On Statistical Characteristics of Real-life Knowledge Graphs

Weining Qian

Institute for Data Science and Engineering
East China Normal University
wnqian@sei.ecnu.edu.cn

NSFC key project on Big Data Benchmarks

- Theory and Methods of Benchmarking Big Data Management Systems
  - 2015.1 – 2019.12

East China
Normal University

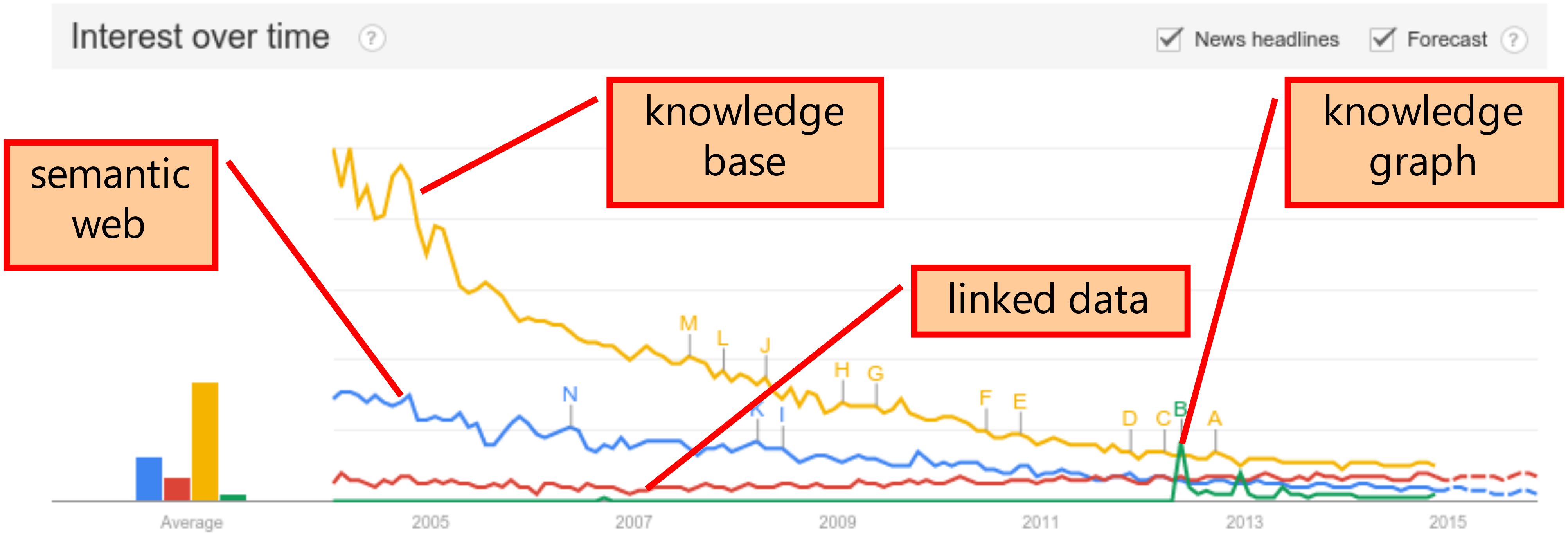
Renmin University
of China

Institute of Computing Technology
Chinese Academy of Sciences

Research focuses

- BigDataBench Suite (@ict.ac)
  - http://prof.ict.ac.cn
- Benchmarking transaction processing in NewSQL systems (@ecnu)
  - SecKillBench
- Benchmarking Big Data systems (@rmu)
- Benchmarks for graph data (@ecnu)
  - Social media, knowledge graphs, ...

Knowledge graphs



Some KG's

- YAGO
  - 10M entities in 350K classes
  - 120M facts for 100 relations
  - 100 languages
  - 95% accuracy
- DBPedia
  - 4M entities in 250 classes
  - 500M facts for 6000 properties
  - live updates
- Kosmix
  - 6.5M concepts, 6.7M concept instances,
  - 165M relationship instances
- Freebase
  - 40M entities in 15000 topics
  - 1B facts for 4000 properties
  - core of Google Knowledge Graph
- Google Knowledge Graph
  - 600M entities in 15000 topics
  - 20B facts
- Probase
  - 2.7 million+ concepts
- And many domain/application-specific knowledge graphs

A natural question

- Knowledge graph can serve as the backbone of many Web-scale applications, such as search engine, question answering, text understanding etc.
- How to effectively and efficiently manage a large-scale knowledge graph?
  - MySQL, Oracle, Neo4j, TITAN, Trinity, or other triple stores???

Social networks vs. Knowledge graphs

- Though there are some benchmarks for social networks exist
  - Facebook LinkBench, LDBC SNB, BSMA, ...
- Knowledge graph is different with social network
  - More semantic labels in both entities and relations
  - Topic or domain sensitive
  - Contains various kinds of knowledge
  - Hard to define a unified schema

Why study their statistical characteristics?

- To better understand knowledge graphs
- To help the selection of seeding data sets in benchmarks
- To help the development of data generators

Characteristics of large-scale graphs

- Previous research works on analyzing structural properties of large scale graphs, e.g.
  - [Broder et al. Comput. Netw. 2000] studied the web structure as a graph via a series of metrics, e.g degree, diameter, component.
  - [Kumar et al. KDD, 2006] studied the dynamic social network's structure properties, e.g. degree, hop etc.
  - [Boccaletti et al. Phys. Rep. 2006] surveyed the studies of the structure and dynamics of complex network.

Real-life knowledge graphs

- YAGO2
  - A huge semantic knowledge graph based on WordNet, Wikipedia and GeoNames
  - 10+ million entities, 120+ million facts
- Separate the YAGO2 into three sub-graphs
  - YagoTax: Taxonomy tree of YAGO2
  - YagoFact: Facts in YAGO2
  - YagoWiki: Hyperlink relations in YAGO2 based on Wikipedia

Real-life knowledge graphs

- WordNet
  - A lexical network for the English language.
  - Synonym set as node and semantic relation as edge.
  - 98,000 entities, 154,000 relationships
- DBpedia
  - A multi-language knowledge base extracted from Wikipedia info-boxes
  - English version of DBPedia
  - 4.58 million things and 2,795 different kinds of properties

Real-life knowledge graphs

- Enterprise Knowledge Graph (EKG)
  - Describes an enterprise relationships in Chinese
  - Extracted from reports from enterprises in Shanghai Stock Market
  - Used for credit and risk analysis in financial companies
  - A domain specific knowledge graph
  - Seven kinds of relationships between two entities
    - Assignment, hold, subcompany, changename, **manager**, cooperate, merge
  - 51,853 entities and 430,973 relationships.

Plus two social networks

- SNRand
  - 0.2 million randomly selected users
  - 5 million fellowship relations between users
- SNRank
  - 0.2 million most active users.
  - 36+ million fellowship relations between users
- The raw data is collected from a famous social media platform named Sina Weibo in China

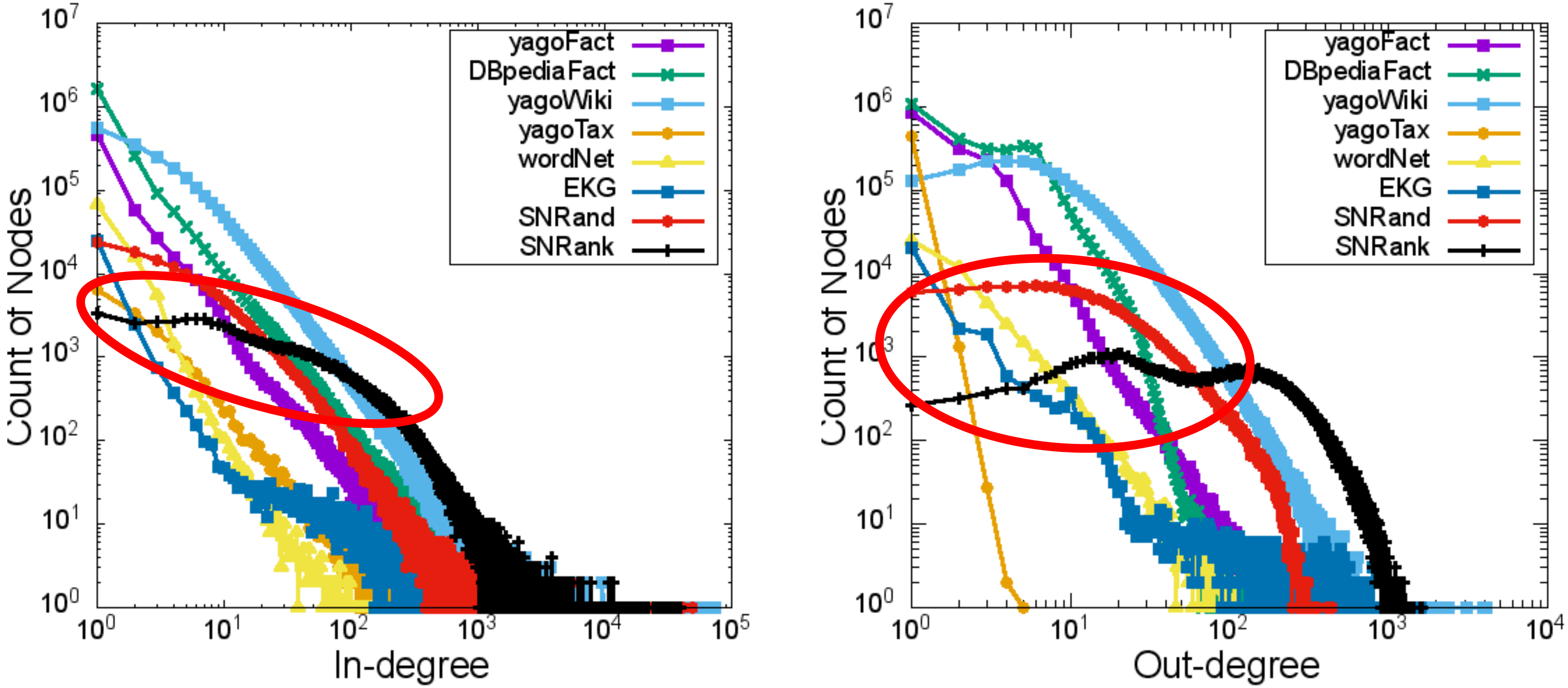
Statistical characteristics

Statistics	Description
#Nodes	Number of nodes.
#Edges	Number of edges.
#Density	The sparsity of a graph, which is formulated as $D(G) = \frac{\lvert E \rvert}{\lvert V \rvert (\lvert V \rvert - 1)}$
#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.

Four kinds of distributions

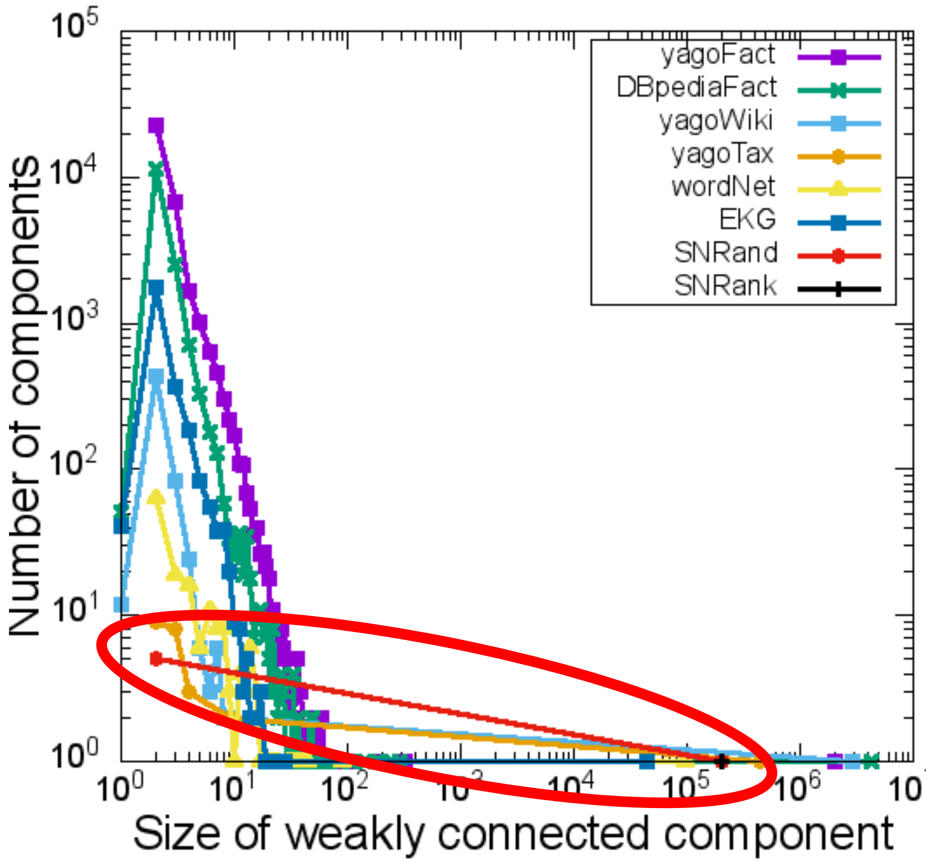
- Distribution of degrees
  - In-degree and out-degree
  - Power-law distribution
- Distribution of hops
  - Reflects the connectivity cost inside a graph
- Distribution of connected components
  - Strongly and weakly connected components
  - Reflects the connectivity of a graph
- Distribution of clustering coefficients
  - Measures the nodes' tendency to cluster together

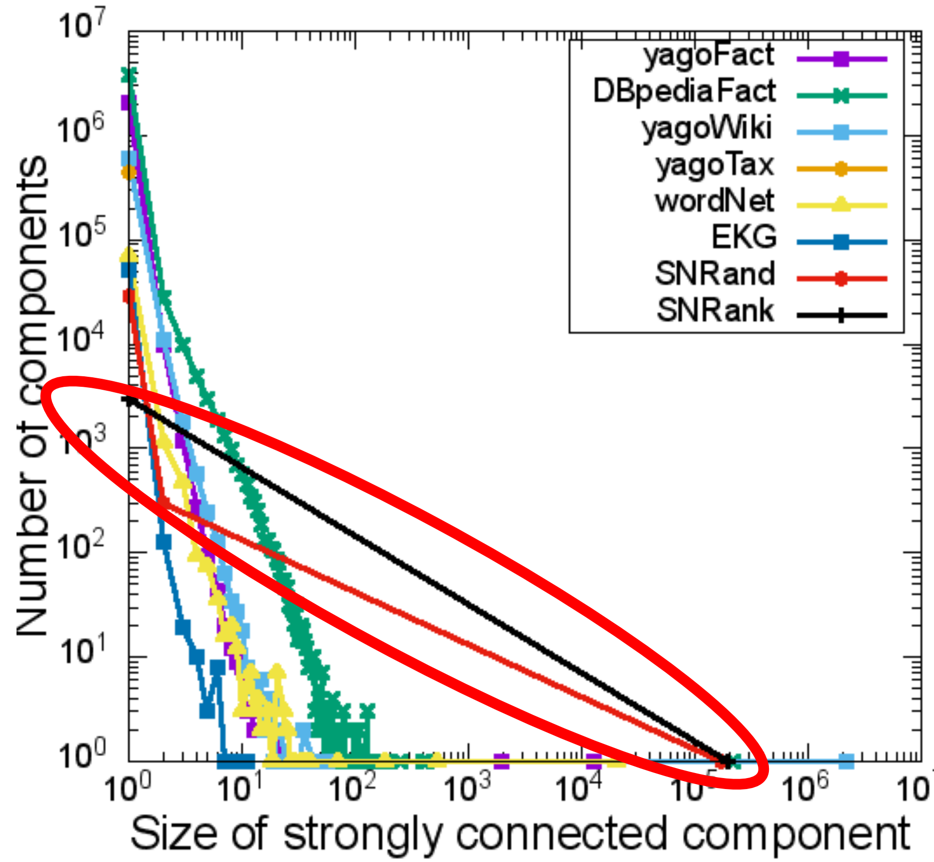
In-degrees and out-degrees



All the in/out-degree distributions exhibit the power-law (or piece-wise power-law), except for some initial segments that deviate the power-law.

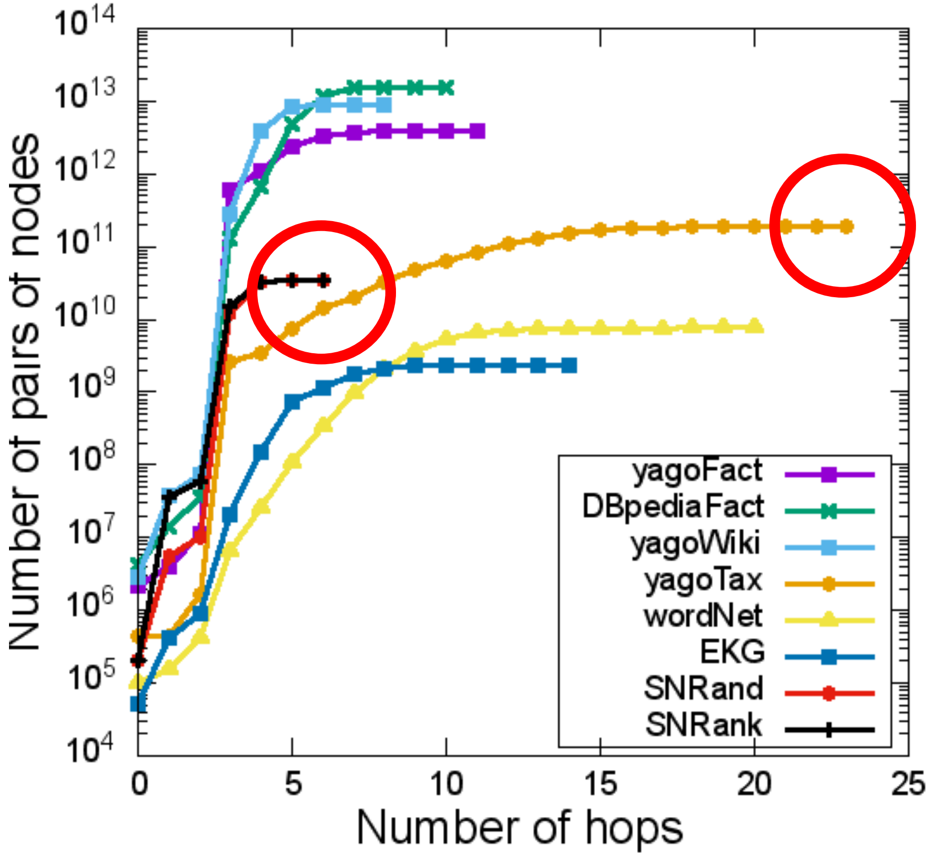
Size of connected components



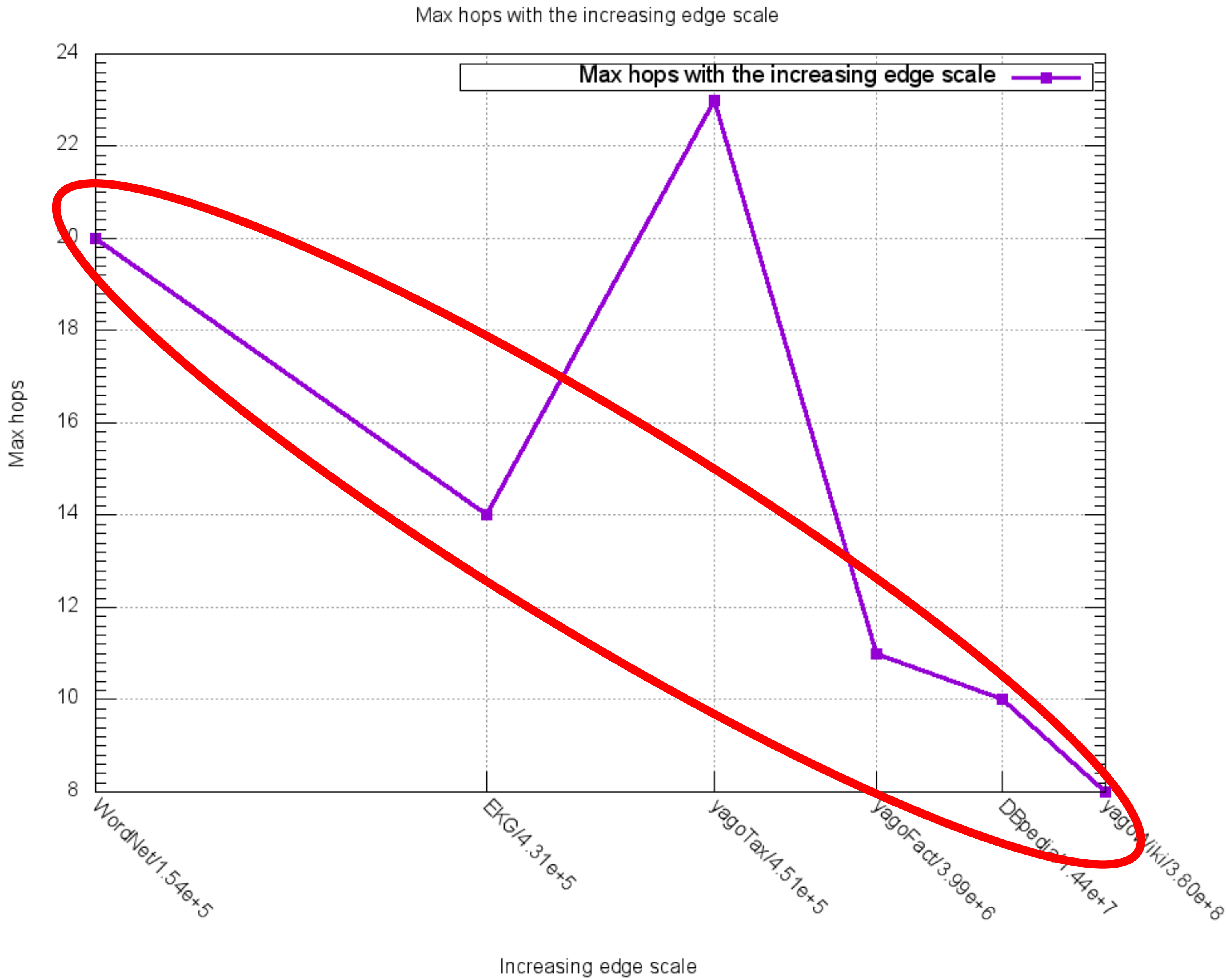


Both the strongly and weakly connected component distributions of knowledge graphs exhibit the power-law distribution in general. While the social networks are nearly in a whole strongly connected component.

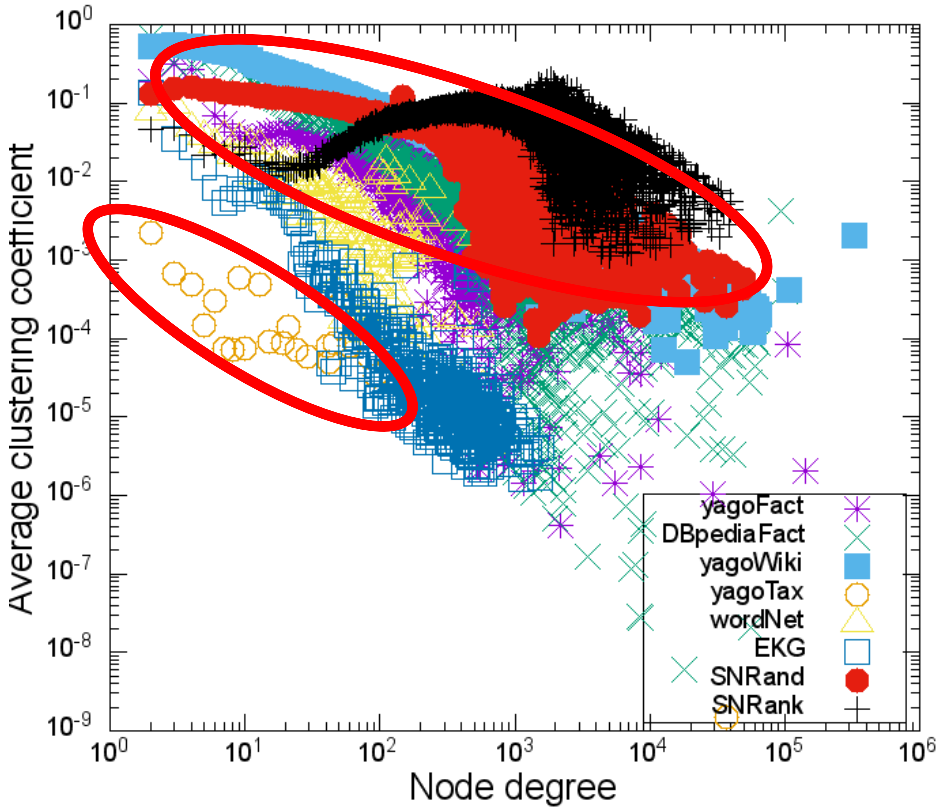
Distance between two vertices



- Social networks are "small worlds"
- Taxonomy's diameter is large (tree-alike)
- Basically, the larger the network is, the smaller the diameter is

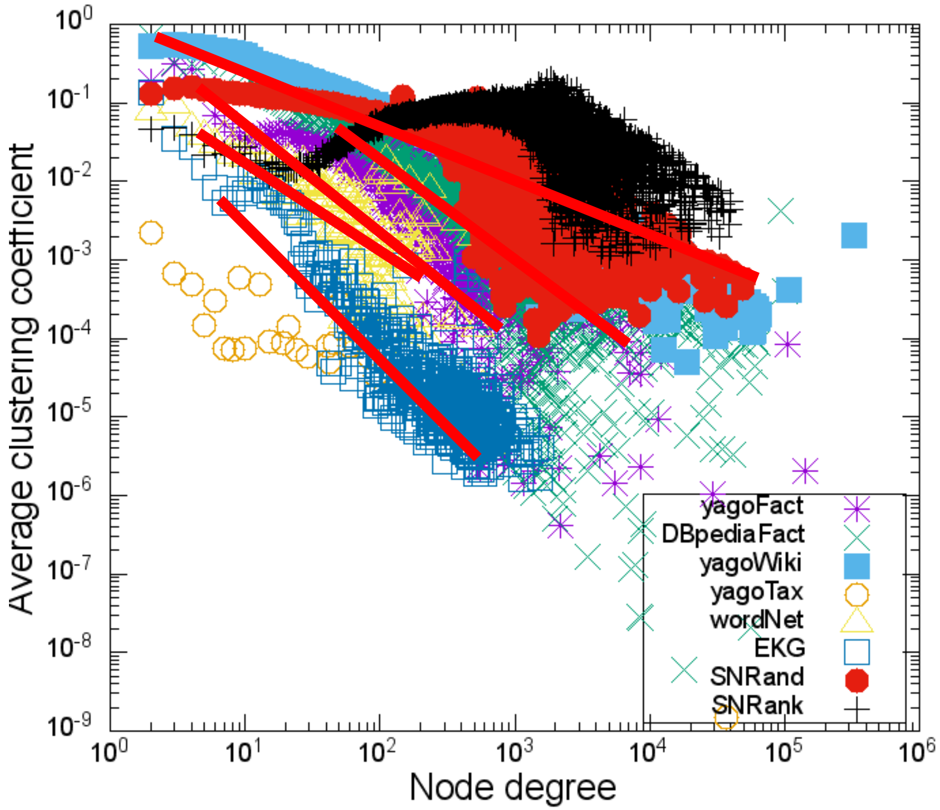

Max hops with the increasing edge scale
Max hops with the increasing edge scale
Max hops
24
22
20
18
16
14
12
10
8
WordNet/1.54e+5
EKG/4.31e+5
yagoTax/4.51e+5
yagoFact/3.99e+6
DBpedia/1.44e+7
yagoWiki/3.80e+8
Increasing edge scale

Cluster coefficient



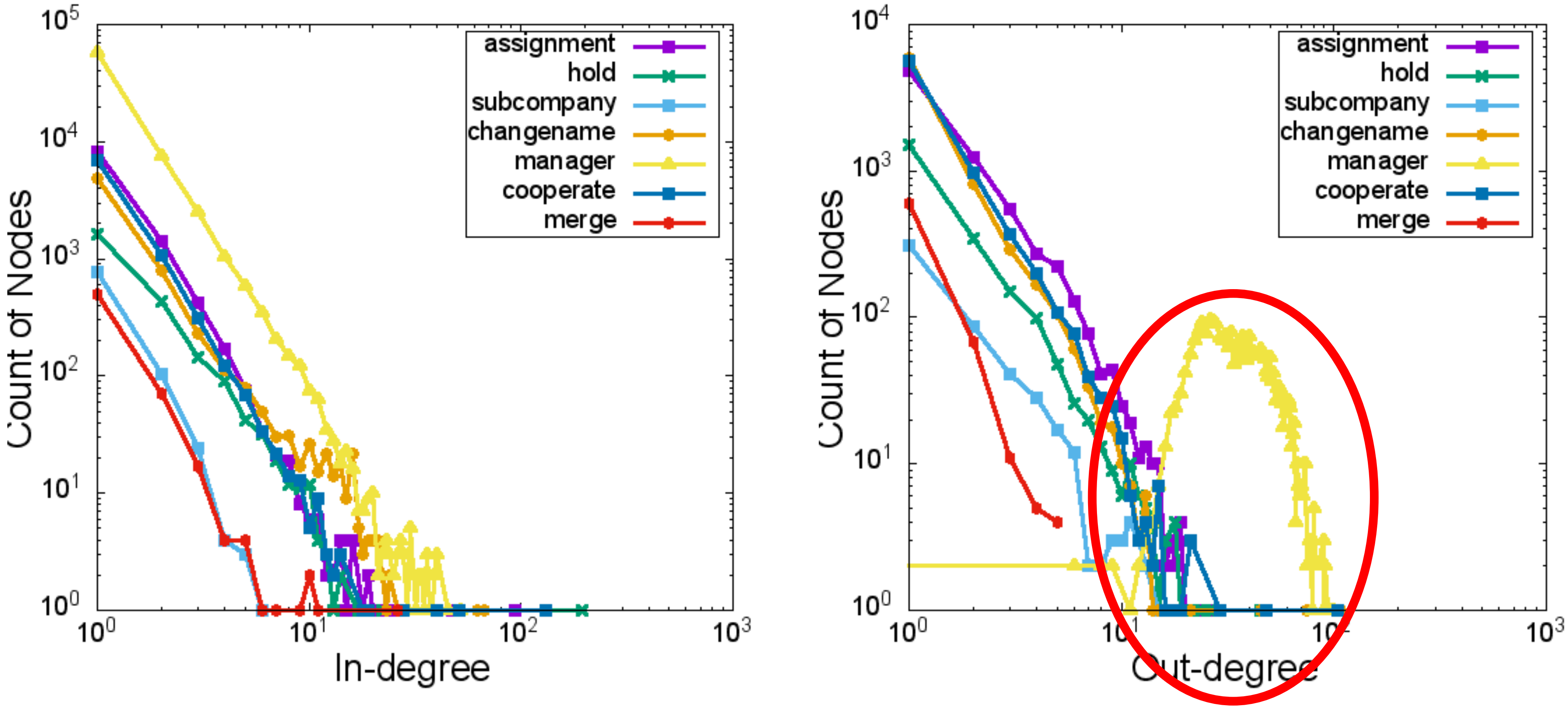
- Taxonomy and social networks are different
- All other KG's are of power-law distributions

Cluster coefficient



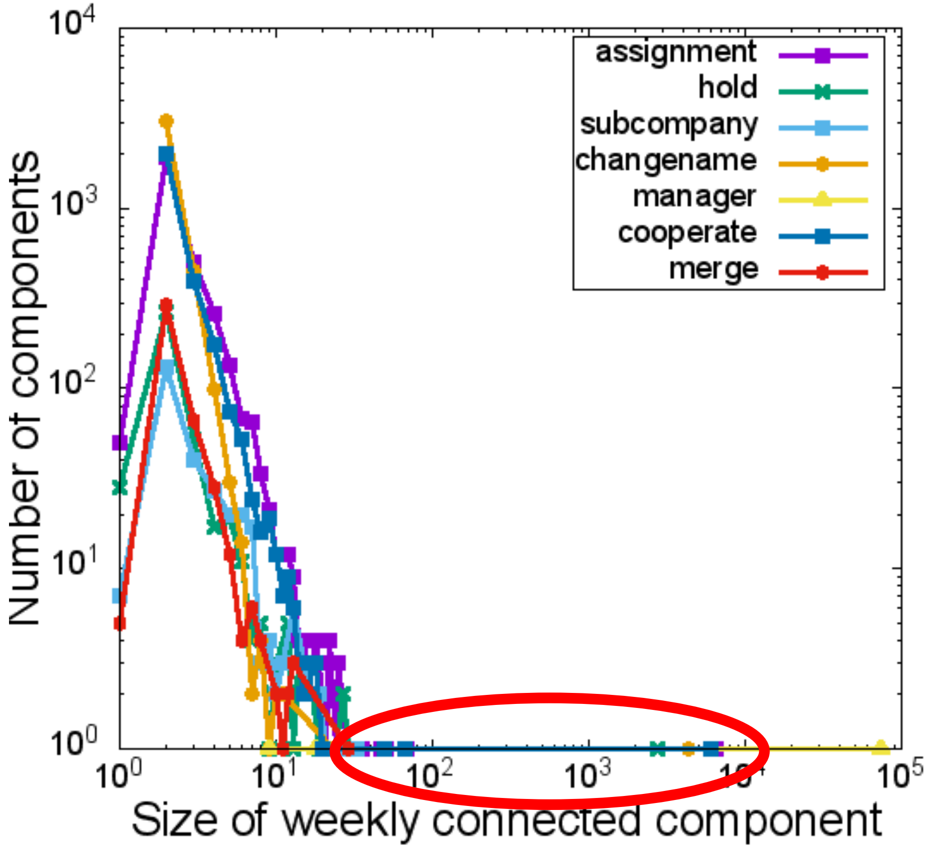
- Taxonomy and social networks are different
- All other KG's are of power-law distributions

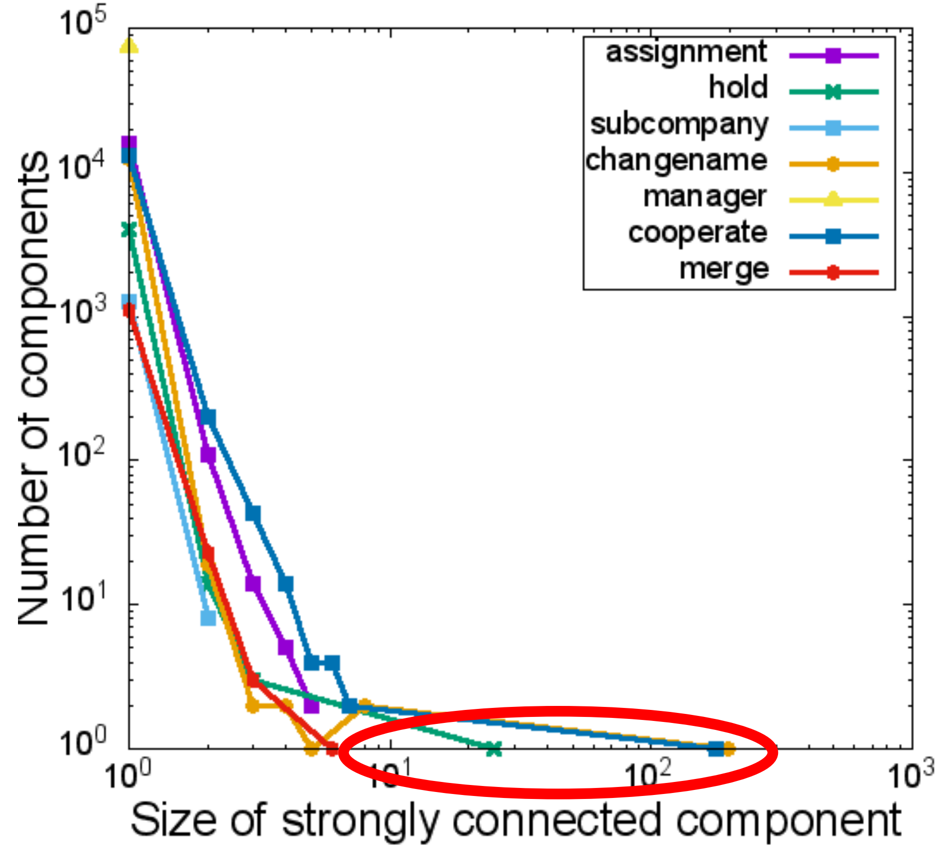
Node degrees of different parts in EKG



Different relationships show different out-degree distributions

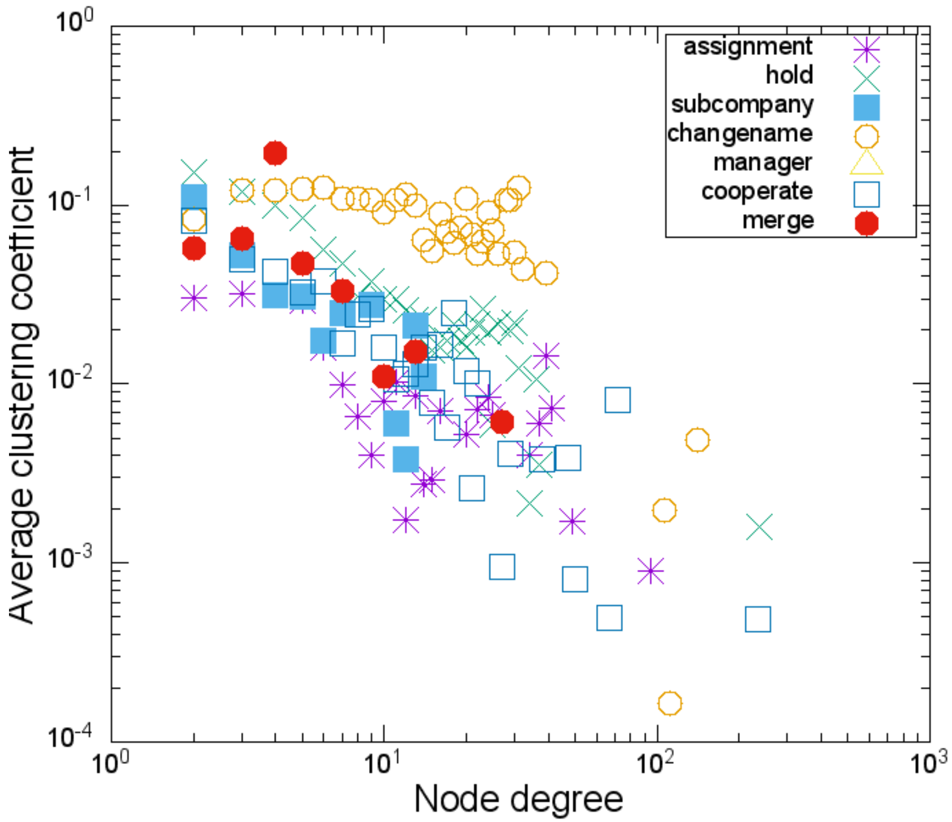
Size of connected components of different parts in EKG



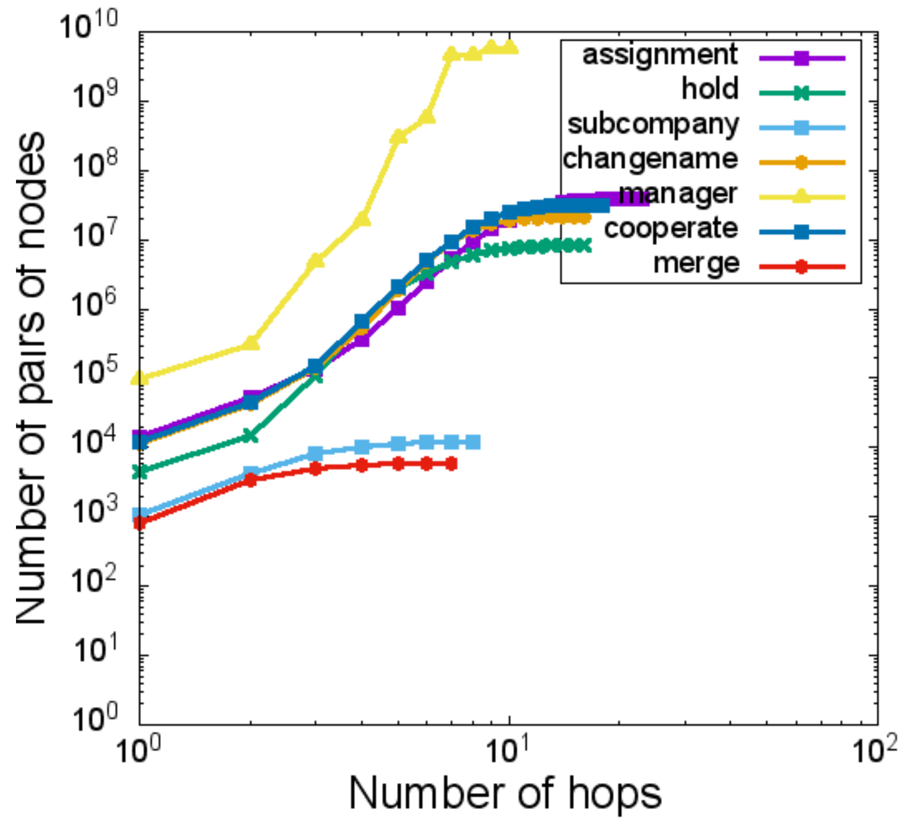


Few connected components are much larger than others

Different parts in EKG



Highly clustered



Small-world network

Conclusions and discussions

- KG's are different to SN's
  - Taxonomy/ontology + fact
    - Following different statistical distributions
  - KG's are labeled
    - Different subgraphs are of different sizes and characteristics
- Both triple stores and relational databases have reasons to be used
  - The key is to avoid joins over power-law distributed data

Wenliang Cheng, Chengyu Wang, Bing Xiao, Weining Qian, Aoying Zhou:On Statistical Characteristics of Real-Life Knowledge Graphs. BPOE 2015: 37-49

谢谢！

Thanks!

http://dase.ecnu.edu.cn

wnqian@sei.ecnu.edu.cn

Statistical characteristics

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