Выделение сообществ в графе взаимодействующих объектов*

М. И. КОЛОМЕЙЧЕНКО
Федеральный исследовательский центр «Информатика и управление» Российской академии наук

И. В. ПОЛЯКОВ
Национальный исследовательский университет «Высшая школа экономики»
e-mail: igorp86@mail.ru

А. А. ЧЕПОВСКИЙ
Национальный исследовательский университет «Высшая школа экономики»
e-mail: c4hapa@gmail.com

А. М. ЧЕПОВСКИЙ
Национальный исследовательский университет «Высшая школа экономики»
e-mail: achepovskiy@hse.ru

УДК 004.421.2:519.178

Ключевые слова: анализ социальной сети, структура графа, алгоритм выделения сообществ.

Аннотация

Рассматривается задача анализа графа социальной сети и других взаимодействующих объектов. Описываются алгоритмы выделения сообществ в социальных сетях, приводится их классификация и анализ. Обсуждается применимость алгоритмов к прикладным задачам анализа больших графов социальных сетей.

Abstract

M. I. Kolomeychenko, I. V. Polyakov, A. A. Chepovskiy, A. M. Chepovskiy, Detection of communities in graph of interactive objects, Fundamentalnaya i prikladnaya matematika, vol. 21 (2016), no. 3, pp. 131—139.

This article describes the problem of analysis of social network graphs and other interacting objects. It also presents community detection algorithms in social networks, their classification and analysis. In addition, it considers applicability of algorithms for real tasks in social network graph analysis.

*Работа выполнена при поддержке РФФИ, гранты № 16-29-09546 офи_м и 16-07-00641.

Фундаментальная и прикладная математика, 2016, том 21, № 3, с. 131—139.

1. Введение

Граф взаимодействующих объектов — это модель социальной сети, анализ структуры которой является одной из актуальных задач современных информационных технологий [1, 2, 6].

Под социальной сетью [3] на качественном уровне понимается социальная структура, состоящая из множества агентов (субъектов, индивидуальных или коллективных, например: индивидов, семей, групп, организаций) и определённого на нём множества отношений (совокупности связей между агентами, например: знакомства, дружбы, сотрудничества, коммуникации). Актуальность исследования социальных сетей объясняется тем, что методы изучения социального взаимодействия субъектов, основанные на использовании социальных сетей, базируются на понимании социальных феноменов главным образом через свойства взаимосвязей между объектами сети.

Активно разрабатываются различные методы анализа графов социальных сетей [2, 6, 8, 9, 23], которые позволяют выявлять структуры графа взаимодействия субъектов, такие, например, как группы общения социальных сетей. При этом визуальные средства являются одним из методов анализа графа [5], в том числе и в совокупности с алгоритмами выявления сообществ [4], которые могут выступать как дополнительные средства автоматического размещения при визуализации графов в прикладном программном обеспечении.

Выделяют два типа графов социальных сетей. Обычно анализируют граф с исходно ненаправленными рёбрами, описывающими постоянные связи объектов социальной сети, который иногда называют «графом друзей». Существует граф контактов объектов социальной сети, описывающий пересылку различного рода информации между объектами. Этот второй тип графа содержит по природе своей направленные рёбра. Но для интересующего нас структурного анализа и выделения сообществ взаимодействия объектов сети граф второго типа можно рассматривать как ненаправленный, заменив направленные рёбра на ненаправленные.

Основной целью алгоритмов выделения сообществ в графах является установление внутренних свойств и взаимосвязей между вершинами, которые недоступны посредством прямого анализа.

Подробные обзоры по методам анализа социальных сетей, включая и структурный анализ графов, содержатся в [1—3, 9]. В данной работе рассматриваются несколько групп наиболее популярных алгоритмов выделения сообществ в социальных сетях на основе анализа структуры графа.

2. Алгоритмы, основанные на оценке характеристики «модулярность»

Для оценки качества разбиения графа на сообщества в ряде работ [14, 18—22] было введено понятие «модулярность» (далее «модулярность Ньюма-

на—Гирвана»), которое должно, по мнению их авторов, описывать качество разбиения графа на сообщества. При этом качество разбиения предполагает оценку того, насколько при заданном разбиении графа на группы плотность внутригрупповых связей больше плотности межгрупповых связей.

В качестве оценки модулярности Ньюмана—Гирвана используется величина, описывающая не то, насколько для данного разбиения внутригрупповые связи более плотные, чем межгрупповые, а то, насколько они более плотные по сравнению с некой начальной плотностью. Поэтому происходит сравнение с «нулевой гипотезой», заключающейся в том, что дуги распределены случайно, т. е. нет закономерностей в распределении плотности дуг внутри групп.

Показатель модульности оценивается следующим образом:

$$Q = \sum_{i,j} \left[\frac{A_{ij}}{2m} - \frac{k_i k_j}{4m^2} \right] \sigma(c_i, c_j), \tag{1}$$

где m — количество связей, A — матрица смежности графа, A_{ij} — наличие ребра между вершинами i и j,

$$\sigma(c_i, c_j) = \begin{cases} 1, & c_i = c_j, \\ 0, & c_i \neq c_j, \end{cases}$$

k_i — степень вершины i, c_i — номер класса, к которому принадлежит вершина i.

Наиболее эффективным из группы алгоритмов, основанных на активном использовании модулярности Ньюмана—Гирвана, является алгоритм Блонделя быстрой оптимизации модулярности [7]. Данный алгоритм находит разбиения больших графов с высокой степенью модулярности за короткое время, кроме того, предоставляет информацию о полной иерархической структуре сообществ, тем самым давая доступ к различным расширениям выделенных сообществ.

Предложенный в [7] метод разбивается на два этапа, которые повторяются итерационно. Предположим, что мы начали с взвешенного графа с N вершинами. В первую очередь каждой вершине сети назначается своё сообщество, т. е. каждая вершина является отдельным сообществом. Затем для каждой вершины i рассматриваются соседние вершины j и вычисляется прирост модулярности, который может иметь место при удалении вершины i из своего сообщества и добавлении её в сообщество вершины j. Вершина i переносится в то сообщество, где достигается максимальный положительный прирост значения модулярности. Если положительного прироста не существует, то вершина i остаётся на исходном месте. Данный процесс повторяется итерационно и последовательно для всех вершин графа до тех пор, пока не будет достигнуто улучшение показателя модулярности (1), после этого первая фаза алгоритма заканчивается.

Вторая фаза алгоритма заключается в построении нового графа, вершинами которого будут сообщества, найденные на первом этапе алгоритма. Веса связей между новыми вершинами представляют сумму весов связей между вершинами, которые соответствуют двум сообществам. Связи между вершинами одного

сообщества становятся петлями в новом графе. Как только вторая фаза алгоритма завершится, станет возможным применить снова первую фазу алгоритма к полученной новой взвешенной сети, и так далее.

Этапы алгоритма повторяются до тех пор, пока не будет достигнут локальный максимум модулярности. Данный алгоритм, по существу, напоминает внутреннюю природу сложных сетей и естественным образом включает в себя понятие иерархии, так как сообщества сообществ строятся в процессе работы алгоритма. Глубина полученной иерархии сообществ определяется числом итераций и обычно является небольшим числом, как было установлено из экспериментов с графами социальных сетей различных размеров.

Различные модификации алгоритмов Ньюмана [13, 14, 18—22] и алгоритм Радикки [24] основаны на похожих идеях. Это методы иерархического разбиения, где связи итеративно удаляются на основе информации о дугах, которые определяются значениями некоторых коэффициентов, введённых для описания частоты встречаемости этих дуг. Рассматриваются обычно либо коэффициент промежуточности, который является мерой того, насколько часто связь входит в кратчайшие пути между различными парами узлов, либо коэффициент группировки дуг, который определяет количество циклов, в которых состоит данная дуга. При этом вычисление коэффициента группировки дуг, как в алгоритме Радикки [24], не столь трудоёмко, как, например, коэффициента промежуточности. За счёт использования данного показателя асимптотическая сложность алгоритма составляет $O(n^2)$ на разреженных графах, где n — количество вершин в графе.

3. Метод, основанный на спектральных свойствах графа

Топология сети взаимодействующих объектов может быть описана строгими математическими методами. Структура графа сети может быть представлена симметричной матрицей Лапласа, диагональные элементы которой определяются степенью соответствующей вершины (количеством связей, исходящих из данных вершин), а недиагональные элементы определяются как -1, если существует связь между парой вершин, и как 0 в противном случае. Заметим, что сумма элементов любой строки или столбца такой матрицы равна нулю. Собственные значения матрицы Лапласа являются неотрицательными вещественными числами, а кратность нулевого собственного значения равна количеству компонент связности. Наименьшее положительное собственное значение определяет алгебраическую связность рассматриваемого графа. Соответствующий этому собственному значению собственный вектор содержит всю информацию об интересующей нас структуре графа. По значениям (величинам) компонент данного собственного вектора можно сгруппировать вершины графа по отдельным группам, которые мы называем сообществами.

На спектральных свойствах графа основан алгоритм Донетти—Муньоса [11]. Метод предусматривает получение некоторых характеристик для каждой вершины графа из решения задачи на собственное значение матрицы Лапласа. Естественно, что получаются достаточно объёмные данные для графов. Затем вершины группируются классическими иерархическими методами кластерного анализа. Из результирующих разбиений выбирают то, которое максимизирует модулярность Ньюмана—Гирвана. Метод работает за время $O(n^3)$, где n — количество вершин в графе, за счёт вычисления только нескольких собственных векторов с помощью итеративного алгоритма Ланцоша.

4. Алгоритмы, основанные на оценке энтропии системы

Структурный алгоритм Росвалля—Бергстрома [26] сводит задачу нахождения наилучшего структурного разбиения графа на сообщества к задаче оптимального сжатия информации о структуре графа, при котором после декодирования результат будет максимально близок к исходной структуре графа. Это достигается путём вычисления минимума функции, которая выражает наилучший компромисс между минимумом разницы исходной и сжатой информацией (максимальной точностью к исходной информации) и максимальным сжатием.

Динамический алгоритм Росвалля—Бергстрома [25, 27] основан на тех же принципах, что и структурный алгоритм Росвалля—Бергстрома. Разница заключается в том, что предыдущий алгоритм сжимал информацию о структуре графа, в то время как данный метод сжимает информацию о динамическом процессе, проходящем в графе (случайное блуждание). Оптимальное сжатие снова достигается оптимизацией показателя качества (минимальной длины описания случайного блуждания).

Для вычисления показателя качества заданного разбиения используется энтропия, описывающая среднюю длину кодового слова, взятого для кодирования вершины. Применение понятия энтропии для описания соответствующей оценки в методе случайного блуждания обеспечивает оптимизацию разбиения сети при случайном движении по вершинам в каждом из сообществ. Показатель качества полученного разбиения, выраженный через энтропию, может быть легко подсчитан для любого разбиения, обновление и пересчёт этого показателя является быстрой операцией.

Данные методы развиваются в работе [12], посвящённой модификации описанного выше алгоритма с целью нахождения пересекающихся сообществ. На первом шаге выполняется поиск непересекающихся сообществ с помощью динамического алгоритма Росвалля—Бергстрома. Авторы переопределяют функцию качества разбиения с учётом того, что вершина может входить сразу в несколько сообществ. На втором шаге формируется список всех граничных вершин

для всех найденных на первом шаге сообществ. Далее каждая такая вершина примыкает к новому сообществу и выполняется пересчёт нового показателя качества разбиения, в результате выбираются вершины с максимальным уменьшением показателя качества. Процедура итеративно продолжается до тех пор, пока выполняется оптимизация функции.

В [28] описывается метод улучшения алгоритма нахождения пересекающихся сообществ с помощью добавления «памяти» в моделирование процесса случайного блуждания по графу. Если раньше при кодировании процесса учитывалось только предыдущее состояние, то теперь учитывается заданное количество последних переходов. Отметим предложенный в [10] метод, способный выделять пересекающиеся сообщества в многослойных сетях. В основе данного метода также лежит оценка энтропии системы. Многослойные сети можно представлять как одну сеть со связями разных типов. Данную методику можно применять в анализе социальных сетей, так как в них имеются связи разных типов: отношение дружбы, публикация постов, проставление «лайков» и т. д.

5. О тестировании

Самый известный тест на выделение сообществ использует определённый класс графов, предложенный Гирваном и Ньюманом (тест GN) [18]. Каждый граф имеет 128 вершин, которые разбиваются на 4 сообщества по 32 вершины в каждом. Средняя степень вершины составляет 16. Вершины имеют примерно одинаковую степень, как в случайном графе. Очевидны недостатки таких тестов: все вершины графа имеют примерно одинаковую степень, все сообщества одинакового размера, у тестовых графов очень небольшие размеры.

Указанные недостатки частично исправляются при тестировании на часто применяемых LFR-моделях [15, 17] генерации случайных графов, обладающих структурой сообществ. Данный подход является наиболее подходящим для тестирования алгоритмов автоматического выделения сообществ, направленных на работу с большими графами. Для оценки и тестирования алгоритмов выделения сообществ необходимы графы с заданной структурой сообществ, которую алгоритмы будут выделять. С помощью LFR-модели можно протестировать алгоритм как на разных конфигурациях сетей, так и на различных распределениях количества и размеров сообществ в них.

В [16] приведены результаты для большинства представленных выше алгоритмов на тестах Гирвана—Ньюмана, LFR, а также на случайных графах. Кроме того в проведённых тестах учитывались такие параметры, как направление связи, веса связей и возможность сообществ пересекаться. Для остальных алгоритмов авторы в своих работах приводят результаты на нескольких графах и развёрнутый анализ полученных результатов. С учётом всех имеющихся тестов среди рассматриваемых алгоритмов наилучшим образом себя показали

алгоритм Блонделя и динамический алгоритм Росвалля—Бергстрома. Эти алгоритмы удовлетворяют требованиям к скорости работы, требуемой памяти и к качеству полученного разбиения, поэтому их часто применяют при работе с большими графами реальных социальных сетей. У остальных алгоритмов наблюдались различные недостатки, например неспособность быстрой работы на больших объёмах данных и нестабильность качества получаемых разбиений при тестировании.

В целом методы, основанные на оценке модулярности, имеют довольно небольшую точность. Исключение составляет алгоритм Блонделя с достаточно хорошими результатами. Для большинства представленных алгоритмов на точность выделения сообществ негативно влияют скорее размеры сообществ (в случае с большими сообществами это видно хуже, тогда как для небольших размеров сообществ это видно отчётливо), нежели размеры графа. Динамический алгоритм Росвалля—Бергстрома показал наилучшую точность как относительно изменения размеров графа, так и относительно изменения размеров сообществ.

Алгоритмы Блонделя и Росвалля—Бергстрома показали наилучшие результаты на LFR-тесте для неориентированных невзвешенных графов. Они продемонстрировали высокую скорость работы (линейную относительно размеров сети) и лучшее качество выделения сообществ на больших графах.

Упомянутое выше тестирование алгоритмов выделения сообществ осуществляется на двух типах тестов: на тесте Гирвана—Ньюмана и на LFR-тестах. Отметим, что все эти тесты искусственно генерируются и скорее всего не моделируют реальных социальных сетей, для которых применение данных алгоритмов наиболее интересно. Реальные графы характеризуются неравномерным распределением степеней вершин и неравномерным распределением самих сообществ, причём законы таких распределений могут быть различными в разных сетях. Поэтому остаётся актуальной задача сравнительного тестирования алгоритмов на графах реальных социальных сетей.

6. Заключение

Наши оценки позволяют сделать вывод о том, что наиболее эффективным и перспективным для реализаций является динамический алгоритм Росвалля—Бергстрома. Развитие данного подхода [10, 12, 28] позволяет учитывать возможные значения атрибутов вершин и рёбер графа, что даёт возможность выявить сообщества, основанные не только на формальном взаимодействии объектов, но и на иных информационных характеристиках социальных сетей. Поэтому его целесообразно применять для решения задач выделения сообществ в графах больших размеров, характерных для массовых социальных сетей и сетей телекоммуникационного взаимодействия. Данные алгоритмы могут оказаться интересны и для задач, возникающих в биологии, экономике, социологии и маркетинге.

Литература

[1] Базенков Н. И., Губанов Д. А. Обзор информационных систем анализа социальных сетей // УБС. — 2013. — Вып. 41. — С. 357—394.

[2] Батура Т. В. Методы анализа компьютерных социальных сетей // Вестн. НГУ, Сер. Информационные технологии. — 2012. — Т. 10, вып. 4. — С. 13—28.

[3] Губанов Д. А., Новиков Д. А., Чхартишвили А. Г. Социальные сети: модели информационного влияния, управления и противоборства. — М.: Физматлит; МЦНМО, 2010.

[4] Коломейченко М. И., Чеповский А. А., Чеповский А. М. Алгоритм выделения сообществ в социальных сетях // Фундамент. и прикл. матем. — 2014. — Т. 19, вып. 1. — С. 21—32.

[5] Коломейченко М. И., Чеповский А. М. Визуализация и анализ графов больших размеров // Бизнес-информатика. — 2014. — № 4 (30). — С. 7—16.

[6] Aggarwal C. Social Network Data Analytics. — New York: Springer, 2011.

[7] Blondel V. D., Guillaume J.-L., Lambiotte R., Lefebvre E. The Louvain method for community detection in large networks // J. Statist. Mech. Theory Experiment. — 2008. — P. 108—121.

[8] Borgatti P., Everett G., Johnson C. Analyzing Social Networks. — SAGE Publ., 2013.

[9] Clauset A., Newman M. E., Moore C. Finding community structure in very large networks // Phys. Rev. — 2004. — Vol. E 70, no. 6. — 066111.

[10] Domenico M., Lancichinetti A., Arenas A., Rosvall M. Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems // Phys. Rev. — 2015. — Vol. X 5. — 011027.

[11] Donetti L., Muñoz M. A. Improved spectral algorithm for the detection of network communities. — 2005. — arXiv:physics/0504059.

[12] Esquivel A., Rosvall M. Compression of flow can reveal overlapping modular organization in networks // Phys. Rev. — 2011. — Vol. X 1. — 021025.

[13] Fortunato S. Community detection in graphs // Phys. Rep. — 2010. — Vol. 486. — P. 75—174.

[14] Girvan M., Newman M. E. Community structure in social and biological networks // Proc. Natl. Acad. Sci. USA. — 2002. — Vol. 99. — P. 7821—7826.

[15] Lambiotte R., Rosvall M. Ranking and clustering of nodes in networks with smart teleportation // Phys. Rev. — 2012. — Vol. E 85, no. 5. — 056107.

[16] Lancichinetti A., Fortunato S. Community detection algorithms: a comparative analysis // Phys. Rev. — 2009. — Vol. E 80, no. 5. — 056117.

[17] Lovasz L. Random walks on graphs: a survey // Combinatorics, Paul Erdős is Eighty / D. Miklós, V. T. Sós, T. Szőnyi, eds. — (Bolyai Soc. Math. Stud.; Vol. 2). — Budapest, 1996. — P. 1—46.

[18] Newman M. E. J. The structure and function of complex networks // SIAM Rev. — 2003. — Vol. 45, no. 10. — P. 167—256.

[19] Newman M. E. Fast algorithm for detecting community structure in networks // Phys. Rev. — 2004. — Vol. E 69. — 066133.

[20] Newman M. E. Modularity and community structure in networks // Proc. Natl. Acad. Sci. USA. — 2006. — Vol. 103. — P. 8577—8582.

[21] Newman M. E. Networks: An Introduction. — Oxford: Oxford Univ. Press, 2010.

[22] Newman M. E., Girvan M. Finding and evaluating community structure in networks // Phys. Rev. — 2004. — Vol. E 69. — 026113.

[23] Palla G., Derenyi I., Farkas I., Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society // Nature. — 2005. — Vol. 435. — P. 814—818.

[24] Radicchi F., Castellano C., Cecconi F., Loreto V., Parisi D. Defining and identifying communities in networks // Proc. Natl. Acad. Sci. USA. — 2004. — Vol. 101. — P. 2658—2663.

[25] Rosvall M., Axelsson D., Bergstrom C. T. The map equation // Eur. Phys. J. Special Topics. — 2009. — Vol. 178, no. 1. — P. 13—23.

[26] Rosvall M., Bergstrom C. T. An information-theoretic framework for resolving community structure in complex networks // Proc. Natl. Acad. Sci. USA. — 2007. — Vol. 104, no. 18. — P. 7327—7331.

[27] Rosvall M., Bergstrom C. T. Maps of information flow reveal community structure in complex networks // Proc. Natl. Acad. Sci. USA. — 2008. — Vol. 105, no. 4. — P. 1118—1123.

[28] Rosvall M., Esquivel A., Lancichinetti A., West J., Lambiotte R. Memory in network flows and its effects on spreading dynamics and community // Nature Commun. — 2014. — Vol. 5. — 4630.