Санкт-Петербургский государственный университет

Фундаментальная информатика и информационные технологии
Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Чугаева Татьяна Васильевна

Поиск связей между сущностями в криминалистическом анализе источников данных

Магистерская диссертация

Научный руководитель:
д. ф.-м. н., профессор Терехов А. Н.

Рецензент:
Тимофеев Н. М.

Санкт-Петербург
2016

SAINT-PETERSBURG STATE UNIVERSITY

Fundamental Computer Science and Information Technologies
Software of Computers, Complexes and Networks

Chugaeva Tatiana

Searching for connections between entities in forensic analysis of digital data sources

Graduation Thesis

Scientific supervisor:
professor Andrey Terekhov

Reviewer:
Nikita Timofeev

Saint-Petersburg
2016

Оглавление

Введение

Анализ цифровых устройств уже давно стал стандартной процедурой при расследовании преступлений. История мгновенных сообщений, удалённые файлы, фотографии сомнительного содержания – всё это может пригодиться для доказательства вины или невиновности подозреваемого в суде [5].

Специальное программное обеспечение для криминалистического анализа данных облегчает задачи нахождения и анализа артефактов, оставшихся от работы пользователя на компьютере, мобильном телефоне и прочих цифровых устройствах. Артефакты могут представлять собой как файлы (например, текстовые документы, файлы баз данных браузера со списком посещённых страниц в интернете), так и данные оперативной памяти компьютера (например, данные, оставшиеся после посещения сайтов — личные сообщения, письма почтовых сервисов и т.п.). Данные могут быть найдены в нераспределённых областях жёсткого диска, где могли сохраниться удалённые файлы или куда специально могли быть спрятаны временные данные программ, представляющие интерес для экспертов. Используемое криминалистами ПО также гарантирует, что во время анализа устройства данные не были изменены.

В анализе могут участвовать источники данных одного или нескольких лиц. Во втором случае эксперту может понадобиться исследовать взаимодействие группы лиц между собой, установить связи между подозреваемыми и жертвами, выявить наиболее активных участников, основные пути передачи информации и т.д. Однако, так как извлекаемые данные разнообразны и их количество может быть довольно объёмным (например, при анализе жёсткого диска могут найтись десятки и тысяч электронных писем и миллионы сообщений), эксперту будет непросто увидеть целостную картину взаимодействий.

Другая сложность состоит в том, что человек может иметь несколько почтовых адресов, номеров телефонов и учётных записей, поэтому не всегда можно сразу обозначить список всех собеседников и взаимодействий. Эксперту приходится вручную искать похожие учётные записи,

которые могут быть созданы одним человеком, и рассматривать их как единое целое. Это занимает немалое время, хотя вполне может быть автоматизировано.

Также было бы полезно выявлять наиболее крепкие связи, то есть те, которые длятся достаточно долго и содержат большое количество сообщений, писем, звонков и прочих видов взаимодействия. Это позволило бы увидеть приоритетные направления взаимодействия, на которые стоит обратить внимание в первую очередь.

Немаловажным является представление результатов анализа цифрового устройства. Обработанные данные должны быть представлены в виде графа связей, чтобы эксперт быстрее вникнул в суть дела и эффективно взаимодействовал с данными. Таким образом, разрозненные данные превращаются в практическую информацию, продвигающую расследование.

В полученном графе могут быть обнаружены группы тесно связанных между собой вершин, которые часто называют сообществами. Задаче разбиения вершин графа взаимодействий на группы, которую также называют выделением сообществ, посвящена работа Куликова Е.К. "Выделение сущностей в криминалистическом анализе источников данных", представленная на конференции "СПИСОК-2016".

Предложенная модель должна быть реализована и интегрирована в продукт Belkasoft Evidence Center [2], поэтому при разработке программного решения использовался язык C# и программная платформа Microsoft .NET 4.0.

Постановка задачи

Целью данной работы является создание модели для нахождения связей между людьми или группами лиц в криминалистическом анализе цифровых источников данных. Для достижения данной цели были выделены следующие задачи.

- Создать модель для нахождения связей между людьми или группами лиц в криминалистическом анализе источников данных.
- Реализовать модель.
- Внедрить модель в коммерческий продукт цифровой криминалистики.
- Полученный граф связей должен использоваться в дальнейшем для выделения сообществ.
- Выполнить апробацию модели.

1. Обзор существующих решений

На данный момент задача отображения информации в терминах отдельных личностей или групп лиц является одной из приоритетных в компаниях, занимающихся цифровой криминалистикой. Ниже представлены программы, которые предоставляют возможности по анализу социальных взаимодействий.

1.1. Мобильный криминалист

Программный комплекс "Мобильный криминалист" компании Oxygen Forensics [8] предоставляет разнообразные возможности для исследования мобильных устройств, извлечения данных из облачных хранилищ и анализа биллингов операторов сотовой связи.

Имеются возможности автоматического или ручного объединения контактов из различных источников (телефонной книги, сообщений, приложений для обмена мгновенными сообщениями и т.п.), просмотра статистики касательно взаимодействий владельца устройства с другими людьми (предпочитаемый тип связи, первую и последнюю дату связи, общее время разговоров и количество принятых и отправленных сообщений). Также в продукте представлен граф, в виде объединённых контактов и их взаимодействий.

Однако данный продукт рассматривает лишь определённый круг устройств, а именно мобильные телефоны, смартфоны и планшеты.

1.2. Forensic Toolkit

Другой продукт – Forensic Toolkit (FTK) компании AccessData – содержит отдельный компонент Social Analyzer для визуализации общения по электронной почте. Можно просматривать связи на уровне доменов и на уровне конкретных адресов [1].

Однако это является лишь частным случаем анализа социальных взаимодействий, учитывая всю извлекаемую из устройств информацию.

1.3. UFED Link Analysis

Более широкими возможностями обладает продукт UFED Link Analysis компании Cellebrite [3]. Граф взаимодействий состоит из вершин, которые созданы по учётной записи, номеру телефона, электронному адресу почты или по источнику данных. Связь в графе может содержать различные типы взаимодействий: звонки, сообщения, письма и т.п. При этом данные в графе можно фильтровать по типу связи, количеству взаимодействий, времени.

1.4. Nuix

Компания Nuix предоставляет криминалистические продукты для различных потребностей. Например, Nuix Incident Response содержит визуализацию данных в виде графа [7]. Помимо взаимодействий между людьми, показаны истории браузеров, события операционной системы, взаимодействия с файлами и их удаление. Также возможен анализ по ключевым словам и разнообразная фильтрация данных.

1.5. IBM i2 Analyst's Notebook

Также стоит упомянуть о продукте IBM i2 Analyst's Notebook [6]. Он является наиболее мощным решением и предлагает систематизацию разрозненных данных в едином согласованном представлении.

Ключевым понятием является централизованность сети взаимодействий, то есть интерес представляет поиск ключевых личностей, потеря которых нарушит процесс обмена информацией в сети. Также учитывается сила связи между узлами. Например, родственная связь является более сильной по сравнению со связью между коллегами.

Возможности продукта огромны и разнообразны, однако в области компьютерной криминалистики данный продукт фактически не используется по причине высокой сложности и стоимости. Основными пользователями продукта являются крупные компании, содержащие огромные объемы информации о взаимодействиях сотрудников.

1.6. Выводы

Все рассмотренные выше продукты предлагают некоторые возможности анализа социальных взаимодействий. Некоторые из них специализируются на узком круге анализируемых устройств или данных, используемых для графа связей. Другие продукты предоставляют широкие возможности, но в то же время не используются в криминалистике.

Для анализа социальных взаимодействий наиболее подходящими решениями на данный момент являются UFED Link Analysis и Nuix. Однако не стоит рассчитывать только на эти продукты, поскольку данные на которых проводится анализ взаимодействий, могут различаться. Во-первых, криминалистический продукт может не найти какие-то данные по ошибке или же вообще не поддерживать их извлечение. Во-вторых, в настоящее время новые системы мгновенного обмена сообщениями появляются очень быстро, уже существующие программы постоянно обновляются и меняют форматы хранения данных, что также влияет на множество извлекаемых данных.

Решение, предложенное в данной работе, интегрировано в криминалистический продукт Evidence Center, разработанный компанией Belkasoft. Данный продукт обнаруживает более 700 типов артефактов, поддерживает все основные программы мгновенного обмена сообщениями, браузеры, почтовые клиенты, социальные сети, пиринговые программы и т.п.

2. Описание модели для нахождения связей между людьми или группами лиц

Данный раздел описывает модель для нахождения связей между людьми или группами лиц в криминалистическом анализе источников данных. В начале рассмотрены исходные данные, на которых будет строиться модель. Далее следует описание понятия “сущность”, являющееся ключевым в данной работе. В конце раздела рассказано о связях между сущностями и вычислении весов.

2.1. Исходные данные

Анализ источника данных может привести к обнаружению различных артефактов, таких, как сообщения, письма, звонки, изображения, текстовые документы, видео, данные реестра и т.д. Среди них может найтись список контактов, полученных, например, из истории мгновенных сообщений.

Контакт – информация, характеризующая человека или группу лиц. Контакт может быть получен из следующих источников:

- звонки;
- голосовые сообщения;
- короткие текстовые сообщения;
- мгновенные сообщения;
- электронные письма;
- адресная книга мобильного телефона;
- информации об анализируемом устройстве.

Соответственно, контакт может содержать название учётной записи, псевдоним, адрес электронной почты, телефонные номера, имя, фамилию, название компании и так далее. Например, из электронного

письма извлекаются контакты отправителя, получателей, получателей копии письма и скрытых получателей копии письма.

Взаимодействие – факт передачи информации между контактами. Например, звонок, голосовое сообщение, короткое текстовое сообщение, мгновенное сообщение или электронное письмо. У некоторых контактов может не быть взаимодействий с другими контактам, если они, к примеру, получены из адресной книги мобильного телефона. Также заметим, что взаимодействия существуют только между контактами одного типа. Например, извлечённые данные не предоставляют информации о взаимодействиях между контактом, полученным из электронного письма, и контактом из истории голосовых сообщений.

Эксперт может предположить, что пара контактов принадлежит одному и тому же человеку, а значит существует взаимодействие между людьми, представленными контактами разных типов, но наверняка утверждать этого нельзя.

2.2. Понятие “сущности”

Для автоматизации анализа социальных связей необходимо заранее находить и объединять контакты, принадлежащие одним и тем же лицам. Таким образом эксперт сможет анализировать не отдельные сообщения и звонки, а социальные взаимодействия между людьми в целом.

Сущность – человек или группа лиц, представленные идентифицирующими данными (каждая характеристика может встречаться несколько раз):

- название учётной записи;
- адрес электронной почты;
- номер телефона;
- фамилия;
- имя;
- псевдоним.

2.3. Связи между сущностями

Для пары контактов могут существовать взаимодействия между ними, следовательно между сущностями, как объединениями контактов, также могут быть взаимодействия.

Между двумя сущностями существует связь, если состоялся хотя бы один факт взаимодействия между ними.

2.4. Веса для связей между сущностями

Каждая связь имеет вес, характеризующий её значимость. Значение веса зависит от типа, количества и времени взаимодействий. Ниже указаны взаимодействия в порядке убывания значимости:

- звонок;
- голосовое сообщение;
- короткое текстовое сообщение;
- мгновенное сообщение;
- электронное письмо.

Например, звонок считается более значимым типом взаимодействия, нежели электронное письмо. Использование телефона предполагает общение тет-а-тет и большую вовлечённость в диалог, тогда как у письма могут быть несколько получателей, и ответ на него может прийти через несколько часов или даже дней.

Количество взаимодействий также указывает на степень близости между людьми. Так связи с наибольшим количеством взаимодействий зачастую указывают на родственников, близких друзей или обсуждение какого-то дела.

Помимо контактов, принадлежащих реальным людям, существуют электронные адреса компаний, рассылающих рекламные предложения или новости, спам-боты и т.п. Они как правило отправляют множество писем или сообщений за короткий временной промежуток, но это

не означает, что связь с ними имеет высокую ценность. Поэтому для оценки значимости стоит также учитывать время взаимодействий между сущностями.

3. Реализация модели

В данном разделе описана реализация предложенной модели. В первой части описано создание сущностей для источника данных. Определение связей между сущностями и вычисление весов представлены во второй и третьей частях, соответственно. В четвёртой описаны создание графа связей и визуализация полученных результатов. Информация об известных ограничениях реализации находится в конце раздела.

3.1. Выделение сущностей

Алгоритм выделения сущностей работает со списком контактов, найденных во время анализа цифрового источника данных.

3.1.1. Создание контактов для источника данных

Уже на этапе анализа известны почтовые ящики и учётные записи, в которые был выполнен вход с анализируемого устройства, а также телефонные номера источника данных. В рамках данной работы предполагается, что такие контакты принадлежат владельцу устройства. Например, при анализе ноутбука считается, что все найденные на устройстве учётные записи Skype, с которых был выполнен вход, принадлежат хозяину ноутбука.

Чтобы не терять полезную информацию, при добавлении источника данных для него создаётся отдельный контакт. Далее во время анализа разнообразных типов данных извлекаются прочие контакты. Все контакты владельца ссылаются на контакт источника данных – родительский контакт.

Может показаться, что родительский контакт и есть сущность, однако это не совсем так. Сущность может объединять контакты нескольких источников данных. Например, при анализе резервных копий, созданных в разное время, но с одного устройства, будут созданы два родительских контакта, объединённых в одну сущность. Наборы контактов владельца мобильного устройства могут различаться для каж-

дой из копий, так как информация о старых учётных записях может быть удалена, а со временем могли появиться новые контакты.

3.1.2. Создание сущностей

Для создания сущности необходимо найти все контакты, которые предположительно принадлежат одному человеку или группе лиц. Для этого проводится сравнение некоторых характеристик контактов (в том виде, в котором они были извлечены):

- название учётной записи (peter_petrov);
- адрес электронной почты (peter.petrov@gmail.com);
- номер телефона (123-45-67);
- фамилия и имя (Петров Петя);
- псевдоним (sunny);
- имя почтового ящика и название учётной записи;
- имя почтового ящика и псевдоним.

Фамилия и имя сравниваются вместе, чтобы уменьшить количества ошибок. Если совпадает хотя бы одна пара указанных выше характеристик, контакты будут объединены в одну сущность.

В первую очередь сравниваются родительские контакты, далее контакты, принадлежащие владельцу источника данных, и в конце прочие оставшиеся контакты. При этом контакты владельца входят в ту же сущность, что и родительский контакт.

Таким образом, для каждого множества похожих контактов создаются сущности. Каждая сущность состоит из названия (имя первого добавленного контакта) и списка всех её контактов.

3.2. Вычисление связей между сущностями

Исходные данные, полученные после анализа источника данных, могут содержать взаимодействия между контактами: звонки, текстовые сообщения, письма и т.п. Наличие взаимодействий означает, что между сущностями существует связь.

3.3. Вычисление весов для связей

Для того чтобы эксперт мог исследовать наиболее значимые отношения между сущностями, определим вес связи между сущностями. Перед тем как вычислить веса для связей между сущностями необходимо получить веса для связей между их контактами.

Определим вес направленной связи W_oriented для пары контактов (C1, C2) как среднее арифметическое трёх параметров:

- тип связи: звонок, голосовое сообщение, короткое текстовое сообщение, мгновенное сообщение, письмо (указаны в порядке убывания значимости);
- процент количества взаимодействий C1 с C2;
- процент времени взаимодействия C1 с C2.

Для устранения случаев незначительных взаимодействий между контактами, добавлены минимальные пороговые значения для общего количества и общего времени взаимодействий (контакта C1).

Вес ненаправленной связи W_undirected(C1, C2) будем считать средним арифметическим направленных весов.

$$W\_undirected(C1, C2) = (W\_oriented(C1, C2) + W\_oriented(C2, C1))/2$$

Пусть есть две сущности E1 с контактами C1..Cn и E2 с контактами C1..Cm, где n и m – количества контактов сущностей E1 и E2 соответ-

ственно. Тогда вес связи W между E1 и E2 будет равен следующему.

$$W(E1, E2) = (\sum_{i=0}^{n} \sum_{j=0}^{m} W\_undirected(Ci, Cj))/nm$$

У весов для связей между сущностями также существует минимальное пороговое значение.

3.4. Визуализация результатов

Для визуализации была использована библиотека GoDiagram [4], которая представляет собой набор элементов управления и классов, построенных на платформе .NET и предназначенных для создания двумерных графов.

Результатом предложенной модели является неориентированный граф (граф связей), в котором вершины и рёбра представляют контакты и связи, соответственно (см. Рис. 1).

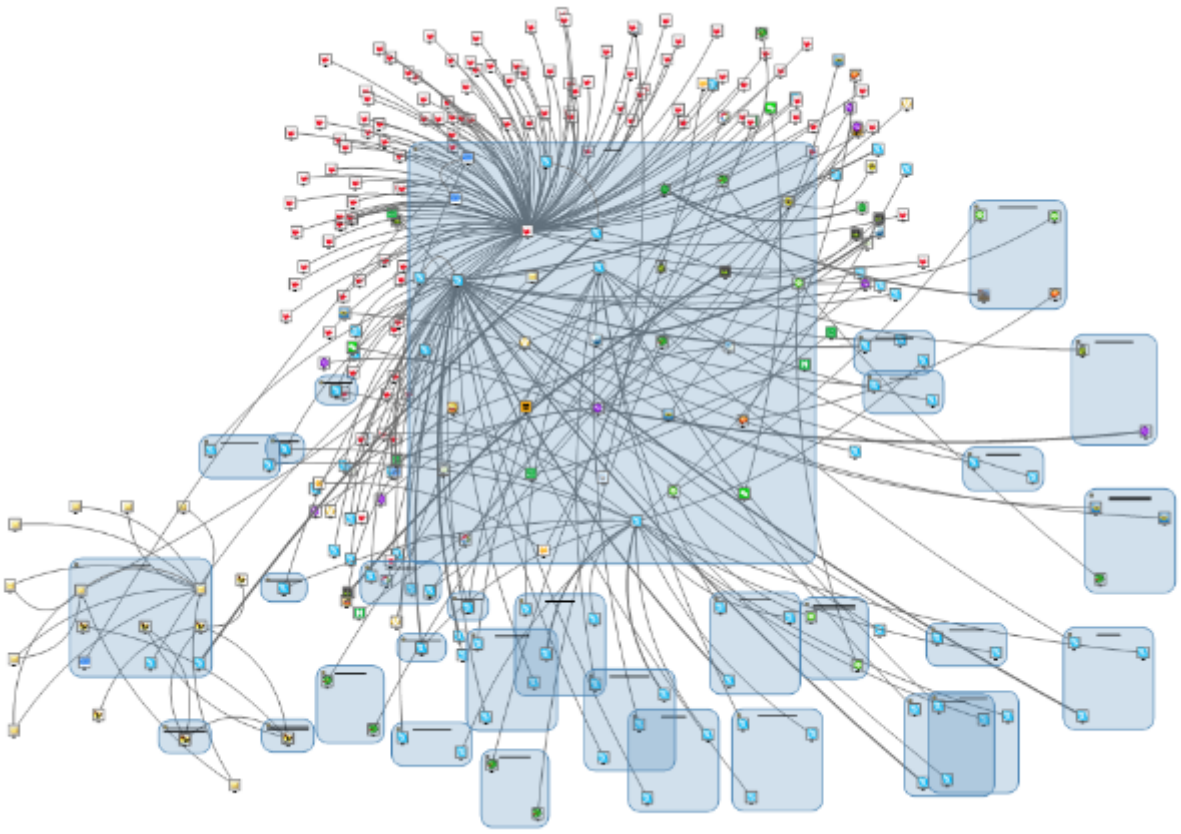

Рис. 1: Пример графа связей

Сущность представлена в виде подграфа, объединяющего соответствующие контакты и связи между ними (см. Рис. 2). Подграфы сущностей можно сворачивать, упрощая граф связей, и разворачивать, показывая контакты, принадлежащие сущностями. Таким образом можно увидеть взаимодействия как между сущностями, так и между отдельными контактами.

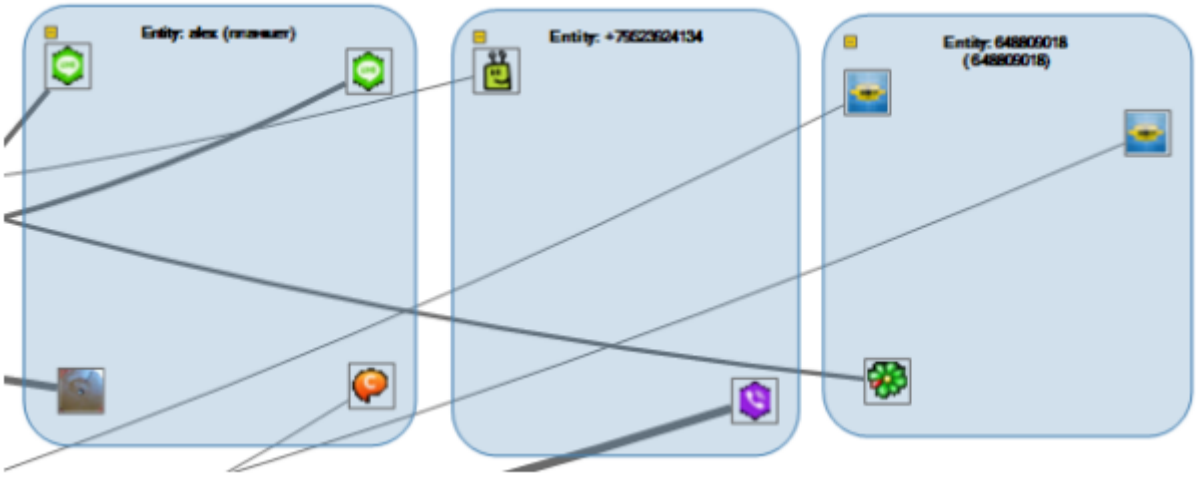

Рис. 2: Пример сущностей, содержащих различные контакты

При выделении вершины или подграфа есть возможность подсветки смежных узлов и инцидентных рёбер (см. Рис. 3). Также для каждого узла и ребра можно посмотреть более подробную информацию.

Для наглядности было решено устанавливать толщину ребра в зависимости от веса и не отображать само численное значение (см. Рис. 4).

Представленная модель была реализована и интегрирована в современный продукт компьютерной криминалистики Belkasoft Evidence Center, разрабатываемый компанией Belkasoft.

3.5. Ограничения реализации модели

Описанная выше реализация модели имеет ряд известных ограничений.

1. Не всегда возможно объединить все контакты, принадлежащие

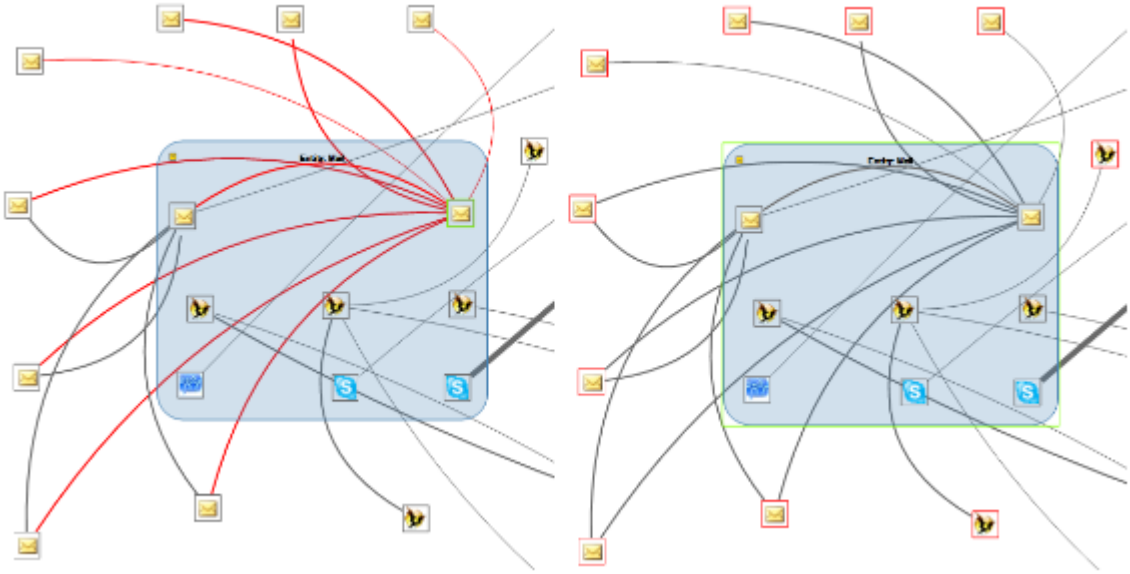

Рис. 3: Подсветка инцидентных рёбер для контакта (слева) и смежных узлов для сущности (справа)

одному человеку или группе лиц из-за недостатка данных.

2. Также возможны ложно положительные объединения контактов в одну сущность. Например, если контакты, имеющие одинаковые названия учётных записей, на самом деле принадлежат разным людям. В случае совпадения имени и фамилии контакты также могут быть объединены.

3. Формула вычисления веса не учитывает, что некоторые характеристики связи могут быть более важными, чем другие (например, можно считать неважным тип взаимодействия).

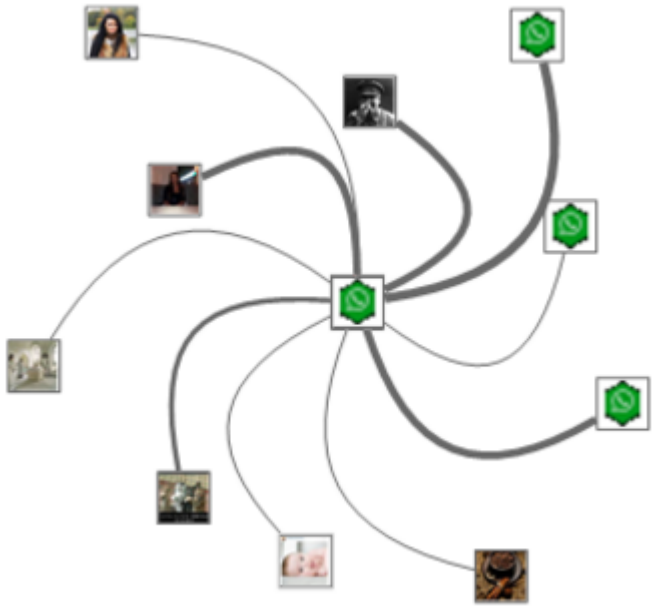

Рис. 4: Толщина рёбер в графе зависит от веса: самые тонкие рёбра имеют наименьший вес

4. Использование результатов модели

На основе полученного графа эксперт может сделать выводы о структуре взаимодействий, о характерном поведении некоторых сущностей. Например, могут быть обнаружены сущности, которые только распространяют информацию, но не принимают её.

Рассматривая граф в целом могут быть видны группы сущностей, которые общаются друг с другом значительно больше, чем с остальными. Например, это могут быть люди с общими интересами или занимающиеся одним делом. Такие группы группы тесно связанных вершин в графе называют сообществами, а задачу разбиения вершин графа на группы – выделением сообществ.

Граф, полученный в ходе работы модели, содержит необходимые данные для обнаружения сообществ, а именно веса для рёбер. Основой для достоверного выделения сообществ считается качественный подсчёт весов.

5. Апробация модели

В рамках работы была проведена апробация предложенной модели на различных цифровых данных.

5.1. Учётные записи Skype

Были проанализированы данные программы мгновенного обмена сообщениями Skype полученные от трёх разных людей. Программой Belkasoft Evidence Center было найдено 49713 различных артефактов, том числе 471 контакт. На Рис. 5 представлен полученный граф связей. Сущности источников данных имеют наибольшее количество связей. Из графа видно, что некоторые сущности общались с каждым из хозяев источников данных.

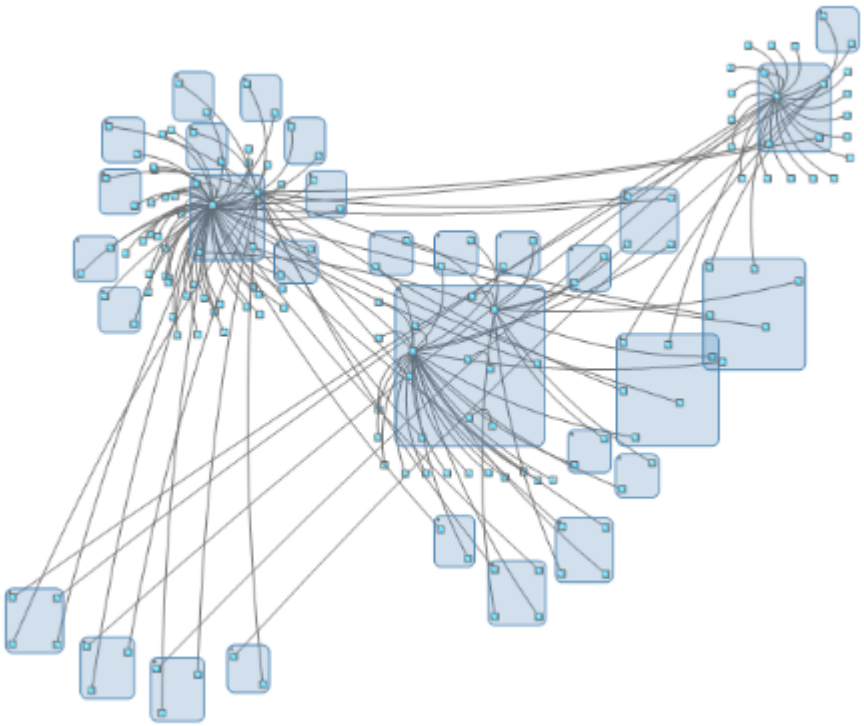

Рис. 5: Граф связей, полученный для трёх источников данных Skype

5.2. Резервные копии мобильных телефонов с операционной системой Android

Были проанализированы две резервные копии мобильных телефонов. В процессе анализа было обнаружено 12963 различных артефакта, том числе 336 контактов. На Рис. 6 представлен полученный граф связей. В одной резервной копии были обнаружены только 2 контакта для Viber и ВКонтакте, в другой копии были найдены контакты для Skype, FireChat и других систем мгновенного обмена сообщениями. Можно заметить, что в данном случае были обнаружены общие собеседники в ВКонтакте.

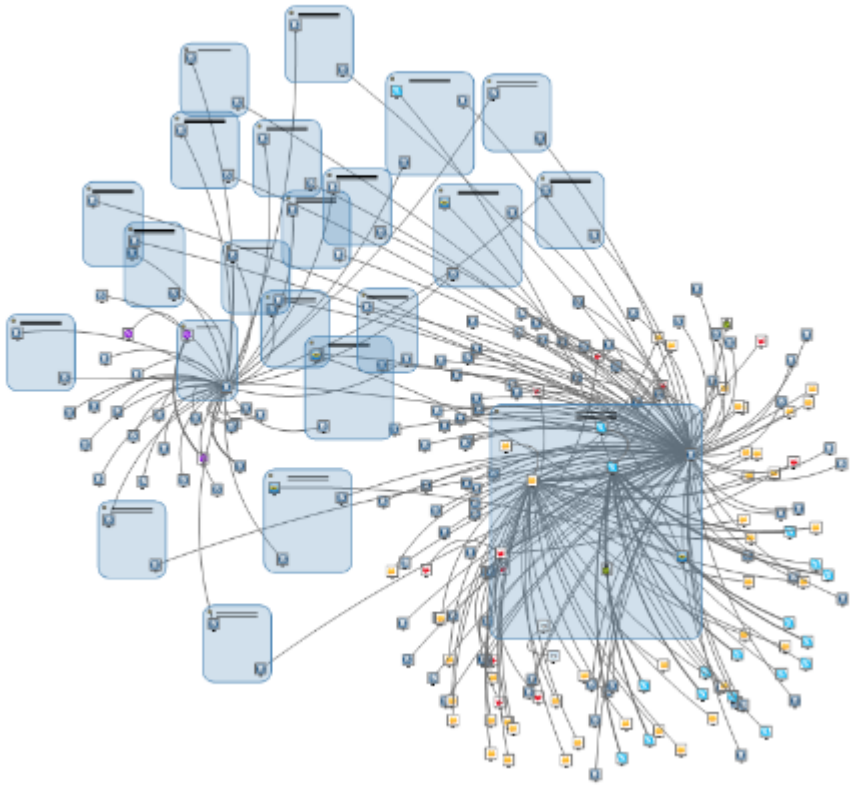

Рис. 6: Граф связей, полученный для двух резервных копий мобильных телефонов с операционной системой Android

5.3. Резервные копии мобильных телефонов с операционной системой iOS

Было проанализировано две резервные копии мобильных телефонов с операционной системой iOS. Было найдено 11945 различных артефактов, том числе 561 контакт. На Рис. 7 представлен полученный граф связей. Две сущности, содержащие наибольшее количество контактов, принадлежат хозяевам устройств. Большинство сущностей в графе состоят из двух контактов одного типа, причём каждый контакт связан с сущностью одного из хозяев. Следовательно у хозяев устройств существует определённое количество общих знакомых.

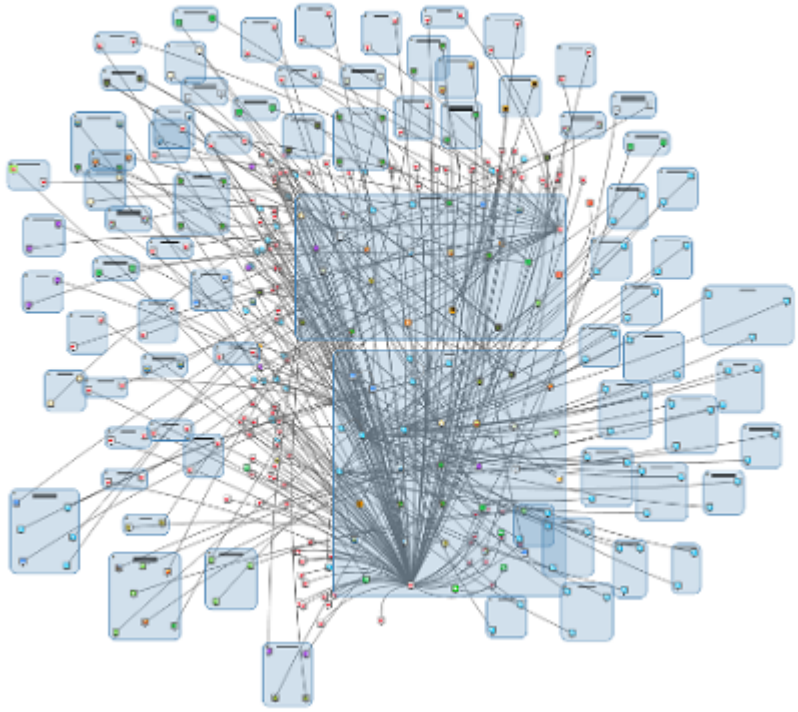

Рис. 7: Граф связей, полученный для двух резервных копий мобильных телефонов с операционной системой iOS

Заключение

В рамках данной работы были получены следующие результаты.

- Создана модель для нахождения связей между людьми или группами лиц в криминалистическом анализе источников данных.
- Предложенная модель реализована.
- Модель внедрена в продукт Belkasoft Evidence Center.
- Полученный граф связей используется для выделения сообществ.
- Выполнена апробация модели.

По результатам работы был сделан доклад “Поиск связей между сущностями в криминалистическом анализе цифровых источников данных” на конференции “СПИСОК-2016”, тезисы опубликованы в сборнике материалов конференции.

В дальнейшем планируется провести эксперименты с использованием других формул для вычисления весов. Также будут исследованы возможности для ускорения работы алгоритмов выделения сущностей и подсчёта весов.

Список литературы

[1] AccessData. Summation Feature Friday with Tim Leehealey: Visualization. — 2016. — https://www.youtube.com/watch?v=k6sPnzSF6O8&index=3&list=WL [дата просмотра: 13.05.2016].

[2] Belkasoft Evidence Center 2016. — https://belkasoft.com/ec [дата просмотра: 13.05.2016].

[3] Cellebrite Mobile Forensics: Link Analysis Identify Connections Between Suspects. — 2014. — https://www.youtube.com/watch?v=3f5hA3SwTVo [дата просмотра: 13.05.2016].

[4] GoDiagram for WinForms. — http://www.nwoods.com/products/godiagram/ [дата просмотра: 13.05.2016].

[5] Gubanov Yuri. Retrieving Digital Evidence: Methods, Techniques and Issues. — 2012. — https://articles.forensicfocus.com/2012/07/11/retrieving-digital-evidence-methods-techniques-and-issues/ [дата просмотра: 13.05.2016].

[6] IBM. Анализ и визуализация данных для эффективной аналитики. — 2016. — http://www-03.ibm.com/software/products/ru/analysts-notebook [дата просмотра: 13.05.2016].

[7] Nuix Incident Response. — http://www.nuix.com/products/nuix-incident-response [дата просмотра: 13.05.2016].

[8] Программный комплекс “Мобильный криминалист” компании Oxygen forensics. — http://www.oxygensoftware.ru/ru/ [дата просмотра: 13.05.2016].