На правах рукописи

Коршунов Антон Викторович

Исследование структуры сообществ пользователей в графах онлайновых социальных сетей

Специальность 05.13.11 — математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Автореферат
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва — 2015

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институт системного программирования Российской академии наук

Научный руководитель: **Кузнецов Сергей Дмитриевич**
доктор технических наук,
главный научный сотрудник ИСП РАН

Официальные оппоненты:

Воронцов Константин Вячеславович,
доктор физико-математических наук,
старший научный сотрудник, Федеральное государственное бюджетное учреждение науки Вычислительный центр им. А. А. Дородницына Российской академии наук (ВЦ РАН)

Фильченков Андрей Александрович,
кандидат физико-математических наук,
ведущий инженер, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт проблем управления им.В.А.Трапезникова РАН (ИПУ РАН)

Защита состоится «28» мая 2015 года, в 16 часов на заседании диссертационного совета Д 002.087.01 при Федеральном государственном бюджетном учреждении науки Институт системного программирования Российской академии наук по адресу: 109004, Москва, ул. Александра Солженицина, д. 25.

С диссертацией и авторефератом можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки Институт системного программирования Российской академии наук.

Автореферат разослан «27» апреля 2015 года.

Ученый секретарь
диссертационного совета
кандидат физ.-мат. наук Зеленов С. В.

Общая характеристика работы

Актуальность

Онлайновые социальные сети (Facebook, вКонтакте, LiveJournal и другие) являются одним из наиболее популярных типов Интернет-сервисов[1]. Социальные графы пользователей таких сервисов обладают модульной структурой, которая во многом определяется склонностью пользователей объединяться в сообщества.

Сообщество — это группа пользователей, выполняющая общую роль или функцию и обладающая общими свойствами, ценностями и целями. С точки зрения сетевого анализа сообщества представляют собой кластеры пользователей, связанных между собой сильнее, чем с другими пользователями сети.

Знание структуры сообществ пользователей находит применение в ряде практических приложений анализа социальных данных: определение значений скрытых атрибутов пользователей, оптимизация передачи сообщений в коммуникационных сетях, ограничение распространения вредоносного программного обеспечения, идентификация распространителей спам-сообщений, рекомендация товаров, услуг и контента пользователям социальных сетей.

Исследования структурных особенностей групп пользователей в социальных графах показали, что множества пользователей в сообществах имеют тенденцию к значительному пересечению. Более того, количество общих сообществ у пары пользователей увеличивает вероятность образования социальной связи между ними[2]. Эти особенности ограничивают применимость к задаче определения структуры сообществ классических методов кластеризации графов, которые способны идентифицировать в социальном графе только компактные подграфы с незначительным или нулевым пересечением вершин.

В последние годы были предложены методы как для определения структуры сообществ, так и для тестирования качества таких методов путём генерации случайных социальных графов с заданной структурой сообществ, с которой впоследствии сравниваются алгоритмически найденные сообщества[3]. Однако большинство методов обладают значительной вычислительной

[1] В марте 2015 года социальная сеть Facebook сообщает об 1,39 миллиарде, а Twitter — о 288 миллионах пользователей, которые совершают какие-либо действия в сети хотя бы 1 раз в месяц.

[2] Yang Jaewon, Leskovec Jure. Structure and Overlaps of Ground-Truth Communities in Networks // ACM Transactions on Intelligent Systems and Technology (TIST). 2014. Т. 5, № 2. С. 26.

[3] Xie Jierui, Kelley Stephen, Szymanski Boleslaw K. Overlapping community detection in networks: The state-of-the-art and comparative study // ACM Computing Surveys (CSUR). 2013. Т. 45, № 4. С. 43.

сложностью и плохой масштабируемостью, что ограничивает их применимость для анализа социальных графов из сотен миллионов пользователей.

Известные методы генерации тестовых данных не учитывают перечисленных выше свойств структуры сообществ, что требует пересмотра требований к таким методам и разработки новых подходов для более достоверной оценки качества методов определения структуры сообществ. Кроме того, неизвестны методы, способные синтезировать графы из сотен миллионов вершин, что обуславливает дополнительные требования к масштабируемости подобных методов.

Немногочисленные масштабируемые методы определения структуры сообществ либо обладают значительной вычислительной сложностью, либо неспособны находить пересекающиеся сообщества, либо имеют тенденцию к значительному ухудшению качества с увеличением количества сообществ у пользователей.

Целью диссертационной работы является разработка моделей, методов и программных средств для исследования структуры сообществ пользователей в графах онлайновых социальных сетей. Разрабатываемые модели, методы и программные средства должны сочетать низкую вычислительную сложность, хорошую масштабируемость и высокое качество работы вне зависимости от количества сообществ у пользователей.

Для достижения поставленной цели были поставлены и решены следующие **задачи**:

1) исследовать структурные свойства сообществ пользователей в графах онлайновых социальных сетей, методы определения структуры сообществ пользователей, а также методы генерации случайных графов, обладающих свойствами социальных графов и заданной структурой сообществ пользователей;
2) разработать и реализовать метод генерации случайных социальных графов с заданной структурой сообществ пользователей;
3) разработать и реализовать метод определения структуры сообществ пользователей в социальном графе;
4) провести экспериментальное исследование качества, производительности и масштабируемости разработанных методов, а также оценку их применимости для решения прикладных задач.

Основные положения, выносимые на защиту:

1) разработан распределённый метод СКВ для генерации случайных социальных графов с заданной структурой сообществ пользователей;

2) разработан распределённый метод EgoLP для определения структуры сообществ пользователей в социальном графе;

3) для экспериментального подтверждения эффективности предложенных методов реализованы прототипы систем для определения структуры сообществ пользователей и генерации случайных социальных графов с заданной структурой сообществ пользователей[4]. Реализованные прототипы позволили подтвердить высокое качество предложенных методов и соответствие экспериментальных оценок производительности теоретическим оценкам вычислительной сложности.

Научная новизна

В диссертационной работе предложены два новых метода исследования структуры сообществ пользователей в социальных графах.

Метод CKB позволяет осуществлять распределённую генерацию случайных социальных графов с заданной структурой сообществ пользователей, обладающей характерным для реальных социальных сетей набором свойств. Параметрами метода являются количество пользователей и параметры распределения размеров сообществ и распределения количества сообществ у пользователя. Кроме того, предусмотрена возможность управления вероятностью ребра в сообществе в зависимости от его размера, а также регуляции среднего коэффициента кластеризации вершин в сообществе. Экспериментально продемонстрировано, что синтезируемые графы обладают всеми описанными в работе свойствами социальных графов с сообществами. Программная реализация метода обладает масштабируемостью, близкой к линейной, что при достаточном размере вычислительного кластера позволяет генерировать графы из сотен миллионов вершин за несколько часов. Таким образом, предложенный метод превосходит известные методы по совокупности масштабируемости и количества поддерживаемых свойств социальных графов с сообществами.

Метод EgoLP позволяет определять структуру сообществ пользователей в социальном графе. Основой метода является итеративная пересылка меток сообществ по рёбрам графа в соответствии с установленными правилами взаимодействия вершин. Экспериментально продемонстрировано, что предложенный метод превосходит известные методы по совокупности критериев: а) близость определённой структуры сообществ с заранее известной; б) точность решения прикладной задачи определения скрытых атрибутов пользователей с использованием информации о сообществах; в) вычислительная сложность; г) масштабируемость.

[4]Веб-демонстрация прототипа метода CKB доступна по адресу: http://ckb.at.ispras.ru/home/

Теоретическая и практическая значимость

Теоретическая значимость работы заключается в следующем:

- предложены способы вычисления свойств используемых моделей случайных графов: вероятность ребра кратности ≥ 2 в случайном двудольном графе “пользователь-сообщество”, а также средняя степень вершины в случайном социальном графе с сообществами;
- косвенно подтверждена гипотеза о значительном пересечении сообществ контактов индивидуального пользователя с сообществами социального графа, в которых состоит данный пользователь.

Разработанный в диссертационной работе метод `EgoLP` позволяет определять структуру сообществ пользователей в масштабе всей популяции социальной сети (сотни миллионов пользователей), обеспечивая при этом возможность решения практических задач, связанных с использованием знаний о сообществах. Одним из этапов предложенного метода является определение структуры сообществ среди непосредственных контактов каждого пользователя. Полученные сообщества могут использоваться пользователями в качестве замены ручной группировки контактов для оптимизации потоков информации на персональных страницах пользователей.

Кроме того, разработанный метод `CKB` для генерации случайных социальных графов с заданной структурой сообществ пользователей позволяет, в отличие от известных аналогичных методов:

- создавать в случайном социальном графе структуру сообществ, обладающую характерными свойствами сообществ пользователей реальных социальных сетей;
- исследовать качество методов определения структуры сообществ на графах из сотен миллионов вершин.

Таким образом, можно ожидать, что генерируемые с помощью предложенного метода тестовые данные будут применяться исследователями для оценки качества и усовершенствования методов определения структуры сообществ пользователей в социальных графах.

На основе предложенных методов были поданы заявки на патенты:

- заявка на патент P20140009930 “Fast and Distributed Detection for Overlapping Community”, подана в Республике Корея 27.01.2014 г.;
- заявка на патент 2014117945 “Способ и устройства для распределённой генерации случайных социальных графов со структурой пересекающихся сообществ пользователей”, подана в РФ 05.05.2014 г.

Апробация работы

Основные результаты диссертационной работы докладывались в рамках следующих мероприятий:

- сто шестьдесят третье (30 мая 2013 года) заседание Московской секции ACM SIGMOD (ВМК МГУ, г. Москва);
- 15-я Всероссийская научная конференция “Электронные библиотеки: перспективные методы и технологии, электронные коллекции” — RCDL’2013 (14-17 октября 2013 года, г. Ярославль);
- 5-й симпозиум по сложным сетям CompleNet-2014 (12-14 марта 2014 года, г. Болонья, Италия);
- 10-я Международная конференция “Интеллектуализация обработки информации-2014” (4-11 октября 2014 года, о. Крит, Греция);
- 4-й симпозиум по интеллектуальному анализу сетевых данных DaMNet-2014 (14 декабря 2014 года, г. Шэньчжень, КНР).

Кроме того, результаты работы обсуждались в рамках семинаров “Распределенные объектно-ориентированные системы” в Институте системного программирования РАН, а также на семинаре по анализу социальных сетей Института проблем управления РАН.

Диссертационная работа выполнена при поддержке гранта РФФИ №13-07-12134 офи_м “Исследование и разработка методов распределенной обработки больших баз графовых данных”.

Личный вклад

Все выносимые на защиту результаты получены лично автором. Программные реализации выполнены совместно с Кириллом Чихрадзе и Назаром Бузуном.

Публикации

Основные результаты по теме диссертации опубликованы в 8 печатных работах, из которых 2 статьи опубликованы в рецензируемых журналах, рекомендованных ВАК РФ [1,2], 4 статьи включены в реферативную базу данных Scopus [2,3,4,8].

В работах [4,5,6] автору принадлежат обзорные разделы и описание основных элементов разработанных методов. В статьях [7,8] автором написаны обзорные разделы. В статье [1] автору принадлежит раздел, посвящённый генерации случайных социальных графов с сообществами пользователей, а также поиску сообществ пользователей. В работе [2] автору принадлежит раздел, посвящённый исследованию применимости эго-сообществ

пользователей для решения задачи рекомендации пользователям получателей электронных сообщений.

Структура и объём диссертации

Диссертация состоит из введения, четырёх глав, заключения и двух приложений. Полный объем диссертации **134** страницы текста с **92** рисунками и **9** таблицами. Объём приложений составляет **30** страниц. Список литературы содержит **78** наименований.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, формулируется цель, ставятся задачи работы, аргументированы научная новизна и практическая значимость полученных результатов, представлены выносимые на защиту научные положения.

Первая глава содержит обзор предметной области, вводится определение сообщества пользователей, приводится описание структурных свойств социальных графов с сообществами, а также используемого в работе способа оценки соответствия алгоритмически найденных сообществ интуитивным представлениям о группах пользователей, связанных социальными связями.

В современном понимании *онлайновая социальная сеть* — это Интернет-сервис, позволяющий пользователям публиковать на своих страницах персональные и иные данные и служащий для упрощения коммуникации и обмена информацией между пользователями сети Интернет.

Помимо коммуникационной функции, сервисы социальных сетей играют роль баз пользовательских данных, в которых с каждым пользователем ассоциирован набор персональной информации, составляющий его "виртуальную личность". Данные всех пользователей одного сервиса образуют его *социальный граф*, — динамическую структуру, полностью описывающую состояние и поведение составляющих её пользователей, а также их отношения между собой и объектами внешнего мира в некоторый момент времени. Вершинами графа принято считать пользовательские аккаунты, а рёбрами — социальные связи типа “дружба”, “подписка”, “следование”, которыми пары пользователей явно связывают свои аккаунты.

Являясь по сути виртуальным отражением части человеческого социума, социальные сети во многих случаях наследуют характерную для него структуру социальных групп, или *сообществ*. Наличие сообществ пользователей является распространённым свойством современных онлайновых социальных сетей вне

зависимости от состава аудитории, природы связей и преобладающих сценариев использования: общение, обмен контентом, поиск информации, развлечения и т.д. Например, часто можно встретить сообщества, в которых участников объединяют общие интересы, политические и религиозные предпочтения, географическая близость и т.д.

Таким образом, с точки зрения сетевого анализа сообщества представляют собой *кластеры* пользователей, связанных между собой сильнее, чем с другими пользователями сети. С функциональной точки зрения *сообщество* — это группа пользователей, выполняющая общую роль или функцию и обладающая общими свойствами, ценностями и целями.

Ввиду отсутствия общепринятой формализации приведённого определения сообщества, важно выделить некоторые особенности сообществ пользователей, характерные для социального графа на уровне связей между пользователями. Формализация задачи и метода поиска сообществ на уровне сетевых данных позволяет использовать общепринятую терминологию теории графов и соответствующие инструменты.

Рассмотрим неориентированный граф $G(V, E)$, где $|V| = n$ и $|E| = m$. Пусть V' —- некоторое подмножество V и пусть E' —- подмножество всех ребер графа G, концевые вершины которых входят в V'. Тогда граф $G' = (V', E')$ называется *вершинно-порождённым подграфом* графа G. В случае социального графа V' соответствует группе пользователей, а E' — всем связям между ними в социальном графе.

Сообществом пользователей $Z_c(V_c, E_c)$ будем называть любой вершинно-порождённый подграф социального графа $G(V, E)$.

Покрытием $\mathbb{C}$ социального графа $G(V, E)$ будем называть множество сообществ пользователей, заданных для G: $\mathbb{C} = \{Z_c\}_{c=1}^{K}$, причём $\forall c : V_c \subseteq V$, $E_c \subseteq E$.

Рассмотрим двудольный граф $B(V, \mathbb{C}, M)$, где V соответствует множеству вершин социального графа G, $\mathbb{C}$ — покрытие, а ребро $(u, c) \in M$ соединяет вершину $u \in V$ с сообществом $Z_c \in \mathbb{C}$, если $u \in V_c$. Степень m_u вершины u равна количеству сообществ, в которых состоит пользователь u. Степень $n_c = |V_c|$ вершины Z_c называется *размером сообщества* и равна количеству пользователей, которые состоят в сообществе Z_c.

Mislove et al[5] и Yang et al[6] исследовали реальные данные о публичных пользовательских группах из социальных сетей Friendster, LiveJournal, Orkut и

[5]Measurement and analysis of online social networks / Alan Mislove, Massimiliano Marcon, Krishna P Gummadi [и др.] // Proceedings of the 7th ACM SIGCOMM conference on Internet measurement / ACM. 2007. С. 29–42.

[6]Yang Jaewon, Leskovec Jure. Structure and Overlaps of Ground-Truth Communities in Networks // ACM Transactions on Intelligent Systems and Technology (TIST). 2014. Т. 5, № 2. С. 26.

YouTube. По результатам исследований были выявлены следующие свойства социальных графов с сообществами:

1. Групповые свойства сообществ:
    - множества вершин сообществ могут пересекаться;
    - распределение размеров сообществ подчиняется степенному закону;
2. Связь пользователей и сообществ:
    - распределение количества сообществ, в которых состоит пользователь, подчиняется степенному закону;
    - количество сообществ у пользователя прямо пропорционально количеству его связей с другими пользователями;
3. Связи между пользователями:
    - вероятность ребра между парой вершин увеличивается с ростом количества общих сообществ, которым принадлежат обе вершины;
    - для пары сообществ пересечение их более плотно, чем непересекающаяся часть;
    - количество рёбер в сообществе растёт суперлинейно с размером сообщества;
    - *связующие вершины* сообщества (имеющие среди всех вершин сообщества наибольшее количество связей с другими вершинами из этого сообщества) более вероятно находятся в пересечениях с другими сообществами, чем в непересекающейся области сообщества;
    - средний коэффициент кластеризации вершин в сообществе обратно пропорционален размеру сообщества.

Интуитивная ассоциация сообщества пользователей с кластером вершин социального графа приводит к естественному предположению о том, что вершины “хорошего” сообщества компактно связаны между собой и одновременно хорошо отделены от остальных вершин графа.

В диссертационной работе используется предложенный Yang et al подход к оценке соответствия произвольного подграфа интуитивным свойствам сообщества пользователей. Вводится набор *метрик качества сообществ*, каждая из которых оценивает некоторое желаемое свойство отдельного сообщества: отделимость, плотность, сплочённость, средний коэффициент кластеризации вершин.

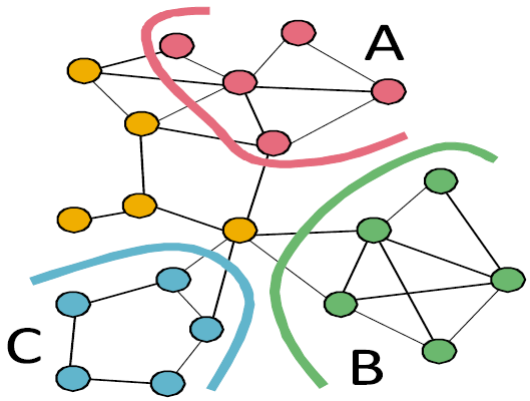

Рис. 1: Сообщества пользователей как кластеры вершин социального графа.

На рисунке 1 изображены 3 подграфа, которые интуитивно идентифицируются как кластеры вершин представленного социального графа, которые могут соответствовать сообществам пользователей. Подграф A имеет 6 внутренних и 5 внешних рёбер, в силу чего обладает более низкой *отделимостью* по сравнению с другими кластерами. В подграфе B представлены 8 из 10 возможных рёбер между его вершинами, что говорит о его высокой *плотности* по сравнению с подграфом A и особенно подграфом C. При этом в подграфе C достаточно удалить 2 ребра, чтобы разделить его на компоненты связности, что говорит о его низкой *сплочённости*. Данный подграф также обладает самым низким *коэффициентом кластеризации*, поскольку ни одна тройка его вершин не образует 3-клику. Следовательно, подграфы C и B по совокупности значений метрик качества являются “худшим” и “лучшим” сообществами соответственно.

Подграф с большим значением какой-либо метрики качества не обязательно соответствует сообществу. Однако сообщество, найденное алгоритмически, должно иметь высокие показатели по одной или более метрик качества.

Во **второй главе** приводится обзор методов определения структуры сообществ, а также способов оценки качества таких методов.

В первом разделе второй главы дан обзор существующих методов определения структуры сообществ пользователей и дана оценка их применимости к реальным социальным графам.

По результатам исследования в рамках диссертационной работы были выделены три основных класса алгоритмов, которые обеспечивают хорошую точность при определении структуры значительно пересекающихся сообществ. Эти классы включают методы, основанные на вероятностных моделях, методы локальной оптимизации и методы распространения меток. Однако большинство методов из первых двух классов обладают значительной вычислительной

сложностью, а распределённая реализация их затруднена в силу различных причин.

В рамках работы был детально исследован метод SLPA (*Speaker-listener Label Propagation Algorithm*)[7] из класса распространения меток как обладающий оптимальным сочетанием качества, временной сложности и масштабируемости. Метод заключается в итеративном распространении по рёбрам меток сообществ, которые накапливаются в памяти каждой вершины и по завершении итераций определяют её принадлежность к сообществам. Таким образом, временная сложность метода линейно зависит от количества рёбер в исходном графе, а возможность распределять вычисления по независимым машинам с периодической синхронизацией облегчает масштабируемую реализацию.

Вместе с тем, детальное исследование результатов работы SLPA выявило неспособность метода разделять значительно пересекаюшиеся сообщества. В частности, были обнаружены следующие проблемы, которые обуславливают ухудшение качества при увеличении степени пересечения сообществ в исследуемом графе:

1) в случае наличия в графе вершин со степенью, существенно (>3 раз) превышающей среднюю степень вершины в графе, их метки активно распространяются, что ведёт к формированию очень больших сообществ (по сравнению с известными заранее сообществами);

2) недостаточное количество найденных сообществ по сравнению с заранее известными: с увеличением количества итераций количество сообществ в графе уменьшается, предположительно из-за доминирования одной или нескольких популярных меток сообществ среди меток большинства вершин;

3) наличие сообществ с низкой связностью между вершинами, включая несвязные сообщества;

4) если инициализировать вершины в соответствии с заранее известными сообществами (т.е. заранее задать корректное покрытие графа сообществами), то в процессе обмена метками вершины добавляют в свои множества меток много “шумовых” меток, что ведёт к ухудшению качества результатов с увеличением числа итераций.

В процессе разработки предложенного в работе метода `EgoLP` для определения структуры сообществ пользователей (глава 4) целью было

[7]Xie Jierui, Szymanski Boleslaw K, Liu Xiaoming. SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process // 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW) / IEEE. 2011. С. 344–349.

устранение перечисленных недостатков, повышение качества для значительно пересекающихся сообществ, а также сохранение низкой вычислительной сложности и хорошей масштабируемости SLPA.

Во втором разделе второй главы рассматриваются способы оценки качества методов определения структуры сообществ в социальном графе.

Современные исследователи в области идентификации социальных сообществ широко применяют социальные графы с эталонной структурой сообществ для оценки методов определения структуры сообществ. Такие *шаблонные сети* (англ. *benchmark networks*) состоят из набора вершин и рёбер социального графа (пользователи и связи между ними), а также списка сообществ, в которых состоит каждый пользователь. Графы и соответствующие им покрытия могут быть как синтезированы, так и получены из данных реальных социальных сервисов.

В качестве непосредственного критерия качества методов определения структуры сообществ принято использовать некоторую оценку близости двух покрытий для некоторого графа: найденного алгоритмом и *референтного*, то есть заранее заданного или известного. Такой подход позволяет исследовать способность различных методов восстанавливать структуру сообществ, заданную особым способом, зависящим от конкретного приложения или исследовательской задачи.

Для того, чтобы установить близость известного и найденного покрытий, в диссертационной работе используется значение *нормализованной взаимной информации* (англ. *Normalized Mutual Information, NMI*), которое показывает, в какой степени информация о структуре одного покрытия уменьшает неопределённость о другом покрытии. Значения NMI лежат в промежутке $[0; 1]$. Минимальное значение соответствует абсолютно разным покрытиям, максимальное — полностью совпадающим покрытиям.

Наиболее известным хранилищем данных социальных сетей со структурой сообществ является Stanford Large Network Dataset Collection[8]. Однако сбор тестовых данных из реальных социальных сетей является трудоёмким, а свойства полученных наборов данных часто далеки от желаемых. Поэтому принято использовать программные средства для генерации случайных социальных графов со структурой сообществ пользователей, заданной в соответствии с некоторой моделью.

Наиболее распространён метод *LFR*[9], способный создавать неориентированные, ориентированные и взвешенные графы заданного размера

[8] `http://snap.stanford.edu/data/index.html#communities`

[9] Lancichinetti A., Fortunato S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities // Phys. Rev. 2009. T. 80(1).

и с нужными свойствами сообществ (размеры, количество сообществ у пользователей). Вместе с тем, данный метод в недостаточной степени учитывает специфику социальных сетей и неспособен создавать графы с покрытиями, удовлетворяющими всем перечисленным в главе 1 свойствам сообществ пользователей. В частности, в генерируемых графах количество сообществ, к которым принадлежит пользователь, распределено не по степенному закону, а в соответствии с бинарным дискретным распределением, что нехарактерно для социальных сетей. Кроме того, метод LFR имеет существенные ограничения в плане производительности при генерации графов с $> 10^6$ вершин, что затрудняет оценку применимости методов определения структуры сообществ к социальным графам большой размерности.

В конце главы перечислены возможные приложения методов определения структуры сообществ пользователей в социальном графе: определение значений скрытых атрибутов пользователей, оптимизация передачи сообщений в коммуникационных сетях, ограничение распространения вредоносного программного обеспечения, идентификация распространителей спам-сообщений, рекомендация товаров, услуг и контента пользователям социальных сетей и др.

Другим аспектом практического использования полученных сообществ является упрощение анализа сложных сетей больших размеров. Большинство методов анализа не предназначены для обработки графов размером $> 10^6$ вершин. Поэтому для таких графов важно сначала определить один или несколько подграфов, представляющих интерес для дальнейшего анализа.

В этом случае масштабируемый алгоритм определения структуры сообществ может помочь идентифицировать сообщества, соответствующие естественным модулям исследуемой сети (страны, организации), которые могут быть без труда идентифицированы аналитиком. В дальнейшем каждое из выбранных на предварительном этапе сообществ может быть исследовано с помощью других методов анализа структуры сетей. Например, становится возможным анализ структуры связей всего социального графа, определение целевых сообществ пользователей и детальный анализ путей распространения информации в соответствующих им подграфах.

Результаты второй главы опубликованы в работах [1, 4, 5, 8].

В **третьей главе** приведено описание разработанного метода генерации случайных графов, обладающих свойствами социальных графов и заданной структурой сообществ пользователей.

Требовалось разработать метод генерации пар графов $G(V, E)$ и $B(V, \mathbb{C}, M)$ с перечисленными в главе 1 свойствами, характерными для

социальных графов с сообществами пользователей. Кроме того, распределение степеней вершин V должно следовать степенному закону.

Метод должен позволять разбивать задачу генерации рёбер графа $B(V, \mathbb{C}, M)$ на независимые подзадачи, каждая из которых может решаться независимо от других с последующей агрегацией результатов решения всех подзадач. Аналогичное требование предъявляется к способу решения задачи генерации рёбер графа $G(V, E)$. Как следствие, метод должен позволять распределённую программную реализацию с близкой к линейной масштабируемостью в зависимости от $|E|$.

Разработанный метод генерации случайных социальных графов с заданной структурой пересекающихся сообществ пользователей получил название СКВ[10] и состоит из двух последовательных этапов:

1) пользователи распределяются по сообществам;

2) генерируются связи между пользователями внутри каждого сообщества.

На этапе **генерации двудольного графа "пользователь-сообщество"** производится генерация графа $B(V, \mathbb{C}, M)$. Для этого требуется вычислить количество генерируемых рёбер M, а также сгенерировать последовательности степеней для V и $\mathbb{C}$.

Согласно модели генерации, количество сообществ $m_j \in \mathbb{N}$ пользователя j и размер $x_c \in \mathbb{N}$ сообщества Z_c являются случайными величинами со степенным распределением с экспонентами β_1 и β_2:

$$m_j \sim m^{-\beta_1}, \forall m_j : m_{min} \leq m_j \leq m_{max}, \tag{1}$$

$$x_c \sim x^{-\beta_2}, \forall x_c : x_{min} \leq x_c \leq x_{max}. \tag{2}$$

Сначала генерируются ожидаемые последовательности степеней в долях графа $B(V, \mathbb{C}, M)$: $\vec{d^1} = (d^1_1, ..., d^1_{N_1})$ и $\vec{d^2} = (d^2_1, ..., d^2_{N_2})$ для количества сообществ у пользователя и размеров сообществ соответственно. Элементы последовательностей независимо семплируются из распределений $m^{-\beta_1}$ и $x^{-\beta_2}$.

Количество пользователей N_1 задаётся в виде параметра, а количество сообществ N_2 сначала должно быть вычислено исходя из остальных параметров генерации:

$$N_2 = \left\lfloor \frac{N_1 \cdot \mathbb{E}[m_j]}{\mathbb{E}[x_c]} \right\rfloor, \tag{3}$$

где при условии $\beta_1 \neq 2$ и $\beta_2 \neq 2$:

$$\mathbb{E}[m_j] = \frac{(1-\beta_1)(m_{max}^{2-\beta_1} - m_{min}^{2-\beta_1})}{(m_{max}^{1-\beta_1} - m_{min}^{1-\beta_1})(2-\beta_1)}, \tag{4}$$

[10]По первым буквам фамилий разработчиков в латинской транскрипции.

$$\mathbb{E}[x_c] = \frac{(1-\beta_2)(x_{max}^{2-\beta_2} - x_{min}^{2-\beta_2})}{(x_{max}^{1-\beta_2} - x_{min}^{1-\beta_2})(2-\beta_2)}. \quad (5)$$

Далее вычисляется количество генерируемых рёбер $|M|$ между долями графа B:

$$|M| = \lfloor (1 + P_{(c,j)}^{\geq 2}) M_0 \rfloor, \quad (6)$$

где $P_{(c,j)}^{\geq 2}$ — вероятность ребра (c, j) кратности ≥ 2, а M_0 соответствует ожидаемому количеству рёбер:

$$M_0 = \lfloor N_1 \cdot \mathbb{E}[m_j] \rfloor = \lfloor N_2 \cdot \mathbb{E}[x_c] \rfloor. \quad (7)$$

Поскольку рёбра в предложенном методе генерируются независимо, то после удаления кратных рёбер из общего списка по окончании генерации количество рёбер несколько отличается от ожидаемого значения M_0, заданного в соответствии с входными параметрами. В разработанном методе в качестве поправки при расчёте количества генерируемых рёбер используется вероятность генерации ребра кратности ≥ 2. Это позволяет уменьшить погрешность количества сгенерированных рёбер, связанную с удалением кратных рёбер.

Теорема 1. *Вероятность ребра (c, j) кратности ≥ 2 в случайном двудольном графе "пользователь-сообщество" при независимой генерации M_0 рёбер между его долями удовлетворяет*

$$P_{(c,j)}^{\geq 2} \to \frac{\mathbb{E}[x_c^2]\mathbb{E}[m_j^2]}{2M_0^2} \quad (8)$$

при условии $\frac{x_{max} m_{max}}{M_0} \to 0$,
где x_c — случайная величина, соответствующая размеру сообщества Z_c, m_j — случайная величина, соответствующая количеству сообществ у пользователя j.

Для генерации рёбер между долями графа согласно последовательностям степеней $\vec{d_1}$ и $\vec{d_2}$ вычисляются значения

$$D_1^1 = d_1^1, D_2^1 = D_1^1 + d_2^1, ..., D_{k+1}^1 = D_k^1 + d_{k+1}^1, ..., D_{N_1}^1 = D_{N_1-1}^1 + d_{N_1}^1,$$

$$D_1^2 = d_1^2, D_2^2 = D_1^2 + d_2^2, ..., D_{k+1}^2 = D_k^2 + d_{k+1}^2, ..., D_{N_2}^2 = D_{N_2-1}^2 + d_{N_2}^2.$$

Далее для последовательности натуральных чисел

$$[M] = \{1, 2, 3, \ldots, |M|\},$$

выполняется в цикле для $t = 1$ до $|M|$:

1) выбираются случайные числа p и q из $[M]$ равномерно;

2) находится интервал $[D_i^1, D_{i+1}^1]$, которому принадлежит p;

3) находится интервал $[D_j^2, D_{j+1}^2]$, которому принадлежит q;

4) в M добавляется ребро (i, j).

Предложенный алгоритм генерации рёбер позволяет каждому из s узлов вычислительного кластера выполнять независимую генерацию $\frac{|M|}{s}$ рёбер, имея лишь информацию об ожидаемых последовательностях степеней в долях графа. Таким образом, процесс генерации разбивается на независимые подзадачи, по завершении которых все сгенерированные рёбра агрегируются, сортируются по одному из концов, а кратные рёбра удаляются.

Временная сложность этапа генерации двудольного графа $O(|M| \log(N_1 N_2))$.

На этапе **генерации рёбер внутри сообществ** выполняется генерация рёбер в графе $G(V, E)$, которая проходит независимо в каждом из сгенерированных на предыдущем этапе сообществ согласно модели AGM.

Графовая модель принадлежности пользователей к сообществам (англ. *community-affiliation graph model, AGM*) — вероятностная генеративная модель для графов, которая представляет организацию сложных сетей в виде пересекающихся сообществ. Она была разработана Yang et al путём анализа данных о реальных сообществах пользователей в социальных сетях (глава 1). Модель представляет принадлежность вершин к сообществам как двудольный граф, в котором рёбра от пользователя идут к сообществам, которым этот пользователь принадлежит.

Другая часть модели основана на том, что люди принадлежат многим сообществам (друзья, члены семьи, коллеги), но связи между ними часто возникают как результат одной доминирующей причины. Такой характер связей моделируется заданием для каждого сообщества Z_c вероятности p_c, с которой вершина будет соединена ребром с другой вершиной из этого сообщества. В результате каждое сообщество, которому принадлежит пара вершин, имеет независимый шанс создать ребро между этими двумя вершинами. Следовательно, чем большему количеству сообществ принадлежит пара пользователей, тем больше вероятность того, что между ними будет образовано ребро.

Для генерации графа $G(V, E)$ по модели AGM требуется, прежде всего, задать двудольный граф $B(V, \mathbb{C}, M)$, а также параметры $\{p_c\}$ и ϵ (вероятность появления ребра между парой вершин в графе безотносительно сообществ). В разработанном методе используется случайный двудольный граф “пользователь-сообщество”, созданный на предыдущем этапе. Затем нужно сгенерировать необходимое количество ребёр между вершинами из V.

Рёбра в графе $G(V, E)$ генерируются независимо в каждом сообществе Z_c с вероятностью p_c, что позволяет выполнять генерацию одновременно во всех сообществах и упрощает распределённую реализацию:

$$p_c = \frac{\alpha}{x_c^{\gamma}}, \tag{9}$$

где x_c — размер сообщества, $\gamma \in (0; 1)$ — параметр модели, $\alpha > 0$ задаётся в виде параметра либо определяется заданной средней степенью вершины.

Теорема 2. *Ожидаемая средняя степень вершины случайного социального графа из N вершин при независимой генерации его рёбер с вероятностью $p_c = \frac{\alpha}{x_c^{\gamma}}$ в каждом из сообществ удовлетворяет*

$$\mathbb{E}[d_v] \to (N-1) \sum_{r=1}^{k} (-1)^{r+1} \left[\alpha^r \sum_{c_1 < \ldots < c_r} \mathbb{E}_{\xi_{ijc}} \left[\prod_{k=\{1,\ldots,r\}} \frac{x_{c_k}^{2-\gamma} m_i m_j}{M^2} \right] \right] \tag{10}$$

при условии $\frac{x_{max} m_{max}}{M} \to 0$,
где k — максимальная кратность ребра, ξ_{ijc} — индикаторная случайная величина, соответствующая появлению ребра (i, j) в сообществе Z_c, x_c — случайная величина, соответствующая размеру сообщества Z_c, m_i и m_j — случайные величины, соответствующие количеству сообществ у произвольных пользователей i и j, M — случайная величина, соответствующая количеству рёбер в двудольном графе "пользователь-сообщество", $\alpha > 0$ и $\gamma \in (0; 1)$ — константы.

После решения уравнения 10 k-ой степени с переменной α и фиксированной γ можно вычислить вероятности $\{p_c\}$, необходимые для достижения заданной средней степени.

Для генерации рёбер внутри сообществ производится семплирование случайных графов из модели $\mathcal{G}(x_c, p_c)$ Эрдёша-Реньи для каждого сообщества. При этом каждое ребро между x_c вершинами появляется с вероятностью p_c. С целью повышения коэффициента кластеризации генерируемых сообществ используется *модель триплетов*[11], которая в некотором смысле является генерализацией модели Эрдёша-Реньи. Вместо пар вершин рассматриваются все возможные комбинации троек вершин. Каждая тройка вершин может находиться в восьми *конфигурациях* (рисунок 2) в зависимости от наличия или отсутствия рёбер между упорядоченными вершинами.

Положим, что возможные конфигурации независимо распределены со следующими вероятностями: $P(000) = p_0$ (нет рёбер), $P(001) = P(010) =$

[11] Wegner Anatol. Random graphs with motifs. Preprint // Max Planck Institute for Mathematics in the Sciences. 2011.

Рис. 2: Варианты конфигурации для тройки вершин.

$P(100) = p_1$ (одно ребро), $P(011) = P(101) = P(110) = p_2$ (два ребра), $P(111) = p_3$ (клика размера 3).

Для генерации каждого ребра внутри триплета с одинаковой вероятностью p примем $p_0 = (1-p)^3$, $p_1 = p(1-p)^2$, $p_2 = p^2(1-p)$, $p_3 = p^3$. Отметим, что при этом модель триплетов становится эквивалентной исходной модели $\mathcal{G}(x_c, p_c)$ Эрдёша-Реньи для сообщества.

Следовательно, заданная вероятность p_c ребра внутри сообщества связана с вероятностью p ребра в одной из троек сообщества следующим образом:

$$p_c = 1 - (1-p)^{x_c - 2}, \tag{11}$$

где $x_c = |V_c|$.

Получаем вероятность ребра внутри триплета:

$$p = 1 - e^{\frac{\ln(1-p_c)}{x_c-2}} = 1 - e^{\frac{\ln\left(1-\frac{\alpha}{x_c^\gamma}\right)}{x_c-2}}. \tag{12}$$

Для генерации графа по описанной модели T_c раз выбирается случайная тройка вершин в сообществе и генерируется некоторая конфигурация рёбер в каждой выбранной тройке. T_c является случайной величиной, имеющей биномиальное распределение со следующими параметрами:

$$T_c \sim Bin(C^3_{x_c}, 3p_1 + 3p_2 + p_3), \tag{13}$$

Однако при независимом выборе троек вершин существует вероятность повторного выбора одной и той же тройки. Поэтому в разработанном методе выбирается $T_c(1+\delta)$ троек, где поправка δ находится следующим образом:

$$\delta = \frac{\log(1-p_c)}{T_c \log(1 - p_\Delta(1-\bar{p}))} - 1, \tag{14}$$

где $\bar{p}$ — вероятность отсутствия ребра в тройке вершин:

$$\bar{p} = p_0 + 2p_1 + p_2, \tag{15}$$

а p_Δ — вероятность выбора некоторой тройки для заданных вершин i и j:

$$p_\Delta = \frac{x_c - 2}{C^3_{x_c}}. \tag{16}$$

Описанный алгоритм генерации позволяет разбивать задачу генерации конфигураций рёбер в $T_c(1+\delta)$ тройках вершин на независимые подзадачи. В итоге одно, два или три ребра создаются внутри каждой тройки вершин с соответствующими нормализованными вероятностями конфигураций. В завершение кратные ребра удаляются.

Временная сложность этапа генерации рёбер внутри сообществ равна $O(\sum_{c=1}^{N_2} T_c)$.

Предложенный метод был реализован на языке программирования Scala с использованием Apache Spark[12] — фреймворка для распределённых вычислений в распределённой среде[13].

Экспериментальное исследование методов LFR и CKB показало, что CKB-графы имеют наиболее схожую структуру с реальными сетями. Все требуемые структурные свойства социального графа с сообществами (глава 1) выполняются в синтезированных с помощью предложенного метода шаблонных сетях. Кроме того, преимуществом метода CKB является возможность регулировать коэффициент кластеризации вершин в сообществах.

Анализ метрик качества синтезированных сообществ показывает, что CKB-сообщества обладают в среднем более высокими показателями по всем метрикам качества сообществ по сравнению с LFR-сообществами. Вместе с тем, минимальное значение отделимости для LFR-сообществ превышает аналогичный показатель метода CKB. Данный факт можно объяснить большим количеством значительно пересекающихся сообществ в результатах работы предложенного метода. При этом становится сложнее провести чёткие “границы” между отдельными сообществами, что ведёт к ухудшению их отделимости в некоторых случаях. Это обуславливает низкую эффективность некоторых методов определения структуры сообществ, которые рассматривают их как хорошо отделимые плотные подграфы с незначительным пересечением множеств вершин

Для оценки масштабируемости алгоритма был использован кластер Amazon EC2. Показано, что алгоритм имеет близкую к линейной масштабируемость, что озволяет создавать синтетические сети больших размеров за приемлемое время. Так, например, генерация графа с 1 миллиардом вершин заняла 150 минут на 150 машинах кластера Amazon EC2.

Результаты третьей главы опубликованы в работах [1, 4, 5, 6].

В **четвёртой главе** приведено описание EgoLP — разработанного метода определения структуры сообществ пользователей.

[12] `https://spark.apache.org/`

[13] Веб-демонстрация разработанного прототипа доступна по адресу: `http://ckb.at.ispras.ru/home/`

Требовалось разработать метод определения структуры сообществ пользователей в графах онлайновых социальных сетей. Метод должен удовлетворять следующим критериям:

1) высокое качество восстановления заранее известной структуры сообществ;
2) высокая точность решения прикладной задачи с использованием информации о сообществах;
3) вычислительная сложность, не превышающая линейную по числу рёбер графа;
4) близкая к линейной масштабируемость;
5) способность обрабатывать социальные графы с $> 10^6$ вершин и средней степенью > 100, характерной для социальных сетей.

Далее описываются основные этапы разработанного метода `EgoLP`: определение структуры эго-сообществ каждого пользователя, определение структуры сообществ путём распространения меток сообществ, определение подсообществ.

В некоторых задачах социального анализа в качестве входных данных используется не весь социальный граф, а так называемая *эго-сеть* $Ego_v(v, V_v^{ego}, E_v^{ego})$, которая состоит из центрального пользователя v, всех его соседей V_v^{ego} и всех рёбер E_v^{ego} между участниками этой сети. Сообщества непосредственных контактов центрального пользователя v называются *эго-сообществами* $\{\mathcal{E}_{vk}(V_{vk}, E_{vk})\}$, $V_{vk} \subseteq V_v^{ego}$.

В основе метода лежит гипотеза о взаимосвязи мезо- и микроскопического уровней организации социальной сети, на которой основаны некоторые методы определения структуры сообществ. Суть гипотезы состоит в том, что эго-сообщества каждого пользователя значительно пересекаются с сообществами социального графа, в которых состоит данный пользователь. Иными словами, объединение эго-сообществ различных пользователей ведёт к формированию глобальных сообществ, что можно рассматривать как одну из причин образования связей в социальном графе. Данная гипотеза не была проверена экспериментально в рамках диссертационной работы, однако эффективность разработанного метода в применении к реальным и синтетическим данным косвенно подтверждает её справедливость.

На этапе **определения структуры эго-сообществ** для каждого пользователя строится эго-сеть и выполняется модификация алгоритма SLPA для поиска эго-сообществ.

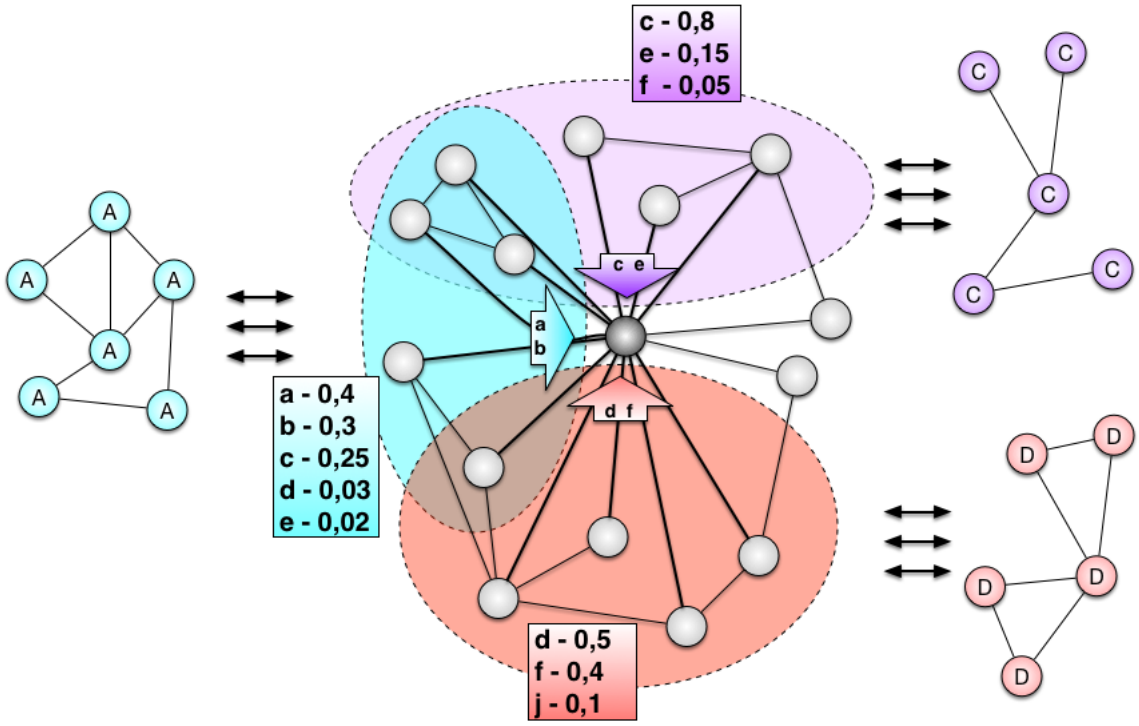

Рис. 3: Распространение меток с использованием эго-сообществ. “Слушающий” узел расположен в центре и получает метки сообществ от “говорящих” узлов — всех остальных вершин своей эго-сети. На каждой итерации алгоритма каждый “слушающий” узел выбирает несколько наиболее популярных меток из каждого сообщества окружающих его “говорящих” узлов. Выбранные на данной итерации метки указывают на связь “слушающего” узла с сообществами A, C и D.

Важно отметить, что полученные эго-сообщества для пользователей социальных сетей являются ценными сами по себе. В частности, они могут быть использованы в любом приложении, которое опирается на данные о социальных кругах среди контактов целевого пользователя (например, в качестве замены ручной группировки контактов в Facebook и Google+ для оптимизации потоков информации на персональных страницах пользователей).

В частности, в работе [2] автором демонстрируется применимость информации об эго-сообществах пользователя для решения задачи рекомендации получателей электронных сообщений[14].

На этапе **распространения меток** осуществляется итеративное распространение меток сообществ для определения глобальной структуры сообществ входного графа. На каждой итерации каждый узел поочерёдно выполняет роль “говорящего” и “слушающего узла” согласно заданным стратегиям взаимодействия узлов:

[14]Веб-демонстрация разработанного прототипа доступна по адресу: `http://rs.at.ispras.ru/`

- **стратегия “говорящего узла”**: для каждого смежного ребра “говорящий узел” случайно выбирает ls элементов из своего текущего набора меток сообществ и удаляет дубликаты, после чего отправляет метки соседней вершине. Таким образом, узел с большим количеством сообществ посылает в среднем больше меток;

- **стратегия “слушающего узла”**: входящие метки распределяются между эго-сообществами в соответствии с индексами отправивших их “говорящих узлов”. Затем в каждом эго-сообществе выбираются lr наиболее частых меток и добавляются к множеству меток “слушающего узла”. Размер множества ограничен mx.

Таким образом, входящие метки сообществ для узла агрегируются не по всему множеству его соседей (как в методе SLPA), а по каждому эго-сообществу. В результате в память “слушающего” узла добавляются высокочастотные метки из каждого эго-сообщества (рисунок 3).

Предложенная стратегия “слушающего узла”, основанная на эго-сообществах, обеспечивает следующие преимущества по сравнению с другими методами на основе распространения меток:

- более точно оценивается число меток, которые необходимо принять и добавить в массив меток “слушающего узла” (частично решает проблемы недостатка сообществ и наличия шумовых меток);

- наиболее популярные метки рассчитываются независимо в разных эго-сообществах, что позволяет выявить сообщества с различной силой связи с узлом (частично решает проблему доминирующего сообщества).

На этапе **определения подсообществ** производится выявление возможных подсообществ в каждом сообществе, полученном на этапе распространения меток.

Рассматривается матрица Лапласа L_c для каждого сообщества:

$$L_c = A_c - diag(\vec{d_c}), \tag{17}$$

где A_c – матрица смежности, $\vec{d_c}$ – степени узлов.

Второе наименьшее собственное значение L (также известное как *алгебраическая связность* λ_{n-1}) позволяет оценить внутреннюю связность сообщества: нулевое значение означает наличие нескольких компонент связности, тогда как низкие значения ($< \lambda_x$) указывают на низкую связность. Предполагается, что для таких сообществ низкая связность обуславливается наличием подсообществ, которые не были разделены на предыдущих этапах. С целью уточнения результатов на данном этапе к каждому сообществу с низкой

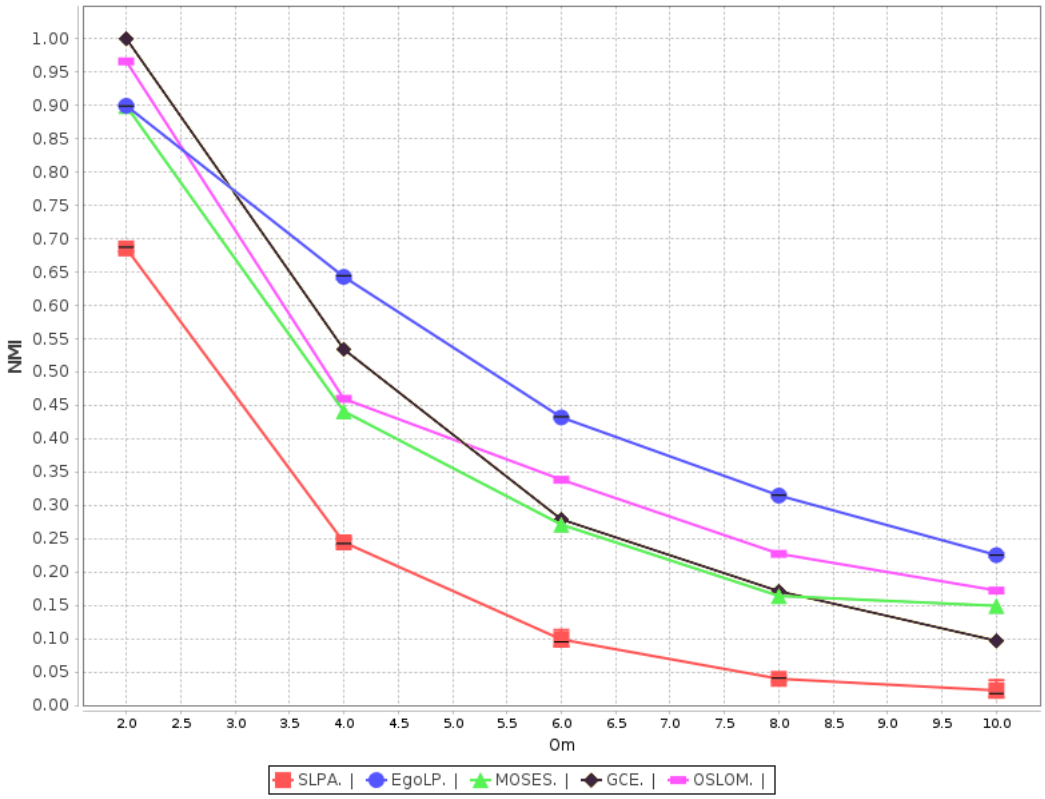

Рис. 4: Сравнение качества EgoLP и других методов с помощью LFR-графов с различными значениями O_m — числа сообществ, к которым принадлежит каждая из $O_n = 0.5N$ вершин.

связностью применяется модификация алгоритма SLPA с дополнительным параметром k, который регулирует количество меток, отправляемых и получаемых каждой вершиной на каждой итерации.

Общая вычислительная сложность `EgoLP` определяется этапом определения структры эго-сообществ и составляет $O(d_{mean} \cdot |E|)$, где d_{mean} соответствует средней степени вершины. Однако максимальная степень вершины в графе d_{max} ограничена параметром метода (вершины со степенью $> d_{max}$ не участвуют в наиболее вычислительно сложных этапах), что позволяет при $|E| \to \infty$ в случае обработки больших графов пренебречь и средней степенью вершины. Таким образом, общая сложность метода $O(|E|)$.

Для оценки качества разработанного метода использовались генераторы LFR и CKB, а также социальные графы, полученные из сетей Facebook и LiveJournal.

С помощью выбранного метода оценки качества было проведено сравнение EgoLP с другими методами определения структуры сообществ пользователей. Для сравнения были выбраны следующие алгоритмы: OSLOM и GCE из класса методов локальной оптимизации, SLPA из класса методов распространения меток, MOSES из класса методов, основанных на вероятностных моделях. Каждый из методов является одним из наиболее

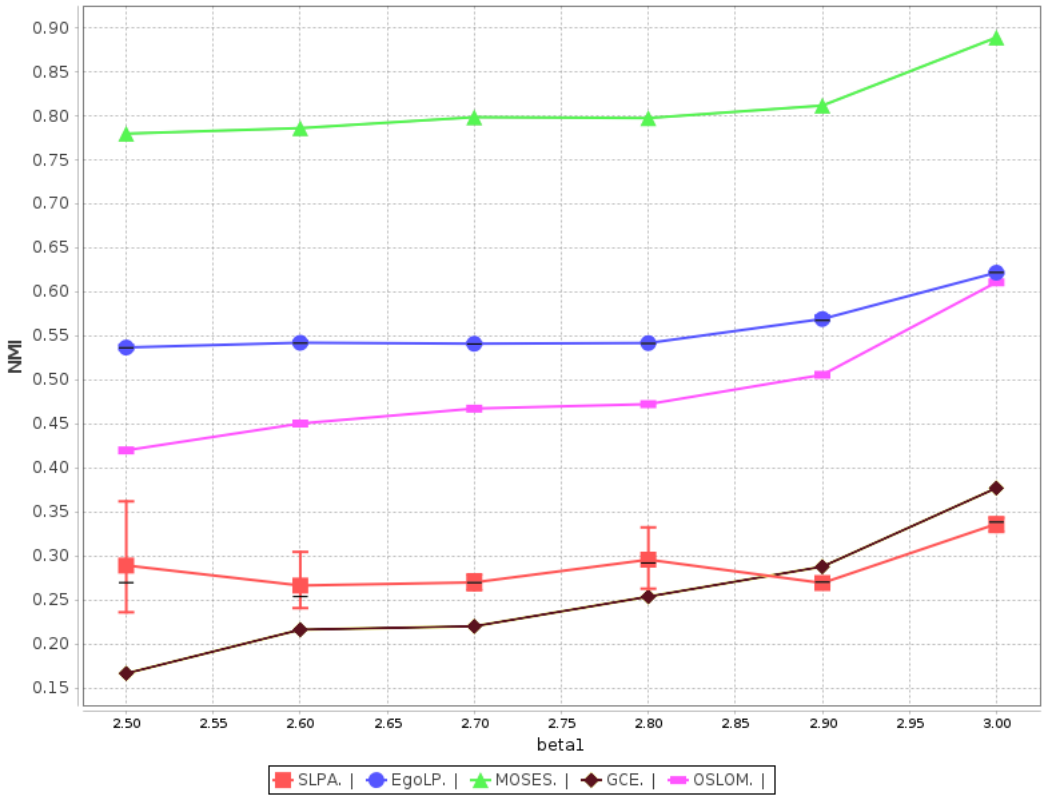

Рис. 5: Сравнение качества EgoLP и других методов с помощью CKB-графов с различными значениями β_1 — экспоненты степенного распределения количества сообществ у пользователя.

популярных в своём классе и обеспечивает относительно высокую точность определения структуры сообществ.

В случае CKB-графов была исследована зависимость NMI от параметра β_1 — экспоненты степенного распределения количества сообществ у пользователя. Уменьшение β_1 при неизменных значениях остальных параметров приводит к увеличению среднего числа сообществ у пользователя, что делает структуру сообществ более сложной для определения. Для LFR-графов увеличение пересечения сообществ достигается за счёт увеличения параметра O_m, который регулирует количество сообществ, к которым относится каждая из O_n вершин.

По результатам тестирования (рисунки 4 и 5) значение NMI для предложенного метода в большинстве случаев превосходит аналогичные показатели выбранных для сравнения методов. В остальных случаях лучшее качество показывают методы, обладающие большей вычислительной сложностью и неприменимые к графам большой размерности (сотни миллионов вершин).

Кроме того, методами EgoLP и SLPA был проанализирован граф LiveJournal с референтными сообществами. Значение NMI составило 0.29 для SLPA и 0.35 для EgoLP.

Для исследования применимости информации о найденных сообществах к прикладной задаче был использован метод косвенного оценивания качества структуры сообществ с помощью задачи определения скрытых атрибутов пользователей.

Для эксперимента использовался набор данных `Facebook100` — коллекция социальных сетей учащихся 100 американских колледжей и университетов. Все сети являются неориентированными и содержат указанные пользователями значения атрибутов: факультет, пол, профилирующая дисциплина, общежитие, год выпуска, школа. Согласно результатам исследований Traud et al[15], год выпуска и общежитие в наибольшей степени ассоциированы со структурой сообществ.

Метки сообществ для каждого пользователя, найденные с помощью различных методов определения структуры сообществ, были использованы в качестве векторов признаков для обучения классификатора по методу *градиентного бустинга*. Качество оценивалось по точности определения года выпуска и общежития для студентов в тестовых подвыборках. В этом приложении MOSES демонстрирует лучший результат (средняя точность 0.77), а EgoLP находится на второй позиции (средняя точность 0.69). Таким образом, основываясь только на конфигурации рёбер социального графа, найденные EgoLP покрытия позволяют со средней точностью 0.69 определять значения атрибутов, близость по которым оказывает влияние на формирование сообществ пользователей.

В приложении А содержатся результаты экспериментального исследования структурных свойств социальных графов с сообществами (глава 1). По результатам исследования можно заключить, что среди рассмотренных методов EgoLP и MOSES наиболее точно воспроизводят свойства референтных сообществ в продуцируемых покрытиях.

В приложении В содержатся результаты экспериментального исследования сообществ с помощью метрик качества (глава 1). По результатам исследования можно заключить:

- большинство методов находит сообщества с большей отделимостью по сравнению с референтными;

[15] Traud Amanda L, Mucha Peter J, Porter Mason A. Social structure of Facebook networks // Physica A: Statistical Mechanics and its Applications. 2012. Т. 391, № 16. С. 4165–4180.

- большинство методов находит сообщества с большими значениями плотности, сплочённости и коэффициента кластеризации по сравнению с LFR-сообществами и с меньшими значениями — по сравнению с CKB-сообществами;

- среди всех методов GCE и SLPA показывают худшие результаты по совокупности метрик, а MOSES и EgoLP показывают лучшие результаты.

Предложенный метод был реализован на языке программирования Scala с использованием Apache Spark. Основная часть реализации EgoLP основана на вычислительной парадигме Pregel, позволяющей оптимизировать время обработки графовых данных.

Для оценки производительности разработанного метода было проведено тестирование метода в параллельном режиме на кластере из потребительских компьютеров без разделяемых ресурсов. По результатам тестирования метод показал близкую к линейной масштабируемость. Обработка синтетического социального графа из 920 миллионов вершин со средней степенью 100 заняла около 300 часов на кластере из 18 машин.

Результаты четвёртой главы опубликованы в работе [6].

В **заключении** приведены основные результаты работы:

1) проведено исследование современных методов определения структуры сообществ пользователей в графах онлайновых социальных сетей, показавшее ограниченную применимость и недостаточную эффективность большинства рассмотренных методов, особенно в приложениях, требующих определения структуры значительно пересекающихся сообществ в социальных сетях с сотнями миллионов пользователей;

2) проведено исследование современных методов генерации случайных социальных графов с заданной структурой сообществ пользователей, выявившее существенные различия между структурными свойствами синтезируемых графов и реальными социальными сетями с сообществами, а также отсутствие масштабируемых методов, позволяющих синтезировать графы из сотен миллионов вершин;

3) разработан метод `CKB` для распределённой генерации случайных социальных графов с заданной структурой сообществ, учитывающий некоторые из недавно доказанных свойств модульной структуры социальных сетей. Реализация модели с помощью Apache Spark позволяет осуществлять распределённую генерацию случайных социальных графов из сотен миллионов вершин с различными наборами параметров. Синтезированные графы с заданной структурой сообществ могут

применяться для оценки применимости методов определения структуры пересекающихся сообществ к социальным графам различной природы;

4) разработан метод `EgoLP` для определения структуры значительно пересекающихся сообществ пользователей. Основой метода является итеративная пересылка меток сообществ по рёбрам графа в соответствии с установленными правилами взаимодействия вершин. Отличительной особенностью является идентификация эго-сообществ в сети непосредственных соседей каждого пользователя. В дальнейшем с помощью особых правил взаимодействия вершин поощряется объединение эго-сообществ в глобальные. Метод имеет распределённую реализацию на основе Apache Spark с использованием парадигмы распределённых вычислений Pregel. Экспериментально продемонстрировано, что предложенный метод превосходит известные методы по совокупности критериев: а) близость определённой структуры сообществ с заранее известной; б) точность решения прикладной задачи определения скрытых атрибутов пользователей с использованием информации о сообществах; в) вычислительная сложность; г) масштабируемость. Сложность алгоритма линейно зависит от количества рёбер графа, что позволяет применять его к социальным сетям с сотнями миллионов пользователей;

5) для экспериментального подтверждения эффективности предложенных методов реализованы прототипы систем для определения структуры сообществ пользователей и генерации случайных социальных графов с заданной структурой сообществ пользователей. Реализованные прототипы позволили подтвердить высокое качество предложенных методов и соответствие экспериментальных оценок производительности теоретическим оценкам вычислительной сложности

Разработанные методы позволяют исследовать структуру сообществ социальных сетей из сотен миллионов пользователей и применять полученные знания для решения исследовательских и бизнес-задач, а также для оптимизации решения других задач анализа больших социальных графов.

Публикации автора по теме диссертации

1. Анализ социальных сетей: методы и приложения / Антон Коршунов, Иван Белобородов, Назар Бузун [и др.] // ТРУДЫ ИНСТИТУТА СИСТЕМНОГО ПРОГРАММИРОВАНИЯ РАН. 2014. Т. 26, № 1.

2. Гомзин А.Г., Ипатов С.А., Коршунов А.В. Рекомендация получателей электронных сообщений с использованием различных типов локальных данных социальных сетей // Вестник НовГУ. 2014. № 81.

3. Коршунов Антон. Задачи и методы определения атрибутов пользователей социальных сетей // Selected Papers of the 15th All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections Yaroslavl, Russia, October 14-17, 2013. CEUR Workshop Proceedings 1108. 2013. С. 183–193.

4. Distributed Generation of Billion-node Social Graphs with Overlapping Community Structure / Kyrylo Chykhradze, Anton Korshunov, Nazar Buzun [и др.] // Complex Networks V. Springer, 2014. С. 199– 208.

5. Использование модели социальной сети с сообществами пользователей для распределенной генерации случайных социальных графов / К.К. Чихрадзе, А.В. Коршунов, Н.О. Бузун [и др.] // Машинное обучение и анализ данных. 2014. Т. 1s, № 8. С. 1027–1047.

6. EgoLP: Fast and Distributed Community Detection in Billion-node Social Networks / Nazar Buzun, Anton Korshunov, Valeriy Avanesov [и др.] // Proceedings of DaMNet-2014 - The Fourth IEEE ICDM Workshop on Data Mining in Networks. December 14, 2014, Shenzhen, China. С. 533 – 540.

7. Бузун Назар, Коршунов Антон. Выявление пересекающихся сообществ в социальных сетях // Доклады Всероссийской научной конференции «Анализ изображений, сетей и текстов»-АИСТ. 2012.

8. Buzun Nazar, Korshunov Anton. Innovative methods and measures in overlapping community detection // Proceedings of International Workshop on Experimental Economics in Machine Learning. 2012. С. 20–32.