На правах рукописи

Сорокин Иван Витальевич

МАТЕМАТИЧЕСКИЕ МОДЕЛИ И АЛГОРИТМЫ РАСПОЗНАВАНИЯ УПАКОВАННЫХ ВРЕДОНОСНЫХ ПРОГРАММ

Специальность 05.13.19 – Методы и системы защиты информации, информационная безопасность

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург
2013

Работа выполнена на кафедре информатики федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Российский государственный педагогический университет им. А.И.Герцена»

Научный руководитель:

Копыльцов Александр Васильевич - д.т.н., профессор

Официальные оппоненты:

Коробейников Анатолий Григорьевич - д.т.н, профессор, зам. директора по науке, Санкт-Петербургский филиал Института Земного магнетизма, ионосферы и распространения радиоволн им. Н.В. Пушкова РАН

Емельянов Александр Александрович - к.т.н, доцент кафедры «Прикладные информационные технологии», Санкт-Петербургский государственный университет сервиса и экономики

Ведущая организация:

Санкт-Петербургский институт информатики и автоматизации РАН

Защита диссертации состоится «11» декабря 2013 г. в 15:50 на заседании диссертационного совета Д 212.227.05 при Санкт-Петербургском национальном исследовательском университете информационных технологий, механики и оптики (НИУ ИТМО) по адресу: 197101, г. Санкт-Петербург, Кронверский проспект, д. 49.

С диссертацией можно ознакомиться в библиотеке НИУ ИТМО.

Автореферат разослан «8» ноября 2013 г.

Ученый секретарь
диссертационного совета Поляков В. И.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В настоящее время имеется большое количество различных вредоносных программ. Для их обезвреживания разработаны различные подходы (Christodorescu, 2003; Song, 2005; Kruegel, 2006; Stolfo, 2006; Perdisci, 2008; Shafiq, 2008; Rieck, 2008; Lakhotia, 2010). Среди всего многообразия угроз в последние годы существенно увеличилось количество вредоносных программ, в которых используются методы упаковки и запутывания программного кода (Sun и др., 2010; Ugarte-Pedrero и др., 2011; Cesare и др. 2012; Tahan и др. 2012; Naval и др. 2012; Kim и др. 2012; Бабенко и др. 2012; Подловченко и др., 2012). Применение известных методов обезвреживания не позволяет эффективно бороться с такими вредоносными программами. Поэтому актуальным является разработка новых моделей, методов и средств выявления подобных угроз.

Целью работы является повышение качества выявления упакованных вредоносных программ за счет создания новых математических моделей и алгоритмов распознавания, позволяющих уменьшить вычислительные ресурсы антивирусного программного обеспечения. Для достижения цели были поставлены следующие **задачи**:

1) классифицировать методы обнаружения вредоносных программ с учетом использования энтропийного подхода к анализу характеристик файлов для выявления эффективных подходов к представлению и распознаванию образов;

2) разработать модели и методы представления файлов упакованных вредоносных программ, позволяющие создавать компактную форму описания характеристик файлов;

3) разработать методы идентификации и классификации упакованных вредоносных программ, позволяющие сравнивать файлы между собой на основе анализа энтропийных характеристик с минимальным использованием вычислительных ресурсов.

Методы исследования. Для решения поставленных задач использовались методы теории распознавания образов, теории информационной безопасности, теории цифровой обработки сигналов и компьютерной лингвистики.

Объект исследования - упакованные вредоносные программы для операционной системы Microsoft Windows.

Предмет исследования - характеристики файлов упакованных вредоносных программ.

Основные положения, выносимые на защиту:

1) алгоритм сегментации бинарных файлов, основанный на вейвлет-анализе значений энтропии файла в виде ряда, построенного методом скользящего окна;

2) три способа описания характеристик сегментов файлов, учитывающих размер (количество байтов) и энтропию сегмента, частоту появления байтов в сегменте и модель неокогнитрона;

3) подход по использованию векторной модели для классификации упакованных вредоносных программ, с учетом трех способов описания характеристик сегментов файлов;

4) метод идентификации упакованных вредоносных программ, основанный на взвешенном расстоянии редактирования и учитывающий две характеристики сегмента (размер и информационную энтропию).

Новые научные результаты:

1) Предложена классификация методов анализа и распознавания вредоносных программ, использующих энтропийный подход. Классификация позволила выявить недостатки предлагаемых ранее подходов, неспособных в компактной форме представить энтропийные характеристики отдельных вредоносных программ.

2) Разработан алгоритм сегментации бинарных файлов, основанный на дискретном вейвлет-преобразовании с учетом локальных экстремумов вейвлет-коэффициентов на разных масштабах преобразования. Выбор соответствующего типа

преобразования обоснован с точки зрения эффективности использования при анализе вредоносных программ. Предлагаемый алгоритм сегментации позволяет проводить поиск границ итоговых сегментов при минимальных вычислительных ресурсах.

3) Разработано и обосновано использование трех способов описания сегментированных участков файлов. Во-первых, учет размера сегмента и его информационной энтропии. Во-вторых, учет частоты появления различных байтов в сегменте. В-третьих, использование нейросетевой модели неокогнитрон для запоминания расположения байтов в сегменте. Каждый из трех подходов позволяет выявлять характеристики присущие отдельным вредоносным программам.

4) Предложен подход к сравнению файлов между собой, основанный на векторной форме представления с учетом трех способов описания сегментированных участков файлов. Такой подход позволяет использовать аппарат векторной алгебры при классификации вредоносных программ.

5) Разработан метод идентификации упакованных вредоносных программ, основанный на взвешенном расстоянии редактирования для последовательностей, состоящих из элементов (сегментов) с двумя числовыми характеристиками: размер сегмента и его энтропия.

Теоретическая значимость разработанных методов и алгоритмов заключается в том, что они могут использоваться в теории информационной безопасности.

Практическая значимость полученных результатов состоит в том, что они могут использоваться при разработке антивирусного программного обеспечения.

Достоверность результатов работы подтверждается корректным использованием математического аппарата и соответствием теоретических результатов результатам экспериментов, полученных при тестировании на различных выборках упакованных вредоносных программ.

Реализация и внедрение результатов работы. Результаты работы были использованы при разработке антивирусного программного обеспечения компании ООО «Доктор Веб», что подтверждается актом об использовании.

Личный вклад автора. Все основные результаты диссертации получены лично соискателем.

Апробация работы. Результаты работы докладывались и обсуждались на Санкт-Петербургской межрегиональной конференции «Информационная безопасность регионов России (ИБРР)» в 2009 и 2011 гг., на Санкт-Петербургской международной конференции «Региональная информатика (РИ)» в 2010 и 2012 гг. и на 20-ой ежегодной конференции European Institute for Computer Anti-Virus Research (EICAR) в 2011 г.

Публикации. По теме диссертации опубликованы 10 научных работ, из них 3 статьи опубликованы в журналах, рекомендованных ВАК.

Структура и объем работы. Диссертация состоит из введения, шести глав, заключения и списка литературы, включающего 103 наименования. Работа изложена на 123 страницах, содержит 54 рисунка и 10 таблиц.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность выбранной темы, поставлена цель и сформулированы решаемые задачи, определены объект и предмет исследования, перечислены методы, научная новизна, положения, выносимые на защиту, теоретическая и практическая значимость полученных результатов.

В первой главе представлен обзор публикаций в области распознавания вредоносных программ. Все подходы рассмотрены в контексте двух проблем, возникающих при решении задач распознавания образов: представление исходных данных и выбор решающих процедур. В некоторых

случаях сложно провести четкую границу между используемыми алгоритмами. Тем не менее, предложенная классификация позволила сравнить существующие подходы и показала недостаточную изученность в вопросе использования энтропийных характеристик файлов.

Во второй главе описывается система распознавания упакованных вредоносных программ, состоящая из двух основных частей: модуль сегментации файлов и модуль сравнения файлов. В первом модуле осуществляется предварительная обработка и вейвлет-анализ входных файлов. Во втором модуле сравниваются сегментированные файлы между собой с использованием одного из двух подходов либо на основе векторной модели, либо на основе использования взвешенного расстояния редактирования. В последующих главах эти модули рассмотрены более подробно.

В третьей главе рассматривается модуль сегментации входных файлов. Сначала рассматривается алгоритм предварительной обработки файла, а затем вейвлет-анализ и сегментация файла. Этап предварительной обработки заключается в использовании метода скользящего окна, позволяющего представить содержимое файла в виде ряда $Y = \{y_i : i = 1,..., N\}$, где N – количество всех окон, y_i - информационная энтропия подсчитанная в i-ом окне файла (Рис. 1). Подсчет энтропии в i-ом окне осуществляется с учетом частоты появления различных байтов по следующей формуле:

$$y_i = -\sum_{j=1}^{n} p(j) \log_2 p(j) \qquad (1)$$

где $p(j)$ – частота появления j-го байта в i-ом окне, n – количество различных байтов в окне.

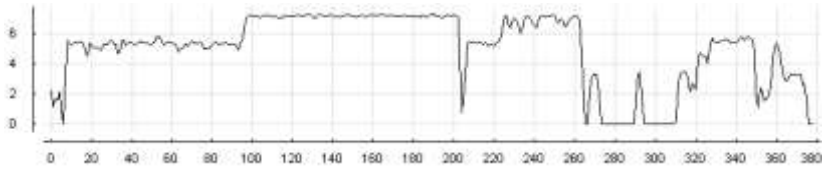

Рис. 1. Энтропия файла (по оси абсцисс - номера окон в файле, по оси ординат - информационная энтропия в окне).

Этап вейвлет-анализа является основным в этом модуле и служит для разбиения ряда на сегменты, обладающие однородной структурой. Для этого предлагается использование дискретного избыточного варианта вейвлет-преобразования, вычисляемого по формуле:

$$W(a,b) = \frac{1}{\sqrt{a}} \sum_{i=b}^{b+a} y_i \cdot \psi_{HAAR}\left(\frac{i-b}{a}\right), \qquad (2)$$

где a – параметр масштаба, b – параметр сдвига, y_i – информационная энтропия в i-ом окне. В качестве базисного вейвлета используется вейвлет Хаара ψ_{HAAR}, обладающий несимметричной формой и нулевым моментом равным нулю. Изменение параметра сдвига b осуществляется последовательно от 1 до N. Это приводит к избыточности вычислений, которая позволяет решать проблему инвариантности при сдвиге исходных данных. Параметр масштаба a изменяется, как 2 в степени от 1 до 5: 2, 4, 8, 16, 32. Благодаря этому появляется возможность на каждом следующем масштабе преобразования использовать значения коэффициентов, вычисленные на предыдущем шаге. Вейвлет-коэффициенты $W(a,b)$ используются для анализа и сегментации анализируемого ряда (Рис. 2).

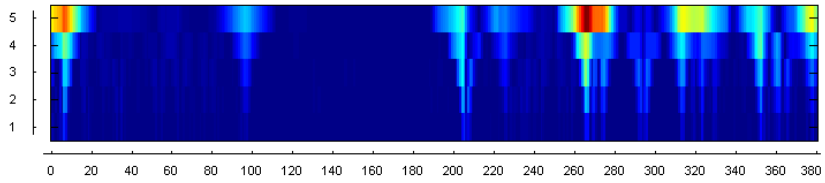

Рис. 2. Вейвлет-коэффициенты дискретного избыточного вейвлет-преобразования (по оси абсцисс – номера членов ряда, по оси ординат - масштаб преобразования, яркость точек характеризует значения вейвлет-коэффициентов).

Суть алгоритма сегментации заключается в поиске значений локальных экстремумов на каждом шаге преобразования. Границы итоговых сегментов определяются найденными локальными экстремумами (Рис. 3).

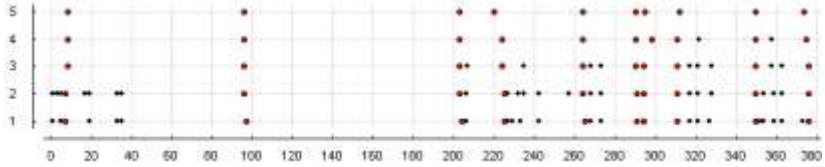

Рис. 3. Точки локальных экстремумов вейвлет-коэффициентов (по оси абсцисс – номера членов ряда, по оси ординат - масштаб преобразования, выделенные точки - локальные экстремумы, определяющие границы сегментов).

В итоге, на выходе модуля сегментации файлов получается последовательность сегментов, каждый из которых характеризуется определенной длиной (количество байтов). Дальнейшее использование характеристик сегментов зависит от выбранной модели сравнения: либо векторной модели, либо модели основанной на расстоянии редактирования Левенштейна.

В четвертой главе для сравнения вредоносных программ описывается векторная модель, при которой каждый компонент вектора обозначает характерный сегмент или набор сегментов файла. Для того чтобы использовать такую

векторную модель необходимо решить вопрос определения всех возможных сегментов, которые могут встречаться в файлах упакованных вредоносных программ. Для уменьшения общего количества сегментов предлагается выявлять похожие сегменты в разных файлах и выделять их в одну группу. В качестве критериев для выявления похожих сегментов предлагается использовать один из трех подходов, которые учитывают: либо размер и энтропию сегмента, либо частоту появления различных байтов в сегменте, либо нейросетевую модель (Рис. 4).

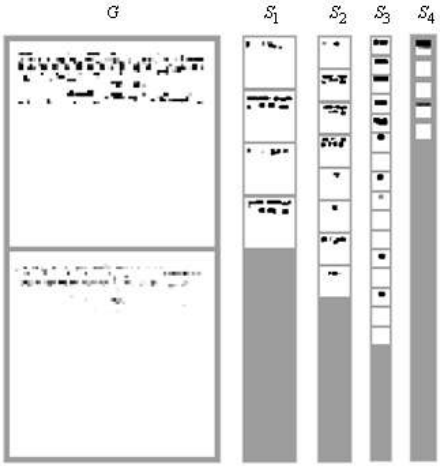

Рис. 4. Активация нейронов в слоях (G, $S_1 - S_4$) неокогнитрона на предъявленный образ сегмента.

Из трех перечисленных подходов, наиболее трудоемким является последний, предполагающий использование нейронной сети неокогнитрон. В этом подходе на вход сети подается изображение, интенсивность окраски точек которого определяется значением (от 0 до 255) байтов в сегменте. Нейронная сеть обучается таким образом, чтобы осуществлялось запоминание взаимного расположения байтов в отдельном сегменте.

В пятой главе предлагается модель сравнения файлов, в которой файлы описываются последовательностью сегментов, а близость файлов между собой определяется с использованием взвешенного расстояния редактирования, основанного на метрике Левенштейна. Ключевым моментом в предлагаемом подходе является функция стоимости g, которая определяет величину штрафа за несовпадение двух сегментов, каждый из которых характеризуется двумя величинами (размером и энтропией). Обозначим размеры (количество байтов) двух сегментов через l_1 и l_2, а соответствующие значения энтропии – через h_1 и h_2. Тогда общий штраф между двумя сегментами вычисляется следующим образом:

$$\mathrm{g} = \alpha \cdot \frac{|l_1 - l_2|}{l_1 + l_2} + (1-\alpha) \cdot \left(\frac{1}{1+\exp(\beta \cdot |h_1 - h_2| + \gamma)} - \delta \right), \qquad (3)$$

где первое слагаемое обозначает штраф за разницу в размерах, а второе - за разницу в энтропии двух сегментов. Коэффициенты $\alpha = 0.3$, $\beta = -4.0$, $\gamma = 6.5$, $\delta = 0.001501$ подобраны экспериментальным путем.

Для сравнения двух вредоносных программ, с учетом выражения (3) для функции стоимости g, предлагается алгоритм, основанный на глобальном выравнивании последовательностей. Алгоритм заключается в заполнении двумерного массива $C[i][j]$ следующим образом:

$$\begin{aligned}
&C[0][0] = 0,\\
&C[i][0] = C[i-1][0] + \alpha \cdot \lg(l_{A,i-1}),\\
&C[0][j] = C[0][j-1] + \alpha \cdot \lg(l_{B,j-1}),\\
&C[i+1][j+1] = \min \begin{cases} C[i][j] + g(A_i, B_j) \cdot \lg((l_{A,i} + l_{B,j})/2) \\ C[i][j+1] + \alpha \cdot \lg(l_{A,i}) \\ C[i+1][j] + \alpha \cdot \lg(l_{B,j}) \end{cases},
\end{aligned} \qquad (4)$$

где коэффициент $\alpha = 0.3$ подобран экспериментальным путем, $1 \le i \le |A|$ и $1 \le j \le |B|$ - индексы сегментов двух сравниваемых последовательностей A и B, количество сегментов в которых равно $|A|$ и $|B|$ соответственно, $l_{A,i}$ и $l_{B,j}$ - размеры сегментов A_i и B_j соответственно. На каждом шаге заполнения массива $C[i][j]$ осуществляется одна из трех операций преобразования: замещение, удаление или вставка сегмента файла. Алгоритм позволяет определить общий штраф при сравнении последовательностей сегментов двух файлов и произвести их выравнивание, т.е. найти похожие сегменты в файлах. Также появляется возможность определить степень похожести двух файлов (выраженную в процентах), вычисляемую как доля фактического штрафа от максимально возможного штрафа за несовпадение двух файлов.

В шестой главе рассматривается эффективность предложенных подходов для выявления упакованных вредоносных программ. Сначала оценивается количество операций при реализации предлагаемых алгоритмов, а затем приводятся результаты тестирования системы распознавания упакованных вредоносных программ. Для тестирования составлены выборки пяти различных семейств упакованных вредоносных программ (443 файла) и набор чистых (не вредоносных) программ (100 файлов). Под семейством подразумевается совокупность вредоносных программ, относящихся к одной из пяти угроз: Backdoor.Tdss (100 файлов), BackDoor.Maxplus (ZeroAccess) (63 файла), BackDoor.Butter (Mariposa) (100 файлов), Trojan.Mayachok (Cidox) (80 файлов), BackDoor.Slym (Kelihos) (100 файлов).

Результаты тестирования показывают, что с помощью алгоритма, построенного на основе расстояния редактирования, можно отделять упакованные вредоносные

программы от чистых программ и распознавать отдельные семейства вредоносных программ между собой (Рис. 5).

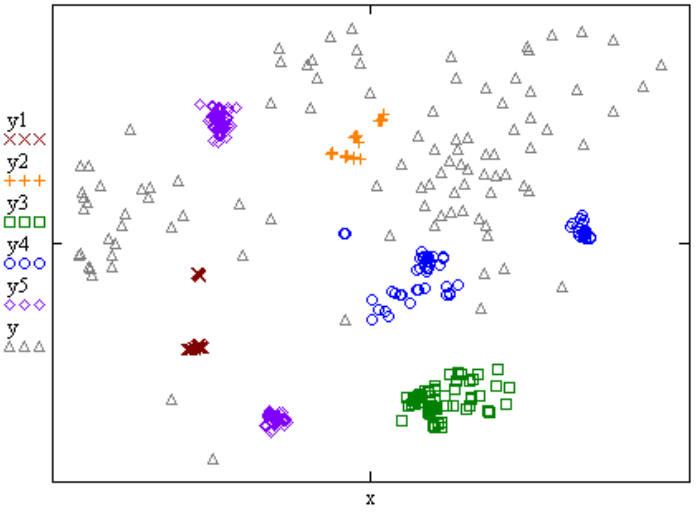

Рис. 5. Взаимное расположение пяти тестовых выборок вредоносных программ (y1-y5) и набора чистых программ (y), полученное методом многомерного шкалирования с весовыми коэффициентами. По осям x и y – расстояния.

Сравнение алгоритма, основанного на взвешенном расстоянии редактирования, с двумя другими подходами (сигнатурно-байтовый и модель поведения под эмулятором), используемыми в антивирусных программных продуктах ООО «Доктор Веб», показало, что при распознавании большинства тестовых файлов (373 из 443) предлагаемый алгоритм показывает лучшие характеристики. Во-первых, лучше устойчивость к защитным методам обфускации и антиэмуляции, чем у детектирования по модели поведения. Во-вторых, при отсутствии ошибок 1-го рода, нормированный уровень ошибок 2-го рода составил 15,8%, что меньше, чем у сигнатурного подхода с 26,4%. Для сравнения подходов была создана база данных антивирусных записей, содержащая

информацию о вредоносных программах. В качестве показателя качества использовалось минимальное количество записей необходимых для детектирования вредоносных программ. Такая эмпирическая оценка характеризует обобщающую способность распознавания. Чем меньше количество записей необходимых для тестовой выборки, тем больше вероятность обнаружить вредоносную программу, не вошедшую в тестовый набор. В следующей таблице показаны результаты сравнения для детектирования 100 вредоносных программ Backdoor.Tdss (Рис. 5, y1).

Способ обнаружения	**Количество файлов детектируемых одной из 6 антивирусных записей**						**Итого записей**
	1	**2**	**3**	**4**	**5**	**6**	
Сигнатурно-байтовый	9	9	12	13	28	29	6
Модель поведения	9	15	35	41			4
Энтропийный	50	50					2

Из результатов проведенного эксперимента видно, что предлагаемый подход для распознавания упакованных вредоносных программ, требует меньшее количество антивирусных записей, чем другие подходы. В некоторых случаях, предлагаемый подход позволяет существенно сократить объем антивирусной базы данных. Например, для детектирования 63 вредоносных программ BackDoor.Maxplus (Рис.5, y2), собранных из одного источника за 5 дней, требуется 63 сигнатурных записи (защита от выявления по модели поведения), занимающих в базе 974 байта. При использовании предлагаемого подхода понадобится 3 энтропийные записи, которые займут 190 байтов. Уменьшение количества антивирусных записей уменьшает не только объем базы данных, но и время обработки проверяемых файлов. Все это подтверждает эффективность предлагаемого подхода для распознавания упакованных вредоносных программ.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1) Предложена классификация подходов по распознаванию упакованных вредоносных программ. Сравнительный анализ позволил выявить недостатки подходов по использованию энтропийных характеристик файлов.

2) Разработан алгоритм сегментации файлов, основанный на вейвлет-анализе энтропийных характеристик файлов. Используемый способ вычисления коэффициентов вейвлет-преобразования позволяет проводить поиск границ итоговых сегментов при минимальных числовых расчетах.

3) Предложено три способа описания сегментов файлов. В первом случае учитывается размер (количество байтов) и энтропия сегмента файла. Во втором случае используется частота появления байтов в сегменте. В третьем - модель неокогнитрона, позволяющая запоминать расположение отдельных байтов внутри сегмента файла.

4) Предложен подход по использованию векторной модели для представления и сравнения файлов между собой, в которой компоненты вектора определяются одним из трех способов описания сегментированных участков файлов.

5) Разработан метод идентификации упакованных вредоносных программ, основанный на взвешенном расстоянии редактирования для последовательностей сегментов, характеризующиеся двумя значениями: размером и информационной энтропией.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в журналах, рекомендованных ВАК:

1. Копыльцов А.В., Сорокин И.В. Вейвлет-анализ структурной энтропии файлов // Известия РГПУ им. А.И.Герцена. - 2011. - № 138. - С. 7-15.

2. Копыльцов А.В., Сорокин И.В. Алгоритм выравнивания последовательностей сегментов файлов // Вестник

ИНЖЭКОНа, серия “Технические науки”. - 2011. - № 8(51), - С. 31-37.

3. Сорокин И.В., Копыльцов А.В. Использование неокогнитрона для распознавания вредоносных файлов // Известия СПбГЭТУ «ЛЭТИ». - 2013. - № 3. - С. 45-51.

Публикации в других изданиях:

4. Копыльцов А.В., Сорокин И.В. Энтропийный анализ файлов для выявления и классификации вредоносного программного обеспечения // Материалы конференции «Информационная безопасность регионов России 2009». - СПб., 2009. - С. 59-60.

5. Сорокин И.В., Копыльцов А.В. Основные методы выявления зомби-сетей // Материалы конференции «Информационная безопасность регионов России 2009». - СПб., 2009. - С. 140.

6. Копыльцов А.В., Сорокин И.В. Энтропийный анализ файлов для выявления и классификации вредоносного программного обеспечения // Труды конференции «Информационная безопасность регионов России 2009». - СПб., 2010. - С. 135-140.

7. Сорокин И.В., Копыльцов А.В. Сегментация временных рядов с использованием вейвлетов для анализа структурной энтропии вредоносных файлов // Материалы конференции «Региональная информатика 2010». - СПб., 2010. - С. 143.

8. Сорокин И.В., Копыльцов А.В. Моделирование структуры вредоносных файлов с использованием стохастических грамматик // Материалы конференции «Информационная безопасность регионов России 2011». - СПб., 2011. - С. 131.

9. Sorokin I. Comparing files using structural entropy // Journal in computer virology. - 2011. - Vol. 7. - N 4. - pp. 259-265.

10. Сорокин И.В. Описание файлов для распознавания вредоносных программ // Материалы конференции «Региональная информатика 2012». - СПб., 2012. - С. 127-128.