ЭНТРОПИЯ КАК ИНДИКАТОР ВОЗНИКНОВЕНИЯ АНОМАЛИЙ СЕТЕВОГО ТРАФИКА

Борисов Д.Н.
Донецкий национальный технический университет, г. Донецк
кафедра автоматики и телекоммуникаций
E-mail: dborisov@ints.net

Abstract

Borisov D.N. Entropy as an indicator of rising traffic anomaly. *The questions of traffic anomaly detection by are discussed. Network traffic as an research object is analized. Mathematic model based on entropy of network traffic packet headers for traffic anomaly detection is proposed.*

Описание предметной области. Глобальная сеть передачи данных Интернет заняла прочное место в современной жизне. Без этого средства связи невозможно представить себе современные телекоммуникации. Такие средства, казавшиеся экзотикой еще несколько лет назад, как видеоконференции, обмен мгновенными сообщениями, дешевые (относительно классических) средства междугородней и международной голосовой связи, прочно вошли в повседневную жизнь. Следует отметить все большую конвергенцию сетей: уже сейчас многие национальные и транснациональные операторы связи осуществляют передачу международного трафика через каналы глобальной сети. Сети с коммутацией пакетов постепенно вытесняют сети с коммутацией каналов, а самой большой такой сетью и является Интернет (или просто Сеть).

В основу сети Интернет при разработке были положены принципы децентрализации и самоуправляемости. Это связано с тем, что сеть разрабатывалась военным ведомством США как средство связи на случай ядерной войны. Однако, аналитики не предполагали повсеместного распространения сети. Согласно некоторым исследованиям [1] количество пользователей сети Internet в мире на 2005 год составило 167 на 1000 человек населения Земли. Аналогичная статистика на 2000 год составила 69 пользователя на 1000 человек, а к 2010 году эта цифра должна увеличиться до 262 человек. Сейчас стали проявляться негативные аспекты такого подхода. Одна из проблем, связанных с таким подходом и рассматривается в этой работе.

DDoS атаки как объект исследования. Объектом исследования данной работы являются возникающие в работе Сети аномалии трафика. К ним можно отнести внезапные всплески активности пользователей (примером служит перегрузка основных магистральных каналов и информационных сайтов 11-12 сентября 2001 года), атаки хакеров на сервера гигантов IT-индустрии (таких как Sun Microsistems, Microsoft и др). Сущность этих возмущений состоит в изменении структуры трафика таким образом, что бы занять максимальное количество ресурсов серверов, обрабатывающих запросы пользователей. Такой тип атаки называют «отказом в обслуживании» (DoS – Denial of Service). Рассматриваемый в этой работе вариант такой атаки основывается на принципе, который получил название «распределенной атаки типа отказ в обслуживании» (DDoS – Distributed Denial of Service). Во время такой атаки множество хостов-агентов из Сети генерируют трафик, нацеленный на хосты-жертвы. Если ресурсы, необходимые для обработки такого объема трафика превышают ресурсы хоста-жертвы, происходит отказ в работе жертвы, что и является целью атаки. Такая атака может быть вызвана как действиями легальных

пользователей, так и намеренно причиняющих вред.

Определимся с основными терминами:

DDoS атака – это такой вид вредоносной сетевой активности, при котором все ресурсы жертвы поглощаются атакующими хостами и легальные пользователи не могут получить предоставляемые хостом-жертвой услуги.

Сетевой трафик, сетевой поток – это множество сетевых пакетов, проходящих через канал передачи данных и обладающих каким-то общим признаком.

Канал передачи данных – физическая среда, предназначенная для передачи информации, путем кодирования её электромагнитными импульсами. Для этой работы существенно, что канал передачи данных обладает в том числе следующими характеристиками: максимальная полоса пропускания канала передачи данных; используемая полоса пропускания канала передачи данных; количество пакетов в секунду, проходящее через канал передачи данных. Мы выбрали именно эти характеристики, поскольку именно они влияют на способность предоставлять услуги пользователям Сети – способность маршрутизатора обрабатывать пакеты.

Маршрутизатор – устройство, обрабатывающее (принимающее и передающее в канал передачи данных) сетевые пакеты. Собирает и передает на обработку информацию о сетевых потоках. Наиболее уязвимое место сети при DDoS атаке. Следует отметить, что в данном случае определение маршрутизатора несколько отличается от общепринятого.

Хост-жертва – хост в Сети, являющийся целью атаки.

Исходя из всего вышеописанного, сформулируем объект и предмет исследования. Объектом исследования будут каналы передачи данных и сетевой трафик в них.

Предметы исследования: общая энтропия заголовков сетевых пакетов как индикатор возникновения аномалий сетевого трафика; выявление потоков, которые могут привести к отказу в работе маршрутизатора или его клиентов (“вредоносных”).

Основными задачами данного исследования являются:

- идентификация DDoS;
- определение признаков, характерных для потоков DDoS;
- идентификация потоков трафика, вызывающих перегрузку хостов-жертв.

Для решения поставленных задач предлагается использовать общее значение энтропии заголовков сетевых пакетов для определения степени “нормальности” сетевого трафика в канале передачи данных. В случае обнаружения аномалии сетевого трафика, выделять пакеты, дестабилизирующие ситуацию в канале передачи данных и строить правила фильтрации таких пакетов.

Актуальность проблемы. Возраст большинства научных разработок по этой теме невелик – 1-3 года, так как ранее эта проблема не имела практической ценности. В данный же момент большинство производителей сетевого оборудования и поставщики крупных сетевых решений ищут эффективные пути решения проблем, связанных с DDoS. С этим и связан взрывной рост количества работ и публикаций, посвященных этой теме. Наиболее полно данная проблема освещается в книге «Internet Denial of Service: Attack and Defense Mechanisms» (Отказ в обслуживании в Интернет: механизмы атаки и защиты) [2]. Авторы этой книги активно разрабатывают направление сетевой безопасности, связанное с DDoS. Так, перу Елены Мирковик и Питера Райхера принадлежит работа по систематизации механизмов атаки и защиты, рассмотренная в статье «A Taxonomy of DDoS Attack and DDoS Defense Mechanisms» (Систематика механизмов распределенных атак типа «отказ в обслуживании» и механизмов защиты) [3]. Эта систематика наиболее комплексно освещает

проблемы организации и реализации таких атак. Ведущие мировые поставщики сетевого оборудования, консалтинговые фирмы, специализирующиеся на IT, поставщики сетевого ПО, специализирующиеся на сетевой безопасности, предлагают решения, снижающие риск и ущерб от таких атак. Однако, до сих пор не предложено решения со 100%-й вероятностью идентифицирующего данный тип сетевой активности. В этой работе предлагается один из возможных методов, повышающих вероятность обнаружения такой атаки.

Методология исследования. В качестве исходных данных в данном исследовании будем использовать информацию, предоставляемую протоколом NetFlow v.5

Рассмотрим поля IP пакета, которые мы можем получить из предоставляемых протоколом NetFlow v.5 данных, их назначение и теоретическое влияние на общую энтропию заголовка пакета (табл.1).

Таблица 1: Исходная информация, предоставляемая протоколом NetFlow версии 5

№,i	*Байты*	*Поле (F)*	*Описание*
1	0-3	srcaddr	Исходный IP адрес.
2	4-7	dstaddr	IP адрес назначения.
3	8-11	nexthop	Адрес следующего шлюза для потока..
4	12-15	input and output	SNMP индекс входного и исходящего интерфейсов.
5	16-19	dPkts	Число пакетов в потоке.
6	20-23	dOctets	Число байт, относящихся к 3 уровню модели OSI во всех пакетах потока.
7	24-27	First	Значение SysUptime на момент появления потока.
8	28-31	Last	Значение SysUptime на момент прихода последнего пакета потока.
9	32-35	srcport and dstport	TCP/UDP исходный порт и порт назначения либо их аналог (код ICMP сообщения, например).
10	36-39	pad1, tcp_flags, prot, and tos	Неиспользуемое поле, кумулятивное «ИЛИ» TCP флагов, № IP протокола, и IP type-of-service.
11	40-43	src_as and dst_as	Исходная AS и AS назначения, либо источника, либо полученная от соседа по BGP
12	44-47	src_mask, dst_mask, and pad2	Длинна маски подсети источника и адресата, 2 неиспользуемых байта.

F_1 - Исходный адрес и адрес назначения пакета. Как описано в [2] все множество адресуемых хостов в Сети обладает разной популярностью. Назовем популярностью частоту, с которой к этому хосту обращаются другие хосты, а так же частоту его ответов. В случае нормальной работы сети, каждый хост обладает некоторой относительно постоянной популярностью. Это связано с пристрастиями пользователей к определенным WEB-ресурсам, относительным постоянством потоков электронной почты, новостей, мультимедиа информации и других наиболее востребованных ресурсов Сети. Таким образом, появления множества пакетов, не являющихся характерными для данного канала передачи данных является весомым аргументом в определении “нормальности” сетевого трафика. Однако, следует учитывать, что появление таких пакетов абсолютно не обязательно является DDoS атакой, а может быть вызвано вполне легальными причинами, такими как: появление новых поставщиков, клиентов, переносом популярных ресурсов, изменением конфигурации

сетевых сервисов и пр. Это поле является качественной случайной величиной – каждый адрес является отдельной сущностью, количество которых ограничено, и каждая обладает определенной вероятностью возникновения в канале передачи данных.

F_2 - Адрес следующего шлюза на пути пакета не является значимой информацией, поскольку не несет информации о природе и свойствах потока, а отражает ситуацию в таблице маршрутизации.

F_3 - SNMP индекс входного и исходящего интерфейсов. Ситуация такая же, как и с адресом следующего шлюза.

F_4 - Число пакетов в потоке является значимой информацией, поскольку вкупе с полями 7-8 позволяет узнать часть, вносимую в общее число пакетов в секунду в канале передачи данных. Эта величина отражает ресурсы маршрутизатора, которые тратятся на обработку потока. Чем выше это значение, тем больше вероятность отказа в работе маршрутизатора именно из-за этого потока. Это поле можно считать непрерывной случайной величиной, поскольку это более соответствует её природе.

F_5 - Число байт, относящихся к 3 уровню модели OSI во всех пакетах потока является значимой информацией поскольку отражает свойства потока. Это поле можно считать непрерывной величиной, поскольку это более соответствует её природе.

F_6 - Значение SysUptime на момент появления потока. Используется как вспомогательное поле для определения значимых характеристик анализируемого потока. Так же непрерывная величина, по сути - время.

F_7 - Значение SysUptime на момент прихода последнего пакета потока. Используется как вспомогательное поле для определения значимых характеристик анализируемого потока. Так же непрерывная величина, по сути - время.

F_8 - TCP/UDP исходный порт и порт назначения либо их аналог (код ICMP сообщения, например). Большинство сервисов, предоставляемых Сетью, использует точно определенные порты для обмена данными как минимум с одной стороны. Так, HTTP использует для обмена информацией 80-й порт протокола TCP, активное FTP-соединение – 20-й и 21-й порты. Однако, более молодые сервисы используют либо случайные порты (как, например, ICQ), либо порты в некотором достаточно большом диапазоне, например пассивный FTP использует для передачи данных порты из верхнего диапазона всех доступных (обычно, с номером больше 40000). То есть, это поле характеризует поток с определенной стороны. Поле считается качественной характеристикой, поскольку несет информацию о сервисе, генерирующим поток данных.

F_9 - Кумулятивное «ИЛИ» TCP флагов. По этому параметру можно определить статус и историю сетевого соединения. Известно, что для инициирования соединения используются флаги SYN и ACK TCP заголовка. Для завершения - FIN и ACK, при перезапуске соединения – RST и ACK [RFC793]. То есть, это поле характеризует поток с определенной стороны. Качественная случайная величина, поскольку несет информацию о состоянии сетевого потока.

F_{10} - № IP протокола. Позволяет рассчитать распределение трафика по протоколам. При DDoS эта характеристика должна резко измениться, поскольку обычно распределение достаточно стабильно. Однако, к резкому изменению могут привести и иные факторы, описанные в предыдущих пунктах. То есть, это поле характеризует поток с определенной стороны. Так же качественная случайная величина, поскольку отражает тип протокола, используемого потоком.

F_{11} - IP type-of-service. Используется для определения приоритета обслуживания

пакетов. Система расстановки приоритетов построена таким образом, что высший приоритет получают пакеты управления сетью, которых значительно меньше, чем пакетов данных, имеющих низший приоритет. Резкое изменение этого соотношения может быть вызвано определенными типами DDoS атак, так же как и сугубо легальными техническими причинами. То есть, это поле характеризует поток с определенной стороны. Так же качественная случайная величина, несущая информацию о качестве обслуживания.

F_{12} - Исходная AS и AS назначения, либо источника, либо полученная от соседа по BGP. Ситуация та же, как и с адресом следующего шлюза.

F_{13} - Длинна маски подсети источника и адресата. Не будем учитывать эту информацию, поскольку она определяет значимую для процесса маршрутизации часть IP-адреса источника и назначения, а непосредственной информации о природе пакета не несет.

Описанные выше данные несут полную информацию о сетевом трафике в канале передачи данных, обслуживаемом маршрутизатором.

Общий смысл подхода к выявлению степени «нормальности» сетевых потоков изложен ниже и основывается на применении понятия энтропии к информации о сетевых потоках, обрабатываемых маршрутизатором.

Согласно теории информации, энтропия “служит мерой неопределённости сообщений данного источника (сообщения описываются множеством величин x1, x2,..., xn, которые могут быть, например, словами какого-либо языка, и соответствующих вероятностей p1, p2,..., pn появления величин x1, x2,..., xn в сообщении). В качестве такой "меры неопределённости" в теории информации принимается число двоичных знаков, необходимое для фиксирования (записи) произвольного сообщения данного источника. Для определённого (дискретного) статистического распределения вероятностей информационной Энтропией называют величину не меньшую, чем

$$H = -\sum_i p_i \log_2 p_i$$

и равную среднему числу двоичных знаков, необходимых для записи сообщений.” [4].

Предлагаемый метод базируется на предположении, что энтропия параметров – случайная величина, при чем её плотность распределения подчиняется какому-то закону. А её резкое изменение можно рассматривать как аномалию, в большинстве случаев вызванную аномалией трафика.

Промежуток времени, на основе данных которого будут строиться измеряемые параметры, предполагается определить эмпирически. Однако, этот промежуток времени должен предоставлять полную информацию о состоянии канала передачи данных. Поэтому определим границы этого промежутка.

Минимальное время, которое проходит между экспортом данных об одном и том же потоке на маршрутизаторах Cisco по умолчанию равняется 30 секундам. Логично принять этот промежуток времени, как минимальный возможный, поскольку за 30 секунд маршрутизатор экспортирует данные обо всех сетевых потоках, проходящих через него. По описанным выше причинам, любой взятый промежуток времени для анализа должен быть кратен 30 секундам.

Максимальный промежуток времени, за который целесообразно обрабатывать информацию, определим исходя из практических соображений: DDoS атака обнаруживается в ручном режиме (в зависимости от квалификации персонала) не более чем за пол часа. Это время включает в себя время реакции на снижение качества обслуживания или отказ в работе канального оборудование, коммуникации между пострадавшим и персоналом провайдера, и время, необходимое на локализацию и устранение причины неполадки. Таким образом, система обнаружения и локализации DDoS атаки будет полезна, если поможет

снизить время обнаружения и реакции. Поэтому установим верхнее ограничение промежутка анализа трафика в 30 минут.

Следует отметить, что чем меньший промежуток времени анализируется, тем больше будет разрыв в параметрах между соседними промежутками. Практика показывает, что среднее время жизни одного потока не превышает 30 секунд. На существующих полосах пропускания каналов передачи данных основные всплески сглаживаются очередями. Но стандартная длинна очереди на канале передачи данных в 1Мбит/сек такова, что трафик сглаживается на промежутке времени менее 1 секунды (не более 1000 пакетов, не более 25600 байт).

Для определения энтропии воздействия потоков в канале на маршрутизатор нам необходимы сообщения и вероятности их возникновения в канале. Будем считать, что кроме существующих сообщений, никаких других в канале передачи данных быть не может. В качестве сообщения возьмем информацию о потоке. В качестве вероятности сообщения будем считать количество ресурсов маршрутизатора, используемых потоком, деленных на общее количество используемых маршрутизатором ресурсов:

$$p_i = \frac{F_{5i} / (F_{8i} - F_{7i})}{\sum_k (F_{5k} / (F_{8k} - F_{7k}))}$$

где p_i - вероятность сообщения;

F_{5i} - число пакетов в потоке;

F_{8i} - время прихода последнего учтенного пакета в потоке;

F_{7i} - время прихода первого учтенного пакета в потоке.

На основе этих данных вычисляется энтропия сетевого потока в канале передачи данных:

$$H = -\sum_i p_i \log_2 (p_i)$$

где H - значение энтропии.

Примем энтропию канала передачи данных за случайную величину. На основе реальных данных выберем закон распределения. Предположим, что она подчиняется выбранному закону распределения. В случае резкого её возмущения, определим, какие именно поля вносят наибольшее возмущающее воздействие.

Одним из возможных методов определения потоков, вносящих возмущение, может быть вычисление энтропии по разному количеству полей. Каждое поле, включая дополнительное – случайная величина с определенной вероятностью возникновения.

Дополнительным полем, участвующим в вычислении энтропии, будем считать часть ширины пропускания канала, используемая потоком, деленное на всю используемую полосу пропускания канала:

$$V_i = \frac{F_{6i} / (F_{8i} - F_{7i})}{\sum_k (F_{6k} / (F_{8k} - F_{7k}))}$$

где V_i - часть полосы пропускания, занимаемая потоком;

F_{6i} - число пакетов в потоке,

F_{8i} - время прихода последнего учтенного пакета в потоке,

F_{7i} - время прихода первого учтенного пакета в потоке.

Полями в этом случае являются изначальные значения полей (за исключением полей 5-8 и неиспользуемых полей). За сообщение примем составную случайную величину, вероятность возникновения которой рассчитывается как произведение вероятностей

значений каждого составляющего её поля. За вероятность примем суммарные ресурсы маршрутизатора, используемые потоками с равными значением поля деленные на общие занятые ресурсы маршрутизатора. Рассчитаем энтропию для каждой комбинации полей. Опять примем каждое вычисленное значение энтропии за случайную величину. Наибольший возмущающий фактор вносят те поля, энтропия комбинации которых максимально возмущена.

Еще один предлагаемый метод состоит в хранении сигнатур законов распределения значений энтропии в момент атак. В текущее время, в связи со снижением квалификации интернет-нарушителей (назовем их «хакерами»), большинство DDoS атак производится дилетантами с низким уровнем подготовки. Для этого ими используется доступное в Сети ПО. Атаки, генерируемые одним и тем же ПО аналогичны друг другу по структуре трафика. Таким образом, получив величину энтропии, значительно отличающуюся от нормальной, но подходящей под одну из сигнатур можно с большей уверенностью говорить о наличии вредоносного трафика в канале передачи данных.

Выводы

В этой статье нами был предложен метод выявления аномалий сетевого трафика на основе информации об упорядоченности трафика в канале передачи данных. Основная идея этого метода (идея об относительной стационарности параметров сетевого трафика) была почерпнута из работы [5]. Однако, автор этого труда акцентируется на корпоративных сетях и в качестве основы для анализа сетевых потоков использует базу знаний о поведении сетевых потоков. На наш взгляд, этот подход слишком ресурсоемок для применения в магистральных сетях передачи данных, поскольку требует хранения больших объемов информации и больших объемов вычислений. Поэтому была найдена такая характеристика сетевых потоков, которая не требует для вычисления больших объемов вычислений и предварительного накопления больших объемов информации. Однако, полностью отказаться от хранения данных не удастся, поскольку сети передачи данных являются динамически изменяемыми объектами и без постоянного обучения предложенные методы не будет работать.

Литература

1. "Europe #1 in Per Capita Cell Phone Usage", 2006, http://www.c-i-a.com/pr0206.htm.

2. Jelena Mirkovic, Sven Dietrich, David Dittrich, Peter Reiher;"Internet Denial of Service: Attack and Defense Mechanisms",Prentice Hall PTR,2004,ISBN: 0131475738.

3. Jelena Mirkovic, Peter Reiher "A Taxonomy of DDoS Attack and DDoS Defense Mechanisms",2004.

4. Яглом А. М., Яглом И. М., Вероятность и информация, 2 изд., М., 1960.

5. Seong Soo Kim, «Real-time analysis of aggregate network traffic for anomaly detection», 2005.