

How Do Organizations Publish Semantic Markup? Three Case Studies Using Public Schema.org Crawls

Daye Nam and Mayank Kejriwal, University of Southern California

Jointly launched in mid-2011 by major search engines like Google and Bing, Schema.org is designed to facilitate structured and knowledge graph-centric search applications on the Web. The Web Data Commons project has crawled increasing amounts of Schema.org data in recent years, providing a golden opportunity for socio-technological data studies that consider the semantics of content. The authors present empirical studies of organizations in three economic sectors that expose semantically linked Schema.org annotations.

Due to a strong push by search engines like Google and Bing, many websites continue to publish and expose increasing amounts of semantic markup in the form of microdata, RDFa, and Schema.org annotations (<http://schema.org>).

The last has become especially popular, and is displayed directly in search results by Google for some categories. Annotations, published in topical domains such as Organization and Person using the Schema.org vocabulary, are known to influence results of the Google Knowledge

Graph, in addition to influencing the search rankings, which further incentivizes public-facing entities to mark up their webpages.

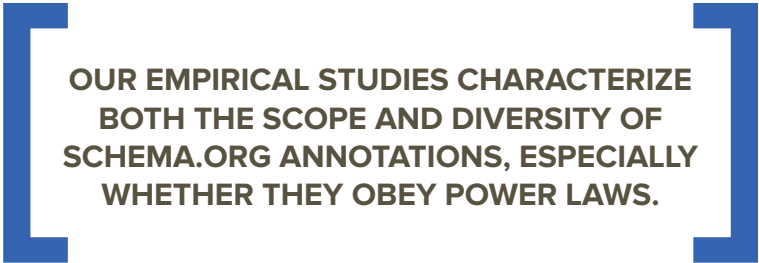
Recently, the Web Data Commons (WDC) project¹ has crawled and made publicly available large amounts of Schema.org annotations published in various domains. With careful processing, this data can be used to study different Web domains in terms of their semantic and topical properties. Particularly pertinent is the study of organizations in different economic sectors. Currently, WDC Schema.org crawls cover several organizational subdomains such as Hospital, School, Museum, Restaurant, and Hotel. Each such subdomain corresponds to a class in the Schema.org vocabulary. Although the WDC data has limitations like any large-scale crawling project, the consistent, public availability of this data over a period of at least three years provides a rich opportunity—and accompanying challenges—for socio-technological data studies that might not otherwise have been possible.

In this article, we present a principled methodology for studying organizational Schema.org annotations on the general Web. We draw on interdisciplinary techniques inspired by both Semantic Web research and network theory, and place our methodology in a broad context such that it can be used to study aspects of other topical domains as well. An important challenge arises from the observation that, unlike the hyperlinked (“regular”) Web or Semantic Web ecosystems like Linked Open Data,² Schema.org annotations gathered by the WDC project do not have a natural graph-theoretic structure. Also, microdata is still in its infancy, having only “taken off” since 2011, with the full pace of growth not

fully documented or known due to its inherently decentralized nature. We address both issues by showing a principled way to interlink Schema.org knowledge fragments and analyze the subsequent link structures.

Our empirical studies characterize both the scope and diversity of Schema.org annotations, especially whether they obey power laws inspired by classic network science.³ We examine the digital Schema.org footprint of three organizational domains—Hospital, Museum,

analyzing its growth and evolution. A (noncomprehensive) set of Schema.org studies include those on its necessity and design philosophy,⁴ normative analysis,⁵ extraction and identity resolution,⁶ and deployment analysis,⁷ among many others. The work of Robert Meusel⁶ is closest to ours in spirit, but because his analysis is purely statistical, he does not actually construct networks from the data or attempt to analyze measures of diversity or dynamicity.



OUR EMPIRICAL STUDIES CHARACTERIZE BOTH THE SCOPE AND DIVERSITY OF SCHEMA.ORG ANNOTATIONS, ESPECIALLY WHETHER THEY OBEY POWER LAWS.

and School—that are economically and functionally distinct in the non-digital world. Trends and differences that we observed both across the organizations (cross-sectional analysis) and across time (longitudinal analysis) illustrate the importance of structure and semantics in fine-grained Web science research with underlying socio-technological motivations. Finally, we describe potential future studies made feasible by our proposed methodology, especially network-theoretic studies that were previously applicable only to explicitly connected regions of the Web, such as Linked Open Data and observable aspects of social media.

RELATED WORK

As Schema.org has continued to increase in popularity, several researchers have proposed critically

Network theory has been closely linked to the Web sciences since at least the early 2000s. An influential work by Albert-László Barabási, Réka Albert, and Hawoong Jeong³ on the preferential attachment model gained significant traction, accounting for many of the structural power-law properties and “physics” observed for real-world hyperlink networks, especially the presence of hubs. Although Schema.org data is not naturally structured as a network, we devise experiments to verify if power-law properties also hold for the annotations published by various organizations, and the cross-sectional and longitudinal variance of these properties.

Within Semantic Web research is a movement that advocates a set of principles known as Linked Open Data that governs how data should be published,

linked, and accessed.² The movement espouses open standards and data interlinking: the fourth Linked Open Data principle, for example, states that new datasets should be linked to existing RDF datasets already published as Linked Open Data to qualify as Linked Open Data themselves. The movement has been quite successful, growing by orders of magnitude since its founding in 2007. Unlike Linked Open Data, the data considered in this article does not have the benefit of interlinking, which is challenging for structural techniques. However, with our proposed methodology, Linked Open Data-centric analyses become feasible on applicable subsets of Schema.org data.

We note that Schema.org differs in both its incentives and adoptees from Linked Open Data despite surface similarities: unlike Linked Open Data, Schema.org is ontologically constrained (all markup is roughly typed according to the Schema.org vocabulary⁴), has a much broader audience than the publishers of Linked Open Data, and is directly consumed by search engines like Google and Bing in both ranking and display of retrieved items. In general, the Schema.org ecosystem is more concerned with search rather than explicit data interlinking or strong semantics.

MOTIVATION

Our work is influenced by various subfields, some of which have historically been instrumental in the Web sciences, while others are emerging, typically industrially motivated research areas.

Resource Description Framework (RDF)

RDF is a graph-theoretic data model that is useful for publishing structured

data on the Web, and has emerged as the de facto standard on the Semantic Web, serving as the basis for more expressive languages like RDF Schema (RDFS) and the Web Ontology Language (OWL). Descriptions and standards for these are available in depth on the World Wide Web Consortium (W3C) standards page for the Semantic Web (www.w3.org/standards/semanticWeb). An RDF graph can be defined as a set of triples, where each triple is of the form (subject, property object). Subjects and properties must necessarily be Uniform Resource Identifiers (URIs), and even more generally, Internationalized Resource Identifiers (IRIs), whereas objects can be either URIs or literals, such as a string or a number. A triple may also be viewed as an edge in a directed, labeled graph, with the property serving as the edge label.

Importantly, subjects and objects can be *blank nodes*, which are special (in the sense of semantics) abstract identifiers. This identifier is used for an entity that exists but is primarily identified through its properties, since it is an anonymous resource to which a URI has not been explicitly allocated.

RDF triples are defined in the context of a graph, which is just the default (unnamed) graph in many cases. Each Schema.org annotation in the WDC crawls is serialized as a blank node, defined in the context of the webpage URL from which the annotation was scraped. Annotations are thus serialized not as triples but as quads. For example, if the (simplified) triple (:wonder_woman, sch:rating, 7.8) is extracted from two webpages, imdb.com/wonder_woman and moviereviews.com/wonder_woman, the crawl would contain two quads: (:wonder_woman,sch:rating,7.8,imdb

.com/wonder_woman) and (:wonder_woman,sch:rating,7.8,moviereviews.com/wonder_woman).

Graphs and network theory

Historically, and in many current fields of computer science research that overlap the Web sciences, graphs have played an important mathematical and computational role. The simplest definition of an undirected graph G is as a tuple (V, E) where V is a set of nodes and E (the set of edges) is a subset of $V \times V$, with an element $\{u, v\}$ indicating that an edge exists between nodes u and v . Similar definitions can be devised for directed graphs, attributed graphs, and weighted graphs.

Network theory, particularly in computer science, draws heavily on graph theory and representation, including undirected and directed graphs, weighted graphs, multi-relational graphs, and time-evolving graphs. The analyses used in this article are principally concerned with an undirected tripartite graph, whose construction, motivation, and structure are detailed later.

DATA COLLECTION

The studies described in this article use publicly available Schema.org annotations released by the WDC project. We consider annotated instances scraped from webpages and released as N-Quads files for three organizational domains in the Schema.org vocabulary: School, Museum, and Hospital.

Table 1 profiles the data for these domains collected over three years, 2014–2016. We draw special attention to the concept of the *pay-level domain* (PLD), which can be thought of as a “website” or “Web domain,” although in some cases the domain is further constrained by location (for

example, chicago.backpage.com vs. backpage.com). Table 2 provides some examples of the difference between a PLD and a webpage URL for several common scenarios for the School domain. For those who are interested in replicating extractions of PLDs from URLs, a Python function that takes a URL as input and outputs its PLD can be found at https://drive.google.com/drive/folders/OB7YcfP_4gRhmYjVNa2lWUnk1Vkk?usp=sharing.

Note that the trend in Table 1 that PLDs have continued to increase is consistent with increased adoption, though the data also shows that the crawl itself has changed: total file sizes significantly decline from 2015 to 2016 in the general case, despite evidence of increased adoption.

It is also important to note that each Schema.org annotation is a blank node from the WDC perspective. Figure 1

TABLE 1. Data collected for three topical domains in the Schema.org vocabulary over three years: 2014/2015/2016.

Data category	School domain	Museum domain	Hospital domain
Total file size in Mbytes (G-zip compressed N-Quads files)	154/248/97.1	2.7/4.1/92.2	98/203/77
URLs (×1,000)	403/319/926	7/24/96	233/407/191
Quads (×1,000,000)	7.79/16.43/3.83	1.32/2.54/4.27	4.68/10.86/4.01
Pay-level domains (PLDs)	165/200/297	43/69/123	174/223/335
Blank-node entities (×1,000)	668/1,240/217	8.5/29.1/100.6	270/514/215

shows some examples. Each blank node is thus associated with a set of triples describing the entity itself and with a single URL from which it was scraped. Formally, the representation is N-Quads; we provided an example describing Wonder Woman earlier. All studies in this article are at the

abstraction of the PLD rather than an individual webpage URL.

TRIPARTITE NETWORK CONSTRUCTION

The data described earlier may be thought of as a *forest of RDF graphlets* since, by default, two blank nodes

TABLE 2. Examples of School pay-level domains (PLDs).

Scenario	School name(s) with identifier (in this case, phone number)	Webpage(s)	PLD(s)
Same school, different webpages, same PLD	Tory Urban Retreat (6443844329)	http://www.wellingtonnz.com/discover/things-to-do/shopping/yoga-unlimited-at-tory-urban-retreat/ http://www.wellingtonnz.com/discover/things-to-do/shopping/tory-urban-retreat/	http://www.wellingtonnz.com/
Same school, different webpages, different PLDs	Rose Valley Elementary School (13014494990)	https://www.ziprealty.com/schools/detail/360383/FORT-WASHINGTON,MD/PRINCE-GEORGE'S-COUNTY-PUBLIC-SCHOOLS/ROSE-VALLEY-ELEMENTARY-SCHOOL http://www.publicschoolreview.com/rose-valley-elementary-school-profile/20744	http://www.ziprealty.com/ http://www.publicschoolreview.com/
Different schools, different webpages, same PLD	Rose Valley Elementary School (13014494990) Damascus High School (13012537030)	http://www.publicschoolreview.com/rose-valley-elementary-school-profile/20744 http://www.publicschoolreview.com/damascus-high-school-profile	http://www.publicschoolreview.com/

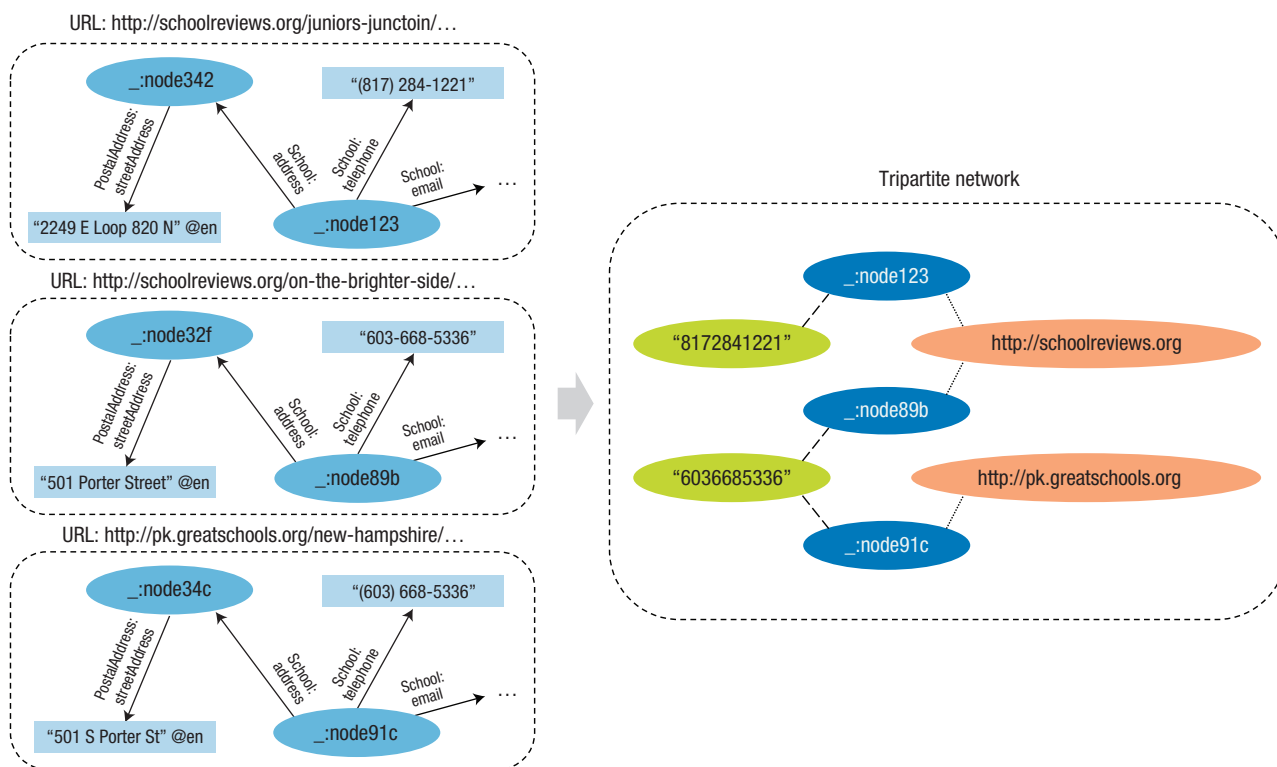


FIGURE 1. Examples of Schema.org annotations scraped from three independent URLs (left) and of tripartite network construction (right). The network has two types of links: resolution (between the blue and green nodes) and semantic (between the blue and pink nodes).

potentially referring to the same underlying organization are not linked. This more-than-five-decades-old research problem⁸ of automatically resolving entities into a single underlying entity is known as *entity resolution* (ER). Although much progress has been achieved in ER, particularly in the Semantic Web community for graphs published as RDF Linked Open Data,^{9,10} the problem has not been solved and only limited work¹¹ has been done on difficult large-scale datasets.

Without ER, it is difficult applying network theory of any kind in a study because the RDF graphlets forest is heavily disconnected. A critical

real-world observation that we can draw on to address this issue is that organizations almost always publish their phone numbers (and in many cases, addresses), since a primary incentive in publishing Schema.org annotations is search optimization by engines like Google. As we are only concerned with organizations in this article, we consider an organization's phones as its *pseudo-identifiers*. Specifically, two blank nodes are said to represent the same organization if they share at least one pseudo-identifier. We refer to such a link as a *resolution link*. In the same vein, two blank nodes are said to be *semantically linked* if they co-occur within the context of a PLD.

We can concretize and represent both resolution and semantic links as a *tripartite undirected network*. Figure 1 shows an example of this construction for the School domain. A multi-partite network is defined to not allow links between nodes in the same partition, only across partitions. We also note that it is possible to “close” the graph in several ways by using paths. For example, we could derive a bipartite network that abstracts away blank nodes completely by connecting a pseudo-identifier to a PLD if a blank node extracted from that PLD contains the pseudo-identifier as a property. In the graph-theoretic representation in Figure 1, this would be equivalent to

connecting a node from the pseudo-identifier partition to a node from the PLD partition if a path exists between the two nodes in the original tripartite network.

We consider phone numbers instead of addresses primarily because the former are easier to normalize and more readily available in the corpus than the latter, and also because it is not self-evident that addresses can serve the role of pseudo-identifiers. It is also highly unlikely after normalization for any two distinct organizations to have the same phone number, which makes the network very precise. Moreover, this manner of interlinking has precedent in previous Semantic Web research, particularly in using email addresses to interlink fragments typed according to the Friend-of-a Friend (FOAF) ontology (www.foaf-project.org).

It is possible for organizations to be linked in the real world (for example, one might be a subsidiary of another) and have different phone numbers. This kind of information, unfortunately, is difficult to acquire and often requires a paid service like that offered by Thomson Reuters (www.thomsonreuters.com/en/products-services.html). This is one reason why the phone number is only a pseudo-identifier. An interesting goal of future work would be to try completing the graph further and empirically assess if that changes any of the results we present. While we do not address this issue further, we note that it is not uncommon for Web sciences researchers to perform analysis of partially complete (albeit precise) graphs.

CASE STUDIES

To provide a conceptually simple illustration of how constructed tripartite

networks can be used to analyze Schema.org markup published by different organizations on the Web, here we analyze the results from three sets of experiments using our collected data on the topical domains Hospital, Museum, and School. Bear in mind that the scope of the analysis is necessarily limited to these domains and the 2014–2016 time frame, along with inherent bias or incompleteness that might be present in the WDC crawls. Later we discuss more ambitious studies that become possible due to our proposed tripartite network-theoretic methodology.

Sharing and representation

In early network-theoretic studies of the hyperlinked Web, the degree distribution of the nodes obeyed the power law, indicating the presence of hubs in the network.³ In other words, some nodes—Wikipedia, for example—received a disproportionate number of incoming links while most nodes had few, if any, incoming links. The first set of our experiments aimed to determine whether the sharing and representation of Schema.org annotations obey similar laws. Are some annotations overrepresented in the data? Is this finding consistent across years and topical domains? What do the hubs in such a distribution correspond to in the real world?

To answer these questions, we constructed a tripartite network in which each pseudo-identifier is represented as a unique Schema.org “item”—the equivalent of a node in a Web hyperlink graph. The degree of each node is computed as the number of blank nodes that have an edge incident on the node. In Figure 1, for example, the degrees of the two phone-number nodes are 2 and 1, respectively.

If the empirical frequency F of the degree k obeys a power law, F would be proportional to k^{-a} for some positive constant a . On a log-log plot, a perfect power-law distribution would be a straight line with a negative slope. As in previous studies, we plotted the empirical frequency of the degree on the y-axis and the degree itself on the x-axis to observe if this is so.

Figure 2a illustrates the results. Hospital seems to show remarkable stability and exhibits (along with Museum, to a more limited extent) the familiar power-law distributions observed on the hyperlinked Web,⁴ but School does not seem to exhibit a power-law distribution. In previous years the curve was almost quadratic, which suggests a normal distribution. Both Museum and School appear to be evolving, not being as stable as Hospital. Museum seems to be evolving in a more regular way than School, which is the least regular of the three domains.

To identify the specific organizations that correspond to high-degree nodes (“hubs”), we noted the pseudo-identifiers representing the highest-degree nodes and manually searched for the identifiers online.

For Museum, there are several examples including Stadel Museum in Frankfurt, Penn Museum in Philadelphia, and the American Folk Art Museum in New York City. Although it is impossible to manually verify the pseudo-identifiers of all high-degree nodes, this finding provides some evidence that higher-degree nodes correspond to museums that are well visited but do not necessarily reflect ranking in terms of endowment or influence. Slightly lesser-known museums might be higher in ranking by advertising more aggressively through semantic markups.

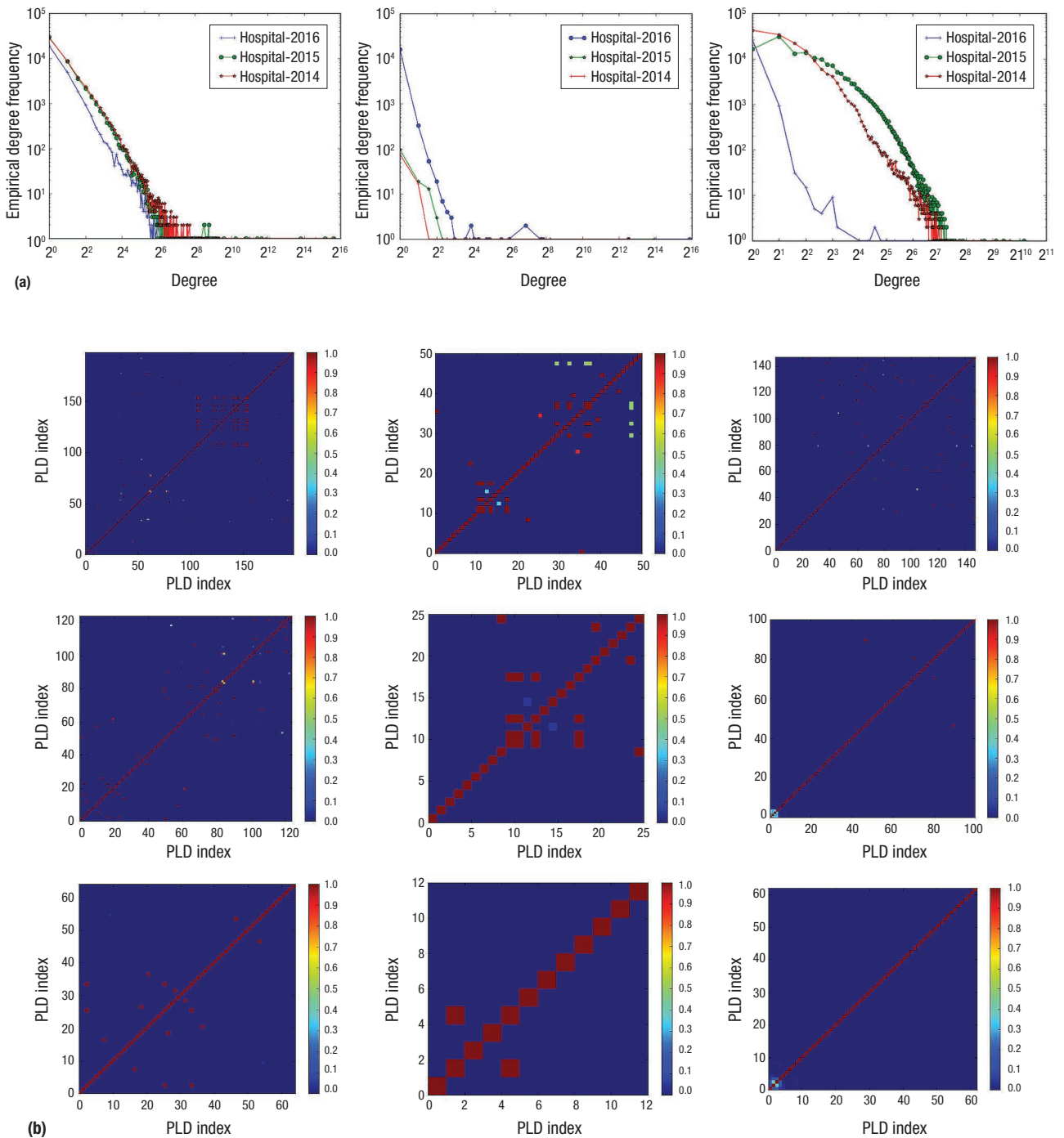


FIGURE 2. Using tripartite networks to analyze Schema.org markup for the topical domains Hospital, Museum, and School. (a) Log-log illustration of the sharing and representation of Schema.org pseudo-identifiers as a function of their degree for all three domains. (b) Heat-map representations of the Jaccard similarity matrices (with 1.0 indicating perfect similarity and minimum diversity). In the general case, PLD indexes will map to different PLDs across any two maps, and the number of PLDs (axes scaling) will also be different. Interpretations must be similarly cautious.

Our analysis yielded similar findings for both School and Hospital. Examples of high-degree nodes for the former include Albuquerque Public

Schools and the American Graphics Institute and for the latter include St. Jude Hospital and Queen of the Valley Medical Association.

Diversity

The set of experiments described above did not take the PLDs into account. Furthermore, it did not

present a clear picture of whether Schema.org is diverse. Put simply, did every PLD roughly describe the same set of items? We addressed this question visually in a second set of experiments by first closing the tripartite network into a bipartite network, with one partition consisting of pseudo-identifiers and another partition consisting of PLDs, and then linking two nodes (across partitions only) if a path existed between them in the original network.

We represent the bipartite network as a dictionary D of key-value pairs, where the key is a PLD and the value is the set of pseudo-identifiers associated with the PLD. We denote the set $\{p_1, \dots, p_{|p|}\}$ of PLDs as P . Next, using an inverted index methodology, we efficiently compute the Jaccard similarity matrix on the cross product $P \times P$. Specifically, the value of the cell positionally denoted by (i, j) is computed using the formula $|D[p_i] \cap D[p_j]| / |D[p_i] \cup D[p_j]|$. Note that the elements on the diagonal are always 1.0. For a given domain and year, we illustrate the matrix using a heat map as shown in Figure 2b, with higher values indicated with redder colors.

Although there are small pockets of high heat on the maps, indicating significant commonalities, much of the map is characterized by lower Jaccard scores and hence higher diversity. Across time, the results are harder to compare because of the flux and growth in PLDs (especially for Museum). For School and Hospital, the evidence indicates that diversity might be decreasing, as there are more hot spots in 2016 for both domains than in 2014, but the effect is quite limited. Overall, semantic mark-ups remain diverse in organizational domains.

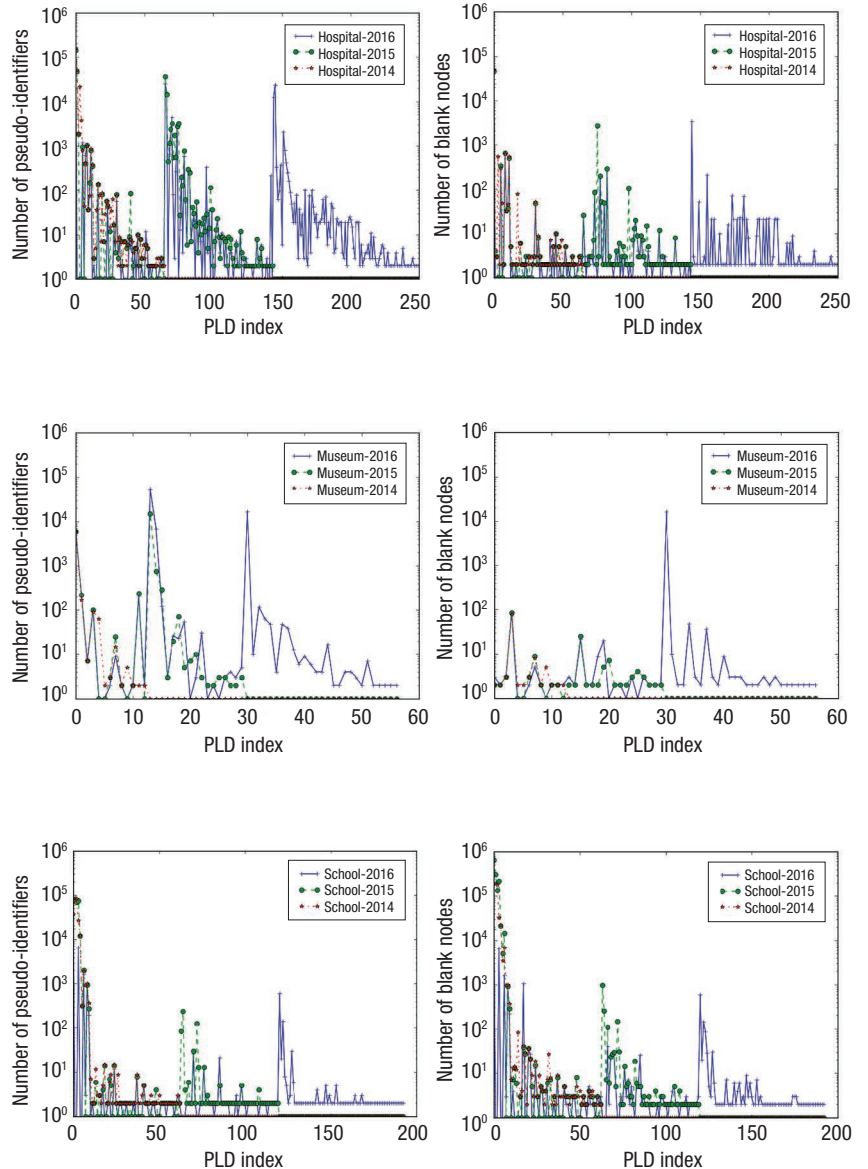


FIGURE 3. Per-PLD longitudinal illustrations of the three organizational domains both as a function of the number of pseudo-identifiers and the number of blank nodes. The x-axis has no intrinsic ordering.

Dynamic properties of PLDs

In the final set of experiments, we mapped the dynamic activity of PLDs according to the WDC crawl by computing and plotting the number of blank nodes and the number of pseudo-identifiers recorded for each PLD (for a fixed topical domain) over the period 2014–2016. If a PLD did not have Schema.org annotations for the topical domain for a year, we set these values to 0. A nonzero value in the blank node plot indicates that at least one

Schema.org annotation must have been present and got extracted. For greater readability, we assigned each PLD an integer (PLD index) and placed these integers on the x-axis of the plots shown in Figure 3.

A conservative interpretation of the plots is that the WDC crawl is becoming “more complete” since a nonzero value in either 2014 or 2015 strongly implies a nonzero value in 2016. It could also mean that more Schema.org data is getting published

(and thus extracted). There is external evidence for both interpretations, but more careful crawls will be needed to determine either contribution.

The plots also illustrate interesting cross-sectional differences. The organizational domain Hospital, which earlier had the smoothest scale-free distributions, also shows greater distributional coverage compared to the other domains in Figure 3. More studies are needed on the nature of the irregularity of School and Museum, and we describe some possible directions in the next section. To facilitate such studies, all raw data, including the PLD mappings, are available as supplemental spreadsheets at https://drive.google.com/drive/folders/OB7YcfP_4gRhMYjVNa2lWUnk1Vkk?usp=sharing.

FUTURE WORK

The three sets of experiments described above were initial empirical responses to three broad socio-technological questions we formulated. The first set of experiments addressed the question of whether Schema.org growth and evolution for the three organization types mirrored that of the hyperlinked Web in a semantically well-defined way, the second set examined the diversity of Schema.org entities, and the third set was a preliminary longitudinal study. All three studies were feasible because, with our proposed tripartite network construction, we were able to transform Schema.org from a forest of RDF graphlets to a conceptual network where nodes represent entities (for example, specific schools) and links are analogous to entity-level hyperlinks in that a link is forged between two entities if they share a PLD—that is, they are contextually co-located on the Web.

ABOUT THE AUTHORS

DAYE NAM is completing her MSc in computer science at the University of Southern California (USC) Viterbi School of Engineering and will begin her doctoral studies this fall at Carnegie Mellon University. She conducted this analysis as a directed research project under Dr. Kejriwal's supervision. Contact her at dayenam@usc.edu.

MAYANK KEJRIWAL is a research assistant professor in the Department of Industrial and Systems Engineering and a research scientist at the Information Sciences Institute (ISI) at the USC Viterbi School of Engineering. His research focuses on challenging datasets that cover various genres, such as tables, graphs, HTML documents from both the Web and Dark Web, natural language documents, and social media documents like tweets, and that span challenging domains such as human trafficking, crisis informatics, causal reasoning, and geopolitical forecasting. Kejriwal received a PhD in computer science from the University of Texas at Austin. He is a member of IEEE, ACM, AAAI, AAAS, SIAM, and AGU. Contact him at kejriwal@isi.edu.

In this way, a global picture of specific Schema.org types (of which the instances can be interlinked using pseudo-identifiers) emerges.

These representational transformations make other studies possible, including centrality and node importance analyses as well as global community detection on each Schema.org type's network in the context of the overall Web as well as within a local (PLD-specific) context. In addition, the research questions we formulated could be addressed in alternative ways. For example, diversity could be explored using homophily metrics instead of Jaccard similarity matrices.

All such studies can leverage our openly shared supplementary data. Also, Semantic Web researchers can now undertake studies traditionally limited to Linked Open Data.¹² For example, we are currently comparing the growth,

semantics, and structural metrics of the more standardized and connected Linked Open Data ecosystem with the decentralized Schema.org ecosystem.

This article proposed a novel methodology to characterize three organizational domains that already have a significant Schema.org presence. To do so in a principled manner, we construct networks using domain-specific linking keys (in our case, phone numbers) and process and analyze the networks using the framework of network theory. Our analysis shows that while the distribution of Schema.org annotations has clear, relatively stable power-law dependencies, there is considerable evidence of both diversity and dynamic behavior. Most importantly, the behavior across topical domains is quite different

quantitatively, although some common trends stand out. We conclude that the structure and semantics in Schema.org annotations are rich resources for future, more fine-grained data studies in the Web sciences. **□**

DISCLAIMER

This work was not funded by any agency, organization, or grant.

REFERENCES

1. H. Mühleisen and C. Bizer, "Web Data Commons—Extracting Structured Data from Two Large Web Corpora," *Proc. WWW 2012 Workshop Linked Data on the Web (LDOW 12)*, 2012; <http://events.linkeddata.org/ldow2012/papers/ldow2012-inv-paper-2.pdf>.
2. C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data—The Story So Far," *Int'l J. Semantic Web and Information Systems*, vol. 5, no. 3, 2009, pp.1-22.
3. A.L. Barabási, R. Albert, and H. Jeong, "Scale-Free Characteristics of Random Networks: The Topology of the World-Wide Web," *Physica A: Statistical Mechanics and Its Applications*, vol. 281, no. 1, 2000, pp. 69-77.
4. R.V. Guha, D. Brickley, and S. Macbeth, "Schema.org: Evolution of Structured Data on the Web," *Comm. ACM*, vol. 59, no. 2, 2016, pp. 44-51.
5. P.F. Patel-Schneider, "Analyzing Schema.org," *Proc. 13th Int'l Semantic Web Conf. (ISWC 14)*, 2014, pp. 261-276.
6. R. Meusel, "Web-Scale Profiling of Semantic Annotations in HTML Pages," doctoral dissertation, Univ. of Mannheim, 2016; https://ub-madoc.bib.uni-mannheim.de/41884/1/thesis_final_rm_20170322-1.pdf.
7. C. Bizer et al., "Deployment of RDFa, Microdata, and Microformats on the Web—A Quantitative Analysis," *Proc. 12th Int'l Semantic Web Conf. (ISWC 13)*, 2013, pp. 17-32.
8. P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Springer, 2012.
9. M. Kejriwal and D.P. Miranker, "Semi-supervised Instance Matching Using Boosted Classifiers," *Proc. 12th European Semantic Web Conf. (ESWC 15)*, 2015, pp. 388-402.
10. J. Volz et al., "Silk—A Link Discovery Framework for the Web of Data," *Proc. WWW 2009 Workshop Linked Data on the Web (LDOW 09)*, 2009; http://events.linkeddata.org/ldow2009/papers/ldow2009_paper13.pdf.
11. M. Kejriwal, "Populating Entity Name Systems for Big Data integration," *Proc. 13th Int'l Semantic Web Conf. (ISWC 14)*, 2014, pp. 521-528.
12. M. Schmachtenberg, C. Bizer, and H. Paulheim, "Adoption of the Linked Data Best Practices in Different Topical Domains," *Proc. 13th Int'l Semantic Web Conf. (ISWC 14)*, 2014, pp. 245-260.

**SUBMIT
TODAY**

IEEE TRANSACTIONS ON
**MULTI-SCALE
COMPUTING
SYSTEMS**

▶ **SUBSCRIBE
AND SUBMIT**

For more information on paper submission, featured articles, calls for papers, and subscription links visit:

www.computer.org/tmscs

TMSCS is financially cosponsored by IEEE Computer Society, IEEE Communications Society, and IEEE Nanotechnology Council

TMSCS is technically cosponsored by IEEE Council on Electronic Design Automation

 **IEEE**

IEEE
 **computer
society**

myCS Read your subscriptions through the myCS publications portal at <http://mycs.computer.org>