Санкт-Петербургский государственный университет

Математическое обеспечение и администрирование информационных систем
Системное программирование

Куликов Егор Константинович

Выделение сущностей в криминалистическом анализе источников данных

Бакалаврская работа

Научный руководитель:
ст. преп. Губанов Ю. А.

Рецензент:
к. ф.-м. н., доцент Бугайченко Д. Ю.

Санкт-Петербург
2016

SAINT-PETERSBURG STATE UNIVERSITY

Software and Administration of Information Systems
Software engineering

Egor Kulikov

Determination of entities in the forensic analysis of digital data sources

Graduation Thesis

Scientific supervisor:
Sen. Lect. Yu .A. Gubanov

Reviewer:
Ass. Prof. D. Yu. Bugaichenko

Saint-Petersburg
2016

Оглавление

Введение

Всё чаще для установления причастности определённого лица к совершению преступления или доказательства его вины прибегают к помощи компьютерно-технической экспертизы. В соответствии с судебным постановлением у подозреваемого конфискуются имеющиеся электронные устройства, от ноутбуков и планшетов до автомобильных навигаторов, которые в дальнейшем анализируются с помощью специальных инструментов и программ.

Криминалистический анализ отдельно взятого устройства позволяет получить определённую информацию о его владельце: поисковые запросы, часто посещаемые страницы в интернете, а также активность в социальных сетях и программах мгновенного обмена сообщениями, то есть сферу интересов пользователя устройства и круг его общения. Возможно также и извлечение геолокационных данных. Лицо, наделённое совокупностью этих данных, мы будем в дальнейшем называть сущностью и рассматривать как атомарный объект с точки зрения криминалистического анализа. Собранная информация может помочь установить, владел ли подозреваемый определёнными сведениями о правонарушении, находился ли в момент совершения преступления поблизости, контактировал ли с жертвой или другими подозреваемыми. Собранная путём компьютерно-технической экспертизы информация может в дальнейшем быть предъявлена в суде в качестве доказательства.

Имея санкцию, следователь получает возможность изъять у подсудимого цифровые устройства и провести экспертизу содержащейся на них информации. В то же время представителю органов правопорядка может быть необходимо установить, не содержали ли схожих данных ранее исследуемые устройства. Такие сведения могли бы оказаться весьма полезными в расследовании совершённых или готовящихся преступлений. Например, попадание в руки следователей нескольких устройств рядовых членов диверсионной группировки, возможно, позволило бы установить одного из её руководителей путём сравнения глобального списка контактов из конфискованных устройств, так как участники незаконного формирования, скорее всего, регулярно поддерживают связь со своим руководством.

Установление взаимосвязей между полученными уликами и материалами ранее расследованных уголовных дел также является важной составляющей расследования противоправных деяний. Так, например, согласно сведениям [30] Федеральной службы исполнения наказаний России (ФСИН), лишь около 36% заключённых впервые находятся в местах лишения свободы, причём число осуждённых три и более раза превышает те же 36%. Это показывает, что при расследовании очень важно проверить, не является ли преступление и его обстоятельства схожими с ранее совершёнными.

Для представления и анализа взаимодействий участников расследуемого дела оказывается удобным использование математического аппарата теории графов. Сущно-

сти обозначаются вершинами, а взаимодействия между ними — взвешенными рёбрами. Вес каждого ребра рассчитывается по специальному алгоритму в зависимости от того, насколько тесным было взаимодействие. Построению метода подбора весов посвящена магистерская работа Т. В. Чугаевой «Поиск связей между сущностями в криминалистическом анализе источников данных».

В таком случае взаимосвязанные сущности представляются группой вершин, таких, что в число и суммарный вес рёбер, исходящих в другие его вершины, значительно превышает количество и вес тех, что связывают вершины этой группы с остальными вершинами графа. Такие группы тесно связанных между собой сущностей часто называют сообществами (англ. community), а задачу разбиения графа взаимодействий на группы активно взаимодействующих сущностей — выделением сообществ [6].

1. Постановка задачи

Задачей дипломной работы является построение алгоритма выделения сообществ в графах, возникающих при проведении компьютерно-технических экспертиз, показывающего достаточно точные результаты и приемлемую скорость работы.

Для сравнения производительности и качества выдаваемых алгоритмами разбиений графа на сообщества был построен тестовый набор графов, возникавших в реальных экспертизах.

Алгоритм считается допустимым с точки зрения производительности, если он анализирует любой граф тестового набора (в том числе наибольший, содержащий 764 вершины и 14750 рёбер) не более чем за десять минут на стандартной вычислительной машине отечественного эксперта-криминалиста[1]. Критерии качества результата подробно рассматриваются в разделе 3.1.

Предложенное решение планируется в дальнейшем интегрировать в современный отечественный продукт компьютерной криминалистики Belkasoft Evidence Center, разрабатываемый компанией Belkasoft с 2010 года. Поэтому при разработке программного решения должен быть использован язык C# и программная платформа Microsoft .NET версии 4.0.

[1]Как правило, используются вычислительные машины с четырёхъядерным 64-разрядным процессором Intel Core i7 частотой 4 ГГц, вместимостью оперативного запоминающего устройства 32 Гб и установленной операционной системой семейства Windows.

2. Обзор существующих решений

Структура обзора устроена следующим образом: сначала рассматриваются возможности по построению и анализу графов социальных взаимодействий, предоставляемые современными средствми компьютерной криминалистики, далее проводится обзор некоторых теоретических работ по выделению сообществ в графах и предлагаемых в них алгоритмов, после чего анализируются статьи, авторы которых проводят сравнительный анализ существующих методов выделения сообществ.

2.1. Инструменты компьютерной криминалистики

2.1.1. Мобильный криминалист

«Мобильный криминалист» — приложение, разработываемое отечественной компанией Oxygen Forensics. Эта программа предназначена для судебно-технической экспертизы сотовых телефонов, смартфонов и планшетных компьютеров, используется правительственными учреждениями, полицией, армией, таможенными и налоговыми службами [27].

Возможности приложения, относящиеся к установлению взаимосвязей между владельцами исследуемых устройств, таковы: эксперту-криминалисту предоставляется функциональность по выделению списка контактов пользователей устройства, причём поддерживается автоматическое объединение контактов из различных источников (телефонной книги, сообщений, журнала событий, приложений для обмена мгновенными сообщениями, таких как WhatsApp, социальных сетей) в один мета-контакт. Возможно также объединять контакты и вручную.

На основе набора сущностей, представляющих из себя объединёные контакты, и взаимосвязей между ними, приложение позволяет построить граф взаимодействий. Тем не менее, эксперту-криминалисту не предоставляются какие-либо возможности по автоматизированному анализу этого графа.

2.1.2. Forensic Toolkit

Forensic Toolkit — один из наиболее известных инструментов компьютерной криминалистики, разрабатываемый компанией AccessData. Один из его компонентов, называющийся Social Analyzer, позволяет графически представить общение по электронной почте: визуализируются связи на уровне доменов и на уровне конкретных адресов [1].

Таким образом, компонент поддерживает представление данных о взаимодействиях участников в виде графа, однако возможности по выделению сообществ в этом графе каким-либо способом не предоставляются.

2.1.3. Nuix

Nuix Investigator, разработка компании Nuix, также является достаточно известным инструментом компьютерной криминалистики. Это приложение позволяет выбрать несколько устройств из разных расследуемых дел и провести их совместный анализ, в том числе выявить различные взаимодействия владельцев устройств посредством электронной почты или программ мгновенного обмена сообщениями и представить их затем в виде графа [15].

Однако интеллектуальный анализ получаемого графа, в том числе выделение сообществ, не поддерживается.

2.1.4. IBM i2 Analyst's Notebook

IBM i2 Analyst's Notebook — визуальная аналитическая среда, которая позволяет максимально эффективно использовать огромные объёмы информации, накопленные государственными службами и предприятиями [7].

Этот продукт, в отличие от всех предыдущих, рассматриваемых в этот обзоре, поддерживает не только возможность построения графа взаимодействий, но и некоторый его анализ: так, например, для вершины возможно вычислить значения её центральностей (англ. centrality): betweeness, clotheness, eigenvector. Тем не менее, выделение сообществ в графах взаимодействий не поддерживает и это приложение.

2.1.5. Выводы

Проведённый обзор функциональности некоторых наиболее известных средств компьютерной криминалистики, связанной с анализом социальных взаимодействий, показывает, что все они поддерживают лишь визуализацию социальных графов. Лишь одно приложение предоставляет минимальный набор средств для анализа получаемого графа, однако выделение сообществ не поддерживается и здесь.

Таким образом, современные средства цифровой криминалистики не предлагают никаких решений данной задачи, несмотря на её безусловную значимость для пользователя.

2.2. Исследования по выделению сообществ

Исследования в данной области начались в середине прошлого века и продолжаются по сей день, достигнув максимальной активности в середине 2000-х годов [6]. Несмотря на то, что эти исследования на данный момент не получили широкого применения в компьютерной криминалистике, выделение сообществ в графе активно применяется специалистами по анализу социальных сетей, имеет приложения в социологии [6], может быть использовано при визуализации больших графов [28].

В 2009 году итальянский учёный Santo Fortunato в своей работе "Community detection in graphs" [6] собрал описания большого числа известных на тот момент алгоритмов выделения сообществ, некоторые из которых были реализованы и подробно изучены в рамках этой работы.

2.2.1. Алгоритм Кернигана-Лина

Метод Кернигана-Лина — один из первых алгоритмов выделения сообществ, разработанный [8] в 1970 году, активно применяется по сей день [6]. Идея алгоритма состоит в том, чтобы из начального произвольного разреза графа на две части перейти к такому разрезу, который будет максимально качественно представлять разбиение исходного графа на два сообщества путём максимизации целевой функции. В дальнейшем алгоритм был усовершенствован другим учёными [22], так что стало возможным задать число k и построить таким образом разбиение не на два, а на k сообществ.

К сожалению, этот алгоритм неприменим к задаче выделения сообществ в компьютерной криминалистике, потому что в данном случае предсказать число сообществ в исследуемом графе практически невозможно даже приблизительно. По этой же причине в дальнейшем не будет рассматриваться такой известный метод выделения сообществ, как k-means clustering [12] и некоторые другие.

2.2.2. Алгоритм Гирвана-Ньюмана

Этот алгоритм, как и два последующих, требует введения понятия модулярности. Модулярность — функция, зависящая от разбиения графа на сообщества, которая была предложена [13] авторами как метрика качества этого разбиения.

Определение. Модулярностью Гирван-Ньюмана [13] (англ. Girwan-Newman modularity) называется функция вида

$$Q = \frac{1}{4E} * \sum_{i=1}^{V} \sum_{j=1}^{V} (A_{i,\,j} - \frac{s_i s_j}{2E}) \sigma(c_i, c_j),$$

где V, E — количество вершин и сумма весов рёбер графа соответственно,
A — его матрица смежности,
s_i — сумма весов исходящих рёбер i-ой вершины графа,
c_i — номер сообщества i-ой вершины.

Замечание. В дальнейшем Райхард и Борнхольд предложили [20] немного изме-

нить эту функцию и привести её к виду

$$Q = \frac{1}{4E} * \sum_{i=1}^{V} \sum_{j=1}^{V} (A_{i,\,j} - \gamma \frac{s_i s_j}{2E}) \sigma(c_i, c_j),$$

где $\gamma > 0$ — некая константа, позволяющая в некотором смысле управлять размером сообществ: утверждается [11], что выбор $\gamma < 1$ приводит к увеличению размера сообществ в разбиении графа, а при $\gamma > 1$ сообщества уменьшаются, а их количество возрастает.

Метод, предложенный Мишелем Гирваном и Марком Ньюманом в 2004 году, устроен [14] следующим образом:
1) Для исходного графа вычисляется модулярность Гирван-Ньюмана
2) Для каждого ребра графа вычисляется его центральность по посредничеству (англ. edge betweeness centrality), определение которой приведено ниже.
3) То ребро, центральность которого максимальна, удаляется из графа. В случае, когда таких рёбер в графе несколько, одно из них выбирается случайным образом. Если других рёбер в графе нет, то алгоритм завершает свою работу.
4) Для получившегося после удаления ребра нового графа пересчитывается функция модулярности. Если её значение увеличилось, то возвращаемся к шагу 2, иначе завершаем работу алгоритма.

Четвёртый шаг алгоритма иногда бывает устроен несколько иначе: так, алгоритм завершают не после первого падения модулярности, а в точке достижения её максимума или в одной из точек локального максимума [13].

Сообществами считаются компоненты связности графа, получившегося после завершения работы алгоритма Гирвана-Ньюмана.

Определение. Центральностью по посредничеству ребра графа называется[5] величина, определяемая по формуле

$$C(e) = \sum_{s,t \in V} \frac{\sigma(s,t|e)}{\sigma(s,t)},$$

где V — множество вершин графа,
$\sigma(s,t|e)$ — количество кратчайших путей между вершинами s и t, проходящих через ребро e графа,
$\sigma(s,t)$ — количество кратчайших путей между вершинами s и t.

Алгоритм имеет высокую вычислительную сложность, порядка $O(VE + V^2 logV)$ на итерацию по удалению ребра, которая обуславливается необходимостью каждый

раз пересчитывать центральность по посредничеству, а самый быстрый известный на сегодняшний день алгоритм — алгоритм Брандса [5] имеет именно такую асимптотику.

2.2.3. Алгоритм Радиччи

Определение Коэффициентом кластеризации ребра между вершинами a и b с параметром p (англ. edge clustering coefficient) называется [19] величина, определяемая по формуле

$$C_p = \frac{G_p}{FG_p},$$

где G_p — число циклов длины p, проходящих через данное ребро,
FG_p — максимально возможное, исходя из степеней вершин, число циклов длины p в графе.

Замечание При $p = 3$ формула принимает вид $C_p = \frac{G_3}{(\deg i - 1)(\deg j - 1)}$

Алгоритм, предложенный [19] в 2007 году Филиппо Радиччи, является модификацией метода Гирвана-Ньюмана. Как уже было отмечено выше, последний практически неприменим для анализа больших графов ввиду большой трудоёмкости вычислений. Автор данного алгоритма заметил, что вместо центральности по посредничеству можно использовать коэффициент кластеризации ребра с параметрами 3 или 4. Такой подход позволяет значительно снизить вычислительную сложность.

2.2.4. Louvain-метод

Этот алгоритм, опубликованный [4] в 2008 году, предлагает при решении задач выделения сообществ действовать следующим образом:
1) Изначально каждая вершина графа считается отдельным сообществом.
2) Далее вершины последовательно перемещаются в одно из тех сообществ, с которым имеют рёбра. Сообщество для перемещения выбирается так, чтобы был максимален прирост модулярности. Если ни одно из потенциальных перемещений не приводит к приросту модулярности, то вершина остаётся в том же сообществе.
3) Если ни одна из вершин не переместилась в иное сообщество, то работа алгоритма заканчивается.
4) На основании имеющегося графа строится мультиграф (граф, допускающий кратные рёбра и петли), вершинами которого являются сообщества, полученные на втором шаге. Ребро между вершинами A и B наделяется весом, равным сумме весов рёбер между сообществами, «стянутыми» в эти вершины. Кроме того, в каждой вершине строится петля с весом, равным удвоенной сумме весов всех рёбер сообщества, если она не равна нулю.

5) Для построенного мультиграфа выполняется второй шаг этого алгоритма.

Одна итерация метода, включающая в себя первый и второй шаги, изображена на рис. 1.

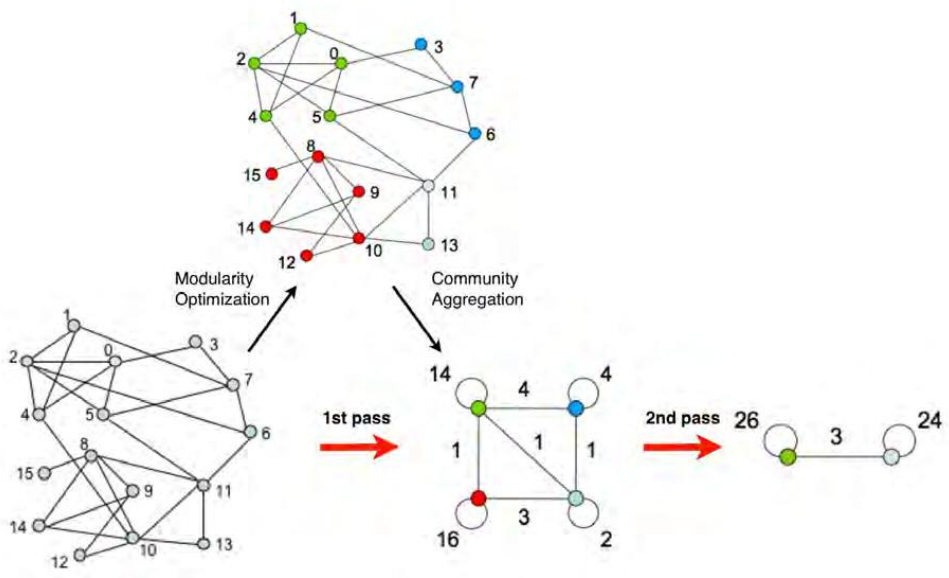


Рис. 1: Схема работы Louvain-метода

Создатели метода особо отмечают [4] крайне низкую вычислительную сложность такого подхода: так, за 152 минуты на машине со среднестатистическими техническими характеристиками, которые приводятся в [4], был проанализирован граф, содержащий 118 миллионов вершин.

2.2.5. Алгоритм Prat-Perez et al.

В 2014 году группой учёных из Каталонии был предложен [18] новый алгоритм выделения сообществ. Этот метод разрабатывался для анализа больших социальных сетей, поэтому акцент был сделан на высокую производительность, достигаемую путём распараллеливания. Иногда в англоязычной литературе фигурирует под названием SCD (Scaled community detection) algorithm.

Сначала строится первичное разбиение графа на сообщества. Для этого вершины упорядочиваются по коэффициенту кластеризации (в случае равенства — по степени) в порядке убывания. Далее идёт последовательный перебор вершин, и сообщества образуются из текущей, а также смежных с ней и не включённых в другие сообщества ранее.

Далее для каждой вершины выбирается наиболее оптимальное действие: удаление этой вершины из её сообщества (Remove), перемещение в одно из сообществ вершин соседей (Transfer) или же оставление на прежнем месте (NoAction). Критерием выбора оптимального действия служит специальная разработанная авторами весовая

функция, формула которой здесь явно не выписывается ввиду её громоздкости и необходимости введения нескольких дополнительных понятий, однако приводится в статье [18].

Следующий шаг алгоритма заключается в том, что к имеющемуся разбиению на сообщества применяются изменения, полученые на предыдущем шаге. Методология вычислительного процесса позволяет применять изменения к вершинам графа параллельно.

Алгоритм заканчивает свою работу, когда оптимальным действием для каждой из вершин является NoAction.

2.2.6. Марковский алгоритм кластеризации

Алгоритм, в зарубежной литературе называемый Markov Cluster Algorithm или сокращённо MCL, был предложен [23] в 2000 году в Нидерландах.

Определение. Правой стохастической матрицей графа (англ. right stohastic matrix) называется матрица, получаемая из деления всех строчек матрицы смежности этого графа на степень вершины, соответствующей этой строчке.

Алгоритм имеет несколько вариаций, в данной работе описание приводится подобно тому, как это сделано в статье [29].

На первом шаге алгоритма (Expansion, распространение) правая стохастическая матрица графа возводится в целую степень (как правило, вторую). На втором шаге (Inflation, накачивание) каждый элемент получившейся матрицы, возводится в некоторую степень α по правилу Адамара (качество получаемого результата сильно зависит от грамотности выбора этого параметра) и приводится к правому стохастическому виду.

Первый и второй шаги последовательно чередуются до тех пор, пока матрицы распространения и накачивания одного и того же шага не совпадут. Вершины, образующие компоненты слабой связности ориентированного графа, задаваемого получившейся матрицей как матрицей смежности, считаются принадлежащими одному сообществу.

Серьёзными недостатками алгоритма являются сильная зависимость качества результата от выбора параметра α и высокая вычислительная сложность, порядка $O(V^3)$.

2.2.7. Алгоритм выделения перекрывающихся сообществ

Ещё один способ выделения сообществ в графах был предложен Джеффри Баумсом, Марком Гольдбергом и Маликом Магдон-Исмаилом в их совместной статье [2]. Его принципиальным отличием от всех остальных описанных в этой работе алгоритмов является возможность распознавать перекрывающиеся сообщества, то есть допускать принадлежность одной вершины сразу нескольким сообществам. Такие ситуации нередко встречаются на практике, поэтому важно уметь их обрабатывать.

Идея алгоритма состоит в следующем:
1) Выбрать начальное разбиение графа на сообщества, в простейшем случае — каждая вершина есть отдельное сообщество.
2) Для каждого начального сообщества запустить процесс, названный разработчиками Iterative Scan Algorithm [2]: в сообщество в некотором порядке происходит добавление смежных вершин и, при необходимости, удаление некоторых изначальных. Процесс останавливается, когда весовая функция алгоритма, которую можно определить несколькими способами [3], перестаёт возрастать.

В статье, описывающей алгоритм [2], разработчики сами указали его довольно существенный недостаток: точность результата сильно зависит от выбора изначального разбиения графа. К безусловным преимуществам относится возможность распознавать перекрывающиеся сообщества, которые нередко встречаются в реальных социальных сетях.

2.3. Сравнительный анализ алгоритмов

В рамках данного обзора будут рассмотрены три работы, в которых производится сравнительный анализ алгоритмов выделения сообществ. О существовании других серьёзных научных работ схожего содержания автору дипломной работы неизвестно.

В первой работе [10] авторы тестируют двенадцать алгоритмов выделения сообществ, перечисленных с приведением краткого описания, на классическом тесте Гирвана-Ньюмана и тесте Лансичинетти, описание которых приводятся в этой же статье. Кроме того, исследуется поведение алгоритмов на случайных графах, то есть таких, что степени любых двух вершин друг от друга не зависят.

Авторы приходят к выводу: наиболее качественные разбиения на сообщества демонстрируют методы Infomap [21] и Louvain. Также в заключении работы они отмечают, что их публикация есть лишь первый шаг в поиске оптимального алгоритма разбиения графа на сообщества.

Другая статья [17], написанная турецкими специалистами, также сравнивает ал-

горитмы на модели Лансичинетти со специально подобранными параметрами, максимально моделирующими реальные социальные сети. Рассматриваются пять алгоритмов. В качестве критерия оценки качества используется вводимый в статье NMI (Normalized Mutual Information), и наиболее высокой оценки удостаивается Louvain-метод. Однако авторы указывают на ряд недостатков NMI как критерия качества разбиения и отмечают, что хорошим продолжением работы могло бы быть усовершенствование этого критерия.

Третья статья [28] появилась год назад и представляет собой выпускную квалификационную работу студента МГУ К.А. Славнова. В ней автор сравнивает результаты семи алгоритмов выделения сообществ на тех же моделях социальных сетей, что и авторы первой из обозреваемых работ. Своими результатами тестирования Славнов подтверждает выводы итальянских учёных о высоком качестве Louvain-метода, однако приходит к выводу о достаточно низком качестве метода Infomap.

Несмотря на существование трёх исследований, посвящённых сравнительному анализу алгоритмов выделения сообществ в графе, их результаты нельзя принимать окончательными и точными для рассматриваемой нами задачи, потому что:

1) Модельные графы, используемые во втором и третьем из вышеперечисленных исследований, являются невзвешенными, в то время как в задачах компьютерной криминалистики крайне важно учитывать, насколько тесным было взаимодействие, то есть вес ребра, вычисленный при построении графа по специальному алгоритму.

2) Ни одна из статей не анализирует алгоритм Prat-Perez et al. и некоторые другие, разработанные после 2011 года.

3) Авторы работ приходят к несколько различным выводам: А. Лансичинетти в своей работе утверждает, что Infomap — один из лучших методов, тогда как по мнению К.А. Славнова, качество результатов этого алгоритма весьма невысокое.

Выявление указанных недостатков объясняет необходимость дополнительного исследования по сравнения алгоритмов выделения сообществ, которое было проведено в рамках данной дипломной работы.

3. Описание проводимого исследования

В рамках проводимого исследования были реализованы алгоритмы выделения сообществ в графах, перечисленные во второй части обзора, и их качество было оценено по двум параметрам: скорости работы и корректности предоставляемого разбиения. После этого рассматривались возможности совместного использования алгоритмов.

Важно отметить, что производительность алгоритма в контексте решаемой задачи играет значительно меньшую роль, чем качество предоставляемого результата разбиения графа на сообщества по следующим причинам.

Во-первых, криминалистические графы имеют, как правило, сравнительно небольшой размер (не более 1000 вершин), а значит использование алгоритмов, имеющих, например, сложность порядка $O(V^2)$, вполне допустимо, тогда как для анализа социальных сетей вроде Facebook такая асимптотика совершенно недопустима. Во-вторых, современные средства компьютерной криминалистики нередко допускают достаточно длительный процесс анализа данных (до нескольких часов) и на роль системы реального времени никоим образом не претендуют.

Поэтому результаты разбиения графов на сообщества в первую очередь оценивались с точки зрения качества.

3.1. Критерии качества

Устоявшегося определения, что есть сообщество, до сих пор не сформировалось [9], поэтому невозможно определить идеальное разбиение для произвольного графа или предложить универсальную метрику качества выдаваемого алгоритмом разбиения.

При решении поставленной задачи было принято решение считать наиболее удачным то разбиение на сообщества, которое будет максимально удобно пользователю. Поэтому несколько графов экспертиз были предоставлены разным экспертам-криминалистам, которые указали такие разбиения этих графов, которые были бы максимально удобны и информативны для проведения расследования. А результаты работы алгоритмов выделения сообществ оценивались не по некоторой абсолютной величине, а относительно предложенных экспертами «идеальных» разбиений.

Выделения сообществ, при условии, что для некоторого набора графов известно наиболее качественное разбиение, можно рассматривать как задачу машинного обучения с учителем, поэтому для оценки выдаваемого тем или иным алгоритмом результата естественно использовать стандартные для этого класса задач метрики: точность (англ. precision) и полноту (англ. recall), которые вычисляются [16] по следующим формулам:

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN},$$

где TP — число сообществ экспертного разбиения, выделенных алгоритмом,
FP — число сообществ, отсутствующих в экспертном разбиений, выделенных алгоритмом,
FN — число сообществ экспертного разбиения, отсутствующих среди выделенных алгоритмом.

Сообщество экспертного разбиения S будем считать найденным алгоритмом, если среди тех сообществ, которые им выделены, найдётся такое , что оно содержит более половины вершин S и менее половины вершин любого другого экспертного сообщества.

Точность и полноту будем вычислять по тем же формулам и для каждого из экспертных сообществ, при этом под TP будет пониматься максимальное число вершин экспертного сообщества, выделенных алгоритмом в одно сообщество, под FP и FN — число вершин, добавленных и исключённых алгоритмом из экспертного сообщества соответственно.

3.2. Сравнительный анализ алгоритмов

В рамках проводимого исследования алгоритмы выделения собществ в графах, рассмотренные в обзоре, были реализованы и протестированы на имеющемся наборе графов компьютерно-технических экспертиз.

Структура сравнительного анализа устроена следующим образом: сначала описываются особенности алгоритмов, выявленные в процессе тестирования, после чего приводятся показатели выбранных метрик качества для разбиений, выдаваемых ими на графах тестового набора. В заключение для одного из таких графов изображаются структуры сообществ, выделенные всеми рассматриваемыми алгоритмами.

3.2.1. Марковский алгоритм кластеризации

Марковский алгоритм кластеризации (разработанный хронологически первым из всех рассматриваемых), показывает в целом достаточно неплохие разбиения, визуально похожие на экспертные. Однако на некоторых графах наблюдается проблема, ранее в литературе уже описанная [29] [24]: выдаваемая алгоритмом структура состоит из одного большого сообщества и некоторого количества маленьких, содержащих две-три вершины, тогда как экспертное разбиение содержит несколько крупных сообществ. Подобные разбиения не позволяют сформировать представление о структуре и ключевых свойствах исследуемого социального графа.

Таким образом, достаточно часто алгоритм выделяет правдоподобную структуру сообществ, однако нередки случаи и малоинформативных для эксперта-криминалиста разбиений.

3.2.2. Алгоритм Гирвана-Ньюмана

Для метода Гирвана-Ньюмана (также одного из наиболее ранних алгоритмов выделения сообществ) в разных источниках приводятся различные условия окончания работы алгоритма. Предлагаются следующие варианты:
1) *Первое уменьшение модулярности.* На тестовых графах довольно часто модулярность сначала незначительно падает, а уже потом начинает расти. Таким образом, практически весь граф трактуется как одно сообщество, что не даёт представления о его структуре.

2) *Достижение максимума модулярности.* Результаты тестирования показали, что максимум модулярности достигается, как правило, в тот момент, когда в графе остаётся лишь несколько рёбер. Выдаваемый набор сообществ в таком случае практически полностью состоит из отдельных вершин и также не отражает реальной структуры сообществ графа.

3) *Достижение локального максимума модулярности.* Точек локального максимума достаточно много, поэтому в отсутствие явных методов одной из них приходится делать случайный выбор, что также делает качество результата работы алгоритма крайне низким.

3.2.3. Алгоритм Радиччи

Алгоритм Радиччи отличается от алгоритма Гирвана-Ньюмана лишь тем, что вместо центральности по посредничеству в качестве способа определения удаляемого на данной итерации ребра используется коэффициент его кластеризации. Соответственно, все описаннные выше проблемы алгоритма Гирвана-Ньюмана наследуются. Более того, выбор иной метрики подбора удаляемого ребра приводит к ещё менее качественному результату результирующего разбиения, давая выигрыш лишь в эффективности вычислений.

3.2.4. Louvain-метод

Результаты тестирования показали, что этот алгоритм неслучайно был признан в трёх статьях, посвящённых сравнительному анализу методов выделения сообществ, одним из лучших: наряду с высокой производительностью алгоритм показал наиболее качественные результаты разбиения тестовых графов на сообщества среди всех

исследуемых.

Тем не менее, предоставляемые результаты визуально достаточно сильно отличались от тех, которые хотели бы видеть эксперты, поэтому к применению Louvain-метода в его классическом варианте было принято решение не прибегать.

3.2.5. Алгоритм Prat-Perez et al.

Этот алгоритм выдаёт правдоподобные разбиения, однако к нему можно также отнести замечание о недостаточном визуальном сходстве с экспертной версией, сделанное выше для Louvain-метода.

3.2.6. Алгоритм выделения перекрывающихся сообществ

Качество выдаваемого алгоритмом разбиения оказывается сильно зависимым от начального разбиения. в случае его самостоятельного использования (то есть считая изначальными сообществами отдельные вершины) низкое: сообщества, содержащие более трёх-четырёх вершин практически не выделяются, число выделенных «перекрытий» в сообществах значительно превышает визуально просматривающееся.

3.2.7. Сравнение алгоритмов по критериям качества

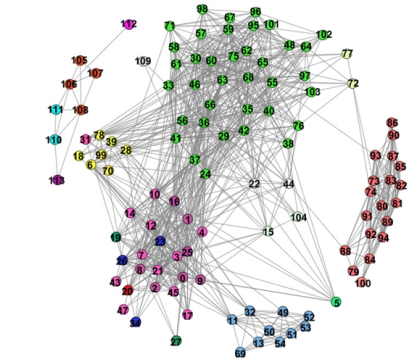
Рассмотрим сначала результаты работы алгоритмов на одном из графов тестового набора: на рис. 2 изображены разбиения на сообщества, выдаваемые каждым из методов, а также экспертное разбиение. В таблице 1 приводятся показатели метрик: точности и полноты для разбиения в целом и для отдельных сообществ. Под точностью и полнотой в сообществах понимается среднее значение этих метрик по всем экспертным сообществам.

Алгоритм	Точность	Полнота	Точность в сообщ.	Полнота в сообщ.
Гирвана-Ньюмана	0.250	0.800	1	0.793
Радиччи	0.250	0.800	1	0.793
Марковский	0.800	0.800	0.730	0.796
Prat-Perez	0.250	0.600	1	0.744
Louvain	0.310	0.800	1	0.790

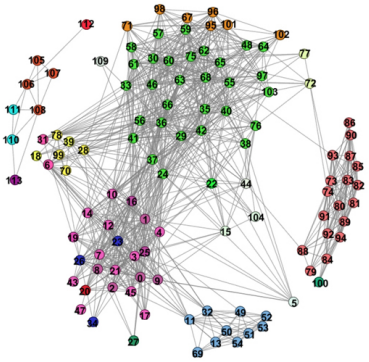
Таблица 1: Значения метрик качества на графе, изображённом на рис. 2

Для всего же набора тестовых графов средние значения показателей приведены в таблице 2.

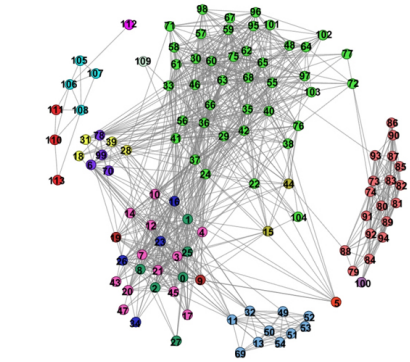
Полученные характеристики позволяют сделать вывод: некоторые из рассмотренных алгоритмов выделения сообществ показывают достаточно точные результаты на



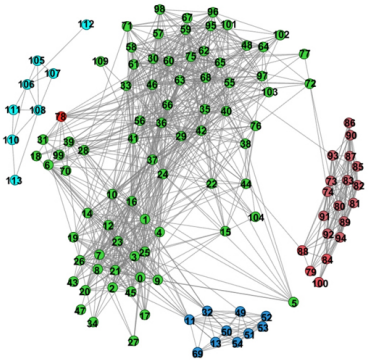
Алгоритм Гирвана-Ньюмана



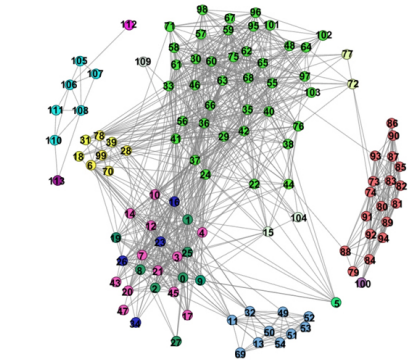
Алгоритм Радиччи



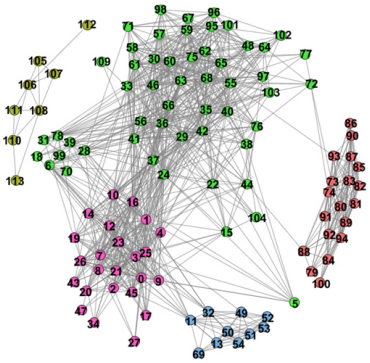
Алгоритм Prat-Perez et al.



Марковский алгоритм кластеризации



Louvain-метод



Экспертное разбиение

Рис. 2: Результаты работы алгоритмов на одном из графов тестового набора

графах, возникавших в реальных криминалистических экспертизах, однако ни один из них не показал на подавляющем большинстве тестов показателей выбранных метрик, достаточно близких к единице, и разбиений, слабо отличимых визуально от экспертных.

Алгоритм	Точность	Полнота	Точность в сообщ.	Полнота в сообщ.
Гирвана-Ньюмана	0.221	0.342	0.930	0.642
Радиччи	0.214	0.387	0.942	0.516
Марковский	0.620	0.445	0.515	0.657
Prat-Perez	0.389	0.616	0.817	0.711
Louvain	0.443	0.815	0.872	0.850

Таблица 2: Средние значения метрик для тестового набора

3.3. Предлагаемый алгоритм

3.3.1. Совместное использование алгоритмов

Для повышения качества структуры сообществ графа и приближения её к образцовой с точки зрения пользователя автором данной дипломной работы был предложен способ совместного использования результатов работы нескольких алгоритмов.

Пусть $G = (V, E)$ - граф,
$A = (A_1, A_2, ..., A_n)$ - алгоритмы выделения сообществ,
$P = (P_1, P_2, ..., P_n)$ - их результаты работы на графе G, где
$P_i = (P_1^i, P_2^i, ..., P_{k_i}^i)$ - список сообществ i-го разбиения.

Построим начальное разбиение анализируемого графа на сообщества следующим образом: вершины $u, v \in V$ будем полагать принадлежащими одному сообществу, если

$$\forall P_i \in P \ \exists\, k \in 1..k_i \ : \ u \in P_k^i \wedge v \in P_k^i$$

и относящимися к разным сообществам в противном случае.

Далее на основе построенного разбиения создаётся мультиграф в точности так, как это делается на втором этапе работы Louvain-метода, после чего для этого мультиграфа проделываются все стандартные шаги его работы.

Были изучены различные варианты совместного использования рассмотренных в данной работе алгоритмов для повышения качества результатов и выяснено, что с помощью одного разбиения, созданного по методу Prat-Perez et al., и последующего анализа мультиграфа, сгенерированного на основе этого разбиения, Louvain-методом можно добиться результата, близкого к экспертному с точки зрения как показателей выбранных метрик, так и визуального восприятия. В то же время такая схема уточнения оказывается наименее трудоёмкой с точки зрения производительности.

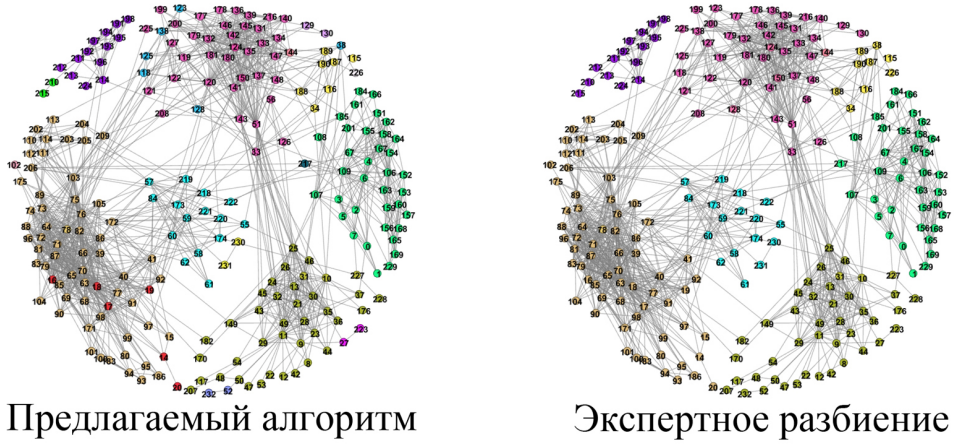


Рис. 3: Сравнение получаемого разбиения с экспертным

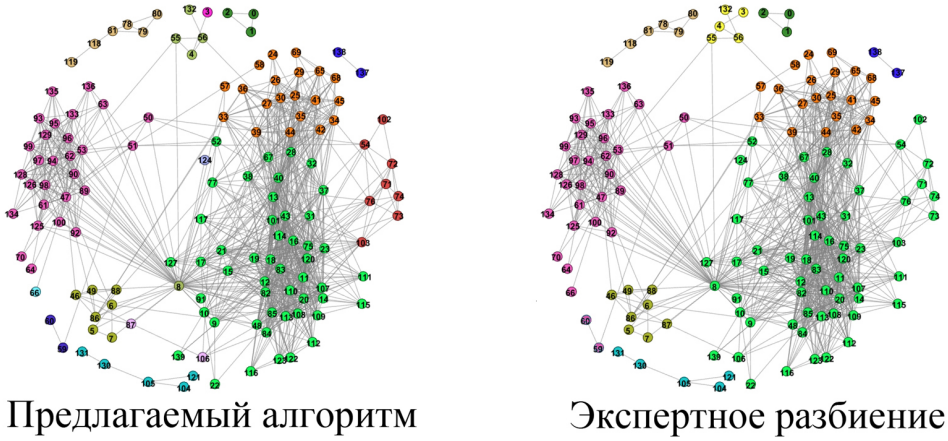


Рис. 4: Сравнение получаемого разбиения с экспертным

3.3.2. Оценка качества

На рис. 3, 4, 5 приводятся три примера сопоставления экспертного разбиения с тем, которое предоставляет предлагаемый алгоритм. Сильное визуальное сходство структур сообществ подтверждается и высокими значениями используемых метрик качества, приведёнными в таблице 3

Алгоритм	Точность	Полнота	Точность в сообщ.	Полнота в сообщ.
Рис. 3	0.615	0.880	0.959	0.865
Рис. 4	0.460	1	1	0.894
Рис. 5	0.625	1	1	0.945

Таблица 3: Значения метрик качества на графах, изображённых на рисунках

Для всего тестового набора средние значения показателей указаны в таблице 4.

Алгоритм	Точность	Полнота	Точность в сообщ.	Полнота в сообщ.
Предлагаемый	0.562	0.933	0.956	0.910

Таблица 4: Средние значения метрик качества для тестового набора

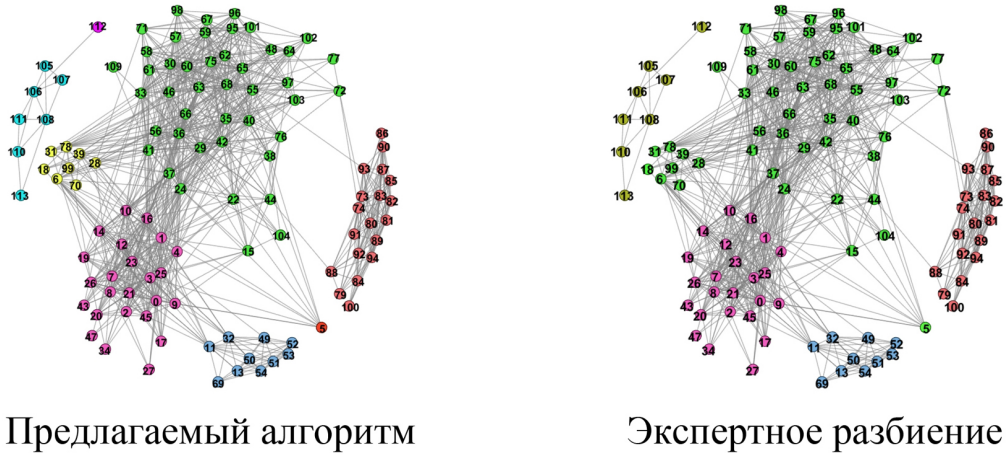


Рис. 5: Сравнение получаемого разбиения с экспертным

Таким образом, предложенная схема совместного использования алгоритмов позволила улучшить качество получаемого разбиения на сообщества. Три из четырёх выбранных показателей принимают значения, близкие к единице, более низкая точность для разбиения в целом обуславливается главным образом особенностью алгоритма «отщеплять» от крупных сообществ одну-две наименее связанные с ним вершины и трактовать их как отдельные сообщества. Это сильно снижает значение точности, однако не оказывает существенного влияния на качество предоставляемого разбиения.

3.3.3. Выводы

Проведённое исследование по сравнению методов выделения сообществ в графах показало, что совместное использование алгоритмов Prat-Perez et al. и Louvain-метода по описанной схеме агрегирования результатов позволяет получать разбиения, достаточно близкие к тем, которые хотели бы использовать в своей работе эксперты-криминалисты. Справедливость этого утверждения подтверждается высокими показателями выбранных метрик качества структуры сообществ графа, а также визуальным сходством предоставляемых разбиений с экспертными образцами.

Предлагаемый алгоритм анализирует все графы тестового набора за интервал времени, не превосходящий указанный в постановке задачи (десять минут), на вычислительной машине с 64-разрядным двухъядерным процессором Intel Core i7-451OU

частотой 2.6 ГГц, вместимостью оперативного запоминающего устройства 8 Гб и установленной операционной системой Windows 8.1, то есть обладающей более низкими техническими характеристиками, чем стандартная машина отечественного эксперта-криминалиста. При этом анализ небольших графов, содержащих около ста вершин, занимает несколько секунд. Таким образом, алгоритм удовлетворяет поставленным требованиям по производительности.

Качественные результаты и приемлемая скорость работы позволяют использовать предложенный алгоритм в качестве решения задачи данной дипломной работы.

Заключение

В результате проведённого исследования был реализован алгоритм выделения сообществ, показывающий качественные результаты и приемлемую скорость работы на графах, возникающих в результате проведения компьютерно-технических экспертиз.

Решение в настоящий момент интегрируется в отечественный продукт компьютерной криминалистики Belkasoft Evidence Center и будет представлено пользователям в одной из ближайших версий продукта.

Результаты работы докладывались на Всероссийской научной конференция по проблемам информатики «СПИСОК-2016» [25] и конференции «Современные технологии в теории и практике программирования» [26], где были удостоены диплома II степени.

Список литературы

[1] AccessData. Summation Feature Friday with Tim Leehealey: Visualization. — 2016. — URL: https://www.youtube.com/watch?v=k6sPnzSF6O8&index=3&list=WL (online; accessed: 28.03.2016).

[2] Baumes J., Goldberg M., Magdon-Ismail M. Efficient identification of overlapping communities. — Rensselaer Polytechnic Institute, 2005.

[3] Baumes J., Krishnamoorthy M.S. Finding communitites by clustering a graph into overlapping subgraph. — Rensselaer Polytechnic Institute, 2005.

[4] Blondel V. et al. Fast unfolding of communities in large networks. — An IOP and SISSA journal, 2008.

[5] Brandes U. On variants of shortest-path betweeness centrality on their generic computation. — University of Konstanz, 2007.

[6] Fortunato S. Community detection in graphs. — Physics Reports, 2010.

[7] IBM. Анализ и визуализация данных для эффективной аналитики. — 2016. — URL: http://www-03.ibm.com/software/products/ru/analysts-notebook (дата обращения: 28.03.2016).

[8] Kernighan B.W., Lin S. An Efficient Heuristic Procedure for Partitioning Graphs. — Bell System Tech. Journal, 1970.

[9] Krings G. Extraction of information from large networks. — Louvain University, 2012.

[10] Lancichinetti A., Fortunato S. Community detection algorithms: a comparative analysis. — Physical review, 2009.

[11] Le Martelot E., Hankin C. Fast milti-scale detection of relevant communities in large-scale networks / Ed. by Brian Skyrms. — Oxford University Press, 2013.

[12] MacQueen J.B. Some methods for classification and analysis of multivariate observations / Ed. by L.M. Cam, J. Neyman. — University of CAlifornia Press, 1967.

[13] Newman M.E.J. Modularity and community structure in networks / Ed. by Brian Skyrms. — The National Academy of Sciences of the USA, 2006.

[14] Newman M.E.J., Girwan M. Finding and evaluating community structure in networks. — Physical review, 2004.

[15] Nuix. ADF Solutions and Nuix Investigator. — 2016. — URL: https://www.nuix.com/sites/default/files/Fact_Sheet_Nuix_and_ADF_Solutions_WEB_US.pdf (online; accessed: 28.03.2016).

[16] Olson D.L., Delen D. Advanced Data Mining Techniques. — Springer, 2008.

[17] Orman G., Labatut V., Cherifi H. Qualitative Comparison of Community Detection Algorithms. — International Conference on Digital Information and Communication Technology and its Applications, 2011.

[18] Prat-Perez A., Domingues-Sal D., Larriba-Pei J. High Quality, Scalable and Parallel Community Detection for Large Real Graphs. — Proceedings of the 23rd international conference on World wide web, 2014.

[19] Radicchi F. et al. Defing and identifying communities in networks / Ed. by Giorgio Parisi. — The National Academy of Sciences of the USA, 2004.

[20] Reichardt J., Bornholdt S. Statistical mechanics of community detection. — University of Bremen, 2008.

[21] Rosvall M., Axelsson D., Bergstrom C.T. The map equation. — The European Physical Journal-Special Topics, 2009.

[22] Suaris P.R., Kedem G. An algorithm for quadrisection and its application to standard cell placement. — IEEE Transactions on Circuits and Systems, 1988.

[23] van Dongen S. Graph Clustering by Flow Simulation. — University of Utrecht, Netherlands, 2000.

[24] Кластеризация графов и поиск сообществ. Часть 1: введение, обзор инструментов и Волосяные Шары // Компания DCA. — 2015. — URL: https://habrahabr.ru/company/dca/blog/265077/ (дата обращения: 06.05.2016).

[25] Куликов Е.К. Выделение сообществ в графах в задачах компьютерной криминалистики. — СПИСОК-2016: материалы всероссийской научной конференции по проблемам информатики, 2016.

[26] Куликов Е.К., Тимофеев Н.М., Губанов Ю.А. Выделение сообществ в графах в задачах компьютерной криминалистики. — Сборник материалов конференции «Современные технологии в теории и практике программирования», 2016.

[27] Оксиджен Софтвер. Мобильный Криминалист Детектив: начало работы. — 2016. — URL: http://www.oxygensoftware.ru/download/articles/Oxygen_Forensic_Detective-Getting_started_RU.pdf (дата обращения: 28.03.2016).

[28] Славнов К.А. Анализ социальных графов. — 2015. — URL: http://www.machinelearning.ru/wiki/images/6/60/2015_417_SlavnovKA.pdf (дата обращения: 06.02.2016).

[29] Федоренко Ю.С. Кластеризация данных на основе нейронного газа и марковских алгоритмов // Молодёжный научно-технический вестник. — 2014. — URL: sntbul.bmstu.ru/file/out/730616 (дата обращения: 03.04.2015).

[30] Характеристика лиц, содержащихся в исправительных колониях для взрослых // Федеральная служба исполнения наказаний России. — 2015. — URL: http://fsin.su/structure/inspector/iao/statistika/Xar-ka%20lic%20sodergahixsya%20v%20IK/ (дата обращения: 06.11.2015).