On Statistical Characteristics of Real-Life Knowledge Graphs

Wenliang Cheng, Chengyu Wang, Bing Xiao, Weining Qian(✉), and Aoying Zhou

Institute for Data Science and Engineering,
ECNU-PINGAN Innovative Research Center for Big Data,
East China Normal University, Shanghai, China
{wenliangcheng,chengyuwang,bingxiao}@ecnu.cn,
{wnqian,ayzhou}@sei.ecnu.edu.cn

Abstract. The success of open-access knowledge graphs, such as YAGO, and commercial products, such as Google Knowledge Graph, has attracted much attention from both academic and industrial communities in building common-sense and domain-specific knowledge graphs. A natural question arises that how to effectively and efficiently manage a large-scale knowledge graph. Though systems and technologies that use relational storage engines or native graph database management systems are proposed, there exists no widely accepted solution. Therefore, a benchmark for management of knowledge graphs is required.

In this paper, we analyze the requirements of benchmarking knowledge graph management from a specific yet important point-of-view, i.e. characteristics of knowledge graph data. Seventeen statistical features of four knowledge graphs as well as two social networks are studied. We show that through these graphs depict similar structures, their tiny differences may result in totally different storage and indexing strategies, that should not be omitted. Finally, we put forward the requirements to seeding datasets and synthetic data generators for benchmarking knowledge graph management based on the study.

Keywords: Benchmark · Knowledge graph · Social network

1 Introduction

Recent years have witnessed an enthusiasm on the construction and application of large-scale knowledge graphs in both academia and industry communities. A knowledge graph can serve as the backbone of Web-scale applications such as query expansion [6], question answering [11] etc. For example, there are over 500 million entities and 3.5 billion facts in Google Knowledge Graph [18], which is employed to enhance Google search engine's search results. A natural question arises that how to efficiently manage such large-scale knowledge graphs. Though systems and technologies, such as using relational database management systems with specific indexing, or building native graph databases, are developed,

© Springer International Publishing Switzerland 2016
J. Zhan et al. (Eds.): BPOE 2015, LNCS 9495, pp. 37–49, 2016.
DOI: 10.1007/978-3-319-29006-5_4

it is still a research problem. Therefore, a benchmark for knowledge graph management is required for better understanding the problem of knowledge graph construction and utilization, and helping users to select appropriate systems or technologies while using knowledge graphs.

Existing benchmarks for graph data management, such as LinkBenck from Facebook [2], Social Network Benchmark (SNB) from LDBC [8] and our previous work BSMA [16], are mainly designed for social graph management applications. Note that the inherent difference between knowledge graphs and social networks is that the vertices and edges in knowledge graphs are usually typed or annotated with rich attributes or semantic labels, while in social graphs, the number of semantic labels of vertices and edges is limited. We argue that these benchmarks should not be used for benchmarking knowledge graph management, not to mention that they are accessed by different kinds of workload.

In this paper, we analyze the requirements for benchmarking knowledge graph management from a basic yet important point-of-view, i.e. statistical characteristics that depict structures of knowledge graphs. We show that the semantic nature of knowledge graphs results in different characteristics from knowledge graphs to social networks, as well as between different knowledge graphs, and different parts within a knowledge graph.

The contributions of this paper are as follows:

- Thirteen statistics and four statistical distributions are introduced for characterizing structures of graphs. Their intuitive meaning and effectiveness on managing graphs are analyzed.
- Studies over four knowledge graphs, including both common-sense knowledge graphs and domain-specific ones, and two social networks, are conducted. The empirical studies show that though these graphs are similar in certain structure characteristics, such as power-law distribution of vertex degrees, they are different in other features. Detailed analysis on how and why these differences exist is provided.
- The requirements for seeding datasets and synthetic data generator for benchmarking knowledge graphs are analyzed. We show that the synthetic data generator cannot be trivially adapted to an existing social network generator.

The rest of this paper is organized as follows. In Sect. 2 we introduce the related works on statistical characteristics and benchmarks in large scale graph. In Sect. 3, we introduce the statistical metrics used to evaluate the graphs. In Sect. 4 we describe four knowledge graphs and two social networks we experiment on. In Sect. 5, we present our empirical studies on this issue. Finally, we conclude with a summary and propose the future work in Sect. 6.

2 Related Work

A benchmark for knowledge graph management requires a clear understanding of the statistical characteristics of knowledge graphs. Research works which are focused on analyzing graph structural properties such as complex network, have

proposed structural metrics and distributions such as *node degree*, *hop* and *diameter* for modelling the structural properties of a graph. Benchmarks on big graph systems benefit a lot from them. For example, a benchmark generates synthetic data according to the distributions or topology of a graph. In this section, we give a brief introduction to these research fields.

For the past years, researchers have devoted themselves to analyzing structural properties of large scale graphs. Broder et al. [4] study the web as a graph via a series of graph structural metrics such as *diameter*, *nodes* and *degree*. For social network, Kumar et al. [12] study the evolutions of Flickr and Yahoo! 360 by analyzing their dynamic time-graph's structure properties, for example, *diameter*, *degree*, *community size*, etc. Boccaletti et al. [3] survey the studies of the structure and dynamics of complex network.

At the same time, benchmarks for big graph analytical have also been developed rapidly in recent years. Lancichinetti et al. [13] propose a benchmark for graphs, which pays attention to the heterogeneity in the distributions of node degrees and community sizes. As for social networks, Linkbench [2] characterizes the Facebook graph workload and constructs a realistic synthetic benchmark. Social Network Benchmark (SNB) from LDBC [8] models a synthetic social network which is similar to Facebook. It is the first LDBC benchmark based on the *choke-point* analysis [8], which identifies the technique challenges to evaluate in a workload. BSMA [16] is another applicable benchmark aimed to analyze the social media data based on Sina Weibo (microblog) in China.

The research works of the graph structural properties and benchmarks inspire us to have a close observation of the knowledge graphs' characteristics in order to study the problems of designing a benchmark for knowledge graphs from the point of statistical characteristics.

3 Statistical Characteristics

Both knowledge graphs and social networks can be modelled as a directed graph $G = (V, E)$, where V is the set of nodes (entities or users) and E is the set of directed edges (semantic relations). The thirteen statistics are illustrated in Table 1 and the four distributions are introduced as follows:

Degree Distribution. The degree distribution of G is $p(d) = \frac{n_d}{|V|}$, where n_d is the number of nodes whose degree is d and $|V|$ donates the number of nodes in G. In many graphs, the degree exhibits a *power-law* distribution [5] which has the form: $p(d) \propto L(d)d^{-\alpha}$, where $\alpha > 1$ and $L(d)$ is a slowly varying function. We study in-degree and out-degree separately in this paper.

Distribution of Hops. For a path $P = \{v_1, v_2, ..., v_h\}$ in G. The hop of a path P is defined as $Hops(P) = h - 2$, where h is the number of nodes in P. The distribution of hops reflects the connectivity cost inside a graph.

Distribution of Connected Components. There are *strongly* and *weakly* connected component in graph theory. A strongly connected component is a

Table 1. Description of statistical characteristics

Statistics	Description
#Nodes	Number of nodes
#Edges	Number of edges
#Density	The sparsity of a graph, which is formulated as $D(G) = \frac{\|E\|}{\|V\|(\|V\|-1)}$
#ZIDNodes	Number of nodes with zero in-degree
#ZODNodes	Number of nodes with zero out-degree
#BiDirEdges	Number of bidirectional edges
#CTriads	Number of closed triangles. A closed triangle is a trio of vertices each of which is connected to both the other two vertices.
#OTriads	Number of open triangles. An open triangle is a trio of vertices each of which is connected to at least one of the other two vertices.
AvgCC	Average clustering coefficient. The average clustering coefficient of a graph is defined as $C = \frac{3 \times \#Closed\ triads}{\#Open\ triads}$ [19].
FMWcc	Fraction of nodes in max weakly connected component
FMScc	Fraction of nodes in max strongly connected component
AppFdiam	Approximately full diameter
90 %EffDiam	The 90 percentile effective diameter, measures minimum number of hops in which 90 % of all connected pairs of nodes in a graph are reachable.

community in which any pair of nodes are reachable. A weakly connected component is a set of nodes in which any two nodes are reachable regardless of the edges' direction. The connected components reflects the connectivity of a graph.

Distribution of Clustering Coefficient. The definition of the clustering coefficient w.r.t. a node v_i is $C_i = \frac{|\{e_{jk}: v_j, v_k \in N_i, e_{jk} \in E\}|}{|N_i|(|N_i|-1)}$ [20], where e_{jk} is the edge between v_j and v_k $(j \neq k)$ and N_i is the set of neighbour nodes of v_i. The clustering coefficient measures the nodes' tendency to cluster together.

4 Data Description

In this section, we describe two social networks and four knowledge graphs we study in this paper.

Social Network. Sina Weibo is a famous social media which provides microblog service in China. In this paper, we generate two graphs consisting of *persons* and *fellowship* relations. In the first graph, we generate 0.2 million users **randomly** (SNRand) from the entire user set, which have more than 5 million relations between them. In the second graph, we select 0.2 million **most active users** including 36 million edges among them (SNRank). Note that the graphs are neither synchronized nor complete. However, as the comparison to knowledge

graphs, the SNRand can simulate the real-life data and SNRank can simulate the most critical situation where activities in social networks are very intense.

WordNet [9] is a lexical network for the English language designed in Princeton University. In WordNet, English words are grouped into sets of cognitive synonym (e.g. nouns, verbs, adjectives and adverbs), in which every synonym stands for a distinct concept. We utilize the real-life *WordNet* directly in our experiment. The words or concepts are nodes and semantic relations are edges.

YAGO2 [10] is a huge semantic knowledge graph which harvests knowledge from *WordNet*, *Wikipedia* and *GeoNames*[1]. We generate three subgraphs from YAGO2, named *YagoTax*, *YagoFact* and *YagoWiki*. *YagoTax* is the taxonomy of YAGO2, consisting of *subClassOf* relations and reflecting the taxonomic knowledge. *YagoFact* contains all the *factual* relations of YAGO2, standing for the factual knowledge. *YagoWiki* consists of the hyperlink relations (*linkedTo*) in YAGO2, reflecting the natural hyperlink structure of Wikipedia.

DBpedia [14] is a multi-language knowledge base extracted from the Wikipedia. The English version of DBpedia describes 4.58 million entities and 2,795 different properties. It utilizes the mapping-based technique to extract facts from Wikipedia info-boxes. We generate the synthetic data *DBpediaFact* from all the factual knowledge of DBpedia.

Enterprise Knowledge Graph(EKG) is a *domain specific knowledge graph* constructed by us. It models the relationships among people, companies, products for customer relation management (CRM). We extract relations from 2 million news articles from Sina Finance News[2] using a bootstrapping strategy similar to *snowball* [1] to iteratively detect relation tuples from entities.

Note the fact that all the knowledge graphs data we generated are real-life, which makes our empirical studies on these graphs more convincing.

5 Empirical Studies

In this section, we evaluate the graphs with a toolkit SNAP [15] and conduct an association rule mining experiment to conduct a series of empirical studies.

5.1 Analysis for Statistics

We analyze the graphs in three groups based on different objectives: 1) In order to study the different aspects of a knowledge graph, we compare the three subgraphs of YAGO2 and make an in-depth analysis of them. 2) We take the four knowledge graphs into consideration and make a series of horizontal comparisons, trying to reveal the differences between each knowledge graph in detail. 3) We compare the knowledge graphs with social networks, attempting to explain why and how the

[1] http://www.geonames.org/.
[2] http://finance.sina.com.cn/.

differences exist between them. The evaluation results are listed in Table 2. Those statistics are normalized by $\#Nodes$ in order to make them be comparable. Notice that the symbol # before the statistics in Table 1 are replaced by % in Table 2 and only #Nodes and #Edges are retained.

Table 2. Normalized statistics of graphs

Statistics	YagoTax	YagoFact	YagoWiki	DBpedia	WordNet	EKG	SNRand	SNRank
#Nodes	4.49e+5	2.14e+6	2.85e+6	4.26e+6	9.79e+4	9.45e+3	2.00e+5	2.02e+5
#Edges	4.51e+5	3.99e+6	3.80e+7	1.44e+7	1.54e+5	1.21e+4	5.45e+6	3.68e+7
Density	2.02e-6	1.75e-6	9.38e-6	1.59e-6	3.21e-5	2.72e-4	2.72e-4	1.80e-3
%ZIDNs	0.958	0.706	0.184	0.461	0.056	0.240	0.128	0.003
%ZODNs	5.78e-5	0.215	0.010	0.198	0.492	0.515	0.010	0.011
%BDEdges	0.000	0.019	2.940	0.129	0.487	0.498	6.984	81.29
%CTriads	0.000	0.365	26.02	2.115	0.043	0.093	59.92	2,167
%OTriads	2,982	93.62	616.9	371.4	30.66	14.82	5.94e+4	2.26e+5
AvgCC	0.000	0.095	0.331	0.325	0.032	0.029	0.105	0.067
FMWcc	0.998	0.953	0.999	0.989	0.988	0.655	1.000	1.000
FMScc	0.000	0.006	0.778	0.051	0.204	0.162	0.854	0.985
AppFdiam	11.00	15.00	14.00	40.00	25.00	18.00	15.00	7.000
90 %EDiam	6.740	5.340	3.830	5.920	10.800	6.770	5.090	3.350

Comparison Between YAGO2's Subgraphs. The three subgraphs *YagoTax*, *YagoFacts* and *YagoWiki* describe three aspects of YAGO2 respectively. From Table 2 we can see, the $\%CTriads$ and $\%OTriads$ of the three subgraphs are different significantly. The $\%OTriads$ of *YagoTax* is highest while the $\%CTriads$ is 0.00. In *YagoFacts* and *YagoWiki*, the differences between $\%CTriads$ and $\%OTriads$ are relatively smaller. The *AvgCC* of *YagoWiki* is higher than *YagoTax* and *YagoFacts*, indicates that the nodes in *YagoWiki* are more likely to be clustered via the relation *linkedTo* than the taxonomic relation *subClassOf* in *YagoTax* and *factual* relations in *YagoFact*. The relative differences of the three subgraphs' $\%ZODNs$ are greater than that of $\%ZIDNs$ in general, shows that the *out-degree* distributions of the three subgraphs are more diverse than their *in-degree* distributions.

Comparison Between Knowledge Graphs. To make the comparisons among the four knowledge graphs meticulously, we further divide them into two groups according to their contents: **taxonomic level** (e.g. *YagoTax* and *WordNet*) and **factual level** (e.g. *YagoFact*, *DBpediaFact* and *EKG*).

In **taxonomic level** group, the $\%ZIDNs$ and $\%ZODNs$ of *YagoTax* show almost every node in *YagoTax* has *out-degrees* while 95.8 % of the nodes have no *in-degree*. Because the *YagoTax* is a taxonomy tree with few hierarchies and tremendous unconnected leaves. In *WordNet*, the $FMScc$ shows that 20.4 %

of the nodes are in the max *strongly connected component* and the $\%ZODNs$ shows half of the nodes in *WordNet* have no *out-degree*. We imply the topology of *WordNet* is a star structure with a max *strongly connected component* (20.4 % nodes) in the centre and the other half of nodes (49.2 %) are distributed outside.

In **factual level** group, the *YagoFact* and *DBpediaFact* are both *common-sense knowledge graphs* (CSKG) and *EKG* is a *domain-specific knowledge graph* (DSKG). We compare *YagoFact* and *DBpediaFact* first. The scale and *density* of *DBpediaFact* are greater than *YagoFact*, indicates that the automatically generated *DBpediaFact* [14] contains more entities than them in *YagoFact*, where the relation extraction method is based on hand-written rules [10]. The $\%CTriads$ and $\%OTriads$ show nodes in *DBpediaFact* are more likely to form *triangles* than *YagoFact*. As for the *CSKG* (*YagoFact*, *DBpediaFact*) and *DSKG* (*EKG*), the *density* of *DSKG* is higher than *CSKG*. The $\%CTriads$, $\%OTriads$ as well as $AvgCC$ show that the *DSKG* have more *triangles* than *CSKG*, too.

Comparison Between Knowledge Graphs and Social Networks. The *density* shows that social networks are denser than knowledge graphs. Due to the activeness of people, social networks contain more *bidirected* edges than knowledge graphs. The $\%CTriads$ and $\%OTriads$ of social networks are greater than knowledge graphs. Because in social networks, a person's friends tend to be friends due to the *triadic closure* property [7]. The $FMWcc$ and $FMScc$ of social networks show that most of nodes in social networks are in a max *strongly connected component*, while in knowledge graphs, most of the nodes tend to form *strongly connected components* within a small range. Another evidence from $FMWcc$ shows that the nodes in knowledge graphs are connected by a max *weakly connected component*. We conclude that the *strongly connected components* in knowledge graphs are connected by a series of *bridges* [19] (also known as cut-edge) actually. In short, there exist gaps between the *strong connected components* in knowledge graphs, while the social network is a whole *strong connected component*.

Conclusion. Table 3 summarizes the statistics which have significant different performances in the three comparison groups. As we can see, the differences between the subgraphs of YAGO2 in the first group are mainly embodied in (*open or close*) *triangle*, *clustering co-efficient* and *strongly connected component* three aspects. Then we can not treat the knowledge graph as a whole graph when generating the synthetic data. We should generate it separately according to different parts or semantic topics.

In the second group, the *density*, $\%CTriads$, $\%OTriads$, $AvgCC$, $\%ZIDNs$ and $\%ZODNs$ are the prominent statistical characteristics which perform diversely between each knowledge graphs on both *taxonomic* and *factual* levels. It implies that when generating synthetic data or designing workloads for a knowledge graph in a special domain, the data characteristics should be considered first and the workloads should emphasize these characteristics in special.

Table 3. The prominent statistics which are different in each comparison group

Experiment groups		Density	%CTriads	%OTriads	AvgCC	FMScc	%BDEdges	%ZDNs
YAGO subgraphs			√	√	√	√		
KGs	Taxonomic	√	√	√				√
	Factual	√	√	√	√			
KGs and SNs		√	√	√		√	√	

[1] %ZDNs donates for %ZIDNs and %ZODNs.
[2] KGs and SNs are short for the "Knowledge Graphs" and "Social Networks".

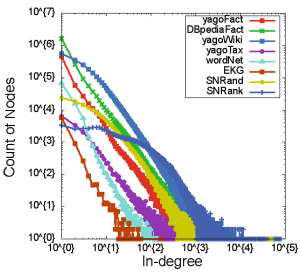

Fig. 1. Dist. of indegree

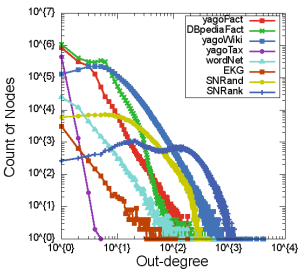

Fig. 2. Dist. of outdegree

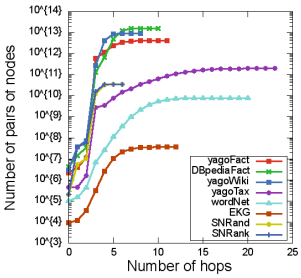

Fig. 3. Dist. of hops

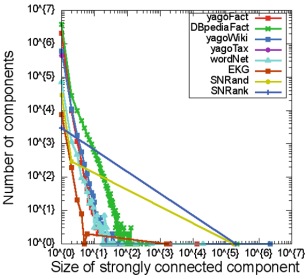

Fig. 4. Dist. of SCC

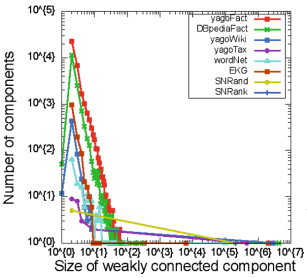

Fig. 5. Dist. of WCC

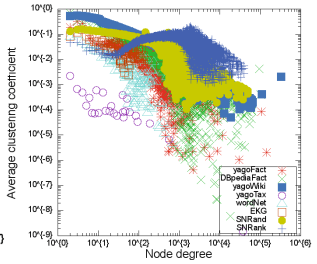

Fig. 6. Dist. of ACC

In the last group, the *density*, *triangle*, *strongly connected component* and *%BDEdges* are the main characteristics which perform different between knowledge graphs and social networks. There are few bidirectional edges in knowledge graphs (but not none). Thus data generator should control their existences appropriately in knowledge graphs. The gaps between *strong connected components* in knowledge graphs remind us that more facts should be extracted and added to the existing knowledge graphs to bridge them in the future. And with the development of them, the benchmarks for knowledge graph data management techniques should focus on different properties dynamically.

5.2 Analysis for Distributions

In this section, we give an in-depth analysis of the graphs' distribution metrics. The distributions are illustrated in Figs. 1, 2, 3, 4, 5 and 6.

Table 4. Fitted parameters for all the distributions

Graphs	InDeg		OutDeg		Hop			SCC		WCC	
	$L(d)$	α	$L(d)$	α	a	b	c	$L(d)$	α	$L(d)$	α
yagoFact	2.07e+5	1.859	4.80e+5	2.245	4.00e+12	5.71	1.22	2.66e+5	4.940	2.13e+5	3.228
YagoTax	1.70e+4	1.855	4.48e+5	8.412	1.95e+11	5.45	0.47	-	-	1.98e+1	1.063
yagoWiki	3.92e+6	1.914	5.61e+8	3.000	9.40e+12	7.84	1.89	1.20e+5	3.849	1.48e+2	0.597
DBpedia	6.77e+5	1.827	2.08e+14	7.697	1.58e+13	10.62	1.96	1.30e+5	2.361	1.55e+5	3.769
WordNet	2.86e+4	2.455	7.86e+4	2.379	7.60e+10	21.40	2.03	7.63e+3	2.745	2.54e+2	2.045
EKG	1.14e+3	1.913	2.96e+2	2.719	3.64e+7	7.49	1.31	7.41e+3	5.236	1.39e+4	3.834
SNRand	5.27e+5	1.884	2.44e+5	1.441	3.55e+10	10.29	3.20	2.85e+4	6.555	-	-
SNRank	1.95e+5	1.293	3.12e+7	2.078	3.43e+10	14.39	4.72	-	-	-	-

[1] The "-" represents that the distribution's points are not enough to fit the parameter.
[2] Due to the divergency of the points, the *average clustering co-efficient* distributions are not fitted.

In Fig. 1, the *in-degree* distributions of all knowledge graphs and social networks exhibit the power-law distributions. The estimated parameters are detailed in Table 4. Nearly all the exponents α are consistently around (1.8,2.4) except that of *SNRank*, which is different with all the others. The initial segment of *SNRank* distribution deviates from the power law greatly until the *in-degree* increases up to around 560.

The *out-degree* distributions are shown in Fig. 2. As we can see, there exist not only significant distinctions between knowledge graph and social network but also between knowledge graphs. The *out-degree* distributions of the three YAGO2 subgraphs are different significantly, which is consistent with the analysis in previous section. All the distributions are deviated from power law initially, and they are diverse with each other as well. The descent rates of the distributions also vary widely. The fitted parameters α in Table 4 fluctuate from 1.4 to 8.4.

Figure 3 illustrates distributions of *hops* in those graphs. All the distributions are in "**S**" shape. In order to fit the data to some curve, we introduce a variant of *sigmoid* function with the form $f(x) = \frac{a}{1+e^{b-cx}}$. The parameters are fitted very well in Table 4. The max hops of *yagoTax* and *WordNet* are larger than the others in general. The *SNRand* and *SNRank* have the minimal max *hops*, close to 6, which is in consistent with *six degrees of separation* theory [20] in social networks. Another interesting discovery is that with the hop added from 2 to 3, all the distributions increase explosively in general.

Figures 4 and 5 reflect the distributions of *connected components*. Both the distributions of *strongly* and *weakly connected components* of knowledge graphs are in power-law distribution uniformly except *yagoTax* which is a flat tree with a lot of unconnected leaves. While in *SNRand* and *SNRank* as social networks, there only have one max *strongly connected component* and a small part of isolated nodes. The pow-law distributions of *strongly* and *weakly connected components* in knowledge graphs show that the nodes in knowledge graphs are clustered in several small ranges (actually most of the *strongly connected components* are connected by *bridges* according to our analysis in previous section). However, nodes in social networks are organized into one big *strongly connected component*.

Figure 6 presents the distributions of *average clustering coefficient* (ACC). In this experiment setting, we treat the graphs as undirected and the *degree* is the total of *in-degree* and *out-degree*. Figure 6 shows, all the other graphs are trend to perform the power law initially except *SNRank*, with the value of x increases, the curves start to diverge. The ACC of social networks is higher than knowledge graphs in general, which also reflects that the *local clustering property* of knowledge graphs is not as strong as social networks.

Table 5. Relatedness of *in* and *out* relations in YAGO2

InRelation1	InRelation2	$R(r1, r2)$	OutRelation1	OutRelation2	$R(r1, r2)$
rdf:type	subClassOf	0.9948	rdf:type	hasWikipediaUrl	0.9999
playsFor	isAffiliatedTo	0.9797	linksTo	hasWikipediaUrl	0.9894
hasChild	isMarriedTo	0.9182	rdf:type	linksTo	0.9815
wasBornIn	isLocatedIn	0.8712	exports	imports	0.9220
graduatedFrom	worksAt	0.7545	playsFor	isAffiliatedTo	0.7668
created	directed	0.7456	imports	dealsWith	0.7475
actedIn	created	0.6943	exports	dealsWith	0.7454
diedIn	wasBornIn	0.6730	imports	hasTLD	0.6553
wroteMusicFor	directed	0.6041	hasTLD	dealsWith	0.6334
isCitizenOf	dealsWith	0.5590	isConnectedTo	hasAirportCode	0.5984

Conclusion. Figures 1 and 2 illustrate that the *in-degree* and *out-degree* distributions of knowledge graphs and social networks are of great differences. The initial segment of the *out-degree* distributions follow a different kind of distributions (e.g. *poisson* or a combination of *poisson* and *power-law*). Users can generate the synthetic data sectionally according to the different *in* or *out degrees*. Figure 3 shows the distributions of *hops* that exhibit the "**S**" shape distribution, which fits a *sigmoid* function very well. Users can utilize the *sigmoid* function to generate the data. Figures 4 and 5 illustrate that the knowledge graphs are separated naturally into a number of *strongly connected components* and isolated *weakly connected components*, in which the size of the components displays power law distributions. The natural partition ability of the knowledge graph allows us to manage the data distributively, which will potentially reduce the cost of *join* operations significantly. However, in social networks, users should divide them by some special graph partition algorithms. Figure 6 shows the distributions of *average clustering co-efficient* are in the power law distribution, but with the incasement of *degree*, the *clustering co-efficient* starts to diverge, indicating the data generator should not only obey the power law distribution in overall but also embody the data divergence at the same time.

5.3 Analysis for Labels' Relatedness

We conduct an association rule experiment on semantic labels of YAGO2 as a case study to discover the relatedness between labels.

For a relationship, there are two kinds of relations, namely, *out* and *in* respectively. To compute the relatedness between labels r_i and r_j, we first define the support of $r_i \rightarrow r_j$ as $supp(r_i \rightarrow r_j) = \frac{|r_i \cap r_j|}{|r_i|}$, which is inspired by the definition in [17], where $|r_i|$ denotes the number of nodes that have a relation r_i and $|r_i \cap r_j|$ denotes the number of nodes that have both relations r_i and r_j. Obviously, the *support* function is not symmetric, inspired by the definition of F-measure, the relatedness of r_i and r_j is defined as $R(r_i, r_j) = \frac{2 \times supp(r_i \rightarrow r_j) \times supp(r_j \rightarrow r_i)}{supp(r_i \rightarrow r_j) + supp(r_j \rightarrow r_i)}$.

Table 5 lists top-10 pairs of relations with highest relatedness. From Table 5, we find the semantic relations are topic related. For example, *hasChild* and *isMarriedTo* indicate children and marriage belong to the same topic. Some semantic relations have no intersection with others. For example, the relatedness between *hasGender* and *isLocatedIn* is 0.00. However, some relations (e.g. *rdf:type*) almost have co-occurrence with any other relations, which indicates the semantic labels are distributed differently due to the semantics. Thus, in the point of designing benchmarks, we conclude the data generator for knowledge graphs cannot be trivially adapted to an existing social network generator. Note that in many information extraction systems, new kinds of relations should be extracted easily by many automatic methods, for example, the *snowball* [1] system learns new generated patterns to discover new tuples.

6 Conclusion and Discussion

In this paper, we observe four knowledge graphs and two kinds of social networks closely on their statistical characteristics. After our in-depth analysis on the experiments, it is shown that:

(1) Different parts of a knowledge graph have different properties in some certain statistical characteristics, such as *clustering co-efficient*, *strongly connected component*.
(2) The different types of knowledge graphs have different properties in several statistical characteristics, and their data distributions are different either, such as *out degree distributions* and *hops distributions*.
(3) Knowledge graphs are different with social networks in several distributions, such as *strongly* and *weakly connected components*. With their development, new kinds of relationships could be discovered easily.

These empirical observations show that the existing synthetic graph generators can not generate the real-life knowledge graph data. Therefore, to build a benchmark for testing and evaluating the different knowledge graph management systems and techniques, an important task is to build generators that meet the following requirements:

- The generator should generate synthetic data of a knowledge graph in different aspects, such as *taxonomic knowledge* and *factual knowledge*.
- The generator should take the semantic labels in knowledge graphs into consideration and preserve the statistical characteristics of the real-life data.
- The generator should be able to not only generate the static synthetic data of a certain knowledge graph, but also the different stages of knowledge graph's construction.

The empirical study conducted in this paper is the first effort on modelling statistical characteristics of knowledge graphs. Our future work includes the design and implementation of the data generators of knowledge graphs.

Acknowledgement. This work is partially supported by National Hightech R&D Program (863 Program) under grant number 2015AA015307, and National Science Foundation of China under grant numbers 61432006 and 61170086. The authors would also like to thank Ping An Technology (Shenzhen) Co., Ltd. for the support of this research.

References

1. Agichtein, E., Gravano, L.: Snowball: extracting relations from large plain-text collections. In: DL 2000, pp. 85–94. ACM (2000)
2. Armstrong, T.G., Ponnekanti, V., Borthakur, D., Callaghan, M.: Linkbench: a database benchmark based on the facebook social graph. In: SIGMOD, pp. 1185–1196. ACM (2013)
3. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.-U.: Complex networks: structure and dynamics. Phys. Rep. **424**(4–5), 175–308 (2006)
4. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the web. Comput. Netw. **33**(1–6), 309–320 (2000)
5. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. SIAM Rev. **51**(4), 661–703 (2009)
6. Dalton, J., Dietz, L., Allan, J.: Entity query feature expansion using knowledge base links. In: SIGIR, pp. 365–374. ACM (2014)
7. David, E., Jon, K.: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, New York (2010)
8. Erling, O., Averbuch, A., Larriba-Pey, J., Chafi, H., Gubichev, A., Prat, A., Pham, M.-D., Boncz, P.: The ldbc social network benchmark: interactive workload. In: SIGMOD, pp. 619–630. ACM (2015)
9. Fellbaum, C. (ed.): WordNet: an electronic lexical database. MIT Press, Cambridge (1998)
10. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: Yago2: a spatially and temporally enhanced knowledge base from wikipedia. Artif. Intell. **194**, 28–61 (2013)
11. Joshi, M., Sawant, U., Chakrabarti, S.: Knowledge graph and corpus driven segmentation and answer inference for telegraphic entity-seeking queries. In: EMNLP, pp. 1104–1114. ACL (2014)
12. Kumar, R., Novak, J., Tomkins, A.: Structure and evolution of online social networks. In: KDD, pp. 611–617. ACM (2006)

13. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E **78**(4), 046110 (2008)
14. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia: a large-scale, multilingual knowledge base extracted from wikipedia. Semant. Web J. **6**(2), 167–195 (2015)
15. Leskovec, J., Sosič, R.: SNAP: a general purpose network analysis and graph mining library in C++, June 2014. http://snap.stanford.edu/snap
16. Ma, H., Wei, J., Qian, W., Yu, C., Xia, F., Zhou, A.: On benchmarking online social media analytical queries. In: GRADES, p. 10 (2013)
17. Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press, New York (2011)
18. Singhal, A.: Introducing the knowledge graph: things, not strings. Official Google Blog, May 2012
19. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications, vol. 8. Cambridge University Press, New-York (1994)
20. Watts, D., Strogatz, S.: Collective dynamics of 'small-world' networks. Nature **393**, 440–442 (1998)