

Automatic Creation and Analysis of a Linked Data Cloud Diagram

Alexander Arturo Mera Caraballo¹, Bernardo Pereira Nunes^{1,4}, Giseli Rabello Lopes²,
Luiz André Portes Paes Leme³, Marco Antonio Casanova¹

¹Department of Informatics – Pontifical Catholic University of Rio de Janeiro, RJ, Brazil
{acaraballo, bnunes, casanova}@inf.puc-rio.br

²Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil
giseli@dcc.ufrj.br

³Fluminense Federal University, Niteroi, RJ, Brazil
lapaesleme@ic.uff.br

⁴Federal University of the State of Rio de Janeiro, RJ, Brazil
bernardo.nunes@uniriotec.br

Abstract. Datasets published on the Web and following the Linked Open Data (LOD) practices have the potential to enrich other LOD datasets in multiple domains. However, the lack of descriptive information, combined with the large number of available LOD datasets, inhibits their interlinking and consumption. Aiming at facilitating such tasks, this paper proposes an automated clustering process for the LOD datasets that, thereby, provide an up-to-date description of the LOD cloud. The process combines metadata inspection and extraction strategies, community detection methods and dataset profiling techniques. The clustering process is evaluated using the LOD diagram as ground truth. The results show the ability of the proposed process to replicate the LOD diagram and to identify new LOD dataset clusters. Finally, experiments conducted by LOD experts indicate that the clustering process generates dataset clusters that tend to be more descriptive than those manually defined in the LOD diagram.

Keywords: Linked Data Cloud Analysis, Automatic Clustering, Domain Identification, Community Detection Algorithms.

1 Introduction

The Linked Data principles established a strong basis for creating a rich space of structured data on the Web. The potentiality of such principles encouraged the government, scientific and industrial communities to transform their data to the Linked Data format, creating the so-called Linked Open Data (LOD) cloud. An essential step of the publishing process is to interlink new datasets with those in the LOD cloud to facilitate the exploration and consumption of existing data. Although frameworks to help create links are available, such as LIMES [1] and SILK [2], the selection of da-

tasets to interlink with a new dataset is still a manual and non-trivial task. One possible direction to facilitate the selection of datasets to interlink with would be to classify the datasets in the LOD cloud by domain similarity and to create expressive descriptions of each class. Thus, the publisher of a new dataset would select the class closest to his dataset and try to interlink his dataset with those in the class.

The *LOD diagram* [3,4], perhaps the best-known classification of the datasets in the LOD cloud, adopted the following categories: “Media”, “Government”, “Publications”, “Life Sciences”, “Geographic”, “Cross-domain”, “User-generated Content” and “Social Networking”. However, the fast growth of the LOD cloud makes it difficult to manually maintain the LOD diagram. To address this problem, we propose a community analysis of the LOD cloud that leads to an automatic clustering of the datasets into communities and to a meaningful description of the communities. The process has three steps. The first step creates a graph to describe the LOD cloud, using metadata extracted from dataset catalogs. The second step uses community detection algorithms to partition the LOD graph into *communities* (also called *clusters*) of related datasets. The last step generates descriptions for the dataset communities by applying dataset profiling techniques. As some of the datasets may contain a large number of resources, only a random sample of each dataset is considered. For each dataset community, this step generates a *profile*, expressed as a vector, whose dimensions correspond to relevance scores for the 23 top-level categories of Wikipedia.

The resulting partition of the LOD graph into communities, with the descriptions obtained, may help data publishers search for datasets to interlink their data as follows. Consider a new dataset d to be published as Linked Data; the same profiling technique used in the process we propose may be used to generate a profile for d , expressed as a vector, as in Step 3. Then, a similarity measure (e.g., cosine-based) may be used to compute the similarity between the profile of d and the profile of each dataset community. Finally, the data publisher may receive recommendations for the community with the highest similarity value. This suggested recommendation process is not the focus of this paper, but it is one of the major motivations of this work.

To summarize, the main contributions of this paper are: (i) an automatic clustering of the LOD datasets which is consistent with the traditional LOD diagram, taken as ground truth; and (ii) an automatic process that generates descriptions of dataset communities.

The remainder of this paper is structured as follows. Section 2 presents background concepts. Section 3 presents the proposed process. Sections 4, 5 and 6 describe the experimental setup, the results and an extensive analysis of the generated communities and their descriptions, respectively. Section 7 reviews the literature. Finally, Section 8 summarizes the contributions and outlines further work.

2 Background

2.1 LOD Concepts

A *dataset* is simply a set t of RDF triples. A resource, identified by an RDF URI reference s , is *defined in* t iff s occurs as the subject of a triple in t . Given two datasets t

and u , a *link* from t to u is a triple of the form (s,p,o) , where s is an RDF URI reference identifying a resource defined in t and o is an RDF URI reference identifying a resource defined in u . A *linkset* from t to u is a set of links from t to u .

The set of RDF datasets publicly available is usually referred to as the *LOD cloud*.

The *LOD graph (or the LOD network)* is an undirected graph $G=(S,E)$, where S denotes a set of datasets in the LOD cloud and E contains an edge (t,u) iff there is at least one linkset from t to u , or from u to t .

A *LOD catalog* describes the datasets available in the LOD cloud. Datahub¹ and the Mannheim Catalog² are two popular catalogs. LODStats³ collects statistics about datasets to describe their internal structure (e.g. vocabulary/class/property usage, number of triples, linksets). The LOD Laundromat⁴ generates a clean version of the LOD cloud along with a metadata graph with structural data.

A *LOD diagram* is a visual representation of the structure of the LOD cloud. At least three versions of the structure of the LOD cloud are currently available [3]. Schmachtenberg et al. [4] provides the most comprehensive statistics about the structure and content of the LOD cloud (as of April 2014). This version of the LOD cloud comprises 1,014 datasets, of which only 570 have linksets. In total, 2,755 linksets (both in- and outlinks) express a relationship between the datasets contained in this version of the LOD cloud. The datasets are divided into eight topical domains, namely, “Media”, “Government”, “Publications”, “Life Sciences”, “Geographic”, “Cross-domain”, “User-generated Content” and “Social Networking”. The datasets are not uniformly distributed per topical domain: “Government” and “Publication” are the largest domains, with 23.85% and 23.33% of all datasets, respectively; “Media” is the smallest domain, containing only 3.68% of all datasets. Table 1 presents the number of datasets in each topical domain, for which linksets are defined. We highlight that the wide variation of the size among the domains represents an additional challenge to community detection/clustering algorithms [5].

Table 1. Number of datasets and linksets per topical domain.

Topical domain	#Datasets	#Inlinks	#Outlinks
Media	21	55	39
Government	136	271	330
Publications	133	772	862
Geographic	24	171	56
Cross-Domain	40	345	180
Life Sciences	63	144	161
Social Networking	90	912	986
User-generated content	42	85	141

¹ <http://datahub.io/>

² <http://linkeddatacatalog.dws.informatik.uni-mannheim.de/>

³ <http://stats.lod2.eu/>

⁴ <http://lodlaundromat.org>

2.2 Communities and Community Detection Algorithms

Let $G=(S,E)$ be an undirected graph and $G_C=(S_C,E_C)$ be a subgraph of G (that is, $S_C \subseteq S$ and $E_C \subseteq E$). Let $|s|$ denote the cardinality of a set s .

The *intra-cluster density* of G_C , denoted $\delta_{int}(G_C)$, is the ratio between the number of edges of G_C and the number of all possible edges of G_C and is defined as follows:

$$\delta_{int}(G_C) = \frac{|E_C|}{|S_C| \cdot (|S_C| - 1) / 2}$$

Let $\gamma(G_C)$ denote the set of all edges of G that have exactly one node is in S_C . The *inter-cluster density* of G_C , denoted $\delta_{ext}(G_C)$, measures the ratio between the cardinality of $\gamma(G_C)$ and the number of all possible edges of G that have exactly one node is in S_C and is defined as follows:

$$\delta_{ext}(G_C) = \frac{|\gamma(G_C)|}{|S_C| \cdot (|S| - |S_C|)}$$

The average link density of $G=(S,E)$, denoted $\delta(G)$, is the ratio between the number of edges of G and the maximum number of possible edges of G :

$$\delta(G) = |E| / (|S|(|S|-1)/2)$$

For the subgraph G_C to be a community, $\delta_{int}(G_C)$ has to be considerably larger than $\delta(G)$ and $\delta_{ext}(G_C)$ has to be much smaller than $\delta(G)$.

The *edge betweenness* [6] of an edge (t,u) in E is the number of pairs (w,v) of nodes in S for which (t,u) belongs to the shortest path between w and v .

Community detection algorithms search, implicitly or explicitly, for the best trade-off between a large $\delta_{int}(G_C)$ and a small $\delta_{ext}(G_C)$. They are usually classified as *non-overlapping* and *overlapping*. In *non-overlapping* algorithms, each node belongs to a single community. An example is the *Edge Betweenness Method* (EBM) [6], which finds communities by successively deleting edges with high edge betweenness. In *overlapping* algorithms, a node may belong to multiple communities. An example is the *Greedy Clique Expansion* algorithm (GCE) [7], which first discovers maximum cliques to be used as seeds of communities and then greedily expands these seeds by optimizing a fitness function. Another example is the *Community Overlap Propagation Algorithm* (COPRA) [8], which follows a label propagation strategy (where the labels represent the communities).

2.3 Clustering Validation Measures

Clustering validation measures are used to validate a clustering (or community detection) strategy against a ground truth.

Let U be the *universe*, that is, the set of all elements. Let $C = \{C_1, C_2, \dots, C_m\}$ and $T = \{T_1, T_2, \dots, T_n\}$ be two sets of subsets of U .

The definitions that follow are generic, but the reader may intuitively consider U as the set of all datasets in the LOD cloud, C as a set of dataset clusters, obtained by one of the clustering algorithms, and T be a set of sets of LOD datasets taken as the ground truth (i.e., the topical domains of the LOD diagram).

Purity [9] is a straightforward measure of cluster quality that is determined by simply dividing the number of elements of the most frequent domain contained in each cluster by the total number of elements. Purity ranges from 0 to 1, where higher values indicate better clusters with respect to the ground truth, and is defined as follows:

$$purity(\mathbf{C}, \mathbf{T}) = \frac{1}{|\mathbf{U}|} \sum_{i=1, \dots, m} \max_j (|C_i \cap T_j|)$$

Unlike purity, the *normalized mutual information* (NMI) [9] offers a trade-off between the number of clusters and their quality. Intuitively, NMI is the fraction of mutual information that is contained in the current clustering representation. NMI ranges from 0 to 1, where higher values indicate better clusters with respect to the ground truth, and is defined as follows:

$$NMI(\mathbf{C}, \mathbf{T}) = \frac{I(\mathbf{C}, \mathbf{T})}{(H(\mathbf{C}) + H(\mathbf{T})) / 2}$$

where $I(\mathbf{C}, \mathbf{T})$ represents the *mutual information* between \mathbf{C} and \mathbf{T} and is defined as:

$$I(\mathbf{C}, \mathbf{T}) = \sum_{i=1, \dots, m} \sum_{j=1, \dots, n} \frac{|C_i \cap T_j|}{|\mathbf{U}|} \log \left(\frac{|\mathbf{U}| \cdot |C_i \cap T_j|}{|C_i| \cdot |T_j|} \right)$$

and $H(\mathbf{C})$ is the *entropy* of \mathbf{C} and is defined as:

$$H(\mathbf{C}) = - \sum_{i=1, \dots, m} \frac{|C_i|}{|\mathbf{U}|} \log \left(\frac{|C_i|}{|\mathbf{U}|} \right)$$

and likewise for $H(\mathbf{T})$, the entropy of \mathbf{T} .

The *Estimated Mutual Information* (EMI) [10] measures the dependence between \mathbf{C} and \mathbf{T} (intuitively, the identified clusters and the topical domains in the LOD diagram). EMI is an $m \times n$ matrix, where each element is defined as follows:

$$EMI_{i,j} = \frac{m_{i,j}}{M} \cdot \log \left(M \cdot \frac{m_{i,j}}{\sum_{a=1}^n m_{i,a} \cdot \sum_{b=1}^m m_{b,j}} \right)$$

where

- $[m_{i,j}]$ is the *co-occurrence matrix* of \mathbf{C} and \mathbf{T} , with $m_{i,j} = |C_i \cap T_j|$, for $i \in [1, m]$ and $j \in [1, n]$
- $M = \sum_{i=1}^m \sum_{j=1}^n m_{i,j}$

2.4 Dataset Profiling Techniques

Profiling techniques address the problem of generating dataset descriptions. We will use in this paper the profiling technique described in [11], that generates *profiles* or *fingerprints* for textual resources. The method has five steps:

1. Extract entities from a given textual resource.
2. Link the extracted entities to English Wikipedia articles.
3. Extract English Wikipedia categories for the articles.

4. Follow the path from each extracted category to its top-level category and compute a vector with scores for the top-level categories thus obtained.
5. Perform a linear aggregation in all dimensions of the vectors to generate the final profile, represented as a histogram for the 23 top-level categories of the English Wikipedia.

3 The Dataset Clusterization and Dataset Community Description Processes

The proposed process has three main steps (see Figure 1):

1. Construction of the LOD graph.
2. Dataset clusterization.
3. Dataset community description.

The first step of the process creates a graph that describes the LOD cloud, using metadata extracted from metadata catalogs (c.f. Section 2.1).

The second step clusters the datasets represented as nodes of the LOD graph. It applies community detection algorithms to partition the LOD graph into *communities* (also called *clusters* or *groups*) of related datasets. Intuitively, a set of datasets forms a community if there are more linksets between datasets within the community than linksets interlinking datasets of the community with datasets in rest of the LOD cloud (c.f. Section 2.2).

The last step generates descriptions for the dataset communities by applying a dataset profiling technique to the datasets in each community C_i identified in the previous step. As some of the datasets may contain a large number of resources, only a random sample of each dataset is considered. Furthermore, to generate the labels that describe C_i , the profiling technique considers the literals of the datatype properties `rdfs:Label`, `skos:subject`, `skos:prefLabel` and `skos:altLabel` of the sampled resources. We recall that this step adopts the profiling technique described in Section 2.4 to generate community descriptions.

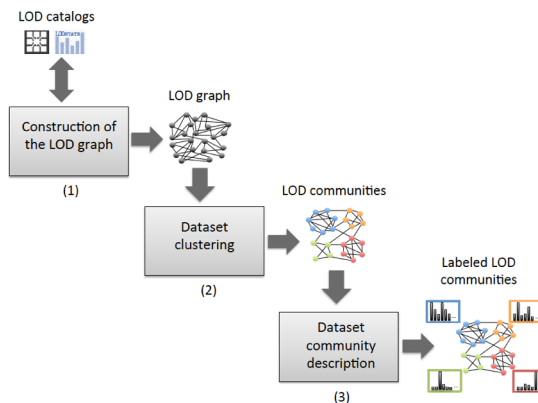


Fig. 1. Community analysis process of the LOD.

4 Evaluation Setup

This section details the evaluation setup of the proposed process. Section 4.1 covers the construction of the LOD graph and describes the ground truth. Section 4.2 introduces the community detection algorithms used and discusses how the resulting communities are evaluated by taking into account the clustering validation measures described in Section 2.3. Finally, Section 4.3 analyses the labels assigned to the resulting communities, considering the expressiveness and the ability to represent the content of the datasets belonging to each community.

4.1 Construction of the LOD graph and Description of the Ground Truth

To construct a LOD graph, we extracted all datasets from the Mannheim Catalog, along with their content metadata: title, description, tags and linksets. For the sake of simplicity and comparison between the ground truth and the proposed approach, we refer to the topical domains also as communities.

As ground truth, we adopted the LOD diagram described in [4] (see Section 2.1).

4.2 Setup of the Dataset Clusterization Step

Three algorithms traditionally used in community detection and clustering problems were considered as an attempt to reproduce the LOD diagram: Greedy Click Expansion (GCE), Community Overlap PPropagation Algorithm (COPRA) and the Betweenness Method (EBM) (see Section 2.2). The choice of these three algorithms was based on their previously reported performance in real world scenarios [12, 13]. We used Purity, Normalized Mutual Information (NMI) and Estimated Mutual Information (EMI) (see Section 2.3) as clustering validation measures, which were estimated by comparing the results obtained by algorithms and the ground truth.

A brief description of parameterization of the three algorithms goes as follows:

- EBM: Table 2 shows the top 10 best configurations for EBM in order to reproduce the results found in the ground truth. Very briefly, the number of edges with the highest betweenness that must be removed from the LOD graph in order to detect the communities was used as stopping criterion.

Table 2. Top 10 best configurations for EBM by decreasing order of NMI.

Number of removed edges	Purity	NMI
600	0.60291	0.49287
550	0.60109	0.48619
300	0.56831	0.47381
650	0.54645	0.46870
700	0.51730	0.45848
500	0.60474	0.45061
750	0.49545	0.44958
450	0.58106	0.44949
800	0.46812	0.44551
850	0.39891	0.42707

Table 3. Top 10 best configurations for GCE by decreasing order of NMI.

Clique size	Overlapping rate	Alpha	Phi	Purity	NMI
3	0.0	0.8	0.2	0.42076	0.57263
3	0.0	1.0	0.8	0.36430	0.55509
3	0.0	1.0	0.2	0.38251	0.54227
3	0.1	0.8	0.6	0.49362	0.51040
3	0.0	1.2	0.2	0.46630	0.51022
3	0.1	1.2	0.2	0.48816	0.50926
3	0.0	0.8	0.4	0.34426	0.50534
3	0.0	1.2	0.8	0.50820	0.50148
3	0.2	1.0	0.2	0.56648	0.49747
3	0.3	0.8	0.2	0.48452	0.49542

- GCE: Table 3 shows the top 10 best configurations for GCE used to reproduce the results found in the ground truth. Very briefly, the *Alpha* and *Phi* parameters were used to control the greedy expansion and to avoid duplicate cliques/communities, respectively.
- COPRA: Table 4 shows the best configuration for COPRA. As COPRA is non-deterministic, the tuning of its parameters was obtained by the average of 5-cycle runs.

Unlike EBM, GCE and COPRA are capable of finding overlapping communities. However, as the ground truth defines non-overlapping communities, these algorithms obtained the best results when the overlapping rate/parameter was set to 0 (no overlap between datasets) and 1 (one label per dataset), respectively.

4.3 Setup of the Dataset Community Description Step

Although the Mannheim Catalog lists 1,014 datasets, only a fraction of the listed datasets has SPARQL endpoints available. At the time of this evaluation, approximately 56% of the SPARQL endpoints were up and running. For each available dataset, a sample of 10% of its resources were extracted and used as input to the *fingerprints* algorithm (see Sections 2.4 and 3), which assigned labels to the communities automatically generated by the best performing parameterization of the GCE algorithm.

Table 4. Best quality results for the community detection/clustering algorithms.

Algorithm	#Clusters	Purity	NMI
GCE	6	0.42	0.57
COPRA	4	0.30	0.32
EBM	18	0.60	0.49

5 Results

The first part of the discussion addresses the performance of the dataset clusterization step. The second part presents the results for the dataset community description step.

5.1 Performance of the Dataset Clusterization Step

Quality of the generated communities. As shown in Table 4, GCE obtained the highest NMI value, 0.57, and EBM the highest purity value, 0.60. The high NMI value achieved by GCE indicates a mutual dependence between the communities found by the algorithm and those described in the ground truth. Despite the highest purity value obtained by EBM, this technique was not consistent with the communities in the ground truth. COPRA obtained low values for both purity and NMI, indicating that the resulting communities and those induced by the ground truth do not match.

Communities detected. Table 5 shows the co-occurrence and estimated mutual information matrices for the best performing parameterization of the GCE algorithm. The first column shows the communities (domains) of the ground truth, whereas columns labeled 0-5 represent the communities found by GCE. The light gray cells mark the highest dependencies between the topical domains extracted from the ground truth and the communities generated by GCE. Note that, due to the low level of dependency between the ground truth categories “Cross-Domain” and “User-Generated Content” (UGC) and the clusters found by GCE, datasets in these ground truth categories communities are possibly split over several clusters.

Table 5. Co-occurrence and EMI matrices of the GCE result.

Domain/Community	0	1	2	3	4	5	0	1	2	3	4	5
Social Networking	0	88	0	0	0	0	0	0.262	0	0	0	0
UGC	0	4	0	0	0	0	0	-0.005	0	0	0	0
Geographic	0	2	4	0	0	0	0	-0.003	0.013	0	0	0
Publications	37	4	1	1	0	0	0.092	-0.013	-0.00	-0.002	0	0
Cross-Domain	1	2	0	0	0	0	-0.002	-0.005	2	0	0	0
Life Sciences	0	2	0	13	24	0	0	-0.007	0	0.046	0.095	0
Government	1	1	10	1	0	59	-0.004	-0.006	0	-0.003	0	0.150
Media	0	2	0	1	0	0	0	-0.003	0.018	0.001	0	0

5.2 Performance of the Dataset Community Description Step

Table 6 shows the labels generated by the dataset community description method adopted (see Section 2.4). These labels were assigned to the communities found by the best performing parameterization of the GCE algorithm. The first column shows the 23 top-level categories of Wikipedia, whereas columns labeled 0-5 represent the communities found by GCE. To facilitate a comparison between the labels in different communities, we normalized the scores assigning 1.0 to the category with the highest score. The light gray cells mark the strongest relations between the categories from the generated labels and the communities generated by GCE. We recall that Table 1 shows the labels assigned to the communities in the ground truth.

Table 6. Histograms of top-level categories for each community structure.

Category / Community	0	1	2	3	4	5
Agriculture	0	0	0.39	0.03	0.02	0.03
Applied Science	0.80	0.34	0.37	0.06	0.11	0.03
Arts	0.03	0.11	0	0	0.01	0.03
Belief	0.03	0.04	0	0	0	0.02
Business	0.59	0.53	0.11	0.03	0.03	0.27
Chronology	0.04	0.15	0.02	0.01	0	0.06
Culture	0.13	0.19	0.27	0	0.03	0.11
Education	0.20	0.06	0.06	0.04	0.12	0.08
Environment	0.01	0.03	0.40	0.02	0.02	0.10
Geography	0.05	0.11	1.00	0.13	0.03	0.70
Health	0.05	0.06	0.41	0.18	0.65	0.03
History	0.06	0.03	0.11	0.06	0.02	0.13
Humanities	0.04	0.08	0	0	0.2	0
Language	0.20	0.10	0.01	0	0.02	0.03
Law	0.04	0.45	0.10	0.01	0.02	0.24
Life	0.08	0.03	0.96	1.00	1.00	0.02
Mathematics	0.60	0.03	0.03	0.03	0.03	0.02
Nature	0.29	0.08	0.24	0.03	0.06	0.03
People	0.02	0.52	0.02	0.01	0.03	1.00
Politics	0.05	0.35	0.12	0.03	0.01	0.65
Science	1.00	0.16	0.26	0.03	0.10	0.03
Society	0.32	1.00	0.14	0.06	0.05	0.32
Technology	0.61	0.37	0.11	0.01	0.02	0.08

6 Discussion and Analysis

6.1 An Analysis of the Dataset Clusterization Results

This section analyses the dataset clusterization results. The analysis compares the dataset communities found in the clustering step – referred to as *Community 0* to *Community 5* – with the dataset topical domains defined in the LOD diagram [4] – “Media”, “Government”, “Publications”, “Geographic”, “Cross-Domain”, “Life Sciences”, “Social Networking” and “User-generated content” – taken as ground truth.

As shown in Section 5, the GCE algorithm did not recognize as communities the datasets classified in the “Cross-domain” and “Media” domains. A possible reason for the lack of a cross-domain community lies in its own nature, that is, cross-domain datasets tend to be linked to datasets from multiple domains, acting as hubs for different communities. Another (interesting) reason is that cross-domain datasets do not contain a large number of links between themselves. The lack of links between cross-domain datasets results in a subgraph with low density, which GCE does not consider a new community. Nevertheless, if overlapping rates are considered, datasets that belong to several communities may generate a cross-domain community. Likewise, the “Media” community presented a low density due to its low number of linksets.

Community 0 presents a high concentration of datasets from the “Publications” domain, including datasets from the ReSIST project⁵, such as `rkb-explorer-acm`, `rkb-explorer-newcastle`, `rkb-explorer-pisa` and `rkb-explorer-budapest`. This led us to assume that this community is equivalent to the “Publications” domain.

Community 1 is the largest community among those recognized and contains mostly datasets from the “Social Networking” domain. This community includes datasets such as `statusnet-postblue-info`, `statusnet-fragdev-com`, `statusnet-bka-li` and `statusnet-skilledtestes-com`.

Contrasting with the previous communities, *Community 2* includes datasets from two different domains, “Government” and “Geographic”. Note that datasets in these two domains share a considerable number of linksets, which led GCE to consider them in the same community. Government datasets often provide statistical data about places, which may justify such a large number of linksets between them. *Community 2* includes datasets from the “Government” domain, such as `eurovoc-in-skos`, `gemet`, `umthes`, `eea`, `eea-rod`, `eurostat-rdf` and `fu-berlin-eurostat`. It also includes datasets from the “Geographic” domain, such as `environmental-applications-reference-thesaurus` and `gadm-geovocab`.

Communities 3 and *4* are equivalent to only one domain, “Life Sciences”. Intuitively, the original “Life Sciences” domain was split into *Community 3*, containing datasets such as `uniprot`, `bio2rdf-biomodels`, `bio2rdf-chembl` and `bio2rdf-reactome`, and into *Community 4*, containing datasets such as `pub-med-central`, `bio2rdf-omim` and `bio2rdf-mesh`. A distinction between these two communities becomes apparent by inspecting the datasets content: *Community 3* is better related to Human Biology data (about molecular and cellular biology), whereas *Community 4* is better related to Medicine data (about diagnosis and treatment of diseases).

Finally, *Community 5* groups datasets from the “Government” domain. Examples of datasets in this community are `statistics-data-gov-uk`, `reference-data-gov-uk`, `opendatacommunities-imd-rank-2010` and `opendatascotland-simd-education-rank`.

6.2 An Analysis of the Dataset Community Description Results

This section analyses the dataset community description results (see Table 6). For each dataset community, the analysis compares the 23-dimension vector description automatically assigned by the fingerprint approach with the labels manually assigned by the ground truth. In what follows, we say that a vector v has a peak for dimension i iff $v_i \geq 0.50$.

Community 0, which is equivalent to the “Publications” domain, is described by a vector with peaks for “Applied Science”, “Business”, “Mathematics”, “Science” and “Technology”. The presence of five categories shows the diversity of the data in this community. We consider that the label “Publications” assigned by the ground truth

⁵ <http://www.rkbexplorer.com/>

classification is better related to the tasks developed in this community than the semantics of the data itself. The rationale behind this argument is that the data come from scholarly articles published in journals and conferences.

Community 1, which is equivalent to the “Social Networking” domain, is described by a vector with peaks for “Business”, “People” and “Society”. Clearly, the vector was able to capture the essence of social data, covering topics related to the society in general.

Community 2, which has datasets from two different domains, “Government” and “Geographic”, is described by a vector with peaks for “Geography” and “Life”. Geographic data are available in various domains and, for this reason, the data cannot be described by a single category.

Community 3, which is partially equivalent to the “Life Sciences” domain, is described by a vector with a single peak for “Life”, which is similar to the manually assigned domain. *Community 3* is complemented by *Community 4*, whose vector has peaks for “Health” and “Life”. Taking into account these two vectors, we may identify datasets in this community with two different content profiles.

Community 5, which is equivalent to the “Government” domain, is described by a vector with peaks for “Geographic”, “People” and “Politics”. The vector also has significant values for “Business”, “Law” and “Society”. In general, datasets in this community are related to government transparency. For this reason, the vector for *Community 5* shows an interesting presence of “People”, “Society” and “Politics”.

7 Related Work

Analyses of the LOD cloud structure followed a wide variety of strategies, ranging from the use of community detection algorithms [12, 13], statistical techniques [4, 14] to dataset profiling techniques [11, 15-17]. Similarly to previous approaches, we combined and applied several techniques from different fields to analyze and generate an automatic version of the manually created LOD diagram. As already know by the LOD community, every couple of years a manual analysis of the state of the LOD cloud is performed and a new LOD diagram is published (see [3, 4]). At the time of the experiments described in this paper, the most recent report was conducted by Schmachtenberg et al. [4] in late 2014 showing the increasing adoption of the LOD principles, the most used vocabularies by data publishers, the degree distribution of the datasets, an interesting manual classification of datasets by topical domain (media, government, publications, geographic, life sciences, cross-domain, user generated content and social networking), among others.

Although such sequence of analyses was widely accepted and adopted by the LOD community, other works presented similar analyses under different perspectives, as those presented by Rodriguez [14]. Based on community detection algorithms, he identified more clusters/communities (Biology, Business, Computer Science, General, Government, Images, Library, Location, Media, Medicine, Movie, Music, Reference and Social) in the LOD cloud. We remind that the main purpose of this work is not only to assign labels to clusters of LOD cloud but to automatically identify and gener-

ate a more up to date version of the LOD diagram alleviating the arduous task of data publishers to link their data to others and finding popular vocabularies and others relevant statistics of the actual state of the LOD cloud.

Community detection algorithms are crucial towards an automatic method to generate LOD diagrams. A number of techniques for identifying communities in graph structures were studied by Fortunato [12]. Basically, a community is represented by a set of nodes that are highly linked within a community and that have a few or no links to other communities. Fortunato also presented techniques to validate the clusters found, which we also adopted (see Section 4). Xie et al. [13] also explored community detection algorithms. Unlike Fortunato’s work, they also considered in their analysis the overlapping structure of communities, i.e., when a community (of datasets) belongs to more than one category. From the 14 algorithms examined by Fortunato, we used the top two best performing overlapping algorithms, GCE and COPRA, in our experiments, as well as a non-overlapping algorithm, which we called the Edge Betweenness Method [6].

As community detection algorithms essentially analyze graph structures to find communities, profiling techniques also play an important role in the identification, at a content level, of the relatedness between datasets. For instance, Emaldi et al. [17], based on a frequent subgraph mining (FSM) technique, extracted structural characteristics of datasets to find similarities among them. Lalithsena et al. [16] relied on a sample of extracted instances from datasets to identify the datasets topical domains. Topics were extracted from reference datasets (such as Freebase) and then ranked and assigned to each dataset profile. Analogously, Fetahu et al. [15] proposed an automated technique to create structured topic profiles for arbitrary datasets through a combination of sampling, named entity recognition, topic extraction and ranking techniques. A more generic approach to create profiles on the Web was presented by Kawase et al. [11]. Kawase’s approach generates a histogram (called *fingerprints*) for any text-based resource on the Web based on the 23 top-level categories of the Wikipedia ontology. In this paper, we evaluated Kawase’s technique, which demonstrated to be suitable to determine the topical domain of dataset communities. The drawback of Fetahu’s approach in our scenario is the large number of categories assigned to a given dataset, which hinders the identification and selection of the most representative topics of a dataset and, consequently, of a community.

8 Conclusions

This paper presented a novel, automatic analysis of the Linked Open Data cloud through community detection algorithms and profiling techniques. The results indicate that the best performing community detection algorithm is the GCE algorithm, with NMI and purity values of 0.57 and 0.42, respectively. Although the EBM algorithm obtained the highest purity value, the high number of communities led to a low NMI value. The mutual dependence between the communities generated using GCE and those from the ground truth is also not high, but, as discussed in Section 6, the lack of linksets between datasets in some domains, such as “Cross-Domain”, implies

a need for the re-organization of datasets as well as the merging and splitting of communities.

The next part of the evaluation focused on comparing the labels manually assigned by the ground truth with the description automatically generated by the profiling technique. The manual labeling process considered as its classification criterion the nature of the data, whereas the automatic process relied on the contents of the datasets to generate the community labels. The experimental results showed that the proposed process automatically creates a clusterization of the LOD datasets which is consistent with the traditional LOD diagram and that it generates meaningful descriptions of the dataset communities. Moreover, the process may be applied to automatically update the LOD diagram to include new datasets. For additional information, including graphical visualizations and detailed results, we refer the reader to the Web site available at <http://drx.inf.puc-rio.br:8181/Approach/communities.jsp>

As for future work, we plan to define a recommendation approach based on previous works [18-21], which includes the proposed process, to help data publishers search for datasets to interlink their data.

Acknowledgments

This work was partly funded by CNPq under grant 444976/2014-0, 303332/2013-1, 442338/2014-7 and 248743/2013-9 and by FAPERJ under grant E-26/201.337/2014. The authors would also like to thank the Microsoft Azure Research Program by the cloud resources awarded for the project entitled “Assessing Recommendation Approaches for Dataset Interlinking”.

References

1. Ngomo, A.-C.N., Auer, S.: LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. Proc. 22nd International Joint Conference on Artificial Intelligence (2011).
2. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: SILK - A Link Discovery Framework for the Web of Data. Proc. Workshop on Linked Data on the Web colocated with the 18th International World Wide Web Conference (2009).
3. Jentzsch, A., Cyganiak, R., Bizer, C.: State of the LOD Cloud, <http://lod-cloud.net/state/>.
4. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the Linked Data Best Practices in Different Topical Domains. Proc. 13th International Semantic Web Conference, Trentino, IT (2014).
5. Ertöz, L., Steinbach, M., Kumar, V.: Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. Proc. SIAM International Conference on Data Mining, San Francisco, CA (2003).
6. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. PNAS. 99, 7821–7826 (2002).
7. Lee, C., Reid, F., McDaid, A., Hurley, N.: Detecting highly overlapping

- community structure by greedy clique expansion. Proc. 4th International Workshop on Social Network Mining and Analysis colocated with the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining February 9 (2010).
8. Gregory, S.: Finding overlapping communities in networks by label propagation. *New Journal of Physics*. 12, 103018 (2010).
 9. Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press (2008).
 10. Nunes, B.P., Mera, A., Casanova, M.A., Fetahu, B., Paes Leme, L.A.P., Dietze, S.: Complex Matching of RDF Datatype Properties. Proc. 25th International Conference on Database and Expert Systems Applications, Berlin, Heidelberg (2013).
 11. Kawase, R., Siehndel, P., Nunes, B.P., Herder, E., Nejdl, W.: Exploiting the wisdom of the crowds for characterizing and connecting heterogeneous resources. Proc. 25th ACM Conference on Hypertext and Social Media, New York, New York, USA (2014).
 12. Fortunato, S.: Community detection in graphs. *Physics Reports*. 486, (2010).
 13. Xie, J., Kelley, S., Szymanski, B.K.: Overlapping community detection in networks: The state-of-the-art and comparative study. *CSUR*. 45, (2013).
 14. Rodriguez, M.A.: A Graph Analysis of the Linked Data Cloud. ArXiv e-prints. (2009).
 15. Fetahu, B., Dietze, S., Nunes, B.P., Casanova, M.A., Taibi, D., Nejdl, W.: A Scalable Approach for Efficiently Generating Structured Dataset Topic Profiles. Proc. 11th European Semantic Web Conference, (2014).
 16. Lalithsena, S., Hitzler, P., Sheth, A.P., Jain, P.: Automatic Domain Identification for Linked Open Data. Presented at the International Conference on Web Intelligence and International Conference on Intelligent Agent Technology (2013).
 17. Emaldi, M., Corcho, O., López-de-Ipiña, D.: Detection of Related Semantic Datasets Based on Frequent Subgraph Mining. Proc. Workshop on Intelligent Exploration of Semantic Data colocated with the 14th International Semantic Web Conference (2015).
 18. Lopes, G.R., Paes Leme, L.A.P., Nunes, B.P., Casanova, M.A., Dietze, S.: Two Approaches to the Dataset Interlinking Recommendation Problem. Proc. 15th International Conference on Web Information System Engineering, Cham (2014).
 19. Caraballo, A.A.M., Nunes, B.P., Lopes, G.R., Paes Leme, L.A.P., Casanova, M.A., Dietze, S.: TRT - A Tripleset Recommendation Tool. Proc. 12th International Semantic Web Conference (2013).
 20. Paes Leme, L.A.P., Lopes, G.R., Nunes, B.P., Casanova, M.A., Dietze, S.: Identifying Candidate Datasets for Data Interlinking. Proc. 13th International Conference on Web Engineering, Berlin, Heidelberg (2013).
 21. Lopes, G.R., Paes Leme, L.A.P., Nunes, B.P., Casanova, M.A., Dietze, S.: Recommending Tripleset Interlinking through a Social Network Approach. Proc. 14th International Conference on Web Information System Engineering, Berlin, Heidelberg (2013).