Link Analysis of Life Science Linked Data

Wei Hu[1](✉), Honglei Qiu[1], and Michel Dumontier[2]

[1] State Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing, China
whu@nju.edu.cn, hlqiu@smail.nju.edu.cn

[2] Stanford Center for Biomedical Informatics Research,
Stanford University, Stanford, USA
michel.dumontier@stanford.edu

Abstract. Semantic Web technologies offer a promising mechanism for the representation and integration of thousands of biomedical databases. Many of these databases provide cross-references to other data sources, but they are generally incomplete and error-prone. In this paper, we conduct an empirical link analysis of the life science Linked Data, obtained from the Bio2RDF project. Three different link graphs for datasets, entities and terms are characterized using degree distribution, connectivity, and clustering metrics, and their correlation is measured as well. Furthermore, we analyze the symmetry and transitivity of entity links to build a benchmark and preliminarily evaluate several entity matching methods. Our findings indicate that the life science data network can help identify hidden links, can be used to validate links, and may offer the mechanism to integrate a wider set of resources for biomedical knowledge discovery.

Keywords: Link analysis · Bio2RDF · Life sciences · Linked data

1 Introduction

Semantic Web (SW) technologies such as Linked Data provide a salient mechanism by which human and machine can navigate across large and heterogeneous data sources [6]. Links in distributed datasets [14] usually occur between entities (a.k.a. instances) or terms (i.e. classes and properties), and can be not only manually curated but also automatically generated [28]. Due to their complexity and descriptive nature, the life science and health care domains have long been used to assess the feasibility of advanced knowledge management systems. With over 1,500 published biomedical databases, numerous efforts have been directed towards establishing Linked Data for the life sciences, including Bio2RDF [5,8], Chem2Bio2RDF [9], Neurocommons [25], the EBI RDF Platform [22], and W3C HCLS Linked Open Drug Data.[1] They contain millions of links (e.g. owl:sameAs relations) over hundreds of datasets that partially overlap in content. Such rich networks can yield insights into the basic structures demanded to express data

[1] http://www.w3.org/wiki/HCLSIG/LODD

M. Arenas et al. (Eds.): ISWC 2015, Part II, LNCS 9367, pp. 446–462, 2015.
DOI: 10.1007/978-3-319-25010-6_29

types, facilitate large-scale data integration, and help improve the overall quality of biomedical data. To the best of our knowledge, however, there is no such study at present.

In this paper, we conduct an empirical link analysis of the life science Linked Data, obtained from the Bio2RDF project, in three perspectives:

- *Dataset link analysis*, which provides the statistics of datasets and their links to other datasets based on the RDF data model;
- *Entity link analysis*, which captures the status and intended semantics of links between entities using a special kind of cross-references in Bio2RDF;
- *Term link analysis*, which measures the overlap of topics between terms by ontology matching.

For each perspective, we investigate the graph features of Bio2RDF vis-à-vis what has been previously reported, e.g. [12,18]. Specifically, we represent datasets (entities and terms respectively) and their links by a graph, and measure the degree distribution, connectivity and clustering metrics. Furthermore, we examine the symmetry and transitivity of entity links, and establish a benchmark to preliminarily evaluate several entity matching approaches. In addition to study each perspective alone, we also analyze their correlation. The data and results shown in this paper are available at http://ws.nju.edu.cn/bio2rdf-analysis/.

Our analytical results and findings are expected to be useful in many areas. For biomedical data exploration [4], our entity link analysis can help create multiple sets of links according to different equivalence criteria and interpretations, e.g. “truly identical” or “close match”. Our dataset link analysis can help identify hidden links between hundreds of biomedical datasets and enable federated SPARQL query processing. Our analysis can also be used to identify error links and poorly annotated datasets, which require more manual or automated curation. Moreover, our empirical analysis of Bio2RDF may reveal some widespread trends in the life sciences and even in the SW, which provide evidences for applications using Linked Data and guide future research.

The rest of this paper is organized as follows. Section 2 provides the preliminaries used in the paper. In Section 3, we introduce the dataset link analysis. In Section 4, we describe the entity link analysis and evaluate entity matching approaches. In Section 5, we present the term link analysis. Section 6 measures the correlation between the three different link structures. We introduce related work in Section 7 and discuss our findings in Section 8. Finally, we conclude this paper with future work in Section 9.

2 Preliminaries

Let $\mathbf{U}$ be the set of URIs, $\mathbf{L}$ be the set of literals, and $\mathbf{B}$ be the set of blank nodes. A triple $\langle s, p, o \rangle \in (\mathbf{U} \cup \mathbf{B}) \times \mathbf{U} \times (\mathbf{U} \cup \mathbf{L} \cup \mathbf{B})$ is called an *RDF triple*. Following VoID [2], an RDF *dataset* is a set of RDF triples that are published, maintained or aggregated by a single provider. Typically, a dataset is accessible on the Web, for example through resolvable HTTP URIs or through a SPARQL endpoint, and is identified by a namespace.

In a dataset, named classes, properties and instances are uniquely identified using URIs. Classes and properties together are referred to as *terms*, and terms sharing a common namespace constitute a vocabulary. In this paper, instances are particularly referred to as *entities*.

A *graph* comprises nodes and edges, and edges can be either ordered (a.k.a. arcs) for a directed graph or unordered for an undirected graph. The *degree* of a node is the number of edges incident to it. For a directed graph, we distinguish between the outgoing degree and incoming degree of a node. The outgoing degree of a node is the number of edges directed from it, while the incoming degree of a node is the number of edges directed to it. A *sink* node is a node with outgoing degree equal to 0, while a *source* node has its incoming degree equal to 0. The degree of a node in a directed graph is the sum of its outgoing and incoming degrees. A node with a degree of 0 is called an *isolated* node.

A random variable x is distributed according to a *power law* when its probability density function $p(x)$ is in the form of $p(x) \propto x^{-\alpha}$, where α is a positive constant called *power law exponent*. Power law functions are *scale-free*, in the sense that if x is rescaled by multiplying it by a constant, $p(x)$ would still be proportional to $x^{-\alpha}$. Clauset *et al.* [11] designed a well-known maximum-likelihood method to estimate α for both discrete and continuous values.

A *weakly connected component* (WCC) for a directed graph is a subgraph in which any two nodes can reach each other through some undirected path and to which no more nodes or edges can be added while still preserving its reachability. The number of nodes in a connected component is called its *size*.

The *average distance* for a WCC is the average shortest path length between all nodes in the WCC. The clustering coefficient for a node in a WCC quantifies how close its neighbors are to be a clique (complete graph), while the *clustering coefficient* for the WCC is the average of the clustering coefficients of all nodes. A graph demonstrates the *small world* phenomenon, if its clustering coefficient is significantly higher than that of a random graph on the same node set, and if the graph has a shorter average distance. Degree distribution, average distance and clustering coefficient are considered as the three most robust measures of network analysis.

3 Dataset Link Analysis

Bio2RDF [8] is an open source project that uses SW technologies to build and provide the largest network of life science Linked Data. Particularly, Bio2RDF defines a set of convention scripts to create RDFS compatible Linked Data from a diverse collection of heterogeneously formatted sources obtained from multiple data providers. In this analysis, we use Bio2RDF Release 3 (July 2014), which is the latest version of Bio2RDF and contains about 11 billion RDF triples, 1 billion entities, 2 thousand classes and 4 thousand properties from 35 datasets. For more information, please visit http://download.bio2rdf.org/release/3/release.html. To conduct the dataset link analysis of Bio2RDF, we define the dataset link graph as follows:

Definition 1 (dataset link graph). *A dataset link graph, denoted by* $(\mathbf{D}, \mathbf{A})$*, is a directed graph, where* $\mathbf{D}$ *is the node set, and each node* $D_i \in \mathbf{D}$ *denotes a dataset;* $\mathbf{A}$ *is the arc set, and each arc* $(D_i, D_j) \in \mathbf{A}$ *exists iff there are at least* k *RDF triples* $\langle s, p, o\rangle \in D_i$*, where* s, o *are two URIs in* D_i *and* D_j *respectively.* k *is a non-negative integer to adjust the sparseness of arcs in the graph.*

Since Bio2RDF has assigned unique names to the data lacking a source identifier, blank nodes are not existent in the datasets. The original dataset of a URI is obtained by dereferencing the URI, because Bio2RDF includes a few datasets, e.g. BioPortal and iRefIndex, which are themselves aggregates of other datasets [8]. Also, meta-level URIs in RDF(S) and OWL are excluded as every Bio2RDF dataset has RDF triples involving them.

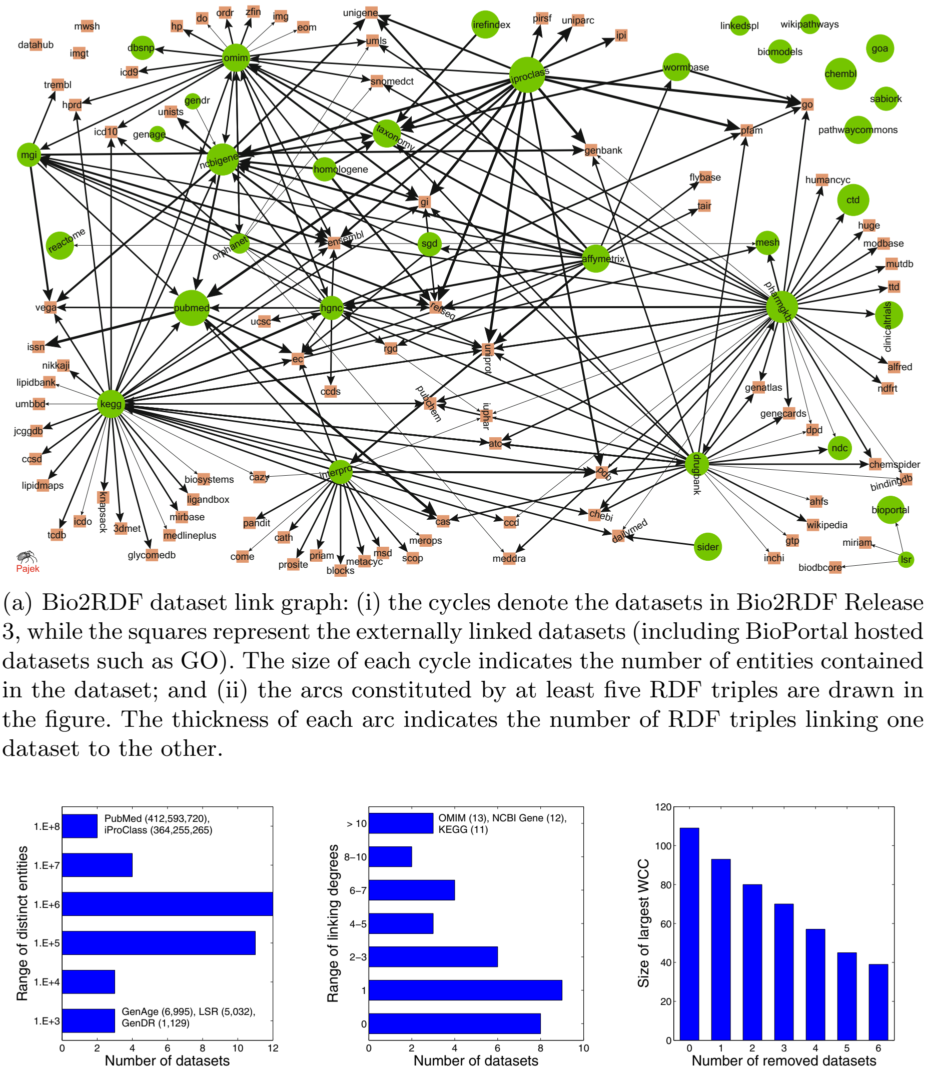
Fig. 1(a) shows the generated dataset link graph for Bio2RDF. We observe that the majority of the datasets is well linked and the largest connected component contains 28 Bio2RDF datasets and 81 *external* datasets that have not been converted in Bio2RDF. The upper right corner depicts seven isolated Bio2RDF datasets that have not linked with others yet, while the upper left corner shows three isolated external datasets linked by less than five triples. In consideration of at least thousands of URIs in each dataset, we regard this little number as a mistake. The lower right corner lists four connected datasets.

Due to most externally linked datasets do not support SPARQL queries, it may be more fair to not consider the directionality of dataset links. The average distance of the largest WCC in the figure is 2.77 and the clustering coefficient is 0.22. The average distance and clustering coefficient for a random graph with the same numbers of nodes and edges are 6.6 and 0.013, respectively. Thus, it indicates very good connectivity among the datasets and reveals the small world phenomenon. Additionally, Fig. 1(a) gives us several hints about the external datasets that are direly needed in the next release of Bio2RDF, such as UniProt and Ensembl, due to many Bio2RDF datasets linking to them (a.k.a. authorities on the Web [7]).

More specifically, Fig. 1(b) illustrates that entities in the Bio2RDF datasets are approximately normally distributed, where 23 datasets have hundred thousands to millions of entities. We also show in this figure the datasets with most or least entities.

Fig. 1(c) illustrates the link distribution between the Bio2RDF datasets only, where OMIM has the most links with the other datasets (including 6 outgoing links and 7 incoming links), followed by NCBI Gene (12 links) and KEGG (11 links). If we took the external datasets into account, the three datasets with most out-going links (a.k.a. hubs [7]) would be KEGG (42 outlinks), PharmGKB (36 outlinks) and DrugBank (24 outlinks).

Fig. 1(d) shows the size variation of the largest WCC by keep removing the datasets holding most links. The sequence of removal is KEGG, PharmGKB, OMIM, DrugBank, InterPro and iProClass. We find that the size of the largest WCC decreases slowly, which demonstrates good resilience among the datasets.



(a) Bio2RDF dataset link graph: (i) the cycles denote the datasets in Bio2RDF Release 3, while the squares represent the externally linked datasets (including BioPortal hosted datasets such as GO). The size of each cycle indicates the number of entities contained in the dataset; and (ii) the arcs constituted by at least five RDF triples are drawn in the figure. The thickness of each arc indicates the number of RDF triples linking one dataset to the other.

(b) Entity distribution

(c) Link distribution betw. Bio2RDF datasets only

(d) WCC size variation by keep removing datasets with most links

Fig. 1. Bio2RDF datasets and their links

In overall, this analysis characterizes a landscape of the current Bio2RDF datasets and provides the basis for analyzing entity and term links in the next two sections.

4 Entity Link Analysis

During the dataset link analysis, we observe that the majority of dataset links is generated from a special kind of RDF triples in the form of $\langle s, \text{x-relation}, o\rangle$.[2] X-relations contribute to more than 76% entity links, followed by article (12%), gene (4.3%) and disease (1.8%), but they have under-specified semantics.

As an example, kegg:x-drugbank links a KEGG entity (e.g. kegg:D03455) to a DrugBank entity (e.g. drugbank:DB00002) and its intended meaning is to specify that these two entities are "truly identical" (e.g. both refer to the same drug "Cetuximab"), but kegg:x-drugbank is not defined as a sub-property of owl:sameAs. In another case, kegg:x-pubmed signifies a reference to a scientific article that is indexed in the PubMed dataset. Other meanings that we observed include "part of" and "close match". Actually, due to the design principles of Bio2RDF [18], owl:sameAs would be only used when the URI is precisely another name for an entity in the original dataset, for instance, where Bio2RDF URIs for DrugBank entries coincide with URIs assigned by DrugBank itself.

Since x-relations are key to link entities in Bio2RDF, we seek to examine its role in link structure and determine the extent to which we can use x-relations to create entity links. We define the entity link graph using x-relations:

Definition 2 (entity link graph). *An entity link graph, denoted by* $(\mathbf{E}, \mathbf{X})$*, is a directed graph, where* $\mathbf{E}$ *is the node set, and each node* $e_i \in \mathbf{E}$ *represents an entity;* $\mathbf{X}$ *is the arc set, and each arc* $(e_i, e_j) \in \mathbf{X}$ *exists iff there is an x-relation linking* e_i *to* e_j*, in other words, there exists an RDF triple* $\langle e_i, \textit{x-relation}, e_j\rangle$.

4.1 Degree Distribution

We generate the entity link graph for Bio2RDF. In Fig. 2, we depict the link distributions (incoming and outgoing) and related statistics for three different types of entities from three datasets: OMIM, NCBI Gene, and KEGG. These three datasets exhibit the most links with the other Bio2RDF datasets (as shown in Fig. 1(c)). The selected types, namely Gene, Phenotype and Drug, have the largest numbers of entities in the corresponding datasets.

We observe that the outgoing/incoming degree distributions of entity links in the three datasets do not exhibit the power law pattern characteristic of scale-free networks (except the outgoing degree distribution for ncbigene:Gene). We find that there are fewer entities with an outgoing/incoming degree of 10 than one would expect from a power law distribution. This may be a consequence of overlap among the Bio2RDF datasets such that entities in one dataset are likely to link with at least a certain number of entities in the remaining datasets. Also, the exponents are large (close to 5) and p-values are very small (close to 0).[3] In particular, only four datasets link to KEGG and there is no many-to-one links

[2] These cross-references are created by the original data owners, while Bio2RDF just uniformly converts them to x-relations.

[3] The power law hypothesis should be rejected for p-values below 0.01 [11].

between their entities, thus the incoming degree distribution for kegg:Drug is sparse. Our results therefore differ from the calculated in-degree distribution of owl:sameAs on the 2010 Billion Triples Challenge (BTC) dataset [12]. We argue that this may be the result of link bias from the life science data providers.

In Table 1, we observe that a few entities link to hundreds of other entities, and most of them are widely studied genes and have many related publications or images. Due to the size of NCBI Gene, many entities are not linked by other entities, resulting in a large number (162,018) of source nodes. A direct outcome of our analysis is that we identified one super-connected node (linked to 75,000 nodes), which turned out to be the result of wrong parsing. This bug was fixed immediately by the authors and an updated dataset was released.

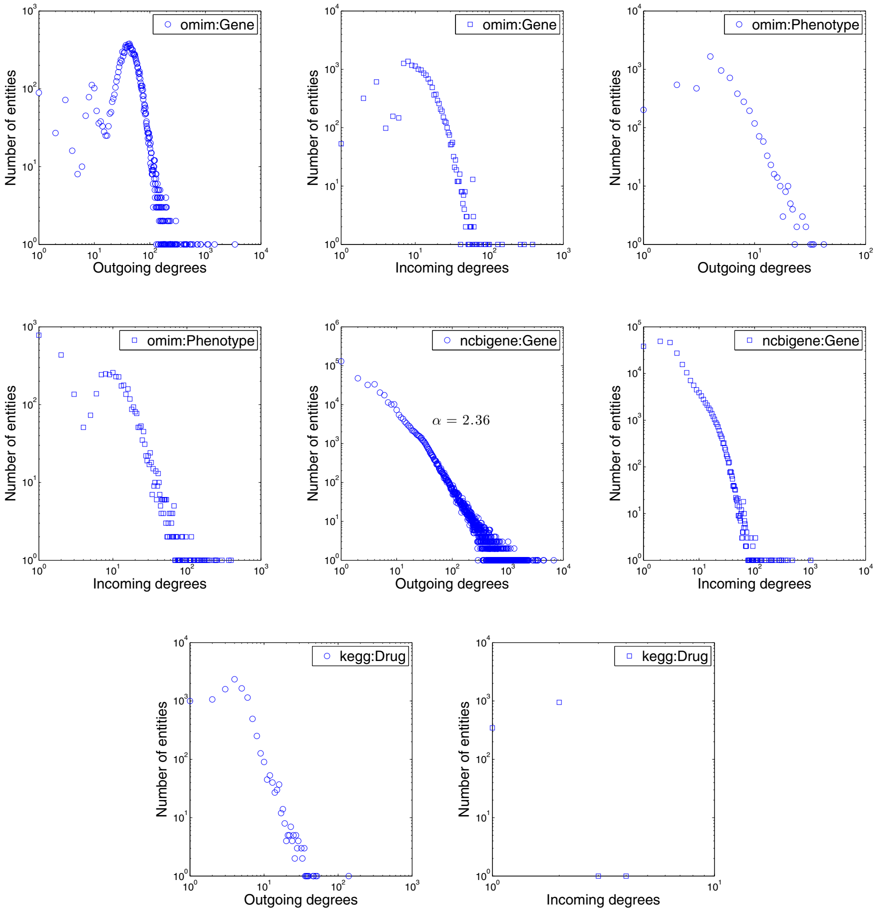


Fig. 2. Bio2RDF entity link distribution: (i) the figures are presented in log-log scale; and (ii) only the datasets in Bio2RDF are considered for computing incoming degrees.

Table 1. Degree analysis of entity links

Entity types	Entity number	Avg. outdegree	Avg. indegree	Max. degree	Isolated nodes	Sink nodes	Source nodes
omim:Gene	14,609	50.3	12.8	3,409	12	15	118
omim:Phenotype	5,825	5.2	10.5	414	34	38	1,027
ncbigene:Gene	394,479	10.8	2.9	6,798	0	0	162,018
kegg:Drug	10,082	4.5	0.2	139	0	0	8,785

4.2 Symmetry and Transitivity of Entity Links

As the entity link graph is directed, we seek to examine the symmetry of entity links. We find that only four pairs of datasets link to each other bi-directionally in Bio2RDF, which are DrugBank—KEGG, DrugBank—PharmGKB, OMIM—HGNC and OMIM—Orphanet.

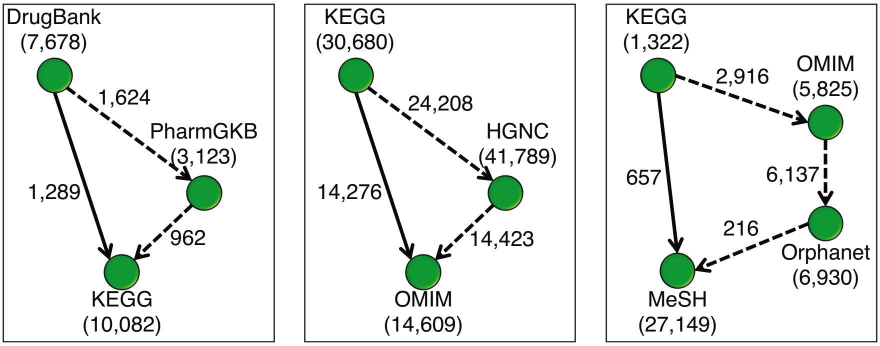
Table 2 lists the results on the symmetry of entity links in the four dataset pairs, where a reciprocal link indicates that two entities e_i, e_j are linked from both directions, a malposed link represents that e_i, e_j are linked in one direction (e.g. $e_i \rightarrow e_j$) but in the other direction e_j links to someone else (e.g. $e_h \leftarrow e_j$), and a missing link implies that either of the two directions is missing.

We observe that the symmetry of entity links varies between different pairs of datasets. For DrugBank—PharmGKB and OMIM—HGNC, a large proportion (99%) of entity links are reciprocal. A possible explanation is that one dataset just borrows the links from the other dataset and simply reverses them. On the other hand, DrugBank—KEGG and OMIM—Orphanet have different numbers of entity links from different directions and are mainly caused by their modeling divergence. For example, OMIM only creates the class omim:Phenotype instead of "Disease" and use it to link to orphanet:Disorder, which causes many links lost in the other direction, since a disorder may have many different phenotypes.

Also, we analyze the transitivity of entity links, which means that a direct entity link $e_i \rightarrow e_j$ may also be inferred from a transitive path through entity links $e_i \rightarrow e_k \rightarrow \ldots \rightarrow e_j$. We find three transitive examples in the Bio2RDF datasets and show them in Fig. 3, where an identical (or different) ending entity indicates that the same entity (different entities) can be achieved through a direct link and a transitive path from the same beginning entity. If the ending entity from the direct link is missing, it is called "missing direct", while the ending entity from the transitive path is missing, it is called "missing transitive".

Table 2. Symmetry analysis of entity links

	Forward	Backward	Reciprocal	Malposed	Missing	Total
DrugBank—KEGG	1,289	2,155	1,964	485	995	3,444
DrugBank—PharmGKB	1,624	1,619	3,210	4	29	3,243
OMIM—HGNC	14,274	14,423	28,514	6	177	28,697
OMIM—Orphanet	6,137	2,600	4,464	2,523	1,750	8,737



(a) Drugs (b) Human genes (c) Diseases

	Direct	Transitive	Identical	Different	Missing	Missing	Total
	links	paths	ending entities		direct	transitive	
Drugs	1,289	954	946	6	2	343	1,297
Human genes	14,276	14,250	14,236	5	9	40	14,290
Diseases	657	33	8	18	7	649	682

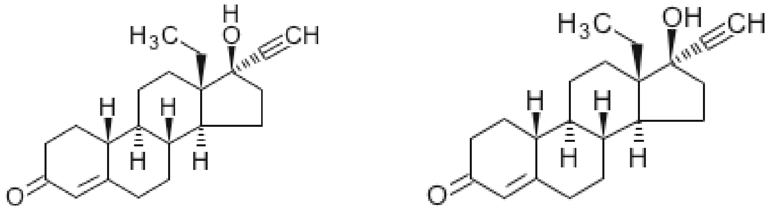
Fig. 3. Transitivity analysis of entity links: (i) the value in each parenthesis denotes the number of entities given a specified topic; and (ii) the solid arcs represent direct links between entities while the dashed arcs form transitive paths. The value on each arc denotes the number of entity links from one dataset to the other.

Our analysis reveals that most links are confirmed through transitivity among the human gene link network only. In the other two examples, there are some intermediate datasets, such as Orphanet, which affect the transitivity. To improve connectivity in the future, these datasets should be enhanced. Also, the number of links may decrease significantly with the increase of transitive path length. Therefore, the transitivity of entity links is often topic-dependent, and its accuracy varies in different contexts.

We take a deeper look at these transitive entity links. Fig. 4 exemplifies two different ending entities from DrugBank to KEGG, where one is from a direct link and the other is from a transitive path. The two drugs have different names but highly similar chemical structures (a.k.a. isomers), and their medical functions are similar as well. The DrugBank provider thinks that the two drugs are the same, while the KEGG provider uses different URIs to identify them without any equivalence relation. This example illustrates the difficulty of linking entities in the life sciences, caused by modeling divergence.

4.3 Evaluation of Entity Matching Approaches

According to our analysis above, we observe that an x-relation probably represents the owl:sameAs relation between two entities if they have the same or very similar types. Furthermore, although owl:sameAs is not a necessarily symmetric



(a) Levonorgestrel (kegg:D00950) using direct link

(b) Norgestrel (kegg:D00954) using transitive path

Fig. 4. Different ending entities from starting entity drugbank:DB00367

property [12], it is considered strongly equivalent only when reciprocal links exist. These observations guide us to use the reciprocal links between similarly-typed entities to build a benchmark and evaluate entity matching approaches.

For this purpose, we reuse the four pairs of datasets in Table 2. A commonly-used approach to entity matching in the life sciences is by comparing the labels of entities [17]. We develop four different string comparison algorithms based on Levenshtein, Jaro-Winkler, N-gram ($N = 2$) and Jaccard distances respectively to compute the similarity of labels. For each algorithm, we change the similarity threshold from 0.1 to 0.95 (step by 0.05) to achieve the highest F1-score, where F1-score $= \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$. In overall, the best threshold for each algorithm falls into $[0.5, 0.8]$. For instance, the best threshold for Jaro-Winkler is achieved at 0.75 when matching DrugBank and KEGG.

Linear regression and logistic regression are often employed to make use of more properties in entities. We re-implement the approach in [28] to identify five matched property pairs by 10-fold cross-validation and combine them using linear or logistic regression for similarity computation. The threshold is set to 0.25, which achieves the best F1-score.

Our experimental results are shown in Fig. 5. We observe that N-gram and Jaro-Winkler algorithms obtain the best F1-score among the string comparison algorithms. But their results are far from perfect, because there are many other useful properties. For example, by considering the property "chemical formula", the F1-scores achieved by logistic regression consistently rise up on all the drug datasets. For OMIM—Orphanet, the low F1-scores are caused by many-to-one links between the entities in omim:Phenotype and orphanet:Disorder.

Moreover, four small-scale drug datasets are provided in OAEI2010 and two entity matching systems participated in the test [15]. However, due to the reliability of reference links, the test did not make clear conclusions. We published our benchmark on our website and expect that it can help both researchers and practitioners in biomedicine and the SW verify their entity linking approaches and tools in future.

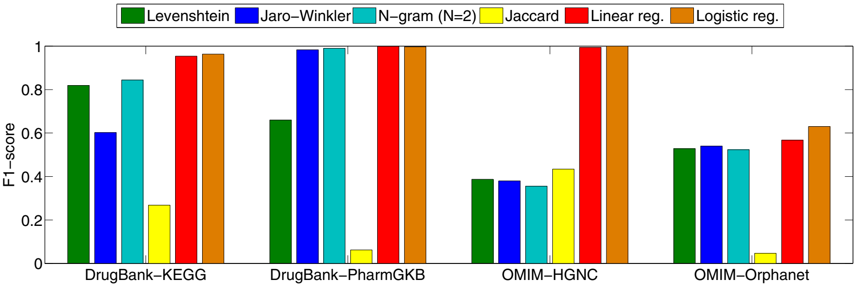


Fig. 5. F1-scores of entity linking approaches

5 Term Link Analysis

Ontology matching aims at creating mappings between terms (classes and properties) from different vocabularies [14], which has already been used for the term link analysis on the SW [17,20,24]. In order to investigate the link structure of terms in Bio2RDF, we define the term link graph as follows:

Definition 3 (term link graph). *A term link graph, denoted by* $(\mathbf{T}, \mathbf{M})$*, is an undirected graph, where* $\mathbf{T}$ *is the node set, and each node* $t_i \in \mathbf{T}$ *denotes a term;* $\mathbf{M}$ *is the edge set, and each edge* $(t_i, t_j) \in \mathbf{M}$ *exists iff there is a mapping between* t_i *and* t_j *with similarity greater than a specified threshold* $\eta \in [0, 1)$*.*

We construct the term link graph for Bio2RDF using Falcon-AO [21], which is a fully automatic ontology matching tool. The strength of Falcon-AO is that it combines various powerful matchers including two linguistic matchers and a structural matcher. We also enhance Falcon-AO with background knowledge to support synonym identification in the life sciences, e.g. "disease" vs. "disorder". It is worth noting that there are many approaches and tools can be used as alternatives for this analysis [14]. Among others, Ghazvinian *et al.* [17] used a simple lexical matching of preferred names and synonyms to generate mappings across all classes in 207 biomedical ontologies, while Nikolov and Motta [24] captured the mappings between classes by analyzing existing entity links. However, both of them did not consider the property matching problem.

For the 35 datasets in Bio2RDF, we create 82,689 mappings between classes, 1,540 mappings between object properties and 858 mapping between data properties, with similarity greater than 0.9. We set this threshold based on our empirical experience to achieve a high precision. Due to the simple structure of the Bio2RDF vocabularies, most mappings are found by linguistic matching (similar to [17]). We also note that the mappings between classes are largely in consistent with those discovered in [8] between SIO (Semanticscience Integrated Ontology) and 19 vocabularies in Bio2RDF Release 2. However, SIO only defines very general level properties (e.g. "has attribute"), and matches the properties in other vocabularies using the super/sub-property relation.

Table 3. Top-5 popular labels for classes and properties

(a) Classes

Labels	Distinct URIs
Resource	35
Gene	10
Drug	6
Enzyme	5
Pathway	5

(b) Object properties

Labels	Distinct URIs
x-uniprot	11
x-ncbigene	10
article	8
gene	8
source	8

(c) Data properties

Labels	Distinct URIs
synonym	25
definition	22
comment	9
chromosome	8
name	8

We extract the label of each term in these mappings and count the times of each label appearing in different terms (by ignoring their string cases). The five most frequently-occurred labels for classes, object and data properties are list in Table 3, where "Resource" is used in all the Bio2RDF datasets to define entities. However, unlike the findings in [17,20], the degree distribution of term links does not obey the power law, because there is a significant overlap between terms in different vocabularies, indicating that most biomedical data providers have very similar topic interests like genes and drugs. Besides, the created mappings can be used to support query rewriting in applications.

6 Correlation of Different Link Graphs

Earlier in this paper, we have showed our link analysis of datasets, entities and terms respectively. It is also natural for us to ask whether the three types of link graphs are correlated or independent. The *Spearman's rank correlation coefficient* (denoted by $\rho \in [-1, 1]$) measures the agreement degree between two rankings [23], which is suitable for answering our question. The sign of ρ indicates positive or negative correlation, while its absolute value assesses relative degree, with a larger absolute value being stronger correlation.

We abstract the entity and term link graphs to the dataset level and order the Bio2RDF dataset pairs based on their correlation values. For the entity link graph, the correlation value between two datasets D_i, D_j is defined as the number of direct entity links between D_i, D_j divided by the total number of entities in D_i, D_j. Note that both directions are involved, i.e. $D_i \rightarrow D_j$ and $D_j \rightarrow D_i$.

Inspired by [20], the correlation value of two datasets derived from the term link graph is defined as the ratio of the number of term mappings between the two datasets to the total number of their terms. Note that term mappings are undirected according to our definition.

For the dataset link graph in Fig. 1(a), the correlation value of two Bio2RDF datasets is obtained by finding the shortest path between them, with a shorter length being more strongly correlated. This measure has also been used in [10]. Therefore, we generate three rankings of all pairs of Bio2RDF datasets from the entity, term and dataset link graphs.

Table 4. Spearman's rank correlation coefficients among link graphs

	Dataset link graph	Entity link graph
Entity link graph	0.51	
Term link graph	0.42	0.16

Table 4 lists the correlation coefficients among the entity, term and dataset link graphs. The signs reflect that all the three graphs are positively correlated, where the dataset link graph has strong correlation with the entity link graph ($\rho = 0.51$) as well as the term link graph ($\rho = 0.42$). It can be explained as closer datasets in distance predicting more linked entities along with more matched classes and properties.

On the other hand, the correlation coefficient between the entity link graph and the term link graph is not strong ($\rho = 0.16$), which demonstrates that the number of linked entities contributes little to the overlap of vocabularies, since linked entities may centralize in a few classes, while entities under other classes have not been interlinked yet.

7 Related Work

Network analysis has long been used to study link structures in biomedicine and the Web. The small world phenomenon and the scale-free characteristic are often observed [1,3,7,11]. Recently, it has been conducted on the SW. Theoharis *et al.* [27] investigated the graph features of 250 ontologies and found that a majority of ontologies with a significant number of properties approximate powers for the total degree, while each ontology owns a few focal classes with considerable properties and subclasses. Ell *et al.* [13] introduced a set of label-related metrics including completeness, accessibility, unambiguity and multi-linguality to measure the current state of labeling the Web of Data. These works did not address the relations across different datasets.

To examine entity links, Ding *et al.* [12] carried out an empirical experiment of the owl:sameAs deployment status and used the statistics to focus discussion on the usage of owl:sameAs in the BTC2010 dataset. Our findings in Bio2RDF do not match their results in some aspects. Halpin *et al.* [18] found that owl:sameAs is widely misused to capture different degrees of equivalence, and its practical use is not limited to the case where two entities are truly identical but instead includes application scenarios where they can be treated as being operationally equivalent. Our investigation on the x-relations in Bio2RDF well confirms their observation. For a more general notion of links, Ge *et al.* [16] defined the object link graph according to the RDF data model and compared the graph features of two object link graphs crawled by the Falcons search engine in 2008 and 2009 respectively, containing some incomplete biomedical data.

Analysis of term links has also been performed. Ghazvinian *et al.* [17] analyzed the morphology of term mappings between 207 vocabularies in BioPortal and UMLS, while Hu *et al.* [20] extended this idea to a larger scale containing four thousand Web ontologies. Nikolov and Motta [24] created term mappings from declared coreference association (e.g. owl:sameAs) and co-typing, where a term mapping can hold either the equivalence or subsumption relation. Tordai *et al.* [27] empirically studied the quality of chains of (almost) equivalent terms in the domains of biomedicine, cultural heritage and library subject headings with multiple languages (English, Dutch, German and French). More generally, Cheng *et al.* [10] presented the declarative, topical and distributional relatedness between three thousand vocabularies and the correlation of these relatedness. Unlike these works, we holistically analyzed the life science Linked Data on the levels of datasets, entities and terms.

8 Discussion of Findings

The analytical results that we have presented in the previous sections allow us to make the following observations:

- Bio2RDF offers the biggest network of the life science Linked Data and also is a significant portion of Linked Open Data, which ensures the significance of our empirical study. Although our hypothesis is that the life science data network should be in consistence with that of the SW, we are surprised that some results turn out to be different than previously reported, e.g. the degree distribution of entity links does not strictly follow the power law.
- A dominated part of entities in Bio2RDF have been linked using x-relations, but the intended semantics of these entity links differs. When the meanings of two classes are identical or equivalent and their belonging datasets also have close topics, the entity links are likely to represent logical equivalence. Additionally, the classes and properties in different Bio2RDF datasets have large overlap and can be identified mainly by linguistic matching.
- Symmetric and transitive entity links exist in Bio2RDF, which can reinforce the correctness of these links, but their effectiveness is currently weakened due to the relatively small number. Adding more symmetric and transitive links should be an important future work for the life science data providers and aggregators (e.g. OpenLifeData[4]). Besides, the meanings of entity links may be shifted during transitive. In consideration of the quality and coverage of the entities and terms in Bio2RDF, we suggest to use KEGG, DrugBank and OMIM as the most prominent knowledge bases for applications in the life sciences.
- Previous work has demonstrated the effectiveness of using string matching to find linked entities or terms [15]. However, according to our benchmark, only considering the labels of entities may fail in some cases, e.g. comparing short-form abbreviations of gene names, while combining different properties and using simple machine learning algorithms like logistic regression achieve a good accuracy. However, discovering many-to-one links between entities is still a difficult problem that needs to be carefully studied.

[4] http://www.openlifedata.org/

9 Conclusion

In this paper, we described our analytical results of the life science Linked Data, obtained from the Bio2RDF project, so as to better inform the development of novel methods for exploring, querying and analyzing this wealth of knowledge. Our link analysis coupled with a benchmark give a first glimpse concerning the structure of the life science Linked Data, and offer new results by which we and others may utilize in future. A question raised from our study is how to make use of the findings to improve applications in the life sciences. Another future work is to repeat analysis on other linked biomedical data and compare the findings.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (Nos. 61370019 and 61223003), and in part by the National Social Science Foundation of China (No. 11AZD121). Wei Hu thanks the Stanford Center for Biomedical Informatics Research for hosting his visit.

References

1. Adamic, L.A., Huberman, B.A.: Power-Law Distribution of the World Wide Web. Science **287**(5461), 2115 (2000)
2. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing Linked Datasets with the VoID Vocabulary. W3C Interest Group Note (2011)
3. Barabási, A.-L., Gulbahce, N., Loscalzo, J.: Network Medicine: A Network-Based Approach to Human Disease. Nature Reviews Genetics **12**, 56–68 (2011)
4. Batchelor, C., et al.: Scientific lenses to support multiple views over linked chemistry data. In: Mika, P., et al. (eds.) ISWC 2014, Part I. LNCS, vol. 8796, pp. 98–113. Springer, Heidelberg (2014)
5. Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., Morissette, J.: Bio2RDF: Towards a Mashup to Build Bioinformatics Knowledge Systems. Journal of Biomedical Informatics **41**(5), 706–716 (2008)
6. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems **5**(3), 1–22 (2009)
7. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph Structure in the Web. Computer Networks **33**(1–6), 309–320 (2000)
8. Callahan, A., Cruz-Toledo, J., Ansell, P., Dumontier, M.: Bio2RDF release 2: improved coverage, interoperability and provenance of life science linked data. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 200–212. Springer, Heidelberg (2013)
9. Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y., Wild, D.J.: Chem2Bio2RDF: A Semantic Framework for Linking and Data Mining Chemogenomic and Systems Chemical Biology Data. BMC Bioinformatics **11**, 255 (2010)
10. Cheng, G., Qu, Y.: Relatedness between Vocabularies on the Web of Data: A Taxonomy and an Empirical Study. Journal of Web Semantics **20**, 1–17 (2013)

11. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-Law Distributions in Empirical Data. SIAM Review **51**(4), 661–703 (2009)
12. Ding, L., Shinavier, J., Shangguan, Z., McGuinness, D.L.: SameAs networks and beyond: analyzing deployment status and implications of owl:sameAs in linked data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 145–160. Springer, Heidelberg (2010)
13. Ell, B., Vrandečić, D., Simperl, E.: Labels in the web of data. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 162–176. Springer, Heidelberg (2011)
14. Euzenat, J., Shvaiko, P.: Ontology Matching, 2nd edn. Springer (2013)
15. Ferrara, A., Nikolov, A., Noessner, J., Scharffe, F.: Evaluation of Instance Matching Tools: The Experience of OAEI. Journal of Web Semantics **21**, 49–60 (2013)
16. Ge, W., Chen, J., Hu, W., Qu, Y.: Object link structure in the semantic web. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010, Part II. LNCS, vol. 6089, pp. 257–271. Springer, Heidelberg (2010)
17. Ghazvinian, A., Noy, N.F., Jonquet, C., Shah, N., Musen, M.A.: What four million mappings can tell you about two hundred ontologies. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 229–242. Springer, Heidelberg (2009)
18. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl:sameAs isn't the same: an analysis of identity in linked data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 305–320. Springer, Heidelberg (2010)
19. Hu, W., Chen, J., Zhang, H., Qu, Y.: How matchable are four thousand ontologies on the semantic web. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011, Part I. LNCS, vol. 6643, pp. 290–304. Springer, Heidelberg (2011)
20. Hu, W., Qu, Y.: Falcon-AO: A Practical Ontology Matching System. Journal of Web Semantics **6**(3), 237–239 (2008)
21. Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S.M., Martin, M., Le Novère, N., Parkinson, H., Birney, E., Jenkinson, A.M.: The EBI RDF Platform: Linked Open Data for the Life Sciences. Bioinformatics **30**(9), 1338–1339 (2014)
22. Myers, J.L., Well, A.D., Lorch Jr., R.F.: Research Design and Statistical Analysis, 3rd edn. Routledge (2010)
23. Nikolov, A., Motta, E.: Capturing emerging relations between schema ontologies on the web of data. In: International Workshop on Consuming Linked Data (2010)
24. Ruttenberg, A., Rees, J.A., Samwald, M., Marshall, M.S.: Life Sciences on the Semantic Web: The Neurocommons and Beyond. Briefings in Bioinformatics **10**(2), 193–204 (2009)
25. Theoharis, Y., Tzitzikas, Y., Kotzinos, D., Christophides, V.: On Graph Features of Semantic Web Schemas. IEEE Transactions on Knowledge and Data Engineering **20**(5), 692–702 (2008)

26. Tordai, A., Ghazvinian, A., van Ossenbruggen, J., Musen, M.A., Noy, N.F.: Lost in translation? empirical analysis of mapping compositions for large ontologies. In: International Workshop on Ontology Matching (2010)
27. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and maintaining links on the web of data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 650–665. Springer, Heidelberg (2009)
28. Xu, M., Wang, Z., Bie, R., Li, J., Zheng, C., Ke, W., Zhou, M.: Discovering missing semantic relations between entities in Wikipedia. In: Alani, H., et al. (eds.) ISWC 2013, Part I. LNCS, vol. 8218, pp. 673–686. Springer, Heidelberg (2013)