На правах рукописи

Матвеева Веста Сергеевна

СТАТИСТИЧЕСКИЙ МЕТОД ОБНАРУЖЕНИЯ ЛОКАЛЬНЫХ НЕОДНОРОДНОСТЕЙ ДАННЫХ ДЛЯ РАССЛЕДОВАНИЯ ИНЦИДЕНТОВ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Специальность: 05.13.19 – методы и системы защиты информации, информационная безопасность

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Автор: _____

Москва – 2015

Работа выполнена в Национальном исследовательском ядерном университете «МИФИ»
(НИЯУ МИФИ)

Научный руководитель:

Кандидат технических наук,
доцент кафедры
«Криптология и дискретная математика»
НИЯУ МИФИ
Епишкина Анна Васильевна

Официальные оппоненты:

Доктор физико-математических наук,
старший научный сотрудник,
заведующий кафедрой
информационной безопасности
НИУ ВШЭ,
Баранов Александр Павлович

Кандидат технических наук,
доцент,
начальник Управления испытаний и сертификации
информационных технологий и систем в
Международном центре по информатике
и электронике «ИнтерЭВМ»
Безкоровайный Михаил Михайлович

Ведущая организация:

Федеральное государственное бюджетное
образовательное учреждение высшего
профессионального образования
«Московский Государственный Технический
Университет имени Н.Э. Баумана»
(МГТУ им. Н.Э. Баумана)

Защита состоится «07» октября 2015 г. в 15 часов 00 минут на заседании диссертационного совета Д 212.130.08 на базе Национального исследовательского ядерного университета «МИФИ» по адресу: 115409, г. Москва, Каширское ш., д. 31. Тел. для справок: +7 (499) 324-87-66, +7 (495) 788-56-99.

С диссертацией можно ознакомиться в библиотеке Национального исследовательского ядерного университета «МИФИ» и на сайте: http://ods.mephi.ru.

Просим принять участие в работе совета или прислать отзыв в двух экземплярах, заверенный печатью организации.

Автореферат разослан «___» _____ 2015 г.

Ученый секретарь диссертационного совета  Горбатов В.С.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. При обнаружении инцидента информационной безопасности (ИБ) согласно ГОСТ Р ИСО/МЭК ТО 18044-2007 встает задача выявления его причин и причастных к нему лиц в рамках расследования. Проводится сбор доказательной базы, основными источниками которой зачастую являются электронные носители информации. Для уничтожения и сокрытия следов работы на электронном носителе информации нарушителями ИБ могут использоваться разные подходы, среди которых одними из самых действенных признаются перезапись и шифрование файлов.

Поиск перезаписанных файлов является важной задачей при проведении исследования электронного носителя информации, поскольку выявление обстоятельств удаления файла и сведений о нем может дать представление об используемых программах, скрываемых данных, времени использования файла и так далее. Без выявления факта перезаписи данных следы использования интересующих сведений на исследуемом электронном носителе информации могут быть пропущены, что приведет к неверным выводам в рамках проведения расследования инцидента ИБ.

Поиск зашифрованных файлов также является важной задачей при проведении исследования электронного носителя информации, поскольку дает возможность выявить скрываемые значимые сведения. Обнаружение таких файлов инициирует процесс выяснения данных для расшифрования, которые могут быть дополнительно собраны в рамках первичного реагирования на инцидент ИБ. Кроме этого, процесс поиска зашифрованных файлов позволяет выявлять зашифрованные модули, настройки и скопированные сведения вредоносных программ, что инициирует поиск самой вредоносной программы, с помощью которой эти зашифрованные данные можно расшифровать.

Генераторы псевдослучайных чисел, используемые средствами перезаписи данных, и алгоритмы шифрования, используемые современными средствами криптографической защиты информации (СКЗИ), обеспечивают свойства выходных последовательностей, близкие к свойствам равномерно распределенной случайной последовательности (РРСП). В работе решается задача поиска файлов с данными, которые представляют собой последовательность байтов со свойствами РРСП. При этом размер последовательности равен размеру файла. Далее такие файлы будем называть файлами, представляющими собой РРСП.

Задачей поиска перезаписанных псевдослучайными числами данных занимались исследователи: Г.К. Кесслер (G.C.Kessler), Г.Х. Карлтон (G.H.Carlton), А. Саволди (A. Savoldi), М. Пицинелли (M. Piccinelli), П. Губиан (P. Gubian), С.Л. Гарфинкел (S.L. Garfinkel), Г. Печерл

(G. Pecherle), К. Джоуроди (C. Győrödi), Р. Джоуроди (R. Győrödi), Б. Андроник (B. Andronic), И. Игнат (I. Ignat) и др.

Задачей поиска зашифрованных данных занимались отдельные отечественные и зарубежные исследователи: И. Сластенова, Т. Плотникова, И. Кесей (E. Casey), И. Джозиак (I. Jozwiak), М. Кейчара (M. Kedziora), А. Мейлинск (A. Melinska), Д. Зигфрид (J. Siegfried), К. Сиедсма (C. Siedsma), Б.-Д. Кантримен (B.-J. Countryman), Ч. Д. Хасмер (C. D. Hosmer) и специалисты, разрабатывающие методы определения типа файла по его содержимому, а также компании: ООО «Национальный центр по борьбе с преступлениями в сфере высоких технологий», Passware Inc. и др.

В результате анализа опубликованных способов обнаружения файлов, перезаписанных псевдослучайными числами, и зашифрованных файлов выявлены такие их недостатки, как: ограничение размера искомого файла, ошибки в обнаружении искомых файлов, необходимость большого количества времени для поиска, отсутствие учета локальных особенностей содержимого файлов. В результате анализа существующих средств обнаружения зашифрованных файлов выявлено, что они работают только на уровне файлов в явном виде и задействуют функции самой операционной системы (ОС) для получения их содержимого, что может использоваться для предотвращения доступа к данным. В открытых источниках отсутствуют сведения о специализированных подходах, а также программных средствах для точного обнаружения файлов, представляющих собой РРСП, без указанных существенных недостатков. Под точностью обнаружения файлов, представляющих собой РРСП, в работе понимается число искомых файлов, которое выявляется оцениваемым способом, относительно общего количества таких файлов на исследуемом электронном носителе информации.

Решение обеих приведенных задач основывается на свойствах выходных данных и особенностях реализации применяемых программных средств, поэтому они идентичны в части свойств самих данных, однако различны в части условий их создания применяемыми для этого средствами. Подходы к выявлению зашифрованных данных, описанные в открытых источниках, могут применяться для поиска перезаписанных псевдослучайными числами файлов.

В связи с вышеуказанными причинами задача поиска файлов, представляющих собой РРСП, является актуальной и требует разработки метода обнаружения таких файлов, который учитывает внутреннюю структуру содержимого файла, снижает ошибки в обнаружении искомых файлов, а также позволяет выявлять локальные неоднородности в содержимом файла.

Объект исследования. Файлы на электронном носителе информации, анализируемом при расследовании инцидента ИБ.

Предмет исследования. Данные, содержащиеся в файлах, и их статистические свойства.

Цель диссертационной работы. Разработка метода, используемого при расследовании инцидентов ИБ, который на основании статистического анализа содержимого файлов позволяет выявлять локальные неоднородности данных, не характерные для РРСП.

Научная задача заключается в анализе статистических свойств содержимого файлов с точки зрения их близости к свойствам РРСП для разработки метода обнаружения файлов, представляющих собой РРСП.

В рамках решения научной задачи необходимо:

– провести сравнительный анализ и классификацию существующих способов обнаружения файлов, представляющих собой РРСП;

– исследовать статистические свойства содержимого файлов в файловой системе с точки зрения их близости к свойствам РРСП;

– получить математическое описание статистических свойств РРСП, которые могут использоваться для выявления локальных неоднородностей в содержимом файлов;

– сформулировать необходимое условие обнаружения файлов, представляющих собой РРСП;

– разработать метод обнаружения файлов, представляющих собой РРСП;

– создать архитектуру средства обнаружения файлов, представляющих собой РРСП, и реализовать средство обнаружения.

Методы исследований. В работе используются методы теории вероятностей и математической статистики, теории информации и вейвлет-анализа.

Научная новизна работы состоит в следующем:

– получено математическое описание статистических свойств РРСП, которые могут использоваться для выявления локальных неоднородностей в содержимом файлов, в том числе сжатых форматов;

– сформулировано и обосновано необходимое условие обнаружения файлов, представляющих собой РРСП, изложенное с использованием математической статистики и вейвлет-анализа, которое позволяет оценить свойства распределения байтов в файле с точки зрения их близости к свойствам распределения, характерного для РРСП;

– разработан метод обнаружения файлов, представляющих собой РРСП, основанный на особенностях выбора параметров для вейвлет-преобразования содержимого файла с целью детального анализа его свойств, в том числе выявления локальных неоднородностей.

Практическая значимость результатов работы заключается в том, что разработанный метод, а также построенное на его основе средство обнаружения файлов, представляющих собой РРСП, используются при проведении исследований электронных носителей информации

в рамках расследования инцидентов ИБ, в том числе при проведении судебной компьютерной экспертизы. Разработанный метод позволяет получить близкую к нулю вероятность того, что файлы, представляющие собой РРСП, будут пропущены. Реализованное на его основе средство обнаружения позволяет провести поиск за время, приемлемое в рамках реагирования на инциденты ИБ, которое в работе принято в пределах от 4 до 8 часов. Также реализованное программное средство обнаружения осуществляет доступ к файлам путем непосредственного считывания содержимого кластеров электронного носителя информации, выделенных под них, поэтому становится возможным получать доступ к содержимому удаленных файлов в файловой системе, к файлам ограниченного доступа, отдельным кластерам, свободной области файловой системы, а также неразмеченной области электронного носителя информации при обращении к кластерам этой области.

Результаты работы представляют практическую ценность для обеспечения ИБ и для реагирования на инциденты ИБ в частности.

Достоверность результатов. Достоверность теоретических результатов обеспечивается корректностью применения математического аппарата для вывода основных утверждений, сформулированных в работе. Теоретические результаты подтверждаются результатами проведенных экспериментов.

Внедрение результатов исследований. Средство обнаружения файлов, представляющих собой РРСП, включено в состав криминалистических средств для проведения судебных компьютерных экспертиз экспертами 12 отдела компьютерных экспертиз и психофизических исследований ЭКЦ ГУ МВД России по г. Москве, Лаборатории компьютерной криминалистики и исследования вредоносного кода ООО «Группа информационной безопасности» и ООО «ТРАСТ», что подтверждается соответствующими актами о внедрении.

Результаты проведенного анализа статистических свойств содержимого файлов в сравнении со свойствами РРСП внедрены в образовательный процесс кафедры «Криптология и дискретная математика» Национального исследовательского ядерного университета «МИФИ» в рамках учебного курса «Криптографические средства обеспечения информационной безопасности».

Публикации и апробация работ. Результаты диссертации изложены в 9 публикациях, 5 из которых опубликованы в рецензируемых журналах ВАК РФ, а 2 опубликованы в рецензируемых журналах Scopus. Результаты работы докладывались на Всероссийской научно-практической конференции «Проблемы информационной безопасности в системе высшей

школы» г. Москва, 2014-2015 гг. и на научно-практическом семинаре в Центре специальных разработок Министерства обороны Российской Федерации.

Основные положения, выносимые на защиту:

– математическое описание статистических свойств РРСП, которые могут использоваться для выявления локальных неоднородностей данных в файлах;

– необходимое условие обнаружения файлов, представляющих собой РРСП;

– метод обнаружения файлов, представляющих собой РРСП, который основан на вейвлет-анализе;

– результаты тестирования средства обнаружения файлов, представляющих собой РРСП.

Структура работы. Работа состоит из введения, четырех глав, заключения, списка сокращений и условных обозначений, списка используемых источников, включающего 98 наименований, и 9 приложений. Текст диссертации изложен на 100 страницах, включая 43 рисунка и 14 таблиц.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность темы диссертации, представляется общий обзор предметной области, формулируются цель и задачи, предмет и объект исследования, описывается структура и логика диссертационной работы.

В **первой главе** систематизируются способы, которые применяются для поиска файлов, представляющих собой РРСП, приводится их классификация, а также указываются недостатки их применения для решения поставленной задачи. Проводится сравнительный анализ программных средств, которые используют существующие способы.

В результате исследования способов обнаружения файлов, представляющих собой РРСП, предложена их классификация (рисунок 1), согласно которой все способы делятся на две группы: основанные на сигнатурном анализе данных и на статистическом анализе данных.

Средства для перезаписи данных, как правило, не вносят специальных сигнатур, расширений и заголовков в перезаписываемые файлы, в то время как СКЗИ могут вносить подобные изменения, что позволяет проводить поиск зашифрованных с их помощью файлов по подготовленному набору сигнатур. Сигнатурный анализ не учитывает внутреннюю структуру содержимого файлов.

Статистический анализ основан на оценке статистических свойств РРСП, а именно равномерности распределения элементов оцениваемой последовательности, построенной по содержимому файла, и независимости элементов между собой. Статистический анализ может

осуществляться путем подсчета оценочной статистической величины или графическим способом путем наглядного анализа.

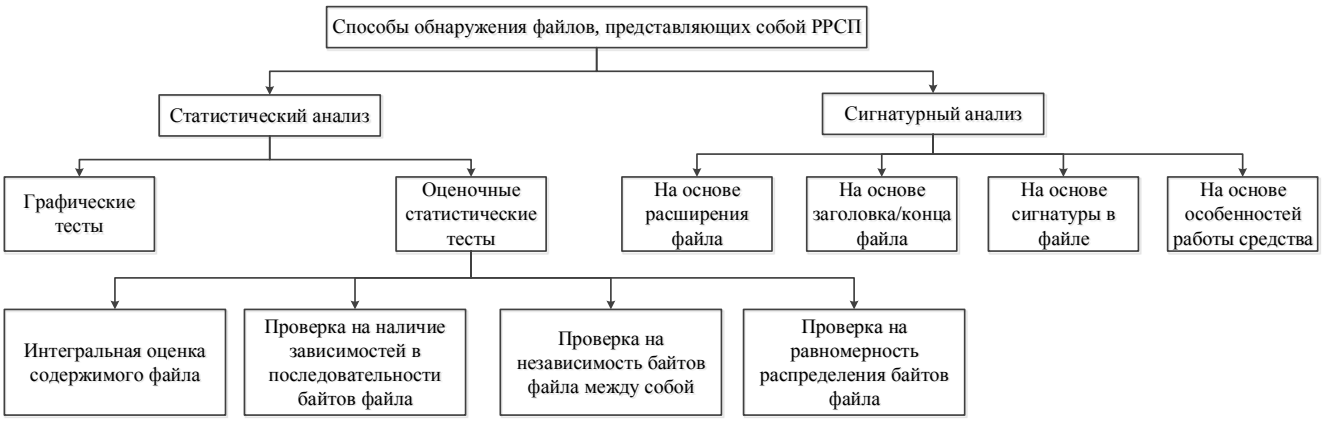

Рисунок 1 – Классификация существующих способов обнаружения файлов, представляющих собой РРСП

В результате проведенного исследования выявлено, что одни способы не учитывают внутреннюю структуру содержимого файлов, а другие разрабатывались для задач, отличных от поставленной, и не учитывают особенностей файловой системы и существующих форматов файлов. Кроме этого, файлы сжатых форматов проходят многие тесты на проверку статистических свойств, поэтому их применение по отдельности приводит к многочисленным ошибкам. Для устранения или уменьшения количества ошибок требуется применение тестов в совокупности, что является затратным по времени. В работе под приемлемым временем реагирования на инциденты ИБ принимаем время реагирования в пределах от 4 до 8 часов.

Также в данной главе проводится сравнительный анализ существующих автоматизированных средств, использующих расчет статистических величин для оценки содержимого файлов и наборы сигнатур для различения файлов известных форматов. В результате анализа выявлено, что основными недостатками существующих средств являются ограничение на минимальный размер искомых файлов, а также работа только на уровне файлов в явном виде. Для доступа к файлам используются функции ОС, поэтому файлы ограниченного доступа не анализируются.

На основе проведенного исследования выявлено, что задача поиска файлов, представляющих собой РРСП, является актуальной и требует разработки метода обнаружения таких файлов, лишенного указанных недостатков.

Во **второй главе** предлагается способ представления содержимого файлов, который положен в основу исследования свойств данных с точки зрения их близости к свойствам РРСП. С помощью вейлет-анализа для РРСП получено математическое описание статистических

свойств, которые могут использоваться для выявления локальных неоднородностей данных в файлах. На основании этого описания формулируется необходимое условие обнаружения файлов, представляющих собой РРСП, для проверки гипотезы различимости файлов, представляющих собой РРСП, и файлов других форматов, в том числе сжатых.

Файл f представляется в виде конечной последовательности байтов $\{y_1, y_2, ..., y_n\}$ длины n, где $y_i \in Y = \{0,1,...,255\}$, $|Y| = k = 256$ – мощность множества Y. Обычно эта последовательность анализируется последовательно байт за байтом или по блокам. Рассмотрим эту последовательность иначе для обеспечения оценки распределения значений байтов методом скользящего окна с учетом их взаимного расположения относительно друг друга.

Последовательность наносится на плоскость (плоскость распределения), по оси абсцисс которой откладываются значения от 0 до $n-1$, а по оси ординат – от 0 до 255. Плоскость заполняется точками (x, y), такими что x – номер байта в последовательности байтов файла, а y – значение этого байта.

Определение. Распределение точек на плоскости назовем равномерным, если для любых $(x, y) \in \{0,..., n-1\} \times \{0,...,255\}$ вероятность появления на плоскости точки (x, y) равна $P_{(x,y)} = 1/k.$

Для РРСП распределение точек на плоскости равномерно. Для изучения свойств распределения точек на плоскости распределения наложим на нее сетку из прямоугольников размером $w \times h$ (фрагмент), где w – ширина прямоугольника, h – его высота, и осуществим отображение f_1 последовательности байтов файла в последовательность плотностей:

$$f_1 : \{y_1, y_2, ..., y_n\} \to \{\rho_{11}, \rho_{12}, ..., \rho_{LC}\}, \tag{1}$$

где ρ_{ij} – плотность точек во фрагменте размера $w \times h$, который находится в i-ой строке, в j-ом столбце, и рассчитывается по формуле (2):

$$\rho_{ij} = \frac{n_{ij}}{w \cdot h}, \tag{2}$$

где n_{ij} – количество точек во фрагменте размера $w \times h$, который находится в i-ой строке и в j-ом столбце, $i = \overline{1, L}$, $j = \overline{1, C}$, $L = \left\lceil \frac{k}{h} \right\rceil$ – количество строк фрагментов, на

которые разбита плоскость распределения, где операция $\lceil\ \rceil$ определяется следующим образом: $\lceil x \rceil$ – наименьшее целое число, для которого выполняется неравенство $x \le \lceil x \rceil$, $C = \left\lceil \frac{n}{w} \right\rceil$ – количество столбцов фрагментов, на которые разбита плоскость распределения. Если n не кратно w, то в качестве последнего фрагмента в строке берется фрагмент, начинающийся со значения $n-w-1$, длины w. Если k не кратно h, то в качестве последнего фрагмента в столбце берется фрагмент, начинающийся со значения $k-h-1$, высоты h.

Теоретическое значение n_{ij} рассчитывается по формуле (3):

$$n_{ij} = \begin{cases} \sum\limits_{m=w\cdot(j-1)}^{w\cdot j-1} 1_{\{h\cdot(i-1)\le y_m < h\cdot i\}}, i = \overline{1, L-1},\ j = \overline{1, C-1} \\ \sum\limits_{m=n-w-1}^{n-1} 1_{\{h\cdot(i-1)\le y_m < h\cdot i\}}, i = \overline{1, L-1},\ j = C \\ \sum\limits_{m=w\cdot(j-1)}^{w\cdot j-1} 1_{\{k-h-1\le y_m < k-1\}}, i = L,\ j = \overline{1, C-1} \end{cases}, \quad (3)$$

где $1_{\{l_1 \le y_m < l_2\}}$ – индикатор выполнения неравенства, заданного в индексе: значение y_m попадает в диапазон значений фрагмента, $m = \overline{c_1, c_2}$, c_1, c_2 – границы фрагмента по оси абсцисс, l_1, l_2 – границы фрагмента по оси ординат. Значение индикатора рассчитывается по формуле (4):

$$1_{\{l_1 \le y_m < l_2\}} = \begin{cases} 0, y_m \ge l_2, y_m < l_1 \\ 1, l_1 \le y_m < l_2 \end{cases}. \quad (4)$$

При осуществлении отображения f_1 движение от фрагмента к фрагменту целесообразно осуществлять методом скользящего окна слева направо, так как скопление значений байтов с одним значением будет располагаться горизонтально. При вертикальном движении захватываются все значения, которые имеются в подпоследовательности длины w, в связи с чем имеющиеся скопления значений будет сложно выявить в построенной таким способом последовательности плотностей.

Для РРСП плотность ρ_{ij} является случайной величиной, которая имеет полиномиальное распределение с количеством исходов, равным количеству фрагментов в столбце $\left\lceil \frac{k}{H} \right\rceil$. Для

случайной величины ρ_{ij} построен закон распределения, функции вероятности и распределения, а также получены теоретические значения математического ожидания и дисперсии:

$$\begin{cases} E\rho_{ij} = \dfrac{1}{k} \\ D\rho_{ij} = \dfrac{k-h}{w \cdot h \cdot k^2} \end{cases}. \quad (5)$$

Математическое ожидание плотности не зависит от размера фрагмента, а дисперсия зависит от выбранных значений ширины и высоты фрагмента.

В работе сформулирована **гипотеза различимости файлов, представляющих собой РРСП, и файлов других форматов**: для файлов, не представляющих собой РРСП, распределение точек на плоскости распределения не является равномерным, т.е. содержит области повышенной или пониженной концентрации точек, которые представляют собой выраженные отклонения элементов последовательности плотностей от соответствующих значений, характерных для РРСП.

На основании полученных значений математического ожидания и дисперсии (5), характерных для РРСП, проводится выделение файлов с близкими к РРСП свойствами. Такими файлами являются файлы сжатых форматов в силу использования выравнивания вероятности встречаемости символов и самой природы сжатия.

Экспериментально автором установлено, что файлы сжатых форматов имеют выраженные локальные отклонения в последовательности плотностей от значений, характерных для РРСП, что может использоваться для различения этих файлов и РРСП.

Для локализации отклонений различного характера в последовательности плотностей автором используется вейвлет-анализ, который применяется для анализа сигналов и позволяет разложить их во временном и в частотном пространстве одновременно. Базисные функции для вейвлет-преобразования локализованы по времени и частоте, поэтому они предоставляют возможность одинаково хорошо выявлять и низкочастотные, и высокочастотные характеристики сигналов, что отличает их от гармонической базисной функции преобразования Фурье. Форма базисной функции должна походить на искомую форму отклонения, т.е. рассчитываемые вейвлет-коэффициенты показывают степень близости анализируемого сигнала и анализирующего вейвлета. Вейвлет-преобразование, в отличие от оконного преобразования Фурье, обладает изменяемым (подвижным) временным окном, узким на малых временных масштабах и широким на больших, а свойства окна (его ширина и перемещение по частоте) присущи самим вейвлетам.

С помощью вейвлет-преобразования анализируется последовательность плотностей $\{\rho_{11},\rho_{12},...,\rho_{LC}\}$, при этом в качестве базисной функции выбирается вейвлет Хаара, так как искомое отклонение может быть выражено в виде единичного значения, а также в виде группового всплеска или спада значений (отклонение нескольких подряд идущих значений плотностей в последовательности плотностей). С помощью выбранной базисной функции проводится поиск резкого перепада значений независимо от его формы. Для упрощения записи дальнейших формул нумеровать плотности в последовательности плотностей будем по порядку: $\{\rho_1,\rho_2,...,\rho_{L\cdot C}\}$. В этом случае вейвлет-преобразование последовательности плотностей с использованием базисной функции Хаара может быть представлено следующим образом:

$$W_{ab}=\frac{1}{\sqrt{a}}\sum_{t=b}^{b+a}\rho_t\psi_{HAAR}(\frac{t-b}{a}), \tag{6}$$

где $a\in N, a\neq 0$ – параметр масштаба (т. е. количество значений элементов последовательности плотностей, используемых для подсчета вейвлет-коэффициента), $b\in N, b\neq 0$ – параметр сдвига (т. е. позиция в последовательности плотностей, от которой начинается отсчет элементов в этой последовательности), ρ_t – значение плотности *t*-го фрагмента, $t=\overline{1,L\cdot C}$.

Таким образом, совершаем отображение f_2 последовательности плотностей в последовательность вейвлет-коэффициентов для выбранного значения параметра масштаба a:

$$f_2:\{\rho_1,\rho_2,...,\rho_{L\cdot C}\}\rightarrow\{W_{a1},W_{a2},...,W_{a(n-a)}\}. \tag{7}$$

Для РРСП каждый элемент последовательности вейвлет-коэффициентов представляет собой случайную величину со следующими значениями математического ожидания и дисперсии:

$$\begin{cases} EW_{ab}=0 \\ DW_{ab}=\dfrac{k-h}{w\cdot h\cdot k^2} \end{cases}. \tag{8}$$

Математическое ожидание вейвлет-коэффициента не зависит от размера фрагмента и параметров вейвлет-преобразования, а дисперсия зависит только от выбранных значений ширины и высоты фрагмента.

В работе выдвинута гипотеза о том, что вейвлет-коэффициент как случайная величина имеет нормальное распределение $\mathrm{N}(0, DW_{ab})$. Гипотеза о виде распределения подтверждена в результате эксперимента с помощью критерия согласия «Хи-квадрат», на основании чего можно провести оценку порогового значения модуля вейвлет-коэффициента при выбранном значении вероятности α через значение квантиля z_α нормального распределения $\mathrm{N}(0,1)$:

$$|W_{ab}| \le z_\alpha \sqrt{DW_{ab}}, \quad \forall W_{ab} \in \{W_{a1}, W_{a2}, ..., W_{a(n-a)}\}. \tag{9}$$

Например, при $\alpha = 0{,}999$: $z_\alpha \approx 3{,}09$.

Будем считать, что последовательность содержит локальную неоднородность, если хотя бы одно неравенство с вейвлет-коэффициентами в (9) для разных параметров вейвлет-преобразования не выполняется. Таким образом, локальная неоднородность данных может быть определена следующим образом: последовательность содержит локальную неоднородность, если существует такое значение параметра сдвига b, что модуль вейвлет-коэффициента при некотором значении параметра масштаба a удовлетворяет неравенству: $|W_{ab}| > z_\alpha \sqrt{DW_{ab}}$. Локальные неоднородности могут быть выражены в виде отклонений различного характера в последовательности плотностей от значений, характерных для РРСП.

Таким образом, научная задача может быть формализована следующим образом: разработать метод обнаружения файлов, статистические свойства содержимого которых описываются выражениями в (5) и (9), при этом метод должен иметь вероятность ошибки I рода, близкую к 0, и вероятность ошибки II рода, превышающую соответствующую вероятность критерия «Хи-квадрат» на равномерность не более, чем на 0,01.

Разработка метода должна быть направлена на проверку гипотезы H_0: «Рассматриваемый файл представляет собой РРСП» и альтернативной ей гипотезы H_1: «Рассматриваемый файл отличен от РРСП».

Для экспериментальной проверки выдвинутой в работе гипотезы различимости файлов, представляющих собой РРСП, и файлов других форматов, формулируется необходимое условие обнаружения файлов, представляющих собой РРСП: если файл представляет собой РРСП, то для него справедливы следующие свойства:

$$\begin{cases} \varepsilon_{1\min} < E\rho_{ij} < \varepsilon_{1\max} \\ \varepsilon_{2\min} < D\rho_{ij} < \varepsilon_{2\max} \\ |W_{ab}| - \varepsilon_3 \le 0, \forall W_{ab} \in \{W_{a1}, W_{a2}, ..., W_{a(n-a)}\} \end{cases}, \tag{10}$$

где $\varepsilon_{1\min}, \varepsilon_{1\max}, \varepsilon_{2\min}, \varepsilon_{2\max}, \varepsilon_3$ определяются экспериментально.

Проверка выдвинутой гипотезы с применением указанного необходимого условия будет осуществляться в последующих главах.

В **третьей главе** проводится исследование зависимости характера выявляемых отклонений в последовательности плотностей от параметров вейвлет-преобразования, а также разрабатывается метод обнаружения файлов, представляющих собой РРСП, основанный на проверке необходимого условия. В главе создается архитектура средства обнаружения файлов, представляющих собой РРСП, и анализируется программная реализация средства обнаружения.

Уменьшая размер фрагмента, получаем более детальный анализ распределения байтов, а значит более явно выраженное скопление или разреженность их расположения на плоскости распределения. В работе сформулировано и проверено экспериментально утверждение о том, что амплитуда плотности в последовательности плотностей возрастает с уменьшением размера фрагмента. Следовательно, для выявления отклонений от значений, характерных для РРСП, необходимо выбрать наименьший размер фрагмента с учетом размера самого файла, чтобы затратить приемлемое время для его обработки. На основании проведенного исследования разработан способ выбора размера фрагмента в зависимости от размера файла, который позволяет выявлять разные виды отклонений в оцениваемой последовательности.

Выявленная зависимость представлена на примере файла-архива формата «zip», использующего алгоритм сжатия LZMA и максимальный уровень сжатия (рисунок 3).

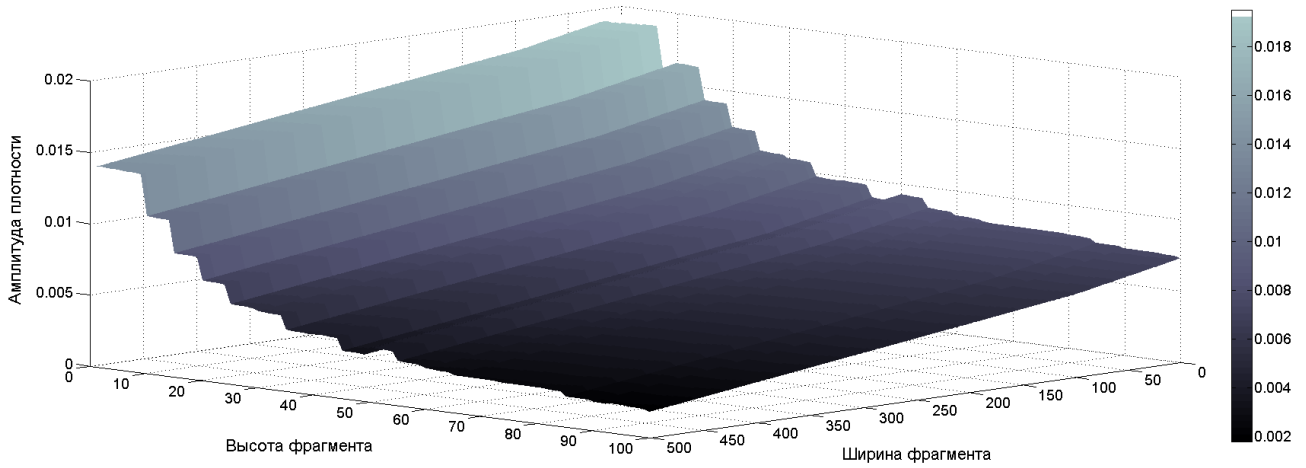

Рисунок 3 – Зависимость амплитуды плотности в последовательности плотностей от размера фрагмента на примере файла-архива формата «zip»

На детализацию характера отклонений с помощью вейвлет-анализа оказывает влияние значение параметра масштаба a. В работе проведено исследование зависимости характера выявляемых отклонений в последовательности плотностей от значения параметра масштаба, так как в значение вейвлет-коэффициента вносит вклад количество суммируемых значений плотностей, равное параметру масштаба. Выявлено, что локальные единичные отклонения

могут быть выявлены небольшим значением ширины окна (a=2, 4, 8), в то время как групповые – большим (a=16, 32, 64).

Полученные выводы экспериментально подтверждены на коллекции файлов разных форматов. Примеры гистограмм вейвлет-коэффициентов для файла формата «pdf», разных значений параметра масштаба и размера фрагмента приведены на рисунках 4 – 6.

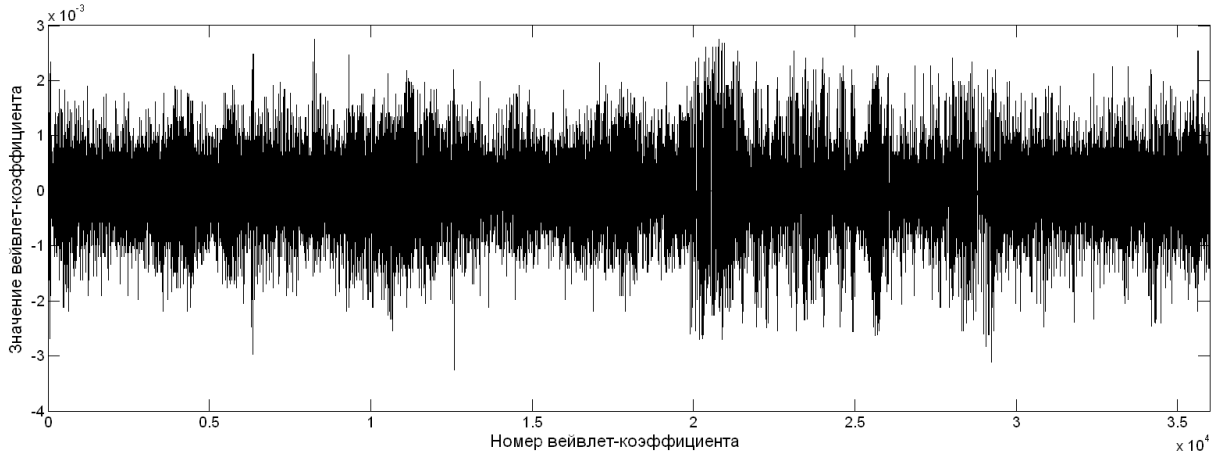

Рисунок 4 – Гистограмма вейвлет-коэффициентов для файла формата «pdf», размер фрагмента 100×100 и a=2

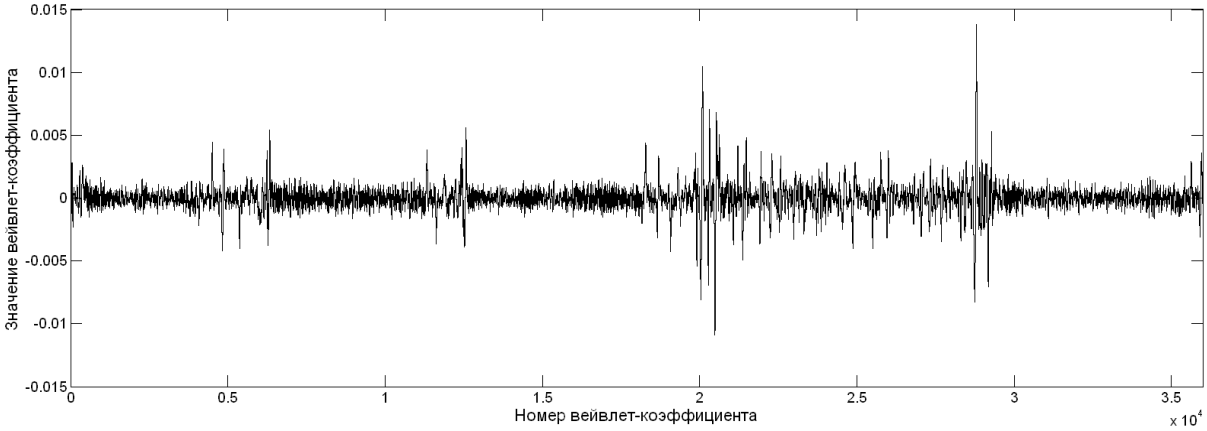

Рисунок 5 – Гистограмма вейвлет-коэффициентов для файла формата «pdf», размер фрагмента 100×100 и a=64

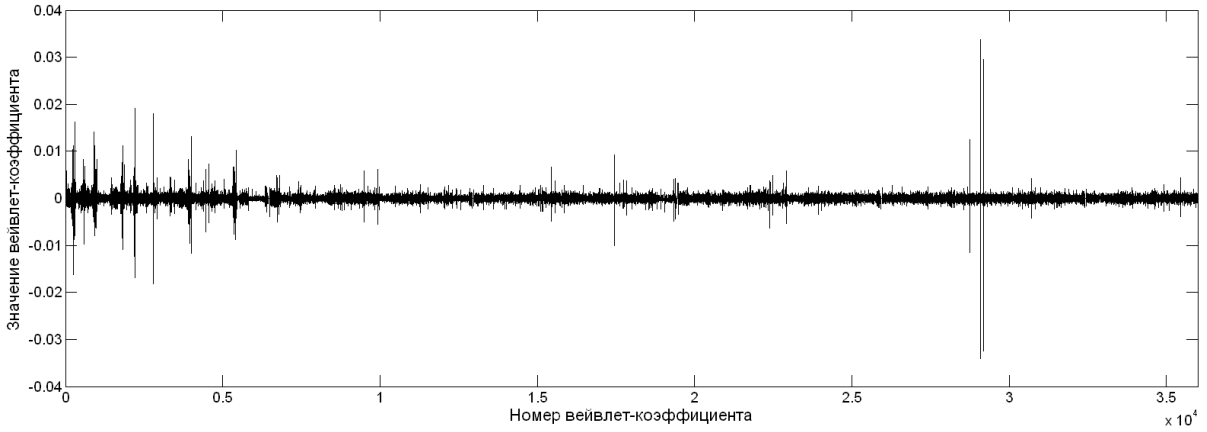

Рисунок 6 – Гистограмма вейвлет-коэффициентов для файла формата «pdf», размер фрагмента 2000×5 и a=2

На основании проведенного исследования разработан способ выбора значения параметра масштаба, который позволяет выявлять локальные и групповые отклонения в оцениваемой последовательности.

Полученные выводы позволили разработать метод обнаружения файлов, представляющих собой РРСП. Метод основан на оценке свойств содержимого файла с помощью статистических характеристик, полученных для РРСП, а также вейвлет-коэффициентов для выявления локальных неоднородностей данных. Рассчитываемые значения должны удовлетворять неравенствам в (10), исходя из математического описания отсутствия неоднородности. Размер фрагмента выбирается на основе разработанного способа выбора размера фрагмента в зависимости от размера файла, а параметры для вейвлет-преобразования выбираются на основе разработанного способа выбора значения параметра масштаба. В результате применения метода файлы делятся на два класса: X_1 – файлы, представляющие собой РРСП, и X_2 – файлы, не представляющие собой РРСП.

Предлагаемый в работе метод использует статистический анализ для обнаружения файлов, представляющих собой РРСП. Он лишен выявленных в главе 1 недостатков существующих способов, так как разрабатывался для применения в файловой системе с учетом статистических особенностей современных форматов файлов, в том числе сжатых.

Для построения архитектуры средства обнаружения формулируются требования к поиску файлов, представляющих собой РРСП, на основе разработанного метода и к программной реализации средства обнаружения для преодоления системных ограничений на доступ. Причем программная реализация должна быть осуществлена для файловой системы NTFS и ОС версии выше «Microsoft Windows Vista».

С учетом требований создана архитектура средства обнаружения, изображенная на рисунке 7, где между блоками на схеме отображены запрашиваемые и передаваемые между подсистемами данные.

Реализация средства осуществлялась на языке программирования C#.

Подсистема взаимодействия с пользователем представляет собой графический интерфейс, через который пользователь имеет возможность отдавать команды подсистемам средства обнаружения. В графическом интерфейсе отображаются результаты работы средства.

Подсистема работы с физическим носителем информации отвечает за разбор структуры выбранного электронного носителя информации, обход дерева каталогов в файловой системе, считывание кластеров, выделенных под файл, и представление их в виде последовательности байтов. На выходе подсистемы получается последовательность байтов для файла или кластера, которая передается на вход подсистемы построения последовательности плотностей.

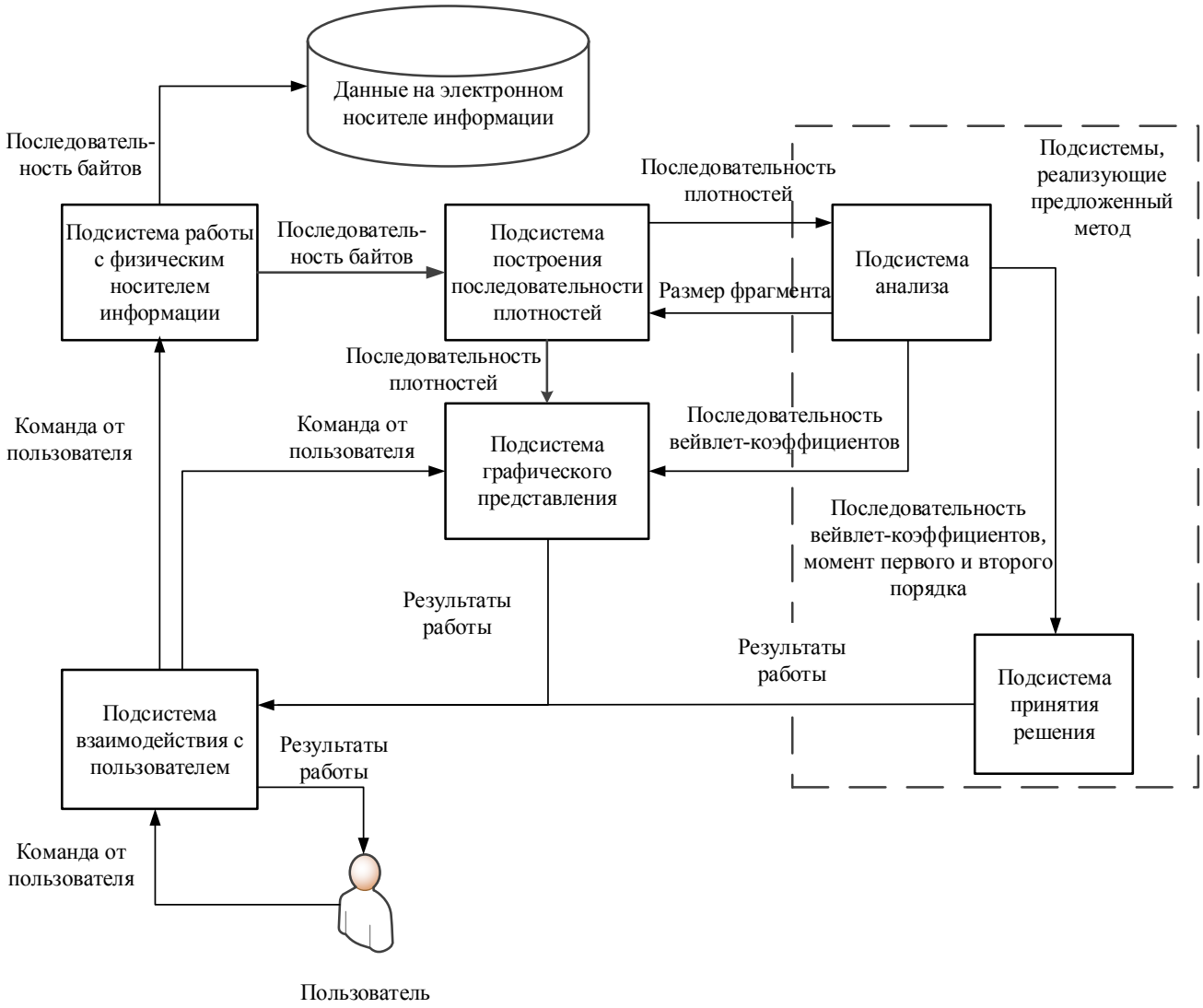

Рисунок 7 – Архитектура средства обнаружения файлов, представляющих собой РРСП

В подсистеме построения последовательности плотностей осуществляется отображение f_1, приведенное в (1), с учетом выбора размера фрагмента. Полученная последовательность плотностей передается на вход подсистеме анализа.

Подсистема анализа и подсистема принятия решения работают вместе в соответствии с предложенным методом. Подсистемой анализа проводится расчет статистических величин и вейвлет-коэффициентов, а подсистемой принятия решения проводится сравнение полученных величин с заданными пороговыми значениями. На основе сравнения подсистемой принятия решения запрашиваются новые величины у подсистемы анализа в случае прохождения условий сравнения и принимается решение о классе, к которому относится файл, в случае окончания проверок или выхода рассчитанных значений за пределы допустимых границ.

Подсистема графического представления реализует требование о возможности визуализации получаемых последовательностей в виде гистограмм для упрощения их анализа пользователем.

В описанную архитектуру средства обнаружения не включена система выработки порогового значения, так как она представляет собой то же средство обнаружения за исключением того, что все тестируемые файлы должны быть отнесены к классу X_1, поэтому ее код меняется в области подсистемы принятия решения.

Средство обнаружения файлов, представляющих собой РРСП, позволяет использовать его в свободной области файловой системы и в неразмеченной области электронного носителя информации на основе анализа кластеров как отдельных файлов. Из описания подсистем средства обнаружения вытекает порядок эксплуатации разработанного средства. Результаты тестирования программной реализации предложенного метода приведены в главе 4.

В **четвертой главе** приводятся результаты получения пороговых значений системой выработки порогового значения и результаты тестирования средства обнаружения файлов, представляющих собой РРСП, на нескольких коллекциях файлов. В главе описываются примеры практического применения полученных автором результатов при решении конкретных прикладных задач, в том числе примеры внедрения результатов в четырех проектах.

В результате обучения системы выработки пороговых значений получены значения $\varepsilon_{1\min}, \varepsilon_{1\max}, \varepsilon_{2\min}, \varepsilon_{2\max}, \varepsilon_3$ для разных размеров фрагментов и параметров масштаба. Обучение проводилось на коллекции файлов разных размеров с псевдослучайными числами. Выработанные пороговые значения использовались для проведения дальнейшего тестирования средства обнаружения файлов, представляющих собой РРСП.

В результате тестирования выявлено следующее:

– вероятность ошибки I рода для разработанного средства обнаружения близка к нулю, так как пороговые значения формировались таким образом, чтобы все файлы из коллекции файлов с псевдослучайными числами были отнесены к классу X_1;

– вероятность ошибки II рода разработанного средства обнаружения при выбранных пороговых значениях превышает соответствующую вероятность для критерия «Хи-квадрат» на равномерность не более, чем на 0,01 (вероятность ошибки II рода обуславливается наличием медиафайлов и файлов-архивов размером до 1 МБ, но может быть снижена путем выбора других пороговых значений, влияющих на вероятность ошибки I рода). Таким образом, можно заключить, что поставленная в главе 2 научная задача решена;

– для сравнения произведено тестирование на тех же коллекциях набора тестов NIST и выявлено, что однозначно трактовать его результаты при применении в файловой системе сложно, и в отличие от тестов NIST разработанное средство обнаружения затрачивает

приемлемое в рамках реагирования на инцидент ИБ время, которое в работе принято в пределах от 4 до 8 часов;

– тестирование программного обеспечения (ПО) для поиска зашифрованных файлов «Passware Kit Forensic» на тех же коллекциях показало вероятность ошибки I рода, достигающую 6% без учета ограничений на размер проверяемого файла, при незначительной вероятности ошибки II рода, связанной с применением сигнатурного анализа;

– тестирование на областях файловой системы, перезаписанных псевдослучайными числами средствами безвозвратного удаления данных «Eraser» и «wipe», показало, что разработанное средство однозначно идентифицирует эти области, как содержащие РРСП.

Таким образом, автором разработан и реализован метод, который позволяет выявлять все имеющиеся в файловой системе файлы и области, представляющие собой РРСП, за приемлемое в рамках реагирования на инциденты ИБ время.

Разработанный метод может использоваться для оценки статистических свойств данных в файловой системе для файлов в явном виде, в свободной области файловой системы, в неразмеченной области электронного носителя информации, в данных сетевого трафика, в данных, перехватываемых системами защиты информации.

Разработанное средство обнаружения файлов, представляющих собой РРСП, может использоваться для поиска файлов и областей с зашифрованной информацией или следов безвозвратного удаления данных при исследовании электронных носителей информации, в том числе в рамках судебной компьютерной экспертизы.

Разработанное средство обнаружения файлов, представляющих собой РРСП, включено в состав криминалистических средств для проведения судебных компьютерных экспертиз экспертами 12 отдела компьютерных экспертиз и психофизических исследований ЭКЦ ГУ МВД России по г. Москве, Лаборатории компьютерной криминалистики и исследования вредоносного кода ООО «Группа информационной безопасности» и ООО «ТРАСТ», что подтверждается соответствующими актами о внедрении.

В рамках внедрения результатов диссертационной работы в ООО «Группа информационной безопасности» проведено сравнительное тестирование средства обнаружения файлов, представляющих собой РРСП, на тридцати объектах исследования, находящихся в Лаборатории компьютерной криминалистики и исследования вредоносного кода для проведения судебной компьютерной экспертизы. В результате тестирования выявлено, что вероятность ошибки II рода в обнаружении искомых файлов составляет около 0,5% от общего числа файлов, большая часть из которых – файлы-архивы высокого уровня сжатия. Кроме этого, выявлено, что разработанное средство обнаруживает зашифрованные модули вредоносных программ (по классификации антивирусного средства «Kaspersky Antivirus»

версии 15.0.2.361): «Trojan-Spy.Win32.Lurk», «Trojan.Win32.Regin», «Trojan-Spy.Win32.Carberp», «Trojan-Downloader.Win32.Andromeda» и зашифрованные конфигурационные файлы «Backdoor.Win32.Zegost».

В рамках внедрения в ООО «ТРАСТ» проведено тестирование средства обнаружения на нескольких образах накопителей на жестких магнитных дисках, созданных с накопителей подозреваемых в компьютерных преступлениях лиц, в результате которого выявлены зашифрованные с помощью ПО «TrueCrypt» и «BestCrypt» файлы. Файлы представляют собой зашифрованные контейнеры, содержащие криминалистически значимые сведения. Время сканирования одного образа составило в среднем 3 часа.

Осуществлено также внедрение результатов работы в образовательный процесс кафедры «Криптология и дискретная математика» НИЯУ МИФИ в виде трех лекций в учебном курсе «Криптографические средства обеспечения информационной безопасности».

В **заключении** приведены основные результаты диссертационной работы.

ОСНОВНЫЕ ВЫВОДЫ И РЕЗУЛЬТАТЫ РАБОТЫ

В работе получены следующие основные выводы и результаты:

1. Проведен анализ существующих способов обнаружения файлов, представляющих собой РРСП, выявлены их недостатки для решения поставленной в работе задачи. Предложена классификация существующих способов обнаружения таких файлов, а также проведено исследование существующих программных средств, которые позволяют выявлять файлы, представляющие собой РРСП. В результате исследования выявлено отсутствие специализированных способов и средств для точного обнаружения искомых файлов, и обоснована необходимость создания нового метода для осуществления поиска файлов, представляющих собой РРСП, в файловой системе, без выявленных недостатков.

2. Для РРСП получено математическое описание статистических свойств, которые могут использоваться для выявления в данных локальных неоднородностей, формализованных граничными значениями рассчитываемых вейвлет-коэффициентов с разными параметрами. На основании теоретических и экспериментальных предпосылок выдвинута гипотеза различимости файлов, представляющих собой РРСП, и файлов других форматов, в том числе сжатых.

3. Сформулировано и обосновано необходимое условие обнаружения файлов, представляющих собой РРСП, основанное на оценке распределения байтов в файле путем интегральной оценки содержимого файла и путем подсчета вейвлет-коэффициентов, значения которых выходят за пределы порогового значения в случае обнаружения локальной неоднородности в распределении оцениваемых данных.

4. Разработан метод обнаружения файлов, представляющих собой РРСП, на основе сформулированного необходимого условия. В результате применения метода файлы в

файловой системе разделяются на два класса: X_1 (файлы, представляющие собой РРСП) и X_2 (файлы, не представляющие собой РРСП). Метод позволяет на основании статистического анализа выявлять локальные неоднородности данных в файлах, что применяется для различения файлов, представляющих собой РРСП, и файлов сжатых форматов. Таким образом, цель диссертационной работы достигнута.

5. Создана архитектура средства обнаружения файлов, представляющих собой РРСП, а также реализовано средство обнаружения таких файлов на основе предложенного метода, которое позволяет получать доступ к содержимому файлов в обход прав на доступ, выставленных в ОС. Также средство обнаружения проводит оценку содержимого отдельных кластеров, а значит может применяться для работы со свободной областью файловой системы и неразмеченной областью электронного носителя информации.

6. Разработанное программное средство обнаружения протестировано на коллекциях зашифрованных файлов, файлов с псевдослучайными числами и нескольких коллекциях файлов распространенных форматов. Результаты тестирования подтверждают, что разработанный метод позволяет выявлять содержащиеся в файловой системе файлы, представляющие собой РРСП, с вероятностью ошибки I рода, близкой к 0, и вероятностью ошибки II, превышающей соответствующую вероятность критерия «Хи-квадрат» на равномерность не более, чем на 0,01. Кроме этого тестирование разработанного средства обнаружения показало, что сканирование файловой системы осуществляется за приемлемое в рамках реагирования на инциденты ИБ время. Таким образом, выдвинутая гипотеза различимости файлов, представляющих собой РРСП, и файлов других форматов доказана экспериментально, а поставленная в главе 2 научная задача решена.

7. Разработанное средство обнаружения файлов, представляющих собой РРСП, используется для поиска файлов и областей с зашифрованной информацией или следов безвозвратного удаления данных при исследовании электронных носителей информации, в том числе в рамках судебной компьютерной экспертизы. Разработанное средство обнаружения включено в состав криминалистических средств для проведения судебных компьютерных экспертиз в 12 отделе ЭКЦ ГУ МВД России по г. Москве, в Лаборатории компьютерной криминалистики и исследования вредоносного кода ООО «Группа информационной безопасности» и в ООО «ТРАСТ», что подтверждено соответствующими актами о внедрении.

8. Результаты проведенного анализа статистических особенностей содержимого файлов с точки зрения их близости к свойствам РРСП внедрены в образовательный процесс кафедры «Криптология и дискретная математика» НИЯУ МИФИ в рамках учебного курса «Криптографические средства обеспечения информационной безопасности», что подтверждено актом о внедрении.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Матвеева В. С., Мамаев А. В. Криминалистический подход к анализу временных атрибутов файлов в операционной системе семейства Microsoft Windows и файловой системе NTFS / В. С. Матвеева, А.В. Мамаев // Безопасность информационных технологий. – 2013 г. – № 1. – С. 114-115. – ‘РИНЦ’.

2. Matveeva V.S., FORENSIC APPROACH TO ANALYSIS OF FILE TIMESTAMPS IN MICROSOFT WINDOWS OPERATING SYSTEMS AND NTFS FILE SYSTEM // eForensics Magazine. – 2013. – № 24.

3. Матвеева В. С. Энтропия и ее использование для решения задач информационной безопасности / В. С. Матвеева // Безопасность информационных технологий. – 2014 г. – № 3. – С. 30-36. – ‘РИНЦ’.

4. Матвеева В. С. Статистические особенности данных, зашифрованных с помощью программных средств криптографической защиты информации, и способ их обнаружения / В. С. Матвеева // Информация и безопасность. – 2015 г. – Т. 18. № 1. – С. 119-122. – ‘РИНЦ’.

5. Матвеева В. С. Критерий оценки содержимого файлов различных форматов на предмет их близости к случайным данным / В. С. Матвеева // Безопасность информационных технологий. – 2015 г. – № 1. – С. 106-108. – ‘РИНЦ’.

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. – ‘РИНЦ’.

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

8. Matveeva V., Epishkina A. Searching for Random Data in File System During Forensic Expertise // V. Matveeva, A. Epishkina // Biosciences Biotechnology Research Asia. – 2015. – V. 12. - № 1. – P. 745-752. – ‘Scopus’.

9. Matveeva V. Assessment of uniformity of byte distribution in a file based on the wavelet transform as an approach to search encrypted data / V. Matveeva // International Journal Electronic Security and Digital Forensics. – 2015. – V. 7. – № 2. – P. 134-146. – ‘Scopus’.

Личный вклад автора в работе, написанной в соавторстве, состоит в следующем: [1] – анализ особенностей подмены временных атрибутов с целью ее выявления в рамках судебных компьютерных экспертиз; [6] – анализ зависимостей выбираемых для вейвлет-анализа параметров от размера файла и выявляемых отклонений; [8] – разработка метода поиска псевдослучайных данных в рамках компьютерной экспертизы.

МАТВЕЕВА ВЕСТА СЕРГЕЕВНА

СТАТИСТИЧЕСКИЙ МЕТОД ОБНАРУЖЕНИЯ ЛОКАЛЬНЫХ НЕОДНОРОДНОСТЕЙ ДАННЫХ ДЛЯ РАССЛЕДОВАНИЯ ИНЦИДЕНТОВ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Подписано в печать ___.___.2015. Формат 60 × 84 $^{1}/_{16}$.
Усл. печ. л. 1,0. Уч.-изд. л. 1,0. Тираж 100 экз. Заказ № ___

Национальный исследовательский ядерный университет «МИФИ» (НИЯУ МИФИ)

115409, Москва, Каширское шоссе, 31