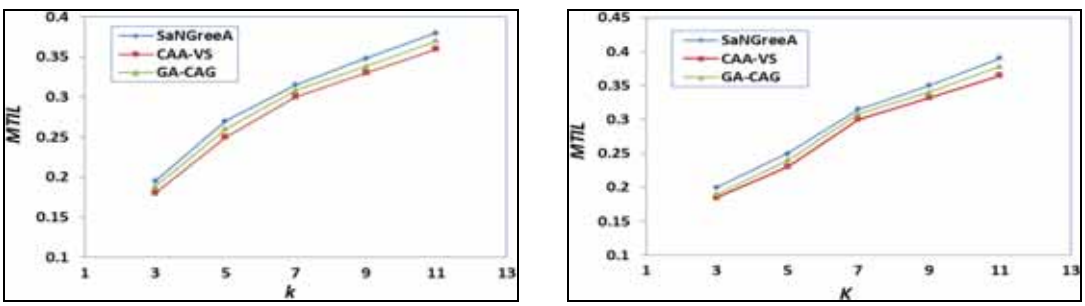
由于 k 值增大带来的聚类变大,即$|V_i|$值变大,使得类中节点属性值更丰富,反映在 $p_{ij}^{\alpha_{jt}}\log p_{ij}^{\alpha_{jt}}$ 值上的各属性值的离散程度也呈变大趋势.故根据聚类熵计算公式,其值随 k 的增大而增大.图 4 也验证了这点.另外,当数据规模变大,不同属性值的个数将更多,则属性值分布的离散程度会变大,也会带来聚类熵增大.与图 4(a)相比,图 4(b)上聚类熵值偏大正证实了这点.图 4 还表明:3 种算法聚类熵值相差较小,在 *DB*1 上,CAA-VS 和 SaNGreeA 分别有最低和最高的熵值;而在 *DB*2 上,GA-CAG 具有更小的熵值.反映当数据量变大时,GA-CAG 强大的搜索优化能力显现一定优势.

5.2 算法性能分析

本节从数据信息损失和算法运行时间两个方面,各通过一组实验讨论算法性能.本文算法构成的信息损失来自超点内子图隐匿带来的结构信息损失,以及超点属性概化带来的属性信息损失,因此,我们以本文第 2 节定义的 MTIL 计算值来代表数据信息损失量.图 5 反映了本文算法分别在两组数据上运行时,MTIL 值随 k 值变化的情况.图中显示,MTIL 值随 k 值正向变化.这是由于 k 值变大,则超点内节点个数增多,在任意属性上的节点属性值将更加丰富,对其概化时,就可能需要取一个更为宽泛的概化属性值.即:各个属性上的概化程度都可能变大,导致平均属性信息损失 NAIL 变大;同时,各超点子图规模变大,子图隐匿时将造成更多的边被隐藏,导致平均结构信息损失 NSIL 值也增大.比较图 5(a)和图 5(b)发现:两者的 MTIL 值非常接近,说明数据规模对 MTIL 值没有多大影响.这是因为 MTIL 表示的是平均信息损失,主要与匿名方法有关;而总体看,图 5(b)的 MTIL 略微偏大.主要是因为 n 变大可能使平均每个聚类中有更大比例的不同属性值被隐匿,以致 NAIL 有细微变大趋势.图 5 也表明,CAA-VS 有更低的信息损失值,反映出本文算法在减小信息损失方面的有效性.

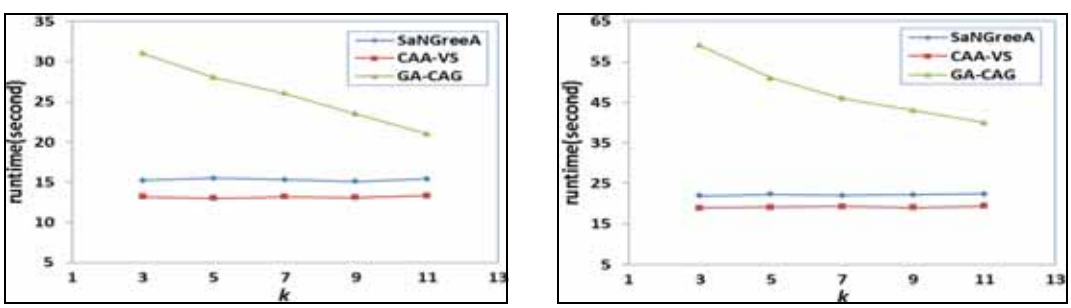


(a) *DB*1 (b) *DB*2

Fig.5 The impacts of changing k on the information loss (MTIL) in the datasets of *DB*1 and *DB*2

图 5 在 *DB*1 和 *DB*2 数据集上的数据信息损失变化情况

为考察本文算法的时间效率,最后一组实验就 3 种算法分别在 *DB*1 和 *DB*2 上的运行时间做出测试分析.图 6 显示了本组实验结果.



(a) *DB*1 (b) *DB*2

Fig.6 The impacts of changing k on the runtime in the datasets of *DB*1 and *DB*2

图 6 在 *DB*1 和 *DB*2 数据集上的运行时间变化情况

图 6 显示:随着 k 值的变大,GA-CAG 的运行时间有明显减小趋势,而 CAA-VS 和 SaNGreeA 的基本没变.这说明对 CAA-VS 和 SaNGreeA 算法来说,k 值的变化基本不会影响运行时间.分析两者的算法过程也可以发现,k 值的变化不会明显影响算法运算量.而对 GA-CAG 算法来说,由于 k 值的增大会减少交叉算子的运算量,故其运行时间有比较明显的减少.图 6 也显示:本文算法用时最少,而 GA-CAG 有最大用时.说明 CAA-VS 一次完成图节点的聚类划分,有着更好的时间效率;而 GA-CAG 利用遗传算法来迭代搜索最优解,需要更多的耗时.比较图 6(a)和图 6(b)可见,各算法在 *DB*2 上的运算时间更高.这是因为 *DB*2 的数据规模更大,算法运行过程需要处理更多的数据,故耗时更多.

6 结束语

在社交网络图中,个体节点或边蕴含的隐私信息可能因遭受到恶意盗取而泄露.考虑同时保护两者信息,防止一切以连接边和属性值为背景知识的攻击,从而保障社交网络图数据发布的隐私安全,本文提出一种基于结构和属性综合相似度的属性图聚类匿名方法.该方法分别定义了节点间的结构相似度和属性相似度,依据两者的综合相似度利用贪心法实现属性图聚类划分,最后对各个聚类进行属性概化和子图隐匿的匿名处理.实验也验证了该方法在实现聚类质量和算法性能方面的有效性.后续工作中,我们将针对目前社交网络中个性化隐私保护不足的问题展开深入研究.

References:

[1] Liu XY, Wang B, Yang XC. Survey on privacy preserving techniques for publishing social network data. Ruan Jian Xue Bao/Journal of Software, 2014,25(3):576−590 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4511.htm [doi: 10.13328/j.cnki.jos.004511]

[2] Sweeney L. *K*-Anonymity: A model for protecting privacy. Int'l Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems, 2002,10(5):557−570. [doi: 10.1142/S0218488502001648]

[3] Gong NZ, Talwalkar A, Mackey L, Huang L, Shin ECR, Stefanov E, Shi E, Song D. Jointly predicting links and inferring attributes using a social-attribute network (SAN). In: Proc. of the SNA-KDD. 2012.

[4] Fu YY, Fu H, Xie X, Sun GZ, Zhang M. Anonymity and privacy preservation for social network. Communications of the CCF, 2014,10(6):51−58 (in Chinese).

[5] Liu P, Li XX. An improved privacy preserving algorithm for publishing social network data. In: Proc. of the IEEE Int'l Conf. on High Performance Computing and Communications & IEEE Int'l Conf. on Embedded and Ubiquitous Computing. 2013. 888−895. [doi: 10.1109/HPCC.and.EUC.2013.127]

[6] Zou L, Chen L, Özsu MT. *K*-Automorphism: A general framework for privacy preserving network publication. Proc. of the VLDB Endowment, 2009,2(1):946−957. [doi: 10.14778/1687627.1687734]

[7] Yuan M, Chen L, Yu PS, Yu T. Protecting sensitive labels in social network data anonymization. IEEE Trans. on Knowledge and Data Engineering, 2013,23(3):633−647. [doi: 10.1109/TKDE.2011.259]

[8] Masoumzadeh A, Joshi J. Preserving structural properties in edge-perturbing anonymization techniques for social networks. IEEE Trans. on Dependable and Secure Computing, 2012,9(6):877−889. [doi: 10.1109/TDSC.2012.65]

[9] Zheleva E, Getoor L. Preserving the privacy of sensitive relationships in graph data. In: Proc. of the Privacy, Security, and Trust in KDD. Berlin, Heidelberg: Springer-Verlag, 2008. 153−171. [doi: 10.1007/978-3-540-78478-4_9]

[10] Campan A, Truta TM. Data and structural *k*-anonymity in social networks. In: Proc. of the Privacy, Security, and Trust in KDD. Berlin, Heidelberg: Springer-Verlag, 2009. 33−54. [doi: 10.1007/978-3-642-01718-6_4]

[11] Tassa T, Cohen DJ. Anonymization of centralized and distributed social networks by sequential clustering. IEEE Trans. on Knowledge and Data Engineering, 2013,25(2):311−324. [doi: 10.1109/TKDE.2011.232]

[12] Skarkala ME, Maragoudakis M, Gritzalis S, Mitrou L, Toivonen H, Moen P. Privacy preservation by *k*-anonymization of weighted social networks. In: Proc. of the IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining. 2012. 423−428. [doi: 10.1109/ASONAM. 2012.75]

[13] Fu YY, Zhang M, Feng DG, Chen KQ. Attribute privacy preservation in social networks based on node anatomy. Ruan Jian Xue Bao/Journal of Software, 2014,25(4):768−780 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4565.htm [doi: 10.13328/j.cnki.jos.004565]

[14] Hay M, Li C, Miklau G, Jensen D. Accurate estimation of the degree distribution of private networks. In: Proc. of the IEEE Int'l Conf. on Data Mining. 2009. 169−178. [doi: 10.1109/ICDM.2009.11]

[15] Sala A, Zhao X, Wilson C, Zheng H, Zhao BY. Sharing graphs using differentially private graph models. In: Proc. of the ACM SIGCOMM Conf. on Internet Measurement Conf. 2011. 81−98. [doi: 10.1145/2068816.2068825]

[16] Jiang HW, Zeng GS, Hu KK. A graph-clustering anonymity method implemented by genetic algorithm for privacy-preserving. Journal of Computer Research and Development, 2016,53(10):2354−2364 (in Chinese with English abstract).

[17] Wang R, Zhang M, Feng D, Fu Y. A clustering approach for privacy-preserving in social networks. In: Proc. of the Information Security and Cryptology (ICISC 2014). Springer Int'l Publishing, 2014. 193−204. [doi: 10.1007/978-3-319-15943-0_12]

[18] Han QL, Zhao HB, Pan HW, Yin GS, Chang JY. Research on spatio-temporal object graph clustering algorithm based on structure and attribute. Journal of Computer Research and Development, 2013,50(Suppl.):154−162 (in Chinese with English abstract).

附中文参考文献:

[1] 刘向宇,王斌,杨晓春.社会网络数据发布隐私保护技术综述.软件学报,2014,25(3):576−590. http://www.jos.org.cn/1000-9825/4511.htm [doi: 10.13328/j.cnki.jos.004511]

[4] 付艳艳,付浩,谢幸,孙广中,张敏.社交网络匿名与隐私保护.中国计算机学会通讯,2014,10(6):51−58.

[13] 付艳艳,张敏,冯登国,陈开渠.基于节点分割的社交网络属性隐私保护.软件学报,2014,25(4):768−780. http://www.jos.org.cn/1000-9825/4565.htm [doi: 10.13328/j.cnki.jos.004565]

[16] 姜火文,曾国荪,胡克坤.一种遗传算法实现的图聚类匿名隐私保护方法.计算机研究与发展,2016,53(10):2354−2364.

[18] 韩启龙,赵洪斌,潘海为,印桂生,常吉羽.基于结构——属性的时空对象图聚类算法的研究.计算机研究与发展,2013,50(增刊):154−162.

姜火文(1974－),男,江西南昌人,博士,副教授,CCF 学生会员,主要研究领域为隐私安全,软件演化,智能计算.

刘文娟(1989－),女,硕士,主要研究领域为系统扩展,大数据处理.

占清华(1979－),男,硕士,讲师,主要研究领域为信息安全,病毒防治和入侵检测.

马海英(1977－),女,博士,副教授,主要研究领域为隐私保护,公钥密码学,网络安全.