

蛋白质作用网络中模体识别技术研究

孔德生, 何洁月

(东南大学 计算机科学与工程学院, 江苏 南京 210096)

摘要:生物网络是利用网络理论对生物系统进行建模,从而借助于网络的概念、属性和复杂网络研究的各种方法来理解生物系统的演化 and 行为。生物网络是生物信息学中一个崭新的研究领域,特别是蛋白质作用网络中网络模体具有很重要的生物意义。网络模体为在某个网络的多个不同部分出现的相互连接的子结构,其表达程度明显高于在随机网络中的表达。文中对模体识别技术进行了研究,系统阐述了模体识别技术的研究现状和各种技术方法,展望了模体识别技术的未来研究方向。识别大模体及将模体跟功能相结合将是该领域的发展方向。

关键词:网络模体;蛋白质作用网络;图挖掘

中图分类号:TP393

文献标识码:A

文章编号:1673-629X(2010)08-0001-04

Survey of Motif Discovery Algorithms in Protein-Protein Interaction Networks

KONG De-sheng, HE Jie-yue

(School of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract: Network theory is used to model biological systems by biological networks. Therefore, the concepts, properties, and various methods of network could be applied to understand the evolution and behavior of biological systems. It is a new field of bioinformatics, in particular motifs in protein interaction network have very important biological significance. Motifs are patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks. In this paper, give a survey of the up-to-date research work and the methods of motif discovery in protein interaction network. Moreover, also describes the future research prospect in the motif discovery. How to identify large motif and combine motif with function is a working direction in the future.

Key words: network motif; protein interaction network; graph mining

0 引言

生物网络是利用网络理论对生物系统进行建模,从而借助于网络的概念、属性和复杂网络研究的各种方法来理解生物系统的演化 and 行为^[1]。生物网络是生物信息学中一个崭新的研究领域。经常讨论的生物网络有蛋白质作用网络、基因调控网络、新陈代谢网络、神经网络等。近年来,描述这些网络的数据有了显著的发展,出现了很多生物网络数据库。蛋白质作用网络是生物网络中很重要的一类网络。

近年来一些研究者^[2]对复杂自然网络进行了深入研究,发现存在一些表达程度比随机网络高得多的子结构,他们称这些重复出现的局部子结构为模体,是基

于“进化过程中所保存的特定功能模块”^[3]。

模体的概念由 Shen, Milo 等人在文献[4]中首次提出,他们定义网络模体为在某个网络的多个不同部分出现的相互连接的子结构,其表达程度明显高于在随机网络中的表达。一些研究者已经将识别的模体应用到实际中,并取得了惊人的效果^[5~8]。

本综述主要致力于蛋白质反应网络的模体识别过程,尽管在文中主要考虑蛋白质反应网络,但是这些方法也适用于其他生物网络,甚至其他领域的复杂网络,例如转录调控网络、社会网络和因特网等。

1 蛋白质作用网络定义及特性

蛋白质作用网络可以用图来表示,蛋白质表示为图的顶点,蛋白质之间的反应表示为图的边。不妨假设该图为 G , G 是由顶点集合 V 和边集合 E 组成,记为 $G = (V, E)$ 。其中 $V(G)$ 是顶点的有限集合,且 $V \neq \emptyset$; $E(G)$ 是边的有限集合,边由 V 中的不同顶点对构

收稿日期:2009-12-08;修回日期:2010-03-24

基金项目:江苏自然科学基金项目(BK2007105)

作者简介:孔德生(1982-),男,江苏泰州人,硕士生,研究方向为数据挖掘、生物信息学;何洁月,教授,研究方向为数据挖掘、生物信息学、数据库技术、信息集成等。

成。

模体有两种判断方式,分别基于频率和统计意义,前者认为模体是在原图中表达次数超过某个阈值的子图,后者认为模体是表达程度在原图中高于在一系列随机图中的子图。这两种判断方式是不等价的。文献[2,9]认为上述的模体定义有个缺点,可能会遗漏那些功能上很重要,但是在统计上没有意义的模体。

子图的频率与计算子图出现次数时采取的重叠策略相关,文献[9]中提出了计算子图频率时的三种重叠策略 F_1, F_2, F_3 。 F_1 允许任意顶点和边的重叠, F_2 只允许顶点的重叠, F_3 不允许顶点和边的重叠。

向下封闭性是指子图的频率随着子图增大单调递减。该性质缩小了候选子图的数量,使得可以使用Apriori^[10]算法对子图进行剪枝。上述的三种重叠策略中, F_2, F_3 保持向下封闭性,而 F_1 则不保持。因此在设计模体搜索算法时如何计算子图频率是一个很关键的问题。在蛋白质反应网络中,同一种蛋白质经常在若干个模块中分别承担不同的作用,因此允许子图任意顶点和边的重叠是合理的。

根据模体的定义,模体是过度表达的子结构。现存的统计方法有 Z -score^[2], P -value^[11]和阈值法^[12]:

Z -score 于文献[2]中提出,公式如下:

$$Z(G_k) = \frac{f_{\text{real}}(G_k) - \bar{f}_{\text{rand}}(G_k)}{\text{std}(f_{\text{real}}(G_k))}$$

其中 $f_{\text{real}}(G_k)$ 是子图 G_k 在真实网络中出现次数, $\bar{f}_{\text{rand}}(G_k)$ 和 $\text{std}(f_{\text{real}}(G_k))$ 分别是子图 G_k 在一系列随机网络中出现次数的平均值和标准差。如果子图是过度表达的, Z -score就会很大,反之, Z -score就会是负数,或者接近于零。文献[11]中提出,如果 Z -score值大于2,则认为该子结构是过分表达的。

P -value 法于文献[11,13]中提出,公式如下:

$$P(m) = \frac{1}{N} \sum_{n=1}^N \delta_{F_{1,r}(m) \geq F_1(m)}$$

其中 $F_1(m)$ 是模体在真实网络中出现的频率, $F_{1,r}(m)$ 是模体在随机网络中出现的频率, N 是随机网络的个数, δ_c 为1,当条件 c 成立时,否则为0。如果 P -value值小于0.01,则认为该子结构是过分表达的。为了合理的计算 P -value,一般至少需要考虑一千个随机网络,然而如果考虑的随机网络比较少,则可以使用 Z -score法。

阈值法就是直接定义一个阈值,算法定义模体为repeated和unique的子结构。repeated是指子结构在原图中表达的次数超过了某个用户自定义的阈值 t_f ,unique则是指在一系列 n 个随机网络中,子结构至少在 t_n 个网络中出现超过 t_f 次,这里 t_f, t_n 和 n 均为给定

的参数, t_n 即为直接定义的阈值。

2 随机网络模型

文献[14]简单介绍了在复杂网络分析中涉及的一些典型性质和概念,例如距离、平均路径长度、聚类系数、度分布、相关系数等,这些性质描述了复杂网络的拓扑结构。

蛋白质反应网络具有较小的平均路径长度,较大的聚类系数,度分布服从幂律分布等拓扑特征。下面基于这些参数介绍一下复杂网络的基本随机模型,介绍三种网络基本模型,分别是ER模型^[15]、WS模型^[16]和BA模型^[17]。

ER模型是随机网络中最基本的模型^[15],其主要特征是:度分布为正态分布,每个结点有大致相同的度,小的平均路径长度和小的聚类系数。

ER模型描述了复杂网络的小世界特性,但是没有能够解释很多现实网络具有的高聚类系数。Watts和Strogatz提出了一种WS网络模型^[16],该模型同时具有高聚类系数和小的平均路径长度的特性。但除此之外,该模型没有能够解释现实网络的其他特征。

这两种网络模型都没有能够解释很多现实网络的“无尺度”特性,鉴于此,文献[17]提出了一种称为BA模型的随机网络,该模型构建了无尺度网络的演化过程,该演化模型基于两个关键机制:生长性和偏爱性选择。

在蛋白质反应网络模体识别过程中,网络模体的识别高度依赖于空模型的选择及其生成方法。在文献中使用得最多的空模型是ER随机模型,因为该模型具有小的聚类系数,即几乎不具有局部结构。

文献[18]提出了一种随机网络生成算法,通过随机重绘现实网络中的边,来生成新的随机网络。该算法生成的随机网络跟原网络相比,具有不同的局部结构,并且每个结点的度与原网络相同。

3 模体识别算法

目前的模体识别方法可以分为以网络为中心的方法和以模体为中心的方法^[19]。第一类方法试图在现实网络中计算给定大小的所有或者大部分子图的统计意义,典型算法有穷举法、抽样法和NeMoFinder。由于现实网络非常复杂,所以这一类方法的时间复杂度和空间复杂度非常高,只能识别小规模模体。第二类方法基于某个特定子图,分析其在某个网络中是否是模体,典型方法可见文献[19],能够识别中等规模的模体,并且性能有很大提高,但不能识别所有的模体。

还有一类基于图的挖掘算法,在图事务中基于

Apriori^[20]和DFS算法挖掘频繁子图,能够为模体识别提供很好的思路,例如AGM^[10],FSG^[21],gSpan^[22],CloseSpan^[23]和FFSM^[24]等。

一个典型的模体识别算法由三步构成,第一步,在原图中找出特定大小的所有子图及其出现次数;第二步,分析第一步中的所有子图,将同构的子图归为同一类;第三步,将这些子图在原图中出现的频率与在一组随机图中出现的频率作比较,判断统计意义。

3.1 穷举法

文献[2]开创性地对各种自然界的真实网络进行了分析,揭开了模体识别研究的序幕。他们提出了一种穷举搜索算法,枚举各种真实网络中指定大小的所有模体,识别了一些具有特定功能的模体,例如前馈回路、单输入模体、多输入模体、子调控模体等。由于真实网络非常复杂,他们仅仅枚举了大小为3和4的模体。

文献[25]提出了一个更快的穷举搜索算法——ESU,该算法对原图中的每个顶点加以标注,并且排序。算法开始于单个结点,每次迭代增加一个结点,直到子图达到指定的大小 k 。在扩展过程中,对于每一个中间子图,维持一个候选结点集合。该算法能够搜索出原图中所有大小为 k 的子图,并且只搜索一次。由于不会重复搜索同一个子图,可以有效地降低算法运行的时间。

由于真实网络非常复杂,并且同构测试是一个NP问题,所有穷举法只能识别小规模模体。

3.2 抽样法

文献[26]提出了一种基于抽样的算法——Mfinder,识别的模体大小达到8。该算法用了逐层搜索的迭代方法,首先从原图中随机选择一条边,然后随机选择一条与这条边联接的边以扩展子图,直到子图达到指定的大小。该算法对于网络大小具有很好的扩展性,然而最大的缺点是抽样偏置,并且还导致了调整偏置的额外计算代价。

文献[25]针对上述的抽样偏置,提出了一个新的抽样算法RAND-ESU,识别的模体大小达到14。该算法基于上文所述的ESU算法,ESU算法完全遍历ESU树,而RAND-ESU树则给每层赋予一个概率值 $0 < p_d \leq 1$,其中 $1 \leq d \leq k$,当 $p_d = 1$ 时,RAND-ESU算法退化为ESU算法。由此可见,RAND-ESU访问每个ESU树叶结点的概率相等,即 $\prod_{1 \leq d \leq k} p_d$ 。Mfinder算法存在抽样偏置,而RAND-ESU算法则解决了此问题,但是却存在如何给ESU树每层分配概率的问题。文献[25]采取了这样的策略:随着 d 的增大, p_d 逐渐减小,这可以保证访问ESU树尽量多的区域。综上

所述,RAND-ESU在ESU的基础上增加了抽样参数,具有很多优点,首先它是无偏置的,其次由于不需要修正偏置,所以易于实现,算法高效。

3.3 NeMoFinder

NeMoFinder能够识别模体的大小达到12。第1节中介绍了该方法中模体的定义,该定义是基于文献[2]提出的,该文献通过穷举法对现实网络进行分析后发现,具有实际意义的模体一般都同时具有repeated和unique的性质。NeMoFinder基于SPIN方法^[27],首先搜索频繁树,然后通过连接操作扩展频繁树为子图,从而可以显著减少候选模体的数量。算法的第一阶段是在原图中搜索repeated子图,第二阶段是生成一系列随机图,并计算第一阶段中每个repeated子图的unique值,从而识别出repeated和unique的子图。

3.4 其他方法

文献[28]提出了一种启发式方法,可以显著加快同构测试过程。算法对每个子图分配一个标号,保证同构的图具有相同的标号,不同构的图则具有不同的标号,该算法的精确度依赖于使用的标号个数。文献对该算法进行了测试,用于穷举法,能够识别的模体大小达到8。跟抽样法结合,识别的模体大小则达到14。

在文献[19]中,Joshua使用了symmetry-breaking技术,消除了由于子图的对称性而引入的重复同构测试,使得算法的运行速度达到指数级的提高。他们将该算法运用了蛋白质作用网络和基因调控网络,识别了一个大小达到15的模体。

Ziv等人提出了一种基于矩阵计算的搜索算法,该方法提出了functionals和scalars的概念。functionals表示在网络中向前移动或者向后移动,scalars则能够表示子图,可惜的是,scalars与子图之间的对应关系不是那么的显而易见,并且它们之间是多对多关系^[29]。该方法跟其他方法相比,在效率上具有压倒性的优势,并且能够作用于大型模体,但尚需要一些改进。

Schreiber等人在文献[9]中提出了一种称为FPF的方法,该方法基于hSigGram和vSigGram方法^[30],能够识别的模体大小达到9。由于候选子图个数随着顶点个数呈指数增长,算法基于正规标号提出了Generating Parent的概念,即每个大小为 i 的子图都唯一地由一个大小为 $i-1$ 的子图生成,这样可以避免同一个子图重复生成多次。该文献还提出了算法的并行版本,极大地缩小了算法的运行时间。

4 结束语

如何从浩瀚的生物网络中识别出跟功能相关的结构是当前的一个研究热点,而如何从生物网络识别出

模体是研究生物网络结构和功能的关键一步。

文中对生物作用网络中的模体识别算法进行了概述,模体识别是一个需要继续深入研究的领域,当前的搜索方法只能识别小规模模体和近似识别中等规模的模体,对于大规模模体的识别根本没有解决方法。将来的研究方向包括改进抽样技术,基于模体对网络进行分析,以及基于矩阵计算来识别模体。

如何应用识别出的模体也是一个热门的研究方向,近年来,一些科学家开始利用模体去分析蛋白质作用网络,例如, Saito 等人^[6,7]利用提取的模体来检测 PPI 网络中的误报率; Albert 等人^[5]利用模体来预测 PPI 网络; Middendorf 等人^[8]利用机器学习方法来对网络进行分类。这表明模体可以为复杂网络的深入研究提供有效的帮助。

模体与功能的结合是一个很有前景的领域, Milo 等人在文献[2]中对若干种现实世界的网络进行了分析,发现具有相似功能的网络具有相同的模体,这是一个非常诱人的领域。

因此,对特定生物网络的模体识别过程,所要追求的目标是能够发现尽可能大的模体,同时还要兼顾较小的时空开销,同时,还要研究模体的应用,尤其将模体跟功能相结合,从而发现蛋白质网络的设计原则和进化准则。

参考文献:

- [1] Barabasi A L, Oltvai Z N. Network biology: understanding the cell's functional organization[J]. *Nature Reviews Genetics*, 2004, 5(2): 101 - 113.
- [2] Milo R, Shen - Orr S, Itzkovitz S, et al. Network motifs: Simple building blocks of complex networks[J]. *Science*, 2002, 298: 824 - 827.
- [3] Vespignani A. Evolution Thinks Modular[J]. *Nature Genetics*, 2003, 35(2): 118 - 119.
- [4] Shen - Orr S S, Milo R, Mangan S, et al. Network motifs in the transcriptional regulation network of *Escherichia coli*[J]. *Nature Genet*, 2002, 31: 64 - 68.
- [5] Albert I, Albert R. Conserved network motifs allow protein - protein interaction prediction[J]. *Bioinformatics*, 2004, 20(18): 3346 - 3352.
- [6] Saito R, Suzuki H, Hayashizaki Y. Interaction generality, a measurement to assess reliability of protein - protein interaction[J]. *Nucleic Acids Res*, 2002, 30: 1163 - 1168.
- [7] Saito R, Suzuki H, Hayashizaki Y. Construction of reliable protein - protein interaction networks with a new interaction generality measure[J]. *Bioinformatics*, 2002, 18: 756 - 763.
- [8] Middendorf M. Inferring network mechanisms: The *Drosophila melanogaster* protein interaction network[J]. *Proc. Natl. Acad. Sci.*, 2005, 102(9): 3192 - 3197.
- [9] Schreiber F, Schwob B, Bermeyer H. Frequency concepts and pattern detection for the analysis of motifs in networks[J]. *Transactions on Computational Systems Biology*, 2005(3): 89 - 104.
- [10] Inokuchi A, Washio A, Okada T, et al. Applying the Apriori - based Graph Mining Method to Mutagenesis Data Analysis[J]. *Journal of Computer Aided Chemistry*, 2001(2): 87 - 92.
- [11] Kashtan N, Itzkovitz S, Milo R, et al. Mfinder tool guide[R]. Israel: Department of Molecular Cell Biology and Computer Science and Applied Mathematics, Weizman Institute of Science, 2002.
- [12] Chen J, Hsu W, Lee M L, et al. NeMoFinder: Dissecting genome - wide protein - protein interactions with meso - scale network motifs[C] // In: KDD 2006. Philadelphia, USA: ACM, 2006: 106 - 115.
- [13] Wernicke S, Rasche F. FANMOD: a tool for fast network motif detection[J]. *Bioinformatics*, 2006, 22(9): 1152 - 1153.
- [14] Junker B H. Analysis of Biological Networks[M]. Oxford: Wiley Blackwell, 2008.
- [15] Erdos P, Rnyi A. On random graphs I[J]. *Publ. Math.*, 1959(6): 290 - 297.
- [16] Watts D J, Strogatz S H. Collective dynamics of small - world networks[J]. *Nature*, 1998, 393: 440 - 442.
- [17] Barabási A - L, Albert R. Emergence of scaling in random networks[J]. *Science*, 1999, 286: 509 - 512.
- [18] Maslov S, Sneppen K. Specificity and Stability in Topology of Protein Networks[J]. *Science*, 2002, 296: 910 - 913.
- [19] Grochow J A, Kellis M. Network motif discovery using subgraph enumeration and symmetry - breaking[C] // RECOMB. [s.l.]: [s.n.], 2007: 92 - 106.
- [20] Inokuchi A, Washio T, Motoda H. An apriori - based algorithm for mining frequent substructures from graph data[C] // In Proc. of the 4th European Conference on Principles of Data Mining and Knowledge Discovery. [s.l.]: [s.n.], 2000: 13 - 23.
- [21] Kuramochi M, Karypis G. Frequent subgraph discovery[C] // In: Cercone N, Lin T Y, Wu X, eds. Proc. of the 2001 IEEE Int'l Conf. on Data Mining. San Jose: IEEE Computer Society, 2001: 313 - 320.
- [22] Yan X, Han J. gSpan: Graph - Based substructure pattern mining[C] // In: Kumar V, Tsumoto S, Zhong N, Yu P S, Wu X, eds. Proc. of the 2002 IEEE Int'l Conf. on Data Mining. Maebashi: IEEE Computer Society, 2002: 721 - 724.
- [23] Yan X, Han J. Closegraph: Mining closed frequent graph patterns[C] // In: Proc. of the 2003 IEEE Int'l Conf. on Data Mining. San Jose: IEEE Computer Society, 2003: 721 - 724.

开销与低带宽需求的特点,使绘制系统的渲染规模得以进一步扩展,并大大缓解了体绘制阶段的性能瓶颈。

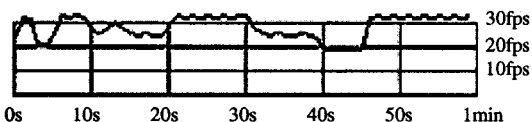


图 3 600x600x6000 数据集的 1 分钟帧率跟踪曲线

4 结束语

提出了一种基于预计算切片序列的动态判断体绘制技术,该技术从切片方向空间中预选取一组切片方案作为备选方案,绘制阶段根据数据体与视点的位置关系从备选方案中选取最适合的切片方案进行绘制。该技术在运行时传输方案索引号而非实际几何数据,从而达到降低带宽需求的目的,同时,该技术还拥有运行时计算开销低,切片结果规则且易于处理等优点。通过理论推导结合实验反馈验证了该算法的普遍适用性。同时,实验结果证明了该算法可以达到提高体绘制性能、拓展体绘制规模的目的。

目前该技术也存在尚待改进之处。首先,该算法对非分块的数据集体绘制效果尚不理想,且不支持视点处于数据体内部的情形;另外,因几何数据基于预计算,步长固定,所以一定程度上减少了多分辨率控制的灵活性。

参考文献:

- [1] Akeley K. Reality Engine Graphics[J]. Computer Graphics, 1993, 27: 109-116.
- [2] Lamar E, Hamann B, Joy K I. Multiresolution techniques for interactive hardware texture-based volume visualization

[C]//Proceedings of IEEE Visualization. Los Angeles, California:[s. n.],1999:355-361.

- [3] Gobbetti E, Marton F, Antonio J. A single pass GPU ray casting framework for interactive out-of-core rendering of massive volumetric datasets[J]. Vis Comput, 2008, 24(7): 797-806.
- [4] Lefebvre S, Dachsbacher C. Tiled trees[C]// Proceedings of SIGGRAPH. San Diego, California:[s. n.],2007: 25-31.
- [5] 马仁安,张二华,杨静宇. 不规则地质体的分割与体绘制方法研究[J]. 计算机研究与发展,2005,42(5):883-887.
- [6] 曹轶,莫则尧,王弘堃. 协同分布式图形硬件的混合并行体绘制[J]. 中国图像图形学报,2008,13(7):1379-1384.
- [7] Gobbetti E, Marton F. Far voxels: a multiresolution framework for interactive rendering of huge complex 3d models on commodity graphics platforms[C]// Proceedings of SIGGRAPH. Los Angeles, California:[s. n.],2005:878-885.
- [8] Crail C, Fabrice N, Sylvain L. GigaVoxels: ray-guided streaming for efficient and detailed voxel rendering[C]//Proceedings of the symposium on Interactive 3D graphics and games. Boston, Massachusetts:[s. n.],2009:15-22.
- [9] 薛健,田捷,戴亚康. 海量医学数据处理框架及快速提绘制算法[J]. Journal of Software,2008,19(12):3237-3248.
- [10] 肖永飞,付宜利,王国树. 硬件加速的大数据量自适应体绘制[J]. 计算机辅助设计与图形学学报,2009,21(5):612-616.
- [11] 宋涛,欧宗瑛,王瑜. 八叉树编码体数据的快速体绘制算法[J]. 计算机辅助设计与图形学学报,2005,17(9): 1990-1996.
- [12] 马晓晨,孔小利,陈建军. 一种大规模三维地震数据体绘制的 LOD 技术[J]. 大庆石油学院学报,2008,32(4):23-26.

(上接第 4 页)

terms[C]//In: Getoor L, Senator T E, Domingos P, Faloutsos C, eds. Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Washington: ACM, 2003:286-295.

- [24] Huan J, Wang W, Prins J. Efficient mining of frequent subgraphs in the presence of isomorphism[C]// In: Kumar V, Tsumoto S, Zhong N, Yu P S, Wu X, eds. Proc. of the 2003 IEEE Int'l Conf. on Data Mining. Melbourne: IEEE Computer Society, 2002:549-552.
- [25] Wernicke S. Efficient detection of network motifs[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics,2006,3(4):347-359.
- [26] Kashtan N, Itzkovitz S, Milo R, et al. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs[J]. Bioinformatics, 2004, 20(11): 1746-1758.
- [27] Huan J, Wang W, Prins J, et al. Spin: mining maximal frequent subgraphs from graph databases[C]//Proceedings of the tenth ACM SIGKDD International Conference on Knowledge discovery and data mining. [s. l.]:[s. n.], 2004:581-586.
- [28] Baskerville K, Paczusi M. Subgraph Ensemble and Motif Discovery Using a New Heuristic for Graph Isomorphism[EB/OL]. 2006-06-19. arXiv.org, q-bio/0606023.
- [29] Ciriello G, Guerra C. A Review on Models and Algorithms for Motif Discovery in Protein-Protein Interaction Network[J]. Briefings in Functional Genomics and Proteomics,2008,7(2): 147-156.
- [30] Kuramochi M, Karypis G. Finding frequent patterns in a large sparse graph[C]// In: SIAM International Conference on Data Mining (SDM'04) 2004. SIAM, Lake Buena Vista, Florida, USA:[s. n.],2004:243-271.