

复杂网络中基于采样的近似三角计数方法研究

黄取治¹ 张军朝²

(福建师范大学信息技术学院 福州 350007)¹ (太原理工大学计算机科学与技术学院 太原 030024)²

摘 要 复杂网络中的三角计数可以用于分析网络的同质性和传递性。为了提高复杂网络中三角计数的性能,提出了一种基于采样的近似三角计数方法。首先,以一定的采样概率对网络中的边进行采样从而得到一个子网络,并在该子网络中统计三角的个数。其次,依据采样的概率思想,应用于网络中的三角个数估计原网络中的三角个数。最后,对采样方法的均值和方差进行了理论分析,并给出了由采样方法得到的加速比。理论分析与实验表明,与传统的节点迭代方法相比,提出的方法在保证高准确性的前提下大大提高了算法的运行效率,因而更适用于大规模网络中基于三角计数的相关应用。

关键词 复杂网络,采样,三角计数,同质性,近似算法

中图分类号 TP319 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.11.039

Approximating Triangle Counting Based on Sampling in Complex Networks

HUANG Qu-zhi¹ ZHANG Jun-chao²

(School of Information Technology, Fujian Normal University, Fuzhou 350007, China)¹

(College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan 030024, China)²

Abstract Counting of triangles in complex networks is the basis for research on homophily and transitivity. In order to improve the performance of triangles counting, this paper proposed a sampling based approximating triangle counting algorithm. Firstly, we sampled every edge in a graph with constant probability, and counted the number of triangles in the sub-graph. Secondly, based on probability theory of sampling, we approximated the number of triangles in the original graph with the induced sub-graph. Finally, we analyzed the mean and variance of our proposed sampling method, and gave its corresponding speedup. Theoretical analysis and experiments show that, compared with the classic node iteration approach, the proposed algorithm improves the execution time a lot while keeping high accuracy, and thus is more suitable for triangle counting based applications in large scale networks.

Keywords Complex networks, Sampling, Triangle counting, Homophily, Approximation algorithm

复杂网络用节点以及节点之间的连接来描述系统的拓扑结构,是一种关于物理、数学、计算机科学和社会科学等相关学科的交叉学科^[1]。有别于简单或者随机网络,复杂网络通常表现出不同的物理特征,如社团结构^[2]、小世界现象^[3]等。现实社会中的许多网络,如 World Wide Web、在线社交网络、基因网络、蛋白质交互网络等,往往都具有复杂的拓扑结构,都可以看成复杂网络。

在复杂网络分析中,网络中的三角^[4](节点个数为 3 的完全图)是一项重要的研究内容。例如在社会网络中,通过对网络中的三角进行分析可以观察网络的同质性^[5]和传递性^[6]。网络的同质性表明社会网络中的用户往往会选择与这两个用户自身相似的用户作为好友;网络的传递性则表明如果两个用户拥有相同的好友,那么这两个用户也可能是或者会成为好友。此外,三角在图数据的挖掘中也起着非常重要的作用。Becchetti 等人^[7]应用网络中的三角来进行垃圾邮件的检测, Eckmann 等人^[8]应用三角来发现 Web 网络中隐藏的主题结

构, Bar-Yossef 等人^[9]应用三角计数来优化数据库中的查询规划。

复杂网络中的三角计数可以分为准确计数方法和近似计数方法两种。在准确计数方法中,最简单的方法是枚举网络中任意 3 个节点的组合,并观察 $\binom{n}{3}$ 个组合中诱导子图为三角形的个数^[10]。该方法的时间复杂度为 $O(n^3)$, 对于大规模网络是不可行的。节点迭代方法^[11]遍历每个节点,取出该节点的所有邻居对组合,并观察邻居对在网络中是否互为邻居。该方法所需要的时间总和为 $\sum_{v \in V} \binom{d_v}{2}$, 对式中的每一项取上界可得 $O(d_{\max}^2 n)$, 其中 $d_{\max}$ 为图中节点的最大度数。边迭代方法^[12]与节点迭代方法有着相同的时间复杂度,但是通过简单的哈希函数可以将计算复杂度减小至 $O(m^{\frac{3}{2}})$ 。基于快速矩阵相乘的方法,网络中的三角计数所需的时间复杂度为 $O(m^{\frac{2\gamma}{\gamma+1}}) \subset O(m^{1.41})$ ^[13]。随着网络规模的增大,准确的三角

到稿日期:2014-10-05 返修日期:2014-12-20 本文受国家自然科学基金(60443004)资助。

黄取治(1982—),女,硕士,讲师,主要研究方向为云计算技术、网络与数据挖掘;张军朝(1974—),男,博士后,副教授,主要研究方向为算法分析、物联网技术等。

计数方法很难满足用户的实际需求,近似方法在牺牲一定准确性的基础上可以大大减小计算复杂度。近似三角计数方法^[14-17]往往将图数据视为流数据,通过对图数据的一次或若干次扫描提取出重要信息,并根据提取出的信息对三角个数进行估计。这些方法通过随机采样的方式提取出与原图结构相似的子图,并在子图上进行分析。分析结果的准确性依赖于子图与原图之间的比例。

为了提高复杂网络中三角计数的性能,本文基于边采样方法研究了如何准确而高效地进行三角个数的计算。本文以一定的采样概率对网络中的边进行采样,在得到的子网络中进行相应的三角计数,并根据子网络中的三角个数估计原网络中的三角个数。

1 近似三角计数方法

在无向无权图 $G=(V,E)$ 中,节点的个数为 $n=|V|$,边的个数为 $m=|E|$ 。图中的三角是一个包含 3 个节点的完全子图。给定图 $G=(V,E)$,其包含的三角的总个数为

$$\text{tri}(G) = \frac{1}{3} \sum_{v \in V} |\{ \langle u, v, w \rangle \mid (u, v), (v, w), (w, u) \in E \}| \quad (1)$$

1.1 采样方法

当复杂网络的规模很大时,准确地计算出式(1)的结果需要很长的时间。因此,通过概率方法对复杂网络进行采样,对采样的子网络进行分析,然后再推导出原网络的三角个数。已知图 $G=(V,E)$ 后,首先采用采样方法对图的边进行压缩。由于对边的数量进行了压缩,为了保证压缩后的子图与原图有着相似的属性,需要将新图中的边的权重进行扩大。基于采样的三角计数算法的细节如下。

算法 基于采样的三角计数算法

输入:无向无权图 $G=(V,E)$,参数 $p; // p \in [0,1]$

输出:图 $G=(V,E)$ 中包含的三角个数 $\text{tri}(G)$;

1. 令 $\text{tri}(G) \leftarrow 0, E' \leftarrow \{\}$;
2. For each $e \in E$
3. 以概率 p , 令 $E' = E' \cup e$ and $w_e = \frac{1}{p}$;
4. 以概率 $1-p$, 令 $E' = E', w_e = 0$;
5. For each $v \in V'$
6. For each pair $u, w \in \text{neighbors}(v)$
7. If $(u, w) \in E'$
8. $\text{tri}(G) \leftarrow \text{tri}(G) + 1$;
9. $\text{tri}(G) \leftarrow \text{tri}(G) * \frac{1}{6p^3}$;
10. Return $\text{tri}(G)$.

上述算法主要包括两部分。第一部分(第 2-4 行)对图进行采样。在采样过程中,以概率 p 选取一条边,将边的权重设置为 $\frac{1}{p}$ 后加入到新的图中;同时以概率 $1-p$ 舍弃掉该边。经过采样过程后,可以得到 $G'=(V',E')$,其中边的条数为 $m'=m * p$,并且每条边的权重都为 $\frac{1}{p}$ 。第二部分(第 5-8 行)采用节点迭代方法计算新图 G' 中的三角个数。对于每一个节点 v ,遍历其邻居对 (u, w) 组合,如果该邻居对 (u, w) 在新图 G' 中仍互为邻居,即 $(u, w) \in E'$,那么三角的个数加 1。最后,将得到的三角个数乘以一个三角的权重,就可以得到原图中三角的个数。

1.2 均值与方差分析

在采样之后,我们希望得到的新图的三角个数等于原图中的三角个数。设原图 G 中含有 K 个三角,对于每一个三角设置一个指示变量 $\delta_i, 1 \leq i \leq K$ 。 $\delta_i = 1$ 表明原图中的第 i 个三角在新图 G' 中也存在;否则, $\delta_i = 0$ 。用 X 表示 G' 中含有的三角个数的随机变量,那么有如下定理成立。

定理 1 G' 中含有三角个数的期望值等于 G 中实际含有的三角个数,即 $E[X] = K$ 。

证明:随机变量 X 为指示变量 $\delta_i (1 \leq i \leq K)$ 的和再乘以三角的权重 $\frac{1}{p^3}, X = \sum_{i=1}^K \frac{\delta_i}{p^3}$ 。于是有 $E[X] = E[\sum_{i=1}^K \frac{\delta_i}{p^3}] = \sum_{i=1}^K E[\frac{\delta_i}{p^3}] = \frac{1}{p^3} \sum_{i=1}^K E[\delta_i] = \frac{1}{p^3} \sum_{i=1}^K p^3 = K$ 。

定理 2 令 G 含有的总的三角个数为 K, p 为采样比例,那么随机变量 X 的方差为

$$\text{Var}(X) = \frac{K(p^3 - p^6) + 2k(p^5 - p^6)}{p^6} \quad (2)$$

其中, k 为含有共享边的三角对的个数。

证明:在上述采样过程中,指示变量 $\delta_i (1 \leq i \leq K)$ 是同分布的,但是 δ_i 和 $\delta_j (i \neq j)$ 不是相互独立的。如图 1 所示,当 δ_i 和 δ_j 含有共享边时, δ_i 和 δ_j 不是相互独立的。如果该共享边没有出现在 G' 中,那么 $\delta_i = \delta_j = 0$;在其它情况下 δ_i 和 δ_j 是相互独立的。

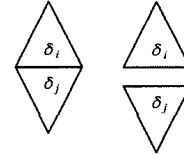


图 1 三角对的独立性示意图

根据随机变量的方差定义及基本性质,可得

$$\text{Var}(X) = \text{Var}(\sum_{i=1}^K \frac{\delta_i}{p^3}) = \frac{1}{p^6} \sum_{i=1}^K \sum_{j=1}^K \text{Cov}(\delta_i, \delta_j) \quad (3)$$

式(3)是含有 K^2 个项的和函数,其中 K 项($i=j$)是指示变量的方差,剩下的 $\binom{K}{2}$ 项是指示变量对。当 $i=j$ 时,可得 $K(p^3 - p^6)$ 。当 $i \neq j$ 时,令 $\binom{K}{2}$ 个指示变量对中有 k 个共享一条边,那么有 $\text{Cov}(\delta_i, \delta_j) = p^5 - p^6$;对于余下的 $\binom{K}{2} - k$ 个指示变量对, $\text{Cov}(\delta_i, \delta_j) = p^6 - p^6 = 0$ 。综上可得, $\text{Var}(X) = \frac{K(p^3 - p^6) + 2k(p^5 - p^6)}{p^6}$ 。

应用切比雪夫不等式,可以得到如下推论。

$$\text{推论 1} \quad \Pr(|X - K| \geq \epsilon K) \leq \frac{p^3 - p^6}{p^6 \epsilon^2 K} + 2k \frac{p^5 - p^6}{p^6 \epsilon^2 K^2}.$$

由推论 1 可知,提出的采样方法计算图中三角个数的准确性依赖于图中三角的个数、含有共享边的三角对的个数以及采样比例。图中含有的三角个数越多,该方法的估计准确性越高;图中含有共享边的三角对越少,该方法的估计准确性越高。同时,该方法的估计准确性随着采样比例 p 的增加而提高。

1.3 加速比分析

当采用节点迭代方法计算图 $G=(V,E)$ 中的三角个数

时,对于图中的每个节点 $v \in V$ (v 的邻居对个数为 $\binom{d_v}{2}$),枚举出 v 的邻居对并判断该邻居对是否互为邻居,因此其计算复杂度为 $\sum_{v \in V} \binom{d_v}{2}$ 。

在本文算法中,第一步是对图进行采样,时间复杂度为 $O(n)$ 。在以采样比例 p 进行采样后,得到的新图中含有的边为原来的 p 倍,每个节点的度数也为原来的 p 倍。根据节点迭代方法,可得采样后的节点迭代的时间复杂度为 $\sum_{v \in V} \binom{p \cdot d_v}{2}$,即 $p^2 \sum_{v \in V} \binom{d_v}{2}$ 。于是,该算法的时间复杂度为 $O(n) + p^2 \sum_{v \in V} \binom{d_v}{2}$,该算法与节点迭代算法的加速比近似为 p^2 。

2 实验分析

实验采用的平台为一台普通的 PC,其包含一个英特尔酷睿双核 2.4GHz CPU,内存为 4GB,磁盘存储空间为 1TB。本实验运行在 Ubuntu 10.04 操作系统下,采用的编程语言为 Java。

实验采用的数据集包含一个合成数据集和一个真实数据集。合成数据集是采用文献[18]的方法随机生成的一个 ER 图,该 ER 图包含 1 万个节点和 8 万条边;采用的真实数据集为公开的 Epinions 数据集^[1],该数据集包含 49290 个节点和 487181 条边。

在实验中,将提出的方法与节点迭代方法进行了对比,主要对比了算法的执行时间和准确性。首先,通过实验对比了两种算法在两个数据集上的执行时间,并观察了算法的加速比,实验结果如图 2 和图 3 所示。在 ER 图和 Epinions 数据集中,节点迭代算法所需要的执行时间分别为 254s 和 487s。在对提出的算法进行测试的过程中,逐渐改变采样参数 p 的大小,将 p 从 0.1 增加到 0.9,每次增加 0.1。图 2 为本文提出的方法在 ER 图上的测试结果,左侧纵坐标表示算法的执行时间,右侧的纵坐标表示算法的加速比。从图 2 可以看出,随着 p 值的不断增加,该算法所需的运行时间不断增多,并且近似呈指数增长。与此同时,算法的加速比也相应地有所减小。图 3 为本文提出的方法在 Epinions 数据集上的测试结果,其坐标轴的意义同图 2。在图 3 中,算法的运行时间随着 p 值的增加不断增加,并且加速比不断减小。该结果与图 2 的结果相似,区别在于,ER 图是一个随机图,当图的规模增大到一定程度时,这种随机性会趋于稳态值。在 Epinions 数据集中,该数据集为真实的社会网络数据,当逐渐增大采样比例时,采样得到的子网络往往具有随机性,因而得到的结果曲线有所波动。

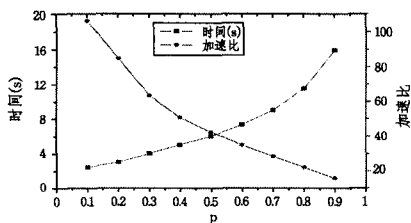


图 2 ER 图上算法的执行时间随 p 的变化

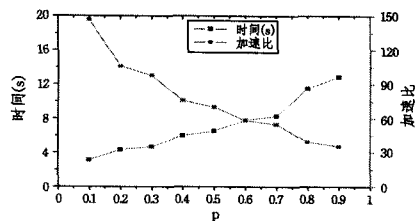


图 3 Epinions 数据集上算法的执行时间随 p 的变化

接下来,分析了本文提出的方法的准确性。将采用节点迭代算法得到的三角个数作为精确值,将由提出的方法得到的三角个数与精确值的比值作为算法的准确性度量标准。图 4 和图 5 分别为提出的方法在 ER 图和 Epinions 数据集上的测试结果。图中左侧的纵坐标为算法的准确性,右侧纵坐标为算法的加速比。这两幅图的相同点是,随着加速比的下降,算法的准确性逐渐提高。这意味着在真实的三角计算应用中,需要对算法的准确性和执行效率进行权衡,高效率往往以牺牲准确性为前提。

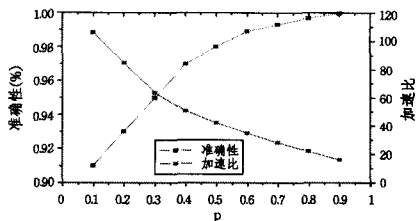


图 4 ER 图上算法的准确性随 p 的变化

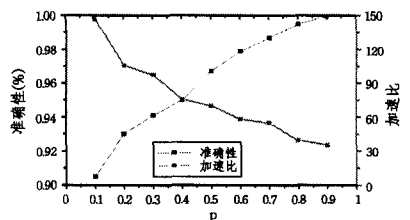


图 5 Epinions 数据集上算法的准确性随 p 的变化

最后,将图 4 和图 5 的结果显示在一个图中,以观察算法的准确性随着加速比的变化分布,结果见图 6。从图 6 可以看出,合成数据集(ER 图)的点构成的曲线在真实数据集(Epinions)的左下侧,这表明提出的方法在处理真实网络时具有更好的性能。此外,在两个数据集下,本文提出的方法在保证准确性大于 90%的前提下,加速比可以提高至 100 甚至更多。

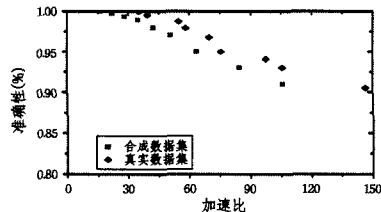


图 6 算法的准确性随着加速比的分布

结束语 三角计数是复杂网络和社会网络研究领域的重要研究内容之一。通过统计网络中的三角个数可以分析网络的同质性和传递性等特征。本文以一定的采样概率对网络中的边进行采样,在得到的子网络中进行相应的三角计数,并根

(下转第 227 页)

- [10] Bojinov H, Sanchez D, Reber P J, et al. Neuroscience Meets Cryptography: Designing Crypto Primitives Secure Against Rubber Hose Attacks[C]//USENIX Security Symposium, 2012; 129-141
- [11] Castelluccia C, Dürmuth M, Perito D. Adaptive Password-Strength Meters from Markov Models[C]//NDSS, 2012
- [12] Burr W E, Dodson D F, Polk W T. Electronic authentication guideline[R]. NIST special publication 800-63, 2006
- [13] de Carnavalet X C, Mannan M. From very weak to very strong: Analyzing password-strength meters[C]//Proceedings of the Network and Distributed System Security Symposium, 2014
- [14] Spafford E H. Opus: Preventing weak password choices [J]. Computers & Security, 1992, 11(3): 273-278
- [15] Davies C, Ganesan R. Bapasswd: A new proactive password checker[C]//16th National Computer Security Conference, 1993; 1-15
- [16] Vijaya M S, Jamuna K S, Karpagavalli S. Password strength prediction using supervised machine learning techniques[C]//International Conference on Advances in Computing, Control, & Telecommunication Technologies, 2009 (ACT'09). IEEE, 2009; 401-405
- [17] Blundo C, D'Arco P, Santis A D, et al. A Novel Approach to Proactive Password Checking [J]. Computer Science, 2002, 2437: 30-39

(上接第 190 页)

据子网络中的三角个数估计原网络中的三角个数。本文对采样方法的均值和方差进行了理论分析,并给出了通过采样方法得到的加速比。理论分析与实验表明,与传统的节点迭代方法相比,本文提出的方法在保证高准确性的前提下大大提高了算法的运行效率。

参 考 文 献

- [1] 张建国,林亚平,周四望,等. 传感器网络中误差有界的小波数据压缩算法[J]. 软件学报, 2010, 21(6): 1364-1377
Zhang Jian-ming, Lin Ya-ping, Zhou Si-wang, et al. Haar Wavelet Data Compression Algorithm with Error Bound for Wireless Sensor Networks[J]. Journal of Software, 2010, 21(6): 1364-1377
- [2] 张聪,沈惠璋. 基于谱方法的复杂网络中社团结构的模块度[J]. 系统工程理论与实践, 2013, 33(5): 1231-1239
Zhang Cong, Shen Hui-zhang. Modularity for community structure in complex networks based on spectral method[J]. Systems Engineering-Theory & Practice, 2013, 33(5): 1231-1239
- [3] 黄萍,张许杰,刘刚. 小世界网络的研究现状与展望[J]. 情报杂志, 2011, 26(4): 66-68
Huang Ping, Zhang Xu-jie, Liu Gang. The Present Research Situation and Forecast of the Small World Network[J]. Journal of Information, 2011, 26(4): 66-68
- [4] Tsourakakis C E. Fast counting of triangles in large real networks without counting: Algorithms and laws[C]//The Eighth IEEE International Conference on Data Mining, 2008. IEEE, 2008: 608-617
- [5] Centola D. An experimental study of homophily in the adoption of health behavior[J]. Science, 2011, 334(6060): 1269-1272
- [6] Wittkop T, Rahmann S, Röttger R, et al. Extension and robustness of transitivity clustering for protein-protein interaction network analysis[J]. Internet Mathematics, 2011, 7(4): 255-273
- [7] Becchetti L, Boldi P, Castillo C, et al. Efficient semi-streaming algorithms for local triangle counting in massive graphs[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2008: 16-24
- [8] Eckmann J P, Moses E. Curvature of co-links uncovers hidden thematic layers in the world wide Web[J]. Proceedings of the National Academy of Sciences, 2012, 99(9): 5825-5829
- [9] Bar-Yossef Z, Kumar R, Sivakumar D. Reductions in streaming algorithms, with an application to counting triangles in graphs [C]//Proceedings of the Thirteenth annual ACM-SIAM Symposium on Discrete Algorithms. 2012: 623-632
- [10] 陈学工,邱华,付金华,等. 基于三角形不规则网模型的快速体系化方法[J]. 计算机应用, 2010, 30(12): 3281-3283
Chen Xue-gong, Qiu Hua, Fu Jin-hua, et al. Fast voxelization based on triangulated irregular network model[J]. Journal of Computer Applications, 2010, 30(12): 3281-3283
- [11] Schank T, Wagner D. Finding, counting and listing all triangles in large graphs, an experimental study[M]//Experimental and Efficient Algorithms. Springer Berlin Heidelberg, 2005: 606-609
- [12] Lind P G, González M C, Herrmann H J. Cycles and clustering in bipartite networks [J]. Physical Review E, 2005, 72(5): 056127
- [13] 彭宇,罗清华,彭喜元. 网络化测试体系中不确定性数据处理方法浅析 [J]. 仪器仪表学报, 2010, 31(1): 229-240
Peng Yu, Luo Qing-hua, Peng Xi-yuan. Analysis of uncertain data processing methods in networking test framework[J]. Chinese Journal of Scientific Instrument, 2010, 31(1): 229-240
- [14] Jowhari H, Ghodsi M. New streaming algorithms for counting triangles in graphs[M]//Computing and Combinatorics. Springer Berlin Heidelberg, 2005: 710-716
- [15] Buriol L S, Frahling G, Leonardi S, et al. Counting triangles in data streams[C]//Proceedings of the Twenty-fifth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. ACM, 2006: 253-262
- [16] Becchetti L, Boldi P, Castillo C, et al. Efficient semi-streaming algorithms for local triangle counting in massive graphs[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2008: 16-24
- [17] Pavan A, Tangwongsan K, Tirthapura S, et al. Counting and sampling triangles from a graph stream[J]. Proceedings of the VLDB Endowment, 2013, 6(14): 1870-1881
- [18] Gong B, Yang L, Yang K. Synchronization on Erdős-Rényi networks[J]. Physical Review E, 2005, 72(3): 037101
- [19] 吴渝,杨涛,肖开洲. BBS 突发舆情分析及基于小世界网络的预测模型[J]. 重庆邮电大学学报(自然科学版), 2010, 22(3): 350-354
Wu Yu, Yang Tao, Xiao Kai-zhou. Analysis of emergent BBS sentiment and its prediction model based on small world network[J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition), 2010, 22(3): 350-354