СИСТЕМНЫЙ АНАЛИЗ
СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ПРОЦЕССОВ

УДК 519.177 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait/1995-5499/2023/3/75-83
Поступила в редакцию 28.03.2023
Подписана в печать 30.09.2023

ПРИМЕНЕНИЕ СПЕКТРАЛЬНЫХ МЕТОДОВ ДЛЯ РАСПОЗНАВАНИЯ СТРУКТУРЫ СООБЩЕСТВ В СЛОЖНЫХ СЕТЯХ

© 2023 Н. В. Гринева✉, П. А. Семёнова

Финансовый университет при Правительстве Российской Федерации
Ленинградский пр-т, 49/2, 125167 Москва, Российская Федерация

Аннотация. В работе исследованы методы спектральной кластеризации для обнаружения сообществ неориентированного графа. Эти алгоритмы получены из задач секционирования графов и стали одним из самых популярных способов определения структуры в последние годы. Реализованы несколько видов традиционных алгоритмов спектрального анализа на языке программирования Python для выделения сообществ в неориентированном графе, а также произведен сравнительный анализ методов, который будет являться уникальной информацией для корректного выбора способа обнаружения структуры сети. Практическая значимость работы заключается в возможности наилучшего выбора реализации алгоритма на основе спектральных методов для выделения сообществ, исходя из свойств конкретной сети и целей разбиения.

Ключевые слова: теория графов, структура сообществ, спектральный анализ, кластеризация, матрица Лапласа, модулярность.

ВВЕДЕНИЕ

Существует большое разнообразие сложных систем, которые могут быть смоделированы и проанализированы в виде графов. Многие сети реального мира имеют структуру модулей или сообществ. Например: если изобразить в виде графа контакты между людьми в компании, где вершины будут представлять сотрудников, а ребра — наличие взаимодействия между ними, то будут заметны области, в которых связей между вершинами будет больше, чем вне этой группы. Они представляют некоторые отделы компании и являются структурой сложной сети. Таким образом, в теории графов под сообществом принято понимать подграф с большим количеством внутренних связей и сравнительно меньшим количеством внешних связей. Подобные группы отражают топологические отношения между элементами базовой системы и функциональными сущностями.

Обнаружение структуры сообщества в сложных сетях имеет решающее значение, поскольку оно обеспечивает понимание подструктур всей сети и позволяет нам вывести особые и скрытые отношения между вершинами, для того чтобы иметь полное представление о взаимодействии различных структур и использовать полученную информацию в реализации проектов. Однако разработка эффективного алгоритма до сих пор остается ключевой задачей теории сложных сетей. Проблема состоит в том, что нет единого критерия измерения структуры сообщества, что является основным недостатком многих реализаций. Но несмотря на большое разнообразие методов, исследователи выделяют спектральную кластеризацию как один самых

✉ Гринева Наталья Владимировна
e-mail: ngrineva@fa.ru

мощных инструментов по выделению сообществ, так как результаты, полученные спектральными методами, очень часто превосходят другие, такие как алгоритм K-средних и, кроме того, спектральный подход очень прост в реализации и может быть решен стандартными методами линейной алгебры.

Целью работы является выработка рекомендаций по применению спектральных методов для выделения сообществ и обоснование наилучшего выбора для конкретной сети.

1. АНАЛИЗ ИСПОЛЬЗОВАННЫХ ТЕСТОВЫХ ВЫБОРОК ДЛЯ ПРОВЕРКИ МЕТОДОВ

Для проверки корректности спектральных методов, а также проведения их сравнительного анализа, целью которого является выделение преимуществ и недостатков исследуемых способов кластеризации, необходимы тестовые данные. Поэтому в качестве тестирующих датасетов работы алгоритмов использовались три сети.

Первая — встроенная в библиотеку networkx сеть членов клуба каратэ Закари, созданная в 1977 году. Это хорошо известный граф, часто используемый в качестве эталона для тестирования алгоритмов обнаружения сообщества. Это хорошо известный граф, часто используемый в качестве эталона для тестирования алгоритмов обнаружения сообщества. Он состоит из 34 человек, членов клуба карате в США, которые занимались в течение нескольких лет. Ребра соединяют людей, которые взаимодействовали вне деятельности клуба. В какой-то момент конфликт между президентом клуба и тренером привел к разделению клуба на две отдельные группы, поддерживающие инструктора и владельца соответственно [1].

Второй достаточно популярной сетью с известным количеством реальных сообществ является набор данных командного турнира по американскому футболу на сезон 2000 года. Вершины в графе представляют команды (идентифицированные по названиям их колледжей), а ребра представляют состоявшееся соревнование между ними.

Особенность этих двух примеров заключается в том, что мы заранее знаем о существующей структуре сообществ. Поэтому в качестве эксперимента была рассмотрена реальная сеть, представляющая сотрудничество ученых в одном из институтов. Структура сети не была изучена ранее. Узлы сети представляют сотрудников института, а ребра — наличие связи между учеными, которые опубликовали хотя бы одну совместную работу в течение некоторого периода. На рис. 1 изображена сеть сотрудничества, на которой отмечены предположительные сообщества.

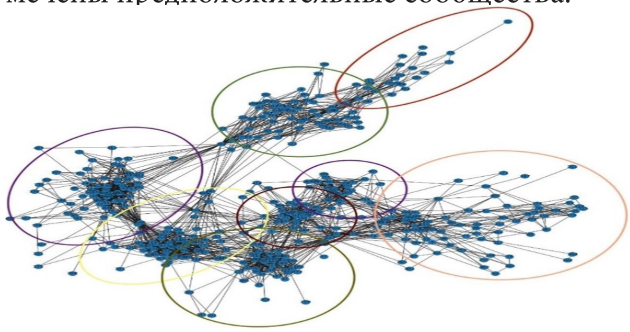

Рис. 1. Представление разбиения сети сотрудничества ученых в институте [Fig. 1.Representation of the breakdown of the network of collaboration between scientists at the institute]

2. ОСНОВНЫЕ ТЕРМИНЫ И ОБОЗНАЧЕНИЯ

Прежде чем дать подробное описание сравниваемых алгоритмов, введем некоторые термины, которые широко используются в следующих разделах.

Под G понимается неориентированная сеть, представленная в виде графа и состоящая из множества вершин $V = \{v_1, v_2, \ldots, v_n\}$ и множества ребер E. Для описания графа в матричной форме используют матрицу смежности A, где элемент a_{ij} обозначает количество ребер, соединяющих узлы v_i и v_j. D является диагональной матрицей, где d_i — степень вершины v_i, вычисляемой по формуле $d_{ii} = \sum_j a_{ij}$. Матрицей L называется матрица Лапласа, вычисляемая как $L = D - A$, а нормализованная матрица Лапласа L_{norm} определяется следующим образом:

$$L_{norm} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}.$$

Вычисленные собственные значения будем обозначать $\lambda_1 \ldots \lambda_n$, а собственные вектора $x_1 \ldots x_n$. Одной из основных характеристик стандартной матрицы Лапласа является вектор Фидлера, используемый для двустороннего разделения сети и соответствующий второму наименьшему собственному значению — значению Фидлера. [6] Оно приблизительно соответствует минимальному срезу графа, необходимому для разделения графа на два связанных подграфа. Каждое значение в векторе Фидлера дает нам информацию о том, какой стороне разреза принадлежит этот узел, относительно знака.

Для распределения узлов на сообщества на основе значений собственного вектора для некоторых алгоритмов будет использовать функция:

$$sign(x_i) = \begin{cases} 1, & x_i \geq 0 \\ -1, & x_i < 0 \end{cases}.$$

3.АНАЛИЗ МЕТОДОВ ВЫДЕЛЕНИЯ СООБЩЕСТВ НА БАЗЕ СПЕКТРАЛЬНОГО АНАЛИЗА

Многие системы могут быть представлены в виде сетей или графов — наборов вершин, соединенных в пары ребрами. Одной из общих характеристик многих сетей является их структура сообществ. Структура сообщества играет важную роль в функциональных свойствах сложной сети, а её обнаружение имеет важное практическое значение. В [2] авторы разделяют все методы по типам в зависимости от способа их реализации. Существуют модели оптимизации модульности, латентного пространства, распространения меток, алгоритмы неотрицательной матричной факторизации, аппроксимации блочной модели и спектральной кластеризации. В статье [3] ученые из Китая выделяют общую идею спектральной кластеризации в использовании базового метода обнаружения сообществ на соответствующих собственных векторах матрицы Лапласа. Существуют разные способы определения лапласиана, имеющие разные математические интерпретации, и поэтому кластеризация также имеет разные интерпретации.

3.1. Классический метод спектральной кластеризации

Первый рассмотренный алгоритм — классический метод спектральной кластеризации. Он основан на нахождении собственных значений матрицы Лапласа. После их вычисления находится вектор Фидлера, соответствующий первому минимальному собственному значению после 0, и строится векторное пространство для каждой вершины. Каждому узлу соответствует вектор, размерность которого равна числу кластеров. То есть, если предполагается, что сеть состоит из четырех сообществ, то каждой ее вершине будет сопоставляться вектор из компонент четырех собственных векторов, отвечающим первым четырём минимальным собственным значениям.

Алгоритм классического метода спектральной кластеризации	
Вход	G: исследуемая сеть k: количество предполагаемых кластеров
Выход	$\{v_i\}$: разбиение V на сообщества 1. Определить матрицу правдоподобия A; 2. Вычислить диагональную матрицу D; 3. Вычислить лапласиан L; 4. Вычислить k первых собственных векторов, соответствующих k наименьшим ненулевым собственным значениям L; 5. Построить матрицу, образованную первыми k собственными векторами; 6. Произвести кластеризацию узлов графа на основе этих признаков (например, с использованием кластеризации k-средних).

Особенностью данного метода кластеризации является то, что мы определяем количество кластеров заранее. Но зачастую невозможно визуально оценить точное число сообществ, особенно если сеть включает в себя более 100 узлов.

Полученные на последнем шаге алгоритма разбиения сети сотрудничество ученых на сообщества были визуализированы в виде графа на рис. 2, для сети научного сотрудничества было не только совершено неверное распределение внутри кластеров, но и не было определено предположительное число сообществ близкое к реальному.

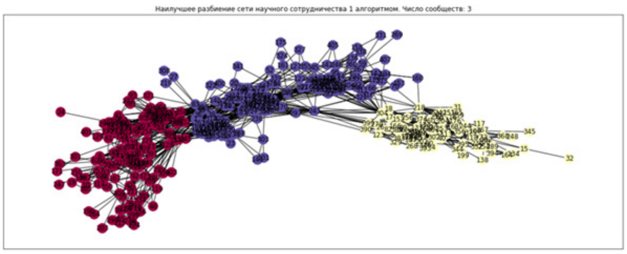

Рис. 2. Результаты разбиения алгоритма классического метода спектральной кластеризации
[Fig. 2. Results of partitioning the algorithm of the classical spectral clustering method]

3.2. Спектральный метод наименьшего разреза, основанный на матрице Лапласа

В спектральном методе наименьшего разреза решается задача поиска условного экстремума методом Лагранжа (при ограничениях $\sum_i x_i^2 = n$ и $\sum_i x_i = 0$):

$$Q(x) = \frac{1}{4} x^T L x - \lambda\,(x^T x - n),$$

при условии $x^T I = 0$,

где I — единичный вектор.

Для нахождения экстремума продифференцируем по x^T и приравняем к 0:

$$Lx = \lambda x, \text{ при условии } x^T \perp I.$$

Используя данное соотношение, получим:

$$Q(x) = \frac{1}{4} x^T L x - \lambda\left(x^T x - n\right) =$$

$$-\frac{1}{4}\lambda x x^T - \lambda\left(x^T x - n\right) = \frac{4n - 3xx^T}{4}\lambda.$$

Пусть решением является собственный вектор x_i, соответствующий собственному значению λ_i и удовлетворяющий условию $x^T x = n$. Тогда задача сводится к нахождению минимального собственного значения λ_i:

$$Q(x_i) = \frac{4n - 3x_i x_i^T}{4}\lambda_i =$$

$$= \frac{4n - 3n}{4}\lambda_i = \frac{n}{4}\lambda_i \to \min.$$

Так как матрица Лапласа для неориентированной является симметричной и положительно определенной, то наименьшее собственное значение будет равно 0, но не будет удовлетворять ограничению $\sum_i x_i = 0$, так как $x_1 = I$. Тогда решением будет второе снизу по величине собственное значение λ_2 и соответствующий ему вектор Фидлера. Таким образом, алгоритм спектрального метода наименьшего разреза выглядит следующим образом:

Алгоритм спектрального метода наименьшего разреза

Вход	G: исследуемая сеть
Выход	$\{v_i\}$: разбиение V на сообщества 1. Определить матрицу правдоподобия A; 2. Вычислить диагональную матрицу D; 3. Вычислить лапласиан L; 4. Определить собственный вектор, соответствующий значению Фидлера; 5. На основе отобранного вектора построить индикаторный вектор f.

В отличие от предыдущего этот метод использует только один собственный вектор и способен обнаружить только два сообщества. В качестве функции, преобразующей собственный вектор x в индикаторный, использовалась функция $sign(x)$. Для того чтобы обнаружить полную структуру сообществ в сети, была применена рекурсия, критерием вызова которой являлось убывание метрики.

На рис. 3 изображено разбиение сети сотрудничество ученых на сообщества.

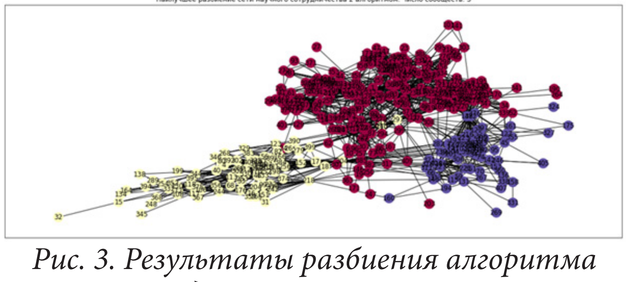

Рис. 3. Результаты разбиения алгоритма метода наименьшего разреза
[Fig. 3. Partition results of the least cut algorithm]

3.3. Спектральный метод наименьшего разреза на основе нормализованного лапласиана

Ученые Сюэ-Ци Чэн и Хуа-Вэй Шэнь, исследуя динамику диффузии в сетях, в [4] доказали, что структура сообщества может быть идентифицирована с помощью собственных значений и собственных векторов нормализованной матрицы Лапласа. Главным преимуществом нормированного лапласиана является то, что он учитывает плотность по двум направлениям: как плотность связей между группами, так и плотность связей внутри группы, в том время как обычный рассматривает только частоту связей между кластерами.

Алгоритм наименьшего разреза на основе нормализованного лапласиана совпадает с предыдущим рассмотренным алгоритмом. Единственное отличие заключает в том, что после нахождения матрицы L происходит ее нормировка. Это позволяет ослабить влияние узлов, которые обладают большей степенью, на конечный результат разбиения. Поэтому лучший эффект нормализованной матрицы Лапласа достигается в сетях, где существует большая неоднородность узлов.

На рис. 4 заметно, что алгоритм становится более чувствительным к связям внутри каждого кластера и старается сгруппировать сообщество таким образом, чтобы плотность внутри каждого была примерно одинаковая. Поэтому по сравнению с предыдущим было обнаружено меньшее число сообществ для больших сетей.

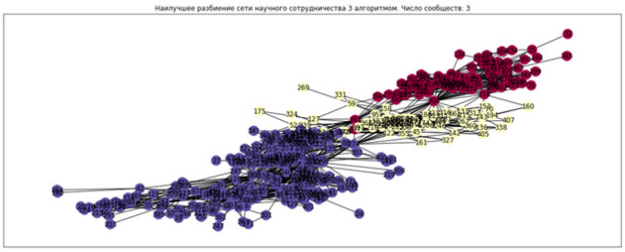

Рис. 4. Результаты разбиения алгоритма метода наименьшего разреза на основе нормализованного лапласиана
[Fig. 4. Results of partitioning the least cut algorithm based on the normalized Laplacian]

3.4. Спектральный метод максимизации модулярности

Модулярность — это мера, определяющая силу разделения сети на кластеры. Сети с высокой модульностью имеют сильные связи между узлами внутри сообщества, но слабые между узлами в разных группах. Основным недостатком метрики является то, что она не распознает маленькие сообщества.

В [5] группа ученных из Сингапура предложила использовать матрицу модулярности в спектральном методе для обнаружения структуры сообществ. Её элементы определяются как $B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$, где $k_i = \sum_j A_{ij}$ — степень узла i, а $2m = \sum_{ij} A_{ij} = \sum_i k_i$ — сумма степеней всех узлов. В данном методе в отличие ранее разобранных методов решается обратная задача на поиск условного экстремума при ограничении $\sum_i x_i^2 = n$:

$$Q(x) = \frac{1}{4m} x^T Bx - \lambda(x^T x - n) \to \max.$$

Чтобы найти решение, продифференцируем по x^T и приравняем к 0:

$$Bx = \lambda x.$$

Используя данное соотношение, получим:

$$Q\left(x\right) = \frac{1}{4m} x^T Bx - \lambda \left(x^T x - n\right) =$$

$$= \frac{1}{4m} \lambda x x^T - \lambda \left(x^T x - n\right) = \frac{4n - 3x^T x}{4m} \lambda.$$

Тогда задача сводится к нахождению максимального собственного значения λ_i, удовлетворяющего условию, что $x_i^T x_i = n$:

$$Q\left(x_i\right) = \frac{4n - 3x_i^T x_i}{4m} \lambda_i =$$

$$= \frac{4n - 3n}{4m} \lambda_i = \frac{n}{4m} \lambda_i \to \max.$$

Алгоритм спектрального метода максимизации модульности	
Вход	G: исследуемая сеть
Выход	$\{v_i\}$: разбиение V на сообщества
	1. Определить матрицу правдоподобия A;

2. Вычислить матрицу Лапласа L;
3. Построить матрицу модулярности B;
4. Определить собственный вектор, соответствующий максимальному собственному значению;
5. На основе отобранного вектора построить индикаторный лапласиан f.

Как и метод минимального разреза алгоритм нацелен на разбиение сети на два сообщества, поэтому для нахождения полной структуры сети использовалась рекурсия, остановка которой вызывалась уменьшением величины модулярности. На рис. 5 изображены результаты выделения сообществ в тестовых данных рассмотренным алгоритмом. Например, для выборки сообщества ученых по рисунку заметим, что вершины {32, 134, 394, 153, 76} отнесены к сообществу, с которым не имеют ни одного общего ребра.

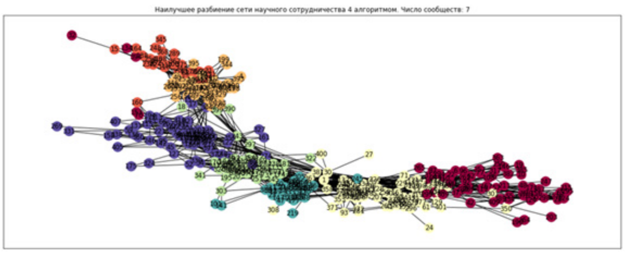

Рис. 5. Результаты разбиения алгоритма максимизации модулярности
[Fig. 5. Results of partitioning the modularity maximization algorithm]

4. СРАВНИТЕЛЬНЫЙ АНАЛИЗ РАССМОТРЕННЫХ МЕТОДОВ

При реализации методов производилось построение графов с выделенными в ходе работы программы сообществами. Граф представляет визуализацию сети, каждая из которых обладала уникальным свойством. Так, например, для сети Закари заранее была известна структура сообществ, образованных при распаде клуба. В наборе данных, представляющем группировку команд в университетском турнире, присутствуют очень плотные связи между кластерами, что усложняет задачу обнаружения всех существующих подмножеств. Сеть связей между учеными в некотором институте состояла из большого количества узлов, превосходящее в несколько раз предыдущие датасеты. Для того чтобы определить, какой алгоритм справляется лучше с поставленной задачей, использовались метрики, которые с разных сторон оценивают оптимальность разбиения определенной сети.

Индекс Калински — Харабаша [6] называют критерием отношения дисперсий, который может использоваться для определения оптимального значения k в кластеризации. Значение индекса является мерой того, насколько узел удален от центроида предположительного кластера и общего центроида сети. Для его нахождения используется отношение суммы межкластерной дисперсии и внутрикластерной дисперсии для всех выделенных групп, где дисперсия определяется как сумма квадратов расстояний. Оценка алгоритма с помощью данной метрики производилась на третьей сети, так как для нее характерно большее количество сообществ с разряженными связями между друг другом. Самое низкое значение было у классического метода, а самое высокое — у метода на основе матрицы Лапласа, что свидетельствует о том, что наилучшее разбиение было достигнуто вторым алгоритмом несмотря на то, что число выделенных сообществ одинаково. Стоит также заметить, что значение для нормализованного лапласиана также близки к максимуму.

Второй использованной метрикой был индекс Дэвиса — Булдина [6], который основан на соотношении внутрикластерных и межкластерных расстояний. Если два кластера расположены близко друг к другу, то есть отношение расстояния между сообществами маленькое, но имеют большое среднее расстояние между узлами в подмножествах и их центроидами, то это индекс Дэвиса — Булдина будет большим, указывая на то, что эти кластеры не очень различимы. То есть, более низкие значения метрики указывают на компактные, хорошо разделенные кластеры. Нахождение оптимального алгоритма согласно данной метрики производилось на основе набора данных о разбиение команд. Наихудший результат определения структуры сообществ показал алгоритм наименьшего разреза основанный на нормализованной матрицы Лапласа, лучший результат был достигнут при использовании классического алгоритма.

Для того чтобы воспользоваться информацией об эталонном разбиении клуба карате была выбрана статистика нормализованной взаимной информации (NMI) в качестве меры сходства между двумя возможными разделами сети. NMI основан на определении матрицы путаницы, показывающей ошибки при выделении кластеров. Таким образом, метрика принимает максимальное значение 1, если эталонное разделение идентично полученному, и NMI = 0, если они статистически независимы. Главным минус этого способа оценки алгоритма является то, что он не может объективно оценить оптимально разбиение сети на сообщества, и так как любой алгоритм обнаружения сообщества зависит от определенной функции выгоды (например, в последнем методе заложена функция основанная на модулярности), то зачастую мы имеем не высокое значение индекса нормализованной взаимной информации. Наилучший результат показали методы наименьшего разреза на основе матриц Лапласа.

Модулярность — одна из важных характеристик сети при выделении кластеров, так как она позволяется оценить, насколько сильны связи между вершинами сообщества и ослаблены вне его. Поэтому реализованные алгоритмы были протестированы на показатель данной метрики. Наилучший результат, как и ожидалось, показал метод максимизации модульности. Самая минимальная модулярность была зафиксирована при втором алгоритме, но стоит отметить, что для всех четырех разбиений наблюдались достаточно высокие значения.

По времени работы программы наиболее трудозатратным оказался последний метод, а самым быстрым — классический алгоритм. Результаты оцененных метрик представлены в табл. 1.

Таким образом, при выборе метода в первую очередь нужно опираться на характеристики датасета и целей разбиения. Если необходимо найти подмножества сети, в которой наблюдаются очень тесные взаимосвязи между предполагаемыми сообществами, то эффективнее использовать классический метод спектральной кластеризации, основанный на K-средних. Если стоит задача выделить все возможные кластеры таким образом, что плотность внутри выделенных групп выше, чем вне, то для этих целей более подходит алгоритм максимизации модулярности. При использовании сетей с большим количеством узлом для выделения структуры сообществ лучше применять алгоритм, основанный на нормализованном лапласиане, так как он позволяет снизить влияние вершин с большим количеством связей и оптимизировать поиск кластеров с помощью минимального разреза, учитывающего плотность связей и обеспечивающего сбалансированное разбиение.

ЗАКЛЮЧЕНИЕ

В работе был проведен анализ спектральных методов обнаружения структуры сообществ и оценок кластеризации. На его основе были отобраны методы спектрального анализа и показатели качества разбиения сети на кластеры, также были выделены особенности реализации алгоритмов и применимость метрик в зависимости от целей интерпретации полученных результатов.

Таблица 1. Сравнительная характеристика реализованных алгоритмов
[Table 1. Comparative characteristics of the implemented algorithms]

Сравнительные характеристики	Метод 1	Метод 2	Метод 3	Метод 4
Индекс Калински — Харабаша	15.9594	18.3554	18.1479	16.2419
Индекс Дэвиса — Булдина	1.6165	2.4536	3.7249	2.0859
NMI	0.2065	0.7324	0.7324	0.5791
Модулярность	0.5546	0.5044	0.5050	0.6514
Время работы программы	460 мс	562 мс	461 мс	41 с

Для проведения сравнительного анализа на основе метрик были отобраны тестовые выборки для оценки методов. Каждая сеть благодаря уникальным характеристикам позволила проанализировать с разных сторон результаты работы алгоритмов.

В процессе работе были реализованы четыре метода выделения сообществ на основе спектра матриц, и описаны результаты работы каждого из алгоритмов. После был проведен сравнительный анализ с использованием оценок качества кластеризации.

По итогам работы выработаны рекомендации по применению спектральных методов для обнаружения структуры сложных сетей в зависимости от их свойств и целей разбиения на сообщества. Таким образом, цель исследования была достигнута.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Fortunato S.* Community detection in graphs / S. Fortunato // Physics Reports. – 2010. – V. 486, Iss. 3-5. – P. 103.

2. *Zhaoa X., Lianga J., Wangab J.* A community detection algorithm based on graph compression for large-scale social networks / Xingwang Zhaoa, Jiye Lianga, Jie Wangab // Information Sciences. – 04.2021 – № 551. – P. 358–372.

3. *Cheng J.* A divisive spectral method for network community detection / Jianjun Cheng, Longjie Li, Mingwei Leng, Weiguo Lu, Yukai Yao, and Xiaoyun Chen // Statistical Mechanics Theory and Experiment. – 2015. – № 4. – P. 24.

4. Spectral methods for the detection of network community structure: A comparative analysis / Hua-Wei Shen and Xue-Qi Cheng // Statistical Mechanics Theory and Experiment. – 2010. – № 5. –P. 21–24.

5. A spectral algorithm of community identification / Xiaofeng Gong, Kun Li, Menghui Li, C.-H. Lai // A Letters Journal Exploring the Frontiers of Physics. – 2013. – № 5 – P. 73–78.

6. Сложные системы и ИИ. Разделение данных. – Режим доступа: https://complex-systems-ai.com/ru. – (Дата обращения: 15.02.2023).

Гринева Наталья Владимировна — кандидат экономических наук, доцент, доцент Департамента анализа данных и машинного обучения Финансового университета при Правительстве Российской Федерации.
E-mail: ngrineva@fa.ru
ORCID iD: https://orcid.org/0000-0001-7647-5967

Семёнова Полина Алексеевна – студентка 4 курса факультета информационных технологий и анализа больших данных Финансового университета при Правительстве Российской Федерации.
E-mail: polli_eee@mail.ru
ORCID iD: https://orcid.org/0009-0000-4835-5319

DOI: https://doi.org/10.17308/sait/1995-5499/2023/3/75-83 ISSN 1995-5499
Received 28.03.2023
Accepted 30.09.2023

APPLICATION OF SPECTRAL METHODS FOR RECOGNIZING THE STRUCTURE OF COMMUNITIES IN COMPLEX NETWORKS

© 2023 N. V. Grineva✉, P. A. Semenova

Financial University under the Government of the Russian Federation,
49/2, Leningradsky Avenue, 125167 Moscow, Russian Federation

Annotation. In this paper, spectral clustering methods for detecting communities of an undirected graph are investigated. These algorithms are derived from graph partitioning problems and have become one of the most popular ways to determine structure in recent years. Several types of traditional spectral analysis algorithms have been implemented in Python programming language to identify communities in an undirected graph, and a comparative analysis of methods has been carried out, which will be unique information for the correct choice of a network structure detection method. The practical significance of the work lies in the possibility of the best choice of the implementation of the algorithm based on spectral methods for identifying communities, based on the properties of a particular network and the goals of partitioning.
Keywords: graph theory, community structure, spectral analysis, clustering, Laplace matrix, modularity.

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

1. *Fortunato S.* (2010) Community detection in graphs. *Physics Reports*. V. 486, Iss. 3-5. P. 103.
2. *Zhaoa X., Lianga J., Wangab J.* (2021) A community detection algorithm based on graph compression for large-scale social network. *Information Sciences*. No 551. P. 358–372.
3. *Jianjun Cheng, Longjie Li, Mingwei Leng, Weiguo Lu, Yukai Yao and Xiaoyun Chen* (2015) A divisive spectral method for network community detection. *Statistical Mechanics Theory and Experiment*. No 4. P. 24.
4. *Hua-Wei Shen and Xue-Qi Cheng* (2010) Spectral methods for the detection of network community structure: A comparative analysis. *Statistical Mechanics Theory and Experiment*. No 5. P. 21–24.
5. *Xiaofeng Gong, Kun Li, Menghui Li, C.-H. Lai* (2013) A spectral algorithm of community identification. *A Letters Journal Exploring the Frontiers of Physics*. No 5. P. 73–78.
6. Complex systems and AI. Data separation. – Available at: http://complex-systems-ai.com/ru (accessed: 02/15/2023).

Grineva Natalia V. — Candidate of Economic Sciences, Associate Professor, Associate Professor of the Department of data analysis and machine learning Financial University under the Government of the Russian Federation.
E-mail: ngrineva@fa.ru
ORCID iD: https://orcid.org/0000-0001-7647-5967

Semenova Polina A. — 4th-year student of the Faculty of Information Technology and Big Data Analysis of the Financial University under the Government of the Russian Federation.
E-mail: polli_eee@mail.ru
ORCID iD: https://orcid.org/0009-0000-4835-5319

✉ Grineva Natalia V.
e-mail: ngrineva@fa.ru