УДК 004.421.2:519.178 DOI: 10.18698/0236-3933-2021-2-4-22

СПОСОБЫ ВЫДЕЛЕНИЯ СООБЩЕСТВ С ОПРЕДЕЛЕННЫМИ ТИПАМИ ОТНОШЕНИЙ В ГРАФАХ НА ОСНОВЕ БИЛЛИНГОВОЙ ИНФОРМАЦИИ

Н.С. Вирцева virtseva@list.ru
И.Э. Вишняков vishnyakov@bmstu.ru
И.П. Иванов ivanov@bmstu.ru

МГТУ им. Н.Э. Баумана, Москва, Российская Федерация

Аннотация

В настоящее время одной из актуальных задач анализа графов является выделение сообществ. Разработано большое число алгоритмов для выделения сообществ в графах. Часто такие сообщества не имеют ничего общего с группами людей (семьей, коллегами, друзьями), а используются для упрощения представления графа. Для большого числа задач полезным является выделение именно группы людей, плотно общающихся друг с другом. Многие алгоритмы выделения сообществ не учитывают того, что один участник может входить в несколько сообществ, а это является необходимым условием при выделении круга общения. Рассмотрены основные подходы к выделению сообществ, среди которых отмечены подходы, основанные на оптимизации функционала, поиске клик, кластеризации и распространении меток. Отдельно рассмотрены подходы, базирующиеся на анализе эго-сетей, т. е. рассматривающие подграф, образованный связями одного участника. Приведены основные алгоритмы, применяемые для выделения в графах сообществ с определенными типами отношений на основе биллинговой информации, и результаты анализа графов, построенных на основе этой информации, полезные для выделения сообществ

Ключевые слова

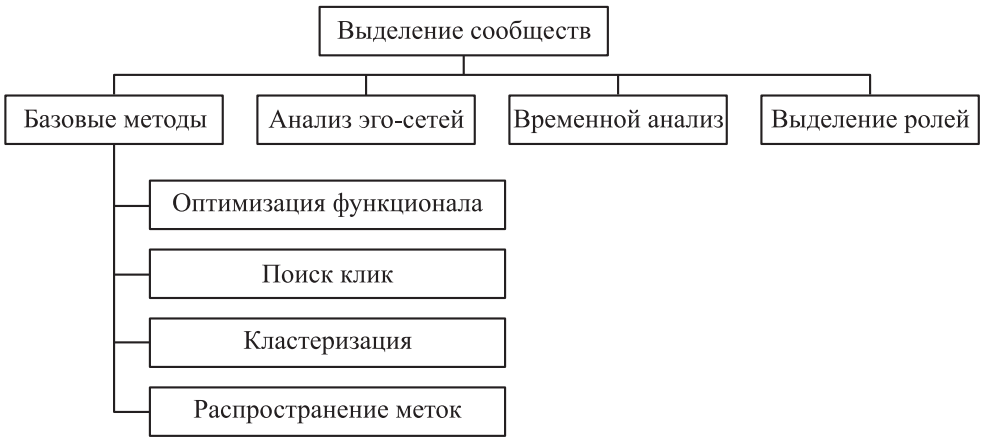
Выделение сообществ, анализ графов, биллинг

Поступила 13.07.2020
Принята 14.08.2020

Введение. Различные связи между людьми представляются в виде графов из тысяч вершин и миллионов ребер, что делает необходимым их анализ и обработку. Одним из способов анализа графов является выделение сообществ. Существует много вариантов определения понятия сообщества [1, 2].

В общем случае под сообществом понимается некоторый подграф, внутри которого вершины связаны между собой более тесно, чем с остальными вершинами графа. Каждый алгоритм выделения сообществ основывается на своем определении понятия сообщества.

Разработано большое число алгоритмов выделения сообществ, общая схема рассматриваемых методов приведена на рисунке. Часть из этих алгоритмов основана на том, что каждая вершина графа относится только к одному сообществу [3–5]. В реальности человек может состоять в нескольких социальных группах. Например, два человека могут быть одновременно коллегами и родственниками. Для выделения таких сообществ созданы алгоритмы поиска перекрывающихся сообществ [1, 2, 6]. Графы связей между людьми постоянно меняются во времени, поскольку люди с кем-то перестают общаться, а с кем-то знакомятся. В то же время такие группы, как семья или лучшие друзья, например, имеют тенденцию оставаться постоянными. Для изучения структуры сообществ и их изменений во времени используются методы временного анализа [3, 7, 8]. Выделение ролей сводится к классификации вершин графов, в которой вершины с общими свойствами относятся к одному классу [2], и может оказаться полезным при выделении сообществ.



Общая структура методов выделения сообществ

Под графом понимается пара (V, E), где V — набор вершин, а E — набор ребер или связей, представленных парами вершин из V. При задании связей между людьми в виде графа люди представляются вершинами, а связи между ними — ребрами в графе [9]. Под сообществом понимается некоторая сплоченная группа или набор вершин графа, тесно связанных между собой. Для компьютерной обработки необходимы

формальные критерии, отражающие это интуитивное понимание сплоченности [3]:

взаимная зависимость (mutuality), члены групп выбирают друг друга для включения в группу; с точки зрения теории графов это означает смежность вершин;

компактность (compactness), члены групп взаимно достижимы, хотя и необязательно смежны; в теории графов это можно выразить двумя способами: взаимная достижимость может интерпретироваться как «небольшое» расстояние (short distance) между вершинами или высокая связность (high connectivity);

плотность (density), члены групп имеют большое число контактов друг с другом; в терминах теории графов это означает, что у членов группы имеется много соседей внутри группы;

разделение (separation), члены групп имеют больше контактов внутри группы, чем снаружи.

Эти критерии по-другому можно назвать так: полная зависимость (complete mutuality), достижимость (reachability), степень вершин (vertex degree) и сравнение внутренней и внешней сплоченности/связности (the comparison of internal versus external cohesion) [1].

Сообщества называются перекрывающимися, если они содержат общие вершины. В работе [10] предложены минимальные требования к понятию сообществ, которые могут перекрываться: связность (connectedness) — сообщество должно быть связанным подграфом; локальная оптимальность (local optimality) — в соответствии с некоторой метрикой, определенной для подмножества вершин, сообщество не может быть улучшено добавлением или удалением вершин.

На основе данных характеристик формулируются различные определения того, что считать группой или сообществом. Соответствующие сообщества, как правило, являются максимальными подграфами анализируемого графа.

Максимальным подграфом является такой подграф, который не может быть увеличен добавлением вершин или ребер с сохранением всех свойств, указанных в определении.

Внутренней степенью сообщества называется число ребер, соединяющих вершины сообщества; соответствующие ребра называются внутренними.

Внешней степенью сообщества называется число ребер, соединяющих вершину сообщества с вершиной графа, не принадлежащей этому сообществу. Соответствующие ребра называются внешними.

Сообщества называются вложенными, если все вершины одного сообщества содержатся в другом сообществе. При анализе графов также используется понятие случайного графа. Под случайным графом понимается граф, построенный по двум параметрам: числу n вершин графа и вероятности p существования ребра между двумя вершинами. Каждая пара вершин с равной вероятностью p соединена ребром независимо от других пар. В некоторых случаях для проверки имеет ли граф структуру, в которой можно выделить сообщества, используют нулевую модель — это случайный граф, совпадающий с исходным графом по некоторым структурным свойствам. Самая известная нулевая модель, предложенная Ньюманом, представляет собой граф, состоящий из вершин исходного графа, ребра в который добавляются случайно в предположении, что степень каждой вершины совпадает со степенью вершины в исходном графе [1].

Большинство способов выделения сообществ рассчитано на то, что требуется упростить представление графа, но не учитывают, что требуется получить сообщества, соответствующие некоторым группам плотно общающихся между собой людей [11]. В настоящей работе приведен обзор существующих методов, которые могут быть применены к выделению сообществ, соответствующих реальным кругам общения людей, таким как семья, коллеги или друзья. Для обзора отобраны алгоритмы, которые позволяют выделять именно перекрывающиеся сообщества. В настоящей работе приведен обзор базовых алгоритмов выделения сообществ. Выделение сообществ также рассмотрено с точки зрения отдельного пользователя, в этом случае применяют анализ эго-сетей. Приведены методы анализа таких сетей.

Алгоритмы и методы выделения сообществ. Существует большое число алгоритмов, позволяющих выделить сообщества в графе. Часть из них основана на том, что каждую вершину графа можно отнести только к одному сообществу [1–4]. К алгоритмам данного класса принадлежат многие алгоритмы кластеризации вершин графа [12], алгоритмы, основанные на оптимизации функции модулярности [1, 13, 14]. В случае выделения таких групп, как семья, друзья, коллеги, каждая вершина может принадлежать нескольким группам. В таком случае необходимы алгоритмы для выделения перекрывающихся сообществ (overlapping communities). Многие из них являются модификациями методов поиска неперекрывающихся сообществ. Среди таких алгоритмов выделяют следующие: выделение клик (clique percolation), оптимизация некоторой функции (function optimization), кластеризация связей (link clustering), алгоритмы распространения меток (label propagation) и др. [15–17].

Алгоритмы, использующие оптимизацию функции (node seeds and local expansion, local expansion and optimization). В данном случае сообщество определяется как подграф, оптимизирующий некоторую функцию, отражающую качество выделенных сообществ [18]. Разные перекрывающиеся подграфы могут быть локально оптимальными, поэтому вершины могут принадлежать нескольким сообществам. Выделение сообществ в графе сводится к поиску локально оптимальных подграфов [1, 15]. Как правило, сначала рассматриваются отдельные вершины или небольшие наборы вершин, к которым добавляются другие вершины так, чтобы улучшить значение функции. Полученные оптимальные с точки зрения рассматриваемой функции группы вершин и являются сообществами. В зависимости от выбора функции такой подход может применяться к различным графам (взвешенным, ориентированным).

Одним из эффективных алгоритмов поиска оптимальных с точки зрения функции сообществ является алгоритм IS^2 [18]. Данный алгоритм применяется к взвешенным неориентированным графам и является комбинацией двух других алгоритмов — последовательного обхода (Iterative Scan, IS) и удаления по рангу (Rank Removal, RaRe). Алгоритм IS начинается с выбора произвольной вершины в графе в качестве начального сообщества. На каждом шаге к сообществу добавляется или удаляется из него одна вершина графа до тех пор, пока значение функции не улучшится. Значением функции в данном случае является отношение сумм весов ребер внутри сообщества и весов всех ребер, хотя бы одна вершина каждого из которых принадлежит сообществу. Когда значение функции невозможно улучшить, выбирается другая вершина для поиска нового сообщества. Процесс останавливается, если в результате выбора новой вершины несколько итераций подряд получается уже найденное сообщество [15]. Алгоритм RaRe начинается с присвоения ранга всем вершинам графа. Ранг вершины определяется с помощью мер центральности (centrality scores) [1, 2, 13]. К таким мерам относятся степень вершин, степень посредничества (betweenness centrality) [13] и ранг страницы (PageRank) [1]. Затем из графа удаляются вершины с высоким рангом, что приводит к разбиению графа на отдельные связные компоненты, которые используются в качестве начальных сообществ. Удаленные вершины добавляют к таким компонентам в том случае, если это улучшает значение функции. Каждая вершина может быть добавлена к нескольким компонентам, что делает возможным получение перекрывающихся сообществ. В комбинации алгоритмов IS^2 алгоритм IS применяется для уточнения результатов выполнения алгоритма RaRe. Изначально стро-

ятся сообщества с помощью алгоритма RaRe. На каждом шаге, как и в IS, выбирается произвольная вершина v в качестве начального сообщества, к которой добавляются другие вершины графа. Однако добавляемые вершины выбираются не из всего графа, а только из сообщества c, полученного с помощью RaRe и содержащего вершину v, а также из соседних с c сообществ. Сложность алгоритма IS^2 равна $O(mk + n)$, где m — число ребер; k — число итераций алгоритма; n — число вершин графа.

Алгоритм применяется к невзвешенным неориентированным графам и использует функцию оценки сообщества (fitness function), являющуюся отношением числа ребер внутри сообщества к суммарному числу его внешних и внутренних ребер, возведенному в степень α [19]. С помощью параметра α контролируется максимальный размер сообщества. В этом случае сначала формируется сообщество, состоящее из одной произвольно выбранной вершины. К этому сообществу добавляется вершина, смежная хотя бы с одной вершиной этого сообщества и имеющая наибольшее положительное значение оценочной функции. Затем выполняется пересчет значений функции для вершин, и из сообщества удаляются вершины с отрицательной оценкой. Таким образом формируются сообщества для каждой вершины, не попавшей еще ни в одно сообщество. Аналогичный подход приведен в [20]. В этом случае добавляемая вершина выбирается, исходя из предположения, если число ребер между вершиной и сообществом сильно отличается от ожидаемого числа в нулевой модели, то связь между вершиной и сообществом сильная. В [18] также предлагаются способы применения данного алгоритма к взвешенным и ориентированным графам. Сложность этих алгоритмов равна $O(n^2)$.

Алгоритмы с поиском клик. Такие алгоритмы тем или иным образом выполняют поиск клик для выделения сообществ. Под k-кликой (k-clique) понимается полный подграф из k вершин (т. е. каждая вершина подграфа смежная со всеми другими вершинами подграфа).

Самым известным алгоритмом для выделения перекрывающихся сообществ является алгоритм CPM (Clique Percolation Method) [1, 21, 22], позволяющий выделять перекрывающиеся сообщества в неориентированных невзвешенных графах. Для описания алгоритма вводятся дополнительные определения. Две k-клики называются смежными, если у них $(k–1)$ общая вершина. Сообществом из k-клик называется такое их объединение, в котором любая k-клика является смежной хотя бы одной k-клике из этого же объединения, и в графе нет других k-клик, удовлетворяющих этому условию. Такие сообщества и являются результатом работы алгоритма. Одной

из эффективных реализаций данного алгоритма является система поиска и визуализации перекрывающихся сообществ CFinder [21]. Алгоритм CPM можно разбить на три части. Сначала находятся все максимальные k-клики, т. е. k-клики, не являющиеся частью большей клики. Затем строится матрица перекрытий (overlap matrix), содержащая информацию о пересечениях найденных k-клик. На последнем шаге по матрице перекрытий находятся k-клики с максимальным пересечением, которые и возвращаются в качестве сообществ. В исходном виде алгоритм не применим к взвешенным или ориентированным графам, однако есть модификации, позволяющие использовать этот подход для таких графов. Сложность алгоритма равна $O\left(m\frac{\ln m}{10}\right)$. Данный подход позволяет работать с графами, содержащими ~10 000 вершин, после чего время работы резко увеличивается.

Предыдущий алгоритм позволяет выделять перекрывающиеся сообщества, однако на практике встречаются также иерархически вложенные сообщества. Для поиска перекрывающихся и иерархически вложенных сообществ в неориентированных невзвешенных графах был предложен алгоритм EAGLE (agglomerative Hierarchical clustering based on maximal clique). В этом случае сначала находятся все максимальные клики исходного графа, затем вводится некоторое пороговое значение k. В качестве исходных сообществ выбираются подграфы двух типов: клики размера не меньше, чем k; отдельные вершины, не принадлежащие таким кликам [23]. Далее наиболее близкие сообщества итеративно объединяются, пока не останется одно сообщество. Для вычисления близости сообществ и «обрезки» полученной дендрограммы используются функции, основанные на модулярности. Сложность алгоритма равна $O(n^2 + (h + 2)s)$, где h — число смежных пар максимальных клик; s — число всех максимальных клик. Приведенный алгоритм, как и предыдущий, применим к графам с десятками тысяч ребер и вершин.

Еще один подход, использующий клики, приведен в [24]. Алгоритм применяется к невзвешенным неориентированным графам. Предлагается формировать сообщества, расширяя клики небольшого размера. К начальным кликам небольшого размера присоединяются вершины, оптимизирующие значение оценочной функции. Таким образом, формируются сообщества, возможно, перекрывающиеся. Этот алгоритм показывает значительно меньшее время выполнения по сравнению с алгоритмом CPM и может применяться к графам с сотнями тысяч вершин. Сложность алгоритма равна $O(mh)$, где h — число клик в графе.

Алгоритмы, использующие кластеризацию связей (link clustering). В предыдущих алгоритмах сообщество рассматривалось как набор тесно связанных между собой вершин. В некоторых случаях сообщества имеют общие вершины, при этом в графе отсутствуют ребра между ними [1]. В методах кластеризации связей сообщество рассматривается как набор тесно связанных ребер [1, 2, 15, 25, 26]. Граф связей (link graph) строится на основе исходного графа следующим образом: вершинами нового графа являются ребра исходного графа, а ребро между вершинами в новом графе добавляется в том случае, когда у соответствующих ребер исходного графа есть общая вершина. Выделяются сообщества, представляющие собой наборы вершин нового графа, т. е. наборы ребер исходного графа. Таким образом, одна вершина исходного графа может оказаться в двух группах, т. е. определяются перекрывающиеся сообщества.

В работе [25] предложено применять иерархическую агломеративную кластеризацию для группировки связей в неориентированном невзвешенном графе. Изначально каждая вершина графа связей является кластером. Метод основан на двух концепциях: сходство связей (link similarity) и плотность разбиения (partition density). Мера, показывающая сходство связей, применяется при кластеризации для выбора наиболее близких связей в целях их объединения в одно сообщество. В качестве такой меры используют отношение числа общих соседей двух сравниваемых вершин к общему числу соседей обеих вершин. Процесс выполняется, пока все кластеры не объединятся в один. Для определения того, как следует «обрезать» построенную дендрограмму, используют формулу плотности разбиения. Плотность разбиения сообщества вычисляется как число связей этого сообщества, нормализованное по минимальному и максимальному числам всех возможных связей между его вершинами, что позволяет определить значимость разбиений связей в дендрограмме. Полученные кластеры и являются сообществами. Сложность алгоритма равна $O(nk_{\max}^2)$, где $k_{\max}$ — максимальная степень вершины графа.

Для поиска сообществ в неориентированных невзвешенных графах используют генетический алгоритм [27]. Идеей является то, что новые наборы формируются заменой элементов — мутацией (mutation), и обменом участков — скрещиванием (crossover), исходного набора, а затем наборов, сформированных указанными операциями. Изначально генерируются последовательности вершин графа так, что каждая следующая вершина последовательности выбирается из смежных с ней вершин. Каждая последовательность является начальным сообществом. Затем

выполняется фиксированное число итераций, и на каждой генерируются новые сообщества. Из текущего набора сообществ выбираются сообщества с наилучшим значением функции модулярности. Каждое новое сообщество получается заменой отдельных вершин на вершины из других сообществ (скрещивание), а затем заменой отдельных вершин на смежные с ними вершины в графе (мутация). Сложность алгоритма равна $O(n^2)$.

Еще один способ выделения сообществ в невзвешенных неориентированных графах предложен в работе [28]. На первом шаге предлагается найти некоторую начальную клику. Предполагается, что клика является частью сообщества, для получения которого используется техника кластеризации — к ней добавляются соседние вершины, оптимизирующие функции силы связи (joint strength) и степени принадлежности (membership degree) ребра, соединяющего вершину с кликой. Затем из графа удаляются все связи полученного сообщества. Данный процесс повторяется до тех пор, пока удается найти хотя бы одну подходящую начальную клику в графе. Преимуществом данного алгоритма по сравнению с другими алгоритмами кластеризации связей является уменьшение числа вычислений с каждой итерацией за счет удаления связей найденного сообщества. Кроме того, становится невозможным получение сообществ с большим числом общих вершин, так как не все связи обязательно будут отнесены к некоторому сообществу. Для кластеризации связей применяют технику кластеризации вершин с помощью меры сходства связей, что позволяет избежать вычислений, затратных по времени. Сложность алгоритма равна $O(nk_{\max})$.

Алгоритмы распространения меток (label propagation). Данные алгоритмы используются для выделения сообществ в ненаправленных графах, в которых ребра могут иметь веса. Каждой вершине присваивается метка, показывающая, к какому сообществу она относится. Изначально каждая вершина помечается своей уникальной меткой. На каждой итерации метки обновляются: в простейшем случае вершине присваивается метка, принадлежащая большинству соседних вершин. Таким образом, через несколько итераций одна и та же метка оказывается у нескольких вершин. Вершины с одинаковыми метками относят к одному сообществу [15, 29].

Для выделения перекрывающихся таким образом сообществ применяют различные способы обновления меток. Например, в [29] предложен алгоритм COPRA, в котором каждой вершине присваивается несколько меток, представленных в виде пары (c, b), где c — идентификатор сообщества;

b — коэффициент принадлежности вершины этому сообществу. Коэффициент имеет тем большее значение, чем сильнее привязана вершина к сообществу. Изначально все коэффициенты всех вершин равны единице. На каждой итерации коэффициенты пересчитываются следующим образом: для выбранного сообщества c коэффициент каждой вершины становится равным сумме коэффициентов смежных с ней вершин, нормализованной по числу этих смежных вершин. Для применения алгоритма к графу со взвешенными ребрами вес ребра учитывается при вычислении значения коэффициента метки. Для того чтобы не хранить в каждой вершине метки всех сообществ, вводится пороговое значение ν, показывающее максимальное число сообществ, которым может принадлежать вершина. В процессе поиска метки с коэффициентом, меньшим l/ν, удаляются. Сложность алгоритма равна $O\left(\nu m \log\left(\frac{\nu m}{n}\right)\right)$, где ν — число выделяемых сообществ, которое является входным параметром алгоритма. Есть и другие методы ограничения числа сообществ, которым может принадлежать одна вершина. Ограничения строят, например, на зависимости числа сообществ, в которые входит вершина, от ее степени [15]. Приведенные методы рассчитаны на графы с десятками тысяч вершин и сотнями тысяч ребер.

Другие методы. Еще один алгоритм для выделения сообществ называется BigClam (Cluster Affiliation Model for Big Networks) [17, 30]. Данный алгоритм позволяет выделять как неперекрывающиеся, так и перекрывающиеся и иерархически вложенные сообщества в неориентированных графах без атрибутов ребер и вершин. Основной задачей при разработке алгоритма являлся поиск сообществ с плотными перекрывающими частями, так как существующие на тот момент алгоритмы основывались на том, что общие части сообществ разреженные. Алгоритм основан на вычислении принадлежности вершины сообществам, которое максимизирует предложенную функцию с помощью неотрицательной матричной факторизации. Функция формируется исходя из того, что вероятность существования ребра между двумя вершинами увеличивается с числом общих для этих вершин сообществ. Это позволяет применять алгоритм к графам с миллионами вершин и ребер. Сложность алгоритма равна $O(nN(u))$, $N(u)$ — число вершин, смежных с вершиной u.

Во многих методах используется анализ связей вершин или анализ атрибутов вершин, но эта информация по отдельности не является достаточной для точного выделения сообществ [31]. Например, информация о связях вершин часто бывает разреженной и зашумленной, а среди

атрибутов встречаются несущественные, отрицательно влияющие на результат выделения сообществ. В [31] рассматривается алгоритм выделения сообществ во взвешенных графах, комбинирующий два типа моделей — построенных на связях и на атрибутах вершин. Авторы описывают собственный подход к комбинированию моделей связей и атрибутов. Чтобы уменьшить влияние атрибутов вершин, которые могут отрицательно повлиять на результат, предлагается дискриминационная модель (discriminative model), позволяющая присвоить вершинам веса с учетом важности атрибутов. Для связей предлагается использовать скрытые параметры «популярности» вершин, основанные на условной вероятности того, что каждая из вершин, соединяемых ребром, будет указывать на другие вершины. Предлагается также способ комбинирования предложенных моделей, основанный на максимизации вероятности принадлежности вершин сообществам. Сложность алгоритма равна $O(M(mC_1 + + nC_2 + T_3))$, где C_1 и C_2 — константные параметры алгоритма; T_3 — время, требуемое для решения построенной задачи максимизации.

Анализ эго-сетей. Некоторые алгоритмы основываются не на поиске самих сообществ в огромном графе, а на рассмотрении связей отдельного человека, решая тем самым задачу выделения круга друзей, коллег и семьи. При анализе социальных сетей часто рассматривают так называемые эго-сети (ego-networks) или эгоцентричные графы (egocentric graphs), т. е. графы, образованные связями одного человека. Такую сеть также называют социальной сетью типа звезда (star social network).

Метод, использующий данные профилей социальных сетей. В работе [11] на примере социальных сетей описывается алгоритм выделения сообществ, таких как семья, коллеги, друзья, на основе данных, извлеченных из пользовательских профилей и общих фотографий. Пользователи представляются векторами признаков, извлеченными из их профилей. Для выделения отдельных кругов общения данного человека задаются правила, т. е. условия, которым должны соответствовать извлеченные признаки, чтобы отнести некоторого другого человека к определенному кругу, например к коллегам. Одним из правил отнесения человека к кругу семьи, например, является наличие фотографии с группой людей, сильно различающихся по возрасту и находящихся в неизвестном месте.

Метод на основе выделения групп с общими свойствами. Еще один подход — это выделение социальных кругов пользователя как групп с общими свойствами [32]. В данном случае информация тоже извлекается из профилей социальных сетей. Используются такие признаки, как пол,

имена, фамилии, должности на работе, институты, университеты, места жительства, дни рождения, политические предпочтения, наличие записей на стенах социальных сетей от других пользователей и др. Для описания отличия между двумя пользователями задают вектор отличий. Каждый элемент вектора показывает, совпадают ли значения данного признака у двух пользователей или нет. Социальные круги пользователя должны удовлетворять двум свойствам. Первое свойство требует, чтобы у всех членов круга были общие отношения друг с другом, т. е. векторы отличий должны быть похожи. Второе свойство состоит в том, что у всех членов круга должны быть более близкие отношения по сравнению с другими людьми, с которыми связан исследуемый человек. Это значит, что для пары пользователей и любого третьего пользователя, не входящего в данный социальный круг, разность векторов отличий первого с третьим и второго с третьим пользователем должна быть достаточно мала. На основе указанных свойств выделяют перекрывающиеся круги общения пользователя [32]. Данный алгоритм отличается тем, что не требует дополнительной исходной информации о том, какими должны получиться искомые сообщества, изначально требуется только произвольный набор признаков.

Заключение. Приведены основные алгоритмы и подходы, которые могут использоваться при выделении сообществ. Для обзора отобраны те из них, которые применимы к социальным графам и позволяют выделять пересекающиеся сообщества, что является необходимым условием для выделения реальных кругов общения людей. Результаты работы алгоритмов отличаются, так как в зависимости от задачи понимание того, что считать сообществом, как правило, варьируется. В связи с этим окончательный выбор алгоритма можно сделать по результатам тестированияи анализа его применимости к решению поставленной задачи.

Выполнено сравнение алгоритмов по сложности и применимости к анализу графов определенных размеров, что позволит сделать обоснованный выбор алгоритмов для конкретной задачи, связанной с получением сообществ в социальных графах. Рассмотренные алгоритмы не дают сразу требуемого результата, поскольку все они выделяют сообщества, но дают разные результаты. Преобразование графа с учетом различных способов обработки и получения признаков из биллинговой информации может улучшить результат. Для улучшения качества выделения сообществ к стандартным алгоритмам выделения может быть добавлен временной анализ, а в качестве альтернативы возможно применение алгоритмов выделения ролей.

ЛИТЕРАТУРА

[1] Fortunato S. Community detection in graphs. *Phys. Rep.*, 2010, vol. 486, no. 3-5, pp. 75–174. DOI: https://doi.org/10.1016/j.physrep.2009.11.002

[2] Lehmann S. Community detection, current and future research trends. In: Encyclopedia of Social Network Analysis and Mining. New York, Springer, 2014, pp. 214–220.

[3] Brandes U., Erlebach T. Network analysis. Berlin, Springer, 2005.

[4] Lancichinetti A., Fortunato S. Community detection algorithms: a comparative analysis. *Phys. Rev. E*, 2009, vol. 80, no. 5, art. 056117.
DOI: https://doi.org/10.1103/PhysRevE.80.056117

[5] Ключарев П.Г., Басараб М.А. Спектральные методы анализа социальных сетей. *Наука и образование: научное издание МГТУ им. Н.Э. Баумана*, 2017, № 5, с. 168–177. URL: https://www.elibrary.ru/item.asp?id=30585833

[6] Xie J., Kelley S., Szymanski B.K. Overlapping community detection in networks: the state-of-the-art and comparative study. *ACM Comput. Surv.*, 2013, vol. 45, no. 4, art. 43. DOI: https://doi.org/10.1145/2501654.2501657

[7] Nanavati A.A., Singh R., Chakraborty D. Analyzing the structure and evolution of massive telecom graphs. *IEEE Trans. Knowl. Data Eng.*, 2008, vol. 20, no. 5, pp. 703–718. DOI: https://doi.org/10.1109/TKDE.2007.190733

[8] Zheleva M., Schmitt P., Vigil M., et al. Community detection in cellular network traces. *Proc. ICDT 13*, 2013, vol. 2, pp. 183–186.
DOI: https://doi.org/10.1145/2517899.2517932

[9] Басараб М.А., Иванов И.П., Колесников А.В. и др. Обнаружение противоправной деятельности в киберпространстве на основе анализа социальных сетей: алгоритмы, методы и средства (обзор). *Вопросы кибербезопасности*, 2016, № 4, с. 11–19. DOI: https://doi.org/10.21681/2311-3456-2016-4-11-19

[10] Goldberg M., Kelley S., Magdon-Ismail M., et al. Finding overlapping communities in social networks. *IEEE 2nd Int. Conf. Soc. Comput.*, 2010.
DOI: https://doi.org/10.1109/SocialCom.2010.24

[11] Raad E., Chbeir R., Dipanda A. Discovering relationship types between users using profiles and shared photos in a social network. *Multimed. Tools Appl.*, 2013, vol. 64, no. 1, pp. 141–170. DOI: https://doi.org/10.1007/s11042-011-0853-7

[12] Eynard D., Javarone M.A., Matteucci M. Clustering algorithms. In: Encyclopedia of Social Network Analysis and Mining. New York, Springer, 2014, pp. 101–114.

[13] Borgatti S.P., Everett M.G. A graph-theoretic perspective on centrality. *Soc. Networks*, 2006, vol. 28, no. 4, pp. 466–484.
DOI: https://doi.org/10.1016/j.socnet.2005.11.005

[14] Everett M.G. Classical algorithms for social network analysis: future and current trends. In: Encyclopedia of Social Network Analysis and Mining. New York, Springer, 2014, pp. 88–94.

[15] Amelio A., Pizzuti C. Overlapping community discovery methods: a survey. In: Social Networks: Analysis and Case Studies. Vienna, Springer, 2014, pp. 105–125.

[16] Mateos P. Demographic, ethnic and socioeconomic community structure in social networks. In: Encyclopedia of Social Network Analysis and Mining. New York, Springer, 2014, pp. 342–346.

[17] Yang J., Leskovec J. Overlapping community detection at scale: a nonnegative matrix factorization approach. *Proc. WSDM 13*, 2013, pp. 587–596.
DOI: https://doi.org/10.1145/2433396.2433471

[18] Baumes J., Goldberg M.K., Krishnamoorthy M.S., et al. Finding communities by clustering a graph into overlapping sub-graphs. *Proc. IADIS Int. Conf. Appl. Comput.*, 2005, vol. 5, pp. 97–104.

[19] Lancichinetti A., Fortunato S., Kertész J. Detecting the overlapping and hierarchical community structure of complex networks. *New J. Phys.*, 2009, vol. 11, no. 3, art. 033015.
DOI: https://doi.org/10.1088/1367-2630/11/3/033015

[20] Lancichinetti A., Radicchi F., Ramasco J.J., et al. Finding statistically significant communities in networks. *PLoS ONE*, 2011, vol. 6, no. 4, art. e18961.
DOI: https://doi.org/10.1371/journal.pone.0018961

[21] Palla G., Derényi I., Farkas I., et al. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 2005, vol. 435, no. 7043, pp. 814–818. DOI: https://doi.org/10.1038/nature03607

[22] Derényi I., Palla G., Vicsek T. Clique percolation in random networks. *Phys. Rev. Lett.*, 2005, vol. 94, no. 16, art. 160202.
DOI: https://doi.org/10.1103/PhysRevLett.94.160202

[23] Shen H., Cheng X., Cai K., et al. Detect overlapping and hierarchical community structure in networks. *Physica A*, 2009, vol. 388, no. 8, pp. 1706–1712. DOI: https://doi.org/10.1016/j.physa.2008.12.021

[24] Lee C., Reid F., McDaid A., et al. Detecting highly overlapping community structure by greedy clique expansion. *Proc. 4th Workshop Soc. Netw. Mining Anal. Held in Conjunction with Int. Conf. Knowl. Discov. Data Mining*, 2010, pp. 33–42.

[25] Ahn Y.Y., Bagrow J.P., Lehmann S. Link communities reveal multiscale complexity in networks. *Nature*, 2010, vol. 466, pp. 761–764.
DOI: https://doi.org/10.1038/nature09182

[26] Coscia M., Giannotti F., Pedreschi D. Extracting and inferring communities via link analysis. In: Encyclopedia of Social Network Analysis and Mining. New York, Springer, 2014, pp. 517–525.

[27] Dickinson B., Valyou B., Hu W. A genetic algorithm for identifying overlapping communities in social networks using an optimized search space. *Soc. Netw.*, 2013, vol. 2, no. 4. DOI: https://doi.org/10.4236/sn.2013.24019

[28] Ding Z., Zhang X., Sun D., et al. Overlapping community detection based on network decomposition. *Sc. Rep.*, 2016, vol. 6, art. 24115.
DOI: https://doi.org/10.1038/srep24115

[29] Gregory S. Finding overlapping communities in networks by label propagation. *New J. Phys*., 2010, vol. 12, no. 10, art. 103018.
DOI: https://doi.org/10.1088/1367-2630/12/10/103018

[30] Yang J., Leskovec J. Community-affiliation graph model for overlapping network community detection. *IEEE 12th Int. Conf. Data Mining*, 2012, pp. 1170–1175.
DOI: https://doi.org/10.1109/ICDM.2012.139

[31] Yang T., Jin R., Chi Y., et al. Combining link and content for community detection. In: Encyclopedia of Social Network Analysis and Mining. New York, Springer, 2014, pp. 190–201.

[32] McAuley J.J., Leskovec J. Learning to discover social circles in ego networks. In: Advances in Neural Information Processing Systems. *NeurIPS*, 2012, vol. 1, pp. 539–547.

Вирцева Наталья Сергеевна — магистр кафедры «Прикладная математика», ассистент кафедры «Теоретическая информатика и компьютерные технологии» МГТУ им. Н.Э. Баумана (Российская Федерация, 105005, Москва, 2-я Бауманская ул., д. 5, корп. 1).

Вишняков Игорь Эдуардович — старший преподаватель кафедры «Теоретическая информатика и компьютерные технологии» МГТУ им. Н.Э. Баумана (Российская Федерация, 105005, Москва, 2-я Бауманская ул., д. 5, корп. 1).

Иванов Игорь Потапович — д-р техн. наук, заведующий кафедрой «Теоретическая информатика и компьютерные технологии» МГТУ им. Н.Э. Баумана (Российская Федерация, 105005, Москва, 2-я Бауманская ул., д. 5, корп. 1).

Просьба ссылаться на эту статью следующим образом:

Вирцева Н.С., Вишняков И.Э., Иванов И.П. Способы выделения сообществ с определенными типами отношений в графах на основе биллинговой информации. *Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение*, 2021, № 2 (135), с. 4–22. DOI: https://doi.org/10.18698/0236-3933-2021-2-4-22

METHODS OF DETECTING COMMUNITIES WITH CERTAIN RELATIONSHIP TYPES IN GRAPHS USING BILLING INFORMATION

N.S. Virtseva virtseva@list.ru
I.E. Vishnyakov vishnyakov@bmstu.ru
I.P. Ivanov ivanov@bmstu.ru

Bauman Moscow State Technical University, Moscow, Russian Federation

Abstract

Currently, one of the urgent tasks of graph analysis is community detection. A large number of algorithms have been developed for detecting communities

Keywords

Community detection, graph analysis, billing

in graphs. Meanwhile, these communities have nothing to do with groups of people, i.e., family, colleagues, friends, and are used to simplify the graph representation. For a large number of tasks, it is useful to detect a group of people who closely communicate with each other. Many algorithms for detecting communities do not take into account that one participant can belong to several communities, and this is a prerequisite for detecting social circles. The paper overviews the main approaches to community detection, and among these emphasizes the approaches based on functionality optimization, clique problem, cluster analysis and label distribution. The approaches based on the analysis of ego-networks, i.e., considering the subgraph formed by the connections of one participant, are considered separately. The study gives the basic algorithms that are applicable for the selection of communities with certain relationship types based on billing information. Findings of research are useful for community detection depending on the task and available input data

Received 13.07.2020
Accepted 14.08.2020

REFERENCES

[1] Fortunato S. Community detection in graphs. *Phys. Rep*., 2010, vol. 486, no. 3-5, pp. 75–174. DOI: https://doi.org/10.1016/j.physrep.2009.11.002

[2] Lehmann S. Community detection, current and future research trends. In: Encyclopedia of Social Network Analysis and Mining. New York, Springer, 2014, pp. 214–220.

[3] Brandes U., Erlebach T. Network analysis. Berlin, Springer, 2005.

[4] Lancichinetti A., Fortunato S. Community detection algorithms: a comparative analysis. *Phys. Rev. E*, 2009, vol. 80, no. 5, art. 056117.
DOI: https://doi.org/10.1103/PhysRevE.80.056117

[5] Klyucharev P.G., Basarab M.A. Spectral analysis methods of social networks. *Nauka i obrazovanie: nauchnoe izdanie MGTU im. N.E. Baumana* [Science and Education: Scientific Publication], 2017, no. 5, pp. 168–177 (in Russ.).
Available at: https://www.elibrary.ru/item.asp?id=30585833

[6] Xie J., Kelley S., Szymanski B.K. Overlapping community detection in networks: the state-of-the-art and comparative study. *ACM Comput. Surv*., 2013, vol. 45, no. 4, art. 43. DOI: https://doi.org/10.1145/2501654.2501657

[7] Nanavati A.A., Singh R., Chakraborty D. Analyzing the structure and evolution of massive telecom graphs. *IEEE Trans. Knowl. Data Eng*., 2008, vol. 20, no. 5, pp. 703–718. DOI: https://doi.org/10.1109/TKDE.2007.190733

[8] Zheleva M., Schmitt P., Vigil M., et al. Community detection in cellular network traces. *Proc. ICDT 13*, 2013, vol. 2, pp. 183–186.
DOI: https://doi.org/10.1145/2517899.2517932

[9] Basarab M.A., Ivanov I.P., Kolesnikov A.V., et al. Detection of illegal activities in cyberspace on the basis of the social networks analysis: algorithms, methods, and tools (a survey). *Vopr. kiberbezop*., 2016, no. 4, pp. 11–19 (in Russ.).
DOI: https://doi.org/10.21681/2311-3456-2016-4-11-19

[10] Goldberg M., Kelley S., Magdon-Ismail M., et al. Finding overlapping communities in social networks. *IEEE 2nd Int. Conf. Soc. Comput*., 2010.
DOI: https://doi.org/10.1109/SocialCom.2010.24

[11] Raad E., Chbeir R., Dipanda A. Discovering relationship types between users using profiles and shared photos in a social network. *Multimed. Tools Appl*., 2013, vol. 64, no. 1, pp. 141–170. DOI: https://doi.org/10.1007/s11042-011-0853-7

[12] Eynard D., Javarone M.A., Matteucci M. Clustring algorithms. In: Encyclopedia of Social Network Analysis and Mining. New York, Springer, 2014, pp. 101–114.

[13] Borgatti S.P., Everett M.G. A graph-theoretic perspective on centrality. *Soc. Networks*, 2006, vol. 28, no. 4, pp. 466–484.
DOI: https://doi.org/10.1016/j.socnet.2005.11.005

[14] Everett M.G. Classical algorithms for social network analysis: future and current trends. In: Encyclopedia of Social Network Analysis and Mining. New York, Springer, 2014, pp. 88–94.

[15] Amelio A., Pizzuti C. Overlapping community discovery methods: a survey. In: Social Networks: Analysis and Case Studies. Vienna, Springer, 2014, pp. 105–125.

[16] Mateos P. Demographic, ethnic and socioeconomic community structure in social networks. In: Encyclopedia of Social Network Analysis and Mining. New York, Springer, 2014, pp. 342–346.

[17] Yang J., Leskovec J. Overlapping community detection at scale: a nonnegative matrix factorization approach. *Proc. WSDM 13*, 2013, pp. 587–596.
DOI: https://doi.org/10.1145/2433396.2433471

[18] Baumes J., Goldberg M.K., Krishnamoorthy M.S., et al. Finding communities by clustering a graph into overlapping sub-graphs. *Proc. IADIS Int. Conf. Appl. Comput*., 2005, vol. 5, pp. 97–104.

[19] Lancichinetti A., Fortunato S., Kertész J. Detecting the overlapping and hierarchical community structure of complex networks. *New J. Phys*., 2009, vol. 11, no. 3, art. 033015.
DOI: https://doi.org/10.1088/1367-2630/11/3/033015

[20] Lancichinetti A., Radicchi F., Ramasco J.J., et al. Finding statistically significant communities in networks. *PLoS ONE*, 2011, vol. 6, no. 4, art. e18961.
DOI: https://doi.org/10.1371/journal.pone.0018961

[21] Palla G., Derényi I., Farkas I., et al. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 2005, vol. 435, no. 7043, pp. 814–818. DOI: https://doi.org/10.1038/nature03607

[22] Derényi I., Palla G., Vicsek T. Clique percolation in random networks. *Phys. Rev. Lett*., 2005, vol. 94, no. 16, art. 160202.
DOI: https://doi.org/10.1103/PhysRevLett.94.160202

[23] Shen H., Cheng X., Cai K., et al. Detect Overlapping and hierarchical community structure in networks. *Physica A*, 2009, vol. 388, no. 8, pp. 1706–1712. DOI: https://doi.org/10.1016/j.physa.2008.12.021

[24] Lee C., Reid F., McDaid A., et al. Detecting highly overlapping community structure by greedy clique expansion. *Proc. 4th Workshop Soc. Netw. Mining Anal. Held in Conjunction with Int. Conf. Knowl. Discov. Data Mining*, 2010, pp. 33–42.

[25] Ahn Y.Y., Bagrow J.P., Lehmann S. Link communities reveal multiscale complexity in networks. *Nature*, 2010, vol. 466, pp. 761–764.
DOI: https://doi.org/10.1038/nature09182

[26] Coscia M., Giannotti F., Pedreschi D. Extracting and inferring communities via link analysis. In: Encyclopedia of Social Network Analysis and Mining. New York, Springer, 2014, pp. 517–525.

[27] Dickinson B., Valyou B., Hu W. A genetic algorithm for identifying overlapping communities in social networks using an optimized search space. *Soc. Netw.*, 2013, vol. 2, no. 4. DOI: https://doi.org/10.4236/sn.2013.24019

[28] Ding Z., Zhang X., Sun D., et al. Overlapping community detection based on network decomposition. *Sc. Rep.*, 2016, vol. 6, art. 24115.
DOI: https://doi.org/10.1038/srep24115

[29] Gregory S. Finding overlapping communities in networks by label propagation. *New J. Phys.*, 2010, vol. 12, no. 10, art. 103018.
DOI: https://doi.org/10.1088/1367-2630/12/10/103018

[30] Yang J., Leskovec J. Community-affiliation graph model for overlapping network community detection. *IEEE 12th Int. Conf. Data Mining*, 2012, pp. 1170–1175. DOI: https://doi.org/10.1109/ICDM.2012.139

[31] Yang T., Jin R., Chi Y., et al. Combining link and content for community detection. In: Encyclopedia of Social Network Analysis and Mining. New York, Springer, 2014, pp. 190–201.

[32] McAuley J.J., Leskovec J. Learning to discover social circles in ego networks. In: Advances in Neural Information Processing Systems. *NeurIPS*, 2012, vol. 1, pp. 539–547.

Virtseva N.S. — Master, Department of Applied Mathematics, Assist. Lecturer, Department of Theoretical Informatics and Computer Technologies, Bauman Moscow State Technical University (2-ya Baumanskaya ul. 5/1, Moscow, 105005 Russian Federation).

Vishnyakov I.E. — Assist. Professor, Department of Theoretical Informatics and Computer Technologies, Bauman Moscow State Technical University (2-ya Baumanskaya ul. 5/1, Moscow, 105005 Russian Federation).

Ivanov I.P. — Dr. Sc. (Eng.), Head of the Department of Theoretical Informatics and Computer Technologies, Bauman Moscow State Technical University (2-ya Baumanskaya ul. 5/1, Moscow, 105005 Russian Federation).

Please cite this article in English as:

Virtseva N.S., Vishnyakov I.E., Ivanov I.P. Methods of detecting communities with certain relationship types in graphs using billing information. *Herald of the Bauman Moscow State Technical University, Series Instrument Engineering*, 2021, no. 2 (135), pp. 4–22 (in Russ.). DOI: https://doi.org/10.18698/0236-3933-2021-2-4-22