УДК 004.054

ПРОБЛЕМА ИСПОЛЬЗОВАНИЯ ЭНТРОПИИ КАК СПОСОБА ОПРЕДЕЛЕНИЯ ДАННЫХ, ПРЕДСТАВЛЯЮЩИХ ИНТЕРЕС ДЛЯ РЕАЛИЗАЦИИ ЗАДАЧ ПРАВООХРАНИТЕЛЬНЫХ ОРГАНОВ

О. А. Слащинин

Главное следственное управление Следственного комитета Республики Беларусь, Беларусь, Минск, mailtempus@mail.ru

Рассмотрена проблема измерения информационной энтропии как одного из способов определения цифровых данных, которые могут представлять интерес для реализации задач правоохранительных органов. Проанализированные в работе положения и сформулированные выводы направлены на совершенствование алгоритма исследования компьютерной информации в рамках осуществления оперативно-розыскной деятельности и предварительного расследования, а также на систематизацию подготовки для указанных органов кадров-специалистов в области форензики.

Ключевые слова: вредоносные программы; компьютерная безопасность; форензика; шифрование; энтропия.

THE PROBLEM OF USING ENTROPY AS A METHOD FOR DEFINING DIGITAL DATA FOR THE IMPLEMENTATION OF THE TASKS OF LAW ENFORCEMENT AGENCIES

O. A. Slashchynin

The Main Investigative Department of the Investigative Committee of the Republic of Belarus, Belarus, Minsk, mailtempus@mail.ru

The problem of measuring information entropy as one of the ways to determine digital data that may be of interest for the implementation of the tasks of law enforcement agencies is considered. Research highlights are aimed at improving the algorithm for the examine of computer information in the framework of operational-search activities and preliminary investigation, as well as at systematizing the training of specialists in the field of digital forensics for law enforcement agencies.

Keywords: computer security; digital forensics; encryption; entropy; malware.

Введение

Сотрудникам правоохранительных органов при исследовании компьютерной информации необходимо обладать специальными знаниями о

методике поиска зашифрованных (упакованных) файлов, следов стеганографии, вредоносных компонентов программ и иных видов цифровых данных. В связи с особенностями распределения и последовательности байт в цифровых данных различных форматов усматривается возможность их идентификации среди всего исследуемого массива информации. Например, алгоритмы шифрования, используемые для маскирования вредоносных компонентов программ, а также сокрытия значимой информации, преобразуют входные данные в совокупность байтов со свойствами равномерно распределенной случайной последовательности. Так, настоящая совокупность может устанавливаться в том числе путем измерения информационной энтропии исследуемых объектов.

Указанный способ идентификации конкретных цифровых данных также может использоваться для иных задач информационной безопасности: определения границ файлов или их фрагментов, пакетного анализа захваченного сетевого трафика, оценки криптостойкости применяемых паролей и алгоритмов шифрования. Применительно к задачам правоохранительных органов измерение информационной энтропии может использоваться для поиска на исследуемых электронных накопителях скрываемых криптоконтейнеров и упакованных файлов, зашифрованных виртуальных машин, файлов кошельков криптовалют (цифровых знаков и токенов), вредоносных компонентов программ.

Однако, измерение энтропии информации как универсальный способ идентификации отдельных видов цифровых данных имеет явную проблему практического применения. Анализ указанной проблемы и предложение возможных путей ее решения позволит усовершенствовать используемый алгоритм исследования компьютерной информации как в целях осуществления правоохранительной деятельности, так и для систематизирования подготовки кадров-специалистов в области компьютерной криминалистики (форензики).

Измерение информационной энтропии (по К. Шеннону)

Понятие энтропии как меры хаотичности (неопределённости) широко используется в естественных и точных науках. В теории информации энтропия может условно интерпретироваться как мера неопределенности отдельной системы, имеющей различные исходы наступления того или иного состояния (непредсказуемость появления одного из символов первичного алфавита). Клод Элвуд Шеннон, сформулировав свойства информационной энтропии и задав требования к её измерению [1], ввёл следующую формулу (1):

$$H = -K \sum_{i=1}^{n} p_i * log_2 p_i \quad (1)$$

или

$$H(x) = \sum_{i=1}^{n} p(x_i) * log_b(\frac{1}{p(x_i)}),$$

где K – положительная константа, n – количество возможных вариаций символьных значений первичного алфавита (состояний), b – единица измерения (бит, нат, трит или хартли), x – случайная величина с диапазоном изменчивости, p_i – вероятность наступления i-состояния.

Применительно к двоичной системе счисления, где бит как единица измерения информации может принимать только два значения (0 или 1), а количество возможных символов входных данных, выраженных в байтах (1 байт = 8 бит), будет равняться 256 (i = 0…255), для вычисления энтропии цифровых данных, соответственно, будет применяться формула (2):

$$H(x) = \sum_{i=0}^{255} p(x_i) * log_2 \left(\frac{1}{p(x_i)}\right) \quad (2)$$

На основании приведенной формулы следует указать, что значение энтропии (x) не может быть отрицательным, а максимальный уровень из-за размера первичного алфавита в 1 байт будет равен 8 ($2^8 = 256$). При любом изменении вероятности (p) наступления i-состояния, соответственно, будет изменяться и значение (x) энтропии. В случае, если состояние системы заранее и достоверно известно, то теоретически ее энтропия близка к 0, а при максимальной степени неопределенности равна 8.

Стоит упомянуть о зависимости уровня энтропии от: 1) размера входных данных (n) и 2) распределения значений байт в этих данных. Соответственно, энтропия цифровых данных небольшого размера будет отличаться от энтропии данных такого же формата, но большего размера. Это связано с тем, что у данных небольшого размера, соответственно, меньшее количество описывающих их бит, чем в данных большего размера, где такую зависимость установить сложнее. Таким образом, значение энтропии данных небольшого размера не будет достигать максимальных величин. Настоящую зависимость легче оценить в результате преобразования формулы (2) при условии $p(i) = \text{кол}\ (i)/n$, а именно (3):

$$H - log_2 n - \frac{1}{n}\sum_{i=0}^{255} \text{кол}(i) * log_2 \text{кол}(i) = log_2 n - \sum_{i=0}^{255} p(i) * log_2 \text{кол}(i), \quad (3)$$

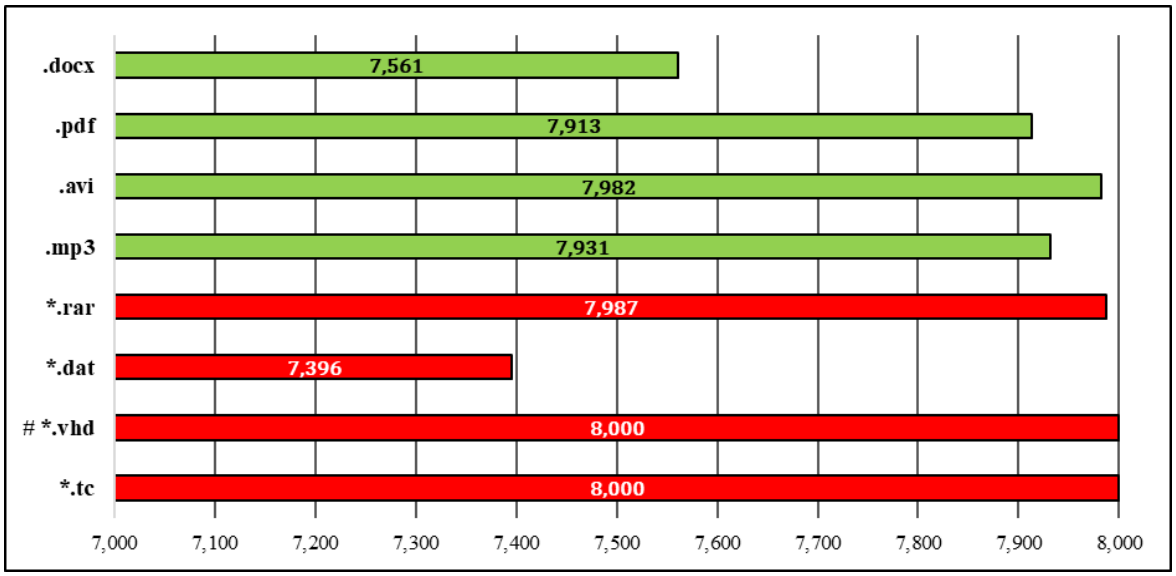
где n – общий размер входных данных в байтах, а кол(i) – количество байт со значением i [2, с. 31-32]. В свою очередь, уровень энтропии также будет зависеть от спецификации структуры (формата) входных данных. Так, структурированные по различным стандартам последовательности байт будут иметь отличные друг от друга значения энтропии.

Практическое применение измерения информационной энтропии и связанные с этим проблемы

Зашифрованные (упакованные) файлы или компоненты программ независимо от их размера или формата представляют собой данные с практически случайной последовательностью описывающих их бит. Это связано с тем, что алгоритмы шифрования преобразуют оригинальный структурированный массив данных в равномерно распределенную случайную последовательность байтов для их конфиденциальности. Алгоритмы сжатия кодируют входной массив согласно тому или иному словарю, в результате чего последовательность описывающих его битов приобретает свойства, близкие к равномерно распределенной случайности. На основании изложенного, указанные файлы и компоненты будут иметь максимальные значения энтропии (от 7 до 8).

Сотрудник при реализации правоохранительных задач исследует массив цифровых данных, содержащийся на изымаемых электронных накопителях. Первичная обработка указанного массива производится программным обеспечением для компьютерно-технических исследований (например, AutoPsy, PALADIN Forensic Suite, CAINE и другие). Их функционал имеет модули по измерению информационной энтропии файлов, размеченных и неразмеченных областей памяти электронных накопителей. Побайтно проанализированные данные могут группироваться по уровню их информационной энтропии, а также отмечаться как «зашифрованные/подозрительные» при уровне энтропии свыше 7. Для этого, также, можно воспользоваться статистическими тестами NIST (англ. – The National Institute of Standards and Technology) для определения упакованных, зашифрованных и случайных данных. В теории расчет эталонных диапазонов и пороговых значений энтропии для отдельных форматов файлов позволит сузить формируемую выборку и усовершенствовать их последующий поиск среди всего массива данных. Упомянутые программные модули могут вручную настраиваться для создания вышеуказанных диапазонов и значений. Отдельными форматами, представляющих интерес для реализации задач правоохранительных органов, могут признаваться криптоконтейнеры (*.tc, и любые иные), зашифрованные (#) образы жестких дисков виртуальных машин (*.vhd и иные),

файлы кошельков (токенов) криптовалют (wallet.dat, *.dat и иные), незашифрованные архивные файлы (*.rar, *.zip и иные). Измерив уровень энтропии выборки файлов вышеуказанных форматов, содержащихся в архивах цифровых доказательств, обнаруженных следователями Следственного комитета Республики Беларусь на электронных накопителях киберпреступников, можно рассчитать их средние показатели. Однако, указанный способ идентификации конкретных форматов данных имеет проблемы, связанные со спецификациями структуры отдельных, но общераспространенных форматов файлов (например, *.docx, *.pdf, *.avi, *.mp3). Средние значения энтропии файлов вышеуказанных форматов коррелируют с диапазоном значений энтропии зашифрованных (упакованных) данных (рисунок).



Средние значения энтропии файлов различных форматов
(выборка: 100 уникальных экземпляров на один формат)

При анализе электронного накопителя использование измерения энтропии как самостоятельного средства идентификации данных конкретного формата приведет к высокому проценту ложных срабатываний программного обеспечения, соответственно, низкому проценту их корректного определения. Решение обозначенной проблемы заключается в комбинировании всех имеющихся в настоящее время способов идентификации следов шифрования или сжатия.

Так измерение уровня энтропии наряду с проверкой соответствия расширений и спецификации структур исследуемых файлов (по базе сигнатур известных форматов) должно служить лишь для создания первичной выборки «подозрительных» объектов из всего исследуемого мас-

сива данных. Далее, характер распределения и последовательности байт в сформированной выборке предлагается определять посредством дополнительных математических критериев, а именно:

1) Использование критерия согласия Пирсона или критерия согласия $\chi 2$ «Хи-квадрат», основанного на оценке равномерности распределения байт без учета их расположения во всем массиве данных. Среднее значение указанного критерия для зашифрованных файлов составляет 255, в отличие от упакованных, значения которых варьируются от 2500 до 6500. Данный критерий чрезвычайно чувствителен к отклонениям равномерно распределенной случайно величины, поэтому его использование для обнаружения зашифрованных файлов является обоснованным [4, с. 710-712]. Однако, рассмотренный критерий имеет недостаток в виде пропуска (невозможности идентификации) зашифрованных данных небольшого размера из-за малого количества описывающих их бит [5, с. 36].

2) Аппроксимация числа π (Пи) методом Монте-Карло, выраженная в вычислении характеристики распределения байтов на координатной плоскости посредством измерения процента ошибки в приближенном значении числа π (Пи). Массив с псевдослучайными данными характеризуются величиной, близкой к значению 3,141592653. В свою очередь, большой процент ошибок (более 0,008 %), измеренный при аппроксимации числа π (Пи), является явным признаком сжатия данных, а в случае точного вычисления (ошибка менее 0,0008 %) – явным признаком шифрования [4, с. 710-712].

3) Определение коэффициента автокорреляции, который позволяет установить степень со-зависимости имеющихся в исследуемом массиве элементов. Зашифрованные данные характеризуются величиной коэффициента автокорреляции, близкой к 0, что говорит об относительном отсутствии со-зависимости всех элементов оцениваемой последовательности [5, с. 36-37].

4) Установление величины арифметического среднего, которая основывается на усредненной сумме значений байт всех элементов оцениваемой последовательности. Совокупность байтов со свойствами равномерно распределенной случайной последовательности, имеющей первичный алфавит, состоящий из 256 различными символов, характеризуется значением арифметического среднего, близким к 127,5 [5, с. 36-37].

5) Оценка плотности распределения байт в исследуемом файле на основании расчета плотности распределения его условных точек на плоскости (номер байта в файле, значение байта). У зашифрованных файлов выявленное посредством вейвлет-анализа (расчета вейвлет-коэффициентов для функции, описывающей сигнал) отклонение в рас-

пределении плотностей на этой плоскости не превосходит эталонные пороговые значения [6, с. 46].

Выводы

Таким образом, измерение энтропии информации как способ определения зашифрованных данных, представляющих интерес для реализации задач правоохранительных органов, может использоваться как первый этап исследования компьютерной информации. Создание упомянутым способом первичной выборки данных позволит сотруднику сэкономить имеющиеся вычислительные и временные ресурсы, которые могли быть затрачены на исследование всего массива информации, в том числе ручным способом. Определив границы потенциально зашифрованных (упакованных) файлов, размеченных и неразмеченных областей памяти, последние могут исследоваться посредством вышеуказанных дополнительных математических критериев. Описанные критерии могут комбинироваться или использоваться по-отдельности в зависимости от имеющихся у сотрудника вычислительных ресурсов или особенностей исследуемого массива данных.

В свою очередь, определение явных криптоконтейнеров (имеющие расширения *.tc, *.vc и иные), архивных файлов (имеющих расширение *.rar, *.zip и иные) или зашифрованных образов жестких дисков виртуальных машин (имеющих расширения *.vhd, *.vdi и иные) на исследуемых электронных накопителях теоретически возможно и с использованием лишь измерения информационной энтропии.

Библиографические ссылки

1. *Shennon C. A.* The Mathematical Theory of Communication / C. Shannon // The Bell System Technical Journal. 1948. V. 27. P. 379-423, 623-656 [Электронный ресурс]. – Режим доступа: web.archive.org/web/19980715013250/http://cm.bell-labs.com/cm/ms/what/ shannonday/shannon1948.pdf. – Дата доступа: 26.02.2023.

2. *Матвеева В. С.* Энтропия и ее использование для решения задач информационной безопасности / В. С. Матвеева // Безопасность информационных технологий – Том 21, № 3 (2014) [Электронный ресурс]. – Режим доступа: bit.mephi.ru/index.php/bit/article/view/163. – Дата доступа: 27.02.2023.

3. SP 800-22 Rev. 1a. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications / National Institute of Standards and Technology: Computer Security Resource Center [Электронный ресурс]. – Режим доступа: csrc.nist.gov/publications/detail/ sp/800-22/rev-1a/final. – Дата доступа: 28.02.2023.

4. *Югансон А. Н.* Метод определения упакованных и зашифрованных данных во встроенном программном обеспечении / А. Н. Югансон // Научно-технический вестник информационных технологий, механики и оптики. 2020. Т. 20. № 5. – С. 708-713.

5. *Матвеева В. С.* Криптография и вредоносные программы / В. С. Матвеева // Information Security/ Информационная безопасность. – 2015. – № 1. – С. 35-38.

6. *Матвеева В. С., Мамаев А. В.* Вейвлет анализ для локализации неоднородностей в распределении байт в файле с целью идентификации зашифрованных данных / В. С. Матвеева // Безопасность информационных технологий. -2015 г. -№ 1. - С. 40-46.