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

0.1. Итеративная схема

В подразделе **??** описана итеративная схема кластеризации основных групп графа. Адаптивную версию алгоритма также можно итерировать, в таком случае после адаптивного создания промежуточного разбиения на его основе запускается новое адаптивное создание промежуточного множества. Это продолжается до тех пор, пока новое промежуточное разбиение не будет не лучше предыдущего, после чего финальный алгоритм выделяет сообщества на основе предыдущего промежуточного разбиения. Такая схема далее называется *итеративная схема кластеризации основных групп графа* (*ACGGCi*).

Такой подход не сильно повышает время работы, так как с каждой итерации количество узлов (сообществ) для разбиения уменьшается. Количество узлов или сообществ не может увеличиться или остаться прежним, так как разбиение P_1 на основе разбиения P_2 имеет не больше сообществ, чем разбиение P_2, и в случае, если P_1 и P_2 имеют одинаковое количество сообществ — они равны, то есть и модулярности их равны.

Таблица 1: Модулярность адаптивной схемы кластеризации основных групп графа и её итеративной версии

	ACGGC	ACGGCi
jazz	0.444739	0.444871
celegans_metabolic	0.439724	0.446973
netscience	0.907922	0.909400
as-22july06	0.671205	0.674992
cond-mat-2003	0.743594	0.746731

Таблица 2: Время работы адаптивной схемы кластеризации основных групп графа и её итеративной версии

	ACGGC	ACGGCi
jazz	23.68	31.51
celegans_metabolic	23.92	77.25
netscience	86.38	96.55
as-22july06	2,329	5,801
cond-mat-2003	9,371	11,654

Как видно из таблиц 1 и 2, модулярность итеративной схемы каждый раз немного выше, хотя и время работы каждый раз заметно выростало, на некоторых графах в полтора раза, а на других — более, чем в три.

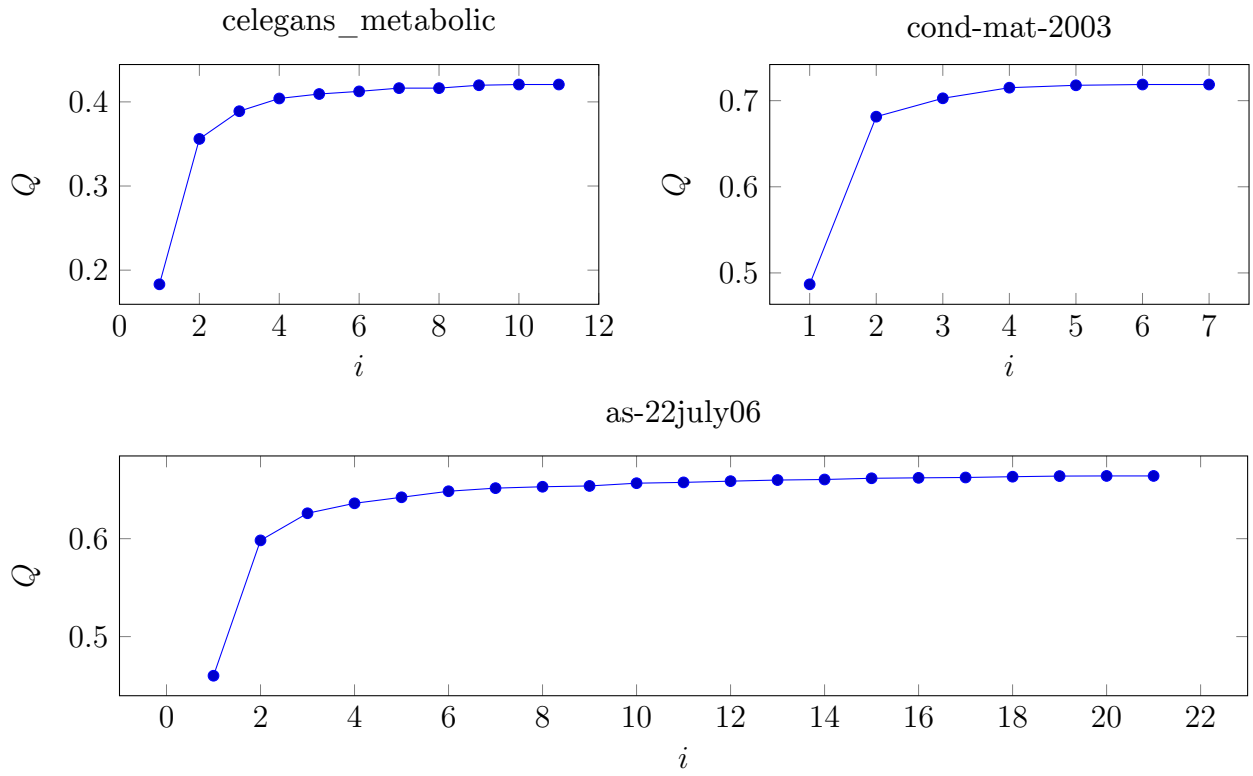

Рис. 1: Модулярности промежуточных разбиений в работе $ACGGCi$ на трёх графах, i — номер итерации

Как можно увидеть на рисунке 1, первое промежуточное разбиение имеет очень небольшую модулярность, а начиная с некоторой итерации модулярность следующего промежуточного разбиения увеличивается очень несильно. Стоит так же заметить, что при $l = 1$, к примеру, модулярность первого промежуточного разбиения была бы выше, так как было бы меньше разбиений, которые участвовали в его создании. Но это бы не принесло глобального выигрыша, так как такое на таком разбиении сложно выделять новые множества. Поэтому небольшая модулярность первого промежуточного разбиения не означает, что оно неудобно для последующей обработки.

Для создания таблиц 1 и 2 и рисунка 1 использовались следующие параметры $ACGGC$ и $ACGGCi$: $d = 2,\ f(Q,k) = -1000 \ln Q,\ \hat{k}_0 = 5,\ l = 6,\ r = 0.05,\ k_{max} = 50$. В качестве финального алгоритма использовался RG_{10}.

0.2. CGGC и ACGGC в качестве финального алгоритма

Одно из преимуществ схемы кластеризации основных групп графа заключается в том, что граф с большим количеством вершин преобразуется в небольшой граф, на котором можно выделять сообщества более точными, но и более долгими алгоритмами, используемыми в качестве финальных алгоритмов. Ранее в работе в качестве финальных алгоритмом рассматривались только RG. Однако существует вариант использования $ACGGC$ в качестве финального алгоритма $CGGC$ или $ACGGC$ с другими параметрами, ровно как и наоборот, использовать $CGGC$ с $ACGGC$ в качестве

финального алгоритма. Это даёт наибольший выигрыш в итеративной схеме кластеризации основных групп графа и её адаптивного аналога, в таком случае одна из версий алгоритма выделяет сообщества в промежуточных разбиениях, пока это возможно. В некоторый момент создание новых промежуточных множеств оказывается неэффективным, однако новые параметры другой версии схемы могут всё ещё быть эффективны.

Так, на графе *celegans_metabolic* $ACGGCi$ с RG_{10} в качестве финального алгоритма имела модулярность 0.446973 как медиану по 3000 запусков, и работала со средним временем 77.25 миллисекунды. С тем же количеством запусков $CGGCi$ с тем же финальным алгоритмом имела модулярность 0.445008 со временем 55.29 миллисекунд. Однако использование $ACGGCi$ с $CGGCi$ в качестве финального алгоритма дало модулярность 0.447324 со временем 89.96 миллисекунд (в качестве финального алгоритма $CGGCi$ при этом использовалось RG_{10}. Таким образом, прирост времени по отношению к использованию в качестве финального алгоритма RG_{10} оказался не очень большим, однако такой подход дал ненулевое увеличение модулярности. Использование же $CGGCi$ стратегии с $ACGGCi$ в качестве финального алгоритма не дало результата на этом графе: модулярность 0.44566 при времени 112.49 миллисекунд.

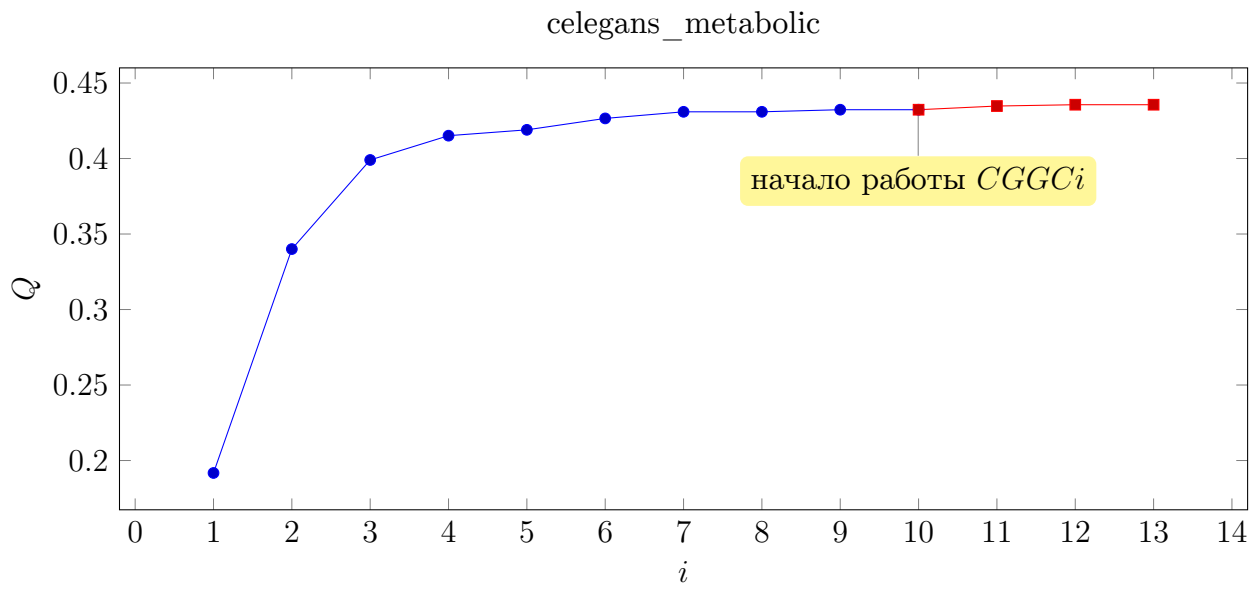

Рис. 2: Модулярности промежуточных разбиений в работе $ACGGCi$ с $CGGCi$ в качестве финального алгоритма на графе *celegans_metabolic*, i — номер итерации, синяя линия — промежутночные разбиения адаптивной версии, красная обозначает промежуточные разбиения неадаптивной версии

В измерениях и на рисунке 2 использовались $ACGGCi$ с параметрами $d = 2,\ f(Q,k) = -1000 \ln Q,\ \hat{k}_0 = 5,\ l = 6,\ r = 0.05,\ k_{max} = 50$ и $CGGCi$ с начальным алгоритмом RG_{10} и параметром $s = 16$.

0.3. Сравнение

Таблица 3: Модулярность разбиений, полученных $ACGGC$ и $CGGC$ на тестовых графах

	$ACGGC^{I}$	$ACGGC^{II}$	$CGGC_{10}^{10}$	$CGGC_{3}^{10}$	$CGGC_{10}^{3}$
karate	0.417242	0.417406	0.415598	0.396532	0.405243
dolphins	0.524109	0.523338	0.521399	0.523338	0.522428
chesapeake	0.262439	0.262439	0.262439	0.262439	0.262370
adjnoun	0.299704	0.299197	0.295015	0.292703	0.290638
polbooks	0.527237	0.527237	0.527237	0.526938	0.526784
football	0.603324	0.604266	0.604266	0.599537	0.599026
celegans_metabolic	0.439604	0.438584	0.435819	0.436066	0.432261
jazz	0.444739	0.444848	0.444871	0.444206	0.444206
netscience	0.907229	0.835267	0.724015	0.708812	0.331957
email	0.573333	0.573409	0.571018	0.572667	0.567423
polblogs	0.424107	0.423208	0.422901	0.421361	0.390395
pgpGiantCompo	0.883115	0.883085	0.882237	0.882532	0.880340
as-22july06	0.671249	0.670677	0.666766	0.669847	0.665260
cond-mat-2003	0.744533	0.750367	0.751109	0.708775	0.413719
caidaRouterLevel	0.846312	0.855651	0.851622	0.858955	0.843835
cnr-2000	0.912762	0.912783	0.912500	0.912777	0.912496
eu-2005	0.938292	0.936984	0.935510	0.936515	0.936420
in-2004	0.979844	0.979771	0.979883		

Таблица 4: Время работы $ACGGC$ и $CGGC$ на тестовых графах

	$ACGGC^{I}$	$ACGGC^{II}$	$CGGC_{10}^{10}$	$CGGC_{3}^{10}$	$CGGC_{10}^{3}$
karate	1.045	1.358	1.381	1.159	1.383
dolphins	2.453	2.982	3.012	2.471	2.998
chesapeake	1.792	2.344	2.244	1.914	2.214
adjnoun	6.193	7.627	7.411	6.342	7.571
polbooks	5.064	6.310	6.136	5.166	6.135
football	7.157	8.441	8.638	7.439	8.547
celegans_metabolic	23.438	29.847	31.842	25.991	31.561
jazz	23.306	27.546	27.580	24.268	27.508
netscience	84.985	61.307	60.831	40.893	59.673
email	71.719	91.856	94.421	80.227	94.101
polblogs	173.03	174.07	177.14	123.35	178.01
pgpGiantCompo	635.50	864.96	1,033.56	832.90	1,030.35
as-22july06	2,330.91	3,152.38	3,714.06	3,071.44	3,716.53
cond-mat-2003	9,165.87	8,172.36	7,485.29	3,443.61	7,422.58
caidaRouterLevel	153,378	154,456	162,787	146,626	162,802
cnr-2000	306,539	413,869	419,965	404,296	420,604
eu-2005	$2.35 \cdot 10^6$	$3.16 \cdot 10^6$	$3.26 \cdot 10^6$	$3.16 \cdot 10^6$	$3.23 \cdot 10^6$
in-2004	$6.59 \cdot 10^6$	$8.32 \cdot 10^6$	$8.91 \cdot 10^6$		

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.

[10] Stefanie Muff, Francesco Rao, and Amedeo Caflisch. Local modularity measure for network clusterizations. *arXiv preprint cond-mat/0503252*, 2005.

[11] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

[12] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On modularity clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 20(2):172–188, 2008.

[13] Mark E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69:066133, Jun 2004.

[14] Michael Ovelgönne and Andreas Geyer-Schulz. A comparison of agglomerative hierarchical algorithms for modularity clustering. In *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*, pages 225–232. Springer, 2012.

[15] Michael Ovelgönne and Andreas Geyer-Schulz. Cluster cores and modularity maximization. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 1204–1213. IEEE, 2010.

[16] Michael Ovelgönne and Andreas Geyer-Schulz. An ensemble learning strategy for graph clustering. *Graph Partitioning and Graph Clustering*, 588:187, 2012.

[17] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[18] Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[19] Julius R Blum. Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, pages 737–744, 1954.

[20] James C Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

[21] James C. Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons, 2005.

[22] Boris T. Polyak. *Introduction to optimization*. Optimization Software New York, 1987.

[23] Oleg Granichin and Natalia Amelina. Simultaneous perturbation stochastic approximation for tracking under unknown but bounded disturbances. *IEEE Transactions on Automatic Control*, 60(5), 2015.

[24] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473, 1977.

[25] David Lusseau, Karsten Schneider, Oliver J Boisseau, Patti Haase, Elisabeth Slooten, and Steve M Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405, 2003.

[26] Daniel Baird and Robert E Ulanowicz. The seasonal dynamics of the Chesapeake Bay ecosystem. *Ecological Monographs*, pages 329–364, 1989.

[27] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

[28] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[29] Jordi Duch and Alex Arenas. Community detection in complex networks using extremal optimization. *Physical review E*, 72(2):027104, 2005.

[30] Pablo M. Gleiser and Leon Danon. Community structure in jazz. *Advances in complex systems*, 6(04):565–573, 2003.

[31] Roger Guimerà Manrique, L Danon, Albert Díaz Guilera, Francesc Giralt, and Àlex Arenas. Self-similar community structure in a network of human interactions. *Physical Review E, 2003, vol. 68, núm. 6, p. 065103-1-065103-4*, 2003.

[32] Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005.

[33] Marián Boguñá, Romualdo Pastor-Satorras, Albert Díaz-Guilera, and Alex Arenas. Models of social networks based on social distance attachment. *Phys. Rev. E*, 70:056122, Nov 2004.

[34] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[35] Paolo Boldi and Sebastiano Vigna. The WebGraph framework I: Compression techniques. In *Proc. of the Thirteenth International World Wide Web Conference (WWW 2004)*, pages 595–601, Manhattan, USA, 2004. ACM Press.

[36] Paolo Boldi, Marco Rosa, Massimo Santini, and Sebastiano Vigna. Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks. In *Proceedings of the 20th international conference on World Wide Web*. ACM Press, 2011.

[37] Paolo Boldi, Bruno Codenotti, Massimo Santini, and Sebastiano Vigna. Ubicrawler: A scalable fully distributed web crawler. *Software: Practice & Experience*, 34(8):711–726, 2004.