще всего заключается в изменении свойств сетевого профиля субъекта с целью обеспечения их несвязности с пользователем, либо же с целью использования более распространенных значений некоторых свойств.

Общее устройство систем анонимизации представлено на рис. 1.

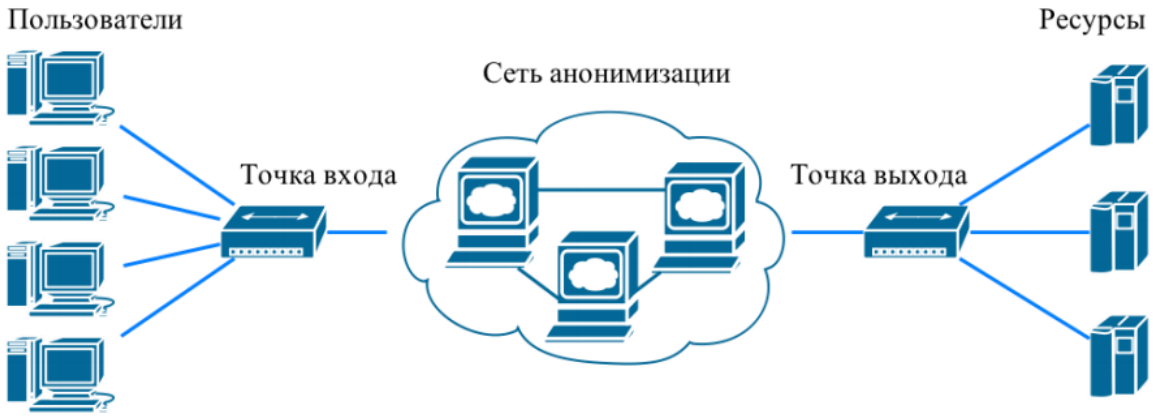

Рис.1. Общая схема систем анонимизации

Деанонимизация — процесс, обратный анонимизации, то есть процесс сопоставления субъекта и его профиля в сети Интернет, либо процесс установления принадлежности двух профилей одному субъекту.

Постановка задачи. В рамках данной статьи предлагается алгоритм деанонимизации на основе анализа статистических характеристик передаваемого трафика. В частности, такой характеристикой является объем передаваемых данных в единицу времени. В течение сеанса связи скорость передаваемых данных от ресурса к клиенту (и обратно) изменяется, образуя функцию, вероятно, уникальную для данного соединения среди множества всех соединений в сети.

Предполагаемый злоумышленник является пользователем сети и обращается к одному из целевых ресурсов. Для использования рассматриваемого в данной статье алгоритма деанонимизации необходимо контролировать канал передачи между злоумышленником и точкой входа (или саму точку входа) и канал передачи между точкой выхода и ресурсом (возможно, саму точку выхода). Такой контроль может быть осуществлен при помощи программных агентов или

средств перехвата и анализа передаваемого трафика, например, СОРМ (система оперативно-розыскных мероприятий).

Предположим, что имеется возможность проводить наблюдение трафика, приходящего к ресурсу и исходящего от абонента. Таким образом, наблюдения представляются двумя наборами данных.

Далее приведена серия экспериментов, в рамках которой была произведена запись статистических характеристик трафика нескольких клиентов при их обращении к различным ресурсам. В частности, проверяется корреляция между трафиком, передаваемым на участках пользователь/прокси-сервер и прокси-сервер/ресурс. В рамках эксперимента каждый пользователь обращался к некоторому закрепленному за ним ресурсу. При этом каждое соединение содержало свои особенности передачи. После проведения эксперимента для анализа сходства двух наборов данных использовалась корреляция Пирсона, рассчитываемая по формуле:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\operatorname{cov}(x, y)}{\sqrt{s_x^2 s_y^2}},$$

где $x^n = (x_1, \ldots, x_n)$ и $y^n = (y_1, \ldots, y_n)$ — две выборки, $\bar{x}$ и $\bar{y}$ — выборочные средние x^n и y^n, а s_x^2 и s_y^2 — выборочные дисперсии, $r_{xy} \in [-1;1]$.

На рисунках ниже изображено значение коэффициента корреляции Пирсона для каждой пары наборов наблюдений. Так, каждый рисунок относится к соответствующему ресурсу и иллюстрирует значения схожести трафика, передаваемого между ресурсом и прокси-сервером, с тремя наборами наблюдений (для каждого пользователя соответственно).

Первый эксперимент — соединение с обычным прокси-сервером для проверки работоспособности алгоритма.

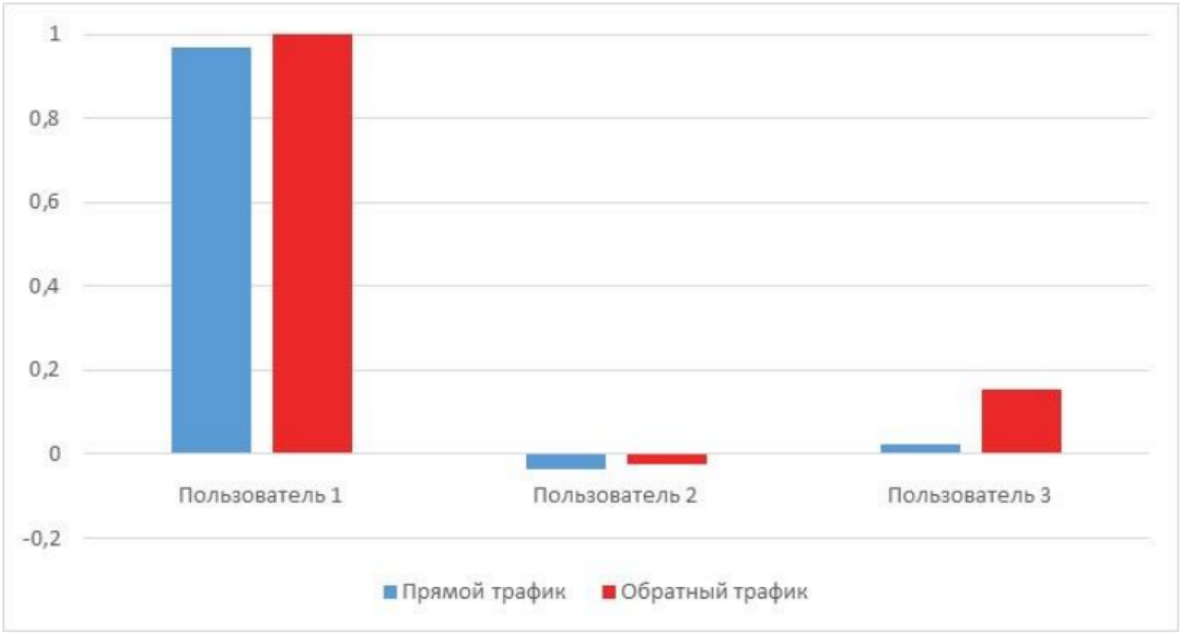

Рис. 2. Взаимодействие пользователей с первым ресурсом

Как видно на графике, анализ показывает устойчивое сходство между статистическими характеристиками трафика на участках до и после прокси-сервера для пользователя 1. При этом происходит полное несовпадение для других пользователей. Этот эффект объясняется отсутствием внешних факторов на соединение.

Известно, что протокол TCP (Transmission Control Protocol) подстраивается под пропускную способность канала со временем. Следовательно, трафик должен быть симметричным даже в том случае, если по пути трафика попадается ограничение пропускной способности. Тем не менее, такого не произойдет, если трафик буферизуется на одном из промежуточных узлов. То же самое происходит в том случае, если используется кэширующий прокси-сервер.

При использовании кэширования данные передаются только на участке прокси-сервер/клиент, но не передаются на участке ресурс/прокси-сервер, так как прокси-сервер не делает запросы для передачи некоторых заранее сохраненных данных. Как правило, кэшируются лишь данные большого объема, например, изображения или исполняемые скрипты. При этом разметка и текст, как правило, на прокси-сервере не сохраняются. Это значит, что трафик на участке у выходного узла будет присутствовать, но меньший, чем у входного узла, что обуславливает наличие корреляции, хоть и меньшей, чем в прошлом случае.

Взаимодействие пользователя 3 с кэширующим прокси проиллюстрировано на рис. 3.

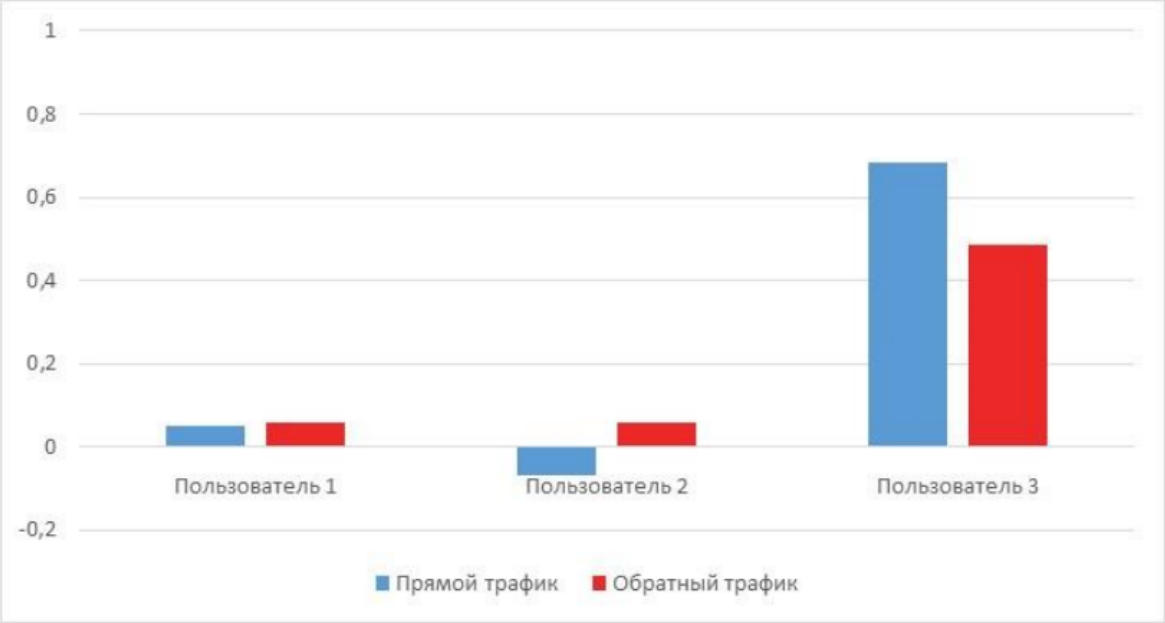

Рис. 3. Взаимодействие клиента с кэширующим прокси

В реальных условиях пользователи могут также обращаться и к другим ресурсам. В случае, если в процессе проксирования не используется мультеплексирование каналов и/или шифрование, то становится возможным фильтровать трафик на основе информации о соединениях. Это сводит случай к уже рассмотренным. Поэтому рассмотрим ситуацию, когда несколько соединений к прокси-серверу мультиплексируются в одно.

На рис. 4 изображен график обращения пользователя 2 к некоторому целевому ресурсу с параллельным подключением к сторонним ресурсам.

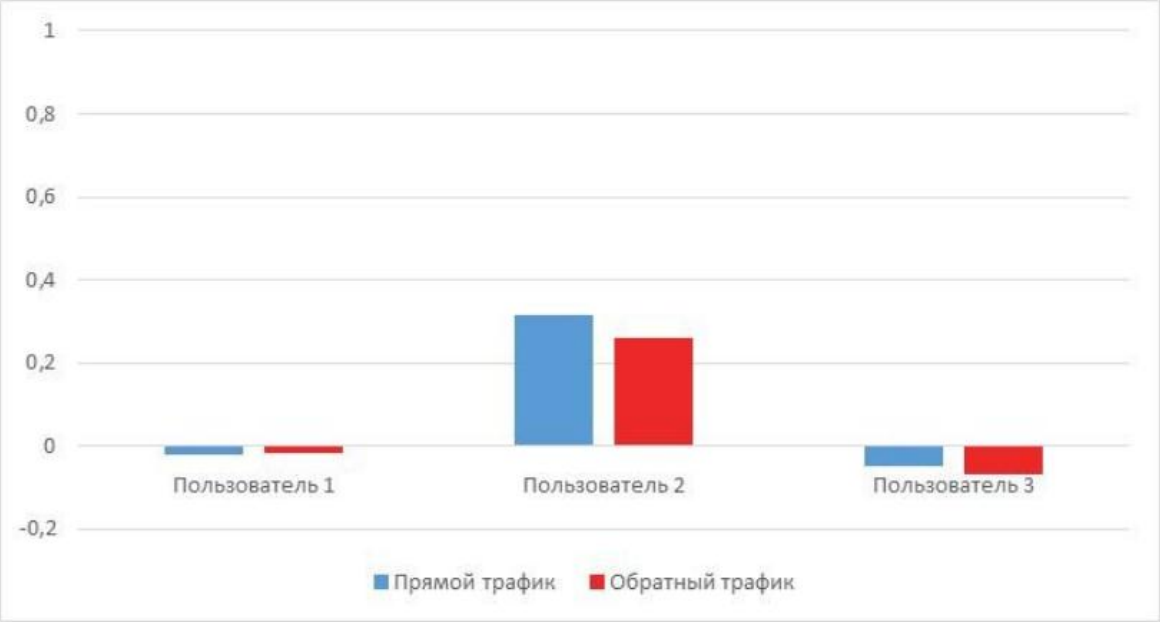

Рис. 4. Взаимодействие клиента при обращении к другим ресурсам

Обращение к другим ресурсам сильно портит статистическую значимость получаемых результатов. Анализ показывает, что пользователь 2 имеет большую вероятность взаимодействия со вторым ресурсом, чем остальные пользователи, однако такие результаты не подходят для использования исследуемого алгоритма в реальных ситуациях.

Заключение. Таким образом, проанализированы существующие методы деанонимизации и предложен алгоритм на основе анализа статистических характеристик трафика. Предложенный алгоритм реализует новый подход к решению задачи деанонимизации. Проведены экспериментальные исследования по анализу характеристик трафика в разных условиях: при обычном соединении, соединении с кэширующим прокси-сервером и параллельном обращении к сторонним ресурсам. Результаты экспериментальных исследований показывают работоспособность алгоритма, однако требуется его доработка для ситуаций, которые могут встречаться в реальных условиях.

В качестве улучшения алгоритмов нахождения сходства между двумя наборами данных могут использоваться: корреляции Спирмена и Кендалла [5], корреляции случайных процессов, а также методы обнаружения сигналов на фоне помех [6].

Библиографический список

1. CAI, X. Systematic Approach to Developing and Evaluating Website Fingerprinting Defenses / CAI, X. [и др.] // In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. — New York, 2014. — С. 227–238.

2. HTTPOS: Sealing Information Leaks with Browser-side Obfuscation of Encrypted Flows / Free Haven. — Режим доступа : http://freehaven.net/anonbib/cache/LZCLCP_NDSS11.pdf (дата обращения 04.04.2016).

3. Anonymity online / Официальный ресурс системы анонимизации Tor. — Режим доступа: https://www.torproject.org/ (дата обращения 05.04.2016).

4. The invisible Internet Project / Официальный ресурс системы анонимизации I2P. — Режим доступа: https://geti2p.net/ (дата обращения 05.04.2016).

5. Гмурман, В. Е. Теория вероятностей и математическая статистика: Учебное пособие для вузов / В. Е. Гмурман. — Москва : Высшая школа, 2004. — 479 с. 6.

6. Свешников, А. А. Прикладные методы теории случайных функций / А. А. Свешников. — Москва : Наука, 1968. — 463 с.