Hyperlink Prediction in Hypernetworks Using Latent Social Features*

Ye Xu[1], Dan Rockmore[1,2,3], and Adam M. Kleinbaum[4]

[1] Computer Science Department, Dartmouth College
[2] Department of Maths, Dartmouth College
[3] The Santa Fe Institute
[4] Tuck School of Business, Dartmouth College
{ye,rockmore}@cs.dartmouth.edu,
adam.m.kleinbaum@tuck.dartmouth.edu

Abstract. Predicting the existence of links between pairwise objects in networks is a key problem in the study of social networks. However, relationships among objects are often more complex than simple pairwise relations. By restricting attention to dyads, it is possible that information valuable for many learning tasks can be lost. The *hypernetwork* relaxes the assumption that only two nodes can participate in a link, permitting instead an arbitrary number of nodes to participate in so-called *hyperlinks* or *hyperedges*, which is a more natural representation for complex, multi-party relations. *However, the hyperlink prediction problem has yet to be studied.* In this paper, we propose HPLSF (Hyperlink Prediction using Latent Social Features), a hyperlink prediction algorithm for hypernetworks. By exploiting the homophily property of social networks, HPLSF explores social features for hyperlink prediction. To handle the problem that social features are not always observable, a latent social feature learning scheme is developed. To cope with the arbitrary cardinality hyperlink issue in hypernetworks, we design a feature-embedding scheme to map the a priori arbitrarily-sized feature set associated with each hyperlink into a uniformly-sized auxiliary space. To address the fact that observed features and latent features may be not independent, we generalize a structural SVM to learn using both observed features and latent features. In experiments, we evaluate the proposed HPLSF framework on three large-scale hypernetwork datasets. Our results on the three diverse datasets demonstrate the effectiveness of the HPLSF algorithm. Although developed in the context of social networks, HPLSF is a general methodology and applies to arbitrary hypernetworks.

Keywords: Hypernetworks, Hyperlink Prediction, Social Features.

* Dan Rockmore was partially supported by AFOSR Award FA9550-11-1-0166 and the Neukom Institute for Computational Science at Dartmouth College. Ye Xu was partially supported by a grant from the Alfred P. Sloan Foundation.

J. Fürnkranz, E. Hüllermeier, and T. Higuchi (Eds.): DS 2013, LNAI 8140, pp. 324–339, 2013.
© Springer-Verlag Berlin Heidelberg 2013

1 Introduction

Networks provide a powerful framework for modeling real world relationships in which vertices represent objects and links between pairs of vertices indicate their interaction [29]. Nevertheless, in many real world problems, the natural relationships encoding the phenomenon may exist among more than two objects or actors. Examples include buyer-broker-seller triads in a market relationship [3], or subsets of co-expressed genes in a genetic network [16]. In such cases limiting the relationships to dyads may obscure valuable information for learning tasks. The *hypernetwork* is a combinatorial structure in which *hyperlinks* or *hyperedges* represent a relationship that can exist among more than three objects (see e.g., [42]) and thus can provide representation for complex relationships (a hyperlink relating only two actors would simply be a link in the usual network sense). Due to its powerful modeling ability, the hypernetwork framework has attracted attention in a variety of application domains, including scene classification[34], bioinformatics[16], finance[3], and sociology[5].

Link prediction techniques [21,22,8,1] aim to predict the existence of links between vertices in a network. It is an important task in many areas, especially social networks. Thus, it is then natural and as useful to consider the analogous *hyperlink prediction problem* in the hypernetwork setting. A significant difference and challenge in the hypernetwork setting is the a priori arbitrary cardinality of each hyperlink (i.e., the number of nodes associated with the hyperlink). To the best of our knowledge, hyperlink prediction remains untouched in the hypernetwork scenario.

In this paper we address the hyperlink prediction problem in the context of social networks. Social networks often exhibit *homophily* [25], wherein people with similar social affiliations or properties show a preference for interacting with each other. For example, in a college, connections are more likely to exist among students who co-enroll in a class or join in the same sports team or group. A few works [12,27] indicate that considering these social affiliations or features can improve the accuracy for link prediction tasks. Unfortunately, these social affiliations or features are not always observable. By ignoring the "latent" social features (as is done in a few current link prediction algorithms [37,22]), it is possible to lose important information for link predictions. Therefore, it is desirable to utilize these latent social features in link prediction methods. However, finding a means of exploring the "latent" social features is a thorny issue and there is limited research on this in the link prediction literature, let alone hyperlink prediction.

In this paper, we propose HPLSF (Hyperlink Prediction using Latent Social Features), a link prediction algorithm for hypernetworks. Although developed in the context of social networks, HPLSF is a general methodology is readily generalized to arbitrary hypernetworks. Following the homophily property of social networks, we utilize social features for hyperlink prediction. To cope with the problem that social features are often unobservable, we design a scheme to learn latent social features for each individual vertex, each dimension of which is indicative of a plausible social affiliation for the vertex. This transforms the hyperlink prediction problem into a classification task, where latent social features and observed features (if available) are utilized together for hyperlink prediction. The fact that hyperlinks can have arbitrary size (given by the number of actors connected in the hyperlink) raises an additional challenge. We attack this by designing a feature embedding method to map the feature set of the nodes associated with each potential hyperlink into an auxiliary space. In this case, uniformly-sized feature sets are learned from the a priori arbitrarily-sized feature sets. In the last step a structural SVM classifier is generalized under the observed features and latent features after feature embedding because interdependent relationships may exist between observed features and latent features.

In summary, the contributions of this paper are as follows: (1) We design an algorithm to predict the existence of hyperlinks in hypernetworks. As far as we know, HPLSF is the first hyperlink prediction work for hypernetworks. HPLSF can be generalized into any type of hyperneworks although in this paper, we employ email hypernetworks for evaluation. (2)We develop a scheme to learn latent social features for each individual vertex in the hypernetwork. In this way, the *homophily* property of social networks can be fully utilized when considering hyperlink prediction for hypernetworks. (3) We propose a novel feature-embedding strategy to cope with the arbitrarily-sized hyperlink cardinality challenge. Contrary to traditional link prediction work [22], we do not consider the feature set extracted from the group of nodes associated with one potential link/hyperlink directly. Instead, we design a scheme to map this feature set into an embedding space, and each dimension of the embedding space reflects the interaction strength of the group of nodes. In this case, the arbitrarily-sized feature extracted from each hyperlink is mapped into a uniformly-sized feature. (4) We propose to employ structural SVM to learn with both observed features and latent features after feature-embedding in case that observed features and latent features are not necessarily independent. (5) We deploy these ideas on three large-scale email hypernetwork datasets from diverse sources: a large university, an urban-centered hospital, and a large IT corporation. It is the first time that these three datasets are considered in the hypernetwork setting. The heterogeneity of these contexts validate the effectiveness of the proposed HPLSF.

The rest of the paper is organized as follows. In Section 2, we briefly introduce related work. The detailed HPLSF framework is proposed in Section 3. In Section 4 we report experimental results. Finally in Section 5, we conclude the paper.

2 Related Work

Hypernetworks (see e.g.,[42]) have drawn significant attention in various domains. For instance, in [10] hypernetworks are used to model DNA interactions wherein they achieve better disease detection accuracy as compared with using traditional networks. In [3], the hypernetwork is employed to model the correlations of daily stock prices, thereby improving the stock price prediction accuracy. Sun et al. [34] model the set of multiple labels along with the labels' correlations under the multi-instance setting via hypernetworks. Because the *high-order relations* in multi-labels [15] can be captured by hyperlinks, the classification performance is competitive. However, all existing works in the current hypernetwork literature assume constant cardinality for the hyperlinks over the whole hypernetwork. Additionally, these previous researches focus on utilizing the hyperlink relationships to infer the labels of individual nodes in hypernetworks, rather than doing hyperlink predictions in hypernetworks. In our paper, we consider the problem of hyperlink prediction for hypernetworks, and allow for arbitrary (and varying) cardinality of the hyperlink over the hypernetwork.

Our work also relates to social feature learning. Hopcroft et al. [13] indicate that social features reflect the *homophily* property and play an important role in social networks. However, social features are not always observable. There have been various efforts exploring methods to learn the "latent" social features. Neville and Jensen [28] utilize a clustering scheme to achieve a membership vector of each person in the network. In [35], a set of social features is learned using a graph cutting method to help classify relational data in networks. It is worth noting that these works all take advantage of social features to improve the classification performance under the *relational learning* setting [23], i.e., classifying each individual datum (vertex) in a network where data are no longer assumed to be independently and identically distributed. Our work however, aims to utilize the latent social features to predict the existence of hyperlinks in hypernetworks.

Recently, there are a few link prediction models proposed for traditional networks that use "latent" social features [12,11,27,43]. These latent feature models assume that each object (vertex) in the network belongs to a set of latent classes. Thus, the latent class membership of each individual object is useful for predicting pairwise links between objects in networks. Note that these works model latent features according to pairwise relations between vertices in the network. Statistical methods such as variational inference or sampling are used in training and inference. These are time-consuming and prone to suffering from the *local maxima problem.* By contrast, our paper explores the "latent" social features for vertices based on the distance information conveyed in hypernetworks, and employs a simple clustering technique.

Feature embedding [19], which maps a fixed set data into a feature space, is a powerful tool in machine learning. Previous feature embedding techniques were designed for a few particular learning tasks. For example, Kondor [19] developed a feature embedding algorithm for image classification. Grangier [9] employed feature embedding to deal with incomplete data in the original dataset. In our work, we design a feature embedding method to address the arbitrary-sized hyperlink cardinality issue in hypernetwork. As far as we known, it is the first time feature embedding techniques have been used in the network/hypernetwork scenario.

Another line of related work is community detection [30]. Community detection focuses on dividing the vertices in a network into several groups by only using the information encoded in the network topology. It is a hot topic in network study and a few methods have been proposed [31,30]. However, there are fundamental differences between community detection and link/hyperlink prediction algorithms. Almost all community detection algorithms only consider topological information from networks, while our proposed hyperlink prediction method aims to utilize both observed and latent social features from nodes (objects) within the network.

3 Hyperlink Prediction Using Latent Social Features

In what follows, we give the description of HPLSF framework, which takes advantage of the observed information as well as the latent social features from each individual vertex in the hypernetwork.

3.1 Hyperlink Prediction Problem

Before presenting the hyperlink prediction problem in detail, we give the formal description of hypernetworks as follows. A hypernetwork is formalized as an ordered pair $H = (V, E)$, where $V = \{\boldsymbol{v_1}, \boldsymbol{v_2}, ..., \boldsymbol{v_n}\}$ is the set of vertices, and $E = \{\boldsymbol{e_1}, \boldsymbol{e_2}, ..., \boldsymbol{e_m}\}$ is the set of hyperlinks (hyperedges). Therein, if $\boldsymbol{e_i} = \{\boldsymbol{v_{i_1}}, \boldsymbol{v_{i_2}}, ..., \boldsymbol{v_{i_k}}\}$ is a hyperlink with $k > 2$, it is then different from a link (edge) in the traditional network setting because the number of associated vertices could be more than 2. An example of hypernetwork [42] is given in Fig.1.

If all the hyperlinks in the hypernetwork H have the same cardinality k, then the H is a k-uniform hypernetwork, otherwise, H is an *arbitrary-sized hypernetwork*. Although most hypernetwork applications [10,34] can only handle k-uniform hypernetworks, in this paper, we propose a hyperlink prediction framework on arbitrary-sized hypernetworks. The task of link prediction for hypernetworks can thus be formulated as follows: Given a training dataset $S = \{(\boldsymbol{e_1}, y_1), (\boldsymbol{e_2}, y_2), ..., (\boldsymbol{e_t}, y_t)\}$, where $\boldsymbol{e_i}$ represents a possible relation among several vertices, and $y_i \in \{-1, +1\}$ represents the label of the $\boldsymbol{e_i}$ (i.e., if $y_i = +1$, there exists a hyperlink among the set of vertices; if $y_i = -1$, there is no hyperlink.), the goal is to learn the labels in the test set $T = \{\boldsymbol{e_{t+1}}, \boldsymbol{e_{t+2}}, ..., \boldsymbol{e_{t+u}}\}$.

3.2 Exploring Latent Social Features

As we have mentioned, homophily (the idea that people with similar attributes are more likely to interact with each other) is an important characteristic in social networks. Several relational learning works (see e.g., [35]) show that utilizing the homophily property of social networks in the course of learning social features can improve the classification accuracy for network data. A few researchers [12,27] suggest that social features also play a significant role in predicting pairwise links for traditional networks. In hypernetwork scenarios, it is then also natural to consider the homophily property, and thus to take advantage of social features for hyperlink prediction.

Unfortunately, social features are often unobservable. Thus it is not trivial to obtain "latent" social features. In hypernetworks, each social feature indicates (to some extent) a particular property or affiliation for objects. Note that objects

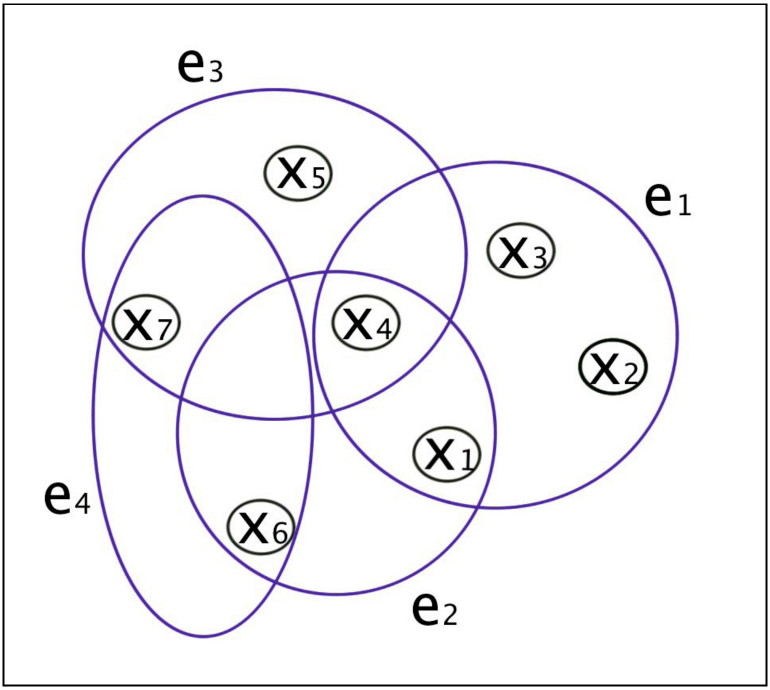

Fig. 1. An example of hypernerwork. The hypernetwork H consists of seven vertices and five hyperlinks. Specifically, $H = (V, E)$, where $V = \{\boldsymbol{x_1}, \boldsymbol{x_2}, ...\boldsymbol{x_7}\}$ and $E = \{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}, \boldsymbol{e_4}\}$. Each of the hyperlink could be associated with more than 2 vertices, namely, $\boldsymbol{e_1} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \boldsymbol{x_4}\}$, $\boldsymbol{e_2} = \{\boldsymbol{x_1}, \boldsymbol{x_4}, \boldsymbol{x_6}\}$, $\boldsymbol{e_3} = \{\boldsymbol{x_4}, \boldsymbol{x_5}, \boldsymbol{x_7}\}$, and $\boldsymbol{e_5} = \{\boldsymbol{x_6}, \boldsymbol{x_7}\}$. Because the number of vertices associated with $\boldsymbol{e_1}$ is 4, we say that the cardinality of hyperlink $\boldsymbol{e_1}$ is equal to 4.

(vertices) sharing similar properties or affiliations in hypernetworks interact at a higher rate than dissimilar objects, and are likely to form groups with more frequent within-group interactions. This is naturally associated with *graph partition* [4], a basic task in *graph theory* [32] which focuses on clustering vertices into groups such that within-group interactions are more frequent than between-group interactions. In this way, the description of affiliation in each particular group is considered as one dimension of social features. Many algorithms [4,33] have been investigated for graph partition, among which clustering-based methods play an important role. Based on the intuition that objects that are similar in group affiliations are very likely to be close in a geometric representation, these clustering-based methods construct a geometric embedding to indicate group affiliations for objects. Therefore, in our work, we employ multidimensional scaling (MDS) [6], a typical geometric embedding learner to explore latent social features based on hypernetwork distance.

Multidimensional scaling (MDS) constructs a geometric embedding (feature vector) preserving as best as possible the original distance among data (objects). Each dimension of the MDS embedding indicates the strength of a certain group affiliation [6] and can be regarded as one social feature. Specifically, for a hypernetwork with N vertices, MDS finds the embedding matrix $\mathbf{Z} \in \mathcal{R}^{N\times p}$ (p is the dimensionality of latent features) whose row vectors are group affiliation descriptions (and thus are treated as latent social features) for the corresponding object in the network as follows,

$$\arg\min_{\mathbf{Z}} \|D - \mathbf{Z}\mathbf{Z}^T\|_F. \tag{1}$$

Therein, $\|\cdot\|_F$ denotes the Frobenius norm, and D is the distance matrix obtained from the hypernetwork (D_{ij} is the length of the shortest path from vertex i to j in the hypernetwork H).

As per [6] we can solve (1) as follows: Let Σ be the matrix of the eigenvectors of D , and Λ be a diagonal matrix with the corresponding eigenvalues. The matrix of the p top eigenvalues is denoted by Λ_p and the corresponding columns of Σ is denoted by Σ_p. Then we can obtain the solution of (1) by,

$$\mathbf{Z} = \Sigma_p \Lambda_p^2 \tag{2}$$

It deserves mentioning that when computing D, we set a shortest path maximum length of five hops in order to avoid the full N^2 computation of all-shortest paths. [6] indicates that this approximation scheme can achieve close result compared with full computation.

3.3 Embedding Features into Uniform-Sized Space

Most existing hypernetwork applications [42,10,34], if not all, simply assume that the cardinality of all hyperlinks in hypernetworks is uniform. Obviously this assumption does not hold under many scenarios and thus limits the application scopes of hypernetworks in real-world. In our work, to address this arbitrary-sized hyperlink cardinality challenge, we propose a feature embedding technique to map feature set from all nodes associated with one potential hyperlink into an embedding space. The dimension of the embedding space is uniform and independent of the cardinality of each particular hyperlink. Therefore, it is convenient to train classifiers under the embedded features for hyperlinks with various cardinality.

Our paper focuses on hyperlink predictions by leveraging social features, where each dimension represents one particular social affiliation for the person(node) in the hypernetwork. We aim to learn a feature embedding from the original social feature set of all persons associated with a potential hyperlink, and the mapped feature in the embedding space is supposed to reflect the discriminability of interaction strength among the group of persons. In this work, we employ *entropy impurity* to measure the similarity strength among a group of people based on the values of each dimension of their social features. If all the group people share similar/close value over one particular social feature, it indicates that those persons are similar over the particular social characteristic and the entropy score would be small. On the contrary, if the values of the particular social feature are diverse, the similarity strength of these people are weak and the entropy score would be large.

In particular, given the social feature set $\{\boldsymbol{z_1}, \boldsymbol{z_2}, ..., \boldsymbol{z_k}\}$ (i.e., features calculated by Eq.(2)) from all nodes associated with one potential hyperlink ($\boldsymbol{z_i} \in \mathcal{R}^{\mathcal{M}}$), we construct a map $f : \mathcal{R}^{M \cdot k} \to \mathcal{R}^{M}$ as follows,

$$f(\boldsymbol{z_1}, \boldsymbol{z_2}, ..., \boldsymbol{z_k}) = [-\sum_{j=1}^{k} p(z_{j1}) logp(z_{j1}), ..., -\sum_{j=1}^{k} p(z_{jM}) logp(z_{jM})] \quad (3)$$

Here, $p(z_{ji})$ is the fraction of feature values at the i^{th} social feature that belong to category z_{ji}, and $-\sum_{j=1}^{k} p(z_{ji}) logp(z_{ji})$ is the entropy score over the total k associated people in the potential hyperlink for the i^{th} social feature.

The designed feature embedding scheme offers great flexibility. First, it can accommodate the hypernetworks with arbitrary-sized hyperlink cardinality. After feature embedding, potential hyperlink features in different dimensionality can be mapped into uniform-sized features (in $\mathcal{R}^{M}$). Second, the embedding technique accommodates both discrete and continuous social features. When the original feature set contains a mix of discrete and continuous values, the entropy score computing scheme naturally handles both of the two types of values. [2] (For continuous values, we can use a threshold based method when calculating the entropy.)

3.4 Learning with Observed and Latent Features

HPLSF aims to conduct hyperlink prediction by leveraging not only observed features but also latent features. In this subsection, we introduce the details of learning with observed and latent features.

After obtaining the embedded latent features (via the method discussed in last subsection) and the observed features[1], we predict whether a hyperlink exists among a set of vertices. The observed features and latent features are not necessarily independent of each other. Therefore, simply combining the observed-features-based classifier and latent-features-based classifier ignores the potential dependence between the output spaces of the two classifiers and might lead to inaccurate prediction results. Different from classic SVM addressing independent output applications [39,40,38], the structural SVM [36] was designed for learning problems involving dependent outputs. Therefore, in our work, a structural SVM based classifier is generalized to capture the potential interdependent relationship between the outputs of the two classifiers.

[1] We assume that the observed features – metadata – are available for vertices in the network.

Structural SVM employs margins between the true structure $\boldsymbol{y}^\star$ and other possible structures $\boldsymbol{y}$:

$$\forall \boldsymbol{y} \in \mathcal{Y}: \ \boldsymbol{w}^\top \phi(\boldsymbol{x}, \boldsymbol{y}^\star) \geq \boldsymbol{w}^\top \phi(\boldsymbol{x}, \boldsymbol{y}) + \varDelta(\boldsymbol{y}^\star, \boldsymbol{y}) - \xi \tag{4}$$

Therein, $\xi > 0$ is a slack variable that controls the tradeoff between satisfying the constraints and optimizing the objectives. $\phi(\boldsymbol{x}, \boldsymbol{y})$ is a joint feature map that characterizes the relation between an input $\boldsymbol{x}$ and an output structure $\boldsymbol{y}$. The loss function $\varDelta(\boldsymbol{y}^\star, \boldsymbol{y})$ quantifies the loss associated with the prediction $\boldsymbol{y}^\star$ when the true output is $\boldsymbol{y}$. In structural SVM, two different structures $(\boldsymbol{y}, \boldsymbol{y}^\star)$ could exhibit similar accuracy, which is reflected in the margin constraint. The violation of margin constraints with high loss $\varDelta(\boldsymbol{y}^\star, \boldsymbol{y})$ should be penalized more severely than the violation involving the output value with smaller loss.

In our hypernetwork scenario, we denote $\boldsymbol{e^o}$ as the embedded observed features of the potential hyperlink $\boldsymbol{e}$ (i.e., set of vertices), and $\boldsymbol{e^l}$ as the embedded latent features of the potential hyperlink $\boldsymbol{e}$. Here, $\boldsymbol{e^o} = f(\boldsymbol{x_1^o}, \boldsymbol{x_2^o}, ..., \boldsymbol{x_k^o})$ is the embedding of observed features from each individual vertex belonging to the vertices set $\boldsymbol{e}$, and $\boldsymbol{e^l} = f(\boldsymbol{x_1^l}, \boldsymbol{x_2^l}, ..., \boldsymbol{x_k^l})$ is the embedding of latent features from each individual vertex. Meanwhile, we define $\boldsymbol{y} = [y^o, y^l]$ where y^o is the output corresponding to embedded observed features $\boldsymbol{e^o}$ and y^l is the output corresponding to embedded latent features $\boldsymbol{e^l}$.[2] We define the loss function of form

$$\varDelta(\boldsymbol{y}^\star, \boldsymbol{y}) = \frac{1}{2}(\mathbf{1}(y^{\star o} \neq y^o) + \mathbf{1}(y^{\star l} \neq y^l)) \tag{5}$$

where $\mathbf{1}(S) = 1$ if S is true; otherwise, $\mathbf{1}(S) = 0$. The defined $\varDelta(\boldsymbol{y}^\star, \boldsymbol{y})$ is non-negative and bounded in $[0, 1]$. This loss function supports flexible notions of structural correctness and has been widely used in many structural SVM work [14,24].

Thus the joint feature map for the structural SVM can be written as follows:

$$\varPhi(\boldsymbol{e}, \boldsymbol{y}) = [\phi_o(\boldsymbol{e^o}, y^o), \ \phi_l(\boldsymbol{e^l}, y^l)] \tag{6}$$

where $\phi_o(\boldsymbol{e^o}, y^o)$ is the feature map describing the relation between observed features of a potential hyperlink and its corresponding output, and $\phi_l(\boldsymbol{e^l}, y^l)$ is the feature map describing the relation between latent features and its corresponding output.

Using the joint feature map defined in Eq.(6), the constraints for the HPLSF can be formulated as the similar form as Eq.(4). By adding the objective function which minimizes the combination of the regularization term and the penalty term for slack variables, the optimization problem can be written as follows:

$$\min_{\boldsymbol{w}, \boldsymbol{\xi}} \ \tfrac{1}{2}\|\boldsymbol{w}\|^2 + \tfrac{C}{N}\textstyle\sum_{i=1}^N \xi_i \tag{7}$$

$$\forall i, \forall \boldsymbol{y} \in \mathcal{Y} \backslash \boldsymbol{y_i} : \boldsymbol{w}^\top \varPhi(\boldsymbol{e_i}, \boldsymbol{y}) \geq \boldsymbol{w}^\top \varPhi(\boldsymbol{e_i}, \boldsymbol{y_i}) + \varDelta'(\boldsymbol{y}, \boldsymbol{y_i}) - \xi_i \tag{8}$$

[2] In our work, for each $\boldsymbol{e_i}$, we use 1-Nearest Neighbor Classifier (1-NNC) as the observed-features-based classifier and the latent-features-based classifier to compute y_i^o and y_i^l respectively.

where C is the parameter controlling the tradeoff between satisfying the constraints and minimizing the regularization term. In all our experiments, we set up $C = 1$. The optimization method discussed in [14] is employed to solve the problem (8).

3.5 Summary of the Proposed Algorithm

```
Input: Hypernetwork $H = (V, E)$, where $V$ is the vertex set and $E$ is the set of hyperlinks.
Construct distance matrix $D$ from $H$, where $D_{ij}$ indicates the length of the shortest path from vertex $i$ to $j$ in the hypernetwork.
Calculate latent features for each object in the hypernetwork using Eq.(2).
Embed the observed feature set and the latent feature set for each hyperlink respectively using Eq.(3).
Construct the joint feature maps as Eq.(6) using embedded latent features and embedded observed features.
Solve the problem (8) via the SVM$^{struct}$.
```

Algorithm 1. Hyperlink Prediction Using Latent Social Features

In this section, we present the detailed HPLSF in Algorithm 1. First, we construct the distance matrix D whose element D_{ij} gives the length of the shortest path between two vertices in the hypernetwork. Then we apply MDS algorithm to calculate the embedding matrix, where each row vector represents the latent feature for the corresponding vertex (object). After obtaining the embedded latent features, we can use them together with the embedded observed features to construct the constraints of the optimization problem (8). Lastly, SVM^{struct} [14] is employed to train the classifier. Note that the proposed prediction framework can be generalized into any other type of hypernetworks although in the experiments we only evaluate it using email hypernetworks.

4 Experiments

In what follows, we introduce three real-world hypernetwork datasets to evaluate the proposed HPLSF framework. The first dataset was collected in a major urban hospital over one year [7]. The second dataset was collected in a large university over 6 semesters (three years) [20,41]. The last dataset was collected from a large IT corporation over three years [18,17]. Because there is no other hyperlink prediction algorithm for hypernetworks with which we can compare our results, we compare HPLSF with three baseline methods. Ob-Model (the classifier trained using embedded observed features) is used as the baseline to demonstrate the importance and necessity of exploring latent social features for hyperlink prediction. We also execute Ex-Model-MDS (the classifier trained using embedded latent features that are learned by the same MDS procedure as HPLSF) and Ex-Model-LFRM

Table 1. The overview of the hospital collaborative hypernetwork datasets

Hyperlink Cardinality	#Hyperlinks
3	735
4	510
5	341
6	245
Arbitrary-Sized	1831

Table 2. The hyperlink prediction accuracy (%) using HPLSF and Ex-Model-LFRM under the hospital collaborative hypernetwork datasets

Hyperlink Cardinality	HPLSF	Ex-Model-LFRM
3	$\mathbf{90.6 \pm 0.4}$	86.2 ± 1.1
4	$\mathbf{88.9 \pm 0.9}$	81.5 ± 2.3
5	$\mathbf{84.3 \pm 2.2}$	74.3 ± 3.5
6	$\mathbf{85.6 \pm 0.4}$	73.7 ± 0.4
Arbitrary-Sized	$\mathbf{80.6 \pm 1.3}$	71.2 ± 1.8

(the classifier trained using embedded latent features that are learned by the Latent Feature Relational Model (LFRM) proposed in [27]) in order to validate the effectiveness of the designed latent social feature learning scheme in HPLSF. Note that LFRM was designed to model latent features under the traditional "two-vertex-link" setting. Thus in our experiments, we transform the hypernetwork into the traditional "two-vertex-link" network (linking all pairs of vertices contained in a particular hyperlink) when using LFRM.

4.1 Hospital Message Hypernetwork Dataset

In this subsection, we consider HPLSF in the context of an email dataset derived from communications in an urban hospital [7]. The hypernetwork contains message communications among patients, their family members, clinicians, and researchers who work on coming up with a cure for particular diseases in the hospital. The messages are collected among 11,944 people over one year via an internal message system in the hospital. The people involved in the message system are treated as the vertex set and all persons that appear in one message are regarded as the set of vertices of a hyperlink. Most hyperlinks in the dataset has a cardinality no larger than 6. In this experiment, we respectively consider $3-$cardinality, $4-$cardinality, $5-$cardinality, $6-$cardinality, and arbitrary-sized cardinality when constructing the hypernetwork. In other word, in the $k-$cardinality hypernetwork, the messages containing exactly k people are considered and others are discarded, while in the arbitrary-sized cardinality hypernetwork, we consider all hyperlinks with any cardinality value. Meanwhile, we randomly sample sets of nodes to form negative hyperlinks. The detailed information about this dataset can be found in Table 1.

Due to the strict privacy regulations in the hospital, the content of each message is discarded. Additionally, any personal information for each person (vertex) in the message system can not be accessed. Therefore, in this dataset, HPLSF only applies latent features to predict hyperlinks. (In this case, HPLSF is equivalent to Ex-Model-MDS.) To demonstrate the effectiveness of the latent feature learning of HPLSF, we execute Ex-Model-LFRM under the hospital collaborative hypernetwork dataset. In this experiment, the latent feature dimension p is determined by this: we use five times 10-fold cross validation to tune this parameter for Ex-Model-LFRM. Then for HPLSF, we use the same value of p as in Ex-Model-LFRM. After obtaining the latent features, SVM^{struct} [14] is employed for training and testing. The 10-fold cross validation scheme is used to achieve the average prediction accuracy, which is listed in Table 2.

The results in Table 2 indicates that HPLSF is able to obtain higher prediction accuracy than Ex-Model-LFRM under all the conditions. Note that as the hyperlink cardinality k increases, the difference between the accuracy of two methods grows. This fact implies that the proposed latent feature learning scheme in HPLSF outperforms LFRM [27] under the hyperlink prediction scenario. It is because that LFRM was designed for the traditional networks and may fail in modeling hyperlink relations. As for the arbitrary-sized hypernetwork, the difference between our HPLSF and the baseline is also significant. Pairwise t-tests at 95% significance level demonstrate the validity of the experiments.

4.2 University Email Hypernetwork Dataset

In what follows, we use a university email hypernetwork dataset [20,41]. The dataset contains email messages delivered to users via the university email system over six separate semesters. The email user population is a mix of students, faculty members, staff, and “affiliates” (a category including postdocs, visiting scholars, and alumni) in the university. Every email record is composed of date, time, sender, and list of recipients. Out of privacy and security concerns, the contents of email messages are discarded and the email addresses are encrypted. However, in this email system, we are allowed to access an email user table that describes the personal information of each user, namely occupation, birth, gender, home country, postal code, years at the university, academic department (for student and faculty), division (for student only), and dormitory building (for student only). Email messages from each of the six semesters are treated as a separate dataset. Each person is treated as a vertex in the hypernetwork, and all the persons that appear in one email are regarded as a set of vertices in a hyperlink. The personal information from every email user is regarded as observed features for each vertex. The average number of nodes for each dataset is 67,736, and the average number of hyperlinks is 253,469. We obtain positive data and negative data using similar scheme as last subsection. Detailed information of the six datasets is listed in Table 3.

We execute HPLSF under each of the six datasets respectively. Ob-Model, Ex-Model-MDS, and Ex-Model-LFRM are used for comparison. In Ob-Model, we use all the accessible user information as observed features. In Ex-Model-LFRM,

Table 3. The overview of the university email hypernetwork datasets

Time	#Vertices	#Hyperlinks
Semester1	57,328	340,717
Semester2	61,451	248,009
Semester3	65,946	131,448
Semester4	73,040	458,273
Semester5	77,256	163,930
Semester6	71,396	178,435

Table 4. The hyperlink prediction accuracy (%) using HPLSF, Ob-Model, Ex-Model-MDS, and Ex-Model-LFRM under the university email hypernetwork datasets

Time	HPLSF	Ob-Model	Ex-Model-MDS	Ex-Model-LFRM
Semester1	$\mathbf{88.7 \pm 0.9}$	82.8 ± 1.1	79.2 ± 1.8	72.5 ± 1.3
Semester2	$\mathbf{89.4 \pm 0.6}$	81.2 ± 2.0	78.7 ± 1.0	73.2 ± 3.0
Semester3	$\mathbf{92.4 \pm 1.0}$	83.2 ± 0.9	83.6 ± 1.4	77.6 ± 3.3
Semester4	$\mathbf{89.7 \pm 1.8}$	80.4 ± 1.7	74.3 ± 2.2	67.7 ± 1.8
Semester5	$\mathbf{90.1 \pm 1.2}$	81.3 ± 1.3	82.3 ± 1.8	79.6 ± 1.1
Semester6	$\mathbf{87.1 \pm 1.2}$	79.4 ± 2.2	83.0 ± 1.5	79.6 ± 1.3

variational inference is used to learn the parameter in the latent feature relational model as described in [26]. The latent feature dimension p is determined using the same scheme as last subsection. SVM^{struct} [14] is applied for training and prediction. In all the methods, we employ the 10-fold cross validation scheme to achieve the average prediction accuracy and list them in Table 4. Table 4 indicates that the prediction accuracy of HPLSF outperforms all the baselines under all the datasets. HPLSF achieves significantly higher accuracy than Ob-Model, which demonstrates that exploring "latent" social features are helpful and necessary for hyperlink predictions because social features are not always observable. Meanwhile, the accuracy of Ex-Model-MDS is much higher than Ex-Model-LFRM under almost all datasets, which implies that HPLSF designs a better way to explore latent features in hypernetworks. Pairwise t-tests at 95% significance level demonstrate the validity of the experiments.

4.3 IT Company Email Hypernetwork Dataset

In what follows, an email hypernetwork dataset collected from a large information technology and electronics company [18,17] is employed to evaluate the proposed HPLSF framework. The dataset contains the complete record, as drawn from the company's servers, of email communications among 30,328 employees from 2006 to 2008.The employees in the company are located in 289 different offices around 50 states in United States and collectively comprise about one quarter of the company's employee population. Each email record comprises the timestamp, sender, lists of receipients, and the size of the message. Privacy laws and corresponding company policies preclude the collection of the content of messages. However, some personal information for each employee is accessible from the HR department of the company, namely work years in the company, employee's job function, office location code, the state of the office, and employee's group ID. Email messages from each of the three years are treated as a separate dataset. Each person is regarded as a vertex in the hypernetwork while all people appearing in one email message are regarded as vertices associated with the hyperlink. The personal information for each employee obtained from HR department is treated as observed features. Detailed information of the three datasets is shown in Table 5.

Table 5. The overview of the IT company email hypernetwork datasets

Year	#Vertices	#Hyperlinks
2006	30,328	992,382
2007	27,134	1,074,507
2008	27,134	473,756

Table 6. The hyperlink prediction accuracy (%) using HPLSF, Ob-Model, Ex-Model-MDS, and Ex-Model-LFRM under the IT company email hypernetwork datasets

Year	HPLSF	Ob-Model	Ex-Model-MDS	Ex-Model-LFRM
2006	$\mathbf{87.4 \pm 0.8}$	74.8 ± 0.2	76.5 ± 2.2	73.7 ± 0.9
2007	$\mathbf{85.1 \pm 0.6}$	81.3 ± 1.3	78.4 ± 2.4	75.3 ± 2.0
2008	$\mathbf{86.7 \pm 1.1}$	77.5 ± 1.3	81.4 ± 1.7	76.8 ± 1.3

We run HPLSF on each of the three datasets sequentially and obtain the hyperlink prediction results. To compare with the proposed algorithm, the three baselines are used as in last section. In Ob-Model, all available personal information from the HR department is used as each employee's observed features. In Ex-Model-LFRM, we still use variational inference to learn the parameter for the LFRM. For all the methods, 10-fold cross validation scheme is employed to calculate the average prediction accuracy.

The results listed in Table 6 show that the proposed HPLSF achieves higher accuracy than any other baseline methods. HPLSF performs much better than Ob-Model, establishing that "latent" social features are helpful for hyperlink predictions. Meanwhile, Ex-Model-MDS still performs better than Ex-Model-LFRM on all the three datasets, which demonstrates again that the latent feature learning scheme designed in HPLSF is better than LFRM under the hypernetwork scenario. Pairwise t-tests at 95% significance level demonstrate the validity of the experiments.

5 Conclusion

In this paper, we propose a link prediction framework for hypernetworks, which is, to the best of our knowledge, the first hyperlink prediction work. The framework, named HPLSF, aims to predict whether a hyperlink exists among a set of vertices in a hypernetwork by leveraging not only observed features but also latent features. By designing a feature-embedding technique, we address the artbitrary-sized hyperlink cardinality challenge in hypernetwork setting. Because observed features and latent features are not necessarily independent of each other, we generalize a structural SVM rather than simply combining the results obtained from classifiers using each of the two types of features. The experimental results under three large email hypernetworks from diverse sources demonstrates the effectiveness of HPLSF.

References

1. Backstrom, L., Leskovec, J.: Supervised random walks: predicting and recommending links in social networks. In: WSDM 2011 (2011)
2. Barros, R.C., Basgalupp, M.P., de Carvalho, A., Freitas, A.A.: A survey of evolutionary algorithms for decision-tree induction. IEEE Trans. SMC 42(3), 291–312 (2012)
3. Bautu, E., Kim, S., Bautu, A., Luchian, H., Zhang, B.-T.: Evolving hypernetwork models of binary time series for forecasting price movements on stock markets. In: IEEE Evolutionary Computation 2009 (2009)
4. Bichot, C.-E., Siarry, P.: Graph Partitioning: Optimisation and Applications. Wiley (2011)
5. Bonachich, P., Holdren, A., Johnston, M.: Hyper-edges and multidimensional centrality. Social Networks 26(3), 189–203 (2004)
6. Cox, T.F., Cox, M.A.A.: Multidimensional Scaling. Chapman and Hall (2001)
7. Gloor, P.A., et al.: Towards growing a coin in a medical research community. Procedia Social and Behavioral Sciences (2010)
8. Gao, S., Denoyer, L., Gallinari, P.: Temporal link prediction by integrating content and structure information. In: CIKM 2011 (2011)
9. Grangier, D., Melvin, I.: Feature set embedding for incomplete data. In: NIPS 2010 (2010)
10. Ha, J.-W., Eom, J.-H., Kim, S.-C., Zhang, B.-T.: Evolutionary hypernetwork models for aptamer-based cardiovascular disease diagnosis. In: GECCO 2007 (2007)
11. Hoff, P.D.: Modeling homophily and stochastic equivalence in relational data. In: NIPS 2007 (2007)
12. Hoff, P.D., Raftery, A.E., Handcock, M.S.: Latent space approaches to social network analysis. J. American Statistical Association 97, 1090–1098 (2001)
13. Hopcroft, J., Khan, O., Kulis, B., Selman, B.: Natural communities in large linked networks. In: SIGKDD 2003 (2003)
14. Joachims, T., Finley, T., Yu, C.J.: Cutting-plane training of structural svm. Machine Learning 77(1), 27–59 (2009)
15. Kang, F., Jin, R., Sukthankar, R.: Correlated label propagation with application to multi-label learning. In: CVPR 2006 (2006)
16. Kim, S., Kim, S.-J., Zhang, B.-T.: Evolving hypernetwork classifiers for microrna expression profile analysis. In: IEEE Evolutionary Computation 2007 (2007)
17. Kleinbaum, A.M.: Organizational misfits and the origins of brokerage in intra-firm networks. Administrative Science Quarterly 57, 407–452 (2012)
18. Kleinbaum, A.M., Stuart, T.E.: Inside the black box of the corporate staff: Social networks and the implementation of corporate strategy. Strategic Management Journal (2013)

19. Kondor, R., Jebara, T.: A kernel between set of vectors. In: ICML 2003 (2003)
20. Kossinets, G., Watts, D.J.: Empirical analysis of an evolving social network. Science (2006)
21. Liben-Nowell, D., Kleinberg, J.: The link prediction problem for social networks. In: CIKM 2003 (2003)
22. Lichtenwalter, R.N., Lussier, J.T., Chawla, N.V.: New perspectives and methods in link prediction. In: SIGKDD 2010 (2010)
23. Macskassy, S.A., Provost, F.: Classification in networked data: A toolkit and a univariate case study. JMLR 8, 935–983 (2007)
24. McFee, B., Lanckriet, G.: Metric learning to rank. In: ICML 2010 (2010)
25. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual Review of Sociology 27(1), 415–444 (2001)
26. Miller, K.T.: Bayesian nonparametric latent feature models. Ph.D. Thesis, University of California, Berkeley (2011)
27. Miller, K.T., Griffiths, T.L., Jordan, M.I.: Nonparametric latent feature models for link prediction. In: NIPS 2009 (2009)
28. Neville, J., Jensen, D.: Leveraging relational autocorrelation with latent group models. In: SIGKDD Workshop 2005 (2005)
29. Newman, M.: The structure and function of complex networks. SIAM Review 45(1), 167–256 (2003)
30. Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103(23), 8577–8582 (2006)
31. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043), 814–818 (2005)
32. Pothen, A.: Graph partitioning algorithms with applications to scientific computing. Technical Report, Norfolk, VA (1997)
33. Shi, J., Malik, J.: Normalized cuts and image segmentation. TPAMI 22(8), 888–905 (2000)
34. Sun, L., Ji, S., Ye, J.: Hypergraph spectral learning for multi-label classification. In: SIGKDD 2008 (2008)
35. Tang, L., Liu, H.: Relational learning via latent social dimensions. In: SIGKDD 2009 (2009)
36. Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector learning for interdependent and structured output spaces. In: ICML 2004 (2004)
37. Wang, C., Satuluri, V., Parthasarathy, S.: Local probabilistic models for link prediction. In: IEEE ICDM 2007 (2007)
38. Xie, L., Gu, N., Li, D., Cao, Z., Tan, M., Nahavandi, S.: Concurrent control chart patterns recognition with singular spectrum analysis and support vector machine. Computers and Industry Engineering 64(1), 280–289 (2013)
39. Xie, L., Li, D., Simske, S.J.: Feature dimensionality reduction for example-based image super-resolution. Journal of Pattern Recognition Research 2, 130–139 (2011)
40. Xu, Y., Ping, W., Campbell, A.: Multi-instance metric learning. In: ICDM 2011 (2011)
41. Xu, Y., Rockmore, D.: Feature selection for link prediction. In: PIKM 2012 (2012)
42. Zhou, D., Huang, J., Schölkopf, B.: Learning with hypergraphs: Clustering, classification, and embedding. In: NIPS 2006 (2006)
43. Zhu, J.: Max-margin nonparametric latent feature models for link prediction. In: ICML 2012 (2012)