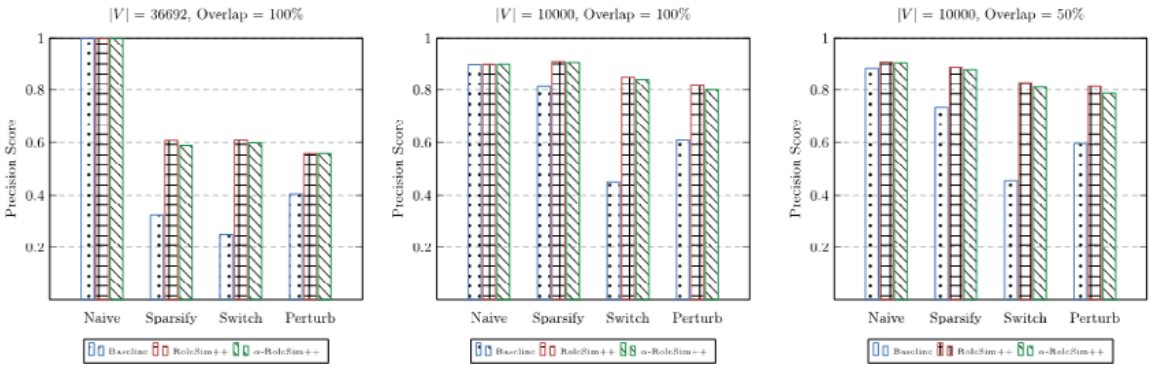
匿名化,统计正确匹配顶点的数量.在 LiveJournal 数据集上,分别按交叠率 50%和 100%实验;在 Enron 数据集上,设定交叠率为 100%,实验的匹配结果如图 4 所示.



(a) Enron 数据集 (b) Livejournal 数据集 (c) Livejournal 数据集

Fig.4 Accuracy over different anonymization algorithms

图 4 不同匿名化算法下的精确度比较

在这两个数据集上的 3 小组实验,结果是一致的.对经过朴素匿名化的目标图,基准算法(Baseline)在交叠率 50%和 100%的图上,效果几乎和本文提出的算法相当.这是因为匿名化算法不改变交叠部分的图结构,给去匿名化降低了难度.当不同的去匿名化算法被应用的时候,本文提出的 RoleMatch 算法精度远远高于基准算法.后者的表现在不同的匿名化方法上差异很大,在交换匿名化上表现最差,当交叠率为 50%时,只去匿名化了 40%的交叠点.RoleMatch(Rolesim++)算法和 RoleMatch(α-Rolesim++)算法在每一种匿名化算法下都保持了较好的健壮性,在 LiveJournal 数据集的实验中,大约能正确匹配 80%以上的交叠点;在 Enron 数据集的结果中,也能正确匹配一半以上.

3.2.3 局部去匿名化

本节实验对比了 3 种算法在非匿名图和匿名图大小不同时的去匿名化的效果.实验中,我们从 LiveJournal 图中爬取一定大小的非匿名图,并应用匿名化算法于实验图中得到匿名图.这样,非匿名图即为匿名图的局部子图.两张图的交叠率从 15%~35%进行匹配实验,多次匹配并求取平均值.实验结果如图 5 所示.

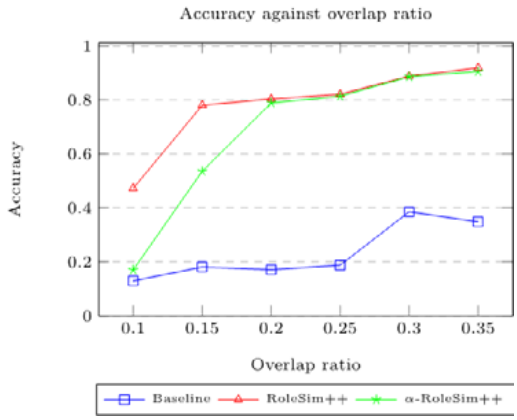


Fig.5 Accuracy against overlap ratio for the local deanonymization problem

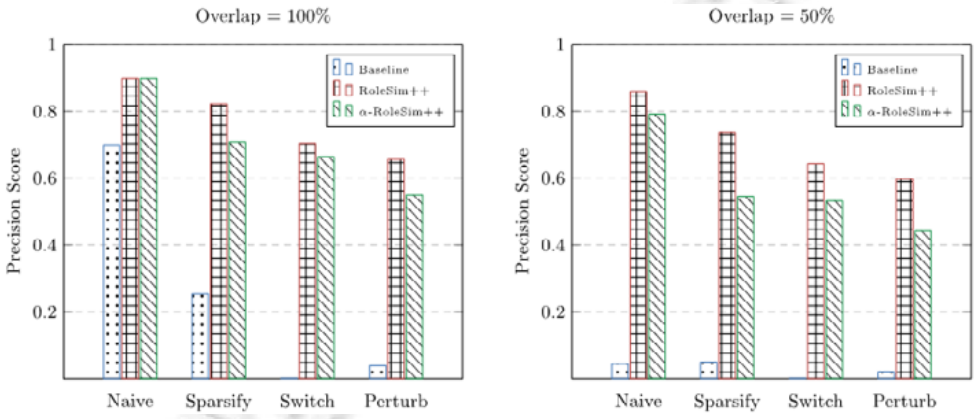
图 5 局部去匿名化中不同交叠率的精确度比较

在图 5 中可以看到:随着交叠率的增加,3 种算法的匹配精度均出现增长.当交叠率低于 0.15 时,3 种算法的去匿名化精度均低于 50%.这是因为当交叠率低于 15%时,匿名图中其他点过多对去匿名化造成影响较大.当交

叠率高于 0.20 后,Rolesim 算法和 Rolesim++算法都取得了 80%以上的去匿名化精度,而 Baseline 算法的去匿名化精度始终低于 40%.

3.2.4 对中间结果的考察

为评估相似度算法的合理性,本节实验对比了 3 种算法的中间结果精度.在实验中,选择第 1 阶段迭代计算之后得到的相似度矩阵,对每个匿名化后的顶点,统计与其相似度最高的顶点恰好是正确匹配的数量.进而又统计了对每个匿名化后的顶点,正确匹配的相似度位于前 1%~10%的比例.比较得到的结果如图 6 和图 7 所示.



(a) Top-1 精度 (b) Top-1 精度

Fig.6 Intermediate results over different anonymization algorithms

图 6 不同匿名化算法下的中间结果比较

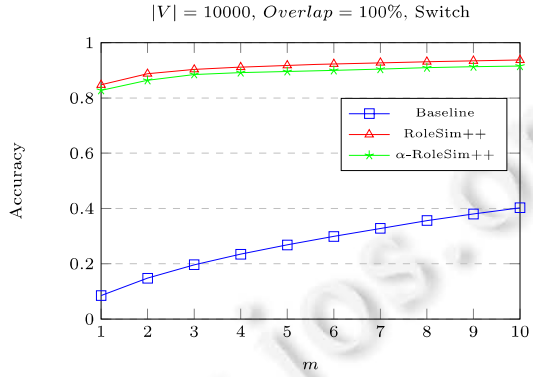


Fig.7 Accuracy of intermediate results (top m%)

图 7 中间结果的精确度(前 m%)

Rolesim++算法和α-Rolesim++算法在不同的匿名化算法下比基准算法有更好的表现,尤其是当交叠率不高的时候.另外,当交叠率逐渐降低时,Rolesim++算法和α-Rolesim++算法之间的差别逐渐明显.与图 4 相对照,更好的相似度度量可以带来更高的去匿名化精度.基准算法在交换匿名化上有最差的精度表现,它的相似度分数也在该匿名化上表现最差(低于 1%).

在这个最相似点对的基础上,可以通过优先匹配相似度高的点对获得更高的精确度,因为它们往往有更大的概率是正确的匹配.这也是为什么很多最相似度不排在第 1 的点对,在结点匹配阶段之后能够被正确匹配,而最后的去匿名化结果比中间结果要好.

3.2.5 执行时间

本节实验比较 Rolesim++算法和α-Rolesim++算法的时间性能.Baseline 算法的时间复杂度和实验中的实际

耗时与 Rolesim++算法并没有显著区别,因此不再与之比较.对于每种去匿名化算法,分别使图中点的数量|*V*|与边的平均密度 *d* 增加.

图 8 展示了算法执行时间的变化.Rolesim++算法的执行时间随着点数与边的密度的增加而显著增加,而 α-Rolesim++算法总是比它更快,而且耗时的增加相对缓慢.由此体现了 α-Rolesim++算法的高效性.

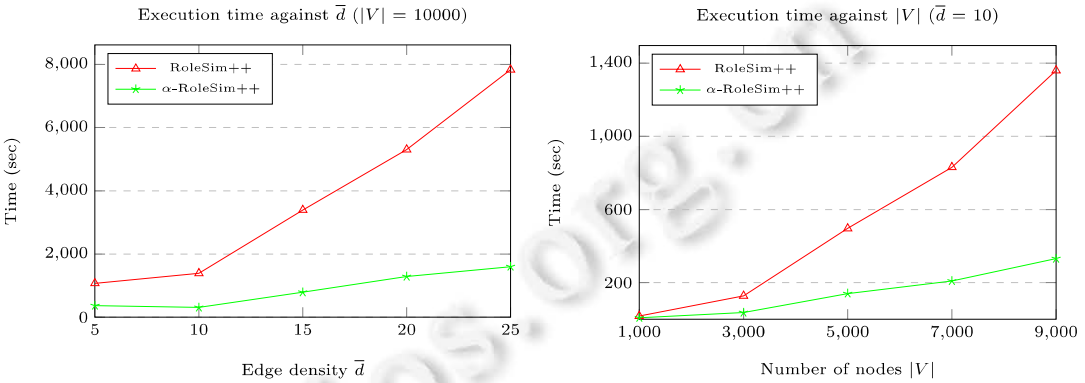


Fig.8 Execution time over |*V*| and *d*

图 8 执行时间与|*V*|和 *d* 的关系

4 结 论

本文提出了一种不需要已知种子的去匿名化算法 RoleMatch,分为结点相似度计算阶段和结点匹配阶段.其中,结点相似度阶段计算的精确而具有容错性的相似度,刻画了不同的图中结点有同一身份的可能性;动态的结点匹配阶段同时考虑了结点的相似度对匹配的影响,以及之前匹配结果的反馈.这一算法仅需要社交网络图的拓扑结构信息,产生非常好的去匿名化结果,并有一个快速计算的版本,能够高效去匿名化.在 LiveJournal 真实数据集上进行的实验,实验方式在传统对称去匿名化的基础上还进行了更贴合应用情景的局部去匿名化实验,各项实验的结果进一步表明了去匿名化的准确性和高效性.接下来,我们将进一步考虑算法的分布式实现,并考虑在富信息图上利用更多的信息进行去匿名化.

References:

[1] Backstrom L, Dwork C, Kleinberg J. Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography. In: Proc. of the 16th Int'l Conf. on World Wide Web. ACM Press, 2007. 181−190. [doi: 10.1145/1242572.1242598]

[2] Wang Y, Zheng B. Preserving privacy in social networks against connection fingerprint attacks. In: Proc. of the 2015 IEEE 31st Int'l Conf. on Data Engineering. IEEE. 2015. 54−65. [doi: 10.1109/ICDE.2015.7113272]

[3] Wu X, Ying X, Liu K, Chen L. A survey of privacy-preservation of graphs and social networks. In: Proc. of the Managing and Mining Graph Data. Springer-Verlag, 2010. 421−453. [doi: 10.1007/978-1-4419-6045-0_14]

[4] Liu K, Terzi E. Towards identity anonymization on graphs. In: Proc. of the 2008 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2008. 93−106. [doi: 10.1145/1376616.1376629]

[5] Zhou B, Pei J. Preserving privacy in social networks against neighborhood attacks. In: Proc. of the 2008 IEEE 24th Int'l Conf. on Data Engineering. IEEE, 2008. 506−515. [doi: 10.1109/ICDE.2008.4497459]

[6] Bonchi F, Gionis A, Tassa T. Identity obfuscation in graphs through the information theoretic lens. In: Proc. of the Information Sciences. Elsevier, 2014. 232−256. [doi: 10.1016/j.ins.2014.02.035]

[7] 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]

[8] Narayanan A, Shmatikov V. De-Anonymizing social networks. In: Proc. of the 2009 30th IEEE Symp. on Security and Privacy. IEEE, 2009. 173−187. [doi: 10.1109/SP.2009.22]

[9] Narayanan A, Shi E, Rubinstein BI. Link prediction by de-anonymization: How we won the kaggle social network challenge. In: Proc. of the 2011 Int'l Joint Conf. on Neural Networks (IJCNN). IEEE, 2011. 1825−1834. [doi: 10.1109/IJCNN.2011.6033446]

[10] Yartseva L, Grossglauser M. On the performance of percolation graph matching. In: Proc. of the 1st ACM Conf. on Online Social Networks. ACM Press, 2013. 119−130. [doi: 10.1145/2512938.2512952]

[11] Kazemi E, Hassani SH, Grossglauser M. Growing a graph matching from a handful of seeds. Proc. of the VLDB Endowment, 2015, 8(10):1010−1021. [doi: 10.14778/2794367.2794371]

[12] Korula N, Lattanzi S. An efficient reconciliation algorithm for social networks. Proc. of the VLDB Endowment, 2014,7(5): 377−388. [doi: 10.14778/2732269.2732274]

[13] Fu H, Zhang A, Xie X. Effective social graph deanonymization based on graph structure and descriptive information. ACM Trans. on Intelligent Systems and Technology (TIST), 2015,6(4):49. [doi: 10.1145/2700836]

[14] Henderson K, Gallagher B, Li L, Akoglu L, Eliassi-Rad T, Tong H, Faloutsos C. It's who you know: Graph mining using recursive structural features. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2011. 663−671. [doi: 10.1145/2020408.2020512]

[15] Jeh G, Widom J. SimRank: A measure of structural-context similarity. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2002. 538−543. [doi: 10.1145/775047.775126]

[16] Blondel VD, Gajardo A, Heymans M, Senellart P, Van Dooren P. A measure of similarity between graph vertices: Applications to synonym extraction and Web searching. Siam Review, 2004,46(4):647−666. [doi: 10.1137/S0036144502415960]

[17] Jin R, Lee VE, Hong H. Axiomatic ranking of network role similarity. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2011. 922−930. [doi: 10.1145/2020408.2020561]

[18] 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]

[19] Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q. Line: Large-Scale information network embedding. In: Proc. of the 24th Int'l Conf. on World Wide Web. ACM Press, 2015. 1067−1077. [doi: 10.1145/2736277.2741093]

刘家霖(1995－),男,福建晋江人,主要研究领域为数据库.

史舒扬(1994－),男,学士,主要研究领域为数据库.

张悦眉(1995－),女,主要研究领域为数据库.

邵蓥侠(1988－),男,博士,主要研究领域为数据库,知识图谱数据管理,并行图计算,知识工程.

崔斌(1975－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为数据库,大数据管理分析.