分冗余删除,SF 策略需要进一步调整社区结构,而 OTF 策略在搜索过程中过滤掉禁止节点,从而避免了另两种策略在效率上的不足.

图 12 展示了 4 种不同方法在 Football 和 DBLP 数据集上所得社区结果的 *F*1-measure.在每次单项条件社区搜索的具体实验中,要求其搜索条件中的必要节点来自同一实际社区,同时禁止节点来自该社区外.这样就可以根据数据集附带的真实社区结果计算 *F*1-measure.从图 12(a)和图 12(b)可以看出:无论是 Football 还是 DBLP 数据集,3 种不同策略下"社区搜索+过滤"方法得到的 *F*1-measure 都基本相同.这表明给定符合真实社区分布的搜索条件,这 3 种策略都能找到相同准确程度的社区结果.WLP 方法得到的社区搜索结果在多数搜索条件下有更好的准确性,在少数条件下准确性不如其他方法,如图 12(a)的 ii 和图 12(b)的 i,iii 和 iv,但是相对差异并不明显.这是由于 WLP 方法的结果受到禁止节点的影响,从而产生波动.例如:当禁止节点离真实社区较近时,WLP 方法一般会得到比真实社区更小的社区.这是因为在真实社区内部,有部分与禁止节点相连的节点会被阈值过滤排除在外.这实际上是为使社区内成员与必要节点更接近而付出的必要开销.

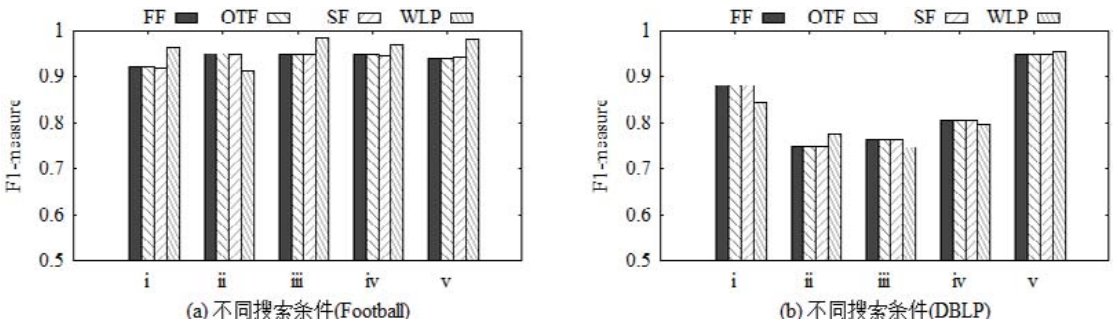

Fig.12 *F*1-measure of different singleconditional community search methods

图 12 不同单项条件社区搜索方法的 *F*1-measure 对比

图 13 对比了通过不同方法所得结果社区的局部模块度 Q_l,其中,图 13(b)的搜索条件中,禁止节点和必要节点来自同一社区,即搜索条件和数据集提供的真实结果有冲突,而图 13(a)则不存在这样的冲突.在有冲突的情形下,容易发现:在"社区搜索+过滤"方法的 3 种不同策略中,SF 所得结果对应的 Q_l 较低,而 FF 和 OTF 的 Q_l 相对较高.这是因为图中显示的 Q_l 为多次实验的均值,在搜索条件存在冲突的情形下,SF 在调整社区结构时可能无法得到社区结果,此时的 Q_l 被记为 0,从而导致 SF 的平均 Q_l 降低.在搜索条件不存在冲突时,这 3 种策略所得社区结果具有相同紧密程度.通过 WLP 方法得到的社区,在紧密程度上由于受禁止节点的影响存在较大波动.此外,图 12 和图 13 的结果也能验证定理 2,即,应用 FF 策略和 OTF 策略的 FindCore 算法得到的结果是相同的.

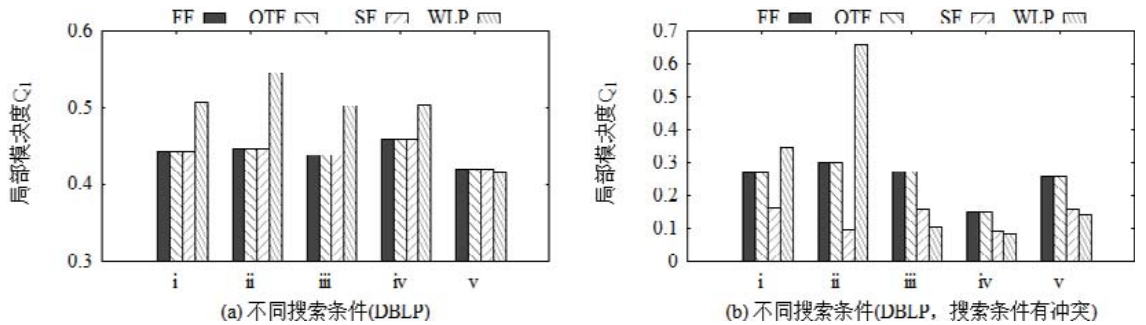

Fig.13 Q_l of singleconditional community search methods

图 13 不同单项条件社区搜索方法的 Q_l 对比

图 14 展示了不同方法所得结果社区的 ASD-ratio.显然,WLP 方法在两个数据集和不同的搜索条件下都具

有最小的 ASD-ratio.这表明 WLP 方法能够充分考虑必要节点和禁止节点的影响,得到社区成员与必要节点更接近的社区.

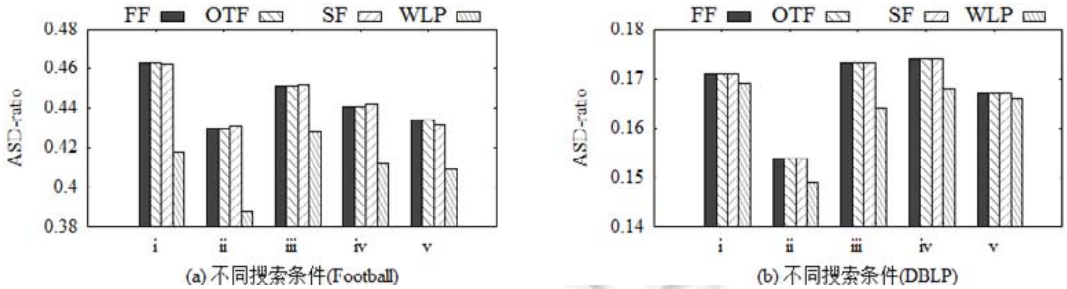

Fig.14 ASD-ratio of different singleconditional community search methods

图 14 不同单项条件社区搜索方法的 ASD-ratio 对比

上述实验结果表明:采用 FF 和 OTF 策略的“社区搜索+过滤”方法可以有效处理 CCS,使用 OTF 策略则可略微提升效率.WLP 方法尽管需要更多的时间开销,所得社区在准确性和紧密程度上存在波动,但是其结果能够体现出社区内成员对于必要节点的倾向性,排除与禁止节点相近的节点,使社区内成员与必要节点更接近,且在 ASD-ratio 上最优,具有较高的应用价值.

5 结束语

条件社区搜索问题是在传统的社区搜索问题基础上,结合实际需求提出的新问题,它包含了现有社区搜索问题,并且考虑了特定节点能否存在于社区中等复杂条件约束.

本文给出了 CCS 的形式化定义,并使用布尔表达式表示搜索条件.在此基础上,本文提出了解决 CCS 的通用框架,将 CCS 分解为多个 SCCS 来处理,并通过简化搜索条件对通用框架进行了优化.对于 SCCS,本文提出了“社区搜索+过滤”的方法(包括 FF,OTF 和 SF 这 3 种策略)和基于标签传播给点加权的 WLP 方法.通过在 4 个真实数据集上的大量实验,比较了这些方法的时间开销和社区结果.实验结果表明,使用 OTF 策略的“社区搜索+过滤”方法在时间开销和社区结果上具有优势.如果考虑用户在使用条件社区搜索时让社区成员远离禁止节点的需求,那么 WLP 方法能够充分考虑必要节点和禁止节点的影响,找到社区内成员与必要节点更接近的社区结果.

条件社区搜索的研究尚处在探索阶段,后续会考虑向 CCS 中引入更多类型的社区定义.对于 SCCS,希望找到一种新的社区搜索算法,使得社区结构在准确性和紧密程度上达到最优的同时也能考虑必要节点和禁止节点的影响.

References:

[1] Fortunato S, Hric D. Community detection in networks: A user guide. Physics Reports, 2016,659:1–44. [doi: 10.1016/j.physrep.2016.09.002]

[2] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. of the National Academy of Sciences, 2002, 99(12):7821–7826. [doi: 10.1073/pnas.122653799]

[3] Dourisboure Y, Geraci F, Pellegrini M. Extraction and classification of dense communities in the Web. In: Proc. of the 16th Int’l Conf. on World Wide Web. ACM Press, 2007. 461–470. [doi: 10.1145/1242572.1242635]

[4] Tang L, Liu H. Community detection and mining in social media. Synthesis Lectures on Data Mining and Knowledge Discovery, 2010,2(1):1–137. [doi: 10.2200/s00298ed1v01y201009dmk003]

[5] Ríos SA, Muñoz R. Dark Web portal overlapping community detection based on topic models. In: Proc. of the ACM SIGKDD Workshop on Intelligence and Security Informatics. ACM Press, 2012. 2. [doi: 10.1145/2331791.2331793]

[6] Chen J, Yuan B. Detecting functional modules in the yeast protein–protein interaction network. Bioinformatics, 2006,22(18): 2283−2290. [doi: 10.1093/bioinformatics/btl370]

[7] Shan J, Shen DR, Kou Y, Nie TZ, Yu G. Approach for hot spread node selection based on overlapping community search. Ruan Jian Xue Bao/Journal of Software, 2017,28(2):326−340 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5117.htm [doi: 10.13328/j.cnki.jos.005117]

[8] Qian J, Wang CK, Guo GY. Community based node betweenness centrality updating algorithms in dynamic networks. Ruan Jian Xue Bao/Journal of Software, 2018,29(3):853−868 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5457.htm [doi: 10.13328/j.cnki.jos.005457]

[9] Cui W, Xiao Y, Wang H, *et al*. Online search of overlapping communities. In: Proc. of the 2013 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2013. 277−288. [doi: 10.1145/2463676.2463722]

[10] Sozio M, Gionis A. The community-search problem and how to plan a successful cocktail party. In: Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2010. 939−948. [doi: 10.1145/1835804.1835923]

[11] Cui W, Xiao Y, Wang H, *et al*. Local search of communities in large graphs. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2014. 991−1002. [doi: 10.1145/2588555.2612179]

[12] Huang X, Cheng H, Qin L, *et al*. Querying *k*-truss community in large and dynamic graphs. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2014. 1311−1322. [doi: 10.1145/2588555.2610495]

[13] Akbas E, Zhao P. Truss-based community search: A truss-equivalence based indexing approach. Proc. of the VLDB Endowment, 2017,10(11):1298−1309. [doi: 10.14778/3137628.3137640]

[14] Huang X, Lakshmanan LVS, Xu J. Community search over big graphs: Models, algorithms, and opportunities. In: Proc. of the 33rd Int'l Conf. on Data Engineering (ICDE). IEEE, 2017. 1451−1454. [doi: 10.1109/icde.2017.211]

[15] Shang J, Wang C, Wang C, *et al*. An attribute-based community search method with graph refining. The Journal of Supercomputing, 2017:1−28. [doi: 10.1007/s11227-017-1976-z]

[16] Fang Y, Cheng R, Luo S, *et al*. Effective community search for large attributed graphs. Proc. of the VLDB Endowment, 2016,9(12): 1233−1244. [doi: 10.14778/2994509.2994538]

[17] Huang X, Lakshmanan LVS. Attribute-driven community search. Proc. of the VLDB Endowment, 2017,10(9):949–960. [doi: 10.14778/3099622.3099626]

[18] Wang M, Wang C, Yu JX, *et al*. Community detection in social networks: An in-depth benchmarking study with a procedure-oriented framework. Proc. of the VLDB Endowment, 2015,8(10):998−1009. [doi: 10.14778/2794367.2794370]

[19] Kernighan BW, Lin S. An efficient heuristic procedure for partitioning graphs. The Bell System Technical Journal, 1970,49(2): 291−307. [doi: 10.1002/j.1538-7305.1970.tb01770.x]

[20] Prat-Pérez A, Dominguez-Sal D, Larriba-Pey JL. High quality, scalable and parallel community detection for large real graphs. In: Proc. of the 23rd Int'l Conf. on World Wide Web. ACM Press, 2014. 225−236. [doi: 10.1145/2566486.2568010]

[21] Zhao F, Tung AKH. Large scale cohesive subgraphs discovery for social network visual analysis. Proc. of the VLDB Endowment, 2012,6(2):85−96. [doi: 10.14778/2535568.2448942]

[22] Newman MEJ. Fast algorithm for detecting community structure in networks. Physical Review E, 2004,69(6):066133. [doi: doi.org/10.1103/physreve.69.066133]

[23] Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Physical Review E, 2004,70(6):066111. [doi: 10.1103/physreve.70.066111]

[24] Radicchi F, Castellano C, Cecconi F, *et al*. Defining and identifying communities in networks. Proc. of the National Academy of Sciences of the United States of America, 2004,101(9):2658−2663. [doi: 10.1073/pnas.0400054101]

[25] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004,69(2):026113. [doi: 10.1103/physreve.69.026113]

[26] Schuetz P, Caflisch A. Multistep greedy algorithm identifies community structure in real-world and computer-generated networks. Physical Review E, 2008,78(2):026112. [doi: 10.1103/physreve.78.026112]

[27] Qiao SJ, Han N, Zhang KF, Zou L, Wang HZ, Gutierrez LA. Algorithm for detecting overlapping communities from complex network big data. Ruan Jian Xue Bao/Journal of Software, 2017,28(3):631−647 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5155.htm [doi: 10.13328/j.cnki.jos.005155]

[28] Donath WE, Hoffman AJ. Lower bounds for the partitioning of graphs. IBM Journal of Research and Development, 1973,17(5): 420−425. [doi: 10.1142/9789812796936_0044]

[29] Shi J, Malik J. Normalized cuts and image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,22(8): 888−905. [doi: 10.1109/34.868688]

[30] Ng AY, Jordan MI, Weiss Y. On spectral clustering: analysis and an algorithm. Proc. of the Nips, 2001,14:849−856.

[31] Mahmood A, Small M, Al-Maadeed SA, *et al*. Using geodesic space density gradients for network community detection. IEEE Trans. on Knowledge and Data Engineering, 2017,29(4):921−935. [doi: 10.1109/tkde.2016.2632716]

[32] Pons P, Latapy M. Computing communities in large networks using random walks. In: Proc. of the Int'l Symp. on Computer and Information Sciences. Berlin, Heidelberg: Springer-Verlag, 2005. 284−293. [doi: 10.1007/11569596_31]

[33] Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proc. of the National Academy of Sciences, 2008,105(4):1118−1123. [doi: 10.1073/pnas.0706851105]

[34] Shang JW, Wang CK, Xin X, Ying X. Community detection algorithm based on deep sparse autoencoder. Ruan Jian Xue Bao/Journal of Software, 2017,28(3):648−662 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5165.htm [doi: 10.13328/j.cnki.jos.005165]

[35] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2014. 701−710. [doi: 10.1145/2623330.2623732]

[36] Cao S, Lu W, Xu Q. Grarep: Learning graph representations with global structural information. In: Proc. of the 24th ACM Int'l Conf. on Information and Knowledge Management. ACM Press, 2015. 891−900. [doi: 10.1145/2806416.2806512]

[37] Raghavan UN, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, 2007,76(3):036106. [doi: 10.1103/physreve.76.036106]

[38] Leung IXY, Hui P, Lio P, *et al*. Towards real-time community detection in large networks. Physical Review E, 2009,79(6):066107. [doi: 10.1103/physreve.79.066107]

[39] Xie J, Szymanski BK, Liu X. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In: Proc. of the 11th ICDMW. IEEE, 2011. 344−349. [doi: 10.1109/icdmw.2011.154]

[40] Davis M, Logemann G, Loveland D. A machine program for theorem-proving. Communications of the ACM, 1962,5(7):394−397. [doi: 10.1145/368273.368557]

[41] Silva JPM, Sakallah KA. GRASP—A new search algorithm for satisfiability. In: Proc. of the Best of ICCAD. Boston: Springer-Verlag, 2003. 73−89. [doi: 10.1007/978-1-4615-0292-0_7]

[42] Zuo XL, Li WJ, Liu YC. Discrete Mathematics. Shanghai: Shanghai Scientific and Technological Literature Press, 1982. 29−39 (in Chinese).

[43] Quine WV. The problem of simplifying truth functions. The American Mathematical Monthly, 1952,59(8):521−531. [doi: /10.2307/2308219]

[44] Khorasgani RR, Chen J, Zaiane OR. Top leaders community detection approach in information networks. In: Proc. of the 4th SNA-KDD Workshop on Social Network Mining and Analysis. 2010.

[45] Yang J, Leskovec J. Defining and evaluating network communities based on ground-truth. Knowledge and Information Systems, 2015,42(1):181−213. [doi: 10.1007/s10115-013-0693-z]

[46] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Physical Review E: Statistical Nonlinear & Soft Matter Physics, 2008,78(4):046110. [doi: 10.1103/physreve.78.046110]

[47] Wang X, Chen G, Lu H. A very fast algorithm for detecting community structures in complex networks. Physica A: Statistical Mechanics and its Applications, 2007,384(2):667−674. [doi: 10.1016/j.physa.2007.05.013]

附中文参考文献:

[7] 单菁,申德荣,寇月,等.基于重叠社区搜索的传播热点选择方法.软件学报,2017,28(2):326−340. http://www.jos.org.cn/1000-9825/5117.htm [doi: 10.13328/j.cnki.jos.005117]

[8] 钱珺,王朝坤,郭高扬.基于社区的动态网络节点介数中心度更新算法.软件学报,2018,29(3):853−868. http://www.jos.org.cn/1000-9825/5457.htm [doi: 10.13328/j.cnki.jos.005457]

[27] 乔少杰,韩楠,张凯峰,等.复杂网络大数据中重叠社区检测算法.软件学报,2017,28(3):631−647. http://www.jos.org.cn/1000-9825/5155.htm [doi: 10.13328/j.cnki.jos.005155]

[34] 尚敬文,王朝坤,辛欣,等.基于深度稀疏自动编码器的社区发现算法.软件学报,2017,28(3):648−662. http://www.jos.org.cn/1000-9825/5165.htm [doi: 10.13328/j.cnki.jos.005165]

[42] 左孝凌,李为鑑,刘永才.离散数学.上海:上海科学技术文献出版社,1982.29−39.

竺俊超(1994−),男,浙江上虞人,博士生,主要研究领域为社会网络,社区发现,社区搜索.

王朝坤(1976−),男,博士,副教授,博士生导师,CCF 专业会员,主要研究领域为图数据管理,社会网络,大数据系统.