

Разделение сообщества двусторонней сети на основе увеличения модульности

И. Н. Полякова, Л. Чэнь*

ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова», г. Москва, Рос-
сийская Федерация

Адрес: 119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1

* chenlong_2022@rambler.ru

Аннотация

Двудольные сети представляют собой особый тип графа, состоящий из двух разных типов узлов. Ребра между узлами существуют только между узлами разных типов. Разделение сообществ является важным звеном в сетевом анализе, которое предполагает разделение узлов сети на несколько кластеров или сообществ, чтобы лучше понять и выявить структурные и функциональные характеристики сети. Для двусторонних сетей разделение сообщества может помочь понять, насколько тесные взаимосвязи формируются между различными типами узлов. Структура двусторонней сети используется во многих областях. Из-за ее уникальных структурных характеристик традиционные методы разделения сообществ часто не могут быть напрямую применены к двусторонним сетям, что делает предлагаемое исследование разделения сообществ двусторонних сетей более значимым. Основная задача этого исследования – разработать новый алгоритм разделения сообщества, основанный на приращении модульности. Этот алгоритм призван преодолеть ограничения существующих методов BRIM и повысить точность и эффективность разделения сообщества в двудольных сетях. Модульность является важным индикатором качества разделения сети, а степень модульности принимает значение от -1 до 1. Когда значение положительное, структура сообщества более плотная, чем случайная сеть, а когда значение отрицательное, структура сообщества менее плотная, чем случайная сеть. В ходе работы решались следующие проблемы: во-первых, как определить и рассчитать приращение модульности в двудольных сетях; во-вторых, как разработать эффективный алгоритм для максимизации приращения модульности для достижения наилучшего эффекта разделения сообщества; наконец, как проверить эффективность и превосходство нового алгоритма в разделении сообществ двудольных сетей. Результаты исследований показывают, что новый алгоритм хорошо работает при разделении двусторонних сетевых сообществ, повышает точность и эффективность разделения сообществ и может более эффективно выявлять структурные характеристики сети, тем самым предоставив мощный инструмент анализа в смежных областях.

Ключевые слова: двудольная сеть, разделение сообществ, сообщества двусторонней сети, увеличение модульности

Финансирование: данная работа выполнена в рамках НИР «Математическое и программное обеспечение перспективных систем обработки символической информации с элементами искусственного интеллекта», проводимой на кафедре алгоритмических языков факультета вычислительной математики и кибернетики Московского государственного университета имени М.В. Ломоносова.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Для цитирования: Полякова И. Н., Чэнь Л. Разделение сообщества двусторонней сети на основе увеличения модульности // Современные информационные технологии и ИТ-образование. 2024. Т. 20, № 2. С. 445-454. <https://doi.org/10.25559/SITITO.020.202402.445-454>

© Полякова И. Н., Чэнь Л., 2024



Контент доступен под лицензией Creative Commons Attribution 4.0 License.
The content is available under Creative Commons Attribution 4.0 License.



Research into Structure of Two-Way Network Community

I. N. Polyakova, L. Chen*

Lomonosov Moscow State University, Moscow, Russian Federation

Address: 1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation

* chenlong_2022@rambler.ru

Abstract

Two-way networks are a special type of graph consisting of two different types of nodes. The edges between nodes exist only between nodes of different types. Community partitioning is an important link in network analysis, which involves dividing network nodes into several clusters or communities in order to better understand and identify the structural and functional characteristics of the network. For two-way networks, community partitioning can help to understand how close relationships are formed between different types of nodes. The structure of two-way network is used in many fields. Due to its unique structural characteristics, traditional community partitioning methods often cannot be directly applied to two-way networks, which makes the proposed study of community partitioning of two-way networks more meaningful. The main objective of this research is to develop a new community partitioning algorithm based on modularity increment. This algorithm aims to overcome the limitations of existing BRIM methods and improve the accuracy and efficiency of community partitioning in two-way networks. Modularity is an important indicator of the quality of network partitioning, and the degree of modularity takes a value from -1 to 1. When the value is positive, the community structure is denser than the random network, and when the value is negative, the community structure is less dense than the random network. The problems addressed in this work were: first, how to determine and calculate the modularity increment in two-way networks; second, how to develop an efficient algorithm to maximise the modularity increment to achieve the best community partitioning effect; finally, how to test the effectiveness and superiority of the new algorithm in community partitioning of two-way networks. The research results show that the new algorithm performs well in separating two-way network communities, improves the accuracy and efficiency of community separation, and can identify the structural characteristics of the network more effectively, thus providing a powerful analysis tool in related fields.

Keywords: bipartite network, community division, increased modularity, two-way network community

Funding: This work was carried out within the framework of the research project «Mathematical methods and software for advanced symbolic information processing systems with artificial intelligence features», conducted at the Department of Algorithmic Languages of the Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University.

Conflict of interests: The authors declares no conflict of interest.

For citation: Polyakova I.N., Chen L. Research into Structure of Two-Way Network Community. *Modern Information Technologies and IT-Education*. 2024;20(2):445-454. <https://doi.org/10.25559/SITITO.020.202402.445-454>



Введение

Сложная сеть – это недавно появившаяся междисциплинарная область исследований, охватывающая математику, информатику, физику и биологию [1]. Изучение структуры сообщества можно разделить на три типа: открытие сообщества, анализ структуры сообщества, а также структура сообщества и сетевая динамика [2, 3]. Большая часть исследований по обнаружению сообществ сосредоточена на одночастных сетях, а исследования по разделению сообществ в двудольных сетях относительно недостаточны [4]. Актуальны алгоритмы для разделения сообщества. Алгоритм разделения сообщества двучастной сети используется для обработки двух типов узлов как одного типа, а затем односторонние узлы извлекаются после разделение сообщества.

Большинство реальных сетей имеют структуру сообщества, а это означает, что большая сеть может быть разделена на несколько подсообществ [5]. Связи внутри этих подсообществ тесны, а связи между этими подсообществами редки. Образующиеся из этих сложных сетей подсообщества и оценка качества сообществ очень важны для анализа реальных сетей.

В последние годы был достигнут большой прогресс в исследовании алгоритмов декомпозиции сообществ для сложных сетей [6, 7]. Чтобы измерить преимущества и недостатки различных алгоритмических декомпозиций, Гирван и Ньюман [5] определили индекс оценки удовлетворенности разделением сети, называемый модульностью. Эта модульность в настоящее время широко используется для измерения качества декомпозиции сообщества [8, 9]. Она применима не только к ненаправленным сетям, но также может быть распространена на направленные сети, взвешенные сети и многоуровневые сети. Это также может помочь определить иерархические структуры в сети, а многоуровневую структуру сообщества сети можно обнаружить путем рекурсивной максимизации модульности. Конечно, у модульности также есть определенные недостатки, такие как невозможность измерения сетей с большими различиями в сообществах и невозможность определить, является ли одно сообщество разумным разделением. Двусторонние сети, как важная ветвь сложных сетей, имеют обширный опыт применения и теоретическую ценность. Двудольная сеть – важная форма выражения в сложных сетях (по сравнению с общими одночастными сетями), состоящая из двух частей с разными типами вершин, причем две вершины одного типа не связаны между собой.

Многие сети в реальном мире демонстрируют естественную дихотомическую структуру [10, 11]. Например, совместные исследования учёных в реальной двусторонней сети – учёные в сети сотрудничества с бумажными документами, совместные исследования актёров в сети сотрудничества актёров кино, болезнь в сети генов болезней, сообщество слушателей сети слушателей-песен и группа терминалов сети P2P данных компьютерных терминалов, а также дихотомическая сетевая структура между читателями и книгами.

При изучении проблемы сегментации сообществ в двусторонних сетях распространенным подходом является их преобразование в одночастные сети, однако это может привести к значительной потере информации и повлиять на точность определения структуры сообществ [12]. Другой подход заклю-

чается в прямом разбиении на двучастные сети. Такие методы могут эффективно сохранить большую часть информации исходной сети. Например, Барбер предложил алгоритм BRIM [13, 14], подходящий для разделения сообщества двудольной сети на основе одномодовых сетей. Алгоритм хорошо работает в небольших двудольных сетях, но имеет ограниченную эффективность в крупномасштабных двудольных сетях. Более того, этим методам часто не хватает признанного стандарта оценки, что затрудняет оценку эффекта алгоритма. Кроме того, относительно недостаточно исследований разделения сообществ в двусторонних сетях, особенно с точки зрения взаимосвязи между структурой сообщества и функциями сети.

В связи с этим целью данного исследования является более глубокое изучение проблемы сегментации сообществ в двусторонних сетях. Рассмотрены основные концепции и свойства двусторонних сетей, проанализированы существующие методы разделения сообществ в двусторонних сетях, включая их преимущества и ограничения, предложен новый алгоритм разделения сообщества, целью которого является преодоление недостатков существующих методов и повышение точности и эффективности разделения сообщества. Кроме того, это исследование сосредоточено на изучении проблемы разделения сообщества на основе увеличения модульности в двусторонних сетях с целью эффективного выявления и анализа структуры сообщества в двусторонних сетях. Описываемый метод не только повышает точность и эффективность разграничения сообществ, но также лучше раскрывает внутренние характеристики модулей сети, сохраняя при этом целостность сетевой структуры.

Современное состояние и существующие проблемы разделения сообществ в двусторонних сетях

С точки зрения графов сообщество можно определить как подмножество узлов, которые тесно связаны друг с другом и слабо связаны с узлами других сообществ в одном графе [15]. В сетевом анализе обнаружение сообществ в сети является одной из наиболее важных задач [16-18]. В крупномасштабной сети, такой как социальная сеть, у нас могут быть миллионы узлов и ребер [19]. Обнаружение сообществ в таких сетях – непростая задача. Поэтому нам нужны алгоритмы обнаружения сообществ, которые смогут разделить сеть на несколько сообществ (рис. 1).

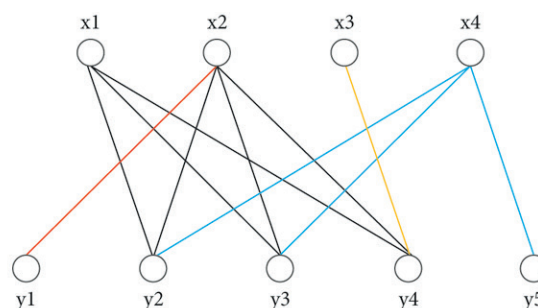


Рис. 1. Двусторонняя сеть

Fig. 1. Two-way network

Источник: здесь и далее в статье все рисунки составлены авторами.

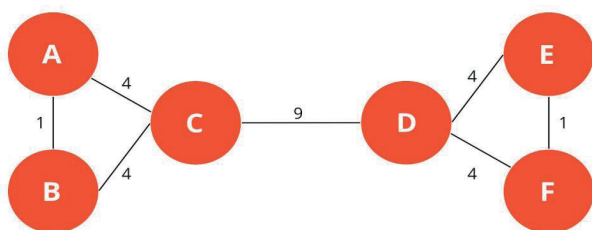
Source: Hereinafter in this article all figures were drawn up by the authors.



Алгоритм Гирвана-Ньюмана

Рассмотрим алгоритм обнаружения сообщества, основанный на расстоянии между ребрами. Его основная идея состоит в том, чтобы вычислить расстояние между ребрами всех ребер в сети, а затем найти ребро с наибольшим значением расстояния между ребрами и удалить его. При удалении изменяется расстояние между некоторыми ребрами. Кратчайший путь, который раньше проходил через удаленное ребро, теперь проходит через другие ребра, поэтому расстояние между ребрами необходимо пересчитывать, затем снова ищется и удаляется ребро с наибольшим значением и так далее. Идентификация сообществ через границы между ними – относительно интуитивный алгоритм обнаружения сообществ.

Алгоритм Гирвана-Ньюмана – это алгоритм обнаружения сообществ, основанный на промежуточности. Его основная идея заключается в непрерывном удалении ребер из сети в порядке от большего к меньшему в соответствии с центральностью промежуточного ребра до тех пор, пока вся сеть не будет разложена на различные сообщества (рис. 2).



Р и с. 2. Метод Гирвана-Ньюмана
F i g. 2. Girvan-Newman method

Можно отметить преимущества алгоритма Гирвана-Ньюмана – когда известно число разделенных сообществ, результаты разделения сообществ могут быть даны более точно, также он может давать результаты разделения сообщества на разных уровнях и раскрывать информацию об иерархической структуре сети. Но можно выделить и недостатки – он страдает от высокой вычислительной сложности и плохих результатов при неизвестном количестве сообществ [2], [4, 5], [20].

Алгоритм BRIM

Алгоритм BRIM (двусторонняя рекурсивная индуктивная модель) – это метод разделения сообщества двусторонней сети, основанный на теории модульности двудольной сети, расширенной из матрицы модульности однодольной сети [21, 22]. Этот алгоритм направлен на повышение точности и эффективности разделения сообщества посредством рекурсивного деления. Он адаптивный и может автоматически выбирать подходящие стратегии и параметры разделения в соответствии с характеристиками сети. Цель алгоритма BRIM – разделить двустороннюю сеть на сообщества с тесными внутренними связями и разреженными связями между сообществами, чтобы лучше отразить структуру и характеристики сети. Хотя этот алгоритм эффективно разделяет двудольную сеть, он требует дополнительных входных параметров [13], [23] на этапе инициализации, что делает его применение весьма ограниченным.

Алгоритм LPA

Алгоритм LPA использует матричную декомпозицию на основе метода распространения меток [18]. Этот алгоритм сначала обеспечивает, что каждой вершине присваивается уникальная метка, а затем каждая вершина выбирает как можно больше соседних меток и, наконец, формирует отдельное сообщество путем итерации вершин с одной и той же меткой. Стоимость делает алгоритм LPA подходящим для больших двудольных сетей и параллельного анализа сообщества в реальном времени больших двудольных сетей. Однако этот алгоритм имеет проблему низкой точности деления. Алгоритм распространения меток применяется к обнаружению сообщества двусторонней сети, но его двусторонняя структура приведет к таким проблемам с метками узлов, как колебание и плохая стабильность.

Позже в исследовании Мураты улучшили алгоритм распространения меток и предложили алгоритм, более подходящий для двусторонних сетей [23]. Этот улучшенный алгоритм облегчает выполнение параллельного анализа сообщества в реальном времени в больших двусторонних сетях. Также они предложили алгоритм LP&RRIM [24], который представляет собой комплексное улучшение алгоритма BRIM и алгоритма LPA. Он основан на алгоритме распространения меток и использует алгоритм BRIM для его улучшения, что приводит к лучшему разделению сообщества, что может достичь хорошие результаты в качестве разделения в крупных двусторонних сетях.

Алгоритм Би-кликковый

В работе Леманна [25] распространили алгоритм К-кликки [19] на двудольные сети и предложили алгоритм разделения сообщества двудольной сети Bi-clique. Алгоритм K-clique изначально был разработан для обнаружения сообществ в сети одного типа. Он основан на поиске К узлов, тесно связанных в сети, и формировании тесно связанного подграфа, который считается сообществом. Однако в двудольной сети ситуация усложняется, поскольку существует два разных типа узлов и связей между ними. Это делает традиционный алгоритм К-кликки неприменимым напрямую к двудольным сетям, поскольку он не учитывает характеристики двудольной структуры. Соответствующим образом была улучшена функция Bi-clique: она может обнаруживать перекрывающиеся сообщества и учитывать сообщества с различной плотностью связей. Это делает его более подходящим для решения проблем разделения сообщества в двусторонних сетях. Несмотря на это, у него все еще есть определенные недостатки: алгоритм анализирует только двудольные сети, плотность которых превышает заданный порог [26].

Алгоритм оптимизации муравьиной колонии

Данный алгоритм обеспечивает сегментацию сообществ с помощью муравьиной модели с высокой точностью [27]. Однако этот алгоритм имеет высокие вычислительные затраты при работе с большими и сложными сетями, процесс настройки параметров сложен и требует много времени, и легко попасть в локальные минимумы. Оптимальное решение влияет на глобальную оптимальность открытия сообщества. Кроме того, на стабильность и последовательность алгоритма влияют эвристические правила и случайность, и при разных запусках могут быть получены разные результаты разделения сообщества.



Проведенное исследование для решения задачи по разделению сообществ двудольных сетей

Цель исследования

Двудольные сети представляют собой особый тип графа, состоящий из двух разных типов узлов. Ребра между узлами существуют только между узлами разных типов. Разделение сообществ является важным звеном в сетевом анализе, которое предполагает разделение узлов сети на несколько кластеров или сообществ, чтобы лучше понять и выявить структурные и функциональные характеристики сети. Для двусторонних сетей разделение сообщества может помочь нам понять, насколько тесные взаимосвязи формируются между различными типами узлов.

Структура двусторонней сети используется во многих областях. Из-за ее уникальных структурных характеристик традиционные методы разделения сообществ часто не могут быть напрямую применены к двусторонним сетям, что делает предлагаемое исследование разделения сообществ двусторонних сетей более значимым.

Основная задача этого исследования – разработать новый алгоритм разделения сообщества, основанный на приращении модульности, для эффективного и точного решения проблемы разделения сообщества в двудольных сетях. Модульность является важным индикатором качества разделения сети, а степень модульности принимает значение от -1 до 1. Когда значение положительное, структура сообщества более плотная, чем случайная сеть, а когда значение отрицательное, структура сообщества менее плотная, чем случайная сеть.

В ходе работы решались следующие проблемы: во-первых, как определить и рассчитать приращение модульности в двудольных сетях; во-вторых, как разработать эффективный алгоритм для максимизации приращения модульности для достижения наилучшего эффекта разделения сообщества; наконец, как проверить эффективность и превосходство нового алгоритма в разделении сообществ двудольных сетей. Решение этих проблем окажет важное влияние на исследования разделения сообществ двусторонних сетей. Оно не только повысит точность и эффективность разделения сообществ, но и поможет более глубоко понять структурные характеристики двусторонних сетей, тем самым предоставив мощный инструмент анализа в смежных областях.

Классический алгоритм Newman

Определение метрики модульности ассоциации Модульность Q используется для описания силы характеристик сетевой ассоциации [2]. Для сети, содержащей n узлов и m ребер, мы определяем модульность, которая рассчитывается следующим образом [28, 29]:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(g_i - g_j) \quad (1)$$

где :

Q : Модульность, мера силы ассоциативной структуры сети. Значение Q не является значимым, если оно находится в пределах $0,3 \sim 0,7$ считается, что имеется четкая ассоциативная структура.

A_{ij} : Элементы матрицы смежности, 1, если есть ребро между узлом i и узлом j , 0 в противном случае.

P_{ij} : Матрица ожидаемого соседства, т.е. ожидаемое значение вероятности существования ребра между узлом i и узлом j в случайной сети.

где:

$P_{ij} = \frac{k_i k_j}{2m}$, k_i и k_j степени узлов i и j , соответственно.

$\delta(g_i - g_j)$: Функция Кронекера, $\delta_{ij} = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}$ используется

для проверки того, находятся ли два узла в одной ассоциации. m : общее количество ребер в сети.

Цель уравнения (1) – сравнить ассоциативную структуру реальной сети со структурой случайной сети. Если $Q > 0$ то реальная сеть имеет более сильную структуру ассоциаций, чем случайная сеть.

Далее запишем в другой матричной форме вычисления степени модульности сети:

$$Q = \frac{1}{2m} \text{Tr}(S^T B S) \quad (2)$$

где:

$B_{ij} = A_{ij} - P_{ij}$: Матрица модульности, которая представляет собой разницу между реальной сетью и случайной сетью.

S : Матрица принадлежности S используется для описания того к какой ассоциации принадлежит каждый узел сети. Для сети, содержащей s ассоциаций, матрица принадлежности может быть определена следующим образом $S \in R^{n \times c}$, n – количество узлов в сети, где если узел i принадлежит к ассоциации c_j , то $S_{ij} = 1$ иначе $S_{ij} = 0$. Таким образом, функция Кронекера $\delta(g_i - g_j)$ может быть выражена в терминах элементов матрицы S . То есть: $\delta(g_i - g_j) = S_{ic} S_{jc}$, подставив это выражение в выражение для Q , получим: $Q = \frac{1}{2m} \sum_{ij} B_{ij} S_{ic} S_{jc}$

Tr : След матрицы, то есть сумма элементов на главной диагонали. Это можно записать в матричной форме:

$$Q = \frac{1}{2m} \text{Tr}(S^T B S)$$

Например: Предположим, что у нас есть небольшая сеть книговыдачи, состоящая из 3 читателей (узлы 1, 2, 3) и 3 книг (узлы A, B, C). Предположим, что записи о выдаче книг выглядят следующим образом:

Читатель 1 взял книги A(1) и B(2).

Читатель 2 взял книгу B(2).

Читатель 3 взял книгу C(3).

Шаг 1: Решите матрицу смежности A

Шаг 2: Решите задачу для ожидаемой матрицы смежности P

Шаг 3: Матрица принадлежности S

Шаг 4: Решите задачу для матрицы B

$$B_{ij} = A_{ij} - P_{ij}$$

Шаг 5: Рассчитайте степень модульности Q

Вычисления на шаге 4 по классическому алгоритму невозможны, так как $Q = \frac{1}{2m} \text{Tr}(S^T B S)$, а здесь B – матрица $n_1 \times n_2$, где n_1 – количество узлов типа I, то есть количество читателей, а n_2 – количество узлов типа II, то есть количество типов книг; суммируя количество точек $n = n_1 + n_2$.

Здесь $S^T B S$ не сможет выполнить вычисление, потому что матрицу $s \times n$ нельзя умножить на матрицу $n_1 \times n_2$ ($n \neq n_1$). Поэтому далее был исследован и опробован алгоритм BRIM.



Алгоритм BRIM

Если число узлов типа 1 равно p , а число узлов типа 2 равно q , то общее число узлов равно $n = p + q$. Матрица связности и матрица ожидаемых вероятностей соответствующей стохастической модели сети имеют следующий вид $\tilde{B}_{ij} = \tilde{A}_{ij} - \tilde{P}_{ij}$.

Где:

$$\tilde{A} = \begin{bmatrix} O_{p \times p} & A_{p \times q} \\ (A^T)_{q \times p} & O_{q \times q} \end{bmatrix} \quad \tilde{P} = \begin{bmatrix} O_{p \times p} & P_{p \times q} \\ (P^T)_{q \times p} & O_{q \times q} \end{bmatrix} \quad (3)$$

Матрица $\tilde{A}$:

В левом верхнем и правом нижнем углах находятся нулевые матрицы, что говорит об отсутствии прямых связей между читателями и книгами.

Другие разделы указывают на фактическую связь между читателем и книгой.

Матрица $n \times n$

Матрица $\tilde{P}$:

$\tilde{P}_{ij}$ обозначает ожидаемую вероятность связи между читателем i и книгой j .

$P_{ij} = \frac{k_i k_j}{2m}$, где k_i – степень читателя i , k_j – степень книги j , а m – общее количество ребер в сети (т.е. общее количество заимствований).

Матрица $n \times n$

Матрица $\tilde{B}_{ij}$:

Матрица $n \times n$

Мы продолжаем использовать пример с алгоритмом Ньюмана:

Шаг 1: Решите матрицу смежности $\tilde{A}$

Шаг 2: Решение для ожидаемой матрицы смежности $\tilde{P}$

Шаг 3: Решите матрицу $\tilde{B}$

$$\tilde{B}_{ij} = \tilde{A}_{ij} - \tilde{P}_{ij}$$

Шаг 4: Рассчитайте модульность дихотомической сети Q

Модульность дихотомической сети рассчитывается следующим образом:

$$Q = \frac{1}{m} \sum_{i=1}^p \sum_{j=1}^q \tilde{B}_{ij} \delta(g_i, h_j) \quad (4)$$

В сочетании с уравнением (2) модульность сети бисекции может быть выражена как:

$$Q = \frac{1}{2m} \text{Tr}(S^T \tilde{B} S) \quad (5)$$

Индикация разделения ассоциаций:

Матрицы R и T :

Для обозначения структуры ассоциаций в сети мы используем матрицы R и T .

R и T соответствуют двум разным типам узлов в сети (например, в сети книгообеспечения один из них – читатель, а другой – книга).

Смысл столбцов матрицы:

Каждый столбец g_i или t_i представляет ассоциацию.

Для каждого столбца g_i в R его элемент указывает, принадлежит ли узел первого типа в сети к ассоциации i или нет.

Аналогично, для каждого столбца t_i в T его элемент указывает, принадлежит ли узел второго типа в сети к ассоциации i или нет.

Исходя из разделения ассоциаций, обозначим матрицы R и T как $R = [r_1 | r_2 | \dots | r_c]$ и $T = [t_1 | t_2 | \dots | t_c]$, подставляя в уравнение (5), получаем:

$$Q = \frac{\sigma_1}{m} (\langle r_1, v_1 \rangle \langle t_1, u_1 \rangle + \langle r_2, v_1 \rangle \langle t_2, u_1 \rangle + \dots + \langle r_n, v_1 \rangle \langle t_n, v_1 \rangle) \quad (6)$$

Матрица R представляет собой ассоциации читателей, кото-

рые мы можем получить: $R = [r_1 | r_2]$, где $r_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ означает, что

читатель 1 принадлежит к обществу 1, $r_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$ указывает, что

считыватели 2 и 3 принадлежат к ассоциации 2.

Матрица T представляет собой ассоциации книг, которые мы

можем получить: $T = [t_1 | t_2]$, где $t_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ указывает, что книги

A и B принадлежат обществу 1, а $t_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ означает, что книга C

принадлежит обществу 2.

Алгоритм BRIM способен выполнять разбиение дихотомических сетей, однако в качестве входного параметра необходимо использовать количество ассоциаций в операции, что имеет некоторые ограничения в применении. Поэтому на этой основе предлагается новый алгоритм связанного класса для обнаружения ассоциаций, основанный на приращении модульности.

Новый алгоритм (максимизация приращений модульности)

Основная идея построения заключается в следующем: сначала предполагается, что все узлы в сети являются независимыми объединениями; затем существующие узлы объединяются путем непрерывной максимизации приращения модульности вдоль перспективы максимизации приращения модульности и в конечном итоге сливаются в одно объединение. И в результате расхождения объединений выбирается разделение с наибольшим значением модульности.

Основные шаги алгоритма следующие:

Шаг 1: Рассчитайте матрицу модульности $\tilde{B}$

Объедините уравнения (3), чтобы вычислить матрицу модульности дихотомической сети $\tilde{B}$.

Шаг 2: Инициализируем матрицу принадлежности S

Инициализируйте матрицу принадлежности S в единичную матрицу E размером $n \times n$ (n – общее количество узлов).

Цель инициализации – рассматривать каждый узел (читатель и книга) как отдельную ассоциацию в начале работы алгоритма.

Например, при общем количестве узлов 6 (3 читателя и 3 книги) и $n = 6$ инициализированная матрица принадлежности S будет представлять собой единичную матрицу 6×6 следующим образом:

$$S = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Шаг 3: Решение для сингулярных значений и сингулярных векторов

Вычислите максимальное сингулярное значение σ_1 матрицы $\tilde{B}$ соответствующие ему левый и правый сингулярные векторы u_1 и v_1 .

Шаг 4: Слияние ассоциаций и расчет прироста модульности

Предположим, что две ассоциации c_i и c_j объединяются.

Рассчитайте приращение модульности ΔQ , используя уравнение (6).



$$\Delta Q = \frac{\sigma_1}{m} (\langle r_{ij}, v_1 \rangle \langle t_{ij}, u_1 \rangle - \langle r_i, v_1 \rangle \langle t_i, u_1 \rangle - \langle r_j, v_1 \rangle \langle t_j, v_1 \rangle) \quad (7)$$

Где $r_{ij} = r_i + r_j$ и $t_{ij} = t_i + t_j$ представляют новые сообщества после слияния сообществ c_i и c_j соответственно.

Выберите лучшее слияние, найдите сообщество с наибольшим увеличением модульности по формуле (7), объедините сообщества c_i и c_j и вычислите статистику модульности.

В объединенной матрице S , если $S_i = S_i + S_j$, то удалите S_j . Каждой агломерации необходимо попытаться объединить все сообщества, соединенные ребрами, вычислить соответствующий ΔQ , а затем найти наибольшее значение для слияния.

Например:

Предположим, что на одной итерации алгоритм рассматривает объединение ассоциации читателя R1 (C1) и ассоциации книги B1 (C4).

Вычислите приращение степени модуля ΔQ :

Алгоритм вычисляет приращение степени модуля для слияния C1 и C4.

Выбор сценария слияния:

Предполагая, что слияние C1 и C4 приводит к наибольшему приращению степени модуля, алгоритм выбирает эту схему слияния.

Далее обновляется матрица принадлежности S

После слияния C1 и C4 матрица принадлежности S должна быть обновлена, чтобы отразить это изменение.

Обновленный вариант S показан ниже:

$$S = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Поскольку читатель 1 и книга 1 теперь объединены в один клуб, то есть C1 и C4 теперь являются одним клубом C, книга 1 в позиции (1,4) должна принадлежать к тому же клубу, что и читатель 1, и 0 становится 1. Аналогично, в позиции (4,1), поскольку читатель 1 и книга 1 являются одним клубом, 0 становится 1.

Шаг 5: Процесс коалесценции

Процесс объединения ассоциаций повторяется непрерывно (шаг 4), каждый раз выбирая схему объединения, которая максимизирует приращение модульности ΔQ .

При каждом слиянии обновляется матрица принадлежности S . Продолжайте объединять ассоциации до тех пор, пока вся сеть не превратится в одну большую ассоциацию.

Шаг 6: Результаты разделения ассоциаций

В качестве результата ассоциативного деления для этой сложной сети берется деление, соответствующее максимальной модульности.

Проиллюстрируем новый алгоритм на примере небольшой сети выдачи книг.

Предположим, что существует небольшая сеть книговыдачи, содержащая 4 читателя (R1, R2, R3, R4) и 7 книг (B1, B2, B3, B4). Начальное состояние:

Изначально каждый читатель и каждая книга были отдельным клубом, поэтому у нас было 8 отдельных клубов.

Первый раунд слияний:

Предположим, алгоритм вычислил, что объединение ассоциаций R1 и B1 дает наибольший прирост модульности, поэтому они объединяются в одну ассоциацию R1-B1.

Аналогично, алгоритм обнаруживает, что объединение R2 и B2 также дает большой прирост модульности, поэтому они также объединяются в R2-B2.

В следующем раунде алгоритм может обнаружить, что объединение ассоциаций R1-B1 с ассоциациями R3 приводит к наибольшему увеличению модульности, поэтому они объединяются в более крупную ассоциацию R1-B1-B3.

Также алгоритм может решить объединить R2-B2 и B4 в более крупную ассоциацию R2-B2-B4.

Последующие слияния:

Процесс слияния продолжается, каждый раз выбирая схему слияния, которая максимизирует инкрементную модульность. Окончательная структура ассоциации:

В конечном итоге этот процесс может привести к формированию двух больших ассоциаций: одной, состоящей из R1, B1, B3, R3, и другой, состоящей из R2, B2, B4, R4.

В этом примере две ассоциации, которые в итоге образуются, отражают две тесно связанные группы читателей и книг в сети книгообеспечения.

Программная реализация

Реализация кода

В исследовании были использованы ключевые пакеты Python для обработки данных и проведения сетевого анализа. Pandas и numpy предоставляют структуры данных и математические операции для эффективной обработки и анализа данных. Matplotlib.pyplot отвечает за визуализацию данных и помогает отображать результаты исследований. В качестве важного инструмента для изучения структуры и характеристик сети networkx используется в анализе сложных сетей. Кроме того, также используется scipy.sparse и его подмодули для эффективной обработки разреженных данных (Pandas, numpy, matplotlib.pyplot, networkx, scipy.sparse). Для вывода изображения результатов заимствования книг сообществом вызывает-ся библиотека pyplot на Python.

Практическая часть

Был проведен эксперимент, используя ситуацию с получением книг 40 учениками определенного класса. Формат данных – файл csv следующего вида, состоящий из читателей и книг (см. рис. 3)

```
Reader,Book
Reader_1,Book_20
Reader_1,Book_19
Reader_1,Book_9
Reader_2,Book_3
Reader_3,Book_8
Reader_3,Book_16
Reader_4,Book_14
Reader_4,Book_5
Reader_4,Book_20
Reader_4,Book_1
Reader_4,Book_6
Reader_5,Book_14
Reader_5,Book_9
Reader_5,Book_18
Reader_5,Book_12
Reader_5,Book_6
Reader_6,Book_9
Reader_6,Book_20
Reader_6,Book_2
Reader_7,Book_19
Reader_8,Book_19
Reader_9,Book_1
Reader_9,Book_12
Reader_9,Book_8
Reader_9,Book_10
Reader_10,Book_20
Reader_11,Book_17
Reader_11,Book_9
Reader_11,Book_1
Reader_12,Book_11
Reader_27,Book_8
Reader_27,Book_6
Reader_27,Book_1
Reader_27,Book_5
Reader_28,Book_19
Reader_28,Book_1
Reader_28,Book_15
Reader_29,Book_14
Reader_30,Book_12
Reader_30,Book_10
Reader_30,Book_20
Reader_31,Book_16
Reader_31,Book_19
Reader_31,Book_20
Reader_31,Book_6
Reader_31,Book_10
Reader_32,Book_3
Reader_32,Book_13
Reader_33,Book_20
Reader_33,Book_18
Reader_33,Book_4
Reader_34,Book_20
Reader_34,Book_13
Reader_34,Book_19
Reader_35,Book_11
Reader_35,Book_6
Reader_36,Book_7
Reader_36,Book_14
```

Р и с. 3. Фрагмент файла csv с экспериментальными данными

Fig. 3. A fragment of the csv file with experimental data



Структура сообщества: [0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
16 17 18 19 20 21 22 23 24 25 26 0 28 29 30 31 32 33 34 35 36 37
38 39 40 40 42 43 44 45 46 47 48 49 40 51 52 53 54 55 40 57 58 40]

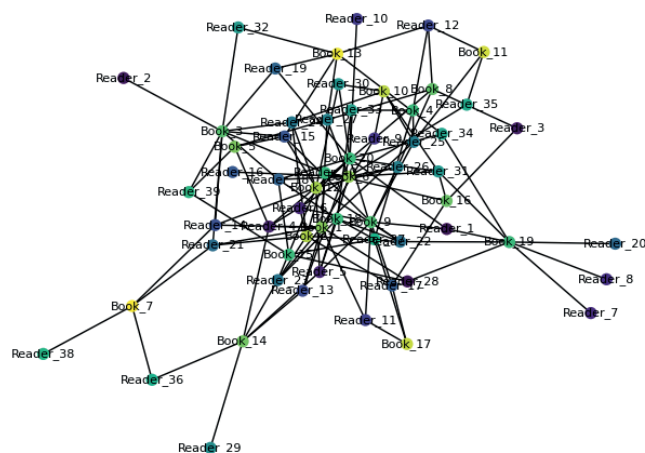


Fig. 4. Community Split Results – Proposed New Algorithm

Окончательная модульность: 0.5875968992248062

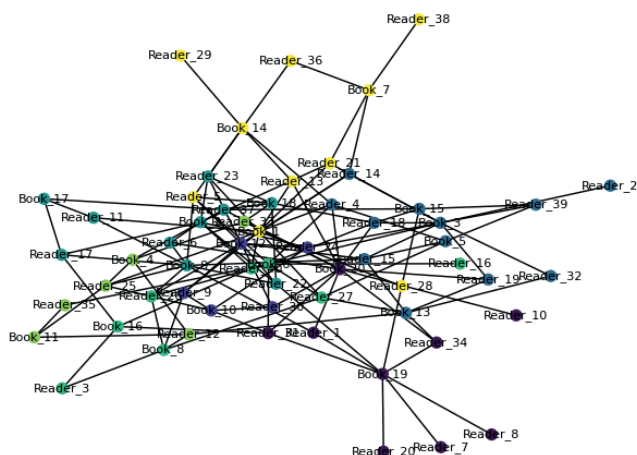


Fig. 5. Community Split Results – BRIM Algorithm

Окончательная модульность вывода равна 0,6535702418777546. Модульность – это показатель от -1 до 1. Вообще говоря, если он находится в диапазоне 0,3 ~ 0,7, считается, что он имеет очевидную структуру сообщества. Здесь мы

В целом полученные результаты показывают, что предлагаемый алгоритм эффективно идентифицирует и разделяет структуру сообществ в сети и достигает цели обнаружения сообществ.

Основная идея алгоритма – использовать модульность как количественный показатель плотности соединений внутри сообщества и разницы в плотности соединений между сообществами. Первоначально каждый узел рассматривается как

независимое сообщество, и алгоритм итеративно оценивает все возможные варианты слияния сообществ и выбирает те варианты слияния, которые максимизируют приращение модульности. Этот процесс не зависит от заданного количества сообществ, что позволяет алгоритму адаптивно идентифицировать сообщества на основе структурных характеристик самой сети.

Хотя это исследование дало хорошие результаты в обнаружении сообществ в двусторонних сетях, все еще существуют некоторые ограничения и возможности для улучшения. Например, вычислительная эффективность алгоритма при обработке крупномасштабных сетей может быть недостаточно высокой, а применимость и эффективность алгоритма в различных типах сетевых структур также нуждаются в дальнейшей проверке. Будущие исследования этого алгоритма могут быть сосредоточены на повышении эффективности вычислений и расширении числа более полезных типов сетей. Таким

образом, предложенный алгоритм может предоставить более мощный инструмент для понимания и анализа сложных сетей. В настоящее время мы в основном фокусируемся на базовом анализе статистических характеристик и структурных условий сети. По сравнению с персонализированными рекомендациями, которые реализованы в системах социальных сетей и искусственном интеллекте при обработке естественного языка, все еще существует большой разрыв.

Таким образом, в дальнейшем можно сосредоточиться на таких ключевых областях, как углубленный анализ поведения пользователей и оптимизация распределения и использования ресурсов, чтобы более полно изучить и применить важные характеристики структуры сообщества, сложную структуру сообщества, сеть. Это предоставит руководство и поддержку для разделения сообществ двусторонних сетей и будет способствовать дальнейшему развитию исследований и практики в этой области.

References

- [1] Calderer G., Kuijjer M.L. Community Detection in Large-Scale Bipartite Biological Networks. *Frontiers in Genetics*. 2021;12:649440. <https://doi.org/10.3389/fgene.2021.649440>
- [2] Newman M.E.J., Cantwell G.T., Young J.-G. Improved mutual information measure for clustering, classification, and community detection. *Physical Review E*. 2020;101:042304. <https://doi.org/10.1103/PhysRevE.101.042304>
- [3] Riolo M.A., Newman M.E.J. Consistency of community structure in complex networks. *Physical Review E*. 2020;101:052306. <https://doi.org/10.1103/PhysRevE.101.052306>
- [4] Newman M.E.J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*. 2006;103(23):8577-8582. <https://doi.org/10.1073/pnas.0601602103>
- [5] Girvan M., Newman M.E.J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*. 2002;99(12):7821-7826. <https://doi.org/10.1073/pnas.122653799>
- [6] Cai Z.M. Network community partition based on intelligent clustering algorithm. *Computer Optics*. 2020;44(6):985-989. <https://doi.org/10.18287/2412-6179-CO-724>
- [7] Zhang H., Wu Y. Optimization and Application of Clustering Algorithm in Community Discovery. *Wireless Personal Communications*. 2018;102(4):2443-2454. <https://doi.org/10.1007/s11277-018-5264-x>
- [8] Peixoto T.P. Revealing Consensus and Dissensus between Network Partitions. *Physical Review X*. 2021;11:021003. <https://doi.org/10.1103/PhysRevX.11.021003>
- [9] Filho D.V., O'Neale D.R.J. Degree distributions of bipartite networks and their projections. *Physical Review E*. 2018;98(2):022307. <https://doi.org/10.1103/PhysRevE.98.022307>
- [10] Arthur R. Modularity and projection of bipartite networks. *Physica A: Statistical Mechanics and its Applications*. 2020;549:124341. <https://doi.org/10.1016/j.physa.2020.124341>
- [11] Valejo A., Góes F., Romanetto L. A benchmarking tool for the generation of bipartite network models with overlapping communities. *Knowledge and Information Systems*. 2020;62:1641-1669. <https://doi.org/10.1007/s10115-019-01411-9>
- [12] Bolorunduro J.O., Zou Z. Community Detection On Multi-layer Graph using Intra-layer and Inter-layer Linkage Graphs (CDMILG). *Expert Systems with Applications*. 2024;238(A):121713. <https://doi.org/10.1016/j.eswa.2023.121713>
- [13] Peng Y., Zhang B., Chang F. Overlapping Community Detection of Bipartite Networks Based on a Novel Community Density. *Future Internet*. 2021;13(4):89. <https://doi.org/10.3390/fi13040089>
- [14] Radicchi F., Castellano C., Cecconi F., Loreto V., Parisi D. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*. 2004;101(9):2658-2663. <https://doi.org/10.1073/pnas.0400054101>
- [15] Liu X., Murata T. Community Detection in Large-Scale Bipartite Networks. In: *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*. Vol. 01 (WI-IAT '09). IEEE Computer Society, USA; 2009. p. 50-57. <https://doi.org/10.1109/WI-IAT.2009.15>
- [16] Xu Y., Chen L. Community Detection on Bipartite Networks Based on Ant Colony Optimization. *Journal of Frontiers of Computer Science and Technology*. 2014;8(3):296-304.
- [17] Calatayud J., Bernardo-Madrid R., Neuman M., Rojas A., Rosvall M. Exploring the solution landscape enables more reliable network community detection. *Physical Review E*. 2019;100:052308. <https://doi.org/10.1103/PhysRevE.100.052308>
- [18] Gregori E., Lenzini L., Mainardi S. Parallel K-Clique Community Detection on Large-Scale Networks. *IEEE Transactions on Parallel and Distributed Systems*. 2013;24(8):1651-1660. <https://doi.org/10.1109/TPDS.2012.229>



- [19] Shao J., Han Z., Yang Q., Zhou T. Community Detection based on Distance Dynamics. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15). New York, NY, USA: Association for Computing Machinery; 2015. p. 1075-1084. <https://doi.org/10.1145/2783258.2783301>
- [20] Gupta S., Kumar P. A constrained agglomerative clustering approach for unipartite and bipartite networks with application to credit networks. *Information Sciences*. 2021;557:332-354. <https://doi.org/10.1016/j.ins.2019.12.085>
- [21] Che S., Yang W., Wang W. A Memetic Algorithm for Community Detection in Bipartite Networks. *IEEE Access*. 2019;7:126897-126912. <https://doi.org/10.1109/ACCESS.2019.2938982>
- [22] Barber M.J. Modularity and community detection in bipartite networks. *Physical Review E*. 2007;76(6):066102. <https://doi.org/10.1103/PhysRevE.76.066102>
- [23] Taguchi H., Murata T., Liu X. BiMLPA: Community Detection in Bipartite Networks by Multi-Label Propagation. In: Masuda N., Goh K.I., Jia T., Yamanoi J., Sayama H. (eds.) Proceedings of NetSci-X 2020: Sixth International Winter School and Conference on Network Science. NetSci-X 2020. Springer Proceedings in Complexity. Cham: Springer; 2020. p. 17-31. https://doi.org/10.1007/978-3-030-38965-9_2
- [24] Staudt C.L., Meyerhenke H. Engineering Parallel Algorithms for Community Detection in Massive Networks. *IEEE Transactions on Parallel and Distributed Systems*. 2016;27(1):171-184. <https://doi.org/10.1109/TPDS.2015.2390633>
- [25] Lehmann S., Schwartz M., Hansen L.K. Biclique communities. *Physical Review E*. 2008;78(1):016108. <https://doi.org/10.1103/PhysRevE.78.016108>
- [26] Setio A.A.A., et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Medical Image Analysis*. 2017;42:1-13. <https://doi.org/10.1016/j.media.2017.06.015>
- [27] Zhao J., You X., Duan Q., Liu S. Multiple Ant Colony Algorithm Combining Community Relationship Network. *Arabian Journal for Science and Engineering*. 2022;47:10531-10546. <https://doi.org/10.1007/s13369-022-06579-x>
- [28] Yen T.C., Larremore D.B. Community detection in bipartite networks with stochastic block models. *Physical Review E*. 2020;102(3):032309. <https://doi.org/10.1103/PhysRevE.102.032309>
- [29] Wu G., Gu C., Yang H. A spectral method of modularity for community detection in bipartite networks. *Europhysics Letters*. 2022;137(3):31001. <https://dx.doi.org/10.1209/0295-5075/ac5506>

Поступила 19.02.2024; одобрена после рецензирования 28.03.2024; принята к публикации 25.05.2024.

Submitted 19.02.2024; approved after reviewing 28.03.2024; accepted for publication 25.05.2024.

Об авторах:

Полякова Ирина Николаевна, доцент кафедры алгоритмических языков факультета вычислительной математики и кибернетики, ФГБОУ ВО «Московский государственный университет имени М.В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1), кандидат физико-математических наук, доцент, ORCID: <https://orcid.org/0000-0003-1432-4906>, polyakova@cs.msu.ru

Чэнь Лун, магистрант кафедры алгоритмических языков факультета вычислительной математики и кибернетики, ФГБОУ ВО «Московский государственный университет имени М.В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1), ORCID: <https://orcid.org/0009-0006-6460-610X>, chenlong_2022@rambler.ru

Все авторы прочитали и одобрили окончательный вариант рукописи.

About the authors:

Irina N. Polyakova, Associate Professor of the Department of Algorithmic Languages, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation), Cand. Sci. (Phys.-Math.), Associate Professor, ORCID: <https://orcid.org/0000-0003-1432-4906>, polyakova@cs.msu.ru

Long Chen, Master Degree Student of the Department of Algorithmic Languages, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation), ORCID: <https://orcid.org/0009-0006-6460-610X>, chenlong_2022@rambler.ru

All authors have read and approved the final manuscript.

