Semantic Web 0 (0) 1
IOS Press

An Authority-flow based Ranking Approach to Discover Potential Novel Associations Between Linked Data

Editor(s): Claudia d'Amato
Solicited review(s): Thomas Scharrenbach, University of Zurich, Switzerland; Ross King, University of Manchester, UK; Anonymous Reviewer

María-Esther Vidal [a] Jean-Carlo Rivera [a] Luis-Daniel Ibáñez [b] Louiqa Raschid [c], Guillermo Palma [a], Héctor Rodriguez [a], Edna Ruckhaus [a]
[a] *Universidad Simón Bolívar, Caracas, Venezuela*
Email: mvidal@ldc.usb.ve
[b] *Nantes University France*
[c] *University of Maryland, College Park, USA*

Abstract. Under the umbrella of the Semantic Web, Linking Open Data projects have made available a large number of semantically intra- and inter-connected links. As an example, in the biomedical domain, data about disorders, disease related genes and proteins, clinical trials, and drugs or interventions are accessible on the Linked Open Data cloud. In addition, domain ontologies have been used to annotate scientific data. For instance, publications in PubMed have been annotated using controlled vocabulary (CV) terms from ontologies such as the Medical Subject Header (MeSH) or the Unified Medical Language System (UMLS). These annotations have been successfully mined to discover associations between drugs and diseases using techniques that have been labeled as Literature-Based Discovery (LBD). Given the large scale of the linked datasets in the Linked Open Data cloud, there is a need to develop scalable techniques that can provide answers in close to real time, to explain a phenomena, to identify anomalies, or to explore a discovery. This paper describes an authority flow based ranking technique that is inspired by LBD methods. The ranking is tailored to a layered graph. The input terms are in the first layer and the ranking will efficiently identify and assign high scores to terms in a third (or subsequent) layer, corresponding to potential novel discoveries. The terms, links and scores are modeled as a Bayesian network. Two sampling techniques are proposed to only traverse the terms that may have high scores. The first technique implements a Direct Sampling reasoning algorithm to approximate the ranking scores of nodes in the Bayesian network; it visits only the nodes with the highest probability. The second technique samples paths in the Bayesian network with the highest conditional probability. An experimental study reveals that the proposed ranking techniques are able to reproduce state-of-the-art discoveries. In addition, the sampling-based approaches are able to reduce execution times and reach high levels of accuracy.

Keywords: Link Prediction and Discovery, Direct Sampling, Bayesian Networks, Literature-Based Discovery, Path Sampling, Authority-flow Ranking Metrics, Probabilistic Logic Sampling

1. Introduction

Emerging technologies such as the Semantic Web and Semantic Grid, and the Linking Open Data initiative have made available a great number of intra- and inter-connected resources. In the context of the Linked Open Data cloud [21], there have been dramatic increases in multiplicity, diversity and the size of the resources. In October 2007, the Cloud consisted of 12 datasets with two billion RDF triples and two million RDF links. By May 2009, the Cloud had 93 datasets, 4.2 billion RDF triples and 142 million RDF links.

Currently, there are approximately 13 billon triples over 200 datasets.

The life science resources are also a good exemplar of both dramatic growth and constant evolution. For example, gene expression data has grown exponentially, and many bibliographic resources have grown at a rate of 300% per year. PubMed [1] and BIOISIS [2], the two largest interconnected bibliographic databases in biomedicine, illustrate the scale of scientific literature. PubMed publishes over 18.5 million references to journal articles, while BIOSIS makes available more than 20 million abstracts. These resources can be queried using query languages such as SPARQL [3]. SPARQL endpoints can be used to recover some of this data [21]; for example, this SPARQL endpoint [4] has been set up to access PubMed.

Many ontologies and controlled vocabularies have become available under the umbrella of the Semantic Web. Ontologies may be specified in different standard languages including XML [5], OWL [6] and RDF [7]. Ontologies provide the basis for the definition of concepts and relationships that support global interoperability among Web resources. The health and life science domain have been particularly successfully in developing and exploiting ontologies. This includes MeSH (Medical Subject Headings) [31], Disease [8], Galen [35], EHR_RM [9], RxNorm [44] and GO [13]. Ontologies can annotate concepts, describe their meaning, capture scientific knowledge. For example, MeSH terms are used by curators to annotate PubMed publications; during the annotation process, the ten or twelve most relevant MeSH terms that describe a publication are selected. Similarly, clinical trials published at Clinical Trials [8] may be annotated with MeSH, SNOMED and RxNorm.

Knowledge encoded in annotations, together with the knowledge represented within the ontologies, may provide the basis for new discoveries. For example, annotations shared by a group of genes have contributed to identify possible relationships between these genes [45,52]. Further, patterns between the MeSH terms annotating a set of publications have been used to discover potential novel associations between drugs and diseases [49]. The techniques to mine annotations to discover associations have been labeled as Literature-Based Discovery (LBD).

Given the large scale of the linked datasets in the Linked Open Data cloud, there is a need to develop scalable techniques that can provide answers in close to real time, to explain a phenomena, to identify anomalies, or to explore a discovery. This paper describes an authority flow based ranking technique [41,54] that is inspired by LBD methods. The ranking is tailored to a layered graph. The input terms are in the first layer and the ranking will assign scores to terms in a third (or subsequent) layer. The terms, links and scores are modeled as a Bayesian network. We devise a ranking technique that is able to assign high scores to potential novel associations.

Furthermore, given the size of the search space, and to reduce the effects of the number of available sources and annotations on the performance, we propose two approximate solutions named graph-sampling and path-sampling. Our proposed sampling techniques rely on the Probabilistic Logic Sampling approach defined by Henrion [22]; they are able to efficiently infer the probability of a potential novel association between a drug and a disease. These techniques sample scenarios in a Bayesian network that models the topology of data connections, where nodes represent data entries. They also estimate ranking scores that measure how important and relevant are the associations between two terms; these scores correspond to the conditional probability of the node in the network that represents one of the two terms. In addition, the approximate techniques exploit information about the topology of the links and their ontology annotations, to guide the ranking process into the space of relevant and important terms. The main difference between the two sampling techniques relies on the search technique used to sample the events in the Bayesian network. In the graph-sampling technique, a breadth-first search strategy is followed to sample the nodes with the highest conditional probabilities; on the other hand, path-sampling follows a depth-first strategy to sample paths with the highest conditional probability.

We show the effectiveness of the ranking techniques as well as their efficiency in two domains. First, we empirically show how the proposed techniques can be used to identify meaningful associations between drugs and diseases. We use publications from PubMed, their MeSH annotations and the semantic types of the

[1] http://www.ncbi.nlm.nih.gov/pubmed/
[2] http://bioisis.net/
[3] http://www.w3.org/TR/rdf-sparql-query/
[4] http://labs.mondeca.com/sparqlEndpointsStatus/details/bio2rdf-pubmed.html
[5] http://www.w3.org/TR/REC-xml/
[6] http://www.w3.org/TR/owl-ref/
[7] http://www.w3.org/TR/rdf-primer/
[8] http://clinicaltrials.gov/

MeSH terms represented in the Unified Medical Language System (UMLS); we also use different ground truth sets to verify the quality of the discovered relationships. We study four different MeSH terms that correspond to drugs or substances and verify that, in some cases, the accuracy of our techniques with regards to the ground truth sets is high, i.e., they may identify more than 70% of the objects discovered by the exact solution or reported as relevant by state-of-the-art techniques or sources of data.

Second, we study the efficiency and effectiveness of the sampling techniques with respect to the exact solution in a bibliographic data domain; we observe that the precision of the sampling techniques may be up to 98%. The empirical results suggest that these ranking techniques provide an efficient and effective solution to the problem of mining annotated datasets in the Linked Open Data cloud.

We summarize our contributions as follows:

- An authority-flow metric able to distinguish terms that may correspond to novel discoveries.
- Two sampling techniques that efficiently traverse a Bayesian network, which represents authority-flows between the Web of terms and identify highly scored paths between terms.
- An extensive empirical study that reveals the benefits of using authority-flow metrics and sampling techniques, to discover or validate links.

This paper is composed of six additional sections. Section 2 illustrates techniques proposed in the area of Literature Based Discovery (LBD) by showing the discovery reported by Srinivasan et al. in [49], where *curcumin longa* was associated with retinal diseases. In Section 3, we compare existing approaches. Section 4 defines the authority-flow based ranking metric, and section 5 describes the sampling techniques that approximate the ranking metric scores. Section 6 reports our experimental results. Finally, we give our conclusions and future work in Section 7.

2. Motivating Example

Consider the area of Literature-Based Discovery (LBD), where by traversing scientific literature annotated with controlled vocabularies like MesH, drugs have been associated with diseases [49,51]. LBD can perform *Open* or *Closed* discoveries, where a scientific problem is represented by a set of articles that discuss an input problem (*Topic A*), and the goal is to

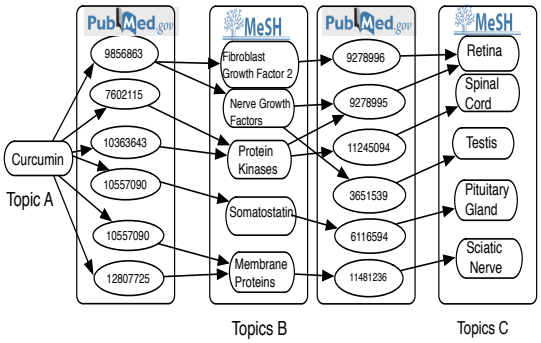


Fig. 1. Open Discovery Graph LBD

prove the significance of the associations between *A* and some other *C* topics discussed in the set of publications reachable from the initial set of publications relevant to *A*. Srinivasan et al. [49] followed this idea and improved the *Open* and *Closed* techniques by recognizing that articles in PubMed have been curated and heavily annotated with controlled vocabulary terms from the MeSH (Medical Subject Headings) ontology. Relationships between publications and terms are annotated with weights or scores that represent the relevance of the term in the document. MeSH term weights are a slight modification of the commonly used *TF-IDF* scores. Figure 1 illustrates a directed graph that models terms and publications visited during the evaluation of an *Open* discovery. Topic A is used to search on the PubMed site and retrieve relevant publications, named Pub_A. Then, MeSH term annotations are extracted from publications in Pub_A, and filtered by using a given set of semantic types of the Unified Medical Language System (UMLS)[9]; only MeSH terms of the UMLS types: (a) Gene or Genome, (b) Enzyme, or (c) Amino Acid are selected. This new set of MeSH terms is named B and is used to repeat the search on the PubMed site. Sets Pub_B, C and Pub_C are built similarly, but, C terms are only of the UMLS types: (a) Disease or Syndrome, or (b) Neoplastic Process.

Srinivasan's algorithm implemented in the Manjal system [49], considerably reduces the space of intermediate results by taking into account just the top-M terms that annotate the curcumin's publications; i.e., only the top-M terms in the set B are considered in the search and the *TF-IDF* scores are used to compute these top-M terms. Nevertheless, it still requires human intervention to create intermediate datasets, and

[9] http://www.nlm.nih.gov/research/umls/

may rank terms that do not conduce to potential novel discoveries. We propose an authority-flow based ranking metric that ranks all the intermediate nodes not only based on a local score, but also by considering the importance or authority of the terms that point to the ranked term. Similar to the Srinivasan's algorithm, the sub-graph that reaches the top-k potential novel nodes is comprised of highly ranked intermediate nodes, but these nodes will be also pointed by highly ranked nodes, and the number of these nodes is relatively small as we will see in our experimental study. Thus, this property provides the basis for our two sampling techniques that approximate the nodes that conduce to novel discoveries by just visiting a reduced number of nodes. We illustrate the usage of these techniques in the context of Literature-based Discovery and present an extensive experimental study that shows the effectiveness and efficiency of these techniques.

3. Related Work

3.1. Linking Open Data

During the last years, communities from different areas have published data in the Cloud of Linked Data following the publication guidelines, and several Linking Open Data projects have been developed. However, the number of triples is still much larger than the number of links between them; by the time that this paper was written, there were billions of triples while there were only millions of links between them.

To specify links between data sources, the ontology voID [56] has been proposed; it enables the discovery and use of Linked Data, and provides the basis for our proposed approach. Additionally, with the goal of supporting the specification of the conditions to be satisfied by each pair elements to be linked, different declarative languages have been defined [19,57]. LinQL [19] extends SQL with the functionality to create links, and to define when two elements are synonyms, hyponyms or similar; it also provides the possibility to access SQL accessible knowledge bases, and to decide the semantic relationship between two terms. Silk-LSL [57] is defined on top of RDF and also supports the functionality of expressing the condition to be satisfied by synonyms, hyponyms or similar terms. Additionally, Silk-LSL offers a larger set of built-in functions as aggregations and weighting. Although these frameworks have been successfully used to link existing Cloud datasets, they do not provide built-in methods to semantically link two terms; however, methods as the one we propose in this paper, could be incorporated as new built-in functions, and thus, enhance their capabilities and effectiveness.

Furthermore, several Linking Open Data projects have been conducted, and different applications and algorithms to discover links between datasets have been developed. Also, a variety of systems to discover links and to use Linked Data to solve real-world problems in diverse domains, have been proposed. The Linking Open Drug Data (LODD) task has connected a list of datasets that includes disorders and disease genes [14], clinical trials [20] and drug banks [60]. The TWC LOGD Portal [7] provides access to Open Government Data [1]; Momtchev et at. [32] implemented PIKB that links pathway, genes, and publications to the Uniprot dataset and the LODD data; Raimond et al. [40] developed a tool to interlink music related data into the Web of Data; Kobilarov et al. [27] interlinked BBC data with DBpedia and MusicBrainz; and Hannemann et al. [18] describe a Linked Data service for data managed at the German National Library system; finally, Cheung et al. [5,47] report a Linked Data based system to support neuroscience research. Although the datasets created or accessed by these systems represent a valuable contribution to the Cloud, the majority of the discovered links represent direct connections, created by applying similarity metrics or named entity resolution techniques. Additionally, none of the existing link discovery techniques make use of information about the link structure to identify connections. In this paper, we propose an alternative approach that relies on authority-flow based ranking techniques and makes use of the topology of the links and their annotations, to assign the highest ranking scores to the paths that correspond to potential novel associations.

3.2. Link Prediction and Ranking Approaches

Several descriptive and predictive inference tasks based on link structure as well as on the semantics encoded in the ontological annotations of the entries, have been proposed to discover potential novel associations between data entries [12,17,23,24,28,36,45]. In general, the idea is to perform random walks in the space of possible associations and discover those that satisfy a particular pattern; correspondences between the discovered patterns are measured in terms of similarity functions. In [16], heuristics are used to discover relevant subgraphs within RDF graphs; relationships among the metadata that describe nodes

are used to discover relevant relationships among entities. To decide if two objects are semantically similar, Jeh et. al. [24] propose a measure that reflects when two objects are similar based on the relationships that they hold with similar objects. Yan et al. [17] developed strategies to efficiently search subgraphs that are similar to a given query graph. Barna et al. [45] describe a greedy algorithm that generates the dense subgraph of the graph that represents known connections between genes to discover potential novel associations. Parundekar et al. [36] propose a learning technique to generate semantic descriptions of available datasets, and link the instances of these datasets to existing data in the Cloud. Hu et al. [23], and Kuramochi and Karypis [28], describe efficient algorithms to discover subgraphs (patterns) that occur in graphs and to aggregate them. Finally, Toupikov et al. [53] propose the usage of source formal descriptions provided in voiD documents, to efficiently implement link analysis ranking metrics like PageRank [2] and HITS [26], and use the ranking scores for measuring popularity of the linked datasets. A weight function is defined to represent user preferences for going from one dataset to another, and this function is included into PageRank and HITS to obtain more precise rankings. In any case, the importance of a node is dissipated whenever the out-degree of the node is high. In this paper we also propose an authority-flow based ranking technique, but this technique ranks paths between data terms, and the authority of one node is computed in terms of score functions that measure how important is a link between two nodes. Additionally, because we want to identify potential novel associations between terms, the importance of a node, not necessarily should be dissipated among all the nodes to which it is linked. For example, the MeSH term *Luteinizing Hormone* indexes 45,068 publications in the curcumin layered graph presented in Section 2. Based on the PageRank metric the importance that this term should transfer to its children, should be reduced proportionally to this number of links; however, in the sub-graph that comprises the top-5 novel MeSH terms identified by the Srinivasan's algorithm, *Luteinizing Hormone* is part of 21,589 paths. Considering that 481 MeSH terms annotate the curcumin's publications and the number of paths that reach these top-5 terms is 2,662,887, this term is one out of 481 and conduces to almost 1% of the paths that reach the top-5 terms; so, it should be highly ranked as well as their descendant nodes. To model this situation, our ranking metric does not penalize the importance of a node when the degree of the node is high, and it is able to highly rank terms as *Luteinizing Hormone*. Finally, the proposed techniques exploit the semantic encoded in the annotations and the topology of the graph induced by the data links, to discover the associations with the highest scores. Annotations as well as the topology of the layered graphs are considered during both the computation of the score functions and the computation of the metric values for each node of the graph. Thus, whenever a node is annotated with a large number of terms, or these terms are close in the ontology used to annotate the publications, the score function will take high values; similar if a node is pointed by a large number of highly ranked nodes, its score will be high whenever these nodes are semantically relevant for it. For example, in case of links between publications and their MeSH terms annotations, a publication will have high values of the metric, if the MeSH terms that index this publication also index a large number of other publications in the same layer of the graph, and they have high scores. On the other hand, a MeSH term will have high metric values if it has been used to annotate a large number of publications, and these publications have been annotated with similar MeSH terms. Note that although PageRank and HITS are able to capture importance and relevance of the terms, they do not reflect the type of information that the score function can reflect.

We illustrate the usage of our techniques in two domains: life science and bibliographic data; however, our hypothesis is that these techniques could also be used in other domains, e.g., in social networks such as Facebook or Twitter, to discover relevant associations or trends. Recently, Moore et al. [33] propose a random walks based approach able to highly rank nodes in a graph that may be useful for a given node, or explain how two nodes are related in the graph. Shortest paths between the input nodes are computed to solve these two problems. The top-k solutions are computed among the shortest paths and based on weights of the nodes which are computed in terms of the degrees of a node; thus, topology of the graph is considered. This metric could be used to explain the results suggested by our approach; nevertheless, because our approach relies on layered acyclic directed graphs where all the paths have the same length and the direction of edges represent different types of connections between concepts, applying this shortest path undirected based approach may not significantly contribute to solve the problem of highly ranking the most potential relevant paths or identifying them efficiently. On the contrary, we are interested in distinguishing portions of the orig-

inal graph that comprise the relevant or novel terms. Empirically, we have observed that our studied layered graphs are comprised of only a small number of target terms that may correspond to potential discoveries and the sub-graphs that reach these target objects are very dense. For example, the layered graph for the term curcumin presented in Section 2 is comprised of 24,455,484 paths that reach 570 target terms in the last layer of the graph; however, the top-5 terms discovered by the Srinivasan's algorithm are part of a subgraph with 2,662,887 paths, i.e., less than 1% of the target terms are reached by over 10% of the paths of the graph. Thus, instead of a metric able to rank among the target nodes, the ones that are reachable from nodes with high degrees, we need a metric able to discriminate the top-k nodes that are part of dense sub-graphs.

3.3. Approximate Inference

A typical inference task in a Bayesian network is to compute the posterior probability of a set of nodes given some observed values of evidence. Several algorithms have been developed to efficiently perform exact inference in a Bayesian network [3,6,15]; however, exact inference in large and complex networks may be intractable, and approximate solutions have been defined [22,30,34,62]. Commonly approximate Bayesian network inference algorithms rely on Monte Carlo algorithms to generate a set of randomly selected nodes according to some known distribution, and then approximate probabilities based on the frequencies of appearances in the sample. The main challenge of these algorithms is to reach estimates that satisfy the required confidence levels and are consistent with the network evidence values, while the size of the sample remains small; several approaches have successfully achieved this goal and depending on the properties of the Bayesian network, their performance and quality can be quite good [4,11,22,30,34,50,62]. Based on these techniques, we devise two sampling techniques that follow two different search strategies, and provide an efficient and effective solution to the Literature-Based Discovery problem.

Finally, sampling techniques have been also applied to the problem of estimating authority-flow metrics. Fogaras et. al. [11] implement a Monte-Carlo based method to approximate personalized PageRank scores. They sample paths whose length is determined by a geometric distribution. Paths are sampled from a Web graph based on a probability that represents whether objects in the paths can be visited by a random surfer. This approach approximates PageRank; however, it is not applicable to our proposed approach because the length of the paths is determined by the number of layers in the results graph, which in our case is fixed and cannot be randomly chosen. In contrast, our techniques sample objects in a layer graph of n layers, and the search is performed layer by layer, until the last layer of the graph is visited. Objects with higher probability to be visited by a random surfer, and links between these objects will have greater chance to be chosen during the sampling process. Thus, the techniques may be able to only traverse relevant paths of length n which may correspond to relevant discoveries.

4. A Ranking-flow based Solution to Discover Semantic Associations

We propose ranking-flow based solutions to the problem of semantic association discovery. The proposed techniques take advantage of existing links between data published on the Cloud of Linked Data, or make use of annotations with controlled vocabularies such as MeSH, GO, PO, etc. We present an exact solution and two approximate sampling-based techniques, which have been implemented in BioNav [55].

The exact ranking technique extends existing authority-flow based metrics like PageRank, ObjectRank and their extensions of layered graphs [41]. This ranking approach assumes that portions of Linked Data comprise a layered graph, named layered Discovery Graph, where nodes represent published data and edges correspond to intra- or inter-dataset links.

A layered Discovery Graph, *lgDG*=(V_{lg}, E_{lg}), is a layered directed acyclic graph, comprised of a finite number k of layers, $L_1, \ldots, L_k$. Layers are composed of data entries that point to data entries in the next layer of the graph; data entries are filtered, and a link between the same two objects is represented at most in one layer of the graph. Data entries in the k-th layer (last layer) are called target objects. Authority-flow based metrics rank the target objects, and these scores are used to identify relevant associations between objects in the first layer and target objects.

Figure 2 illustrates an example of a layered Discovery Graph that models the Open Discovery Graph in Figure 1. In this example, odd layers are composed of MeSH terms while even layers are sets of publications. Also, an edge from a term *b* to a publication *p* indicates that *p* is retrieved by the PubMed search engine when *b* is the search term. Finally, an edge from a publication

p to a term *b* represents that *p* is annotated with *b*. Each edge $e = (b, p)$(resp., $e = (p, b)$) between the layers l_i and l_{i+1} is annotated with the *TF-IDF* score; this value either represents how relevant is a term b in the collection of documents in l_{i+1}, or a document relevance regarding a set of terms. The path of thick edges connects Topic *A* with $C3$; the value 0.729 corresponds to the authority-flow score and represents the relevance of the association between Topic *A* and $C3$. B and C terms are filtered based on different criteria, so, never a link between the same publication and MeSH term will appear more than once in the graph.

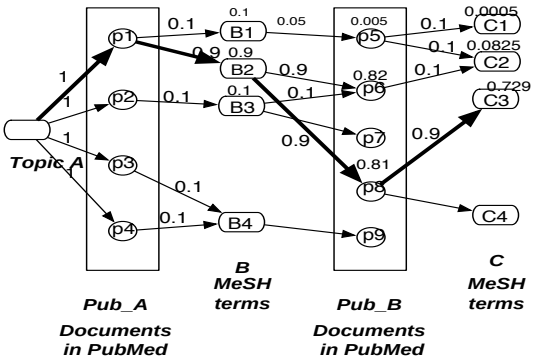


Fig. 2. A layered Discovery Graph

Given a layered Discovery Graph *lgDG*=(V_{lg}, E_{lg}) of k layers, the authority-flow scores of the target objects are formally defined as a ranking vector R:

$$R = M^{k-1} \times R_{ini} = \left(\prod_{l=1}^{k-1} M \right) \times R_{ini} \qquad (1)$$

where, M is a transition matrix and R_{ini} is a vector with the scores of the objects in the first layer of the graph. An entry $M[u, v]$ in the transition matrix M, where u and v are two data objects in *lgDG*, corresponds to $\alpha(u, v)$ or is 0.0. The value $\alpha(u, v)$ is calculated according to a score function that considers semantics encoded in the link, e.g., if the link represents that a publication is annotated with a MeSH term, then this score may reflect relatedness between this annotation and the rest of the annotations of the publication.

$$M[u, v] = \begin{cases} \alpha(u, v), & \text{if } (u, v) \in E_{lg}, \\ 0.0, & \text{otherwise.} \end{cases} \qquad (2)$$

Furthermore, score functions represent information about the topology of the graph in a way that nodes with high values of the metric lgWP are linked by many nodes or linked by highly scored nodes. For example, in Figure 2, $C3$ is pointed by relevant nodes. In the context of LBD, we use this metric to discover novel associations between a topic *A* and MeSH terms in the last layer of the *lgDG*.

4.1. Score Functions

We consider two types of score functions in the *layered graph Weighted Path Count* (lgWP) metric: the *TF-IDF* score function of a link between a and b denotes how important or relevant is a for the type of concepts to which b belongs [46]; the *taxonomic* score function reflects relevance of a to b in terms of how similar are to b the rest of the concepts that are associated with a. Formally, these functions are defined as follows:

Consider an edge e=(a,b) between nodes a and b in layers l_{i-1} and l_i, respectively. The *TF-IDF* score function for e, *tf-idf(a,b,*l_{i-1},l_i*)* is equal to $(w(a, l_{i-1}, l_i) \times C(l_{i-1}, l_i))$, where $w(a)$ is equal to $A(a, l_{i-1}) \times B(a, l_i)$ and:

- *A(a,*l_{i-1}*)*: is the augmented document frequency of a which is defined as

$$A(a, l_{i-1}) = 0.5 + 0.5 \times \left(\frac{tf(a, l_{i-1})}{tf_{max}(l_{i-1})} \right) \qquad (3)$$

  where, $tf(a, l_{i-1})$ is the frequency of a in layer l_{i-1}, and $tf_{max}(l_{i-1})$ is the maximum frequency of any node in l_{i-1}. A value close to 1.0 indicates that node a frequently appears in layer l_{i-1}.
- *B(a,*l_i*)*: inverse term frequency $log_2(\frac{N(l_i)}{Np(l_i)})$, where $N(l_i)$ is a finite number that corresponds to the cardinality of the domain of nodes in l_i, and $N_p(l_i)$ corresponds to the number of nodes in l_i that are associated with node a.
- *C(*l_{i-1}, l_i*)*: is a cosine normalization for all the nodes in layer l_{i-1}, i.e.,

$$C(l_{i-1}, l_i) = \frac{1}{\left(\sum_{a' \in l_{i-1}} w(a', l_{i-1}, l_i)^2 \right)^{1/2}} \qquad (4)$$

Labels of edges of the Layered Discovery Graph in Figure 2 illustrate the values of the *TF-IDF* score function; a value close to 1.0, for example *tf-idf(p1,B2)*, indicates that publication $p1$ has more annotations that the rest of the publications in *PUB_A*; assuming that annotations are done by experts, this is an indication that a large number of MeSH terms properly describe the publication. Furthermore, the *taxonomic* score function assumes that nodes in layer l_i are part of a taxonomy, and captures the taxonomic distance between the nodes in layer l_i that are related to node a. Thus, the *taxonomic* score function for e, *tf-idf(a,b)* is equal to $(w(a, l_{i-1}, l_i) \times C(l_{i-1}, l_i) \times d_{tx}(a, b))$, where, $d_{tx}(a, b)$ corresponds to the average of the taxonomic distance values between b and nodes in layer l_i that are associated with a, i.e.,

$$d_{tx}(a, b) = \frac{1}{t} \sum_{b' \in l_i} d(b, b') \quad (5)$$

where, t is the number of nodes in layer l_i that are related to a, and $d(b, b')$ is the value of the taxonomic distance between b and b'. In the following example, we consider a metric based on the similarity function proposed by Pekar and Staab [38] that captures the ability to represent the taxonomic distance between two vertices with respect to the depth of the common ancestor of these two vertices; nevertheless, any other taxonomic distance metric or semantic similarity function could be used to capture the relatedness of two terms in an ontology [25,29,37,39,42,58]. In addition to the annotation information considered by the *TF-IDF* score function, the *taxonomic* score function reflects the relatedness of the terms used to annotate a publication; a value close to 1.0 indicates that the publication may have a large number of related or similar MeSH annotations. Assuming annotations are done by experts, this is an indication that a large number of related MeSH terms describe the publication. To illustrate the impact on the score values of the topology of the MeSH ontology used to annotate PubMed publications, consider the publication with PMID 15493372 which is annotated with the MeSH terms *Isoenzymes*, *Arachidonate 5-Lipoxygenase*, *Cyclooxygenase*; because this publication is annotated with only three MeSH terms of the UMLS types: (a) *Gene or Genome*, (b) *Enzyme*, and (c) *Amino Acid*, the *TF-IDF* score function value is 0.26. However, if the topology of MeSH is considered, the *taxonomic* score function is able to reflect that the link between this publication and the MeSH term *Cyclooxygenase 2* has a greater score (0.08); this is consistent with the fact that this term is the closest in the taxonomy to the other two terms, and it best represents the content of the publication. Currently, our *taxonomic* score functions are defined for existing specifications of *polyhierarchy* biological ontologies such as, MeSH or SNOMED-CT, where under the *Closed World Assumption*, a node classification is the inference task needed to compute a score function value. Considering a more general approach based on *Open World Assumption* will require first, the adaptation of existing biomedical ontologies and then, the extension of ontology similarity metrics to measure with certain degrees of uncertainty, unknown facts that cannot be inferred from the concepts represented in the ontology. This extension would enhance expressiveness and accuracy of the ontologies and our discovery process; however, it is out of the scope of this paper.

5. Approximate Techniques to Discover Semantic Associations

Although rankings induced by an authority-flow based metric may distinguish relevant associations, the computation of this ranking may be costly. Thus, to speed up this task, we propose sampling-based techniques that extend the Probabilistic Logic sampling approach [22] and traverse only nodes in the layered graph that may conduce to highly ranked objects. We briefly summarize the Probabilistic Logic sampling, next, we define the Estimate Relevant Links problem; finally, we present two approximate solutions.

5.1. Probabilistic Logic Sampling

Bayesian networks are directed acyclic graphs comprised of nodes or random variables and arcs that correspond to direct probabilistic dependencies between them. Bayesian networks encode joint probability distributions over a set of finite nodes or random variables, which are computed as products of the conditional probabilities of the variables given their parents in the network. Nodes are conditionally independent of their predecessors, given their parents, i.e., nodes are conditionally independent of their non-descendants and all other nodes in the network, given their parents, children and children's parents [43]. To overcome intractability of exact inference solutions in Bayesian networks, several approximations have been

proposed [15]. The Probabilistic Logic sampling proposed by Henrion [22] is the simplest approximate approach where a set of randomly generated samples of the network are generated during one iteration of the sampling; approximate probabilities of the query variables are computed according to the frequencies of the nodes sampled during the sampling process; the influence arrows are considered during both the sampling and the computation of the approximate probabilities. This process is repeated r times, where r is a finite number, where different scenarios are generated for each sample; the probability of x, $Pr(x)$, after sampling r scenarios is the average of the probability of x in the scenario s, $Pr(x)^s$, in which the variable x was true, i.e.,

$$Pr(x) = \frac{1}{m}\sum_{s=1}^{r} Pr(x)^s \tag{6}$$

For each conditional probability, $Pr(x|y)$ the Probabilistic Logic sampling approach proposes to generate a sample for each of the independent parameters, i.e., $Pr(x|y)$ and $Pr(x|\bar{y})$, ensuring that samples of $Pr(x)$ can also be done after their parents, $Pr(y)$ and $Pr(\bar{y})$, are sampled. Thus, conditional probabilities represent the conditional dependencies that the sampling process needs to respect. Although Probabilistic Logic sampling has shown to work very well when no prior evidence has been observed [15], the performance and quality of the approach can be impacted by the strategy followed to perform the search. In this work, we propose two sampling techniques that rely on the Probabilistic Logic sampling approach, but implement two different search techniques to sample the nodes in the Bayesian network. In the graph-sampling technique, a breadth-first search strategy is followed to sample the nodes with the highest conditional probabilities; path-sampling respects a depth-first strategy to sample paths with the highest conditional probability.

5.2. Approximating Potential Relevant Associations

Problem Estimate Relevant Associations: Given a layered Discovery Graph, $lgDG = (V_{lg}, E_{lg})$, the computation of highly ranked target objects is reduced to estimating a subgraph $\overline{lgDG}$ of $lgDG$, so that with high confidence (at least δ), the relative error ε of the distance between the top-k target objects in $\overline{lgDG}$, i.e., the expected top-k *E(top$_k$)*, and the exact top-k target objects in $lgDG$, i.e., *top$_k$* is at least δ, i.e.,

$$Pr(|top_k - E(top_k)| \leq \varepsilon) \geq \delta \tag{7}$$

A set SS=$\{lgDG_1, ..., lgDG_m\}$ of independent and identically distributed (i.i.d.) subgraphs of $lgDG$ is generated. Then, $lgDG'$ is computed as the union of the m subgraphs. Each subgraph $lgDG_i$ is generated using a sampling technique on a Bayesian network that models all the navigational information encoded in $lgDG$ and in the transition matrix *M* of the authority-flow metric. We propose two sampling techniques: *graph-sampling* and *path-sampling*. Graph-sampling is based on a Direct Sampling technique over the Bayesian network, that generates the most relevant sub-graph $lgDG_i$ by visiting the most relevant nodes in the Bayesian network. The second sampling approach follows a Monte-Carlo technique on the Bayesian network to just produce the paths with the highest conditional probability to be traversed.

Given a layered graph, a Bayesian network is formally defined as follows:

A Bayesian network *BN=(VB,EB)* for a layered Discovery Graph $lgDG$, is built as follows:

- *BN* and $lgDG$ are homomorphically equivalent, i.e., there is a mapping *f: VB* $\rightarrow V_{lg}$, such that, $(f(u), f(v)) \in E_{lg}$ iff $(u, v) \in EB$.
- Nodes in VB correspond to discrete random variables that represent if a node is visited or not during the discovery process, i.e., *VB* = $\{X \mid X$ takes the value 1 (true) if the node X is visited and 0 (false), otherwise$\}$.
- Each node X in *VB* has a conditional probability distribution[10]:

$$Pr(X \mid Par(X)) = \sum_{j=1}^{n} \alpha\Big(f(Y_j), f(X)\Big) \times Pr(Y_j) \times Y_j \tag{8}$$

where, Y_j is the value of the random variable that represents the j-th parent of the node X in the previous layer of the Bayesian network, Y_j can be 0 or 1; n corresponds to the number of parents of X. The value $\alpha(f(Y_j), f(X))$ represents values of the score function of the edge $(f(Y_j), f(X))$; the score function can be *TF-IDF* or *taxonomic*, and it corresponds to an entry in the transition matrix M. It is seen as the probability

[10] *Par(X)* represents the parents of X in the Bayesian network.

to move from Y_j to X in the Bayesian network. Furthermore, the conditional probability distribution of a node X represents the collective probability that X is visited by a random surfer starting from the objects in the first layer of the layered Discovery Graph. Finally, the probability of the nodes in the first layer of the Bayesian network corresponds to a score that indicates the relevance of these objects with respect to the discovery process; these values are represented in the R_{ini} vector of the ranking metric.

5.3. A Graph-Sampling based Ranking Solution

The first sampling technique traverses the Bayesian network by performing a breadth-first search that visits the nodes with the highest conditional probability. The breadth-first search is based on a Direct Sampling method for a Bayesian network that generates events from the Bayesian network [43].

Given a Bayesian network generated from the layered Discovery Graph $lgDG$, the Direct Sampling generates each subgraph $lgDG_i$. Direct Sampling selects nodes in $lgDG_i$ by sampling the variables from the Bayesian network based on the conditional probability of each random variable or node. Algorithm 1 describes the Direct Sampling algorithm.

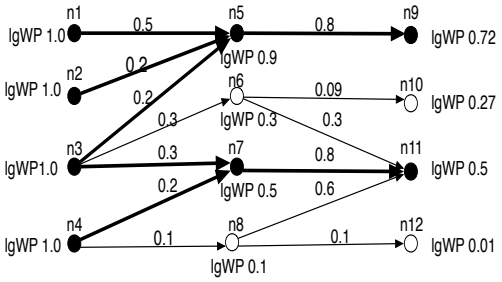
Algorithm 1 The Direct Sampling Algorithm

Input: *BN=(VB,EB)* A Bayesian network for a layered discovery graph.
Output: A subgraph $lgDG_i$

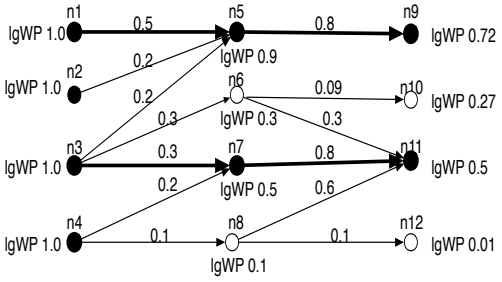
```
TP ← topologicalOrder(BN);
for X ∈ TP do
   Pr(X| Par(X)) ←
                 Σ_{j=1}^{n} α(f(Y_j), f(X)) × Pr(Y_j) × Y_j;
   if ( Pr(X| Par(X)) >= randomNumber) then
      X_i ← 1;
   else
      X_i ← 0;
   end if
end for
```

Variables are sampled, following a topological order starting from the variables in the first layer of the Bayesian network; this process is repeated until variables in the last layer are reached. The values assigned to the parents of a variable define the probability distribution from which the variable is sampled. The conditional probability of each node in the last layer of $lgDG_i$ corresponds to the approximate value of the implemented metric.

Figure 3(a) illustrates the behavior of the graph-sampling technique; highlighted nodes correspond to visited notes and comprise a subgraph $lgDG_i$. Direct Sampling is performed as follows: initially, all the nodes in the first layer have the same probability to be visited and all of them are considered. All their children or nodes in the second layer are also visited and the conditional probability is computed; nodes with the highest scores survive, i.e., $n5$ and $n7$. Then, the children of these selected nodes are also visited, and the process is repeated until nodes in the last layer are reached. Note that nodes $n9$ and $n11$ are the target objects with the highest values of the lgWP metric and with the highest conditional probability. These nodes are pointed by nodes with high lgWP scores or pointed by many nodes; thus, they are very likely to be visited when the Direct Sampling algorithm is performed.



(a) Graph-Sampling



(b) Path-Sampling

Fig. 3. Approximate Bayesian Network Inference Techniques

Once an iteration i of the Direct Sampling is finalized, the sampled layered Discovery Graph $lgDG_i = (V_i, E_i)$ is created. Nodes in V_i correspond to the variables sampled during the Direct Sampling process that are connected to a visited variable in the last layer of the Bayesian network. Additionally, for each edge (u, v) in the Bayesian network that connects nodes $f(u)$ and $f(v)$ in V_i, an edge $(f(u), f(v))$ is added to

E_i. The conditional probabilities of the target objects of each subgraph $lgDG_i$ correspond to the approximate values of the ranking metric. After all the subgraphs $lgDG_1, \ldots, lgDG_m$ are computed, an estimate $lgDG'$ is obtained as the union of these m subgraphs. The approximation of the ranking metric in the graph $lgDG'$ is computed as the average of the approximate ranking metric values of target objects in the subgraphs $lgDG_1, \ldots, lgDG_m$. A bound of the number of iterations or sampled subgraphs is defined in terms of the Chernoff-Hoeffding's bound.

Theorem 1 *Consider a sampled layered graph $lgDG_i$, a term t, and the random variable $X_t(lgDG_i)$ that equals 1 or 0 when the object t is a target object in $lgDG_i$ or not. Let S_t be another random variable that averages the variables $X_t(lgDG_i)$ for the samples $lgDG_1, \ldots, lgDG_m$, i.e.,*

$$S_t \;=\; \frac{1}{m}\sum_{i=1}^{m} X_t(lgDG_i)\,. \tag{9}$$

Let pr be the probability of a term t being a target object of a sample layered graph $lgDG_i$, i.e., $Pr(X_t(lgDG_i) = 1) = pr$. Since the sequence $X_t(lgDG_1)$, $X_t(lgDG_2)$,…,$X_t(lgDG_m)$ represents a sequence of Bernoulli trials, pr corresponds to the probability of success of the trials or the expectation of S_t denoted by $E(S_t)$. Thus, an upper bound for m is the following:

$$m \leq \frac{ln(2) - ln(Pr(|S_t - E(S_t)| \geq \epsilon))}{2\epsilon^2} \tag{10}$$

Proof By using the Chernoff-Hoeffding's bound, the size m of the sample must satisfy the following formula to ensure that the relative error of the estimation of $E(S_t)$ is greater than some given constant ϵ with some probability:

$$Pr(|S_t - E(S_t)| \geq \epsilon) \leq 2e^{(-2m\epsilon^2)} \tag{11}$$

5.4. A Path-Sampling based Ranking Solution

Similarly, the path-sampling technique traverses the Bayesian network and approximates each sub-graph $lgDG_i$ by following a Monte-Carlo based method to generate N random paths that will comprise the subgraph $lgDG_i$. The conditional probability is computed for each visited node by considering the conditional probability of their parents times the authority-flow values of the edges that comprise the path; independence of the event of visiting each edge is assumed.

The process to generate each random path is defined as follows: Let us sample a collection of i.i.d. paths $\xi_1, \ldots, \xi_N \sim P$ from $lgDG$ as follows: starting with a "particle" at $X = s$ in the first layer of the result graph $lgDG$, choose a node X with probability *Pr(X| Par(X))* $\leftarrow \sum_{j=1}^{n} \alpha(f(Y_j), f(X)) \times Pr(Y_j) \times Y_j$; and set X as visited. These statements are repeated until Y becomes a node in the last layer of the layered graph, i.e., a target object is reached. Algorithm 2 describes the path-sampling technique.

Algorithm 2 The Path-Sampling Algorithm

Input: *BN=(VB,EB)* a Bayesian network for a layered discovery graph;
N an integer representing the number of paths to sample;
L the number of layers of *BN*.
Output: A subgraph $lgDG_i$

```
Node ← a random sample from events in the first layer of BN;
Node.Visited← true;
NumberPaths ← 0;
while NumberPaths ≤ N do
  PathLength ← 0;
  while PathLength ≤ L do
    for X ∈ Node.Children do
      Pr(X| Par(X)) ←
            Σ_{j=1}^{n} α(f(Y_j), f(X)) × Pr(Y_j) × Y_j;
    end for
    Node ← a random sample event X from
    Node.Children such that,
    Pr(X| Par(X)) >= randomNumber;
    Node.Visited ← true;
    NumberPaths ← NumberPaths +1;
  end while
end while
```

Figure 3(b) illustrates the behavior of the path-sampling technique; highlighted nodes correspond to visited notes and comprise a subgraph $lgDG_i$. Path-sampling iterates until N paths are generated. To generate one path, path-sampling performs as follows: initially, all the nodes in the first layer have the same probability to be visited and one is randomly chosen, suppose it is n_3. All their children or nodes of this selected node are considered and the conditional probability is computed for all of them; the child node with the highest scores survives, i.e., $n7$. Then, the children of this selected node are also visited, and the process is

repeated until nodes in the last layer are reached. Suppose two paths are required, then nodes $n9$ and $n11$ can be reached. Note that these nodes are sink nodes of the paths with the highest values of the lgWP metric and with the highest conditional probability.

Let $Sp^{'}$={ξ_1,ξ_2,...,ξ_N} be the paths of the layered graph $lgDG_i$ such that each ξ_i, $1 \leq i \leq N$, is randomly chosen, with replacement, from the set of paths in the exact layered graph $lgDG$.

A sampled graph $lgDG^{'}$ corresponds to the minimal sub-graph of $lgDG$ that contains only the nodes in $Sp^{'}$. A bound of the number of sampled paths is defined in terms of the Chernoff-Hoeffding's bound as follows:

Theorem 2 *Let $X_t(\xi_i)$ be an independent identically distributed (i.i.d.) binary random variable that has value 1 if the sink node of the path ξ_i is the term t, and 0 otherwise. Let S_t be another random variable that averages variables $X_t(\xi_i)$ for objects in $Sp^{'}$, i.e.,*

$$S_t = \frac{1}{N}\sum_{i=1}^{N} X_t(\xi_i) \tag{12}$$

Let pr be the probability that t is the sink node of the path ξ_i, i.e., $Pr(X_t(\xi_i) = 1) = pr$; the sequence $X_t(\xi_1)$, $X_t(\xi_2)$,...,$X_t(\xi_N)$ represents a sequence of Bernoulli trials and pr corresponds to the expectation of S_t denoted by $E(S_t)$. Thus, an upper bound for the number of paths to be sampled, i.e., N, is the following and the time complexity of this method is $\Theta(N \times L)$, where L is the number of layers of the layered Discovery Graph.

$$N \leq \frac{ln(2) - ln(Pr(|S_t - E(S_t)| \geq \epsilon))}{2\epsilon^2} \tag{13}$$

Proof The number of paths to sample, N, has to satisfy the following formula to ensure that the relative error of the estimation of $E(S_t)$ is greater than some given constant ϵ with some probability:

$$Pr(|S_t - E(S_t)| \geq \epsilon) \leq 2e^{(-2N\epsilon^2)} \tag{14}$$

6. Experimental Evaluation

In this section we show the quality of our proposed discovery techniques. First, we compare the results obtained by our ranking technique with respect to the results obtained by the Manjal system [49]; then, we validate and compare associations discovered by our approach with information published at specialized websites. Finally, we show the behavior of our proposed ranking techniques on bibliographic data. Experiments were executed on a Sun Fire V440 equipped with two UltraSPARC IIIi processors running at 1.593 GHZ with 16 GB RAM. Results are shown for several iterations of the sampling techniques and only for top-k concepts; the number of iterations and the top-k have been experimentally set up; a trade-off between execution time and quality of the results have considered during the configuration of these two parameters.

Experiments were designed under the assumption of the following hypotheses:

Hypothesis 1: As explained in Section 2, Srinivasan's algorithm relies on MeSH terms that annotate a PubMed publication, to identify the potential novel associations between drugs and diseases, and reduce the space of PubMed publications that need to be traversed during the discovery process. However, many irrelevant publications and MeSH terms can be visited during the search, and a large number of target concepts can be generated. Because our proposed sampling techniques visited only concepts, publications and MeSH terms that are related to highly ranked concepts, and only a small percentage of target concepts have high values of the lgWP metric, we hypothesize: *i*) the concepts highly ranked by our sampling techniques correspond to the ones identified by Manjal, *ii*) the highly ranked target concept will be produced by traversing just a reduced number of the concepts that Manjal will traverse, i.e., by traversing a dense sub-graph that comprises the highly ranked target objects.

Hypothesis 2: Because the sampling techniques identify highly ranked target concepts, and these concepts correspond to MeSH terms used to annotate publications that are: *i*) annotated with a large number of MeSH terms, or *ii*) indexed by MeSH terms associated with publications related to the input drug, and these terms index a large number of publications. Then, these highly ranked target concepts correspond to diseases that possibly can be treated with the input drug.

Hypothesis 3: Because the sampling techniques traverse highly ranked intermediate nodes, irrelevant target nodes that are reachable from poorly ranked intermediate nodes, may be discarded dur-

ing the sampling. Thus, accuracy of the sampling may be higher than the one reached by computing the exact values of lgPW for the whole graph.

6.1. Experiment Configuration

Datasets: We used two datasets. The first dataset is comprised of PubMed publications from the NCBI source[11], all the Medical Subject Headings (MeSH) terms, and all the links between MeSH terms and PubMed publications (indices and annotations). The second dataset is composed of bibliographic data from DBLP[12]. In both cases, we stored the datasets in an Oracle 10g database. For the PubMed database, we downloaded all the PubMed publication ids (circa September 2010), and their corresponding MeSH terms, and stored them in two tables, *PubMed_MeSH* and *MeSH_PubMed*. The former relates a publications with all the MeSH terms that correspond to their annotations; the later relates MeSH terms to all publications that these terms index. Both tables store the *TF-IDF* score function values.

The DBLP data was also stored in four relational tables: *Conferences*, *Year*, *Paper* and *Author*. Relationships between a conference and a paper with a year in which the conference was issued and the paper published, are stored in two tables *C_Y* and *P_Y*, respectively. Similarly, the relationships between a paper and the authors of the paper, and the paper and the conference where the paper is published, are stored in the tables *P_A* and *P_C*, respectively. Tables *C_Y*, *P_Y*, *P_A* and *P_C* have an attribute *score* computed from an authority-flow value assigned to each of the relationships that represent these tables, divided by the number of instances in the corresponding table. Figure 4 illustrates the database schemas of the DBLP; relationships are labelled with the authority-flow values considered in this experiment used to compute the score values.

Query Benchmarks: We ran our ranking-based discovery approach on PubMed data to discover semantic associations between the terms *curcumin*, *gingko*, *aloe*, and *tacrolimus*, and MeSH terms that represent diseases; the sizes of the corresponding layered graphs are reported in Table 1. Layered graphs were built following the criteria proposed by Srinivasan et al. [49] and explained in Section 2; data from tables *PubMed_MeSH* and *MeSH_PubMed* was selected. Additionally, we ran 3 sets of 30 queries against DBLP; layered Discovery Graphs were comprised of 5 layers and at most 876,110 nodes and 4,166,626 edges. Author's names with high, medium and low selectivity were considered; high selectivity means that the author has few publications while low selectivity represents that the author is very productive.

Metrics: We report on performance and quality of our ranking techniques. Performance is measured in terms of runtime, which corresponds to the *user time* produced by the *time* command of the Unix operation system; this value represents the elapsed time between the submission of the query and the output of the target MeSH terms that may correspond to the novel discoveries; time to transfer data from the database to main memory is the dominant contribution to this time metric. Quality is expressed as precision and the values of the normalized top-k Spearman's rho distance metric with ties. Precision measures the percentage of concepts that are produced by the proposed ranking techniques that are present in the ground truth; because we compare lists of the same size, precision and recall have the same values. The normalized top-k Spearman's rho metric with ties, measures how distinct are the orders or permutations of two lists [10]. This metric is defined as follows: let ϕ_1 and ϕ_2 be 2 top-k lists; each set of tied results is called a bucket. Thus, the ranked lists can be viewed as ranked buckets $B_1, B_2, \ldots, B_n$. The position of bucket B_i, denoted $pos(B_i)$ is the average location within bucket B_i. We assign $\phi(x) = pos(B)$ where $\phi(x)$ is the rank of term x, and B is the bucket of x. ρ is the Spearman's rho metric, which is a normalized distance measure that lies in the interval [0,1]. The following formula represents the normalized Spearman's rho distance metric of top-k lists ϕ_1 and ϕ_2:

$$\rho(\phi_1, \phi_2) = \frac{(\sum_{i=1}^{k} |\phi_1(i) - \phi_2(i)|^2)^{1/2}}{(\frac{k\times(k+1)\times(2k+1)}{3})^{1/2}} \quad (15)$$

The maximum value of $(\sum_{i=1}^{k} |\phi_1(i) - \phi_2(i)|^2)^{1/2}$ occurs when list ϕ_1 is the reverse of list ϕ_2 and this value corresponds to $(\frac{k\times(k+1)\times(2k+1)}{3})^{1/2}$ and the normalized value is equal to 1.0; while a

[11] http://www.ncbi.nlm.nih.gov/

[12] http://www.informatik.uni-trier.de/ ley/db/

value of 0.0 represents that ϕ_1 and ϕ_2 have exactly the same elements and in the same order.

Implementations: The ranking and sampling techniques were implemented in Java 1.6.1, and the databases were stored in Oracle 10g. To compute the exact and the sampling methods for a given query, the entire graph and the Bayesian network are both kept in main memory.

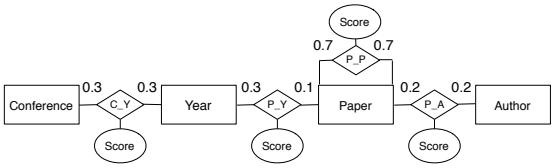


Fig. 4. The DBLP schema

MeSH term	#Nodes	#Edges	#Paths
Curcumin	3,107,901	10,261,791	24,455,484
Gingko	1,906,962	31,909,304	24,548,174
Aloe	4,705,163	16,027,100	22,070,602
Tacrolimus	5,569,784	31,527,717	592,211,756

Table 1
Size of PubMed Layered Graphs

6.2. *Effectiveness of the Ranking Techniques with respect to Manjal*

To reproduce the Manjal's results [49], we ran the metric lgWP on a layered Discovery Graph $lgDG$ of curcumin (5 layers with 3,107,901 nodes and 10,261,791 edges). We ranked the target objects in the graph and observed that our ranking technique was able to produce 4 of the top-5 semantic associations identified by Manjal [49]; our ranking technique exhibits a precision of 80%. Table 2 compares the top-5 target objects discovered by Manjal [49] and the ones discovered by our ranking technique.

Benefits of the graph-sampling technique were also studied; we ran this sampling process for 5 iterations, i.e., 5 sampled subgraphs were computed. Table 3 highlights 4 out of the top-5 MeSH terms identified by Manjal [49], that are also identified by graph-sampling. Table 4 reports on the number of target MeSH terms produced by Manjal and the ones produced during each iteration of graph-sampling; graph-sampling is able to discover 80% of the top novel

k	Manjal Ranking	lgWP
1	**Retina**	**Testis**
2	**Spinal Cord**	**Retina**
3	**Testis**	**Spinal Cord**
4	**Pituitary Gland**	Obesity
5	Sciatic Nerve	**Pituitary Gland**

Table 2
Curcumin Top-5 MeSH terms- Manjal Ranking [49] versus lgWP Ranking

MeSH terms, while the number of target terms is reduced by up to one order of magnitude. Additionally, the number of nodes visited by the exact solution and by graph-sampling during 5 iterations is reported; also it shows the execution time of each iteration. The exact solution ran in 207.3 secs. while one iteration of the graph-sampling consumed around 60 secs; less than 3% of the graph nodes were visited by one iteration.

Finally, we measured the normalized Spearman's rho distance metric between the top-k MeSH discovered by the Manjal system [49], the exact implementation of the lgWP ranking technique, and the top-k terms produced during each iteration of graph-sampling; we report on values of k equal to 5, 10, and 20. In Figure 5(a) we can observe that the normalized Spearman's rho distance is 0.14 for iteration "i=5" indicating that both rankings are very similar; in fact this ranking is even better than the one provided by the exact lgWP solution. However, the similarity between these rankings is lower as k increases; Spearman's rho is almost 0.7 when k is 20.

Similarly, we ran the path-sampling technique to approximate the most relevant links between curcumin and the MeSH terms corresponding to diseases; the sampling process generated between 10 and 60 paths. Table 5 reports the top-10 MeSH terms identified by path-sampling. We can observe that of the top-5 MeSH terms identified by the Manjal system [49] (column 1 in Table 2), up to 4 are also identified among the top-10 MeSH terms identified by the path-sampling.

We also report on the number of target MeSH terms produced by Manjal and the ones produced during each iteration of path-sampling (Table 4). We can observe that path-sampling is able to discover 80% of the top novel MeSH terms, while the number of target terms is reduced by up to three orders of magnitude. Finally, we measured the normalized Spearman's rho distance metric between the top-k MeSH discovered by the Srinivasan's algorithm, the exact implementation of the lgWP ranking technique, and the

k	i=1	i=2	i=3	i=4	i=5
1	**Spinal Cord**	**Spinal Cord**	**Spinal Cord**	**Spinal Cord**	**Spinal Cord**
2	**Retina**	**Retina**	**Retina**	**Retina**	**Retina**
3	Pulmonary Alveoli	Pulmonary Alveoli	**Testis**	**Testis**	Pulmonary Alveoli
4	**Testis**	Astrocytoma	Glioblastoma	Pulmonary Alveoli	**Testis**
5	Astrocytoma	**Pituitary Gland**	Pulmonary Alveoli	Glioblastoma	**Pituitary Gland**
6	Hypothalamus	Glioblastoma	Peritonitis	**Pituitary Gland**	Glioblastoma
7	Meningitis	Meningitis	Astrocytoma	Astrocytoma	Meningitis
8	Peritonitis	Astrocytoma	Myocarditis	Meningitis	Pulmonary Artery
9	Obesity	Coronary Vessels	Testicular Neoplasms	Peritonitis	Neostriatum
10	Escherichia coli	Escherichia coli	Pituitary Gland	Anemia	Escherichia coli

Table 3
Curcumin-Effectiveness of Graph-Sampling Versus Top-5 Manjal's Terms

	Target MeSH Terms				
Manjal	570				
Graph-Sampling	i=1	i=2	i=3	i=4	i=5
	4.21%	6.66%	8.59%	10.70%	12.45%
Path-Sampling	i=10	i=20	i=30	i=40	i=60
	1.40%	2.98%	4.9%	5.96%	7.01 %
	Visited Nodes				
Exact lgWP	3,107,900				
Graph-Sampling	i=1	i=2	i=3	i=4	i=5
	0.28%	0.59%	0.89%	1.19%	1.49 %
Path-Sampling	i=10	i=20	i=30	i=40	i=60
	0.001%	0.003%	0.004%	0.006%	0.009 %
	Execution Time (secs.)				
Exact lgWP	230.7				
Graph-Sampling	i=1	i=2	i=3	i=4	i=5
	26%	52.44%	77.58%	105.33%	130%
Path-Sampling	i=10	i=20	i=30	i=40	i=60
	10.14%	21.41%	32.68%	43.95%	66.49%

Table 4
Efficiency of the RankingTechniques Curcumin-Sampling Techniques

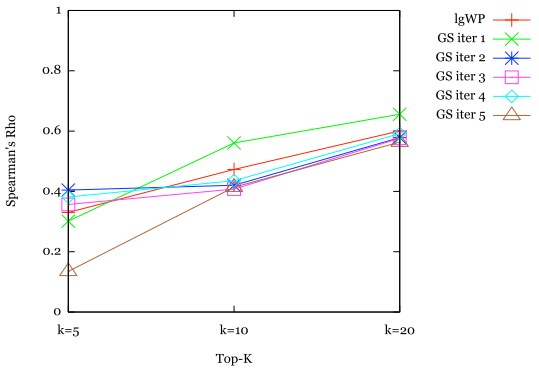
top-k terms produced during each iteration of path-sampling; we report on values of k equal to 5, 10 and 20. In Figure 5(b) we can observe that the normalized Spearman's rho distance is slightly different when 30, 40, 50 or 60 paths are sampled.

6.2.1. Discussion

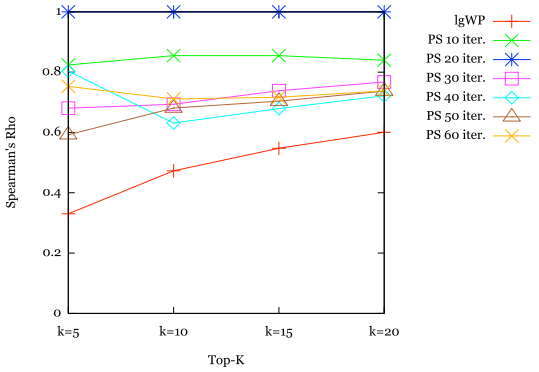
In this experiment we can observe that the lgWP is able to assign the highest scores to 4 of the top-5 terms identified by the Majal system. This indicates that the sub-graph generated by the local ranking performed by the Srinivasan's algorithm where only top-M B are considered during the search, is approximated by the sub-graph comprised of nodes with high values of lgWP. However, as can be seen in Table 4 only a reduced number of nodes comprised this sub-graph; thus, the sampling techniques exploit this property, and are able to accurately approximate this sub-graph and reach a great number of the top-5 novel MeSH terms just by visiting a small number of intermediate nodes. In fact, it can be seen from Figure 5, that graph-sampling seems to better approximate the Srinivasan's algorithm discoveries than the exact computation of

k	i=10	i=20	i=30	i=40	i=50	i=60
1	**Spinal Cord**	Anemia	Obesity	Anemia	**Testis**	**Testis**
2	Hypothalamus	Obesity	**Testis**	Obesity	**Pituitary Gland**	**Pituitary Gland**
3	Hyperinsulinism	Glomerulonephritis	**Pituitary Gland**	**Pituitary Gland**	**Spinal Cord**	Obesity
4	Diaphragm	Graves Disease	**Retina**	**Spinal Cord**	Obesity	Anemia
5	Tonsil	Graft vs Host Disease	Graft vs Host Disease	**Testis**	Coronary Vessels	Coronary Vessels
6	Gills	Astrocytoma	Astrocytoma	**Retina**	Hypothalamus	**Retina**
7	Shigella flexneri	Uremia	Pulmonary Artery	Coronary Vessels	Graft vs Host Disease	Hypothalamus
8	Bursa of Fabricius	Diabetic Angiopathies	Crohn Disease	Pulmonary Alveoli	Chimera	Glomerulonephritis
9		Toxoplasmosis	Pseudomonas Infections	Glomerulonephritis	Ovarian Follicle	Parkinson Disease
10		Teratoma	Parkinson Disease	Parkinson Disease	Celiac Disease	Cystic Fibrosis

Table 5
Curcumin- Effectiveness of Path-Sampling Technique



(a) Quality of Graph-Sampling



(b) Path-Sampling

Fig. 5. Quality of Sampling versus the Manjal System for the MeSH Curcumin. Spearman's Rho Values compare Top-5, Top-10, Top-15 and Top-20 Produced by Manjal and by the Proposed Sampling Techniques.

the lgPW metric. Finally, these results also suggest that these techniques can converge in few iterations; for example, path-sampling just needs to sample a small percentage of paths to reach a high accuracy, less than 1% of all the paths in the layered graph.

6.3. *Effectiveness of the Ranking Techniques to Predict Novel Discoveries*

We have also studied the quality of our proposed approach for the MeSH terms *gingko*, *aloe* and *tacromilus*; because there are no reported results for these three substances for the Manjal system, we consider as ground truths, associations with diseases that have been published by diverse specialized websites or databases.

6.3.1. *Effectiveness of Ranking Techniques for Gingko*

First, we report the results for gingko; ground truths correspond to rankings published by the Mayo Clinic public website[13]. The layered Discovery Graph for gingko is composed of 1,906,962 nodes and 31,909,304 edges, i.e., this graph is more dense than the curcumin layered Discovery Graph. Table 6 shows the ground truth considered in this experiment, while Table 7 presents the top-10 MeSH terms ranked by the exact implementation of the lgWP ranking technique. We can observe that: *i*) one out of the two diseases for which there is a strong scientific evidence of the use of gingko is ranked between the top-10 terms; *ii*) two diseases for which there is an unclear scientific evidence are identified; and *iii*) three diseases where gingko is used based on tradition or theory, are also discovered. We note that our ranking technique was able to highly

[13] http://www.mayoclinic.com/health/gingko-biloba/NS_patient-gingko/DSECTION=evidence

NIH Ranking	Evidence
Intermittent Claudication	A
Alzheimer Disease	A
Brain Diseases	B
Hemorrhoids	C
Memory Disorders	C
Altitude Sickness	C
Asthma	C
Cardiovascular Diseases	C
Chemotherapy	C
Venous Insufficiency	C
Cocaine-Related Disorders	C
Deafness	C
Depressive Disorder	C
Diabetic Neuropathies	C
Diabetes Mellitus	C
Hypertension	C
Neoplasms	T
Carcinoma	T
Colorectal Neoplasms	T
Arthritis	T

Table 6

Ground Truths for Gingko - A: strong scientific evidence; B: good scientific evidence; C: unclear scientific evidence; T: uses based on tradition or theory; NR: no reported.

k	lgWP	Evidence
1	**Carcinoma**	T
2	**Diabetes Mellitus**	C
3	**Neosplasms**	T
4	Leukemia	NR
5	**Hypertension**	C
6	Breast Neoplasms	NR
7	**Alzheimer Disease**	A
8	Liver Neoplasms	NR
9	**Arthritis**	T
10	Adenocarcinoma	NR

Table 7

Effectiveness of lgWP for Gingko. Top-5 MeSH terms Ranked by lgWP using exact solution are compared to Ground Truths for Gingko.

rank Alzheimer Disease for which there seems to be a strong scientific evidence of the usage of gingko.

In addition, Table 8 reports on visited nodes and execution time of 5 iterations of the sampling techniques. The exact solution ran in 536 secs., while one iteration of the graph-sampling techniques consumed around 161 secs. and the path-sampling consumed around 55% of the time consumed by the exact solution; the number of visited nodes by one iteration is at most one order of magnitude less than the number of nodes in the whole layered graph. It is important to notice that although path-sampling just visited a small number of paths, the majority of the runtime was consumed loading the graph in main memory.

Additionally, we could observe that the precision is 60% in each iteration, but something important to highlight is that after iteration "i=2", the MeSH term Alzheimer Disease is among the top-10 terms discovered by the graph-sampling technique; however, in the rest of the iterations, the quality of the ranking does not increase. We hypothesize that this is because the gingko layered Discovery Graph is very dense and there are many different ways to reach important nodes from the terms in the first layer. Furthermore, path-sampling was able to rank Alzheimer Disease as fourth after sampling 30 paths; four of the MeSH terms in Table 8, were identified in this iteration.

6.3.2. Effectiveness of Ranking Techniques for Aloe Vera

Similarly, we ran the ranking techniques for aloe vera; ground truths correspond to rankings published by the Mayo Clinic public website [14]. We executed path-sampling for 10 and 60 paths, graph-sampling for 1 and 2 iterations, and the exact ranking; Table 9 reports on the precision with respect to the ground truths. We observed that the exact ranking techniques could reach a precision of up to 33%. The majority of the terms discovered by these techniques correspond to terms whose evidence is based on tradition or theory; thus, the discoveries may help to support the veracity of these traditions. Finally, we compared the top-5 terms identified by the sampling techniques with respect to the top-5 terms identified by the exact technique; both techniques were able to reach up to 60% of precision with respect to the exact ranking of lgWP, while the number of visited nodes is up to five orders of magnitude less than the exact solution.

6.3.3. Effectiveness of Ranking Techniques for Tacrolimus

Finally, we studied the term tacrolimus; ground truths for this drug were taken from medical publications, and the RDF dataset *LinkedCT* (circa September 2010). Tacrolimus is an immunosuppressive drug which is usually used after liver, kidney and bone transplants to avoid immune system reactions; also it is used to prevent Graft-vs-Host Disease in patients after bone marrow transplantation [59], and to treat Crohn's disease[15]. Furthermore, tacrolimus may have the potential to be used in the treatment of Alzheimer's disease [61], and there are evidences of the development

[14] http://www.mayoclinic.com/health/aloe-vera/NS_patient-aloe/DSECTION=evidence

[15] http://www.nlm.nih.gov/medlineplus/druginfo/meds/a601117.html

Technique	% Visited Nodes					% Execution Time				
Exact	1,906,962					536.00				
Graph-Sampling	Iterations					Iterations				
	i=1	i=2	i=3	i=4	i=5	i=1	i=2	i=3	i=4	i=5
	10.87%	22.81%	34.95%	47.01%	60.66%	30.03%	59.38%	89.55%	118.28%	148.32%
Path-Sampling	Iterations					Iterations				
	i=10	i=20	i=30	i=40	i=50	i=10	i=20	i=30	i=40	i=50
	0.002%	0.004%	0.007%	0.01%	0.01%	55.59%	55.97%	56.34%	56.52%	57.27%

Table 8
Efficiency of lgWP for Gingko. Sampling Techniques are compared to Exact Solution.

	Precision				
Exact	30%				
Graph-Sampling	Iterations		Path-Sampling	# Paths	
	i=1	i=2		i=10	i=60
	20%	26%		33%	20%

Table 9
Effectiveness for Aloe. Precision of the Ranking Techniques

of Diabetes Mellitus in the first 2 months after renal transplants [48].

In our experiments, we can observe that the exact lgWP ranking technique ranked among 2,228 terms: Leukemia as second, Diabetes Mellitus as third, Kidney Failure as 15th, Graft-vs-Host Disease as 22th and Alzheimer Disease as 31st. Although the layered graph for tacrolimus is very large and is composed of 592,211,756 paths, path-sampling was able to rank: Leukemia as the top-1, Diabetes Mellitus as second, Alzheimer Disease as 12th and Graft vs Host Disease as 16th, by just sampling 30 paths. Additionally, graph-sampling ranked among 88 terms: Leukemia as the top-1, Graft-vs-Host Disease as fourth, Kidney Failure as seventh, Autoimmune Disease as 14th and Crohn's disease as 30th, in just one iteration. The number of nodes visited by graph-sampling was 85,087 during iteration 1, while path-sampling only generated 60 paths. Table 10 summarizes these results.

Additionally, we queried the *LinkedCT* dataset[16] and retrieved the clinical trials and the diseases for which the effects of tacrolimus were studied. *LinkedCT* is an RDF dataset of the Cloud of Linked Data, which maintains the trials published by the *ClinicalTrials.gov* web site and their corresponding links to *DBpedia*, *DrugBank*, *Diseasome*, *DailyMed*, *GeoNames*, *PubMed*, etc. First, we ran a SPARQL query against *LinkedCT* to output the diseases (condition_name) associated with a clinical trial whose drug (intervention) is tacrolimus; the answer is comprised of 70 diseases. Then, we compared the top-15 terms identified by our ranking techniques with respect to these 70 diseases, and computed the percentage of the discovered terms, i.e., the precision of the discovered terms with respect to the diseases retrieved from *LinkedCT*; Table 11 reports these values. We can see that among the top-15 terms identified by our ranking techniques, at least 40% correspond to associations already published in *LinkedCT*; for up to 75% of the rest of the top-15 terms, there is a bibliographical evidence of the relationship between the corresponding term and tacrolimus.

Second, we retrieved the references of the condition names in *DBpedia* or *Diseasome* that are associated with a clinical trial whose drug (intervention) is tacrolimus; we obtained 6 diseases. These references were created by using the linking technique proposed by Hassanzadeh et al. in [20]. We notice that our technique is able to detect all least 50% of the links found by the Hassanzadeh's approach for tacrolimus. Additionally, we could identify 9 more links that this technique was unable to find. This indicates that our proposed ranking techniques provide a possible solution to the problem of discovery meaningful links or validating existing links between data in the Cloud of Linked Data. Finally, we report on the number of vis-

[16] http://linkedct.org/index.html

Baseline	Exact lgWP	Graph-Sampling	Path-Sampling
Leukemia	Leukemia (2nd)	Leukemia (2nd)	Leukemia (1st)
Diabetes Mellitus	Diabetes Mellitus (3rd)	Diabetes Mellitus (9th)	Diabetes Mellitus (2nd)
Autoimmune Disease	Autoimmune Disease (18th)	Autoimmune Disease (14th)	Autoimmune Disease (14th)
Liver Transplant	Liver Disease (35th)	Liver Disease (18th)	_
Kidney Transplant	Kidney Disease (20th)	Kidney Disease (3rd)	Kidney Disease (9th)
Bone Marrow Transplant	Bone Marrow Disease (399th)	_	_
Graft-vs-Host Disease	Graft-vs-Host Disease (22th)	Graft-vs-Host Disease (1st)	Graft-vs-Host Disease (15th)
Crohn's Disease	Crohn's Disease (70th)	Crohn's Disease (20th)	Crohn's Disease (27th)
Alzheimer's Disease	Alzheimer's Disease (31st)	Alzheimer's Disease (44th)	_

Table 10
Effectiveness for Tacrolimus. Ranking Techniques are compared to Baseline of Scientific Publications

Technique	Precision w.r.t. LinkedCT Data		Precision w.r.t. LinkedCT Dereferences		#Visited Nodes	
Exact	40%		50%		5,569,784	
Graph-Sampling	Iterations		Iterations		Iterations	
	i=1	i=2	i=1	i=2	i=1	i=2
	40%	66%	**75%**	**75%**	**0.83%**	**1.80%**
Path-Sampling	Iterations		Iterations		Iterations	
	i=30	i=60	i=30	i=60	i=30	i=60
	73%	53%	50%	**75%**	**0.002%**	**0.005%**

Table 11
Tacrolimus-Efficiency and Effectiveness of the Ranking Techniques

ited nodes; as in previous experiments, we observe that the sampling techniques visited up to 4 orders of magnitude less nodes than the exact solution, providing an efficient solution to the problem of estimating novel associations between drugs and diseases.

6.3.4. Discussion

In these three experiments we can observe the effectiveness of the ranking metric lgWP; in many cases high scores are assigned to associations reported by different sources, i.e., scientific publications, specialized web sites, RDF datasets or linked datasets. However, it can also be observed that a large number of publications directly related to the studied drugs or substances, can be irrelevant. Our techniques do not only discover novel potential associations between terms, but they also contribute to filter irrelevant concepts, and provide a solution for searching relevant publications or sources of scientific data that corroborate existing associations between drugs and diseases.

6.4. Effectiveness of the Ranking Techniques on Bibliographic Data

Additionally, we consider DBLP bibliographic data, and ran the exact ranking and the sampling techniques to discover associations between a given author and the most relevant conferences where this author has published at least one paper. We ran 3 sets of 30 queries and compared the ranking produced by the exact solution and the rankings produced by the sampling techniques; layered Discovery Graphs were comprised of 5 layers with at most 876,110 nodes, 4,166,626 edges, and 28,690 paths. Author's names with high, medium and low selectivity were considered; a highly selective name corresponds to an author with less than 10 publications, and a low selective name is associated with more than 300 publications.

The top-5 conferences associated with each author were computed by using the exact ranking and the approximation produced by graph-sampling and path-sampling during 6 iterations; a conference is among the most important conferences of a given author, if the conference has had several editions and the author has published several papers in the conference. Ta-

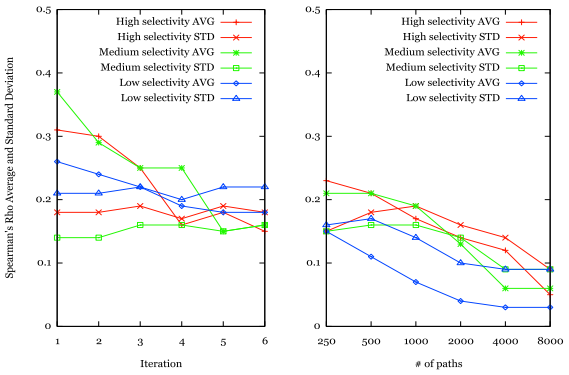


Fig. 6. DBLP- Spearman's rho of Graph-Sampling and Path-Sampling (Average;Standard Deviation)

ble 12 reports the average precision and standard deviation of the approximate top-5 conferences with respect to the exact top-5. We can observe that graph-sampling is able to identify almost 65% of the top-5 conferences after iteration 3, and path-sampling by just producing 250 paths, is able to reach a precision of up to 82%; precision increases with the number of samplings reaching up to 98%. Additionally, we measured Spearman's rho of the top-5 terms produced by the sampling techniques with regards to the top-5 produced by the exact solution; Figure 6 shows the average Spearman's rho and the standard deviation. We can observe that the orderings given by the sampling techniques are close to the ones given by the exact solution, e.g., after producing 8,000 paths, path-sampling produced a top-5 list whose order is almost the same to the one given by the exact solution.

Furthermore, Figures 7(a) and 7(b) report the execution time of graph-sampling and path-sampling; these values include the time to load the graph and the execution time. We can observe that both graph-sampling and path-sampling, are able to reach a precision up to 98%, and produce a ranking relatively close to the exact ranking after the second iteration of these two algorithms; the execution time of the exact technique is one order of magnitude greater than the time of two iterations. These results suggest that the proposed discovery techniques provide an effective and efficient solution to the problem of identifying associations between terms also in the bibliographic domain.

6.4.1. Discussion

In this experiment we can observe that even the bibliographic data is not annotated with controlled vocabularies as MeSH, the authority-flow based metric is able to capture the topology of graph that represents the relationships between concepts, and provide meaningful rankings. As in previous experiments, only a reduced number of associations are possibly relevant; thus, the effectiveness of sampling techniques that only visit the terms that point to these relevant nodes are required. As shown in Table 12 and Figure 7(b), our sampling techniques seem to achieve this requirement and efficiently and effectively identify the top-k terms.

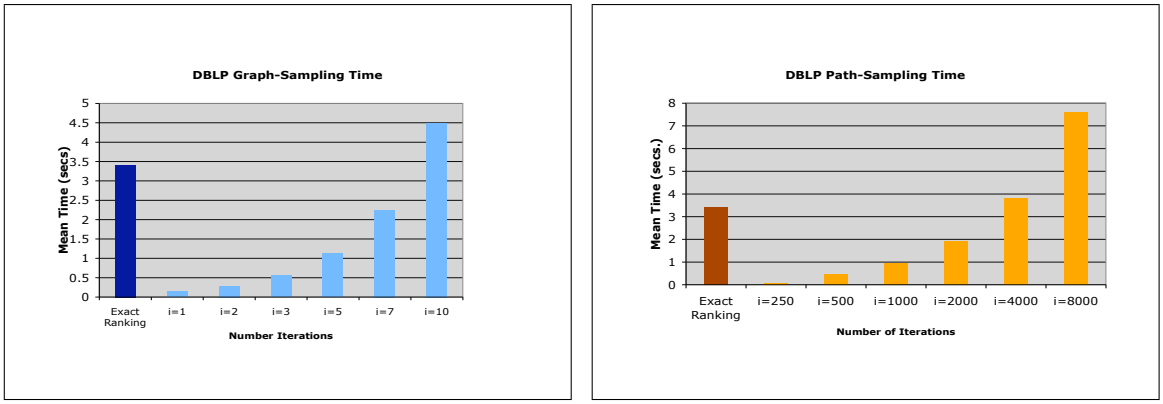
7. Conclusions and Future Work

In this paper we have presented an authority-flow based ranking metric that considers the topology of the data connections as well as the semantic annotations of the data, to identify potential novel annotations. Biological objects (e.g., genes or proteins) or clinical trials are annotated with controlled vocabulary terms from ontologies such as GO, MeSH, SNOMED; many of these datasets have been made available in the Cloud of Linked Data, and their intra- and inter-datasets links induce graphs that capture meaningful knowledge. Thus, techniques that consider the topology of these links may be useful to explain existing phenomena, identify anomalies and potentially lead to a new discovery. This hypothesis was corroborated in this paper with two types of datasets, one comprised of scientific publications and their annotations, and another composed of bibliographic data; target concepts that were part of dense sub-graphs correspond to relevant concepts reported in specialized sources of data.

To identify the nodes that may be part of these dense sub-graphs, we propose an authority-flow ranking metric which ranks target objects in terms of the authority transferred from their parents. We could observe that this ranking technique is able to discriminate among a large number of potential relevant concepts, those that have been shown as relevant by Literature-based approaches as Manjal, or reported in scientific websites, datasets or publications. We also could see that from a large number of possible relevant concepts, a very small number is actually relevant. So, approximate techniques able to efficiently guide the search into the space of these relevant concepts are required; based on our experimental results, we could say that our proposed approximate ranking techniques meet this requirement and are able to efficiently traverse this space and identify the potential relevant concepts reported by other techniques or specialized sources of data. In some cases, the provided approximation is even better than the one found when the exact computation of the

	Selectivity	i=1	i=2	i=3	i=4	i=5	i=6
Graph-Sampling	high	(39;40)	(48;38)	(63;35)	(**81;25**)	(**82;25**)	(**87;19**)
	medium	(34;29)	(56;33)	(68;30)	(72;28)	(**87;19**)	(**89;15**)
	low	(64;35)	(66;36)	(75;31)	(**80;29**)	(**80;28**)	(**81;28**)
	Selectivity	i=250	i=500	i=1,000	i=2,000	i=4,000	i=8,000
Path Sampling	high	(43;38)	(66;33)	(79;30)	(**87;19**)	(**91:13**)	(**96:8**)
	medium	(66;32)	(75;28)	(**83;23**)	(**92;12**)	(**98;5**)	(**98;4**)
	low	(**82;26**)	(**90;25**)	(**94;20**)	(**95;18**)	(**95;18**)	(**95;18**)

Table 12
DBLP- Precision of Graph-Sampling and Path-Sampling DBLP(Average;Standard Deviation)



(a) DBLP Graph-Sampling Execution Time (secs.) (b) DBLP Path-Sampling Execution Time (secs.)

Fig. 7. DBLP- Execution Time Exact Ranking versus Sampling Techniques (Time secs.)

ranking metric lgWP was performed. Additionally, the experimental results suggest that the approximate techniques can converge in few iterations; in many cases, just a sample that corresponds to less than 1% of all population is needed to reach high levels of accuracy. Although we could observe that these techniques are very efficient in terms of the number of visited nodes, the same savings are not observed in the execution time. The cause of this undesirable behavior is that the time to transfer data from the database to main memory is the dominant contribution to the execution time of a discovery request. So, these techniques should be extended to retrieve from the database only the concepts that contribute to the computation of the relevant target objects. Also, the semantics encoded in the controlled vocabularies used to annotate the objects, may play an important role in the discovery process. Thus, score functions able to capture relatedness between annotations should be also considered; these semantic functions could reduce even more the search space and increase the effectiveness of the techniques. We note that in our experimental evaluation the impact of the semantics encoded in the ontologies was not reported; we just focused on showing the effects of the topology of the links. Nevertheless, we also conducted preliminary experiments considering the *taxonomic* score function defined in Section 4.1; the accuracy of the discoveries was increased in cases where the current techniques performed poorly.

In the future, we plan to apply these techniques to inter-links between several datasets in the Cloud of Linked Data, as well as define semantic similarity measures to reflect the semantics encoded in the ontologies used to annotate the biological concepts and support an *Open World Assumption* reasoning process. Further, the sampling techniques should be also enhanced with the capability to reduce not only the traversed nodes, but also to retrieve only the relevant objects from the dataset. So, these sampling techniques will be incorporated to existing semantic management approaches to only retrieve the elements that will be part of the sub-graph that includes the potential novel concepts. Finally, our current approach implements a blocking query execution engine, where results are

produced only after all the data is received from the sources. This decision impacts on both execution time and quality of the answer. In the future we plan to develop adaptive and dynamic approaches able to adapt the ranking process to unexpected data transfers and discontinuous data arrivals.

References

[1] H. Alani, D. Dupplaw, J. Sheridan, K. O'Hara, J. Darlington, N. Shadbolt, and C. Tullo. Unlocking the potential of public sector information with semantic web technology. In *ISWC/ASWC*, pages 708–721, 2007.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.

[3] M. Chavira, A. Darwiche, and M. Jaeger. Compiling relational bayesian networks for exact inference. *Int. J. Approx. Reasoning*, 42(1-2):4–20, 2006.

[4] J. Cheng and M. J. Druzdzel. Ais-bn: An adaptive importance sampling algorithm for evidential reasoning in large bayesian networks. *CoRR*, abs/1106.0253, 2011.

[5] K.-H. Cheung, H. R. Frost, M. S. Marshall, E. Prud'hommeaux, M. Samwald, J. Zhao, and A. Paschke. A journey to semantic web query federation in the life sciences. *BMC Bioinformatics*, 10(S-10), 2009.

[6] A. Darwiche. Conditioning algorithms for exact and approximate inference in causal networks. In *UAI*, pages 99–107, 1995.

[7] L. Ding, D. DiFranzo, A. Graves, J. Michaelis, X. Li, D. L. McGuinness, and J. A. Hendler. Twc data-gov corpus: incrementally generating linked government data from data.gov. In *WWW*, pages 1383–1386, 2010.

[8] Disease Ontology. http://diseaseontology.sourceforge.net.

[9] EHR Ontology. http://trajano.us.es./ isabel/EHR/EHRRM.owl.

[10] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. In *SODA*, pages 28–36, 2003.

[11] D. Fogaras, B. Rácz, K. Csalogány, and T. Sarlós. Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments. *Internet Mathematics*, 2(3), 2005.

[12] L. Getoor and C. P. Diehl. Introduction to the special issue on link mining. *SIGKDD Explorations*, 7(2), 2005.

[13] The Gene Ontology. http://www.geneontology.org/.

[14] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabasi. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104:8685–8690, 2007.

[15] H. Guo and W. Hsu. A survey of algorithms for real-time bayesian network inference. In *In In the joint AAAI-02/KDD-02/UAI-02 workshop on Real-Time Decision Support and Diagnosis Systems*, 2002.

[16] C. Halaschek-Wiener, B. Aleman-Meza, I. B. Arpinar, and A. P. Sheth. Discovering and ranking semantic associations over a large rdf metabase. In *VLDB*, pages 1317–1320, 2004.

[17] J. Han, X. Yan, and P. S. Yu. Mining, indexing, and similarity search in graphs and complex structures. In *ICDE*, page 106, 2006.

[18] J. Hannemann and J. Kett. Linked data for libraries. In *IFLA World Library and Information Congress*, 2010.

[19] O. Hassanzadeh, A. Kementsietsidis, L. Lim, R. J. Miller, and M. Wang. A framework for semantic link discovery over relational data. In *CIKM*, pages 1027–1036, 2009.

[20] O. Hassanzadeh, A. Kementsietsidis, L. Lim, R. J. Miller, and M. Wang. Linkedct: A linked data space for clinical trials. In *Proceedings of the WWW2009 workshop on Linked Data on the Web (LDOW2009)*, 2009.

[21] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers, 2011.

[22] M. Henrion. Propagating uncertainty in bayesian networks by probabilistic logic sampling. *Uncertainty in Artificial Intelligence 2*, page 149Ð163, 1988.

[23] H. Hu, X. Yan, Y. H. 0003, J. Han, and X. J. Zhou. Mining coherent dense subgraphs across massive biological networks for functional discovery. In *ISMB (Supplement of Bioinformatics)*, pages 213–221, 2005.

[24] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *KDD*, pages 538–543, 2002.

[25] J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *CoRR*, cmp-lg/9709008, 1997.

[26] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[27] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee. Media meets semantic web - how the bbc uses dbpedia and linked data to make connections. In *ESWC*, pages 723–737, 2009.

[28] M. Kuramochi and G. Karypis. Finding frequent patterns in a large sparse graph*. *Data Min. Knowl. Discov.*, 11(3):243–271, 2005.

[29] D. Lin. An information-theoretic definition of similarity. In *ICML*, pages 296–304, 1998.

[30] Y. Lin and M. J. Druzdzel. Stochastic sampling and search in belief updating algorithms for very large bayesian networks. In *Working Notes of the AAAI Spring Symposium on Search Techniques for Problem Solving Under Uncertainty and Incomplete Information, Stanford University, Stanford, California,*, pages 77–82, 1999.

[31] Medical Subject Heading (MeSH). http://www.nlm.nih/gov/mesh.

[32] V. Momtchev, D. Peychev, T. Primov, and G. Georgiev. Expanding the pathway and interaction knowledge in linked life data. In *International Semantic Web Challenge*, 2009.

[33] J. L. Moore, F. Steinke, and V. Tresp. A novel metric for information retrieval in semantic networks. In *ESWC Workshops*, pages 65–79, 2011.

[34] S. Moral and A. Salmerón. Dynamic importance sampling in bayesian networks based on probability trees. *Int. J. Approx. Reasoning*, 38(3):245–261, 2005.

[35] O. C. Organization. GALEN common reference model.

[36] R. Parundekar, C. Knoblock, and J. L. Ambite. Linking the deep web to the linked dataweb. In *AAAI Spring Symposium Series*, 2010.

[37] T. Pedersen, S. V. S. Pakhomov, B. T. McInnes, and Y. Liu. Measuring the similarity and relatedness of concepts in the medical domain: Ihi 2012 tutorial overview. In *IHI*, pages 879–880, 2012.

[38] V. Pekar and S. Staab. Taxonomy learning - factoring the structure of a taxonomy into a semantic classification decision. In *COLING*, 2002.

[39] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, and F. M. Couto. Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7), 2009.

[40] Y. Raimond, C. Sutton, and M. B. Sandler. Interlinking music-related data on the web. *IEEE MultiMedia*, 16(2):52–63, 2009.

[41] L. Raschid, Y. Wu, W. Lee, M. Vidal, P.Tsaparas, P. Srinivasan, and A. Sehgal. Ranking target objects of navigational queries. In *WIDM*, pages 27–34, 2006.

[42] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal Of Artificial Intelligence Research*, 11:95–130, 1999.

[43] S. Russell and P.Norvig. *Artificial Intelligence: A Modern Approach. Second Edition*. Princeton Hall, 2003.

[44] An Overview to RxNorm. http://www.nlm.nih.gov/research/umls/rxnorm/.

[45] B. Saha, A. Hoch, S. Khuller, L. Raschid, and X.-N. Zhang. Dense subgraphs with restrictions and applications to gene annotation graphs. In *RECOMB*, pages 456–472, 2010.

[46] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988.

[47] M. Samwald, H. Chen, A. Ruttenberg, E. Lim, L. N. Marenco, P. L. Miller, G. M. Shepherd, and K.-H. Cheung. Semantic senselab: Implementing the vision of the semantic web in neuroscience. *Artificial Intelligence in Medicine*, 48(1):21–28, 2010.

[48] T. Sato, A. Inagaki, K. Uchida, T. Ueki, N. Goto, S. Matsuoka, A. Katayama, T. Haba, Y. Tominaga, Y. Okajima, K. Ohta, H. Suga, S. Taguchi, S. Kakiya, T. Itatsu, T. Kobayashi, and A. Nakao. Diabetes mellitus after transplant: relationship to pretransplant glucose metabolism and tacrolimus or cyclosporine a-based therapy. *Transplantation*, 76(9), 2003.

[49] P. Srinivasan, b. Libbus, and A. Kumar. Mining medline: Postulating a beneficial role for curcumin longa in retinal diseases. In L. Hirschman and J. Pustejovsky, editors, *LT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*, pages 33–40, 2004.

[50] X. Sun, M. J. Druzdzel, and C. Yuan. Dynamic weighting a* search-based map algorithm for bayesian networks. In *Probabilistic Graphical Models*, pages 279–286, 2006.

[51] D. Swanson. Migraine and magnesium: Eleven neglected connections. In *Perspective in Biology and Medicine*, 1988.

[52] A. Thor, P. Anderson, L. Raschid, S. Navlakha, B. Saha, S. Khuller, and X.-N. Zhang. Link prediction for annotation graphs using graph summarization. In *International Semantic Web Conference (1)*, pages 714–729, 2011.

[53] N. Toupikov, J. Umbrich, R. Delbru, M. Hausenblas, and G. Tummarello. Ding! dataset ranking using formal descriptions. In *Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference (WWW 09)*, 2009.

[54] R. Varadarajan, V. Hristidis, L. Raschid, M.-E. Vidal, L. Ibáñez, and H. Rodríguez-Drumond. Flexible and efficient querying and ranking on hyperlinked data sources. In *EDBT*, pages 553–564, 2009.

[55] M.-E. Vidal, E. Ruckhaus, and N. Marquez. BioNav: A System to Discover Semantic Web Associations in the Life Sciences. In *ESWC 09-Poster Session*, 2009.

[56] voiD Guide - Using the Vocabulary of Interlinked Datasets. http://rdfs.org/ns/void-guide.

[57] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the web of data. In *International Semantic Web Conference*, pages 650–665, 2009.

[58] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.

[59] J. Wingard, R. Nash, D. Przepiorka, J. Klein, D. Weisdorf, J. Fay, J. Zhu, R. Maher, W. Fitzsimmons, and V. Ratanatharathorn. Relationship of tacrolimus (fk506) whole blood concentrations and efficacy and safety after hlaidentical sibling bone marrow transplantation. *Biology of Blood and Marrow Transplantation*, 4, 1998.

[60] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34, 2006.

[61] Y. Yoshiyama, M. Higuchi, B. Zhang, S. Huang, N. Iwata, T. Saido, J. Maeda, T. Suhara, J. Trojanowski, and V. Lee. Synapse loss and microglial activation precede tangles in a p301s tauopathy mouse model. *Source Neuron*, 53(3), 2007.

[62] C. Yuan and M. J. Druzdzel. Importance sampling algorithms for bayesian networks: Principles and performance. *Mathematical and Computer Modelling*, 43(9-10):1189–1207, 2006.