© 2023 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes,creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. DOI: 10.1109/ICASSP49357.2023.10096091

HIGHER-ORDER LINK PREDICTION VIA LEARNABLE MAXIMUM MEAN DISCREPANCY

Georgios V. Karanikolas,[†] *Alba Pagès-Zamora,*[‡] *and Georgios B. Giannakis*[†]

[†]Dept. of ECE, University of Minnesota, Minneapolis, MN
[‡] SPCOM Group, Universitat Politècnica de Catalunya, Barcelona, Spain

ABSTRACT

Higher-order link prediction (HOLP) seeks missing links capturing dependencies among three or more network nodes. Predicting high-order links (HOLs) can for instance reveal hyperlinks in the structure of drug substance and metabolic networks. Existing methods either make restrictive assumptions regarding the emergence of HOLs, or, they rely on reduced dimensionality models of limited expressiveness. To overcome these limitations, the HOLP approach developed here leverages distribution similarities across embeddings as captured by a learnable probability metric. The intuition underpinning the novel approach is that sets of nodes whose embeddings are less similar in distribution, are less likely to be connected by a HOL. Specifically, nonlinear dimensionality reduction is effected through a Gaussian process latent variable model that yields nodal embeddings, and also learns a data-driven similarity function (kernel). This kernel forms the core of a maximum mean discrepancy probability metric. Tests on benchmark datasets illustrate the potential of the proposed approach.

***Index Terms*—** Link prediction, probability metrics, Gaussian processes

1. INTRODUCTION

Link prediction refers to the task of predicting edges (links) missing from a graph [13, 20, 10]. These edges may exist, yet remain unobserved due to e.g., privacy concerns; or, they may appear in the future, in the case of temporally evolving graphs [8].

An edge represents an interaction between two nodes. A hyperedge on the other hand, connects three or more nodes, thereby providing a natural representation for the higher-order dependencies amongst them [18]. Higher-order link prediction aims at predicting hyperedges [24, 25]. As a motivating example, consider drug substance networks, where substances are represented by nodes, and a drug can be viewed as the (hyper) edge connecting the substances present in the drug. As it is typical for a drug to contain more than two substances, limiting the prediction scope to edges undermines the potential of forecasting the emergence of new effective drugs.

Prior works. HOLP approaches can be categorized into informal scoring, supervised, and unsupervised learning methods. Informal scoring methods assume that a particular network characteristic, such as the number of common neighbors across nodes comprising a potential HOL, is a good predictor of hyperedge presence [9]. As real-world networks can describe disparate phenomena, it is possible that such network characteristics are not clear HOL indicators of the network at hand. Supervised methods pose HOLP as a classification task of 'present' versus 'absent' hyperedges [24, 21]. Observed hyperedges belong to the 'present' class; while unobserved hyperedges do not necessarily belong to the 'absent' class; that is, there are no labeled training samples for the 'absent' class. Since training a classifier without samples from both classes is not an option, these approaches resort to artificially generated 'absent hyperedges,' meaning sets of vertices assumed not to be connected by a hyperedge. The generation mechanism employed introduces assumptions on hyperedge formation and inherently biases the learned (predictor) model.

Unsupervised HOLP approaches rely on generalized matrix factorization and matching (MFM) [25]. Although the resultant algorithms do not place assumptions on HOL formation, they model embeddings as a linear function of the data employed in the MFM. In addition, all candidate hyperedges are required to be available during the training phase.

To overcome the limitations of existing approaches, our novel approach starts with nonlinear nodal embeddings, and relies on a distribution similarity metric among these embeddings to predict HOLs. The premise is that sets of nodes whose embeddings are less similar in distribution are less likely to be connected by a hyperedge. Unlike informal scoring and supervised approaches, we make no mechanistic assumptions with regards to hyperedge formation. Instead, the distribution similarity is assessed using the maximum mean discrepancy (MMD) metric that relies on a kernel function [6]. The latter is learned jointly with the nodal embeddings by means of a nonlinear dimensionality reduction approach, namely the Gaussian process latent variable model

This work was supported in part by NSF grants 1901134, 2126052, 2212318, 2220292. A. Pages-Zamora was supported by grants PID 2019-104958RB-C41 and RED2018-102668-T funded by MCIN/AEI/ 10.13039/501100011033; and by 2017 SGR 578 funded by Dept. de Recerca i Univ. de la Generalitat de Catalunya 10.13039/501100002809.

(GPLVM) [11]. To score a candidate hyperedge, we consider the possible partitions of its constituent nodes into pairs of sets, compute the MMD of the sets of embeddings corresponding to said sets of nodes, and average over partitions.

Contributions. The use of *distribution similarity* based on a *learnable* MMD metric for HOLP are key novelties of this contribution. *Unsupervised* learning of scoring functions for candidate hyperedges and similarity functions for nodal embeddings, is also novel. HOLP alternatives are either supervised (see e.g., [21] for an approach relying on neural networks) or, in the case of MFM methods, they can only provide scores for candidate hyperedges at the training stage (instead of scoring *functions*); and they additionally rely on fixed (linear) similarity functions in the space of nodal embeddings. Finally, it is worth mentioning the related problem of simplicial closure prediction [2], where hyperedges of fixed cardinality are assumed, alongside with additional constraints on hyperedge formation [1]. Notwithstanding, these modeling assumptions are not made in the HOLP approach here.

Notation. Scalar $[\mathbf{A}]_{ij}$ denotes the (i,j)-th entry of the matrix $\mathbf{A}$, superscript $^\top$ denotes transposition, and $\mathbf{I}$ stands for the identity matrix; $|\cdot|$ denotes set cardinality, $\varnothing$ the empty set, and $\mathbb{1}\{\cdot\}$ is the indicator function. The binomial coefficient is given by $\binom{n}{k}$, the probability of the event A is denoted by $\Pr\{A\}$ and $\lfloor a \rfloor$ stands for the largest integer less than or equal to a. Finally, $\mathcal{N}(x;\mu,\sigma^2)$ represents the value, at x, of the probability density function of a Gaussian random variable with mean μ and variance σ^2.

2. PRELIMINARIES

Consider an undirected hypergraph $\mathcal{G} := (\mathcal{V}, \mathcal{E}_o)$, where $\mathcal{V}$ is the set of vertices, and $\mathcal{E}_o$ denotes the set of *observed* edges linking pairs of vertices, as well as hyperedges that by definition connect three or more vertices. The ambitious goal here is to predict the set $\mathcal{E}_u$ of *unobserved* hyperedges that do not necessarily involve subsets of vertices with fixed cardinality. Given a set $\mathcal{E}_c$ of candidate hyperedges, a HOLP approach assigns a score $S(e)$ per hyperedge $e \in \mathcal{E}_c$, with the premise that higher scores are assigned to candidate hyperedges e that are considered more likely to be present.

The structure of the hypergraph $\mathcal{G}$ can be represented using the incidence matrix $\mathbf{H} \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{E}_o|}$, where $[\mathbf{H}]_{ij} = \mathbb{1}\{i \in e_j\}$, that is $[\mathbf{H}]_{ij} = 1$, if vertex i is involved in (hyper) edge $e_j \in \mathcal{E}_o$, and $[\mathbf{H}]_{ij} = 0$ otherwise [3].

3. GPLVM FOR NODAL EMBEDDINGS

Although $\mathbf{H}$ describes the structure of $\mathcal{G}$, working directly with $\mathbf{H}$ can be challenging for large-scale hypergraphs and a large number of observed hyperedges. Instead, one can rely on some form of the adjacency matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ obtained from $\mathbf{H}$, as in e.g., [25]. Here, we will rely on $\mathbf{A} := \mathbf{D}_v^{-1/2}\mathbf{H}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{H}^\top\mathbf{D}_v^{-1/2}$, where $\mathbf{W}$ is a diagonal weight matrix with diagonal entries $[\mathbf{W}]_{mm} := w(e_m)$; while $\mathbf{D}_v$ is a diagonal vertex degree matrix with diagonal entries $[\mathbf{D}_v]_{nn} := \sum_{\{e\in\mathcal{E}_o\ |n\in e\}} w(e)$; and likewise for the edge degree matrix having diagonal entries $[\mathbf{D}_e]_{mm} := |e_m|$; see also [26]. In temporally evolving hypergraphs, the weight $w(e)$ can represent the number of times the (hyper) edge e is observed in $\mathcal{E}_o$.

To enable HOLP, the adjacency matrix $\mathbf{A}$ is commonly assumed to have a latent low-dimensional structure [25]. Since the i-th column $\mathbf{a}_i$ of $\mathbf{A}$ describes the association of vertex i with the remaining vertices $j \neq i$, we will leverage the structure of $\mathbf{A}$ by obtaining per vertex i a low-dimensional embedding $\mathbf{x}_i \in \mathbb{R}^d$ of $\mathbf{a}_i \in \mathbb{R}^{|\mathcal{V}|}$, where $d \ll |\mathcal{V}|$.

Specifically, the j-th entry of $\mathbf{a}_i$ will be expressed using a Gaussian process latent variable model (GPLVM)[1] [11] as

$$[\mathbf{a}_i]_j \;=\; f_j(\mathbf{x}_i) + \varepsilon_{ij} \tag{1}$$

where f_j is a nonlinear function modeled using a Gaussian process (GP) prior, and $\{\varepsilon_{ij}\}$ are independent and identically distributed from $\mathcal{N}(0,\sigma_\varepsilon^2)$ [23]. The covariance of the said prior is captured by a kernel $\kappa_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ collects the kernel hyperparameters. With $\mathbf{f}_j := [f_j(\mathbf{x}_1)\ldots f_j(\mathbf{x}_{|\mathcal{V}|})]^\top$, the GP prior over functions f_j is tantamount to a multivariate Gaussian $p(\mathbf{f}_j|\mathbf{X};\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}_j;\mathbf{0},\mathbf{K}_{\boldsymbol{\theta}})$, where $\mathbf{X} := [\mathbf{x}_1,\ldots,\mathbf{x}_{|\mathcal{V}|}]^\top$ collects the nodal embeddings viewed as random vectors, and the kernel matrix $\mathbf{K}_{\boldsymbol{\theta}}$ is formed with (l,m)-th entry $\kappa_{\boldsymbol{\theta}}(\mathbf{x}_l,\mathbf{x}_m)$. Letting $\mathbf{F} := [\mathbf{f}_1,\ldots,\mathbf{f}_{|\mathcal{V}|}]$, Bayes rule implies $p(\mathbf{A},\mathbf{F}|\mathbf{X};\boldsymbol{\theta},\sigma_\varepsilon^2) = p(\mathbf{A}|\mathbf{F},\mathbf{X};\boldsymbol{\theta},\sigma_\varepsilon^2)p(\mathbf{F}|\mathbf{X};\boldsymbol{\theta})$, where the form of $p(\mathbf{A}|\mathbf{F},\mathbf{X};\boldsymbol{\theta},\sigma_\varepsilon^2)$ follows from (1) and $p(\mathbf{F}|\mathbf{X};\boldsymbol{\theta}) = \prod_{j=1}^{|\mathcal{V}|} p(\mathbf{f}_j|\mathbf{X};\boldsymbol{\theta})$. Having $p(\mathbf{A},\mathbf{F}|\mathbf{X};\boldsymbol{\theta},\sigma_\varepsilon^2)$ marginalized over $\mathbf{F}$, and supposing independence, we obtain

$$p(\mathbf{A}|\mathbf{X};\boldsymbol{\theta},\sigma_\varepsilon^2) \;=\; \prod_{j=1}^{|\mathcal{V}|} \mathcal{N}(\mathbf{a}_j;\mathbf{0},\mathbf{K}_{\boldsymbol{\theta}} + \sigma_\varepsilon^2\mathbf{I}). \tag{2}$$

The probability density function (pdf) in (2) along with the prior $p(\mathbf{X}) = \prod_{i=1}^{|\mathcal{V}|}\mathcal{N}(\mathbf{x}_i;\mathbf{0},\mathbf{I})$ and Bayes rule, yield the likelihood $p(\mathbf{A},\mathbf{X};\boldsymbol{\theta},\sigma_\varepsilon^2)$. The latter enables learning the wanted embeddings and relevant hyperparameters as [11]

$$\{\mathbf{X},\hat{\boldsymbol{\theta}},\hat{\sigma}_\varepsilon^2\} = \underset{\boldsymbol{\mathcal{X}},\boldsymbol{\theta},\sigma_\varepsilon^2}{\arg\min} -\log p(\mathbf{A}|\boldsymbol{\mathcal{X}};\boldsymbol{\theta},\sigma_\varepsilon^2) - \log p(\boldsymbol{\mathcal{X}}) \tag{3}$$

where the estimated embeddings in (3) turn out to further enjoy maximum-a-posteriori optimality [11].

Remark 1. In addition to low-dimensional nodal embeddings, (3) also learns kernel hyperparameters that specify a similarity function among embeddings. Both will be of benefit to the distribution comparison task that will emerge in

[1] Although here we will rely on the prototypical GPLVM, several variants thereof exist; see e.g., [22].

the ensuing section, which deals with the development of our novel HOLP approach.

Remark 2. The Gaussian pdfs in (3) involve inversion of $|\mathcal{V}| \times |\mathcal{V}|$ covariance matrices. If the cubic complexity of such inversions cannot be afforded, low-complexity approximants are available using either inducing points [12] or random spectral features [7].

4. MMD FOR HOLP

Our idea is to view HOLP as a distribution similarity assessment in the space of embeddings. Intuitively, we expect the formation of a hyperedge to be less likely when the embeddings of its constituent vertices are less similar in distribution. More concretely, let $e := \{v_1, \ldots, v_k\}$ be a candidate hyperedge, and $\mathcal{S} := \{\mathbf{x}_{v_1}, \ldots, \mathbf{x}_{v_k}\}$ the set of associated nodal embeddings. Consider now an arbitrary partitioning of e into two sets π_A and π_B, with $e = \pi_A \cup \pi_B$ and $\pi_A \cap \pi_B = \varnothing$. Let also $\mathcal{S}_A := \{\mathbf{x}_v \mid v \in \pi_A\}$, and similarly for $\mathcal{S}_B$ be the corresponding sets of embeddings. We are interested in comparing the distributions from which the sample sets $\mathcal{S}_A$ and $\mathcal{S}_B$ are drawn from. Although there are multiple ways of partitioning $\mathcal{S}$ into two sets, let us for now focus on comparing an arbitrary pair of sets, and defer the discussion on how comparisons across partitions can be leveraged to yield a score for the candidate hyperedge e.

Aiming to assess similarity of distributions, we will rely on the maximum mean discrepancy (MMD) metric [5, 6]. This (multivariate) integral probability metric is known to perform well even when the sample sets (here $\mathcal{S}_A$ and $\mathcal{S}_B$) are relatively small [4]. For a space of functions $\mathcal{F}$, the MMD between pdfs p and q with respect to $\mathcal{F}$ is defined as

$$\text{MMD}(p, q, \mathcal{F}) := \sup_{f \in \mathcal{F}} \; \mathbb{E}_p[f(X_p)] \; - \; \mathbb{E}_q[f(X_q)] \tag{4}$$

where the random variables X_p, X_q are drawn from p and q, respectively, and with sup denoting the supremum. We can now turn our attention to the space $\mathcal{F}$. A reasonable requisite is for (4) to define a metric. This in turn poses the requirement that $\text{MMD}(p, q, \mathcal{F}) = 0$ if and only if $p = q$. It can be shown that choosing $\mathcal{F}$ to be a unit ball in a universal reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, namely $\mathcal{F} := \{f | \|f\|_{\mathcal{H}} \leq 1\}$, satisfies this requirement [6]. We will adhere to this choice hereafter. Selecting $\mathcal{F}$ amounts to selecting κ; see e.g., [19]. We will revisit the selection of κ in the sequel, but for now let us suppose that κ is given.

Consider now that the sets $\mathcal{S}_A$ and $\mathcal{S}_B$ comprise samples drawn from some arbitrary pdfs p and q, respectively. An empirical estimate of $\text{MMD}(p, q, \mathcal{F})$ is given by [6]

$$\text{MMD}(\mathcal{S}_A, \mathcal{S}_B, \mathcal{F}) = \Bigg(\frac{1}{|\pi_A|^2} \sum_{i \in \pi_A} \sum_{j \in \pi_A} \kappa(\mathbf{x}_i, \mathbf{x}_j) \tag{5}$$

$$- \frac{2}{|\pi_A||\pi_B|} \sum_{i \in \pi_A} \sum_{j \in \pi_B} \kappa(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{|\pi_B|^2} \sum_{i \in \pi_B} \sum_{j \in \pi_B} \kappa(\mathbf{x}_i, \mathbf{x}_j) \Bigg)^{\frac{1}{2}}.$$

It can be shown that the estimator in (5) converges at a rate of $\mathcal{O}(1/\sqrt{|\pi_A| + |\pi_B|})$ to $\text{MMD}(p, q, \mathcal{F})$ [6]. Notice that (5) explicitly highlights the reliance of the probability metric on the kernel choice. To better illustrate this dependence, consider the squared exponential automatic relevance determination (SE-ARD) kernel [23]

$$\kappa(\mathbf{x}, \mathbf{x}') \propto \exp\left(-\frac{1}{2} \sum_{m=1}^{d} \frac{(x_m - x'_m)^2}{\sigma_m^2} \right) \tag{6}$$

that is widely used in the context of GPs, and it can be shown to be universal [16]. We will rely on the SE-ARD kernel hereafter. If we let $\sigma_1 = \sigma_2 = \ldots = \sigma_d := \sigma$, and have $\sigma \to 0$, the MMD tends to zero (cf. (5)) regardless of the locations of the embeddings contained in the sets. More generally, it can be shown that the convergence properties of the empirical MMD depend on the choice of the kernel; see [6, Thm. 8] for rigorous statements.

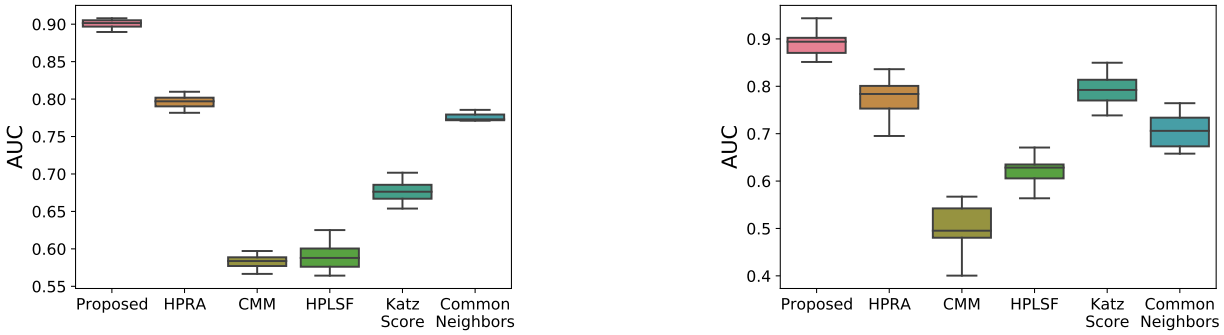
It should be evident that choosing the kernel appropriately is critical to the quality of the probability metric obtained. The kernel selected by the GPLVM is a natural choice. We will thus set $\kappa \equiv \kappa_{\hat{\boldsymbol{\theta}}}$ hereafter, where $\boldsymbol{\theta} := [\sigma_1^2, \ldots, \sigma_d^2]$ for the SE-ARD kernel.

We can now introduce our scoring function for candidate hyperedges. However, we first need to formally describe the assignment of samples to sets. Returning to our candidate hyperedge $e = \{v_1, \ldots, v_k\}$, let $\mathbf{g} := [g_{v_1}, \ldots, g_{v_k}]$ be the set membership indicator vector, that is $g_{v_l} = 0$, if $v_l \in \pi_A$, and $g_{v_l} = 1$, if $v_l \in \pi_B$. Clearly, different values of $\mathbf{g}$ result in different sample sets. To make the connection explicit in our notation, let $\mathcal{S}_{A_\mathbf{g}}$ denote the set $\mathcal{S}_A$ under the assignment described by $\mathbf{g}$, and similarly for $\mathcal{S}_{B_\mathbf{g}}$. By letting $|\mathcal{S}_{A_\mathbf{g}}| = \lfloor \frac{k}{2} \rfloor := k_A$, we will consider splits of (roughly) equal size.

There are $\tilde{k} := \binom{k}{k_A}$ different set membership indicator vectors $\mathbf{g}^{(1)}, \ldots, \mathbf{g}^{(\tilde{k})}$. Using the latter, our score for the candidate hyperedge e is given by

$$S(e) = -\frac{1}{\tilde{k}} \sum_{l=1}^{\tilde{k}} \text{MMD}(\mathcal{S}_{A_{\mathbf{g}^{(l)}}}, \mathcal{S}_{B_{\mathbf{g}^{(l)}}}, \mathcal{F})$$

where the negative sign is introduced to conform with the convention of higher scores being assigned to more likely candidates. In a nutshell, our score reflects the average similarity (negative of 'distance') across the possible splits of observations into pairs of sets.



(a) (b)

Fig. 1: AUCs for hyperedge prediction on the (a) `tags-ask-ubuntu` and (b) `Email-Enron` datasets. Comparison of the proposed HOLP approach against HPRA [9], CMM [25], HPLSF [24], Katz index and number of common neighbors.

5. NUMERICAL TESTS

In order to assess the performance of our novel approach, tests were performed on a) the first $|\mathcal{V}| = 200$ vertices of the `tags-ask-ubuntu` dataset; b) the `Email-Enron` dataset and c) the first $|\mathcal{V}| = 200$ vertices of the `Email-Eu` dataset. Regarding a), nodes correspond to tags and each hyperedge connects the tags associated with a single question on the Stack Exchange website; and with respect to b) and c), nodes correspond to employees and a hyperedge connects employees participating in a single (multi-recipient) email exchange; see [2] for detailed description of these datasets.

A number of competing HOLP alternatives were considered, including coordinated matrix minimization (CMM) [25], that is perhaps the most representative method of its class; the recently proposed hyperedge prediction using resource allocation (HPRA) method [9]; the supervised hyperlink prediction using latent social features (HPLSF) approach [24]; as well as higher-order generalizations of popular link prediction scores, such as the Katz index and the number of common neighbors; see also [25] for a detailed description.

With regards to the proposed approach, the dimensionality of the embeddings was $d = 10$, and the SE-ARD kernel was used. For CMM, all hyperparameters were set as per [25], and we report the best results across the embedding dimensionalities considered therein, that is $\{10, 20, 30\}$.

As in e.g., [15], the set of unobserved hyperedges $\mathcal{E}_u$ was obtained by (randomly) removing 10% of the hyperedges from each dataset. In order to assess hyperedge prediction performance, our candidate set $\mathcal{E}_c$ in the testing phase should also include 'absent hyperedges.' It is worth stressing that $\mathcal{E}_c$ is used here only for evaluation purposes. We relied on the clique negative sampling (CNS) scheme of [17], which is a higher-order counterpart of the widely used approach of [14], to generate the 'absent hyperedge' set $\mathcal{E}_a$ with $|\mathcal{E}_a| = 2|\mathcal{E}_u|$. Our test set is finally obtained as $\mathcal{E}_c = \mathcal{E}_u \cup \mathcal{E}_a$.

The area under the curve (AUC), that is also known as receiver operating characteristic curve, was used as performance metric for the hyperedge prediction task; see also e.g., [15]. The results are depicted using box plots in Figs. 1 and 2, and correspond to the distribution of AUCs obtained across 11 trials. The lower, middle, and upper horizontal lines of each box correspond to the value of AUC below which 25%, 50% and 75% of the trials resulted in. The 25% line is also known as the first quartile, the 50% line as the second quartile (median), and the 75% line as the third quartile. The horizontal lines below (above) each box correspond to the minimum (maximum) AUC value across trials.

These boxes demonstrate that the novel HOLP approach achieves the highest AUC in all three datasets. The proposed approach not only outperforms alternatives at predicting true hyperedges, but also it does so consistently across all datasets. This can be attributed to the adaptation abilities that the learnable nature of the metric brings.

6. CONCLUSIONS

The present work introduced a novel method for higher-order link prediction, that approaches this challenging task from a distribution similarity viewpoint. A learnable probability metric is introduced for assessing the similarity across nodal embeddings, that in effect yields a score for the presence of candidate hyperedges. By combining the merits of GPLVMs with MMD metrics an expressive model emerges, without the need for assumptions on hyperedge formation. Tests on benchmark datasets demonstrated the superior performance of the novel approach relative to existing alternatives.

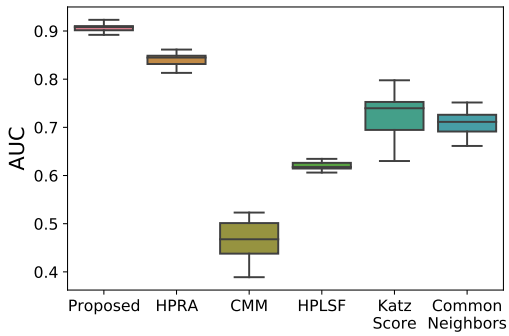


Fig. 2: AUCs for HOLP on the `Email-Eu` dataset.

7. REFERENCES

[1] S. Barbarossa and S. Sardellitti, "Topological signal processing over simplicial complexes," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2992–3007, 2020.

[2] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, "Simplicial closure and higher-order link prediction," *Proc. Natl. Acad. Sci.*, vol. 115, no. 48, pp. E11 221–E11 230, 2018.

[3] C. Berge, *Hypergraphs: Combinatorics of Finite Sets*. North-Holland, 1989, vol. 45.

[4] J. Feydy *et al.*, "Interpolating between optimal transport and MMD using Sinkhorn divergences," in *Proc. of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019, pp. 2681–2690.

[5] R. Fortet and E. Mourier, "Convergence de la répartition empirique vers la répartition théorique," in *Annales scientifiques de l'École Normale Supérieure*, vol. 70, no. 3, 1953, pp. 267–285.

[6] A. Gretton *et al.*, "A kernel two-sample test," *J. of Machine Learning Res.*, vol. 13, no. 1, pp. 723–773, 2012.

[7] G. V. Karanikolas, Q. Lu, and G. B. Giannakis, "Online unsupervised learning using ensemble Gaussian processes with random features," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3190–3194.

[8] E. D. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*. Springer, 2009.

[9] T. Kumar, K. Darwin, S. Parthasarathy, and B. Ravindran, "HPRA: Hyperedge prediction using resource allocation," in *Proc. of 12th ACM conference on Web Science*, 2020, pp. 135–143.

[10] H. Kwak and H. B. K. Jung, "Subgraph representation learning with hard negative samples for inductive link prediction," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4768–4772.

[11] N. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *Journal of Machine Learning Research*, vol. 6, no. 60, pp. 1783–1816, 2005.

[12] N. D. Lawrence, "Learning for larger datasets with the Gaussian process latent variable model," in *Artificial intelligence and statistics*, 2007, pp. 243–250.

[13] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proc. of the 12th ACM International Conference on Information and Knowledge Management*, 2003, pp. 556–559.

[14] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proc. of ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*, 2010, pp. 243–252.

[15] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, 2011.

[16] C. A. Micchelli, Y. Xu, and H. Zhang, "Universal kernels." *Journal of Machine Learning Research*, vol. 7, no. 95, pp. 2651–2667, 2006.

[17] P. Patil, G. Sharma, and M. N. Murty, "Negative sampling for hyperlink prediction in networks," in *Proc. of the 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2020, pp. 607–619.

[18] M. T. Schaub *et al.*, "Signal processing on higher-order networks: Livin' on the edge... and beyond," *Signal Processing*, vol. 187, p. 108149, 2021.

[19] B. Schölkopf and J. A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[20] Y. Tao, Y. Li, and Z. Wu, "Temporal link prediction via reinforcement learning," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3470–3474.

[21] K. Tu, P. Cui, X. Wang, F. Wang, and W. Zhu, "Structural deep embedding for hyper-networks," in *Proc. of AAAI Conference on Artificial Intelligence*, no. 1, 2018.

[22] K. Watanabe, K. Maeda, T. Ogawa, and M. Haseyama, "Summarizing data structures with gaussian process and robust neighborhood preservation," in *Proc. of ECML-PKDD*, 2022.

[23] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*. MIT press, 2006.

[24] Y. Xu, D. Rockmore, and A. M. Kleinbaum, "Hyperlink prediction in hypernetworks using latent social features," in *Proc. of Discovery Science*, 2013, pp. 324–339.

[25] M. Zhang, Z. Cui, S. Jiang, and Y. Chen, "Beyond link prediction: Predicting hyperlinks in adjacency space," in *Proc. of AAAI Conf. on Artificial Intelligence*, 2018.

[26] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Proc. of Advances in Neural Information Processing Systems*, 2006, pp. 1601–1608.