Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК _____

Павлович
Антон Николаевич

Использование машинного обучения для сбора информации о пользователе

АВТОРЕФЕРАТ
на соискание степени магистра информатики
и вычислительной техники

по специальности 1-40 81 04 – «Обработка больших объемов информации»

_____

Научный руководитель
Сиротко Сергей Иванович
к.т.н., доцент

_____

Минск 2017

КРАТКОЕ ВВЕДЕНИЕ

Электронные устройства широко проникли в жизнь людей. Их используют для хранения личной информации и конфиденциальных данных, например, фотографий, паролей и сообщений. Часто данные шифруют, чтобы предотвратить неавторизованный доступ. Однако нет способа защитить данные от кражи до шифрования. Так с помощью перехвата физического сигнала (звука или электромагнитных волн) можно определить либо исходные текстовые данные перед шифрованием (в процессе набора текста) или ключи, которые вводит пользователь для шифрования или расшифрования. История анализа косвенной информации началась в 1918, когда Герберт Ярдли (Herbert Yardley), на тот момент глава отдела MI-8 (США), открыл, что различные электронные устройства для обработки секретной информации имеют побочные излучения, и эти излучения можно использовать для восстановления секретных данных.

Общей целью для атак этого типа являются устройства ввода-вывода, такие как клавиатуры, мыши, сенсорные экраны и принтеры. Примеры атак включают: электромагнитное излучение клавиатур, видео набора текста пользователем на клавиатуре или сенсорном экране, звук набора на клавиатуре. Исследователи приложили много усилий для иследования звука набора и продемонстрировали, что это серьёзная проблема. Успешная атака на этот, казалось бы, второстепенный канал – звук произведённый клавишами – позволит противнику узнать, какой текст набирала жертва. Обычно звук записывали или напрямую с помощью микрофонов, или использовали в своих интересах другие датчики (акселерометры) с целью восстановить исходную акустическую информацию. Собранный аудиопоток затем анализировался с помощью техник машинного обучения с учителем или без, а также триангуляции. Результатом получался частично или полностью восстановленный ввод жертвы.

Похоже, все предыдущие атаки использовали скомпрометированный (контролируемый атакующим) микрофон возле клавиатуры жертвы. Такое требование сильно ограничивает применимость атаки в реальном мире. Возможно, набравшие большую популярность и повсеместно распространённые смартфоны могли бы стать таким компрометирующим устройством (например, если атакующий положит свой смартфон рядом с жертвой), однако атакующему всё ещё будет необходимо контролировать положение смартфона возле жертвы. Кроме того, некоторые работы предполагают даже более строгие ограничения: больше информации для обучения кластера, а это означает большой объём собранных звукозаписей или профилирование стиля набора жертвы и клавиатуры.

В данной работе описывается другой вид атаки с перехватом звука клавиатурного набора, который не требует от атакующего ни рядом расположенного контролируемого микрофона, ни большого объёма данных для анализа. Базовый случай, на котором моделируется атака, – это VoIP-разговор в Skype, самом популярной программе для голосового общения в мире. Модель построена на идее, что во время разговора в Skype человек может заниматься посторонними вещами: написанием электронных писем, чтением новостей, просмотром видео, обновлением статуса в социальных сетях и даже написанием научной работы. Большинство этих второстепенных занятий потребуют активного взаимодействия с клавиатурой (например, ввод пароля от социальной сети). Жертва набирает пароль, VoIP-программа передаёт звук набора атакующему, который затем сможет определить, что набирала жертва.

Таким образом, если атака на звук набора окажется реализованной, то все пользователя VoIP окажутся под угрозой. Предыдущие работы не учитывали условия необходимые для атаки с использованием VoIP-программ. Так программы выполняют ряд преобразований в звуке перед передачей через интернет, например, уменьшают дискретизацию, сжимают аудиопоток, перекодируют стереозвук в моно. Эти трансформации не обсуждались в предыдущих работах и во многом не применимы к условиям описываемой атаки.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Проблема: утечка личных и конфиденциальных данных пользователя при использовании электронных устройств.

Целью диссертационного исследования является восстановление исходного текста по аудиозаписи набора текста, переданной через VoIP, для восстановления применить алгоритмы и методы машинного обучения.

Для реализации поставленной цели необходимо выполнить некоторые промежуточные действия, которые будут являться задачами исследования.

Сформулируем их:

– провести сравнительный анализ возможных решений проблемы;

– описать возможные реалистичные сценарии атаки;

– реализовать атаку на основе предположений сценариев;

– предложить способы защиты от реализованной атаки.

Объектом исследования становится возможность собрать личные и конфиденциальные данные пользователя через удалённые каналы связи. Тогда предметом будет сбор текстовых данных пользователя доступных при VoIP-разговоре.

Гипотеза исследования заключается в положении о том, что возможно реализовать атаку в процессе которой текст набираемый пользователем в процессе VoIP-разговоре будет восстановлен на основе аудиозаписи набора, полученной вместе с основной голосовой информацией.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

В 21-ом веке люди превратились из потребителей контента в производителей. Увеличилось количество каналов связи, доступных широкому кругу населения, как и увеличилось количество и мощность доступных электронных устройств. Как никогда стал важен контроль над доступом к личной информации. В работе рассматривается возможный канал утечки данных или информации о получении доступа к данным.

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя, С. И. Сиротко, заключается в формулировке целей и задач исследования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Пояснительная записка по диссертационной работе включает в себя оглавление, общую характеристику работы, введение, основную часть, состоящую из 5 глав, заключения и списка использованной литературы.

Первая глава содержит обзор и анализ предметной области, определение возможных путей решения, их сравнение и выбор наиболее сбалансированного по таким показателям, как стоимость, простата реализации, точность определения координат и т.д. Также глава содержит раздел постановки задач, исходя из выбранного пути решения.

Вторая глава содержит теоретическое описание устройства Beacon, варианты его взаимодействия с другими устройствами, структуру передаваемых данный и т.д. Также в этом разделе рассматриваются математические алгоритмы определения точных координат по нескольким точкам в пространстве и варианты работы с картой для определения на ней и отображения текущего местоположения пользователя.

В третьей главе происходит обзор программных средств и технологий, необходимых для реализации системы, таких как язык программирования Python.

Четвертая глава описывает процесс создания системы, ее основных функциональных частей и общей схемы работы. В данной главе содержится информация о использовании технологий в данной части системы, описывается ее функциональность, алгоритмы работы и привносимые ею результаты в схему работы системы.

В заключении подводятся итоги и делаются выводы по работе, а также описывается дальнейший план развития проекта.

ЗАКЛЮЧЕНИЕ

Данная работа демонстрирует высокоточную атаку, основанную на подслушивании звука клавиатурного набора удалённого пользователя через VoIP-канал.

В работе были рассмотрены алгоритмы машинного обучения и алгоритмы выделения признаков специфичных для аудиообработки.

Затем были перечислены возможные сценарии атаки, в которых новаторским было использование VoIP, при реалистичных предположения: случайный целевой текст и малый объём выборки для обучения.

После чего на основе предположений была реализована успешная атака и произведено сравнение возможных инструментов для её реализации. Результатом атаки стала добыча конфиденциальной информации жертвы.

В заключение были приведены предложения по возможной защите от подобного рода атак.

В качестве продолжения исследования следует реализовать атаку на большем количестве моделей ноутбуков (клавиатур) и пользователей, что может привести к нахождению новых закономерностей и улучшению качества распознавания.

Или можно провести схожие опыты на других VoIP-клиентах, как Google Hangouts, Viber.

Областью для исследования может стать влияние видом микрофонов на качесто распознавания.

Возможно расширение границ применения атаки с паролей на текст, где существуют различные оптимизации с применением словарей и лингвистических закономерностей для повышения качества распознавания.

СПИСОК ПУБЛИКАЦИЙ АВТОРА

[1-А.] Павлович А. Н. Использование Bluetooth Low Energy Beacons для навигации внутри зданий / Л. А. Лось, А. Н. Павлович // материалы 53-ей научной конференции аспирантов, магистрантов и студентов. – Минск: БГУИР, 2017 – С. 178 – 179.