LOD Network Structure Analysis and Visualization Postprint

Authors: Xia Lixin, Tan Ying

Date: 2017-10-11T00:00:00+00:00

Abstract

[Objective] To conduct structural feature analysis on Linked Open Data (LOD) and guide linked data organization practices using the analysis results. [Method] The LOD network structure is described through metrics including degree distribution, average path length, and clustering coefficient, with comparative examination of two fundamental properties from complex network theory: scale-free characteristics and small-world effects. [Results] The overall LOD network structure exhibits power-law distribution characteristics approximating those of scale-free networks, whereas subnets in library and information science domains display relatively uniform exponential distribution characteristics; both networks concurrently demonstrate small-world effects characterized by short average path lengths and high clustering coefficients. [Limitations] Lack of multi-weight assignment for key nodes. Conclusion The small-world characteristics of LOD can optimize retrieval efficiency, while the scale-free characteristics may compromise overall network stability.

Full Text

Preamble

ChinaXiv Collaborative Journal, Issue 266, 2016, No. 1
Analysis and Visualization of the LOD Network Structure*
Xia Lixin, Tan Ying
(School of Information Management, Central China Normal University, Wuhan 430079)

Abstract

[**Objective**] This study analyzes the structural characteristics of Linked Open Data (LOD) to guide the practical organization of linked data. [**Methods**]

We describe the LOD network structure using metrics such as degree distribution, average path length, and clustering coefficient, comparing two fundamental properties from complex network theory: scale-free characteristics and small-world effects. [**Results**] The overall LOD network structure exhibits power-law distribution features approximating a scale-free network, while the library and information science subdomain shows a relatively homogeneous exponential distribution. Both networks simultaneously demonstrate small-world effects with short average path lengths and high clustering coefficients. [**Limitations**] The study lacks multi-weight assignment for key nodes. [**Conclusions**] The small-world characteristics of LOD can optimize retrieval efficiency, whereas scale-free features may reduce overall network stability.

Keywords: Linked Open Data; Complex network; Network structure; Visualization
Classification Number: G203

Introduction

An increasing number of data owners are publishing their data as linked data on the web, forming a global data space known as the Web of Data [1]. Compared to traditional document networks, the Web of Data is more structured, transforming simple hyperlinks into complex relationship networks that enable web data to be discovered, retrieved, and understood by both humans and machines. In August 2014, the W3C Linked Open Data project released the latest Linked Open Data Cloud diagram, establishing a visual model for the Web of Data. The illustrated open linked datasets grew from dozens to hundreds, covering eight domains: media, government, publications, geography, life sciences, cross-domain, user-generated content, and social networks [2].

From an information science perspective, the LOD Cloud diagram integrates linked open data resources from different domains into an interconnected network and visualizes it, representing a new network morphology following typical knowledge networks such as citation, co-word, and co-authorship networks. What structural properties does the LOD network possess? Are there specific patterns and characteristics in the connections between datasets? Research on these questions helps us understand and evaluate the current state of linked data development and guides the practical publication, interlinking, and retrieval of linked data.

Domestic research on linked data has primarily focused on publication technologies [3-6], interlinking methods [7-9], and resource integration [10-11], with no studies yet examining the entire linked data network structure at the dataset level. Some relevant foreign research exists: Schmachtenberg et al. tracked the growth and interconnection of linked open datasets over the years, arguing that the LOD network has evolved from a DBpedia-centric structure to a more decentralized, non-centralized architecture, with content diversifying as quantities grow geometrically [12]. Auer et al. evaluated dataset quality by counting valid

inbound and outbound links, encountering frequent operational interruptions, access restrictions, and non-standard SPARQL endpoints, concluding that existing LOD statistics are overly optimistic and that actually usable datasets are an order of magnitude lower than reported [13]. Campinas et al. used the semantic web search engine Sindice to statistics on ontologies, predicates, strings, and URIs in linked datasets, providing data support for evaluating entity-oriented semantic search systems [14]. Bizer et al. comparatively analyzed the utilization rates of microdata, microformats, and RDFa, revealing the distribution and development of structured data on web pages [15].

While these studies statistically analyze linked datasets from various perspectives and describe the current state of the linked data network to some extent, the RDF (Resource Description Framework) data model of linked data endows it with typical network topology characteristics. This paper employs complex network theory metrics—degree distribution, average path length, clustering coefficient, and other topological properties—to describe the structure of linked open data, analyzing how to effectively organize linked data from a network connection perspective and revealing potential relationships hidden beneath structural surfaces.

2. Fundamental Properties of Complex Networks

(1) Scale-Free Characteristics

Degree distribution is a crucial parameter for characterizing scale-free networks. ER random networks exhibit approximately Poisson degree distributions, whereas most complex networks follow power-law forms. Networks with degree distributions obeying power-law distributions with exponent r 3 are generally considered scale-free networks, a property known as scale-free characteristics. This property is typically examined separately for degree, out-degree, and in-degree. Research on real-world networks across various domains has found that many networks, including the World Wide Web and citation networks, satisfy power-law degree distributions [16].

(2) Small-World Networks

The average path length of a network refers to the weighted average of the shortest paths between all reachable node pairs. The average clustering coefficient is the mean of all individual node clustering coefficients. Compared to random networks with identical node counts and average degrees, networks exhibiting both short average path lengths and high clustering coefficients are termed small-world networks [17].

3.1 Data Collection

Our data originates from Datahub [18], a free data management platform based on the CKAN data management system. Datahub groups and tags nearly 9,000

datasets, among which those published as linked open data and connected to other datasets are grouped into the LOD Cloud Group—the primary source for datasets in the LOD Cloud diagram. We selected the entire LOD Cloud group as our research object, using the datahub2void software to obtain VoID (Vocabulary of Interlinked Datasets) descriptions [19], with extraction concluding on April 28, 2015.

In linked data, links exist both within and between datasets. When publishing a dataset, external RDF links pointing to the dataset' s URIs must be ensured so that new datasets can be discovered by RDF browsers and crawlers and can supplement resources in existing datasets [20]. This study represents each dataset as a node, with connections between nodes representing dataset links. Since RDF links are directed, the linked open data network constitutes a directed network. We designate the network formed by datasets in the LOD Cloud group as LOD Cloud. Datasets tagged with "Publication" comprise library datasets, scientific publications, conferences, university readings, and citation datasets, which we extract as the Publication network for comparative analysis. Connections within this subnet are defined as links between nodes internal to the subnet. Specific statistics are shown in Table 1 .

3.2 Analysis Methods

(1) Metric Calculation

Cumulative Degree Distribution

For directed networks, the in-degree distribution P(k_in) represents the probability that a randomly selected node has in-degree k_in. The out-degree distribution P(k_out) represents the probability of out-degree k_out. The degree distribution P(k) represents the probability of degree k. To more clearly display degree distribution patterns, this paper uses cumulative degree distribution P_k for plotting, which represents the probability distribution of nodes with degree not less than k.

If the degree distribution follows a power-law distribution, i.e., P(k) k^(-), then the cumulative distribution function P_k follows a power-law with exponent -1: P_k k^(-(-1)). If the degree distribution follows an exponential distribution, P(k) e^(-k/) where > 0, then the cumulative distribution function P_k is also exponential with the same exponent: P_k e^(-k/).

Average Path Length

The distance d_ij between any two nodes i and j in a network is defined as the number of edges on the shortest path connecting them. The average path length L is the average of distances between all node pairs: L = (1/(N(N-1))) Σ_{i j} d_ij. Strictly speaking, the concept of average path length is finite only for connected graphs, but many real-world networks are disconnected. This study employs the classic Dijkstra algorithm for directed networks. For a random graph of equivalent scale with N nodes and average degree k, the average path length is: L_random ln(N)/ln(k).

Average Clustering Coefficient
For a node i with k_i edges connecting it to other nodes, these k_i nodes are called neighbors of i. The clustering coefficient C_i of node i is defined as the ratio of actual edges E_i existing among these neighbors to the maximum possible edges k_i(k_i-1)/2. The clustering coefficient C of the entire network is the average of all individual node clustering coefficients: C = (1/N) Σ_i C_i. For a random graph with N nodes and average degree k, the average clustering coefficient is: C_random k/N.

(2) Correlation Analysis

The variables in this study are ordinal. We calculate Spearman's rank correlation coefficient to analyze correlations between variables. Correlation coefficients are interpreted as follows: 0.00±0.30 indicates weak correlation, ±0.30~±0.50 indicates moderate correlation, ±0.50~±0.80 indicates significant correlation, and ±0.80~±1.00 indicates high correlation. Statistical significance is determined at $p < 0.05$.

(3) Regression Analysis

To determine distribution patterns of metrics, we plot scatter diagrams in Matlab and use the Curve Fitting Tool to add fitting curves. The best-fitting function is selected based on SSE (sum of squared errors, approaching 0 is optimal), R-Square (coefficient of determination, approaching 1 is optimal), Adjusted R-Square (adjusted coefficient of determination, approaching 1 is optimal), and RMSE (root mean square error, approaching 0 is optimal). The distribution patterns are determined according to the fitting functions.

(4) Visualization

We use Gephi to draw structural diagrams of LOD Cloud and Publication networks, employing different colored nodes to represent datasets from different domains, node size to represent degree magnitude, directed connections to represent dataset links, and connection thickness to represent link weight.

4.1 Degree and Correlation

(1) In-degree and Out-degree

Our collected data shows that 89% of nodes in the LOD Cloud network have non-zero degree, and 77% of nodes in the Publication subnet also have non-zero degree. This indicates that linked open datasets are not isolated, while also suggesting considerable room for development in dataset interconnections. Tables 2 and 3 list the top 10 datasets by out-degree and in-degree for both networks.

DBpedia ranks first with an in-degree of 140, meaning it is pointed to by most datasets in the LOD network, indicating its rich data resources and broad do-

main coverage, making it a trusted linking resource for subsequently published datasets. GeoNames, as a global geographic database, also has high in-degree. This phenomenon where nodes preferentially connect to "large" nodes with higher connectivity demonstrates the "preferential attachment" characteristic of the linked open data network. However, DBpedia and GeoNames have much smaller out-degrees compared to their in-degrees, related to their earlier publication dates. This reflects a common problem in the linked data network: many linked datasets lack maintenance after publication, fail to link to newly published datasets in a timely manner, and do not revise broken links, thereby reducing the overall connectivity of the LOD network.

Nearly half of the datasets in Table 3 also appear in Table 2's rankings, indicating that high-degree nodes in Publication also have relatively high degrees compared to core nodes in other domains. However, the degree values in Table 3 do not differ significantly from those in Table 2, suggesting that high-degree nodes in Publication tend to connect within their domain rather than contributing substantially to connecting the entire network.

(2) Correlation Between In-degree and Out-degree

The Spearman correlation coefficient between in-degree and out-degree for the entire LOD network is = 0.6546, with significance level p = 5.98 × 10^(-33) < 0.05, indicating significant correlation between out-degree and in-degree of linked datasets. Table 3 shows that many nodes rank high in both in-degree and out-degree, meaning core nodes simultaneously possess high in-degree and out-degree.

For the Publication network, the Spearman correlation coefficient between in-degree and out-degree is = 0.8939, with significance level p = 1.5 × 10^(-28) < 0.05, indicating high correlation between out-degree and in-degree of library and information science datasets. The positive correlation between in-degree and out-degree suggests that linked datasets tend to connect to datasets that are more frequently connected by others.

4.2 Cumulative Degree Distribution

The fitting results in Figure 1 [Figure 1: see original paper] show that the cumulative in-degree, cumulative out-degree, and cumulative degree distributions of LOD Cloud all approximate power-law distributions with exponent r 3, confirming that LOD Cloud exhibits scale-free network characteristics. Most nodes (28%) have degree 1, with a sparse tail distribution, indicating that a small number of nodes are connected by most nodes. In networks with such structural features, local node failures do not affect overall network stability, but failure of high-degree nodes makes the entire network vulnerable and impedes information flow. In LOD Cloud, a few highly popular nodes play crucial roles in connecting most nodes; identifying these nodes allows new nodes to quickly join the largest connected component of the linked data network and share more re-

sources. However, if new nodes all tend to connect to high-degree central nodes, the failure of central nodes could destroy overall network connectivity.

Figure 2 [Figure 2: see original paper] shows that the cumulative in-degree, cumulative out-degree, and cumulative degree distributions of the library and information science subdomain approximate exponential distributions. As degree increases, cumulative probability does not drop sharply, indicating a relatively uniform degree distribution in the Publication network. Such network structures exhibit stronger stability, as connectivity does not depend on a few extremely high-degree nodes; even local node failures have minimal impact on overall connectivity. The Publication degree distribution does not inherit LOD Cloud' s scale-free characteristics, indicating that network structures vary across different domains in linked open data and are not simply additive.

4.3 Average Path Length and Clustering Coefficient

Table 4 shows that both networks have average path lengths L smaller than those of random networks of equivalent scale (L_random), but clustering coefficients C much larger than C_random, demonstrating clear small-world network features. Short average path length means that even as linked open datasets continue to increase, distances between datasets remain small, accelerating retrieval speed. High clustering coefficient indicates that linked data connections are not random; if datasets B and C both connect to dataset A, they are also likely to connect to each other. This structure enables resources describing the same entity to interconnect, enriching entity description diversity. In short, small-world characteristics ensure both rapid data discovery and data richness, optimizing retrieval efficiency for linked open data.

5. Visualization Analysis

Figure 3 [Figure 3: see original paper] clearly shows that Publication domain datasets cluster tightly together, while life science domain datasets also form a small connected component. Datasets from other domains do not exhibit obvious clustering. This indicates that linked datasets have more interconnections within scientific domains. In the upper-left portion of Figure 3, the media domain features two star topologies centered on BBC Music and MusicBrainz, yet neither connects to the famous American media outlet New York Times; the shortest path between them is realized through DBpedia, suggesting that geographical separation may prevent interconnection among industry giants publishing linked data in the same domain.

Government domain linked datasets confirm this observation: except for several datasets published by the UK government that interconnect, other government datasets remain isolated. Datasets in user-generated content and social network domains are also completely dispersed without any connections. Cross-domain dataset connections are more diverse, most commonly with geographic datasets. This complex network connectivity means linked data cannot be hierarchically

divided by domain. We argue that to make linked data more tightly interconnected, we need to connect the institutions and people publishing linked datasets.

Research shows that the most common linking predicates between datasets are owl:sameAs and rdfs:seeAlso, used to connect two resources describing the same object [12]. Library and information science data primarily consists of bibliographies, papers, authors, and research institutions—information that is duplicated across many datasets, facilitating numerous interconnections and forming tight associations.

Figure 4 [Figure 4: see original paper] illustrates that high-degree nodes in Publication tend to interconnect with other high-degree nodes. The strongly connected component consists of datasets published using RKB Explorer, and even the Library of Congress' s published datasets connect through Consistent Reference Services (CRS), the underlying architecture of the RKB Explorer application, which enables connection of URIs pointing to the same entity [21]. This suggests that technical barriers also exist in linked open data interconnection [7].

Conclusion

The LOD network structure exhibits power-law distribution characteristics approximating a scale-free network at the global level, while simultaneously displaying small-world features with short average path length and high average clustering coefficient. The library and information science subdomain of linked data shows a relatively homogeneous exponential distribution while also possessing small-world network characteristics. The common small-world features can help optimize retrieval efficiency for linked open data; however, the tendency to preferentially connect to high-degree nodes reduces overall network stability. Therefore, careful selection of dataset interconnections is essential during publication. Network structure diagrams reveal that hierarchical structure is not associated with domain content; geographical and technical differences are important factors preventing tight connectivity in the linked data network.

Future research on LOD network structure should address the following: weight is a crucial statistical metric, and assigning weights to key nodes would provide deeper understanding of network structural characteristics. Current research remains at the level of static statistical analysis, but information structures evolve over time. New datasets generate new properties, and the linked open data network is also evolving. Studying this evolution will help us achieve more comprehensive understanding of linked open data.

References

[1] Bizer C, Heath T, Berner-Lee T. Linked Data-The Story So Far[J]. International Journal on Semantic Web and Information Systems, 2009, 5(3): 1-22.

[2] Schmachtenberg M, Bizer C, Paulheim H. State of the LOD Cloud 2014[R/OL]. (2014-08-30). [2015-04-28]. http://linked-datacatalog.dws.informatik.uni-mannheim.de/state/.

[3] Xia Cuijuan, Liu Wei, Zhao Liang, et al. The Current Technologies and Tools for Linked Data: A Case of Drupal[J]. Journal of Library Science in China, 2012, 38(1): 49-57.

[4] Shen Zhihong, Liu Xiaomin, Guo Xuebing, et al. A Research on Publishing Workflow and Key Issues of Linked Data: Experience with Publishing Scientific Literature and Scientific Data as Linked Data[J]. Journal of Library Science in China, 2013, 39(2): 53-62.

[5] Wang Zhongyi, Xia Lixin, Shi Yijin, et al. The Creation and Publishing of Middle Linked Data in Digital Library[J]. New Technology of Library and Information Service, 2013(5): 28-33.

[6] Bai Haiyan, Liang Bing. Semantic Pattern Mapping Between RDBMS and Linked Data Based on Open Source Software[J]. New Technology of Library and Information Service, 2011(7-8): 1-7.

[7] Shen Zhihong, Li Jianhui, Zhang Xiaolin. Research Review on the Interlinking Technology of Linked Data: Applications, Methods and Frameworks[J]. Library and Information Service, 2013, 57(14): 125-133.

[8] Zhu Wenjing, Xia Cuijuan, Liu Wei. Analysis of SILK Linkage Discovery Framework[J]. New Technology of Library and Information Service, 2013(4): 18-24.

[9] Bai Haiyan, Zhu Lijun. Research on Automatic Interlinking of Linked Data[J]. New Technology of Library and Information Service, 2010(2): 44-49.

[10] Ma Feicheng, Zhao Hongbin, Wan Yanling, et al. Integration of Network Information Resource Based on Linked Data[J]. Journal of Intelligence, 2011, 30(2): 167-170, 175.

[11] Ou Shiyan, Hu Shan, Zhang Shuai. An Ontology & Linked Data Driven Semantic Integration Method of Library Information Resources and Its Evaluation[J]. Library and Information Service, 2014, 58(2): 5-13.

[12] Schmachtenberg M, Bizer C, Paulheim H. Adoption of the Linked Data Best Practices in Different Topical Domains[C]. In: Proceedings of the 13th International Semantic Web Conference, Riva del Garda, Italy. Springer International Publishing, 2014: 245-260.

[13] Auer S, Demter J, Martin M, et al. Lodstats-An Extensible Framework for High-Performance Dataset Analytics[C]. In: Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management, Galway, Ireland. Springer Berlin Heidelberg, 2012: 353-362.

[14] Campinas S, Ceccarelli D, Delbru R, et al. The Sindice-2011 Dataset for Entity-Oriented Search in the Web of Data[C]. In: Proceedings of the 1st In-

ternational Workshop on Entity-oriented Search (EOS), Beijing, China. 2011: 26-32.

[15] Bizer C, Eckert K, Meusel R, et al. Deployment of RDFa, Microdata, and Microformats on the Web-A Quantitative Analysis[C]. In: Proceedings of the 12th International Semantic Web Conference, Sydney, Australia. 2013: 17-32.

[16] Wang Xiaofan, Li Xiang, Chen Guanrong. Network Science: An Introduction[M]. Beijing: High Education Press, 2012: 108-115.

[17] Wang Xiaofan, Li Xiang, Chen Guanrong. Complex Networks: Theory and Its Application[M]. The 4th Edition. Beijing: Tsinghua University Press, 2006: 22-34.

[18] About the Datahub[EB/OL]. [2015-04-28]. https://datahub.io/about.

[19] Describing Linked Datasets with the Void Vocabulary[EB/OL]. [2015-04-28]. http://www.w3.org/TR/void/.

[20] Heath T, Bizer C. Linked Data: Evolving the Web into a Global Data Space[M]. San Rafael: Morgan & Claypool Publishers, 2011: 64.

[21] RKB Explorer[EB/OL]. [2015-04-28]. http://www.rkbexplorer.com/explorer.

Author Contributions

Tan Ying: Designed the research scheme, conducted experiments, and wrote the paper;
Xia Lixin: Proposed the research idea and revised the paper.

Dates

Received: July 20, 2015
Revised: October 13, 2015

Note: Figure translations are in progress. See original paper for figures.