Drug-Target Interaction Prediction Using Semantic Similarity and Edge Partitioning

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

[1] Universidad Simón Bolívar, Venezuela
[2] University of Maryland, USA
gpalma,mvidal@ldc.usb.ve, louiqa@umiacs.umd.edu

Abstract. The ability to integrate a wealth of human-curated knowledge from scientific datasets and ontologies can benefit drug-target interaction prediction. The hypothesis is that similar drugs interact with the same targets, and similar targets interact with the same drugs. The similarities between drugs reflect a chemical semantic space, while similarities between targets reflect a genomic semantic space. In this paper, we present a novel method that combines a data mining framework for link prediction, semantic knowledge (similarities) from ontologies or semantic spaces, and an algorithmic approach to partition the edges of a heterogeneous graph that includes drug-target interaction edges, and drug-drug and target-target similarity edges. Our semantics based edge partitioning approach, semEP, has the advantages of edge based community detection which allows a node to participate in more than one cluster or community. The semEP problem is to create a minimal partitioning of the edges such that the cluster density of each subset of edges is maximal. We use semantic knowledge (similarities) to specify edge constraints, i.e., specific drug-target interaction edges that should not participate in the same cluster. Using a well-known dataset of drug-target interactions, we demonstrate the benefits of using semEP predictions to improve the performance of a range of state-of-the-art machine learning based prediction methods. Validation of the novel best predicted interactions of semEP against the STITCH interaction resource reflect both accurate and diverse predictions.

Keywords: Drug-target interaction prediction, vertex coloring graph, community detection, graph partitioning.

1 Introduction

Linked Open Data has important applications across the biomedical enterprise where there is a nexus created by the availability of publicly accessible richly curated scientific collections and the extensive use of ontologies and thesauri. This ability to seamlessly integrate a wealth of human-curated knowledge can benefit many applications including drug-target interaction prediction and drug-drug similarity ranking. Consider that drugs are molecules that participate in some biomolecular reaction associated with a disease related genomic target (protein).

P. Mika et al. (Eds.) ISWC 2014, Part I, LNCS 8796, pp. 131–146, 2014.

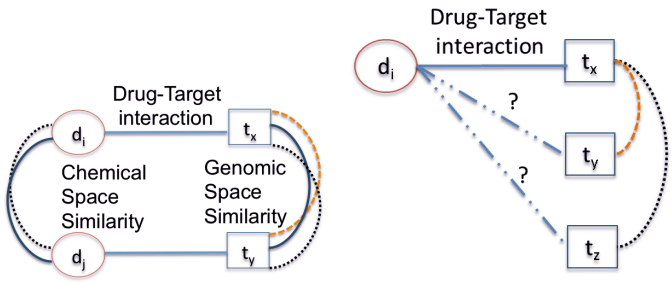
The ability to predict new drug-target interactions can have applications in drug re-purposing to find new targets for drugs. A related application is drug-drug side effect prediction, e.g., to construct the SIDER [16] side effect resource, or to populate ADEpedia [14], a knowledge base of adverse drug events (ADEs) for drug safety surveillance.

Beyond drug-target interaction prediction, drug-drug similarity rankings are an important component of the comprehensive evidence that is used to make clinical or policy recommendations. Consider the following example relevant to a group of monoclonal antibodies (mab) drugs: On November 3, 2010, The New York Times reported that Genentech began offering secret rebates to ophthalmologists in an apparent inducement to get them to prescribe `Ranibizumab` rather than the less expensive `Bevacizumab`. Several studies have shown no superior effect of `Ranibizumab` over `Bevacizumab` for the treatment of macular degeneration, an aging-related eye condition. Subsequently, on April 8, 2014, the Washington Post highlighted the results from analyzing a BIGDATA Medicare collection revealing that one of the largest Medicare billers, an ophthalmologist in West Palm Beach, Fla., earned $20 million in 2012; a large fraction of his earnings came from injecting patients with Lucentis (`Ranibizumab`) instead of Avastin (`Bevacizumab`).

Figures 1(a) and (b) show a schematic overview of drug-target interaction networks; drugs are circles and targets are squares. For interaction prediction, or to determine functionally equivalent drugs, one must exploit drug-drug and target-target similarities; the hypothesis is that similar drugs interact with the same targets, and similar targets interact with the same drugs. The similarities between drugs reflect a chemical semantic space, while similarities between targets reflect a genomic semantic space [8,21]. Within these semantic spaces, pairs of drugs or pairs of targets may have multiple semantics-based similarity scores. For example, drugs can have similarities based on chemical structure or shared side-effects, while gene targets may share sequence based or protein-protein interaction based similarity [21]; this is illustrated by the multiple edge types.

For the purpose of this paper we focus on drug-target interaction edges. However, our method can be applied to a variety of Linked Data collections and ontologies as will be seen in the next section.

There are many approaches for link prediction or similarity ranking, e.g., drug-target interaction networks [29] or citation graphs [20]. The importance of structured knowledge and collective classification for drug-target prediction was discussed in [11]. Structured knowledge include *triads*; in Figure 1(b), the interaction edge (d_i, t_x), the similarity between targets t_x and t_y, and the potential interaction edge (d_i, t_y) form a triad. Similarly, in Figure 1(a), the two interaction edges (d_i, t_x) and (d_j, t_y), the corresponding drug-drug similarity between d_i and d_j, and the target-target similarity between t_x and t_y, form a *tetrad*. Further, collective classification would support the simultaneous reasoning over the edges (d_i, t_x), (d_i, t_y), (d_i, t_z), etc., in Figure 1(b), and their corresponding similarities.



(a) Drug-target interaction

(b) Drug-target interaction prediction through collective classification

Fig. 1. (a) Drug-Target Interaction Network. Drugs are circles and diseases are rectangles. (b) An Example of Collective Classification of Potential Interactions.

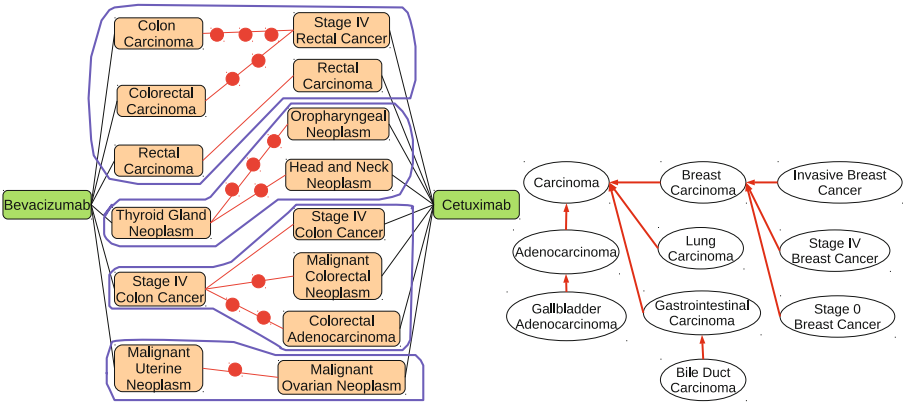
We present semEP, an unsupervised semantics based edge partitioning method; semEP combines a data mining framework for link prediction, semantic knowledge (similarities) from ontologies or semantic spaces, and an algorithmic approach to partition the edges of a heterogeneous graph. For this paper, we consider a graph that includes drug-target interaction edges, and drug-drug and target-target similarities. The semEP problem is to create a minimal partitioning of the edges such that the cluster density of each subset of edges is maximal. An advantage of semEP edge clustering is that it allows a node to participate in more than one cluster or community; this is a natural match with the semantics of drugs that have multiple functions, and thus interact with different targets. We do not limit semEP to triad or tetrad clusters and we consider clusters of varying shape and size. Further, semEP can use semantic knowledge on similarities to specify edge constraints, i.e., specific pairs of drug-target interaction edges that should not occur in the same cluster.

Using a well-known dataset of drug-target interactions [3,8], we demonstrate the benefits of using semEP predictions to improve the performance of all the state-of-the-art machine learning based prediction methods [8]. We also validate the best novel predictions of all the methods (where the interactions are not in the test dataset) against the STITCH drug-target interaction resource [17]. The good performance of semEP reflects its ability to exploit structured semantic knowledge to make accurate and diverse predictions.

This paper is organized as follows: Section 2 provides a motivating example of Linked Data and ontological knowledge and Section 3 describes the semEP edge partitioning problem. Section 4 summarizes related research. Experimental results are reported in Section 5 and Section 6 concludes.

2 Semantics of Annotations and Ontological Relatedness

In this paper, we focus on a specific link prediction use case – the problem of predicting drug-target or drug-disease interaction edges. However, as motivation, we



(a) Bevacizumab and Cetuximab (b) ISA Hierarchy

Fig. 2. (a) Drugs `Bevacizumab` and `Cetuximab` (green rectangles), Disease Annotations (orange rectangles) and NCI Thesaurus Terms (red ovals). Four communities are highlighted in blue. (b) Fragment of an ISA Hierarchy in the NCIt. The red lines indicate ISA relationships

consider the more general problem of drug-drug similarity ranking. `Bevacizumab` and `Cetuximab` are exemplars of monoclonal antibodies that are anti-neoplastic agents used in cancer treatment. We consider the similarity of `Bevacizumab` and `Cetuximab` using their neighborhood graph of shared annotations of disease terms. Figure 2(a) represents (partial) disease annotations associated with each drug; the disease terms are mapped to terms in the NCI Thesaurus (NCIt). Each path between a pair of diseases, e.g., `Colon Carcinoma` and `Stage IV Rectal Cancer`, is identified with red circles representing intermediate NCIt terms.

A simple shared annotation pattern would include the identical term, e.g., `Rectal Carcinoma`. Ontological relatedness indicates that non-identical terms such as `Colon Carcinoma` and `Stage IV Rectal Cancer` are also related to each other. Combining shared annotation and ontological relatedness, we may determine that (`Colon Carcinoma`, `Colorectal Carcinoma`, `Rectal Carcinoma`, `Stage IV Rectal Cancer`), together, form a shared community of ontologically related disease terms. Further, (`Malignant Colorectal Neoplasm`, `Stage IV Colon Cancer`, `Colorectal Adenocarcinoma`) appear to form a (possibly overlapping) community, while (`Thyroid Gland Neoplasm`, `Oropharyngeal Neoplasm`, `Head and Neck Neoplasm`) and (`Malignant Uterine Neoplasm`, `Malignant Ovarian Neoplasm`) form additional distinct communities.

Figure 2(b) shows a fragment of the NCIt ISA hierarchy. `Carcinoma` can be specialized to various organs, e.g., `Lung Carcinoma`; to specific types of disease, e.g., `Adenocarcinoma`; to disease stages, e.g., `Stage IV Breast Cancer`; or to combinations, e.g., `Stage III Colorectal Adenocarcinoma` (not shown).

3 Semantics Based Edge Partitioning Problem (semEP)

3.1 From Structured Knowledge to Link Prediction for Drug Target Interaction Networks

Let $D = \{d_1, d_2, \cdots, d_m\}$ be a drug set and let $T = \{t_1, t_2, \cdots, t_n\}$ be a target set. Let S_d be a drug similarity matrix where the (i,j)-th element denoted $s_d(d_i,d_j)$ is a similarity score (potentially there are multiple scores) between drugs d_i and d_j. Let S_t be a target similarity matrix where the (i,j)-th element denoted $s_t(t_i,t_j)$ is a similarity score between targets t_i and t_j.

Let Y be a binary matrix of `true labels` of drug-target interactions. $Y_{i,j} = 1$ if drug d_i interacts with target t_j; $Y_{i,j} = 0$ otherwise.

The objective is to produce a score matrix F where the (i,j)-th element denoted $F_{i,j}$ is the score or probability that the drug d_i interacts with target t_j.

The hypothesis underlying most solutions is that similar drugs interact with the same targets, and similar targets interact with the same drugs. While this appears to be straightforward, there are many challenges. First, there is no single approach to determine the similarities between drugs or between targets; indeed there are many similarities based on different semantics [21]. Referring to the Linked Data example in the previous section, the NCIt can be used to define a semantic space for drugs and for targets (diseases), while taxonomic metrics can be used to determine similarity scores using the NCIt structure.

A bigger challenge is that the bipartite drug-target interaction network expresses multi-relational or graph structured knowledge. A drug d_i may be complex in its functional behavior and may have multiple targets. Hence, a drug d_j that is similar to d_i based on chemical structure but not on side-effect similarity, may only share some of the targets of d_i.

A state-of-the-art solution for the drug-target interaction prediction problem is presented in [11] where they propose a drug-target prediction framework based on Probabilistic Soft Logic (PSL) [5]. The PSL based solution reasons collectively over interactions using structured rules that capture the multi-relational nature of the network, e.g., the triads and tetrads of Figure 1(a) and (b). Finding the most promising candidates for triad and tetrad based learning is an expensive problem that requires significant tuning [11] and the PSL based program was thus limited to triads and tetrads.

In contrast, semEP can make predictions using larger complex clusters. We can also exploit the drug-drug or target-target similarities to control the shape of the clusters. Figure 3 illustrates a drug-target interaction network on the left, with three drugs DB01100 (Pimozide), DB01244 (Bepridil), and DB00836 (Loperamide), and eight targets. Drugs DB01100 (Pimozide) and DB01244 (Bepridil) share 6 interactions. A node partition may place these two drugs into one community and place DB00836 (Loperamide) in a second community.

Since semEP is an edge partitioning, it can instead consider more complex communities with an overlap of nodes. The broken (dotted) edges in Figure 3 (left) connect each target to its *least similar* target. A visual inspection of these edges reveals that a split of the targets, with 782, 784, and 785 appearing in one

community, while 774, 776, 778, 779, and 8912 are placed in a second community, has the property that no target is placed in a community together with its least similar target. To capture such properties, semEP will consider *edge constraints* as follows: Consider the scenario where targets 784 and 779 have a mutual least-similar-target relationship. Then semEP will guarantee an edge constraint for this pair, i.e., no edge incident to 784 will be placed in the same cluster together with an edge incident to target 779.

Thus, semEP combines the benefit of edge partitioning that allows node overlap in the clusters, and the edge constraints that prohibit (some) pairs of edges to be placed in the same cluster. This accommodates both the semantics of nodes with complex function (node overlap in multiple clusters), and the semantics of separating the edges incident to the least similar pairs of nodes (edge constraints).

Figure 3 (b) shows the two edge communities created by semEP on the right. Community 1 includes drugs DB01100 (Pimozide), DB01244 (Bepridil), and five targets. Community 2 includes those two drugs as well as DB00836 (Loperamide), and has three targets. We note that these communities, with 6 and 7 nodes, respectively, are more complex compared to triads. The predicted drug-target interaction(s) based on these two communities are shown as broken edge(s) in the edge communities on the right. We note that through the use of structured knowledge (edge constraints), edge partitioning and node overlap, semEP predicts an interaction between DB00836 (Loperamide) and target 784.

We summarize the objectives of semEP as follows:

- An edge partitioning that allows the overlap of nodes in multiple clusters; this matches the semantics of complex function associated with nodes.
- Create clusters with high cluster density to improve prediction accuracy.
- Exploit semantic knowledge about the least similar pairs of nodes to identify edge constraints; they will be used to prohibit the placement of incident edges, of the least similar nodes, in the same cluster.
- Balance these competing objectives by creating a minimal number of clusters, each of which has maximal cluster density.

3.2 Problem Definition: semEP

The semantics based edge partition problem (semEP) is the minimal partitioning P of the edges of a graph BG such that the aggregate cluster density over all subsets of edges (clusters) $p \in P$ is maximized. We note that a partitioning P of edges may result in the overlap of nodes across different clusters.

Definition 1 (Cluster (Similarity) Density). *Consider a labeled bipartite graph $BG=(D \cup T, WE)$. Nodes in D represent a set of drugs and nodes in T represent a set of targets. WE is a set of drug-target interactions, i.e., there is an edge $e = (d,t) \in WE$ iff $Y_{d,t} = 1$. Let p be a subset of interactions of WE. Let $D_p \subseteq D$ be the drug set incident on the edges $(d,t) \in p$, and let $T_p \subseteq T$ be the target set incident on the edges $(d,t) \in p$. Let $s_d(i,j)$ represent*

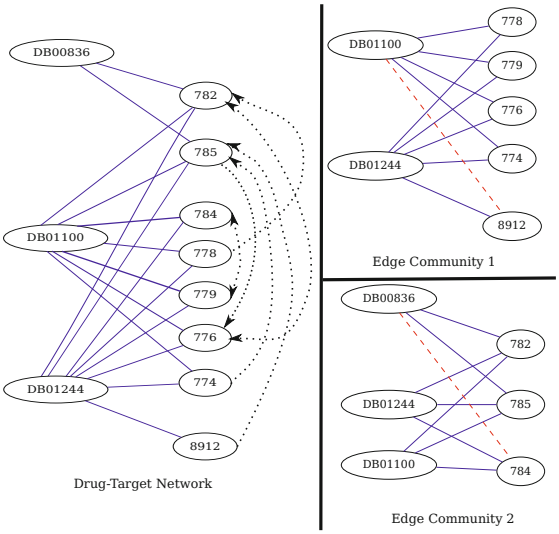


Fig. 3. Using Structured Knowledge for semEP. (a) A drug-target interaction network of three drugs DB01100 (Pimozide), DB01244 (Bepridil) and DB00836 (Loperamide), and eight targets. (b) Two edge communities created by semEP. Community 1 includes drugs DB01100 (Pimozide) and DB01244 (Bepridil) and five targets. Community 2 includes all three drugs and has three targets.

the similarity score between a pair of drugs i *and* $j \in D_p$. *Let* $s_t(i,j)$ *represent the similarity score between a pair of targets* i *and* $j \in D_t$. *Under the condition that* $|Dp| > 0 \wedge |Tp| > 0$, *the cluster (similarity) density of* p $cDensity(p) = \frac{1+\frac{2*\sum_{i,j \in Dp}[i \neq j] s_d(i,j)}{|Dp|(|Dp|-1)}+\frac{2*\sum_{i,j \in Tp}[i \neq j] s_t(i,j)}{|Tp|(|Tp|-1)}}{3}$. *If* $|Dp| = 0$, *or if* $|Tp| = 0$, *then we replace the respective fraction by the value* 0.

To explain, the three terms in the numerator correspond to (1) the average score of the interaction edges in p, (2) the average drug-drug similarity score between all pairs of drugs in p, and (3) the average target-target similarity score between all pairs of targets in p, respectively. We note that the score for interactions is given by $Y_{d,t}$ and is an unweighted score of 1.0 for this special case of drug-target interactions. The cluster (similarity) cDensity penalizes singleton clusters or clusters with a singleton drug or target node.

Definition 2 (The Semantics Based Edge Partition Problem (semEP)). *Given a labeled bipartite graph* $BG=(D \cup T, WE)$ *described as before,* semEP *identifies a (minimal) partition* P *of* WE *such that the aggregate cluster density over all subsets* $p \in P$ $semEP(P) = \frac{\sum_{p \in P}(cDensity(p))}{|P|}$ *is maximal.*

Recall that a solution to semEP corresponds to a partition of the edges where the number of clusters is minimized while the overall cDensity is maximized. We illustrate the impact of these two objectives on drug-target interaction prediction accuracy using the two edge partitions `A` and `B` in Figure 4. Consider the

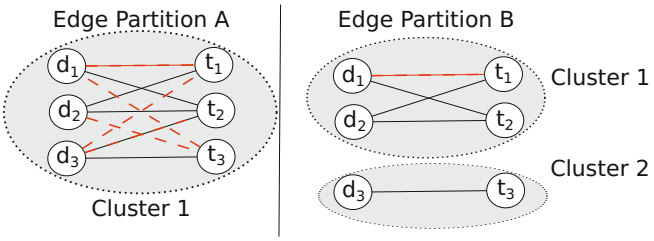


Fig. 4. Two partitions with the same cDensity and red broken predicted edges

following drug-drug and target-target similarity scores: $s_d(d_1, d_3) = s_d(d_2, d_3) = s_t(t_1, t_3) = s_t(t_2, t_3) = 0.1$, and $s_d(d_1, d_2) = s_t(t_1, t_2) = 0.4$. Positive interaction edges are black solid edges while predicted edges are red broken edges. Both partitions have the same cDensity of 0.47. However, partition `A` includes four prediction edges while `B` only includes one prediction edge. Assuming that these are all true positive predictions, then partition `A`, which satisfies the two semEP objectives of maximum aggregate cDensity and minimal number of clusters, has the same precision and greater recall, compared to partition `B`.

Definition 3 (Edge Constraint). *Given nodes i and j, let Inc(i) and Inc(j) correspond to the sets of incident edges to i and j, respectively. Given a real number θ_d or θ_t in the range $[0 : 1]$ and a similarity score $s_d(i, j) < \theta_d$ or $s_t(i, j) < \theta_t$, then there exists an edge constraint EdgeConstraint(i, j, Inc(i), Inc(j), θ).*

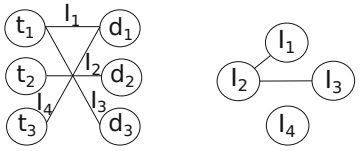
Property 1 (Edge Constraint). *Let P be a solution to the semEP. For a given edge constraint EdgeConstraint(i, j, Inc(i), Inc(j), θ) to hold, there can be no cluster p in P such that $e_i \in Inc(i)$ and $e_j \in Inc(j)$ occur in p.*

We map semEP to the Vertex Coloring Graph (VCG) problem. The Vertex Coloring Graph problem assigns a color to every vertex in a graph such that adjacent vertices are colored with different colors and the number of colors is minimized. Each cluster (component) p in the partition P produced by semEP corresponds to a color in the VCG problem. This will ensure that a minimal number of colors will guarantee a minimal partitioning P.

Definition 4 (Mapping of the Vertex Coloring Problem to the Semantics Based Edge Partition Problem). *Consider a labeled bipartite graph BG=($D \cup T$, WE) and a vertex coloring graph G=(V, F). For each edge or interaction l in WE there is a node v_i in V. Further, there is an edge $l = (v_i, v_j)$ in F, iff there are nodes i and j such that $v_i \in Inc(i)$, $v_j \in Inc(j)$, and EdgeConstraint(i, j, Inc(i),Inc(j), θ)[1] holds. Let P be the (minimal) partition of WE to maximize semEP(P). Let M be a mapping from V to SC, where SC is a set of colors, two vertices from G share the same color if they are in the same partition component p of P and the value cDensity(p) is maximized. The Vertex*

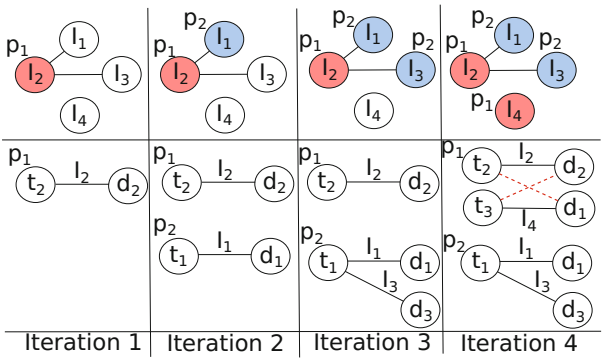
[1] There are thresholds θ_d and θ_t for drugs and targets, respectively.

Coloring Problem for BG is to identify M such that the number of colors used in the coloring of the graph G, namely nc(G), is minimized. Given the set UsedColors of colors in SC that are used in the coloring of the graph, the number of colors corresponds to $nc(G)=\sum_{cl \in UsedColors}(1-cDensity(cl))$*, where cDensity(cl) represents the density of the labels of edges from component p in P, from BG, that are colored with the color cl.*



(a) Drug-Target Interaction Network

(b) Vertex Coloring Graph (VCG)



(c) Iterations of the semEP Algorithm. Colors in VCG correspond to clusters in the interaction network.

Fig. 5. Example Iterations of semEP

Example Iterations of semEP: Consider the drug-target interaction network of Figure 5(a) with four interactions and the following similar scores: $s_t(t_1, t_2) = 0.1$, $s_t(t_1, t_3) = 0.9$, $s_t(t_2, t_3) = 0.8$, $s_d(d_1, d_2) = 0.75$, $s_d(d_1, d_3) = 0.8$, and $s_d(d_2, d_3) = 0.75$. Consider thresholds $\theta_d = \theta_t = 0.6$ below which pairs of drug or pairs of targets are used to specify edge constraints. Figure 5(b) is the Vertex Coloring Graph (VCG) for the interaction network of Figure 5(a). For example the edge (I_1, I_2) is in VCG because the similarity score $s_t(t_1, t_2)$ of targets t_1 and t_2 are below the threshold $\theta_t = 0.6$. Figure 5(c) shows the iterations of *semEP*; in each iteration, the figure on the top assigns a color to a node of the VCG while the figure at the bottom places an edge in a cluster. In the first iteration, semEP chooses vertex I_2 of the VCG since it has the greatest degree, and assigns color

p_1. Simultaneously, the interaction I is placed in the cluster p_1. In the second iteration, vertices I_1 and I_3 have the greatest degree; semEP breaks the tie in favor of I_1. Vertex I_1 is assigned the color p_2 and this creates a new cluster p_2 with interaction I_1. In the third iteration, the vertex I_3 is assigned the feasible color p_2; this adds the interaction T_3 to cluster p_2. In the last iteration, vertex I_4 can be colored with p_1 or p_2; semEP chooses p_1 and interaction I_4 is placed in cluster p_1. The cluster p_1 has a cDensity $1 = (1.0 + 0.8 + 0.75)/3 = 0.85$, and cluster p_2 has cDensity $= (1.0 + 0.0 + 0.8)/3 = 0.6$; thus, the aggregate cDdensity is $(0.85 + 0.6)/2 = 0.73$. If I_4 had instead been placed in p_2, the aggregate cDensity would have been lower and $= (0.33 + 0.9)/2 = 0.62$. Figure 5(c) shows the two predicted edges (broken edges) in the fourth iteration.

An Efficient Implementation of semEP: VCG is NP-hard [15], and many approximate algorithms have been proposed to solve this problem [23]. semEP extends the well-known approximate algorithm DSATUR [4] to solve VCG to obtain the edge partitions. DSATUR is a greedy iterative algorithm that colors each vertex of the graph once by following a heuristic to choose the colors. Given a graph *G=(V,E)*, DSATUR orders vertices in V dynamically based on the number of different colors assigned to the adjacent vertices of each vertex in V, i.e., the vertices are chosen based on the degree of saturation on the partial coloring of the graph built so far. Only colored adjacent nodes are considered. Intuitively, selecting a vertex with the maximum degree of saturation allows one to first color the vertex (vertices) with more restrictions; this is one for which there is a smaller set of colors. Ties are broken based on the vertex degree of the adjacent nodes. As a result of casting the semEP problem to VCG, semEP iteratively adds an edge or interaction to a cluster following the DSATUR heuristic to create clusters that maximize the cluster density. semEP 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 semEP 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 cDensity. Time complexity of DSATUR is $O(|V|^3)$, thus semEP is $O(|WE|^3)$.

4 Related Work

We briefly compare with research in graph data mining, link prediction, clustering, community detection and ranking. Graph data mining [7] covers a broad range of methods dealing with the identification of (sub)structures and patterns in graphs; state-of-the-art approaches include spectral graph clustering [26], RankClus [24], and GNetMine [13]. Spectral graph clustering relies on an unnormalized Laplacian graph representation of a homogeneous network to cluster the graph based on information encoded in its eigenvectors [26]. RankClus [24] and GNetMine [13] interleave link analysis-based ranking with clustering to

place highly ranked entities in highly ranked clusters. These approaches focus on the use of graph properties to partition the graph.

The problem of dealing with multiple types of similarity scores has been modeled as follows: Perform *simultaneous clustering* with multiple heterogeneous networks over an identical set of nodes; the complexity has been shown to be as hard as the *k densest subgraphs* problem [18]. JointCluster [19] is a simultaneous clustering or partition of the nodes such that nodes within each set or cluster in the partition are well connected in each graph, and the total cost of inter-cluster edges (edges with endpoints in different clusters) is low. Khuller et al. presented one of the earliest solutions to a related *K-Center* problem [2].

There has been significant work on community detection [1,9,20,22]; multiple approaches have been identified as follows: [9]: *i*) topology-based techniques that consider network structure; *ii*) topic-based approaches that rely on textual information within nodes; *iii*) hybrid solutions that combine topology- and topic-based approaches. The majority of existing techniques focus on partitioning nodes rather than edge partitioning. Similar to semEP, Ahn et al. [1] introduce a partition density function based on the similarity of nodes; they detect communities that maximize partition density using optimization methods. This may produce a large number of communities, unlike semEP that produces a minimal number. Ereteo et al. [10] tackle the problem of a semantic social network and propose a topology- and topic-based algorithm, SemTagP, to detect communities from the RDF representation of social networks. Osborne et al. [20] present Temporal Semantic Topic-Based Clustering (TST); it uses similarity between research trajectories and a Fuzzy C-Means algorithm.

Ding et al. [8] provides a comprehensive survey of similarity-based machine learning approaches for drug-target interaction prediction. Several machine learning techniques have been evaluated [11,21,28,29]. Approaches presented by Zheng et al. [29] and Perlman et al. [21] consider feature engineering over multiple similarity features. A PSL based solution [11] directly considers multi-relational structured knowledge and learns from multiple similarity metrics.

5 Evaluation of semEP and State-of-the-Art Methods

5.1 Dataset and Evaluation Protocol

Dataset: A well known dataset of over 900 drugs, almost 1,000 targets, and over 5,000 interactions [3] has been used by Ding et al. to compare several state-of-the-art machine learning based interaction prediction methods [8]. This dataset provides a drug-drug chemical similarity score based on the hashed fingerprints from the SMILES resource, and a target-target similarity score based on the normalized Smith-Waterman sequence similarity score. The targets belong to the following four groups: Nuclear receptors, Gprotein-coupled receptors (GPCRs), Ion channels and Enzymes. Dataset statistics are reported in Table 1.

A 10-fold cross validation will randomly select 90% of positive and negative interactions as *training data*, and will use the remaining 10% of elements as *test data*, for each of the four groups of targets in the dataset.

Table 1. Statistics for the Drug-Target Interaction Dataset [3]

Statistics	Nuclear receptor	GPCR	Ion channel	Enzyme
Number of drugs	54	223	210	445
Number of targets	26	95	204	664
Number of drug target interactions	90	635	1,476	2,926
Average interaction count per target	3.46	6.68	7.23	4.4
Average interaction count per drug	1.66	2.84	7.02	6.57
Graph Density[2]	0.028	0.013	0.017	0.005

semEP Prediction: Recall that Y is a binary matrix where $Y_{i,j} = 1$ if drug d_i interacts with target t_j and $F_{i,j}$ is the score or probability of the prediction. Since semEP is not a machine learning method, it works as follows: We represent the training data from Y as a bipartite graph and apply edge partitioning. Table 2 shows the values of the thresholds θ_d and θ_t used to specify edge constraints in Definition 1. For a selected cluster p, all missing interactions are assigned to be positive interactions in Y. The $F_{i,j}$ score assigned to the interactions in p is the normalized graph density $= \frac{|I|}{|D_p|*|T_p|}$, where $|I|$, $|D_p|$ and $|T_p|$ are the cardinalities of the interactions, drugs and targets in p, respectively. We label this density as the interaction prediction density or *iDensity*.

Table 2. Score threshold θ_d and θ_t for edge constraints in Definition 1

Threshold	Nuclear receptor	GPCR	Ion channel	Enzyme
θ_d	0.3421	0.2759	0.2619	0.2333
θ_t	0.1832	0.1416	0.1355	0.0209

State-of-the-Art Methods: We used the code and results from multiple machine learning based prediction methods that are available as supplemental material to the research reported in [8]. Due to space limitations, we simply label and name all the methods as follows: *i*) BLM: Bipartite Local Method [6]; *ii*) LapRLS: Laplacian Regularized Least Squares [27]; *iii*) GIP: Gaussian Interaction Profile [25]; *iv*) KBMF2K: Kernelized Bayesian Matrix Factorization with twin Kernels [12]; and *v*) NBI: Network-Based Inference [6].

5.2 Results

First, we demonstrate the benefits of using semEP predictions to improve the performance of the prediction methods in [8]. We then validate the best *novel* predictions of all the methods against the STITCH drug-target interaction resource [17].

Using semEP to Improve Performance: To measure the impact of semEP predictions on the performance of the methods, we enhance the (initial) interaction prediction matrix Y of each method, over the hold-out test data, with the best predicted interactions of semEP. The best predictions of semEP are those with an *iDensity* prediction score equal or greater than a 0.5 threshold.

[2] Graph Density is defined as $\frac{2\times\#Edges}{\#Nodes\times(\#Nodes-1)}$.

Further, we limit the added predictions to be no more than 30% of the positive interactions in the holdout set. We label this matrix Y_{semEP}. We also create a control binary matrix Y_{cntrl} which enhances the initial predictions of each method, Y with K interactions, where K corresponds to the cardinality of the added predictions in Y_{semEP}. The entries in $Y_{i,j} = 0$ are randomly chosen (K times) without replacement, following a uniform distribution, to create Y_{cntrl}.

We use the metrics Area Under the Curve (AUC) for precision, and Area Under the Precision-Recall curve (AUPR) for the trade-off between precision and recall. Table 3 reports on the AUC and AUPR of each machine learning method Y, the performance when using semEP predictions, Y_{semEP}, and the control predictions Y_{cntrl}, for each of the four target groups.

The AUC for the methods are generally high, representing the robust performance of these methods. Despite this high baseline, Y_{semEP} is able to improve the performance for all of the methods, for all of the target groups. We also observe that the performance of Y_{cntrl} degrades for all of the methods, for all of the target groups.

The impact of Y_{semEP} is noteworthy when considering the AUPR; these values are somewhat low in general, for all methods, reflecting the sparse training data. Again, we observe a major improvement of AUPR, for all of the methods, for all of the target groups. In addition, there is a sharp decrease of performance of Y_{cntrl} for all of the methods / target groups.

Table 3. 10-fold cross validation AUC and AUPR for methods in [8]. Y is the state-of-the-art method; Y_{semEP} is the semEP enhancement; Y_{cntrl} is the random control.

AUC												
Method	**Nuclear receptor**			**GPCR**			**Ion channel**			**Enzyme**		
	Y	Y_{semEP}	Y_{cntrl}	Y	Y_{semEP}	Y_{cntrl}	Y	Y_{semEP}	Y_{cntrl}	Y	Y_{semEP}	Y_{cntrl}
BLM	0.724	0.778	0.665	0.888	0.911	0.798	0.920	0.929	0.879	0.929	0.935	0.838
NBI	0.690	0.825	0.670	0.833	0.900	0.769	0.925	0.947	0.888	0.895	0.915	0.810
GIP	0.861	0.895	0.803	0.943	0.958	0.843	0.975	0.981	0.932	0.968	0.973	0.874
LapRLS	0.848	0.877	0.799	0.941	0.956	0.844	0.967	0.972	0.925	0.962	0.966	0.868
KBMF2K	0.876	0.914	0.822	0.939	0.960	0.845	0.981	0.985	0.936	0.967	0.971	0.869
AUPR												
Method	**Nuclear receptor**			**GPCR**			**Ion channel**			**Enzyme**		
	Y	Y_{semEP}	Y_{cntrl}	Y	Y_{semEP}	Y_{cntrl}	Y	Y_{semEP}	Y_{cntrl}	Y	Y_{semEP}	Y_{cntrl}
BLM	0.242	0.369	0.238	0.472	0.481	0.327	0.599	0.622	0.542	0.499	0.537	0.373
NBI	0.465	0.682	0.342	0.615	0.719	0.467	0.829	0.854	0.744	0.786	0.818	0.616
GIP	0.657	0.749	0.520	0.705	0.764	0.563	0.888	0.897	0.813	0.869	0.878	0.700
LapRLS	0.577	0.676	0.468	0.630	0.704	0.517	0.800	0.818	0.733	0.830	0.838	0.663
KBMF2K	0.557	0.725	0.475	0.673	0.760	0.544	0.879	0.891	0.810	0.796	0.822	0.656

To further explore the benefit of the semEP predictions, Table 4 compares the overlap of the Top 10 positive predictions in Y_{semEP} and the Top 10 positive predictions of each method in Y. The overlap (equal count) is remarkably low, across all methods, and across all target groups. These results suggest that the interactions predicted by semEP are both *accurate and diverse*, compared to the range of state-of-the-art machine learning based prediction methods. The diversity explains the major impact on AUPR by Y_{semEP} and the potential for semEP to exploit structured knowledge in the relevant semantic space(s).

Table 4. Overlap of Top 10 predictions of semEP and each of the methods in [8]. Entries highlighted in bold are cases where predictions are all different.

	Nuclear receptor		GPCR		Ion channel		Enzyme	
Method	Equal	Different	Equal	Different	Equal	Different	Equal	Different
BLM	1	9	0	**10**	0	**10**	0	**10**
NBI	0	**10**	1	9	0	**10**	0	**10**
GIP	2	8	1	9	0	**10**	3	7
LapRLS	4	6	1	9	0	**10**	2	8
KBMF2K	4	6	0	**10**	0	**10**	0	**10**

Validation Using STITCH: We validated the Top 5 *novel* predicted interactions of all methods; novel interactions are those with $Y_{i,j} = 0$ in the hold-out test dataset. The validation was performed against the latest online version of the STITCH [17] drug target interaction portal[3]. Table 5 reports on the number of validated novel predictions. We observe that as before, semEP is able to identify validated novel interactions across all target groups and it identifies the highest number of validated novel interactions for the target groups of GPCRs and Enzymes. We note that the graphs of GPCRs and Enzymes are sparser than the other two graphs (see Graph Density in Table 1). This provides few opportunities for learning in the training data. Nevertheless, semEP can exploit structured knowledge, edge partitioning and node overlap, to make accurate and diverse predictions, even in this sparse learning environment.

Table 5. Top 5 novel interactions manually validated with STITCH. Entries highlighted in bold correspond to the largest number of novel validations.

Method	Nuclear receptor	GPCR	Ion channel	Enzyme
semEP	4	**5**	1	**4**
BLM	2	1	0	0
NBI	1	1	1	2
GIP	3	3	1	1
LapRLS	**5**	3	**2**	2
KBMF2K	3	4	**2**	2

6 Conclusions and Future Work

We defined the semEP problem to create a minimal partitioning of drug-target interaction edges such that the cluster density of each subset of interaction edges is maximal. We map the semEP problem to the Vertex Coloring Graph problem using *Edge Constraints*. semEP combines the benefits of edge partitioning and edge constraints (incident to the least similar drug-drug or target-target pairs) to identify communities. We conducted an extensive evaluation of semEP on a well-known dataset of drug-target interactions. The results suggest that semEP exploits structured knowledge from semantically annotated data, and is clearly able to predict novel interactions and enhance the performance of sophisticated machine learning methods.

[3] `http://stitch.embl.de/`

In future work, we will explore the use of semEP to identify interesting clusters, and combine / compare with the structure learning of the PSL-based method [11]. We will also apply semEP to other domains, e.g., citation graphs, to identify topical and to predict future relationships between researchers.

Acknowledgement. This research has been partially supported by NSF grant 1147144 and DID-USB. We thank Shobeir Fakhrael and Shanfeng Zhu for providing access to the datasets and algorithms that were evaluated in this paper and Shobeir for valuable feedback.

References

1. Ahn, Y.-Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature 466(7307), 761–764 (2010)
2. Bhatia, R., Guha, S., Khuller, S., Sussmann, Y.: Facility location with dynamic distance functions. Journal of Combinatorial Optimization 2(3), 199–217 (1998)
3. Bleakley, K., Yamanishi, Y.: Supervised prediction of drug–target interactions using bipartite local models. Bioinformatics 25(18), 2397–2403 (2009)
4. Brélaz, D.: New methods to color vertices of a graph. Commun. ACM 22(4), 251–256 (1979)
5. Broecheler, M., Mihalkova, L., Getoor, L.: Probabilistic similarity logic. In: Conference on Uncertainty in Artificial Intelligence (2010)
6. Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., Zhou, W., Huang, J., Tang, Y.: Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Computational Biology 8(5), e1002503 (2012)
7. Cook, D.J., Holder, L.B.: Mining graph data. Wiley-Blackwell (2007)
8. Ding, H., Takigawa, I., Mamitsuka, H., Zhu, S.: Similarity-based machine learning methods for predicting drug-target interactions: A brief review. Briefings in Bioinformatics (2013)
9. Ding, Y.: Community detection: Topological vs. topical. Journal of Infometrics 5(4), 498–514 (2011)
10. Erétéo, G., Gandon, F., Buffa, M.: Semtagp: semantic community detection in folksonomies. In: 2011 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 1, pp. 324–331. IEEE (2011)
11. Fakhraei, S., Huang, B., Raschid, L., Getoor, L.: Network-based drug-target interaction prediction with probabilistic soft logic. In: IEEE/ACM Transactions on Computational Biology and Bioinformatics (2014)
12. Gönen, M.: Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization. Bioinformatics 28(18), 2304–2310 (2012)
13. Ji, M., Sun, Y., Danilevsky, M., Han, J., Gao, J.: Graph regularized transductive classification on heterogeneous information networks. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010, Part I. LNCS, vol. 6321, pp. 570–586. Springer, Heidelberg (2010)
14. Jiang, G., Solbrig, H.R., Chute, C.G.: Adepedia: a scalable and standardized knowledge base of adverse drug events using semantic web technology. In: AMIA Annual Symposium Proceedings (2011)
15. Karp, R.: Reducibility among combinatorial problems. In: Miller, R., Thatcher, J. (eds.) Complexity of Computer Computations, pp. 85–103. Plenum Press (1972)

16. Kuhn, M., Campillos, M., Letunic, I., Jensen, L.J., Bork, P.: A side effect resource to capture phenotypic effects of drugs. Molecular Systems Biology 6(1) (2010)
17. Kuhn, M., Szklarczyk, D., Franceschini, A., von Mering, C., Jensen, L.J., Bork, P.: Stitch 3: zooming in on protein–chemical interactions. Nucleic Acids Research 40(D1), D876–D880 (2012)
18. Li, Z., Narayanan, M., Vetta, A.: The complexity of the simultaneous cluster problem. Journal of Graph Algorithms and Applications (2014)
19. Narayanan, M., Vetta, A., Schadt, E.E., Zhu, J.: Simultaneous clustering of multiple gene expression and physical interaction datasets. PLoS Computational Biology 6(4) (2010)
20. Osborne, F., Scavo, G., Motta, E.: Identifying diachronic topic-based research communities by clustering shared research trajectories. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8465, pp. 114–129. Springer, Heidelberg (2014)
21. Perlman, L., Gottlieb, A., Atias, N., Ruppin, E., Sharan, R.: Combining drug and gene similarity measures for drug-target elucidation. Journal of Computational Biology 18(2), 133–145 (2011)
22. Porter, M.A., Onnela, J.-P., Mucha, P.J.: Communities in networks. Notices of the AMS 56(9), 1082–1097 (2009)
23. Segundo, P.S.: A new dsatur-based algorithm for exact vertex coloring. Computers & OR 39(7), 1724–1733 (2012)
24. Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., Wu, T.: Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In: Proceedings of the 12th EDBT. ACM (2009)
25. van Laarhoven, T., Nabuurs, S.B., Marchiori, E.: Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics 27(21) (2011)
26. Von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing 17(4), 395–416 (2007)
27. Xia, Z., Wu, L.-Y., Zhou, X., Wong, S.T.: Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Systems Biology 4(suppl. 2), S6 (2010)
28. Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., Minoru Kanehisa, M.: Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 24(13), i232–i240 (2008)
29. Zheng, X., Ding, H., Mamitsuka, H., Zhu, S.: Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In: KDD, pp. 1025–1033 (2013)