Наталия Кошевая, Наталья Мазниченко
(Харьков, Украина)

ЗАЩИТА АВТОРСКИХ ПРАВ МЕТОДАМИ ТЕКСТОВОЙ СТЕГАНОГРАФИИ

*Введение и постановка задачи***.** Бурное развитие информационных технологий, которое наблюдается в последние годы, привело к тому, что сегодня огромное количество информации, составляющей интеллектуальную собственность, хранится и обрабатывается в компьютерных сетях и/или распространяется в цифровой форме.

Наиболее распространенными нарушениями прав интеллектуальной собственности сегодня являются пиратство, плагиат, подделка информации, изменение информации, недобросовестная конкуренция (промышленный шпионаж и т. п.).

При этом наибольшее внимание уделяется защите прав интеллектуальной собственности мультимедийной информации, распространяемой на цифровых носителях и в сети Интернет, однако упор делается больше на правовое решение проблемы, технические вопросы остаются на втором плане. В то же время нельзя забывать о том, что огромное количество информации представлено в обычном текстовом виде: книги, статьи, электронная переписка, документы, отчеты и многое другое. Причем в области электронного документооборота технические вопросы защиты интеллектуальной собственности не могут быть полностью решены только лишь стандартными средствами защиты информации.

Среди задач, решаемых в рамках систем защиты, особое место занимает задача специального кодирования информации в виде данных, предназначенных для скрытой передачи информации, называемая задачей стеганографии. Построение стеганографических методов привлекает внимание многих специалистов, занятых разработкой новых технологий (например, технологий анализа и фильтрации передаваемой информации в сети), направленных на обеспечение высокой надежности информационных систем. В целом задача стеганографии и противоположная ей задача стегоанализа являются одними из базовых проблем в теории надежности и безопасности информационных технологий. В отличие от криптографии, ограничивающей доступ к информации, содержащейся в передаваемом сообщении с помощью некоторого секретного ключа, задача стеганографии состоит в том, чтобы скрыть сам факт передачи какого-либо сообщения от третьих лиц. Обычно, такая задача решается путем внедрения передаваемого секретного сообщения в безобидный на вид объект данных, так называемый контейнер. Сам контейнер подбирается таким образом, чтобы факты его существования или передачи не вызывали никакого подозрения. Основными характеристиками методов стеганографии следует считать объем внедряемого сообщения и устойчивость к анализу (обнаружению факта наличия внедрения).

В цифровой стеганографии в качестве контейнера используется цифровой объект – компьютерный файл. Современные методы встраивания позволяют внедрять скрытую информацию в файлы аудио, видео, текста, исполняемых программ и т.д. В настоящее время

существует большое количество стеганографических программных пакетов как коммерческих, так и бесплатных, с графическим интерфейсом и в виде консольных приложений.

Цифровая стеганография получила широкое применение в сфере защиты авторских прав. В объект авторского права может быть внедрена специальная метка – отпечаток пальца (fingerprint), которая идентифицирует законного получателя. Например, в каждую продаваемую копию программы может быть внедрена метка, идентифицирующая лицензионного покупателя. В случае обнаружения пиратской копии программы при помощи встроенной метки без труда может быть отслежен пользователь, нарушивший лицензионное соглашение. Еще одной встраиваемой меткой может быть цифровой водяной знак (ЦВЗ, watermark), идентифицирующий автора. Менее проработанным является вопрос защиты текстовой информации при помощи внедрения ЦВЗ. В литературе можно встретить описание синтаксических и семантических методов внедрения информации, однако отсутствует их адаптация для внедрения ЦВЗ.

Следует отметить, что защиты требует только изображение информации, представленное на материальном носителе, причем под ней понимается создание условий, исключающих либо затрудняющих доступ к носителю, внесение изменений или уничтожение носителя, а также восприятие представленных на нем данных, производимое с помощью методов криптографии и стеганографии. И если, образно говоря, криптография делает понятное непонятным, то стеганография делает видимое невидимым (иногда и в прямом смысле слова). Достигается это «растворением» скрываемой информации среди других данных значительно большего объема.

Скрываемая информация называется стеганограммой или просто стего. Данные, среди которых она прячется, играют роль информационного контейнера, а потому так же и именуются. Одна и та же стеганограмма может быть упакована в различные контейнеры подобно тому, как одна и та же криптограмма шифруется различными методами или ключами. Тем не менее, некоторые авторы не склонны придавать значение собственной информации контейнера, считая ее безразличной как для отправителя, так и для получателя стегосообщения [1, с. 26].

Контейнером могут служить любые данные (файлы) достаточно большого объема, например графические или звуковые. Их структура проста и, как правило, обладает большой избыточностью, позволяющей вместить значительный объем дополнительной информации. Однако текстовые файлы все же более распространены, и их структура широко известна. Стеганография, использующая текстовые контейнеры, называется текстовой (text steganography).

В то же время, вопросам текстовой стеганографии посвящено сравнительно мало работ. Авторами проведен анализ основных отечественных и зарубежных источников за более чем 10 последних лет. Целью работы является обзор методов и алгоритмов текстовой стеганографии, применяющихся в сфере защиты авторских прав, анализ надежности и безопасности использования информационных технологий, базирующихся на этих методах.

Основная часть. На сегодняшний день существует множество способов встраивания скрытой информации в тестовые файлы. Их можно условно разделить на следующие группы: синтаксические методы, лексические методы (лингвистическая стеганография) и мимикрия (mimic-function — методы имитирующих функций).

Синтаксические методы основаны на использовании особенностей пунктуации, аббревиатуры и сокращения. Хотя правила пунктуации достаточно строго оговорены правилами используемого языка, существуют случаи, когда эти правила оказываются неоднозначными или же отклонение от них не ведет к существенному искажению смысла скрывающего текста. К синтаксическим методам относят также методы, основанные на изменении стиля и структуры предложения без заметного искажения исходной смысловой нагрузки.

При использовании синтаксических методов в текстовых файлах секретная информация чаще всего кодируется путем изменения количества пробелов, использования невидимых символов, регистра букв, путем изменения межстрочных интервалов, табуляций и т д. Синтаксические конструкции легко встраиваются в любой текст, независимо от его содержания, назначения и языка. Такие системы легко разрабатывать и выполняются они автоматически. Но они легко взламываются и секретная информация легко устраняется путем простейших атак.

В рукописном тексте написание отдельных символов может заметно варьироваться, помимо явных различий в начертании символов, может отличаться высота букв, их ширина, высота средней линии, угол наклона и т. д. Все это может эффективно использоваться для передачи скрытых посланий. Основная сложность методов, основанных на использовании особенностей символов, заключается лишь в формировании правил различения буквы открытого текста от аналогичной буквы скрытого сообщения. В простейшем случае, возле отдельных букв могут встречаться "случайные" точки или едва заметные подчеркивания. Поскольку символы текста в электронном виде идентичны, для целей цифровой стеганографии данный подход малоприменим.

В основу методов кодирования смещением строк положено изменение интервала между строками сообщения. Каждая строка маскирующего текста сдвигается немного вверх или вниз относительно своего исходного положения (базовой линии), соответственно смещением строки вверх можно закодировать, например, единицу, а вниз ноль очередного двоичного символа скрываемого сообщения. Так же может использоваться и сам межстрочный интервал. Метод достаточно часто применяется для целей скрытой маркировки твердых копий электронных документов при печати на сетевых принтерах.

Кодирование с использованием изменения горизонтального интервала между отдельными словами или символами наиболее эффективно при выборе в качестве маскирующего сообщения больших текстов с выравниванием по ширине, так как в данном случае расстояние между словами может меняться в достаточно широких пределах. В ряде случаев применяется кодирование не только длиной символов пробела, но и их числом. Так, два пробела в интервале между предложениями могут кодировать очередной двоичный символ скрытого сообщения со значением, равным единице, а один – со значением нуля. Аналогично могут быть использованы пробельные символы в конце строки.

Число автоматических методов текстовой стеганографии, естественно, не ограничивается рассмотренными примерами. Пополнить запас примеров можно, в частности, разумной комбинацией уже приведенных.

К недостаткам представленных методов следует отнести высокую вероятность разрушения скрытого сообщения при повторном наборе текста или использовании более сложных текстовых редакторов, способных осуществлять ряд автоматических операций над текстом. Такие операции, как форматирование, замена символов табуляции пробелами, удаление лишних пробелов в конце строк и т.д., приведут к порче или же полному уничтожению скрытого сообщения. Значительно большей стойкостью к подобным искажениям обладают методы, оперирующие непосредственно самим текстом, отдельными его предложениями и словами.

Лингвистическая стеганография **(*лексические или семантические методы*),** предполагает использование семантических особенностей языка. Данный подход отличается высокой эффективностью, обусловленной применением различных методов манипулирования непосредственно самими предложениями и словами, а не второстепенными элементами и незначительными особенностями текстов. Ряд методов, относящихся к данному направлению, основан на использовании синонимов. Практически в любом достаточно длинном предложении встречаются слова, которые без потери смысла могут быть заменены синонимами. Если для некоторого слова существует набор более чем из одного синонима, то возможно формирование специальных таблиц замен. В таких таблицах каждому синониму может быть поставлено в соответствие некоторое кодовое

слово, состоящее более чем из одного двоичного символа. Однако необходимо отметить, что в ряде случаев использование методов осложнено определенными нюансами и оттенками ключевых слов в предложениях, что несколько ограничивает их применение.

Рассматривая работы зарубежных специалистов, посвященные лингвистической стеганографии [2, с. 11], можно заметить, что авторы этих работ достаточно четко разграничивают методы и алгоритмы лингвистической стеганографии по защите скрываемой информации от «роботов» (программ автоматического сканирования и анализа текстов) и от людей. Первые направлены на защиту информации при тотальном сканировании всей корреспонденции программными поисковыми роботами. Вторые направлены на защиту информации при внимательном просмотре текста человеком. Не вызывает удивления тот факт, что публикаций и работ, посвященных первому направлению на порядок больше работ, посвященных второму направлению (защите от анализа человеком). Поисковые роботы ищут ключевые слова, фразы, какие-то явные особенности текста. В результате, робота, который не силен в грамматике и не понимает смысла и явного подтекста передаваемых сообщений, обмануть гораздо проще. Задача же скрытой передачи информации в тексте, нацеленная на защиту от анализа передаваемого сообщения человеком, очевидно на порядок сложнее. Поэтому необходимо разработать и реализовать методы скрытой передачи коротких сообщений, использующих в качестве контейнеров текстовые файлы и при этом обеспечить защиту как от визуального анализа, проводимого человеком, так и от статистического анализа, который может быть проведен роботами. Решить поставленную задачу для текстов, например, на русском (а тем более на украинском) языке в действительности значительно сложнее, нежели для текстов на английском языке. Здесь можно выделить два основных фактора, приводящих к усложнению задачи [3, с. 58]. Первым из них является неоднозначное использование слов в русском языке. В различном контексте одни и те же слова могут нести различную смысловую нагрузку. Вторым фактором является широкое использование в русском языке большого количества окончаний слов. Если при построении стеганографической системы не учитывать хотя бы один из этих факторов, результирующий текст будет носить явно несогласованный характер, что является очевидным демаскирующим признаком. Принцип работы базового метода прост. Довольно часто в тексте одно слово может быть заменено другим словом, которое является синонимом исходного слова. В качестве примера можно привести два предложения, несущих одинаковую смысловую нагрузку: «На улице сейчас прекрасная погода» и «На улице сейчас замечательная погода». Так как предложения несут одинаковую смысловую нагрузку, то использование их в тексте эквивалентно. Для того чтобы передать скрытое сообщение первому предложению, мы можем поставить в соответствие двоичный «0», второму – двоичную «1» скрываемого сообщения. Как видно из представленного примера, использование стеганографического метода, основанного на замене синонимов, позволяет сохранить синтаксическую структуру предложения и его смысловую нагрузку. Такую замену слов достаточно легко проделать человеку. В то же время этот метод нельзя реализовать простым машинным алгоритмом, даже если не учитывать необходимость подстановки окончаний и согласования слов.

Методы внедрения, основанные на семантических особенностях текста, являются трудно обнаружимыми. Замена одного слова на соответствующий ему синоним не нарушает синтаксическую структуру предложения и не искажает смысловое содержание. Несмотря на указанную особенность, такой метод внедрения также не лишен недостатков. При замене некоторых слов возможно нарушение стиля языка. Например, во фразе “what time is it?” слово time может быть заменено на синоним duration, но это будет некорректно для английского языка. Также использование некоторых слов в качестве синонимов может нарушать авторский стиль написания текста. На этих фактах базируются многие методы анализа.

Мимикрия. Методы использования имитирующих функций (mimic-function). Метод основан на генерации текстов и является обобщением акростиха. Для тайного сообщения генерируется осмысленный текст, скрывающий само сообщение [4].

Для получения стеготекста используются контекстно-свободные грамматики. Нетерминальные символы могут быть раскрыты по заданным правилам несколькими возможными способами. В зависимости от входного сообщения выбирается правило раскрытия. Сгенерированный стеготекст не содержит грамматических и орфографических ошибок. На сегодняшний день самыми популярными программами, генерирующими искусственный текст, являются Nicetext, Texto и Markov-Chain-Based. Эти программы имеют высокое соотношение размера входного сообщения к размеру генерируемого текста, и получающийся текст максимально похож на естественный. Стоит отметить, что получившийся искусственный текст, как правило, является бессмысленным.

Устойчивость методов, генерирующих стеготекст, подобный естественному, обеспечивается заданными правилами грамматики. Отсутствие грамматических и орфографических ошибок в предложениях делает затруднительным поиск отличий искусственного текста от естественного. Анализ осмысленности текста можно производить только с участием человека, что не всегда возможно из-за огромного объема анализируемой информации. Наиболее эффективный метод анализа использует прогнозирование для выявления искусственной природы текста, порожденного программой Nicetext. Сначала производится анализ слов первой половины текста, и составляется прогноз каждого последующего слова из второй части текста. Если в подавляющем большинстве случаев прогноз оказывается успешным, то это означает, что мы имеем дело с естественным текстом. Частые ошибки при прогнозировании могут свидетельствовать о наличии искусственного текста. Для программ Texto и Markov-Chain-Based используются методы, учитывающие корреляцию слов между предложениями. Так, считается, что предложения, содержащие слова, встречающиеся только в технических текстах, не могут стоять рядом с предложениями, содержащими слова, встречающиеся только в текстах художественной литературы.

Достоинством метода является то, что результирующий текст не является подозрительным для систем мониторинга. К недостаткам можно отнести слабую производительность метода, передачу небольших объемов информации и. низкую степень скрытности сети.

Заключение. Итак, были рассмотрены различные методы обеспечения безопасности использования информационных технологий за счет встраивания скрытой информации в тестовые файлы, каждый из которых имеет свои преимущества и недостатки. На основе предложенного анализа каждый отдельный пользователь может сделать самостоятельный обоснованный выбор приемлемого метода в зависимости от круга решаемых задач. Предложенные алгоритмы, базирующиеся на методах цифровой стеганографии, могут быть использованы, например, для защиты авторских прав собственников и пользователей текстовой информации, представленной в цифровой форме; для анализа и фильтрации передаваемого трафика в сети; с целью пресечения утечки коммерческой информации предприятия; построения систем защиты авторских прав.

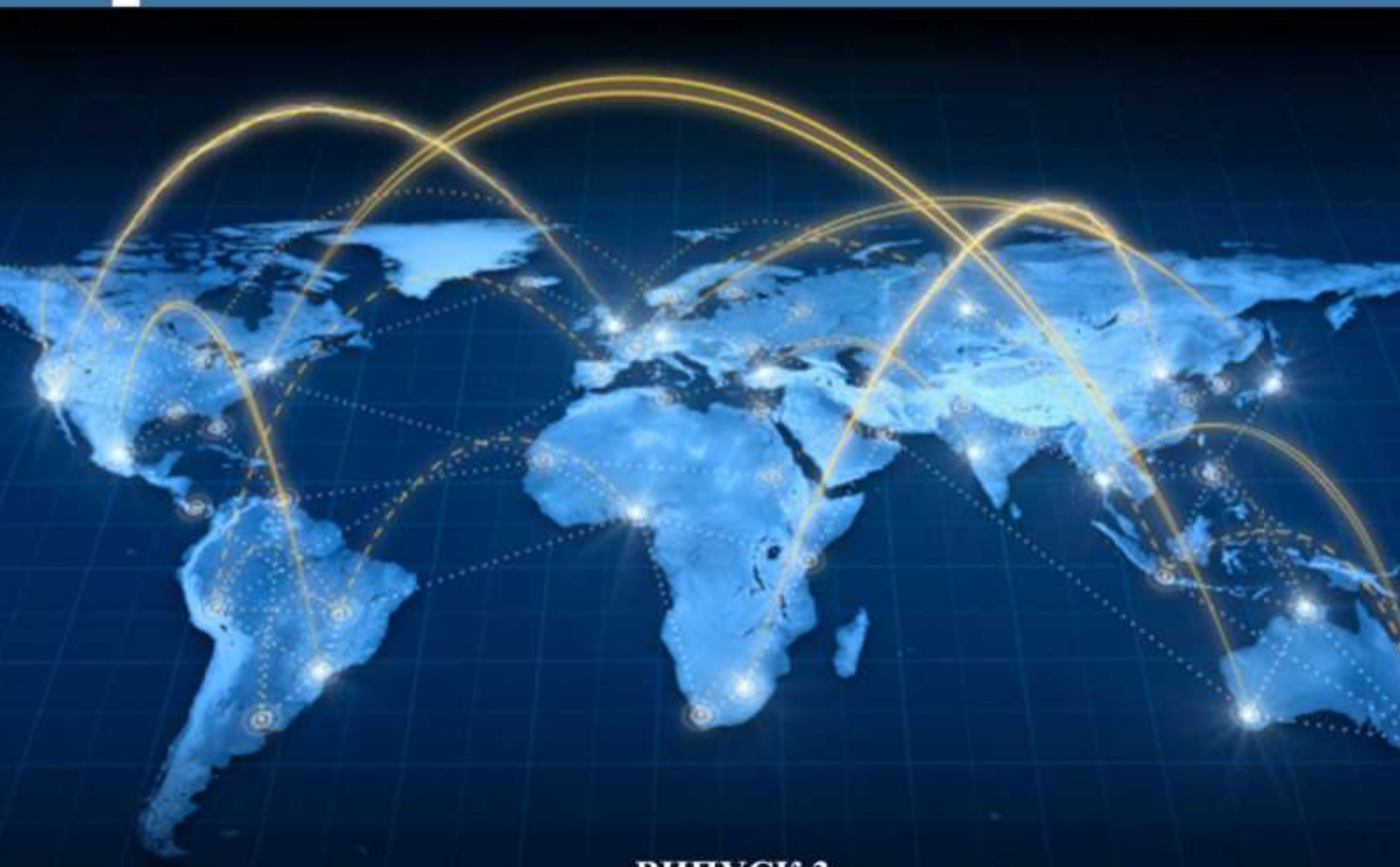
ИСТОЧНИКИ И ЛИТЕРАТУРА:

1. Стеганография, цифровые водяные знаки и стеганоанализ: Монография / А.В. Аграновский, А.В. Балакин, В.Г. Грибунин, С.А. Сапожников. – М.: Вузовская книга, 2009. – 220 с.
2. Bennett K. Linguistic Steganography: survey, analysis, and robustness concerns for hiding information in text, Center for Education and Research in Information Assurance and Security, CERIAS Tech Report 2004-13. – 30 p.
3. Рябко Б.Я., Фионов А.Н. Основы современной криптографии и стеганографии. – М.: Горячая линия – Телеком, 2010. – 232 с.

4. Метод кодирования произвольной двоичной информации на основе лингвистических ресурсов. Ларионова К.Е., Губенко Н.Е. [Электронный ресурс]. – Режим доступа: http://masters.donntu.org/2009/fvti/larionova/library/article11.htm

III МІЖНАРОДНА НАУКОВО-ПРАКТИЧНА
ІНТЕРНЕТ-КОНФЕРЕНЦІЯ

«ТЕНДЕНЦІЇ ТА ПЕРСПЕКТИВИ РОЗВИТКУ НАУКИ І ОСВІТИ В УМОВАХ ГЛОБАЛІЗАЦІЇ»



ВИПУСК 3

2-3 травня 2015 р.

м. Переяслав-Хмельницький

ДЕРЖАВНИЙ ВИЩИЙ НАВЧАЛЬНИЙ ЗАКЛАД
«Переяслав-Хмельницький державний педагогічний університет
імені Григорія Сковороди»

Рада молодих учених університету

Матеріали
ІІІ Міжнародної науково-практичної інтернет-конференції
«Тенденції та перспективи розвитку науки і освіти в умовах глобалізації»

2-3 травня 2015 року

Збірник наукових праць

Переяслав-Хмельницький – 2015