САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Математико-механический факультет

Кафедра Информатики

Проданов Тимофей Петрович

Адаптивный рандомизированный алгоритм выделения сообществ в графах

Бакалаврская работа

Допущена к защите.
Зав. кафедрой:

Научный руководитель:
д. ф.-м. н., профессор О.Н. Граничин

Рецензент:
В.А. Ерофеева

Санкт-Петербург
2015

SAINT-PETERSBURG STATE UNIVERSITY
Mathematics & Mechanics Faculty

Department of Computer Science

Timofey Prodanov

Adaptive randomised algorithm for community detection in graphs

Bachelor's Thesis

Admitted for defence.
Head of the chair:

Scientific supervisor:
Professor Oleg Granichin

Reviewer:
Victoria Erofeeva

Saint-Petersburg
2015

Оглавление

Введение

1. Предварительные сведения

1.1. Выделение сообществ в графах

Исторически, изучение сетей происходило в рамках теории графов, которая начала своё существование с решения Леонардом Эйлером задачи о кёнигсбергских мостах. В 1920-х взял своё начало анализ социальных сетей и лишь последние двадцать лет развивается изучение *сложных сетей*, то есть сетей с неправильной, сложной структурой, в некоторых случаях рассматривают динамически меняющейся во времени сложные сети. От изучения маленьких сетей внимание переходит к сетям из тысяч или миллионов узлов.

В процессе изучения сложных систем, построенных по реальным системам, оказалось, что распределение степеней $P(s)$, определённое как доля узлов со степенью s среди всех узлов графа, сильно отличается от распределения Пуассона, которое ожидается для случайных графов. Также сети, построенные по реальным системам характеризуются короткими путями между любыми двумя узлами и большим количеством маленьких циклов[1]. Это показывает, что модели, предложенные теорией графов, часто будут оказываться далеко от реальных потребностей.

Современное изучение сложных сетей привнесло значительный вклад в понимание реальных систем. Сложные сети с успехом были применены в таких разных областях, как изучение структуры и топологии интернета [2, 3], эпидемиологии [4], биоинформатике [5], поиске преступников [6], социологии [7] и многих других.

Свойством, присутствующим почти у любой сети, является структура сообществ, разделение узлов сети на разные группы узлов так, чтобы внутри каждой группы соединений между узлами много, а соединений между узлами разных групп мало. Способность находить и анализировать подобные группы предоставляет большие возможности в изучении реальных систем, представленных с помощью сложных сетей. Плотно связанные группы узлов в социальных сетях представляют людей, принадлежащих социальным сообществам, плотно сплочённые группы узлов в интернете соответствуют страницам, посвящённым распространённым темам, а сообщества в генетических сетях связаны с функциональными модулями [1]. Таким образом, выделение сообществ в сети является мощным инструментом для понимания функциональности сети.

1.2. Определения и обозначения

Формально, сложная система может быть представлена с помощью графа. В этой работе будут рассматриваться только невзвешенные неориентированные графы. Невзвешенный неориентированный граф $G = (\mathscr{N}, \mathscr{L})$ состоит из двух множеств — множества $\mathscr{N} \neq \emptyset$, элементы которого называются *узлами* или *вершинами* графа, и множества $\mathscr{L}$ неупорядоченных пар из множества $\mathscr{N}$, элементы которого называются *рёбрами* или *связями*. Мощности множеств $\mathscr{N}$ и $\mathscr{L}$ равны N и K соответственно.

Подграфом называется граф $G' = (\mathscr{N}', \mathscr{L}')$, где $\mathscr{N}' \subset \mathscr{N}$ и $\mathscr{L}' \subset \mathscr{L}$.

Узел обычно обозначают по его порядковому месту i в множестве $\mathscr{N}$, а ребро, соединяющее пару узлов i и j обозначается l_{ij}. Узлы, между которыми есть ребро называются *смежными*. Граф часто представляют в матричном виде, задавая для него матрицу смежности A размера $N \times N$, в которой элемент a_{ij} равен единице, если ребро l_{ij} существует, и 0, если не существует. В таком случае степенью узла называют величину $s_i = \sum_{j \in \mathscr{N}} a_{ij}$.

Прогулка из узла i в узел j — это последовательность узлов, начинающаяся с узла i и заканчивающаяся узлом j. Путь — это прогулка, в которой каждый узел встречается единожды. Геодезический путь — это кратчайший путь, а количество узлов в нём на один больше геодезического расстояния.

Сообщество — это подграф, чьи узлы плотно связаны, однако структурная сплочённость узлов определялась по разному. Одно из определений вводит понятие *клик*. Клик — это максимальный такой подграф, состоящий из трёх и более вершин, каждая из которых связана с каждой другой вершиной из клика. n-клик — это максимальный подграф, в котором самое большое геодезическое расстояние между любыми двумя вершинами не превосходит n. Другое определение гласит, что подграф G' является сообществом, если сумма всех степеней внутри G' больше суммы всех степеней, направленных в остальную часть графа [8].

1.3. Модулярность

Однако подобными определениями пользоваться неудобно и их проверка достаточно долгая. В 2004 году была представлена *модулярность* — целевая функция, оценивающая неслучайность разбиения графа на сообщества [9]. Допустим, у нас κ сообществ, определим тогда симметричную матрицу $\mathbf{e}$ размером $\kappa \times \kappa$. Пусть e_{ij} — отношение количества рёбер, которые идут из сообщества i в сообщество j, к полному количеству рёбер в графе (рёбра l_{mn} и l_{nm} считаются различными, m, n — узлы), $a_i = \sum_j e_{ij}$. След такой матрицы $\mathrm{Tr}\mathbf{e} = \sum_i e_{ii}$ показывает отношение рёбер в сети, которые соединяют узлы одного и того же сообщества, и хорошее разбиение на сообщества должно иметь высокое значение следа. Однако если поместить все вершины в одно сообщество — след примет максимальное возможное значение, притом, что такое разбиение не будет сообщать ничего полезного о графе.

Поэтому далее определяется строка $a_i = \sum_j e_{ij}$, которая обозначает долю количества рёбер, идущих к узлам, принадлежащим сообществу i, к полному количеству рёбер в графе. Если в графе рёбра проходят между вершинами независимо от сообществ — e_{ij} будет в среднем равно $a_i a_j$, поэтому модулярность можно определить следующим образом:

$$Q = \sum_i \left(e_{ii} - a_i^2\right) = \mathrm{Tr}\,\mathbf{e} - \|\mathbf{e}^2\|,$$

где $\|\mathbf{x}\|$ является суммой элементов матрицы $\mathbf{x}$. Если количество рёбер внутри сообществ не будет отличаться от случайного взятого количества — модулярность будет примерно равна 0. Максимальным возможным значением функции будет 1, но на практике модулярности графов лежат между 0.3 и 0.7.

Список литературы

[1] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4):175–308, 2006.

[2] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM Computer Communication Review*, volume 29, pages 251–262. ACM, 1999.

[3] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer networks*, 33(1):309–320, 2000.

[4] Cristopher Moore and Mark EJ Newman. Epidemics and percolation in small-world networks. *Physical Review E*, 61(5):5678, 2000.

[5] Jing Zhao, Hong Yu, Jianhua Luo, ZW Cao, and Yixue Li. Complex networks theory for analyzing metabolic networks. *Chinese Science Bulletin*, 51(13):1529–1537, 2006.

[6] Wang Hong, Wang Zhao-wen, Li Jian-bo, and Qiu-hong Wei. Criminal behavior analysis based on complex networks theory. In *IT in Medicine & Education, 2009. ITIME'09. IEEE International Symposium on*, volume 1, pages 951–955. IEEE, 2009.

[7] John Scott. *Social network analysis*. Sage, 2012.

[8] Stanley Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[9] Mark E. J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, Feb 2004.