Exploiting Semantics from Ontologies and Shared Annotations to Find Patterns in Annotated Linked Open Data

Guillermo Palma[1], Maria-Esther Vidal[1], Louiqa Raschid[2], and Andreas Thor[3]

[1] Universidad Simón Bolívar, Venezuela
[2] University of Maryland, USA
[3] University of Leipzig, Germany
gpalma@ldc.usb.ve, mvidal@ldc.usb.ve, louiqa@umiacs.umd.edu,
thor@informatik.uni-leipzig.de

Abstract. Linked Open Data initiatives have made available a diversity of collections that domain experts have annotated with controlled vocabulary terms from ontologies. The challenge is to explore these rich and complex annotated datasets, together with the domain semantics captured within ontologies, to discover patterns of annotations across multiple concepts that may lead to potential discoveries. We identify an annotation signature between a pair of concepts based on shared annotations and ontological relatedness. Formally, an annotation signature is a partitioning of the edges that represent the relationships between shared annotations. A clustering algorithm named *AnnSigClustering* is proposed to generate annotation signatures. Evaluation results over drug and gene datasets demonstrate the effectiveness of using annotation signatures to find patterns.

1 Introduction

Ontologies are developed by domain experts to capture knowledge specific to some domain. They have been extensively developed and widely adopted in the last decade. Simultaneously, Linked Open Data initiatives have made available a diversity of collections that have been annotated with controlled vocabulary (CV) terms from these ontologies. For example, the biomedical community has taken the lead in such activities; every model organism database has genes and proteins that are widely annotated with CV terms from the Gene Ontology (GO). The NCI Thesaurus (NCIt) version 12.05d has 93,788 terms and the LinkedCT dataset of clinical trial results *circa* September 2011 includes 142,207 drugs or interventions, 167,012 conditions or diseases, and 166,890 links to DBPedia, DrugBank and Diseasome. At the opposite end of the domain spectrum, the Financial Industry Business Ontology (FIBO) captures knowledge about the structure, properties and behavior of financial contracts.

The challenge is to explore these rich and complex annotated datasets, together with the domain semantics captured within ontologies, to discover patterns of annotations across multiple concepts that may lead to potential discoveries. For genes, these patterns may involve cross-genome functional annotations, e.g., combining the GO functional annotations of two model organisms such as Arabidopsis thaliana (a plant) and C. elegans (a nematode or worm), to predict new gene function or protein-protein interactions. Drug target prediction, with a goal of finding new targets for existing drugs, has received widespread media attention and has resulted in some notable successes, e.g., `Viagra`. Additional applications include predicting potentially adverse side-effects or providing a comprehensive summary of drug effectiveness so that health professionals may find cost-effective treatments [10].

As a first step to discovering complex annotation patterns, we define an *annotation signature* between a pair of scientific concepts, e.g., a pair of drugs or a pair of genes. The annotation signature builds upon the shared annotations or shared CV terms between the pair of concepts. The signature further makes use of knowledge in the ontology to determine the ontological relatedness of the shared CV terms. The annotation signature is represented by N groups (clusters) of ontologically related shared CV terms. For example, the annotation signature for a (drug, drug) pair will be a set of N clusters, where each cluster includes a group of ontologically related disease terms from the NCIt.

Given a pair of concepts, and their sets of annotations, A_i and A_j from ontology O, elements $a_i \in A_i$ and $a_j \in A_j$ form the nodes of a bipartite graph BG. Between nodes a_i and a_j there may be an edge or a path through O; an edge is the special case where a_i and a_j are identical CV terms from O. There may be a choice of paths between a_i and a_j depending on the the ontology structure and relationship types captured within O. One can use a variety of similarity metrics, applied to the edges and paths through the ontology O, to induce a weighted edge between a_i and a_j in BG; the weight represents the (ontologically related) similarity score in the range $[0.0, 1.0]$ between a_i and a_j.

Our objective is to determine an annotation signature based on the bipartite graph BG. There are many alternatives to create the signature. One could partition the edges of BG with possible overlap of the nodes. Another solution is to cluster the nodes and edges of BG. One may also consider a one-to-one bipartite match [8].

We define a version of the *Annotation Signature Partition* problem as the partitioning of the edges of BG into clusters such that the value of the aggregated cluster density is maximized; we will define the density metric in the paper. We develop *AnnSigClustering*, a clustering solution that implements a greedy iterative algorithm to cluster the edges in BG. We note that such a clustering will result in N clusters of the edges of BG with potential overlap of nodes in different clusters.

We perform an extensive evaluation of the effectiveness of the annotation signature on real-world datasets of genes and their GO annotations, as well as on the LinkedCT dataset of drugs and diseases from NCIt and their associations through the clinical trials.

Our research focuses on exploiting domain specific semantic knowledge. This includes both the ontology structure and relationship types between concepts. We show that by using the ontology structure to tune the (ontologically related) similarity score between node pairs a_i and a_j, we can control the annotation signature to produce clusters of more closely related terms that are more useful to the domain scientist. Further, the choice of specific relationship types can be used to further refine the clusters of CV terms in the annotation signature.

The contributions of this paper can be summarized as follows: *i*) Definition of an *annotation signature* to mine annotated datasets together with domain specific semantic knowledge captured within ontologies. *ii*) A greedy iterative algorithm that exploits knowledge encoded in an ontology to discover the signature of a pair of annotated concepts. *iii*) An empirical study that suggests that annotation signatures represent interesting patterns across drugs and across genes.

This paper is organized as follows: Section 2 presents annotation graphs from different domains and Section 3 defines our approach. Experimental results are reported in Section 4, while related work is summarized in Section 5. Section 6 concludes.

2 Motivating Example

An antineoplastic agent is a substance that inhibits the maturation, growth or spread of tumor cells. Monoclonal antibodies that are also antineoplastic agents have become an important tool in cancer treatments. When used as a medication, the non-proprietary drug name ends in -mab. Scientists are interested in studying the relationships between drugs and the corresponding diseases; drugs are annotated with the NCIt terms that correspond to the conditions that have been tested for these drugs. Figure 1 illustrates `Brentuximab vedotin` and `Catumaxomab` and some of their annotations. Each path between a pair of conditions, e.g., `Colorectal Carcinoma` and `Stage IV Rectal Cancer` through the NCIt is identified using red ovals which represent CV terms from the NCIt. From Figure 1, we may conclude that the shared disease signature for this pair of drugs includes five components. The three terms `Colon Carcinoma`, `Colorectal Carcinoma` and `Stage IV Rectal Cancer` may form one component. Similarly, another component may include `Thyroid Gland Neoplasm`, `Oropharyngeal Neoplasm` and `Head and Neck Neoplasm`.

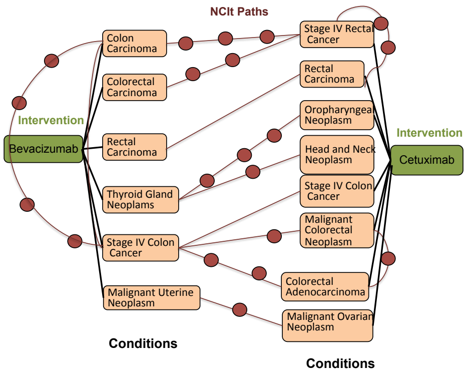


Fig. 1. Annotation graph representing the annotations of Brentuximab vedotin and Catumaxomab. Drugs are green rectangles; diseases are pink rectangles; NCIt terms are red ovals.

Consider a pair of financial contracts representing bonds (corporate, municipal, state, sovereign, etc.) from a repository such as EMMA [1]. Figure 2 shows an example of two bond contracts (green rectangles). These bonds are described by their CUSIP identifier, maturity date, principal, initial offering price, yield, etc. Each contract is also

[1] `http://www.emma.msrb.org/`

associated with a set of FIBO terms (pink ovals). For example, the Financial Contract A is associated with five terms including `Joint Guaranty` and `State Gurantor` while the Financial Contract is associated with seven terms. There is an edge with similarity equal to 1.0 between identical FIBO terms as well as paths through the FIBO ontology and intermediate FIBO terms (red circles).

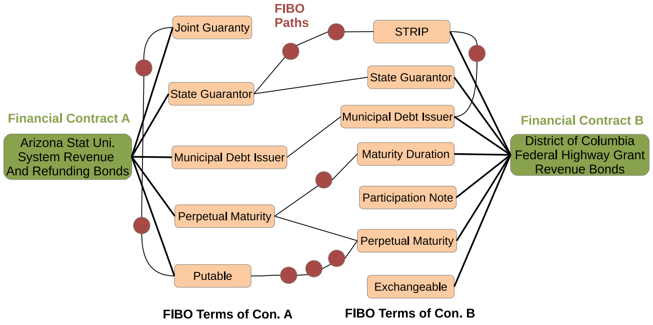


Fig. 2. FIBO terms (pink ovals) annotate a pair of financial contracts (green rectangles). An edge connects identical FIBO terms in the bipartite graph between the two sets of annotations on the left and right. Paths pass through intermediate FIBO terms (red circles).

3 Our Approach

A broad variety of similarity metrics have been proposed in the literature and have been summarized in [2]. Existing similarity metrics include the following: *i*) string-similarity metrics that measure similarity using (approximate) string matching functions; *ii*) path-similarity metrics such as *PathSim* and *HeteSim* that compute relatedness based on the paths that connect concepts in a graph; and *iii*) topological-similarity metrics that measure relatedness in terms of the closeness of CV terms in a given taxonomy or ontology.

We use a taxonomic distance metric d_{tax} [2]. The intuition behind the d_{tax} metric is to capture the taxonomic distance between two vertices with respect to the depth of the common ancestor of these two vertices. Additionally, d_{tax} tries to assign low(er) values of taxonomic distance to pairs of vertices that are: (1) at greater depth in the taxonomy and (2) are closer to their lowest common ancestor. A value close to 0.0 means that the two vertices are close to the leaves and both are close to their lowest common ancestor. A value close to 1.0 represents that both vertices are general or that the lowest common ancestor is close to the root of the taxonomy. Then, (1 - d_{tax}) will be used as the similarity or *ontological relatedness* between the two nodes.

The taxonomic distance metric d_{tax} is as follows, where *root* is the root node in the ontology; lca is the lowest common ancestor, and *pl* denotes path length:

$$d_{tax}(x, y) = \frac{pl(lca(x, y), x) + pl(lca(x, y), y)}{pl(root, x) + pl(root, y)} \tag{1}$$

Recall that we wish to utilize knowledge from the ontology; one option is to fully exploit ontology structure. A CV term that is farther up in the ontology, towards the root, is typically a general concept and its presence in a cluster is less interesting to scientists. This is especially true if the cluster has CV terms at much greater depth. Our goal is to reduce the number of such general concepts that occur in the annotation signature. To do so, we define an extension of d_{tax} named d^{str}_{tax}; it will assign low values of ontological relatedness (similarity) to pairs of CV terms where at least one of the terms is a general concept in the ontology. Let *MaxDepth_Ontology* represent the greatest depth in the ontology.

$$d^{str}_{tax}(A,B) = d_{tax}(A,B) * (1 - pFactor(A,B)) \tag{2}$$

$$pFactor(A,B) = \frac{\max(correctedDepth(A), correctedDepth(B))}{MaxDepth\_Ontology}$$

$$correctedDepth(X) = MaxDepth\_Ontology - Depth(X)$$

Definition 1 (Cluster Density). *Given a labeled bipartite graph BG=($A_i \cup A_j$, WE) with nodes A_i and A_j and edges WE, a distance metric d, and a subset p of WE, the cluster density of p $cDensity(p) = \frac{\sum_{e \in p} 1-d(e)}{|p|}$.*

Definition 2 (The Annotation Signature Partition Problem). *Given a labeled bipartite graph BG=($A_i \cup A_j$, WE), a distance metric d, and a real number θ in the range [0.0:1.0]. For each $a \in A_i$ and $b \in A_j$, if 1-d(a,b) $> \theta$, then there is an edge $e = (a,b) \in$ WE. For each $e = (a,b) \in$ WE, label(e)= 1-d(a,b). The AnnSig Partition Problem identifies a (minimal) partition P of WE such that the aggregate cluster density P $AnnSig(P) = \frac{\sum_{p \in P}(cDensity(p))}{|P|}$ is maximal.*

AnnSigClustering is a greedy iterative algorithm to solve the *Annotation Signature* Partition Problem. *AnnSigClustering* adds an edge to a cluster following a greedy heuristic to create clusters that maximize the cluster density. *AnnSigClustering* assigns a score to an edge e in *WE* according to the number of edges whose adjacent terms are dissimilar to the terms of e, and that have been already assigned to a cluster. Then, edges are chosen in terms of this score (descendant order). Intuitively, selecting an edge with the maximum score, allows *AnnSigClustering* to place first the edges with more restrictions; this is one for which there is a smaller set of potential clusters. The selected edge is assigned to the cluster that maximized the cluster density function. Time complexity of *AnnSigClustering* is $O(|WE|^3)$. To illustrate the behavior of *AnnSigClustering*, lets consider the annotated graph in Figure 2. This graph can be partitioned into 2 groups of edges, e.g., one group includes the edges between `State Guarantor` on the left with two terms `STRIP` and `State Guarantor` on the right; also, the edge between `Municipal Debt Issuer` belongs to this group. The other group is comprised of edges between `Perpetual Maturity` on the left with two terms `Maturity Duration` and `Perpetual Maturity` on the right, as well as the edge between `Putable` and `Perpetual Maturity`. `Exchangeable` that is not ontologically related to any of the FIBO terms associated with the `Financial Contract A` (on the left). These two clusters were created because when each of the edges was assigned to the corresponding cluster, similarity values between the adjacent terms of all the edges in the clusters, were high enough to ensure that cluster density was maximized.

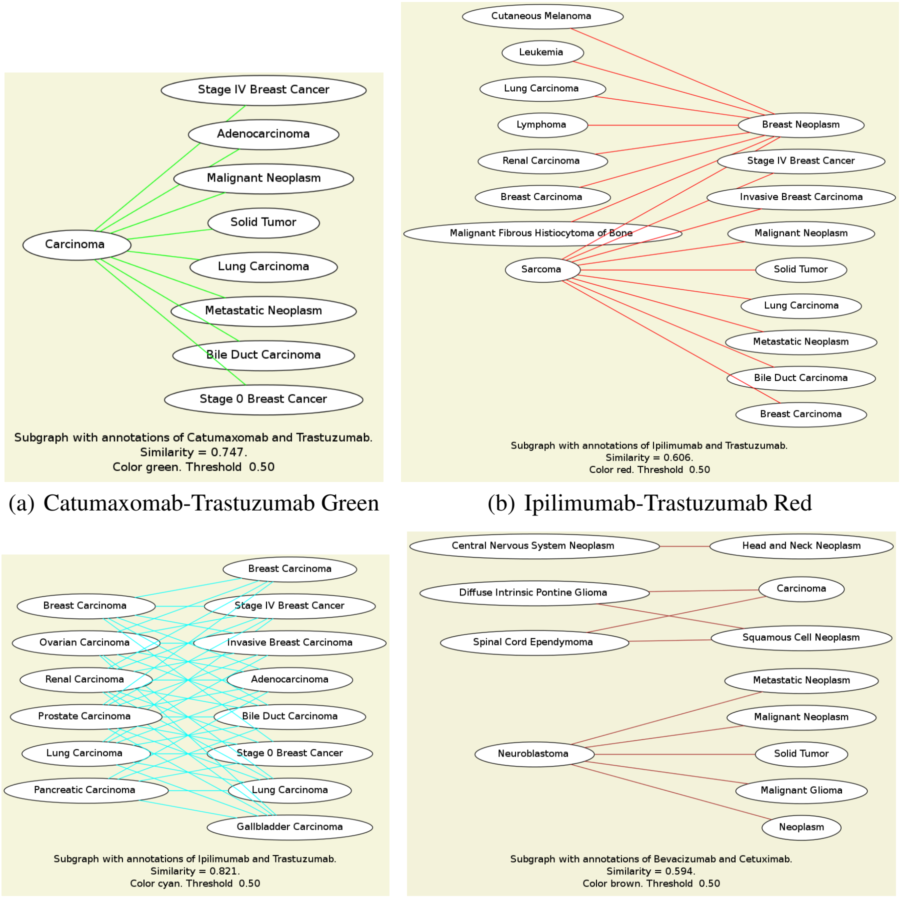
4 Evaluation

The goal of our evaluation is to validate if annotation signatures group together meaningful terms across shared annotations. Additionally, we evaluate the impact of the semantics encoded in the ontologies on the quality of the signature. We study two annotated datasets: *i*) Twelve drugs annotated with NCIt terms that correspond to the diseases associated with these drugs in clinical trials. *ii*) Twenty transporter genes from Arabidopsis thaliana annotated with GO terms. There is no prior *gold standard* solution(s) or ground truth for these two datasets that we can use to evaluate the quality of the annotation signature. Thus, we relied on a team of experts to analyze the annotation signatures. Annotated datasets are included in the supplementary material. All results are available via a Web portal [2].

4.1 Dataset and Evaluators

Drugs: Anti-neoplastic agents and monoclonal antibodies are two popular and independent intervention regimes that have been successfully applied to treat a large range of cancers. There are 12 drugs that fall within their intersection, and scientists are interested in studying the relationships between these drugs and the corresponding diseases. We consider a dataset of the following twelve drugs: Alemtuzumab, Bevacizumab, Brentuximab vedotin, Cetuximab, Catumaxomab, Edrecolomab, Gemtuzumab, Ipilimumab, Ofatumumab, Panitumumab, Rituximab, and Trastuzumab. The protocol to create the dataset is as follows: Each drug was used to retrieve a set of clinical trials in LinkedCT *circa* September 2011 (linkedct.org). Then each disease associated with each trial was linked to its corresponding term in the NCI Thesaurus version 12.05d; annotation was performed by NCIt experts. Our group of evaluators included two experts who develop databases and tools for the NCI Thesaurus and two bioinformatics researchers with expertise on the NCIt and other biomedical ontologies.

Genes: The vacuolar-type H+-ATPase are proton pumps associated with the adenosine triphosphatase (ATP) enzyme. The pump acidifies intracellular compartments and is essential for many processes, including co-transport, guard cell movement, development, and tolerance to environmental stress. Our collaborators in the Sze Lab at the University of Maryland have identified genes encoding subunits of V-ATPase in the Arabidopsis thaliana genome. The pump consists of subunits A through H of the peripheral V1 complex, and subunits a, c, c" and d of the Vo membrane sector. The genes are named AtVHA-n where n represents the code for each subunit. Our dataset included the following twenty genes, AtVHA-A, AtVHA-A1, AtVHA-A2, AtVHA-A3, AtVHA-B1, AtVHA-B2, AtVHA-B3, AtVHA-C, AtVHA-C1, AtVHA-C2, AtVHA-C3, AtVHA-C4, AtVHA-C5, AtVHA-D1, AtVHA-D2, AtVHA-E1, AtVHA-E2, AtVHA-F, AtVHA-c"1 and AtVHA-c"2. We obtained the GO annotations from the TAIR portal[3].



(a) Catumaxomab-Trastuzumab Green

(b) Ipilimumab-Trastuzumab Red

(c) Ipilimumab-Trastuzumab Cyan

(d) Bevacizumab-Cetuximab Brown

Fig. 3. Connectivity Patterns within Each Cluster for $\theta = 0.5$; (a) Catumaxomab-Trastuzumab Green; (b) Ipilimumab-Trastuzumab Red; (c) Ipilimumab-Trastuzumab Cyan; (d) Bevacizumab-Cetuximab Brown.

4.2 Connectivity Patterns within a cluster

The connectivity pattern within each cluster provides insight into the ontological relatedness of the diseases. In Figure 3(a) `Carcinoma` on the left is connected to 8 terms on the right. In Figure 3(b), `Sarcoma` on the left is connected to 9 drugs on the right. Similarly, `Breast Neoplasm` on the right is connected to eight diseases on the left. None of the other drugs has more than one incident edge. In contrast, in Figure 3(c), we see a much more general many-to-many connection pattern between the diseases on the left and right. Finally, Figure 3(d) shows a more complex connectivity pattern where the terms are ontologically related but they are placed within three disconnected graphs. The four terms `Diffuse Intrinsic Pontine Glioma`, `Spinal`

[2] `dynbigraph.appspot.com`

[3] `http://www.arabidopsis.org/,April-May2013`

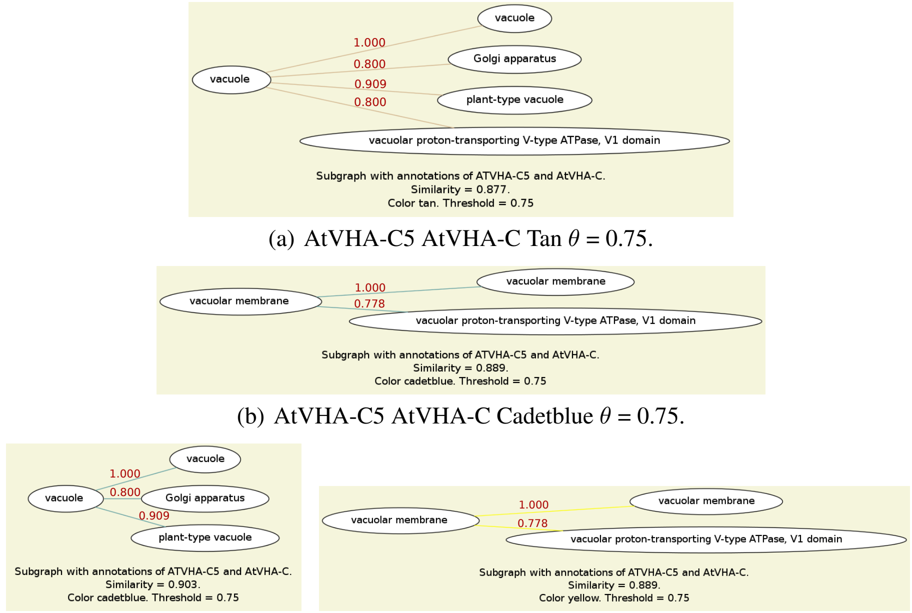
Cord Ependymoma, Carcinoma and Squamous Cell Neoplasm form the most well connected cluster. Comments from the evaluators noted that while groups such as Figure 3(a) that included generic terms such as Carcinoma were valid, they did not convey useful information. In contrast, groups in Figures 3(c) and (d), that had more specific terms and were more densely connected, had the potential to be more meaningful.

4.3 Utilizing Relationship Type Semantics

The goal of this evaluation is to determine the impact of the semantics of the ontology relationships on the annotation signatures. Figure 4 presents an example of exploiting relationship types using the GO ontology. There are five type of relationships captured in the GO ontology: *i)* is_a, *ii)* part_of, *iii)* regulates, *iv)* positively_regulates and *v)* negatively_regulates. Figures 4 (a) and (b) present two components of the gene signature for the genes AtVHA-C5 and AtVHA-C for threshold θ = 0.75. This is a scenario where d_{tax} (ontological relatedness) is computed using paths that consider *all* the GO relationship types. We observe that the term vacuolar proton-transporting V-type ATPase, V1 domain appears in both components of Figures 4(a) and (b). In contrast, Figures 4(c) and (d) present the two components when *only* is_a relationship types are considered. The value for d_{tax} between vacuole and vacuolar proton-transporting V-type ATPase, V1 domain decreases from 0.800 to 0.70. As a result, the term vacuolar proton-transporting V-type ATPase, V1 domain is only present in one component in Figure 4(d). This example illustrates multiple benefits from using ontological knowledge. First, redundancy in patterns is reduced. More important, the modified components represent more precise patterns of relationships between shared annotations and reflect additional semantic knowledge. A summary of this evaluation is described in Section 4.5.

4.4 Utilizing Ontology Structure

Recall that d_{tax}^{str} extended the taxonomic distance metric d_{tax} to consider ontology structure. Figure 5(a) illustrates an example cluster of the annotations for the pair Trastuzumab and Bevacizumab produced by d_{tax}; the threshold θ = 0.50. There are many shortcomings. First, it contains generic CV terms such as Adenocarcinoma and Carcinoma. Further, it is very large and many diverse and unrelated cancers are included. Figure 5(b) shows the result of applying the metric d_{tax}^{str} to exploit ontology structure. The large cluster was partitioned into smaller clusters. Many of the generic CV terms are no longer included and each smaller cluster includes more closely related CV terms. For example, one has a focus on breast cancer related terms, another has a focus on lung cancer, while a third combines terms related to pancreatic, renal and colorectal cancers. This example illustrates benefits from using ontological knowledge to eliminate generic terms from the annotation signatures. Redundancy in patterns is reduced, and the modified annotation signatures are comprised of relationships between more specific terms. Summarized results of the comparison between d_{tax} and d_{tax}^{str} for the dataset of the twelve drugs are presented in next section.


vacuole
1.000
0.800
Golgi apparatus
0.909
plant-type vacuole
0.800
vacuolar proton-transporting V-type ATPase, V1 domain
Subgraph with annotations of ATVHA-C5 and AtVHA-C.
Similarity = 0.877.
Color tan. Threshold = 0.75
vacuolar membrane
1.000
0.778
vacuolar membrane
vacuolar proton-transporting V-type ATPase, V1 domain
Subgraph with annotations of ATVHA-C5 and AtVHA-C.
Similarity = 0.889.
Color cadetblue. Threshold = 0.75
vacuole
1.000
0.800
Golgi apparatus
0.909
plant-type vacuole
Subgraph with annotations of ATVHA-C5 and AtVHA-C.
Similarity = 0.903.
Color cadetblue. Threshold = 0.75
vacuolar membrane
1.000
0.778
vacuolar membrane
vacuolar proton-transporting V-type ATPase, V1 domain
Subgraph with annotations of ATVHA-C5 and AtVHA-C.
Similarity = 0.889.
Color yellow. Threshold = 0.75

(a) AtVHA-C5 AtVHA-C Tan $\theta = 0.75$.

(b) AtVHA-C5 AtVHA-C Cadetblue $\theta = 0.75$.

(c) AtVHA-C5 AtVHA-C Cadetblue $\theta = 0.75$ Only ISA Paths.

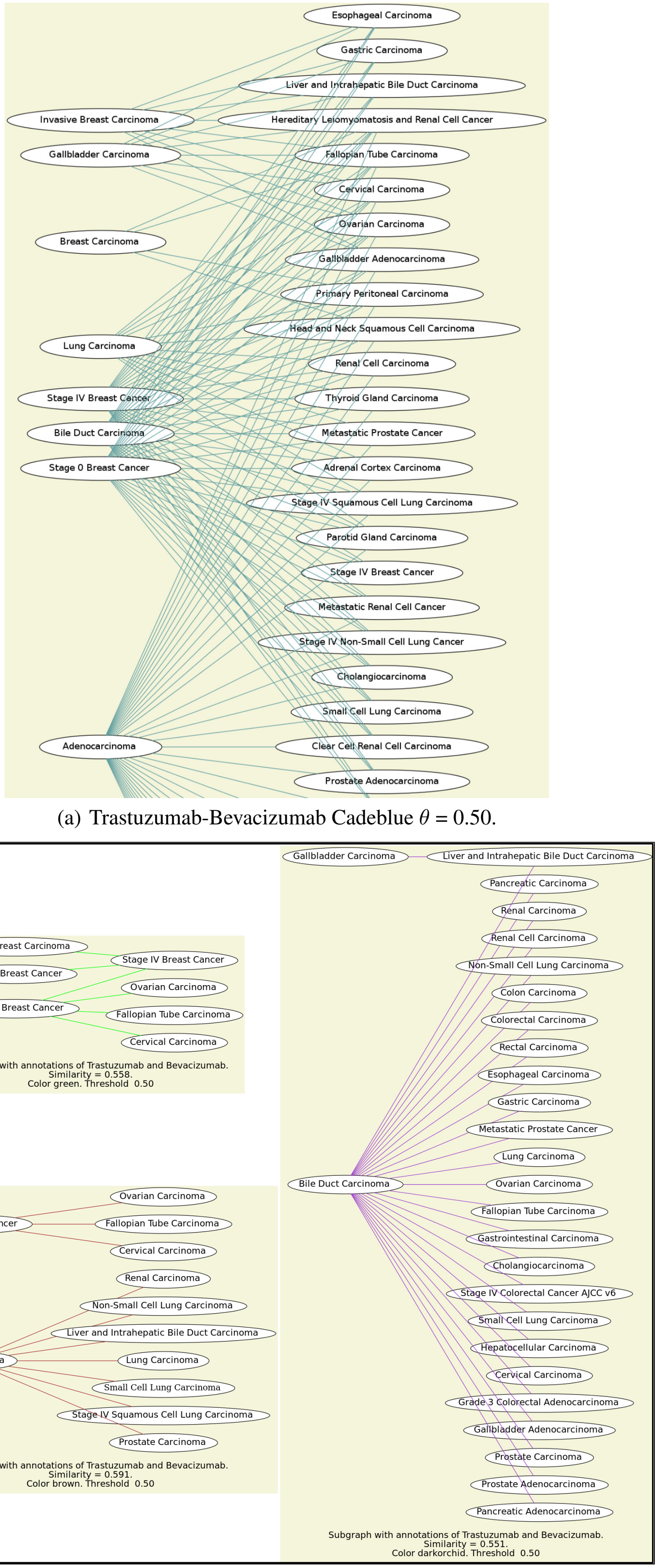
(d) AtVHA-C5 AtVHA-C yellow $\theta = 0.75$ Only ISA Paths.

Fig. 4. Enhancing Discovery Patterns with Semantics for $\theta = 0.75$; (a) and (b) Paths are computed using all five GO relationship types; (c) and (d) Paths are computed using only the `is_a` GO relationship type.

4.5 Summary Statistics

In this section we report on aggregated results of our evaluations. Table 1(a) provides a summary of the gene clustering when d_{tax} (ontological relatedness) is computed using all the GO relationship types and when only `IS_A` relationship types are considered. We compute the annotation signatures for pairwise comparisons of twenty genes; we report on minimum (MIN), maximum (MAX), and average (AVG) number of clusters in these annotation signatures. We consider two different values of threshold $\theta = 0.5$ and 0.75. As the threshold θ increases, the average of the number of clusters decreases. Further, when only `IS_A` relationship types are considered, the values of d_{tax} (ontological relatedness) are affected. The number of paths between two terms decreases, e.g., paths combining `positively_regulates` and `negatively_regulates` are not included in the bipartite graph BG. Additionally, d_{tax} values typically decrease and more edges are deleted from BG. Thus, as observed in Table 1(a), the average of number of clusters decreases. As noted earlier, these refinements also create more closely related clusters.

Table 1(b) provides summary statistics for annotation signatures computed using d_{tax} and d_{tax}^{str} over the pairwise comparisons of twelve diseases. We report on minimum (MIN), maximum (MAX), and average (AVG) number of clusters in these signatures; two values of threshold $\theta = 0.5$ and 0.75 are considered. Because d_{tax}^{str} penalizes generic CV terms, many edges are eliminated from BG. Further, the values of d_{tax}^{str} are lower than the values of d_{tax}. Thus, many of the large clusters computed with d_{tax} are partitioned into smaller clusters by d_{tax}^{str}. At the same time, the number of clusters de-


Esophageal Carcinoma
Gastric Carcinoma
Liver and Intrahepatic Bile Duct Carcinoma
Invasive Breast Carcinoma
Hereditary Leiomyomatosis and Renal Cell Cancer
Gallbladder Carcinoma
Fallopian Tube Carcinoma
Cervical Carcinoma
Breast Carcinoma
Ovarian Carcinoma
Gallbladder Adenocarcinoma
Primary Peritoneal Carcinoma
Lung Carcinoma
Head and Neck Squamous Cell Carcinoma
Renal Cell Carcinoma
Stage IV Breast Cancer
Thyroid Gland Carcinoma
Bile Duct Carcinoma
Metastatic Prostate Cancer
Stage 0 Breast Cancer
Adrenal Cortex Carcinoma
Stage IV Squamous Cell Lung Carcinoma
Parotid Gland Carcinoma
Stage IV Breast Cancer
Metastatic Renal Cell Cancer
Stage IV Non-Small Cell Lung Cancer
Cholangiocarcinoma
Small Cell Lung Carcinoma
Adenocarcinoma
Clear Cell Renal Cell Carcinoma
Prostate Adenocarcinoma
Gallbladder Carcinoma
Liver and Intrahepatic Bile Duct Carcinoma
Pancreatic Carcinoma
Renal Carcinoma
Renal Cell Carcinoma
Invasive Breast Carcinoma
Stage IV Breast Cancer
Non-Small Cell Lung Carcinoma
Stage 0 Breast Cancer
Ovarian Carcinoma
Colon Carcinoma
Stage IV Breast Cancer
Fallopian Tube Carcinoma
Colorectal Carcinoma
Cervical Carcinoma
Rectal Carcinoma
Subgraph with annotations of Trastuzumab and Bevacizumab.
Similarity = 0.558.
Color green. Threshold 0.50
Esophageal Carcinoma
Gastric Carcinoma
Metastatic Prostate Cancer
Lung Carcinoma
Bile Duct Carcinoma
Ovarian Carcinoma
Ovarian Carcinoma
Fallopian Tube Carcinoma
Stage 0 Breast Cancer
Fallopian Tube Carcinoma
Gastrointestinal Carcinoma
Cervical Carcinoma
Cholangiocarcinoma
Renal Carcinoma
Stage IV Colorectal Cancer AJCC v6
Non-Small Cell Lung Carcinoma
Small Cell Lung Carcinoma
Liver and Intrahepatic Bile Duct Carcinoma
Hepatocellular Carcinoma
Lung Carcinoma
Lung Carcinoma
Cervical Carcinoma
Small Cell Lung Carcinoma
Grade 3 Colorectal Adenocarcinoma
Stage IV Squamous Cell Lung Carcinoma
Gallbladder Adenocarcinoma
Prostate Carcinoma
Prostate Carcinoma
Subgraph with annotations of Trastuzumab and Bevacizumab.
Similarity = 0.591.
Color brown. Threshold 0.50
Prostate Adenocarcinoma
Pancreatic Adenocarcinoma
Subgraph with annotations of Trastuzumab and Bevacizumab.
Similarity = 0.551.
Color darkorchid. Threshold 0.50

(a) Trastuzumab-Bevacizumab Cadeblue $\theta = 0.50$.

(b) Trastuzumab-Bevacizumab $\theta = 0.50$ using d_{tax}^{str}

Fig. 5. Enhancing Signatures with Semantics for $\theta = 0.50$. (a) Signature of Trastuzumab-Bevacizumab $\theta = 0.50$; Similarity d_{tax}-Figure has been truncated for readability.; (b) Three clusters of Trastuzumab-Bevacizumab $\theta = 0.50$ when generic terms are penalized using d_{tax}^{str}.

Table 1. Cluster Distribution over the set of annotated drugs and genes. (a) Aggregate Cluster Distribution, all GO relations versus only is_a for AtVHA-n genes ; (b) Aggregate Cluster Distribution, effect of d_{tax} versus d_{tax}^{str} to eliminate relationships with generic terms. MIN, MAX, AVG correspond to the minimum, maximal and average numbers of clusters identified by *AnnSigClustering*, respectively.

(a) Aggregate Cluster Distribution AtVHA-n genes

	MIN	MAX	AVG
0.50	4.00	28.00	**11.96**
0.50 Only is_a	4.00	30.00	**11.75**
0.75	2.00	34.00	**10.99**
0.75 Only is_a	2.00	34.00	**10.45**

(b) Aggregate Cluster Distribution Diseases

	MIN	MAX	AVG
0.50 d_{tax}	1.00	46.00	**6.26**
0.50 d_{tax}^{str}	0.00	28.00	**3.38**
0.75 d_{tax}	0.00	37.00	**4.92**
0.75 d_{tax}^{str}	0.00	9.00	**0.80**

creases. All of these refinements lead to a smaller number of more closely related and meaningful clusters within the annotation signature.

5 Related Work

Graph data mining [5] covers a broad range of methods dealing with the identification of (sub)structures and patterns in graphs. Popular techniques include graph clustering, community detection and cliques. The problem of a 1-to-1 weighted maximal bipartite match has been applied to many problems, e.g, semantic equivalence between two sentences and measuring similarity between shapes for object recognition[1, 3, 11]. These approaches clearly show the benefits of solving a matching problem to identify similarity between terms or concepts. Our research advances prior research in that we consider the relatedness of sets of annotations and identify a many-to-many bipartite match.

A key element in finding patterns is identifying related concepts; we consider ontological relatedness. Similarity metrics (or distance metrics) can be used to measure relatedness; we briefly describe some of the existing metrics. The first class of metrics are string-similarity[4]; they compare the names or labels of the concepts using string comparison functions based on edit distances or other functions that compare strings. This includes the Levenstein distance and Jaro-Winkler [6]. The next are path-similarity metrics that compute relatedness based on the paths that connect the concepts within some appropriate graph. Nodes in the paths can be all of the same abstract types (e.g., PathSim [13]) or they can be heterogeneous (HeteSim [12]). Furthermore, topological-similarity metrics extend the concept of path-similarity and they look at relationships within an ontology or taxonomy that is itself designed to capture relationships (e.g., nan [7], d_{ps} [9] and d_{tax}[2]). We propose an approach that exploits ontological knowledge of scientific annotations to decide relatedness between entities of annotated datasets.

6 Conclusions and Future Work

We have defined the *Annotation Signature* Partitioning problem and the *AnnSigClustering* algorithm to develop the components of a signature based on shared annotations and ontological relatedness. We empirically studied the effectiveness of *AnnSigClustering* to identify potential meaningful signatures of annotated concepts. Further, we have analyzed the effects of considering knowledge encoded in the ontologies used to annotate Linked Data. Our results suggest that the grouping capability of our approach is enhanced whenever the type of relationships are considered as well as when relationships with generic terms are eliminated. Our initial project objective was to validate correctness and utility of components in a signature. Nevertheless, in the future, we will also address performance and scalability. Additionally, we plan to conduct a deeper evaluation study with our collaborators, and thus determine the potential discovery capability of the approach. Finally, we plan to apply our techniques to other domains, e.g., to identify signatures of electoral voters, relationships between financial contracts, and patterns of viral diseases.

References

1. S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(4):509–522, 2002.
2. J. Benik, C. Chang, L. Raschid, M. E. Vidal, G. Palma, and A. Thor. Finding cross genome patterns in annotation graphs. In *Proceedings of Data Integration in the Life Sciences (DILS)*, 2012.
3. S. Bhagwani, S. Satapathy, and H. Karnick. Semantic textual similarity using maximal weighted bipartite graph matching. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 579–585. Association for Computational Linguistics, 2012.
4. W. W. Cohen, P. D. Ravikumar, and S. E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *IIWeb*, pages 73–78, 2003.
5. D. J. Cook and L. B. Holder. *Mining graph data*. Wiley-Blackwell, 2007.
6. M. A. Jaro. Probabilistic linkage of large public health data files. *Statistics in Medicine*, pages 491–498, 1995.
7. B. McInnes, T. Pedersen, and S. Pakhomov. Umls-interface and umls-similarity : Open source software for measuring paths and semantic similarity. *Proceedings of the AMIA Symposium*, pages 431–435, 2009.
8. G. Palma, M.-E. Vidal, E. Haag, L. Raschid, and A. Thor. Measuring relatedness between scientific entities in annotation datasets. Technical report, University of Maryland. UMIACS Technical Report, 2013.
9. V. Pekar and S. Staab. Taxonomy learning - factoring the structure of a taxonomy into a semantic classification decision. In *COLING*, 2002.
10. L. Raschid, G. Palma, M.-E. Vidal, and A. Thor. Exploration using signatures in annotation graph datasets. Technical report, University of Maryland. UMIACS Technical Report, 2013.
11. Y. Shavitt, E. Weinsberg, and U. Weinsberg. Estimating peer similarity using distance of shared files. In *International workshop on peer-to-peer systems (IPTPS)*, volume 104, 2010.
12. C. Shi, X. Kong, P. S. Yu, S. Xie, and B. Wu. Relevance search in heterogeneous networks. In *EDBT*, pages 180–191, 2012.
13. Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB*, 4(11):992–1003, 2011.