面向图数据管理系统基准评测的知识图谱统计特征分析

钱卫宁，孙晨，程文亮，周傲英
华东师范大学数据科学与工程研究院，上海 200062

摘要
近年来，图结构数据在信息安全、科学研究、互联网服务等各个领域被广泛采用，图数据管理系统也随之快速发展。然而，当前最主要的图数据管理系统评测基准都是面向社交网络服务和分析应用而设计和开发的。通过对知识图谱（knowledge graph）这一类快速发展的图结构数据的统计特征进行分析，并和社交网络进行比较，展示知识图谱和社交网络的显著区别，以此说明现有图数据管理系统基准评测无法满足知识图谱管理的需要，进一步展望图数据管理系统基准评测的需求和发展。

关键词
基准评测；图数据；知识图谱；统计特征

中图分类号：TP31 文献标识码：A doi: 10.11959/j.issn.2096-0271.2016049

Statistical characteristics analysis of knowledge graphs for benchmarking graph database management systems

QIAN Weining, SUN Chen, CHENG Wenliang, ZHOU Aoying
Institute for Data Science and Engineering, East China Normal University, Shanghai 200062, China

Abstract
Recently, graph data has been widely used in domains such as information security, scientific research, internet services, etc., that stimulates the fast development of graph data management systems. However, existing benchmarks for graph databases are all designed for applications that manage and analyze social networks. The statistical characteristics of knowledge graphs were analyzed, and compared with two social networks. It was showed that knowledge graphs, as an important and fast growing kind of graph data, were significantly different from social networks. Therefore, existing social network based benchmarks were not suitable for applications that deal with knowledge graphs. Furthermore, the requirements for a new benchmark were analyzed.

Key words
benchmark, graph data, knowledge graph, statistical measurement

1 引言

近年来学术界和工业界都见证了大规模知识图谱构建和应用的高涨热情。知识图谱已成为如扩展查询[1]、问答系统[2]这样的大规模Web应用的核心模块。例如，谷歌知识图谱中就包含了超过5亿个实体以及35亿条事实，用来完善谷歌搜索引擎的搜索结果①。然而，面对如此庞大的知识图谱数据，如何对它进行有效的管理就成了一个很重要的问题。尽管已经存在一些技术，比如使用具有特定索引的关系数据库管理系统或构建本地图数据管理系统，但大规模图数据的管理仍是一个研究难点和热点[3]。因此，需要专门的图数据管理评测基准来更好地理解知识图谱的构建与应用，并帮助用户选择合适的系统或技术。

目前已有的图数据管理基准评测主要是针对社交网络图谱管理应用而设计的。例如，Facebook 的LinkBenck[4]、LDBC的SNB (social network benchmark)[5]以及之前的研究工作BSMA[6]等。但需要注意的是，知识图谱与社交网络之间的本质区别在于知识图谱中的顶点节点和边都是用大量丰富的属性或语义标签标注过的，具有属性和语义标签，而在社交网络中这样的顶点和边却很少。所以，认为现有的面向社交网络的评测基准在对于知识图谱数据的管理问题中是不适用的。况且知识图谱和社交网络的应用场景不同，导致它们的查询负载也不相同。

本文将从描述知识图谱结构的统计特征这一特定视角,分析知识图谱管理基准评测的要求。通过对真实社交网络和知识图谱进行统计特征分析，研究两者之间的差异。

本文的贡献如下。

- 为特征化图谱结构引入了4类分布特征指标，对知识图谱进行了定量和定性的分析，并分析展示了它们在知识图谱的管理数据中的意义。
- 本文对4个知识图谱和2个社交网络进行了深入的分析。其中，知识图谱既包含了通用领域的知识图谱，也包含专门领域的知识图谱。通过实证，尽管展示了这些图数据在某些特定的结构特征上具有相似性，如顶点度数的幂律分布，但在其他特征上却具有明显的差异。同时本文也将差异存在的原因进行了详细分析。
- 分析了知识图谱评测基准中种子数据集以及数据生成器的一些要求，并分析了已有的社交网络数据生成器不能满足面向知识图谱数据管理的评测基准需要的原因。

2 数据集

本文分析2个社交网络和4个知识图谱数据集，简介如下。

社交网络：新浪微博是中国最重要的社交媒体服务之一。使用获取自新浪微博的两个关注关系图数据集。第一个数据集是从整个用户集中随机获取200 000个用户构建的关注关系子图SNRand。它包含了超过5 000 000个用户间关注关系。第二个数据集挑选了200 000个最活跃的用户后所构建的关注关系子图SNRank。它包含了超过36 000 000个关注关系。需要注意的是，SNRand接近于真实的社交网络数据，SNRank则接近很多对热点事件和用户进行分析的社交媒体分析应用中所处理的数据。

WordNet[7]：它是由美国普林斯顿大学设计开发的基于认知语言学的英语词典。其中名词、动词、形容词和副词等各自被组织成一个由同义词词集构成的网络，每个

① https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html

同义词集合都代表一个基本的语义概念，并且这些集合之间也由各种关系连接。在实验中直接利用WordNet，其中单词或概念构成了节点，而语义关系则作为边。

YAGO2[8]：YAGO2是一个继承综合了Wikipedia、WordNet 和GeoNames，经过扩展后得到的大型语义知识图谱。从中生成了3个子图，依次命名为YagoTax、YagoFact 和YagoWiki。其中YagoTax是YAGO2的分类知识，包含了子类实体之间的从属关系（subclassof）、上下位关系（isA）关系等；YagoFact则包含了YAGO2的所有事实关系，代表了事实性知识；YagoWiki包含了YAGO2中包含的Wikipedia页面之间的超链接关系，反映了Wikipedia的内在超链接结构。

DBpedia[9]：DBpedia是一个从Wikipedia中抽取得到的多语言知识库，其英文版本描述了4 580 000个实体和2 795个不同属性。基于DBpedia中所有事实性知识构造了数据集DBpediaFact。

EKG：EKG（enterprise knowledge graph）是根据上海证券交易所上市公司公报构建的特定领域知识图谱，包括企业和企业之间以及企业和人之间的assignment、hold、subcompany、changename、manager、cooperate、merge 7种关系，共包含51 853 个实体和430 973个关系。

3 统计特征

知识图谱和社交网络都可以用有向图$G=(V,E)$表示，其中V是节点的集合，表征实体或用户，E是有向边的集合，表征语义关系。主要考量关于图的4个分布特征，具体如下。

- 度数分布：图G的度数分布是$p(d)=n_d/|V|$，其中，n_d表示度数为d的节点数，$|V|$表示G中所有的节点数。在大多数图中，度数服从幂律分布，即$p(d)\propto L(d)d^{-\alpha}$，其中，$\alpha>1$且$L(d)$是一个慢变函数。本文中将分别研究入度和出度。
- 跳数分布：对于图G中的路径$P=\{v_1,v_2,\cdots,v_h\}$而言，其长度被定义为$Hops(P)=h-2$，其中，h表示P中的节点数。跳数分布反映了图中的连通代价。
- 连通分支的分布：强连通分支中任意两个节点间相互可以到达，弱连通分支则是在忽略边的方向后任意两个节点间可以连通。连通分支反映了图的连通性。
- 聚类系数的分布：一个节点v_i的聚集系数[10]的定义为$C_i=|\{e_{jk}:v_j,v_k\in N_i,e_{jk}\in E\}|/[|N_i|(|N_i|-1)]$，其中，$e_{jk}$是$v_j$和$v_k$之间的边$(j\neq k)$，而$N_i$是$v_i$的邻居节点的集合。聚集系数衡量的是节点聚集的趋势。

4 统计特征分析

利用工具SNAP[11]计算了4个知识图谱与2个社交网络的统计特征。基于不同的目标将图谱分为3组进行分析：为了研究同一知识图谱的不同部分，比较了YAGO2的3个子图；为了研究不同知识图谱之间的差异性，考量了4类知识图谱并对它们进行横向对比；还将知识图谱与社交网络作对比，并解释其间的区别及存在的原因。

图1展示了所有知识图谱以及社交网络的节点入度和出度。所有数据集的入度分布都接近于幂律分布（如**图1**(a)所示）。除了SNRank，这些幂律分布的指数α都在(1.8,2.4)附近。SNRank与其他数据集明显不同，它的初始部分明显偏离幂律，直到入度增加至560左右才服从幂律分布。这是因为SNRank中只包含最活跃的社交网络用户，因此和其他数据集之间有明显偏差。

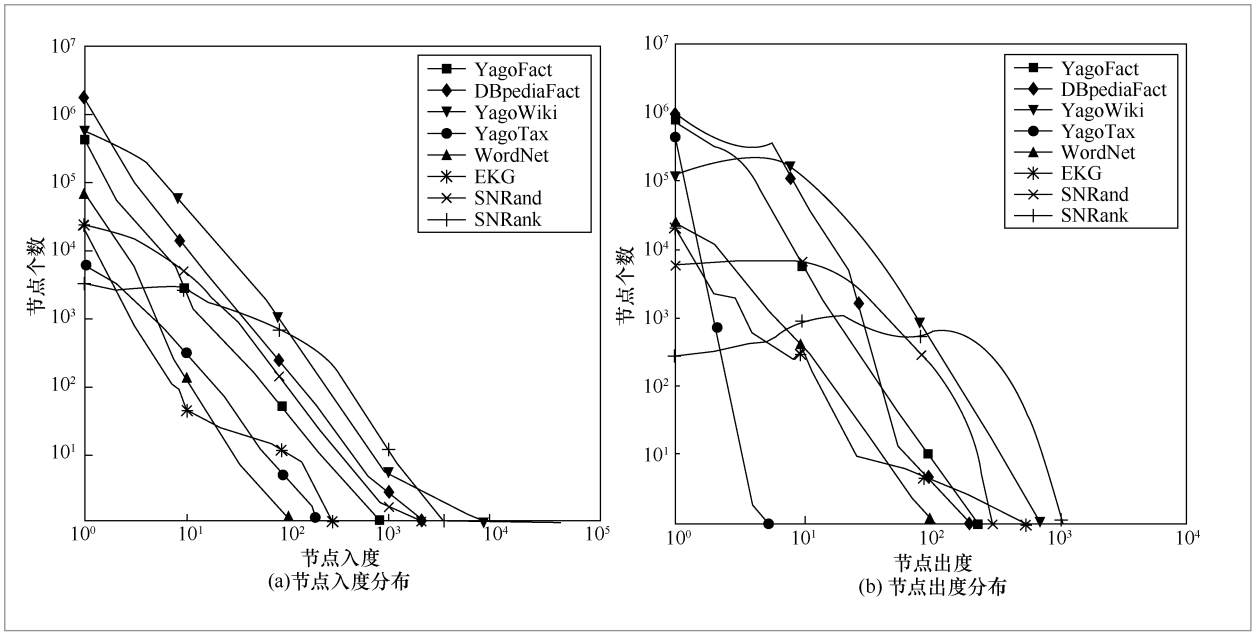

图1 知识图谱与社交网络节点度数分布

出度分布（如图1(b)所示）不同，可以看出，不仅知识图谱与社交网络间，甚至知识图谱之间也存在着明显的区别。3个YAGO2子图的出度分布具有显著不同。所有的分布最初都偏离幂律且彼此间不同，分布的下降率也变化很广，指数参数α波动于1.4~8.4。需要注意，YagoTax作为这些知识图谱数据集中唯一的层次状数据，其分布与其他知识图谱数据和社交网络数据都有着明显的区别。

图2展示了这些图数据的节点间距离分布，均呈现为S型。为了使数据适合于某些曲线，引入Sigmoid函数$f(x)=a/(1+e^{b-cx})$即可对图2(a)的数据进行准确建模。YagoTax

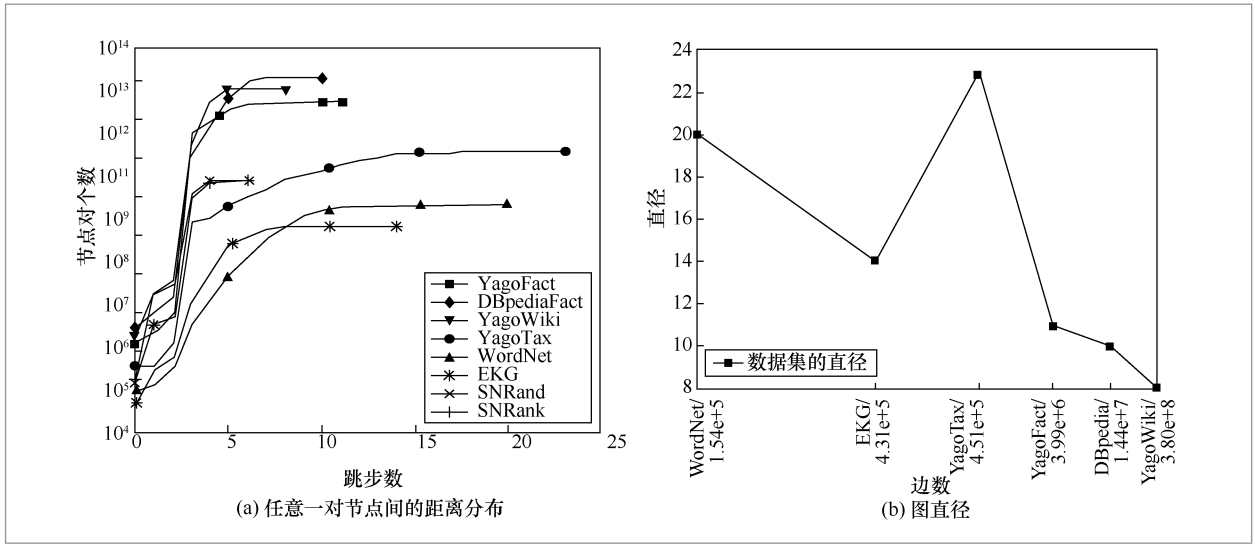

图2 知识图谱与社交网络节点间距离分布与图直径

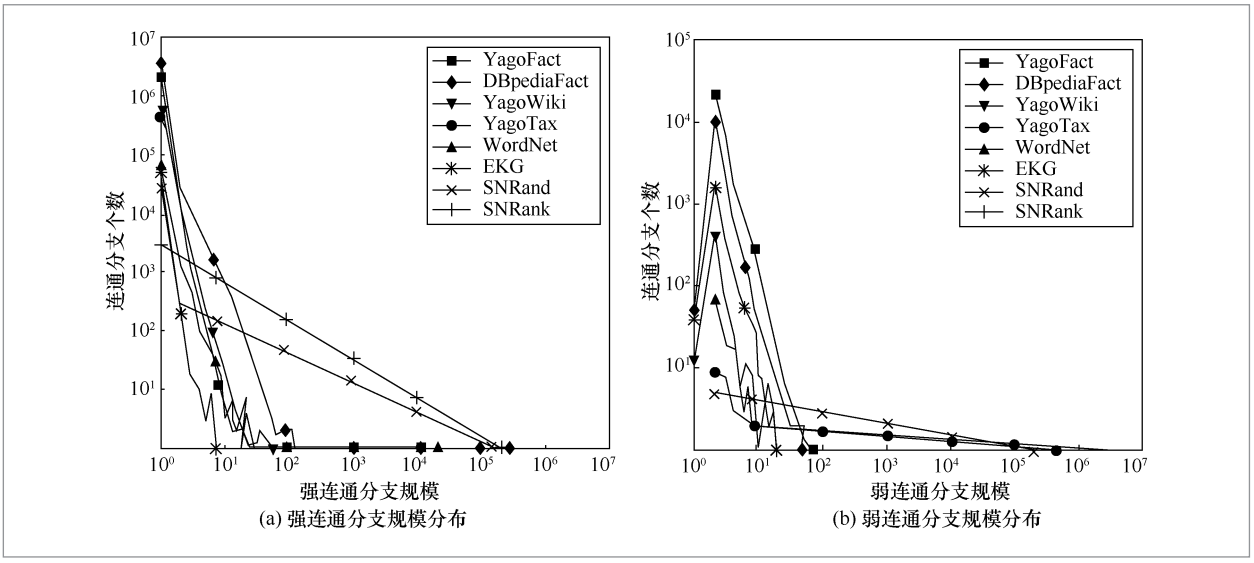

图 3 知识图谱与社交网络连通分支规模分布

和WordNet的直径，即任意一对节点间的最大跳数，普遍比其他数据集都大。而SNRand和SNRank的直径最小，近似于6，这正符合社交网络中的六度分隔现象[10]。另一个有趣的发现是，当跳数由2增加至3时，所有的分布都会产生爆炸式增长。还有一个有趣的现象是除YagoTax以外，所有数据集的直径和数据集节点个数的log值之间呈线性下降关系。这表明，这些数据集在节点间距离的分布上相似。

图3展示了连通分支的分布。除了YagoTax，知识图谱的强、弱连通分支分布都服从幂律分布。这是因为YagoTax是一棵拥有大量非连通叶子节点的扁平的树。在SNRand和SNRank这样的社交网络中，只有一个最大的强连通分支和一小部分的孤立点。知识图谱的强、弱连通分支的幂律分布显示知识图谱中的节点在小范围内聚集，而社交网络中的节点则被组织为一个大的强连通分支。

图4展示了平均聚类系数（average clustering coefficient，ACC）的分布。计算ACC时，将图视为无向图且度数视为出度与入度的总和。图4显示除了SNRank以外，其他的图在最初都倾向于呈现幂律分布，而SNRank则随着x的增大曲线发生偏离。社交网络的ACC一般高于知识图谱，这也反映了知识图谱的局部聚集属性不如社交网络。

进一步将EKG包含的7种关系分为7个子图。子图的节点度数分布如**图5**所示。从中可以看到，“manager”这一关系的出度分布在指数标度上部分呈钟形。这说明，即使去除了分类层次等具有树形或层次状

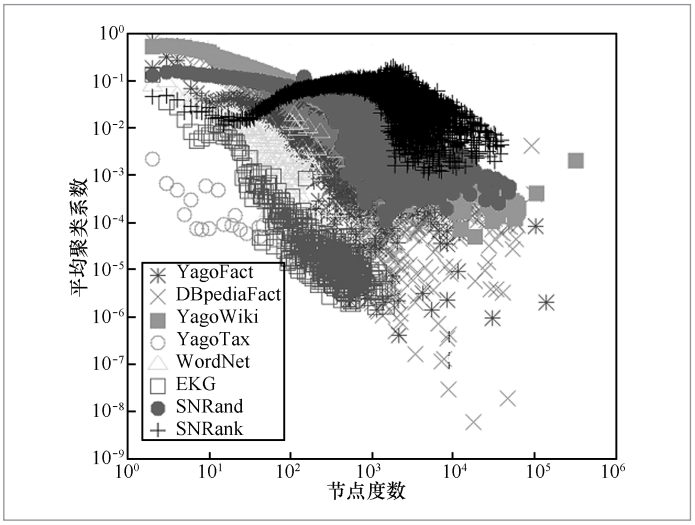

图 4 知识图谱与社交网络节点聚类系数分布

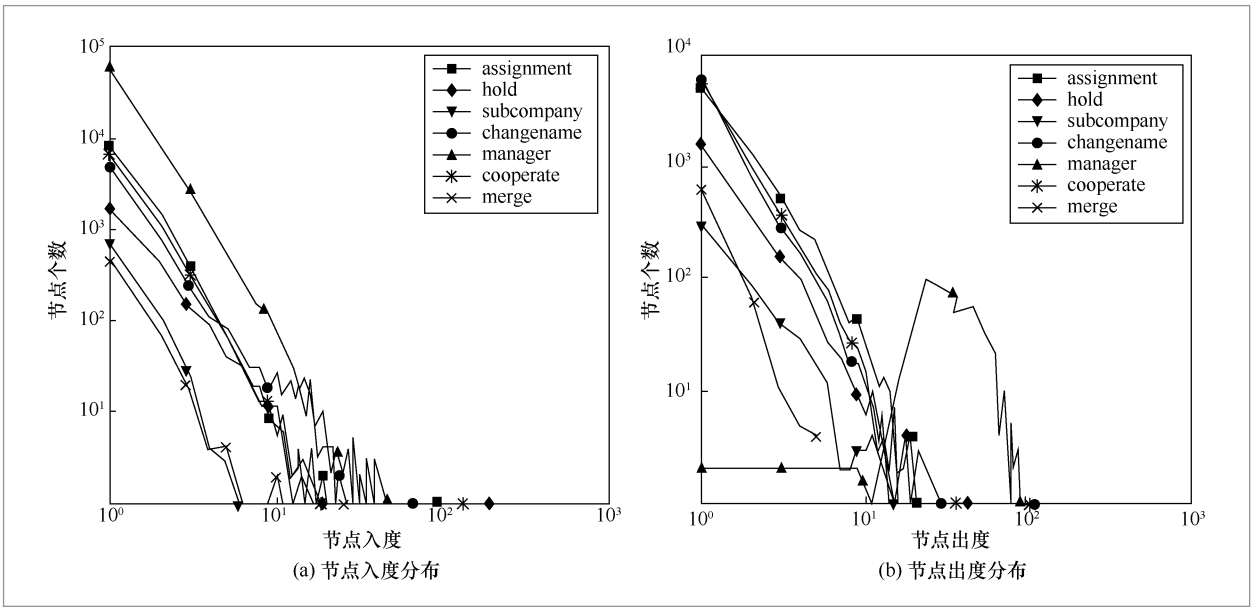

图 5 企业知识图谱各关系子图的节点度数分布

的部分，知识图谱的各部分的分布仍然有较大区别。

进一步分析7个子图的节点间距离分布，如**图6**所示，可以看到，“manager”关系不同于其他子图，而“hold”和“merge”又不同于其余4个子图。这进一步说明了知识图谱各部分之间存在较大区别。

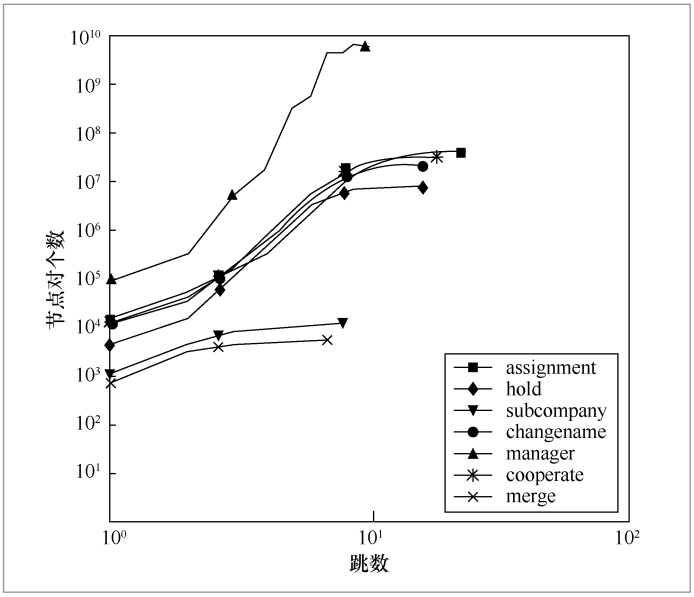

图 6 企业知识图谱各关系子图的节点间距离分布

这7个子图并非在所有统计特征上都不相同。如**图7**所示，在连通分支规模分布方面，这些子图相互间差别较小。同样，如**图8**所示，除“changename”这一关系聚类系数较高以外，剩余6个子图的聚簇特性相近。

5 相关工作

面向知识图谱管理的评测基准的制定需要建立在对知识图谱的统计特征详细分析的基础上。在对复杂网络的研究中，已经提出了很多结构度量标准和分布统计来对图的统计特征建模。这些统计度量包括节点度数、跳数、直径等。

研究人员已花费了大量精力来分析大规模图谱的结构化属性。Broder等人就通过一系列的图结构度量指标（如直径、节点、度等）来研究网络[12]。在社交网络方面，Kumar等人通过研究Flickr以及Yahoo! 360的动态时距图的结构属性诸如直径、度、社区规模等指标来研究其演变[13]。Boccaletti等人则调研了复杂网络的结构

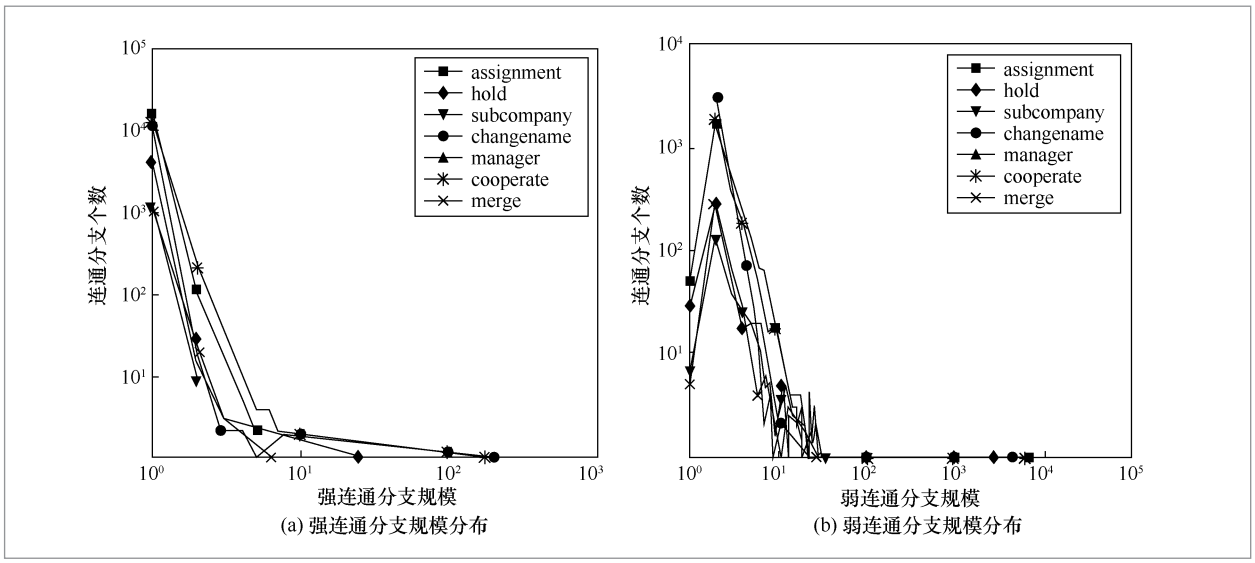

图 7 企业知识图谱各关系子图的连通分支分布

和动态变化[14]。笔者曾对包括EKG的一个早期版本的4个知识图谱和2个社交网络进行统计特征分析[15]。随着EKG的完善和规模增长，本文进一步拓展了原有的分析。

与此同时，大型图谱大规模图谱分析系统也在近年来得以迅速发展。Lancichinetti等人提出了关于节点度数和社区规模的具有分布不均匀性的图数据评测基准[16]。而在社交网络评测基准方面，LinkBench在对Facebook的社交网络工作负载进行特征分析的基础上实现了评测基准[4]。LDBC的社交网络评测基准（SNB）对一个类似于Facebook的综合社交网络进行了建模，它也是第一个基于瓶颈分析的LDBC评测基准，准确刻画了负载处理上的技术挑战[5]。BSMA是另一个社交媒体数据分析评测基准，旨在提供面向新浪微博分析的评测基准[6,17]。

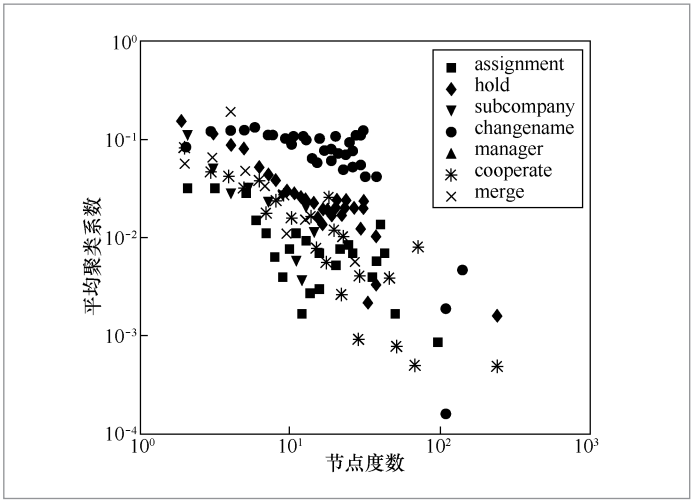

图 8 企业知识图谱各关系子图的节点聚类系数分布

6 结束语

本文分析了4个知识图谱和2个社交网络的统计特征。通过对它们的深入分析后可得出如下结论。

首先，知识图谱与社交网络在节点度数分布、节点间距离和网络直径、连通分支规模分布、节点聚簇程度等方面都不相同。因此，现有的社交网络数据以及相关的数据生成器所产生的模拟数据集都不能被用来模拟知识图谱数据。

其次，与社交网络相比，知识图谱更为复杂。这体现在两个方面。一方面，知识图谱的节点和边上的属性和标签在查询处

理、分析处理时占有重要地位，且属性值和标签值多样；另一方面，在不同属性和标签类别所决定的子图之间，各个统计特征差别巨大。这其中既有概念层次和本体这样的抽象知识的图结构与实体间联系这样的事实知识的图结构的显著区别，也有不同类别的实体间联系子图之间的区别。因此，单纯地用多个同构网络叠加的方式是无法构造出类似于知识图谱的复杂结构的。

第三，不同知识图谱之间仍然具有一定的相似性。在本文研究的4类统计分布特征上，各个知识图谱的事实部分显示出较大的相似性。因此，面向知识图谱应用，设计、开发图数据管理系统基准评测是有意义，也是可能的。

本文的工作仍刚起步。一方面，本文的统计分析只对单一子图进行，并未对知识图谱各子图间的联系进行分析；另一方面，本文只研究了最基本的4类分布特征。但本文的分析结果和结论对于设计评测基准，特别是其中的数据生成器，仍然具有明确的指导意义。

从本文的分析结果可知，面向知识图谱应用的图数据管理系统基准评测的数据生成器应能够生成分布上多样的大规模异构图数据，以弥补现有图数据生成器的不足。

参考文献:

[1] DALTON J, DIETZ L, ALLAN J. Entity query feature expansion using knowledge base links[C]//International ACM Sigir Conference on Research and Development in Information Retrieval, July 6-11, 2014, Gold Coast, QLD, Australia. New York: ACM Press, 2014: 365-374.

[2] JOSHI M, SAWANT U, CHAKRABARTI S. Knowledge graph and corpus driven segmentation and answer inference for telegraphic entity-seeking queries[C]// Conference on Empirical Methods in Natural Language Processing, October 25-29, 2014, Doha, Qatar. [S.l.:s.n.], 2014: 1104-1114.

[3] RAJARAMAN A, ULLMAN J D. Mining of massive datasets[M]. New York: Cambridge University Press, 2011.

[4] ARMSTRONG T G, PONNEKANTI V, BORTHAKUR D, et al. Linkbench: A database benchmark based on the facebook social graph[C]//ACM SIGMOD International Conference on Management of Data, June 22-27, 2013, New York, USA. New York: ACM Press, 2013: 1185-1196.

[5] ERLING O, AVERBUCH A, LARRIBA-PEY J, et al. The ldbc social network benchmark: Interactive workload[C]// ACM SIGMOD International Conference on Management of Data, May 31-June 4, 2015, Melbourne, Victoria, Australia. New York: ACM Press, 2015: 619-630.

[6] MA H, WEI J, QIAN W, et al. On benchmarking online social media analytical queries[C]//International Workshop on Graph Data Management Experiences and Systems, June 23, 2013, New York, USA. [S.l.:s.n.], 2013: 1-7.

[7] FELLBAUM C. WordNet: an electronic lexical database[M]. Massachusetts: MIT Press, 1998.

[8] HOFFART J, SUCHANEK F M, BERBERICH K, et al. Yago2: a spatially and temporally enhanced knowledge base from Wikipedia[J]. Artificial Intelligence, 2013(194): 28-61.

[9] LEHMANN J, ISELE R, JAKOB M, et al. Dbpedia: a large-scale, multilingual knowledge base extracted from Wikipedia[J]. Semantic Web Journal, 2015, 6(2): 167-195.

[10] WATTS D, STROGATZ S. Collective dynamics of 'small-world' networks[J]. Nature, 1998, 393(6684): 440-442.

[11] LESKOVEC J, SOSIC R. SNAP: a general purpose network analysis and graph mining library[J]. ACM Transactions on Intelligent Systems and Technology, 2016, 8(1): 1-20.

[12] BRODER A, KUMAR R, MAGHOUL F, et al. Graph structure in the web[J]. Computer

Networks, 2000, 33(1-6): 309-320.
[13] KUMAR R, NOVAK J, TOMKINS A. Structure and evolution of online social networks[M]. New York: Springer, 2006: 337-357.
[14] BOCCALETTI S, LATORA V, MORENO Y, et al. Complex networks: structure and dynamics[J]. Phys Rep, 2006, 424(4-5): 175-308.
[15] CHENG W, WANG C, XIAO B, et al. On statistical characteristics of real-life knowledge graphs[C]//The Workshop on Big Data Benchmarks, Performance, Optimization, and Emerging Hardware, September 4, 2015, Hawaii, USA. [S.l.:s.n.], 2015: 261-267.
[16] LANCICHINETTI A, FORTUNATO S, RADICCHI F. Benchmark graphs for testing community detection algorithms[J]. Physical Review E Statistical Nonlinear and Soft Matter Physics, 2008, 78(2): 561-570.
[17] 钱卫宁, 夏帆, 周敏奇, 等. 大数据管理系统评测基准的挑战与研究进展[J]. 大数据, 2015, 1(1): 82-96.
QIAN W N, XIA F, ZHOU M Q, et al. Challenges and progress of big data management system benchmarks[J]. Big Data Research, 2015, 1(1): 82-96.

作者简介

钱卫宁(1976-),男,华东师范大学数据科学与工程研究院教授、博士生导师,主要研究方向为互联网环境下的数据管理、大数据管理系统评测基准、社交媒体数据分析、知识图谱构建与应用等。

孙晨(1995-),女,华东师范大学数据科学与工程研究院硕士生,主要研究方向为知识图谱构建。

程文亮(1989-),男,华东师范大学数据科学与工程研究院硕士生,主要研究方向为数据挖掘与知识发现。

周傲英(1965-),男,华东师范大学副校长,长江学者特聘教授,数据科学与工程研究院院长,主要研究方向为Web数据管理、数据密集型计算、内存集群计算、分布事务处理、大数据基准测试和性能优化。

收稿日期: 2016-08-10
基金项目: 国家自然科学基金资助项目(No.61432006)
Foundation Item: The National Natural Science Foundation of China (No.61432006)