Link Prediction in Complex Hyper-Networks Leveraging HyperCentrality.

NANDINI, YV, TANGIRALA, Jaya Lakshmi <http://orcid.org/0000-0003-0183-4093>, KRISHNA ENDURI, Murali <http://orcid.org/0000-0002-9029-2187> and JILANI, Zairul

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 January 2025, accepted 12 January 2025, date of publication 15 January 2025, date of current version 22 January 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3530245

RESEARCH ARTICLE

Link Prediction in Complex Hyper-Networks Leveraging HyperCentrality

Y. V. NANDINI[1], T. JAYA LAKSHMI[1,2], (Member, IEEE), MURALI KRISHNA ENDURI[1], (Member, IEEE), AND MOHD ZAIRUL MAZWAN JILANI[2]

[1]Department of Computer Science and Engineering, SRM University-AP, Amaravati, Andhra Pradesh 522502, India
[2]Department of Computing, Sheffield Hallam University, S1 1WB Sheffield, U.K.

Corresponding author: T. Jaya Lakshmi (j.tangirala@shu.ac.uk)

ABSTRACT In complex networks, predicting the formation of new connections, or links, within complex networks has been a central challenge, traditionally addressed using graph-based models. These models, however, are limited in their ability to capture higher-order interactions that exist in many real-world networks, such as social, biological, and technological systems. To account for these multi-node interactions, hyper-networks have emerged as a more flexible framework, where hyperedges can connect multiple nodes simultaneously. Traditional link prediction methods often treat all common neighbors equally, overlooking the fact that not all nodes contribute uniformly to the formation of future links. Each node within a network holds a distinct level of importance, which can influence the likelihood of link formation among its neighbors. To address this, we introduce a link prediction approach leveraging hypercentrality measures adapted from traditional centrality metrics such as degree, clustering coefficient, betweenness, and closeness to capture node significance and improve link prediction in hyper-networks. We propose the Link Prediction Based on HyperCentrality in hyper-networks (LPHC) model, which enhances traditional common neighbor and jaccard coefficient of hyper-network frameworks by incorporating centrality scores to account for node importance. Our approach is evaluated across multiple real-world hyper-networks datasets, demonstrating its superiority over traditional link prediction methods. The results show that link prediction in hypercentrality-based models, particularly those utilizing hyperdegree and hyperclustering coefficients for common neighbor and jaccard coefficent approaches in hyper-networks, consistently outperform existing methods in terms of both F1-score and Area Under the Precision-Recall Curve (AUPR), offering a more precise understanding of potential link formations in hyper-networks. The proposed LPHC model consistently outperforms the existing HCN and HJC models across all datasets, achieving an overall improvement of 69% compared to HCN and 68% compared to HJC.

INDEX TERMS Hyper-networks, link prediction, centrality measures.

I. INTRODUCTION

In network science, predicting the formation of connections within a network presents a significant research challenge. Traditionally, this problem has been addressed through graph-based models, where nodes represent entities and edges denote pairwise relationships. However, many real-world networks exhibit more intricate, higher-order interactions that cannot be fully captured by pairwise connections alone. Such networks are prevalent in domains like social, biological, and technological systems [1], where interactions often involve multiple entities simultaneously. To model the complexity of these interactions, hyper-networks have emerged as a powerful extension of traditional graphs, where hyperedges connect multiple nodes at once, offering a more comprehensive framework for representing complex relationships. In complex networks, several critical tasks pose significant challenges, such as centrality measures, influence maximization, community detection, and link prediction, all of which have been extended to the domain of

The associate editor coordinating the review of this manuscript and approving it for publication was Hocine Cherifi.

hyper-networks. This paper focuses specifically on the task of link prediction in hyper-networks efficiently utilizing hypercentrality measures. Link prediction [2] in hyper-networks [3] introduces unique challenges due to the multidimensional nature of connections.

Unlike simple graphs, where the prediction task focuses on pairs of nodes, one can predict future hyperedges from hyper-networks, which may involve several nodes at once. However, many real-world complex networks require pair-wise interaction prediction. In a collaborative research hyper-network [4], nodes represent researchers and hyperedges represent groups of -authors who published a paper together. In such networks, recommending meaningful collaborations to a specific author involves predicting pairwise interaction of that author with others. Although they may occasionally work with larger groups, these broader collaborations are less frequent and less stable than their core pairwise relationship. Similarly, in biological hyper-networks [5] like protein-protein interaction (PPI) hyper-networks, pairwise interactions are essential for cellular processes. Direct interactions between two proteins, such as enzyme-substrate binding are critical for functions. While hyperedges capture multi-protein complexes, core biological functions often rely on stable pairwise protein interactions. Hyperedge predictions, though useful, may miss these fundamental direct interactions. Therefore, we intend to predict future interaction between two nodes taking the hypernetwork which provides a rich input.

Chen et al. [6] provided a comprehensive survey on hyperlink prediction, classifying various methods available for link prediction in hyper-networks, including similarity based approach, probabilistic models, matrix-based approaches, and deep learning techniques. In this work, we focus specifically on local similarity measures, which leverage the immediate neighborhood and structural features of nodes to predict future links in hypergraph. In a hyper-network, local similarity measures typically evaluate how likely two or more nodes are to be part of the same hyperedge based on their shared connections. These measures aim to capture the likelihood that nodes, which share neighbors in the hypergraph, will be co-participants in a hyperedge in the future. Some common adaptations of local similarity-based measures for hypergraphs include: common neighbor, Jaccard coefficient along with many others. Nasiri et al. [7] extends the Local Random Walk (LRW) method to multiplex networks, proposing the Multiplex Local Random Walk (MLRW) to predict links by leveraging inter-layer and intra-layer structural information. Berahmand et al. [8] introduces SDAC-DA, a semi-supervised deep clustering method for attributed networks that combines dual autoencoders and end-to-end optimization to enhance clustering performance. Sheikhpour et al. [9] proposes HSDAFS, a semi-supervised feature selection method using hypergraph Laplacian-based discriminant analysis and mixed $l_{2,1}$-norm regularization to capture high-order relationships and enhance feature sparsity. Shen et al. [10] explores the synchronization of fractional uncertain reaction-diffusion complex networks using an adaptive scheme and output-strict passivity lemma.

In similarity based link prediction approaches for hyper-networks, the focus is primarily on the neighborhood structure, where common neighbors between sets of nodes are typically considered to have equal significance [11]. However, this assumption neglects the fact that common neighbors contribute differently to the formation of future hyperedges. The multidimensional nature of hyperedges, which can connect multiple nodes simultaneously, introduces varying levels of influence among nodes. As a result, common neighbors do not equally impact the likelihood of new hyperedges forming. To address this complexity, node importance, quantified through centrality scores can be integrated into the common neighborhood framework, offering a more refined approach to link prediction.

The study presented in this work aims to consider the centrality scores [12] of common neighbors to improve the accuracy of predicting future links. Before introducing the proposed method, we review existing centrality measures for link prediction in graphs [13] and outline their adaptations to the context of hyper-networks.

The structure of this paper is as follows: Sec.II review related work and discuss existing methods related to link prediction and centrality measures, both in graphs and hyper-networks. In Sec.III, we present the methodology, including the hypercentrality measures used, the calculation of average hypercentrality, and the definition of similarity scores. Sec.IV outlines the experimental setup, including hyper-networks sampling and the datasets used, and presents the evaluation results. Sec.V provides a detailed discussion of the results. Finally, Sec.VI concludes the paper and proposes potential directions for future research.

Table. 1 outlines the notations used throughout this work.

The following section briefly describes the related work.

II. RELATED WORK

This section presents the essential technical background, definitions, and relevant information for this work. The definitions are outlined as follows.

Definition 1 (Hyper-Network): *A complex hyper-network [14] is denoted as* $H = (V, E)$, *where* $V = \{v_1, v_2, ..v_n\}$ *represents a set of* $|V|$ *nodes, and* $E = \{e_1, e_2, ..e_m\}$ *represents a set of m hyperedges. Each hyperedge* E_i *is a non-empty subset of the power set of* V, *i.e.,* $E_i \in (2^V - \phi)$

Unlike standard graphs, where an edge connects exactly two nodes, a hyperedge in a hyper-network can connect any number of nodes. In this work, the terms ''hyper-network'' and ''hypergraph'' are used synonymously. Similarly, the terms ''edges,'' ''links,'' and ''connections'' are treated as interchangeable throughout this work.

A hyper-network with V vertices and E hyperedges can be represented using an incidence matrix $I[H]$, where the

TABLE 1. Notations utilized in this research.

Notation	Description
H	Hyper-Network (Also called as Hyper$_graph$)
V	Set of Nodes
E	Set of Hyperedges
i,j,u,v,x	Nodes within the hyper-networks
$\Gamma(u), \Gamma(v)$	Neighbors of nodes u, v
I(H)	Incidence Matrix of hyper-networks
$e_1 e_2$	Hyperedges
s	size of hyperedge h
N	Common Neighbors between the nodes
$\mathbb{HC}$	HyperCentrality
HCN	Common Neighbor in Hyper-network
HJC	Jaccard Coefficeint in Hyper-network
$\mathbb{HC}_d$	HyperDegree Centrality
$\mathbb{HC}_{cc}$	HyperClustering Coefficient
$\mathbb{HC}_b$	HyperBetweenness Centrality
$\mathbb{HC}_{cl}$	HyperCloseness Centrality
$A\mathbb{HC}_{\mathbb{HC}}$	AverageHyperCentrality
$LPHC^{CN}_{\mathbb{HC}}$	Link Prediction based on HyperCentrality using Common Neighbor
$LPHC^{JC}_{\mathbb{HC}}$	Link Prediction based on HyperCentrality using Jaccard Coefficient
AUPR	Area Under the Precision-Recall Curve

rows correspond to the vertices and the columns represent the hyperedges. This matrix captures the relationship between the vertices and hyperedges in the hyper-networks. The incidence matrix of a hyper-networks is mathematically defined as follows:

Definition 2 (Incidence Matrix I(H)): Given a hyper-network $H = (V, E)$, the incidence matrix [14] is a matrix of size $|V| \times |E|$. The rows of the matrix correspond to the vertices $|V|$, and the columns correspond to the hyperedges $|E|$. Each entry I_{ve} in the incidence matrix $I(H)$ is defined as follows:

$$I_{ve} = \begin{cases} 1 : \textit{if node v is part of hyperedge e} \\ 0 : \textit{otherwise} \end{cases} \quad (1)$$

For example, consider a hyper-network from Fig. 1 with nodes $V = \{A, B, C, D, E, F, G\}$ and hyperedges $E = \{e_1, e_2, e_3, \text{and} e_4\}$. Here we define the hyperedges as follows:

- Hyperedge e_1 connects nodes A, B and D.
- Hyperedge e_2 connects nodes B and C.
- Hyperedge e_3 connects nodes A, B, C, and E.
- Hyperedge e_4 connects nodes A, D, E, F and G.

The incidence matrix corresponding to Fig. 1 is constructed by assigning a value of 1 to entries where a node is part of a hyperedge, and 0 otherwise. This matrix is depicted in Fig. 2.

A. LINK PREDICTION IN HYPER-NETWORKS

Link prediction in hyper-networks involves predicting future hyperedges among multiple nodes, where relationships involve more than two nodes, unlike traditional graphs. Hyperedges connect several nodes at once, and the goal is to identify which groups are likely to form new connections based on the hyper-networks's structure. Kumar et al. [15] addresses the challenge of hyperedge prediction, a complex problem with applications in fields like social networks

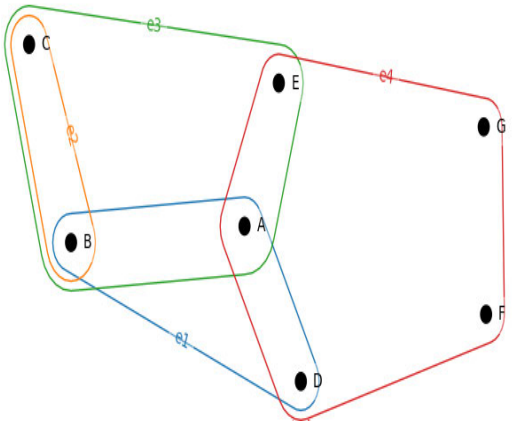

FIGURE 1. Hyper-network with 7 nodes (A, B, C, D, E, F and G) and 4 hyperedges (e_1, e_2, e_3 and e_4).

$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \end{array} \begin{array}{c} \begin{array}{cccc} e_1 & e_2 & e_3 & e_4 \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

FIGURE 2. Incidence matrix of hyper-network $I(H)$.

and metabolic systems. The authors proposed HPRA (Hyperedge-Prediction using Resource-Allocation), a novel algorithm that predicts hyperedges of any size without requiring a predefined candidate set. Extensive experiments demonstrate that HPRA achieves significant improvements over existing methods, effectively recovering missing and predicting future hyperedges.

Though hyper-networks capture complex relationships involving multiple entities through hyperedges, pair-wise interactions remain fundamental in several real-world scenarios. For instance, in academic collaborations, papers often have multiple authors forming hyperedges. Pair-wise interactions model direct collaborations between author pairs, providing insights into author-specific partnerships, influence, and local collaboration patterns. Similarly, in Recommender Systems such as Netflix or Spotify, users interacting with multiple items (movies, songs) form hyperedges. However, the product recommendations to the users are pair-wise relationships between users and items. One more motivating scenario appears in Epidemic Modeling. In disease transmission, groups of individuals participating in shared events like weddings or conferences create hyperedges. But,

pair-wise interactions are significant in identifying who infected whom, supporting effective intervention strategies. Therefore, we aim to develop a framework for pair-wise interaction prediction in hyper-networks.

Our work is majorly inspired from the work of [16] which focuses on local similarity measure. Local similarity-based link prediction leverages the immediate neighborhoods of nodes to compute similarity scores, typically based on shared connections. The authors propose link prediction measures in hypergraphs directly, without converting them to pairwise graphs. The advantage of this method is thhat it preserves the inherent structure of hypergraphs, thereby maintaining the original complexity and information integrity, unlike traditional approaches that rely on transforming hypergraphs into simple graphs. The work in [16] defines two measures *HCN* (Common Neighbor in Hyper-network), *HJC* (Jaccard Coefficient in Hyper-network), which are briefed below.

a) **Common Neighbor in Hyper-network** (*HCN*)**:** The authors [16] generalized common neighbors to hyper-networks by calculating the average of the pairwise *CN* indices between the nodes in each hyperlink. The Link Prediction in hyper-networks using Common Neighbors *HCN* is specified in the below Eq.2.

$$HCN(u, v) = \frac{2}{s_{e_1} * s_{e_2}} \sum_{u \in e_1; v \in e_2; x \in e_1 \cap e_2} |x| \quad (2)$$

where, u, v are nodes with in the Hyper-network, e_1 and e_2 are hyperedges, and s_e is the size of the hyperedge.

b) **Jaccard Coefficient in Hyper-network** (*HJC*)**:** The authors of [16] extended Jaccard Coefficient to hyper-networks setting. The *HJC* is a normalized version of *HCN*, computed by dividing the number of neighbors shared by both nodes by total number of distinct neighbors of either node. Mathematically, *HJC* is expressed as:

$$HJC(u, v) = \frac{HCN(u, v)}{|\Gamma(u) \cup \Gamma(v)|} \quad (3)$$

The Common Neighbor (CN) measure is simple and computationally efficient, making it effective for capturing local structural information in dense networks or tightly-knit communities. However, it has limitations, as it treats all common neighbors equally, failing to account for node importance or influence. CN is less effective in sparse networks or those with long-range dependencies, as it does not consider global structural information. The Jaccard Coefficient (JC), in contrast, provides a normalized similarity score by balancing common neighbors with the union of unique neighbors, making it more robust in networks with diverse connectivity patterns or imbalanced neighborhood sizes. Despite this, JC struggles in sparse networks where large unions lead to low scores and is sensitive to small intersections, reducing its effectiveness in weakly clustered networks. Additionally, JC is computationally more intensive than CN due to the union calculation. To address these challenges, we plan to extend these measures to incorporate global structural properties, improving their predictive accuracy across diverse network scenarios.

In all these measures, all the common neighbors are treated equal. However, these common neighbors do not contribute uniformly to the formation of future links. Each node within a network holds a distinct level of importance, and this significance is believed to influence the likelihood of link formation among neighbors. Accordingly, this study aims to incorporate node significance, represented by centrality scores, to enhance the common neighborhood framework. We begin by briefly reviewing existing centrality measures in graphs, discuss how these measures are adapted to the hyper-networks context. We present our proposal of how this information can be utilized to predict future hyperlinks in the following section.

B. HYPERCENTRALITY MEASURES

In graph theory, centrality measures are commonly used to assess the importance or influence of individual nodes within a network. Centrality in traditional graphs measures the importance of a node based on its connections, focusing on pairwise relationships. Four of the most widely applied centrality measures are degree centrality [17], clustering coefficient [18], [19], betweenness centrality [20], and closeness centrality [21]. These measures offer valuable insights into the structural role of nodes and their contributions to the overall network dynamics. Degree centrality, one of the simplest centrality measures, is defined as the number of direct connections a node has, with higher values indicating nodes that can spread information or exert influence more effectively. The clustering coefficient measures the degree to which a node's neighbors are interconnected, reflecting the node's participation in closely knit groups. Betweenness centrality quantifies the extent to which a node lies on the shortest paths between other nodes, highlighting its role in facilitating information flow across the network. Closeness centrality, on the other hand, is the inverse of the total shortest path distances from a node to all other nodes, indicating how efficiently a node can access other parts of the network. However, these measures do not capture more complex group interactions or multi-node connections found in real-world networks. Hypercentrality measures in hyper-networks address this limitation by considering hyperedges that can connect multiple nodes simultaneously.

Hypercentralities allow for the identification of key nodes and relationships within the more complex topology of hyper-networks, where influence, connectivity, and centrality are defined by higher-order interactions rather than just pairwise connections. Roy et al. [22] proposes utilizing Shapley value-based centrality within the framework of node centrality, while preserving hypergraph structure. Li et al. [23] introduces a novel link prediction approach for social networks utilizing hypergraphs, which effectively capture both pairwise and higher-order relationships, thereby

enhancing the accuracy and performance of link prediction tasks. Ihsan et al. [24] proposed entropy-based centrality measures for hyper-networks, utilizing local similarities to capture centralities.

Aksoy et al. [25] extend traditional graph metrics to s-metrics (high-order hypergraph walks) for hypergraphs by leveraging their s-connected components. This is achieved by first computing the s edge-adjacency matrix, which is then used to construct the corresponding graph representation of the hypergraph. Existing graph metrics can subsequently be applied to this graph representation. Essentially, the authors generate an s-line graph corresponding to the hypergraph, enabling the application of standard graph-based methods. By default, the parameter s is set to 1, though it can be fine-tuned to better suit specific applications.

Centrality measures in graphs and hypergraphs differ due to the structural properties of their connections. While graphs model pairwise relationships (edges), hypergraphs capture higher-order interactions (hyperedges), enabling richer representations of complex systems. Table. 2 contains a detailed comparison of centrality measures in both contexts.

In this study, we intend to utilize hyper-centrality measures; therefore, a brief discussion on centrality measures within hyper-networks is presented below.

a) **HyperDegree Centrality ($\mathbb{HC}_d$):** The hyperdegree of a node is the number of hyperedges the node is part of. Unlike traditional graphs, where edges connect two nodes, hyperedges in hypergraphs can connect multiple nodes simultaneously. However, each hyperedge is counted only once for each node it connects, regardless of how many nodes are part of the hyperedge. The underlying notion is that a node's influence in a hyper-network increases if it participates in more hyperedges, as this indicates a broader range of connections and involvement in various relationships within the network. The Degree Centrality of a node u in a hyper-networks is mathematically expressed as:

$$\mathbb{HC}_d(u) = \sum_{v=1}^{e} I_{ue} \tag{4}$$

where I is the incidence matrix of hyper-networks. I_{ue} represents the existence of node u in hyperedge e. The total Degree Centrality of node u is obtained by summing the edges in which it is participating.

b) **HyperClustering Coefficient ($\mathbb{HC}_{cc}$):** Clustering Coefficient in hyper-networks is a measure of the tendency of nodes to form tightly connected groups, or clusters, within the hyper-networks [25]. Unlike traditional graphs, where clustering focuses on the likelihood that a node's neighbors are also connected to each other, in hyper-networks, clustering involves assessing how nodes participate in hyperedges that create group interactions. The idea is to measure how likely it is for two nodes that share a hyperedge to also be connected by other hyperedges. The mathematical definition of clustering coefficient in hyper-networks is given in Eq.5.

$$\mathbb{HC}_{cc}(u) = \frac{\sum_{e_1,e_2} |e_1 \cap e_2|}{\binom{k(u)}{2}} \tag{5}$$

where $k(u)$ is the set of hyperedges containing u, $\binom{k(u)}{2}$ is the total number of possible pairs of hyperedges involving u and $|e_1 \cap e_2|$ is the size of the intersection of hyperedges containing u.

c) **HyperBetweenness Centrality($\mathbb{HC}_b$):** Betweenness Centrality in hyper-networks identifies nodes that act as key connectors within the network [25]. Nodes with a high number of shortest paths passing through them are more important, as they play a critical role in facilitating communication and interactions across different parts of the hyper-networks. The betweenness in hyper-networks can be computed for nodes or edges. In this work, we consider only centrality of nodes. The *HB* for a node u is calculated using the formula:

$$\mathbb{HC}_b(u) = \sum_{i \neq u \neq j} \frac{\sigma_{i,j}(u)}{\sigma_{i,j}} \tag{6}$$

where $\sigma_{i,j}(u)$ represents the number of shortest paths from node i to node j that pass through node u. and $\sigma_{i,j}$ denotes the total number of shortest paths from node i to node j. In traditional graphs, distance is defined as the number of edges in the shortest path between two nodes. However, in hyper-networks, hyperedges can contain multiple nodes, which changes how we measure distance. In its simplest form, the distance between two nodes is defined as the number of hyperedges you must traverse to connect them. If two nodes are in the same hyperedge, the distance is 1. If they are not directly connected by a hyperedge, you need to traverse intermediate hyperedges, and the distance increases.

d) **HyperCloseness Centrality($\mathbb{HC}_{cl}$)** Closeness Centrality in hyper-networks quantifies how near a node is to all other nodes [25]. Unlike in traditional graphs, where distance is usually defined by the shortest path between two nodes, in hyper-networks, this concept is adapted to account for the fact that hyperedges can link several nodes simultaneously. The closeness centrality is computed for nodes or edges. If edge is set to True, it computes the closeness centrality for edges; otherwise, it computes for nodes. And also we can fix the size of the egdes, within the hyper-networks. The *HC* for a node u is calculated using the formula:

$$\mathbb{HC}_{cl}(u) = \frac{|V| - 1}{\sum_{u \neq v \in V} d(u, v)} \tag{7}$$

Here $d(u, v)$ is the distance between nodes u and v, $|V|$ is the total number of nodes within the hyper-networks.

This study aims to adapt the centrality measures to the more intricate structure of hyper-networks, hypothesizing that centrality-based approaches can provide deeper insights

TABLE 2. Centrality measures in graphs Vs hypergraphs.

Centrality measure of node v	Definition in Graph and Hypergraph	Advantage of hypercentrality	Key property preserved
Degree	Graph: The number of edges incident to node v. Hypergraph: Number of hyperedges containing node v	Hyper degree centrality captures the role of v within multi-member interactions.	Local connectivity and influence. Higher participation implies higher centrality, similar to traditional graphs
Clustering Coefficient	Graph: Measures the likelihood that two neighbors of v are themselves connected, forming a triangle. Hypergraph: Percent of v's neighbors that are neighbors of each other.	Computing this in hypergraphs is more efficient because hypergraphs leverage multi-node connectivity in a single step.	Local density and neighborhood cohesion, which highlights how well-connected a node's neighborhood is, whether through pairs (graphs) or groups (hypergraphs).
Betweenness	Graph: Counts the number of shortest paths passing through a node, indicating its role as a bridge. Hypergraph: Paths are generalized to account for multiple routes passing through hyperedges.	Betweenness in hypergraphs accounts for multi-way flows, which influence information propagation more effectively.	Nodes with higher flow through multi-way paths retain high betweenness.
Closeness	Graph: Measures the reciprocal of the sum of shortest path lengths between a node and all others. Hypergraph: Redefining distances because hyperedges can reduce path lenghts by connecting multiple nodes simultaneously.	Nodes connected through hyperedges may have shorter effective distances, accounting for multi-way interactions.	Nodes with more direct access to others remain central.

into the underlying patterns of connectivity. The authors of [13] proposed a novel link prediction method using similarity scores based on average centrality measures on traditional graphs to improve prediction accuracy. This work is inspired from the work of [13] to extend the methodology to hyper-networks. Metrics like degree, clustering coefficient, betweenness, and closeness are adapted to account for these multi-node connections, offering a more comprehensive understanding of node influence in settings with group interactions. By integrating node centrality metrics into the link prediction framework, the proposed method seeks to capture the multifaceted relationships inherent in hyper-networks, thereby enhancing predictive performance. The following section explains the proposed approach in detail.

III. PROPOSED METHOD

This work extends the traditional concept of link prediction in graphs [2] to hyper-networks, with a particular emphasis on hyperedges of size 2. This choice is motivated by the richer and more informative data provided by hyperedges of size 2, enabling more accurate and interpretable outcomes.In this work, we extend this concept to address the problem of link prediction in hyper-networks as follows.

Definition 3 (Link Prediction in Hyper-networks (LPH)): Given a hyper-network $H = (V, E)$, V representing set of vertices, and E denoting set of hyperlinks, the task of LPH involves forecasting the potential appearance of pairwise links that are not currently present in H but are expected to emerge in the future.

In this work, we propose a novel approach of predicting links in hyper-networks leveraging node centrality scores. This approach is termed Link Prediction Based on Hyper-Centrality (*LPHC*). *LPHC* aims to improve the accuracy of link prediction by focusing on common nodes with higher influence within the network structure. We use Common Neighbor as well as Jaccard Coefficient as link prediction measures in this work. Both of these majorly rely on common neighbors.

The *LPHC* algorithm has the following steps:

1) Calculating the HyperCentrality score for each node within the hyper-networks. In this work we use four centrality measures discussed in later Sec.III-A.
2) Subsequently, the average centrality of all nodes is determined as shown in Eq.8.

$$A\mathbb{HC}_{\mathbb{HC}}(H) = \frac{\sum_{u \in V(H)} \mathbb{HC}(u)}{|V|} \tag{8}$$

In Eq.8, $\mathbb{HC}(u)$ denotes the HyperCentrality score of node u, determined using hyper-networks centrality measures provided in the Sec.II-B. In Eq.8, the generalized form of the average HyperCentrality score $A\mathbb{HC}(H)$ can be adapted based on the specific centrality measure used. For instance: If the HyperCentrality measure corresponds to hyperdegree, then apply $A\mathbb{HC}_d(H)$, if it is hyperbetweenness then apply $A\mathbb{HC}_b(H)$ in place of $A\mathbb{HC}_{\mathbb{HC}}(H)$.

3) To predict a potential link between two nodes, u and v, the method initially identifies their common neighbors in the hyper-networks. In the next step, only those common neighbors whose HyperCentrality scores surpass the average centrality are taken into account in computing link prediction scores. Use these common neighbors to compute link prediction scores.

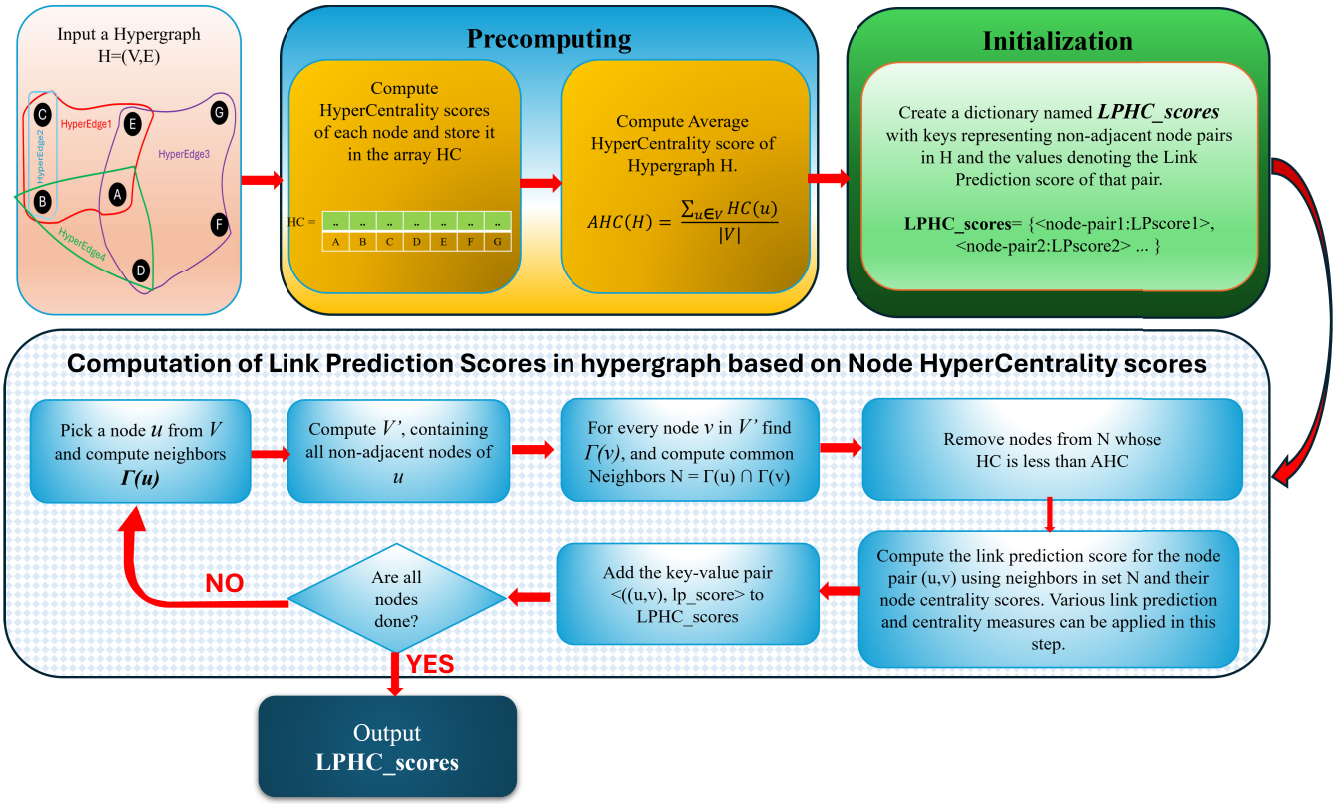

FIGURE 3. ***LPHC*: Link Prediction based on HyperCentralit.**

In this work, we employ two link prediction measures, which are detailed in Sec.III-B.

The proposed method is depicted in Fig. 3 and summarized in Algorithm.1.

A. HYPERCENTRALITY MEASURES

We use various hyper-networks-specific centrality measures, including hyperdegree, hyperbetweenness, hypercloseness, and hyperclustering coefficient, in place of $\mathbb{HC}$. The detailed equations for each of these hypercentrality measures are provided in the Table. 3, illustrating how they are adapted for hyper-networks to effectively capture node influence.

B. LINK PREDICTION MEASURES BASED ON HYPERCENTRALITY

a) **HyperCentrality based Common Neighbor** ($LPHC^{CN}$) : To compute the similarity score between two non-adjacent nodepairs (u, v) in a hyper-networks based on the average centrality, we generalize the concept of common neighbors to hyperedges. The Link Prediction Based Hyper-Centrality for hyper-networks $SCH_{\mathbb{HCM}}$ can be defined in Eq.9:

$$LPHC^{CN}_{\mathbb{HC}}(u, v) = \frac{2}{s_{e_1} * s_{e_2}} \sum_{\substack{u \in e_1; v \in e_2; x \in e_1 \cap e_2 \\ \text{and } \mathbb{HC}(x) \geq A\mathbb{HC}(H)}} |x| \quad (9)$$

where, e_1 and e_2 are hyperedges, and s_e is the size of e, x denotes the common neighbor between hyperedges, $\mathbb{HC}$(x) denotes the hypercentrality score of common neighbor x, $A\mathbb{HC}$(H) is the average hypercentrality which is in defined in Eq.8.

This equation computes the link prediction score by counting the number of common neighbors that has a centrality value greater than or equal to the average centrality of the hyper-network. $\mathbb{HC}$ can be adapted to other HyperCentrality measures such as hyperclustering coefficient $\mathbb{HC}_{cc}$, hyperbetween-ness centrality $\mathbb{HC}_b$, and hypercloseness centrality $\mathbb{HC}_{cl}$, as outlined in rows 2, 3, and 4 of Table. 3, respectively, which leads to the calculation of $LPHC^{CN}$.

b) **HyperCentrality based Jaccard Coefficient** ($LPHC^{JC}$) :

In traditional graphs, the Jaccard Coefficient is calculated based on the neighbors of two nodes, representing the similarity between these nodes as the ratio of their shared neighbors to their total neighbors (union). In hypergraphs, the Jaccard Coefficient is an adaptation of the traditional Jaccard Coefficient used in graphs, but it accounts for the complexity of hyperedges, which can connect multiple nodes simultaneously. For hypergraphs, the Jaccard Coefficient is modified to compare

TABLE 3. The proposed link prediction measure based on HyperCentrality ($LPHC_{\mathbb{HC}}(u, v)$).

	HyperCentrality	$LPHC^{CN}_{\mathbb{HC}}(u, v)$
Degree Centrality ($\mathbb{HC}_d$)	$\mathbb{HC}_d(u) = \sum_{v=1}^{m} I_{uv}$	$LPHC^{CN}_{d}(u, v) = \lvert x \lvert x \in \Gamma(u) \cap \Gamma(v)$ and $\mathbb{HC}_d(x) \geq A\mathbb{HC}_d(H)\rvert$
Clustering Coefficient ($\mathbb{HC}_{cc}$)	$\mathbb{HC}_{cc}(u) = \frac{\sum_{e_1,e_2} \lvert e_1 \cap e_2 \rvert}{\binom{k(u)}{2}}$	$LPHC^{CN}_{cc}(u, v) = \lvert x \lvert x \in \Gamma(u) \cap \Gamma(v)$ and $\mathbb{HC}_{cc}(x) \geq A\mathbb{HC}_{cc}(H)\rvert$
Betweenness Centrality ($\mathbb{HC}_b$)	$\mathbb{HC}_b(u) = \sum_{i \neq u \neq j} \frac{\sigma_{i,j}(u)}{\sigma_{i,j}}$	$LPHC^{CN}_{b}(u, v) = \lvert x \lvert x \in \Gamma(u) \cap \Gamma(v)$ and $\mathbb{HC}_b(x) \geq A\mathbb{HC}_b(H)\rvert$
Closeness Centrality ($\mathbb{HC}_{cl}$)	$\mathbb{HC}_{cl}(u) = \frac{n-1}{\sum_{u \neq v \in V} d(u,v)}$	$LPHC^{CN}_{cl}(u, v) = \lvert x \lvert x \in \Gamma(u) \cap \Gamma(v)$ and $\mathbb{HC}_{cl}(x) \geq A\mathbb{HC}_{cl}(H)\rvert$

the sets of hyperedges that connect two nodes, instead of just pairwise neighbors. In this case, the Jaccard Coefficient between two nodes u and v in a hypergraph is defined as:

$$LPHC^{JC}_{\mathbb{HC}}(u, v) = \frac{|LPHC^{CN}_{\mathbb{HC}}(u, v)|}{|\Gamma(u) \cup \Gamma(v)|} \tag{10}$$

where $\Gamma(u)$ is the set of hyperedges that node u and node v participates in, $LPHC^{JC}_{\mathbb{HC}}(u, v)$ is the number of hyperedges shared by both u and v, and $|\Gamma(u) \cup \Gamma(v)|$ is the total number of hyperedges that contain either u and v, or both. $LPHC^{JC}$ can also be defined using various HyperCentrality measures similar to the ones in Table. 3

The effectiveness of the proposed method is assessed through experimental evaluation on five distinct hypernetworks. Comprehensive details of this evaluation are presented in the subsequent section.

IV. EXPERIMENTAL SETUP

We employed five datasets to demonstrate the effectiveness of the proposed approach, with each dataset sampled from hyper-networks taken from ARB respository ''https://www.cs.cornell.edu/ arb/data/email-Eu/''.

- **National Drug Code Directory** *NDC* **[26]:** In NDC-classes, nodes are class labels assigned to drugs, where each node corresponds to a specific label associated with a drug. Hyperedges are formed by simplices, where each simplex represents a set of nodes (class labels) connected together by a drug.
- **email-Eu:** In the email-Eu dataset, the nodes represent email addresses within a European research institution, and a hyperedge is created by grouping together the sender and all recipients involved in a specific email. ''https://www.cs.cornell.edu/ arb/data/email-Eu/''
- **Drug abuse warning network (DAWN) drugs [26]:** In DAWN, nodes represent the drugs (illicit substances, prescription medications) reported by patients during emergency department visits whereas Hyperedges are the simplices, where each simplex corresponds to a set of drugs used by a patient during a specific visit.
- **cat-edge-geometry-questions:** In this hyper-networks dataset, Nodes correspond to individual geometry-related questions and Hyperedges are created by grouping questions that are conceptually connected, meaning they share common topics, linking multiple questions (nodes) within a single hyperedge.
- **hyperegdes-contact-high-school [27]:** In this nodes represent the people at the high school who were interacting with each other and hyperedges are the maximal cliques of interacting individuals, captured as simplices, where each hyperedge connects all the people who interacted with one another.

The hypernetworks are huge in size. Therefore, we have sampled the networks to reduce the size based on the hyperedge distribution. The hyperedge distribution is defined as a function that gives the number of hyperedges of each possible size (or cardinality). This describes how many hyperedges contain a given number of nodes. Fig. 4 depicts the hyperedge distribution. In the hyperedge distribution visualizations, the x-axis represents the hyperedge size, which corresponds to the number of nodes involved in a single hyperedge. Although the hyperedge size on x-axis can extend beyond 8, we observe that 95% of the nodes participate in hyperedges of size 8 or less. To enhance visualization clarity, the x-axis is therefore restricted to a maximum size of 8, enabling a more focused and interpretable representation

Algorithm 1 ***LPHC***: Link Prediction Based on Hyper-Centrality

Input: $H = (V, E)$: hyper-networks where V is node list and E is hyperedge list.

Output: *LPHC_Scores*: A set of key-value pairs where keys represent non-adjacent node pairs in H and the corresponding values denote the Link Prediction score of the pair.

Initialization: $LPHC\_Scores = \phi$ // *LPHC_scores* contains key-value pairs where keys are the non-adjacent node pairs and values are their LPHC scores.
for every vertex $uinV$ **do**
    Find $\mathbb{HC}(u)$ // Calculate Centrality of node u
end for
Compute $A\mathbb{HC}(H) = \frac{\sum_{u \in V} \mathbb{HC}(u)}{|V|}$ // Average Centrality of hyper-networks H
for every vertex u in V **do**
    Compute the $\Gamma(u)$ // Find the neighbors of node u
    for every vertex $vinV$ **do**
        if $(u, v) \notin E$ **then**
            Compute the $\Gamma(v)$ // Find the neighbors of node v
            Compute $\mathbb{N} = \Gamma(u) \cap \Gamma(v)$
            for every $x \in \mathbb{N}$ **do**
                if $\mathbb{HC}(u) \leq \mathbb{AHC}(\mathbb{H})$ **then**
                    Remove x from $\mathbb{N}$
                end if
            end for
            Calculate *lp_score_uv* using nodes in $\mathbb{N}$ using various link prediction measures discussed in Sec.III-B
            Add $((u, v) : lp\_score\_uv)$ to *LPHC_Scores*
        end if
    end for
end for
    return *LPHC_Scores*
=0

of the most common hyperedge sizes. The y-axis denotes the count of hyperedges for each respective hyperedge size, illustrating how many hyperedges of a given size exist within the dataset. In most datasets, we observe that the majority of hyperedges consist of pairs of nodes (hyperedge size 2), while larger hyperedges (size 3, 4, and beyond) are progressively less frequent. This distribution reflects the dominance of pairwise interactions in hypergraphs.

While our research explores hypergraphs, where hyperedges can connect multiple nodes, we initially concentrate on non-adjacent node pairs (cardinaltiy of hyperedge size 2). This focus is justified by the fact that a significant proportion of the datasets—approximately 60-70% comprises hyperedges that involve exactly two nodes. As demonstrated in the Fig. 4, hyperedge size 2 consistently dominates maximum

TABLE 4. Datasets with nodes and hyperedges after sampling.

Datasets	#Nodes	#Hyperedges
NDC-classes	640	411
email-Eu	745	4868
DAWN	1492	27740
cat-edge-geometry-questions	450	247
hyperedges-contact-high-school	317	781

datasets, suggesting that pairwise interactions are the most prevalent form of connections in these hyper-networks. Given this dominance, it is both efficient and insightful to begin our analysis with hyperedges of size 2.

Sampling the Hyper-Networks: The sampling methods incorporated to the hyper-networks are given below. The edges are first grouped by size, based on the number of nodes they connect. This allows for separate handling of hyperedges of different sizes during the sampling process. A user-controlled fraction of hyperedges from each size group is selected for the final sampled hyper-networks (e.g., 0.5 for 50% sampling). To ensure that the sampled hyper-networks remains meaningful, at least one hyperedge is chosen from each group, and isolated nodes (those no longer connected to any hyperedge) are removed. Given the computational complexity of processing large hyper-networks, this sampling approach reduces the hyper-networks's size, enabling efficient analysis within time and memory constraints, while preserving the key relationships between nodes. Grouping hyperedges by size and sampling proportionally ensures the sampled hyper-networks retains a distribution similar to the original.

The study has been conducted on a PC equipped with an 11th generation Intel(R) Core(TM) i7-8700 CPU, featuring six cores, twelve logical processors, and a base clock speed of 3.20 GHz. The machine ran Windows 10 Education with 16 GB of RAM, and Python was utilized for the investigation.

A. EVALUATION METRICS

To evaluate the performance of the proposed methods, 40% of the network data is reserved for testing, while the remaining 60% is used for training the proposed measures. The effectiveness of the link prediction method, based on hypercentrality measures, is assessed using a range of performance metrics. Although a wide range of metrics is available, this study focuses on the F1-score and the Area Under the Precision-Recall Curve (AUPR). These metrics have been selected due to their ability to provide a comprehensive evaluation of model performance, particularly in the context of imbalanced datasets, where the balance between precision and recall is crucial. The details are outlined below:

- **F1-score:** [28] [29] The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both. It is particularly useful in scenarios

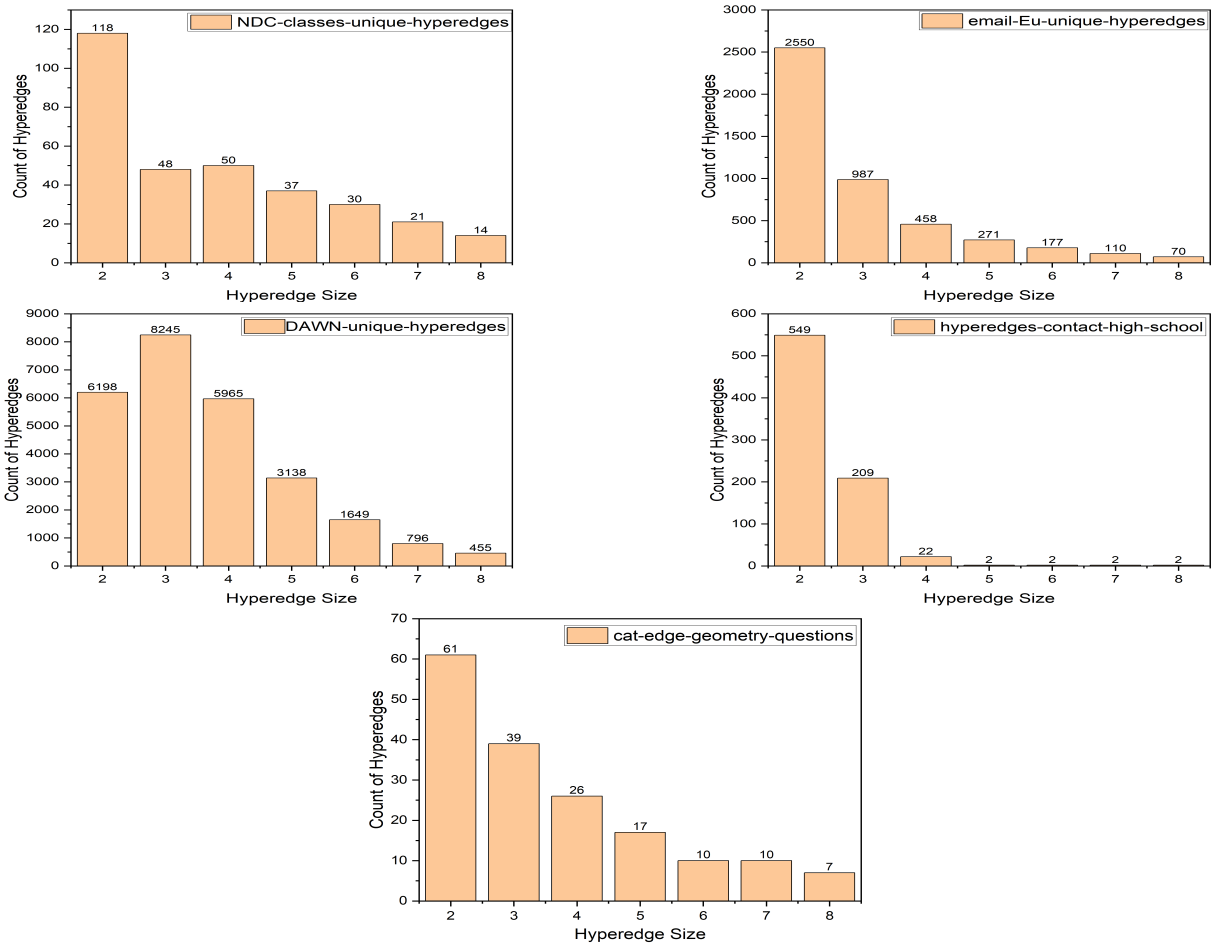

FIGURE 4. Hyperedge distribution across multiple datasets in hyper-networks.

where there is an uneven class distribution or when both false positives and false negatives carry different costs. The F1-score ranges from 0 to 1, with a higher score indicating better performance. It is defined mathematically as:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$

where **Precision** is the ratio of correctly predicted positive instances to the total predicted positives. **Recall** is the ratio of correctly predicted positive instances to the actual positives.

The F1-score is particularly valuable when there is a need to find a balance between precision and recall, such as in imbalanced datasets where one class is much more frequent than another.

- **Area Under the Precision-Recall Curve (AUPR:)** [30] The Area Under the Precision-Recall Curve (AUPR) is a performance metric used primarily for binary classification tasks, particularly in cases with highly imbalanced data where the positive class is much rarer than the negative class. The AUPR quantifies the area under the precision-recall curve, which plots precision on the y-axis and recall on the x-axis at different threshold levels for classifying positive instances. A higher AUPR value indicates better model performance, as it suggests that the model maintains a good trade-off between precision and recall across different thresholds.

There are currently no evaluation measures specifically tailored for HyperNetworks. Commonly used metrics include Area Under the Curve (AUC), F1-Score, Precision, and Recall. Among these, we focus on AUC for Precision-Recall (PR) curves, as it is particularly suited for highly imbalanced real-world graphs, making AUPR a more appropriate choice. While accuracy, the proportion of correctly predicted instances can also be employed, it is less commonly used due to its limitations in handling imbalanced data. This metric assesses the similarity between predicted and actual hyperedges based on node overlap. However, it is not applicable in this study, as our work is restricted to hyperedges of size 2.

V. RESULTS AND DISCUSSION

The proposed *LPHC* model simulations are performed on the top 20,000 node pairs, with results averaged across 10 data

TABLE 5. Average F1-scores of $LPHC^{CN}_{\mathbb{HC}}$ for the top-20000 node pairs.

	$\mathbf{LPHC}^{CN}_{\mathbb{HC}}(u,v)$				LPH without using HyperCentrality	$\mathbf{LPC}^{CN}_{\mathbb{C}}(u,v)$			
Dataset↓ LP measure →	$\mathbb{HC}^{CN}_{d}$	$\mathbb{HC}^{CN}_{cc}$	$\mathbb{HC}^{CN}_{b}$	$\mathbb{HC}^{CN}_{cl}$	**HCN**	$\mathbb{C}^{CN}_{d}$	$\mathbb{C}^{CN}_{cc}$	$\mathbb{C}^{CN}_{b}$	$\mathbb{C}^{CN}_{cl}$
NDC-classes	0.861	**0.898**	0.837	0.798	0.415	0.162	0.211	0.196	0.123
email-Eu	0.838	**0.891**	0.829	0.812	0.501	0.103	0.292	0.182	0.202
DAWN	0.829	**0.868**	0.845	0.836	0.518	0.284	0.298	0.253	0.215
cat-edge-geometry-questions	0.654	**0.702**	0.622	0.512	0.539	0.231	0 .245	0.189	0.192
hyperedges-contact-high-school	0.868	**0.886**	0.841	0.838	0.613	0.258	0.161	0.152	0.156

points to enhance accuracy. The HyperCentrality based link prediction measures examined include Hyperdegree Centrality using Common Neighbor $\mathbb{HC}_d$, HyperClusterring Coefficient based Common Neighbor $\mathbb{HC}_{cc}$, HyperBetweenness Centrality based Common Neighbor $\mathbb{HC}_b$, and HyperCloseness Centrality based Common Neighbor $\mathbb{HC}_{cl}$ and compared our measures with latest Link Prediciton in hyper-networks (HCN and HJC) measures. The F1-score and AUPR values presented in Table. 5 and Table. 7 represent the common neighbor-based hypercentrality measures calculated for node pairs. Similarly, the F1-score and AUPR values in Table. 6 and Table. 8 reflect the Jaccard coefficient-based hypercentrality measures for node pairs across various top 'K' scenarios, where 'K' denotes the number of top-ranking node pairs selected from each dataset. Similarity scores were analyzed for the top 2000, 4000, 6000, 8000, 10000, 12000, 14000, 16000, 18000, and 20000 node pairs, with results averaged over 10 data points. The final results represent the average performance across these selected node pairs.

In Table. 5, the data analysis shows that the proposed $\mathbb{HC}^{CN}_{cc}$ model consistently outperforms other *LPHC* measures across all datasets and scenarios, as well as the latest *HCN* method. For example, in the NDC-classes dataset, the $\mathbb{HC}^{CN}_{cc}$ model demonstrates a 4.2% improvement over $\mathbb{HC}^{CN}_{d}$, a 7.3% improvement over $\mathbb{HC}^{CN}_{b}$, and a 12.6% improvement over $\mathbb{HC}^{CN}_{cl}$. When compared to the HCN, the $\mathbb{HC}^{CN}_{cc}$ shows an impressive 89.5% improvement. In the case of the email-Eu dataset, the $\mathbb{HC}^{CN}_{cc}$ model exhibits significant performance improvements. It demonstrates a 6.3% increase over $\mathbb{HC}^{CN}_{d}$, and a 7.5% improvement over $\mathbb{HC}^{CN}_{b}$. Additionally, the model achieves a 9.7% enhancement over $\mathbb{HC}^{CN}_{cl}$. Most notably, when compared to the HCN method, the $\mathbb{HC}^{CN}_{cc}$ model delivers an impressive 77.8% improvement in performance. For the DAWN dataset, the $\mathbb{HC}^{CN}_{cc}$ model demonstrates an average improvement of 3.73% over $\mathbb{HC}^{CN}_{d}$, $\mathbb{HC}^{CN}_{b}$ and $\mathbb{HC}^{CN}_{cl}$. Additionally, the $\mathbb{HC}^{CN}_{cc}$ model shows a significant 67.4% improvement when compared to the HCN method. For the cat-edge-geometry-questions dataset, the $\mathbb{HC}^{CN}_{cc}$ model demonstrates an average improvement of 19.11% over the other proposed measures. Additionally, the improvement of $\mathbb{HC}^{CN}_{cc}$ over HCN is 30.24%. Similarly, for the hyperedges-contact-high-school dataset, $\mathbb{HC}^{CN}_{cc}$ shows an average improvement of 4.38% over $\mathbb{HC}^{CN}_{d}$, $\mathbb{HC}^{CN}_{b}$ and $\mathbb{HC}^{CN}_{cl}$, with a significant improvement of 44.54% over HCN.

The columns of $LPC^{CN}_{\mathbb{C}}(u,v)$ in Table. 5 present the performance of centrality-based *CN*, where the hypergraph is transformed into a standard pairwise graph, and traditional centrality measures are applied for link prediction. For example, in the NDC-classes dataset, $\mathbb{C}^{CN}_{cc}$ achieves an F1-score of 0.211, considerably lower than the proposed $\mathbb{HC}^{CN}_{cc}$ score of 0.898. This pattern holds across other datasets, with graph-converted centrality measures consistently lagging behind their hypergraph-based counterparts, emphasizing the critical role of preserving hypergraph structures to capture high-order relationships inherent in hyper-networks.

The analysis clearly demonstrates that the proposed $\mathbb{HC}^{CN}_{cc}$ model consistently outperforms other methods. It not only captures direct interactions, as measured by $\mathbb{HC}^{CN}_{d}$, but also incorporates the broader network structure by focusing on common neighbors. This approach enhances its robustness in identifying meaningful relationships between node pairs. Unlike other measures that prioritize local interactions or shortest paths, $\mathbb{HC}^{CN}_{cc}$ emphasizes connectivity through common neighbors, thereby improving its predictive ability, particularly in scenarios where nodes belong to densely connected hyper-networks.

In Table. 6, the Jaccard coefficient-based hypercentralities $\mathbb{HC}^{JC}_{cc}$ offer an alternative approach to link prediction in hyper-networks by focusing on the normalized overlap of common neighbors. While this method provides a useful metric for assessing the relative proportion of shared neighbors, it does not outperform the common neighbor-based

TABLE 6. Average F1-scores of $LPHC^{JC}_{\mathbb{HC}}$ for the top-20000 node pairs.

	$\mathbf{LPHC}^{JC}_{\mathbb{HC}}(u,v)$				LPH without using HyperCentrality	$\mathbf{LPC}^{JC}_{\mathbb{C}}(u,v)$			
Dataset↓ LP measure →	$\mathbb{HC}^{JC}_{d}$	$\mathbb{HC}^{JC}_{cc}$	$\mathbb{HC}^{JC}_{b}$	$\mathbb{HC}^{JC}_{cl}$	**HJC**	$\mathbb{C}^{JC}_{d}$	$\mathbb{C}^{JC}_{cc}$	$\mathbb{C}^{JC}_{b}$	$\mathbb{C}^{JC}_{cl}$
NDC-classes	0.781	**0.828**	0.756	0.743	0.389	0.186	0.189	0.188	0.213
email-Eu	**0.856**	0.821	0.798	0.763	0.436	0.199	0.212	0.151	0.219
DAWN	0.855	**0.876**	0.822	0.816	0.503	0.244	0.211	0.189	0.209
cat-edge-geometry-questions	0.589	**0.612**	0.548	0.516	0.498	0.316	0.212	0.188	0.169
hyperedges-contact-high-school	0.816	**0.829**	0.808	0.812	0.596	0.188	0.156	0.189	0.166

TABLE 7. AUPR of $LPHC^{CN}_{\mathbb{HC}}$ for the top-20000 node pairs.

	$\mathbf{LPHC}^{CN}_{\mathbb{HC}}(u,v)$				LPH without using HyperCentrality	$\mathbf{LPC}^{CN}_{\mathbb{C}}(u,v)$			
Dataset↓ LP measure →	$\mathbb{HC}^{CN}_{d}$	$\mathbb{HC}^{CN}_{cc}$	$\mathbb{HC}^{CN}_{b}$	$\mathbb{HC}^{CN}_{cl}$	**HCN**	$\mathbb{C}^{CN}_{d}$	$\mathbb{C}^{CN}_{cc}$	$\mathbb{C}^{CN}_{b}$	$\mathbb{C}^{CN}_{cl}$
NDC-classes	0.024	**0.033**	0.022	0.027	0.009	0.019	0.022	0.013	0.021
email-Eu	**0.121**	0.059	0.092	0.089	0.058	0.072	0.062	0.076	0.066
DAWN	0.074	**0.093**	0.081	0.063	0.071	0.051	0.073	0.045	0.052
cat-edge-geometry-questions	0.286	**0.322**	0.233	0.111	0.142	0.029	0.061	0.016	0.027
hyperedges-contact-high-school	0.077	**0.086**	0.042	0.051	0.008	0.055	0.043	0.063	0.063

hypercentralities seen in the previous table. For instance, in the email-Eu dataset, the $\mathbb{HC}^{JC}_{d}$ model achieves the highest score of 0.856, yet it still falls short compared to the common neighbor-based $\mathbb{HC}^{CN}_{cc}$ score of 0.891 from the previous analysis from Table. 5. Similarly, across other datasets in Table. 6, $\mathbb{HC}^{JC}_{HC}$ tend to exhibit lower F1-scores, suggesting that the normalization process can dilute the impact of direct relationships, which are more effectively captured by common neighbor-based measures. The normalization effect of the Jaccard coefficient becomes particularly less effective in densely connected hyper-networks, such as DAWN and cat-edge-geometry-questions, where multiple common neighbors more strongly indicate future connections. In such networks, the direct count of common neighbors without normalization appears to be a more accurate predictor of node interactions.

The performance of $LPC^{JC}_{\mathbb{C}}(u,v)$ in Table. 6 corresponds to Centrality-based *JC* after converting the hypergraph into traditional graph. Across all datasets, the converted graph measures, such as $\mathbb{C}^{JC}_{cc}$, consistently under perform compared to their hypergraph-based counterparts, with significant gaps in F1-scores. For example, $\mathbb{C}^{JC}_{cc}$ trails behind $\mathbb{HC}^{JC}_{cc}$ in every dataset, underscoring the advantage of directly leveraging hypergraph structures. This demonstrates that preserving the

TABLE 8. AUPR of $LPHC^{JC}_{\mathbb{HC}}$ for the top-20000 node pairs.

	$\mathbf{LPHC}^{JC}_{\mathbb{HC}}(u,v)$				LPH without using HyperCentrality	$\mathbf{LPC}^{JC}_{\mathbb{C}}(u,v)$			
Dataset↓ LP measure →	$\mathbb{HC}^{JC}_{d}$	$\mathbb{HC}^{JC}_{cc}$	$\mathbb{HC}^{JC}_{b}$	$\mathbb{HC}^{JC}_{cl}$	**HJC**	$\mathbb{C}^{JC}_{d}$	$\mathbb{C}^{JC}_{cc}$	$\mathbb{C}^{JC}_{b}$	$\mathbb{C}^{JC}_{cl}$
NDC-classes	**0.018**	0.016	0.008	0.011	0.002	0.013	0.009	0.005	0.004
email-Eu	**0.109**	0.073	0.071	0.069	0.048	0.062	0.048	0.042	0.048
DAWN	0.072	**0.088**	0.074	0.059	0.069	0.054	0.035	0.047	0.051
cat-edge-geometry-questions	0.255	**0.302**	0.212	0.166	0.112	0.022	0.053	0.019	0.022
hyperedges-contact-high-school	0.015	**0.019**	0.011	0.006	0.003	0.009	0.011	0.008	0.006

high-order relationships in hypergraphs is crucial for accurate link prediction.

Overall, the results from Table. 6 indicate that while Jaccard coefficient-based hypercentralities are valuable for measuring neighborhood overlap, they generally underperform when compared to common neighbor-based hypercentralities. This suggests that in hyper-networks where common neighbors play a key role, using direct common neighbor counts without normalization provides a more reliable framework for predicting links and modeling interactions.

The analysis of the AUPR results, as presented in Table. 7, reveals that in the NDC-classes dataset, the $\mathbb{HC}^{CN}_{d}$ model achieves the highest AUPR score of 0.024. It outperforms $\mathbb{HC}^{CN}_{cc}$ and $\mathbb{HC}^{CN}_{b}$ by 9.1%, and $\mathbb{HC}^{CN}_{cl}$ by 14.3%. Furthermore, $\mathbb{HC}^{CN}_{d}$ demonstrates an improvement of 82.5% over the existing HCN measure. In the email-Eu dataset, the $\mathbb{HC}^{CN}_{d}$ model outperforms all other proposed measures, achieving an average improvement of 57.52%. Additionally, $\mathbb{HC}^{CN}_{d}$ demonstrates a 94.6% improvement over the existing HCN measure. For the DAWN dataset, $\mathbb{HC}^{CN}_{cc}$ surpasses the other proposed measures with an average improvement of 17.15%, and further shows a 30.9% improvement over HCN. In the cat-edge-geometry-questions dataset, $\mathbb{HC}^{CN}_{cc}$ also outperforms the remaining measures, delivering an average improvement of 15.37%, while demonstrating a substantial 97.5% improvement over HCN. Finally, in the hyperedges-contact-high-school dataset, $\mathbb{HC}^{CN}_{cc}$ again leads with an average improvement of 58.33%, and shows an impressive 98.3% improvement over the HCN method. In both the NDC-classes and email-Eu hyper-networks, the interactions are likely dominated by direct, pairwise relationships between nodes, which are best captured by the hyperdegree centrality $\mathbb{HC}^{CN}_{cc}$. This measure focuses on the number of hyperedges a node participates in, which becomes crucial when interactions are primarily based on the presence or absence of direct connections. In DAWN, cat-edge-geometry-questions and hyperedges-contact-high-school datasets, node interactions tend to be more collaborative, involving groups of nodes rather than just pairwise connections. This makes $\mathbb{HC}^{CN}_{cc}$, which measures common neighbors, a better fit for capturing the essence of group-based relationships compared to $\mathbb{HC}^{CN}_{b}$, which focuses solely on individual node participation. These datasets may also exhibit a higher level of redundancy in node connections, where a single node can connect to others through multiple shared neighbors.

Among the converted measures in Table. 7, $\mathbb{C}^{CN}_{cc}$ consistently achieves the highest AUPR across all datasets. However, when compared with the proposed $\mathbb{HC}^{CN}_{cc}$ and other hypergraph-based measures, the proposed methods demonstrate significantly better performance. This highlights the superiority of directly leveraging hypergraph structures over converting them to pairwise graphs, as the latter approach fails to fully capture the inherent high-order relationships within hyper-networks. $\mathbb{HC}^{CN}_{cc}$ effectively captures this redundancy, making it more robust in such network structures compared to other centrality measures.

In Table. 8, the Jaccard coefficient-based hypercentralities ($\mathbb{HC}^{JC}_{HC}$) offer an alternative approach to link prediction by normalizing common neighbors between nodes. The Jaccard coefficient achieves this by dividing the number of shared neighbors by the total number of neighbors, providing a more balanced perspective on shared relationships. However, this normalization can also diminish the influence of highly connected nodes, weakening its predictive power, particularly in datasets where strong, direct interactions are dominant. For example, in the NDC-classes dataset, the $\mathbb{HC}^{JC}_{d}$ model achieves a score of 0.018, which is still lower than the

$\mathbb{HC}_d^{CN}$ score of 0.024, a difference of 33.33%, as observed in Table. 7. The normalization inherent to the Jaccard coefficient weakens its ability to capture the direct connections that are essential in this dataset. Similarly, in the email-Eu dataset, the $\mathbb{HC}_d^{JC}$ model achieves a score of 0.109, but this remains lower than the common neighbor-based score of 0.121, which is 11% higher, again highlighting the limitation of normalization in capturing strong pairwise interactions. In more collaborative hyper-networks, such as DAWN, cat-edge-geometry-questions, and hyperedges-contact-high-school, the $\mathbb{HC}_{HC}^{JC}$ show some improvement, particularly in measuring relative overlap in node neighborhoods. For instance, $\mathbb{HC}_{cc}^{JC}$ achieves competitive scores of 0.302 in cat-edge-geometry-questions and 0.088 in DAWN. However, even in these collaborative networks, the normalization process reduces the impact of nodes with overlapping neighbors, limiting the Jaccard coefficient's ability to effectively capture dense subgraph structures.

In Table. 8, among the converted graph measures, $\mathbb{C}_d^{JC}$ consistently achieves the highest AUPR values across all datasets, closely followed by $\mathbb{C}_{cc}^{JC}$. However, when compared to the proposed $LPHC_{\mathbb{HC}}^{JC}$ hypergraph-based measures, the proposed methods demonstrate significantly superior performance. This highlights the critical advantage of utilizing hypergraph structures directly, as converting them into pairwise graphs results in the loss of essential high-order relationships, thereby reducing prediction accuracy.

In hyper-networks where redundancy and multiple overlapping neighbors play a crucial role (e.g., cat-edge-geometry-questions), the normalization applied by the Jaccard coefficient tends to downplay the importance of nodes with multiple common neighbors. By contrast, common neighbor-based measures more fully capture the extent of shared neighbors, leading to superior performance in networks characterized by dense, redundant structures.

A. EVALUATING THE IMPACT OF VARIOUS HYPERCENTRALITY MEASURES ON LINK PREDICTION PERFORMANCE

In this section, we examine the role of hypercentrality measures like hyperdegree, hyperclustering coefficient, hyperbetweenness, and hypercloseness in link prediction performance, focusing on their effectiveness with Common Neighbors (CN) and Jaccard Coefficient (JC). Results from Table. 5-Table. 8 reveal distinct trends for these measures across datasets.

Hyperdegree Centrality ($\mathbb{HC}_d$) shows consistently strong performance in terms of AUPR (Refer Table. 7), particularly in dense networks like email-Eu and NDC-classes. It effectively captures the influence of frequently participating nodes in smaller, dense hyperedges, contributing to robust link prediction results.

Hyperclustering Coefficient ($\mathbb{HC}_{cc}$) consistently outperforms others in terms of F1-score across all datasets. In Table. 5 $\mathbb{HC}_{cc}$ achieves the highest F1-score across all datasets. Its strength lies in capturing localized clustering dynamics, making it highly effective for networks with strong intra-community structures, such as NDC-classes and hyperedges-contact-high-school. Hyperclustering Coefficient ($\mathbb{HC}_{cc}$) consistently outperforms others in terms of F1-score (Table. 5) across all datasets. Its strength lies in capturing localized clustering dynamics, making it highly effective for networks with strong intra-community structures, such as NDC-classes and hyperedges-contact-high-school.

Hyperbetweenness Centrality ($\mathbb{HC}_b$) and Hypercloseness Centrality ($\mathbb{HC}_{cl}$) demonstrate competitive but slightly lower performance compared to $\mathbb{HC}_d$ and $\mathbb{HC}_{cc}$. $\mathbb{HC}_b$ excels in datasets like DAWN, where long-range dependencies are prominent, as it captures the control nodes exert over communication pathways, making it particularly useful in globally connected networks. In contrast, $\mathbb{HC}_{cl}$ is less impactful in sparse datasets like cat-edge-geometry-questions, as its reliance on shortest paths reduces its efficiency in networks with larger hyperedge sizes.

Overall, the hyperclustering coefficient $\mathbb{HC}_{cc}$ emerges as the best performer across datasets, particularly in terms of F1-score and AUPR, due to its ability to capture dense intra-community relationships. Dataset-specific trends highlight the importance of tailoring hypercentrality measures to the structural characteristics of hypergraphs.

VI. CONCLUSION AND FUTURE WORK

In this paper, we introduced a link prediction framework that leverages hypercentrality measures to address the complexity of multi-node interactions in hyper-networks. By adapting traditional centrality metrics—such as degree, clustering coefficient, betweenness, and closeness to the hyper-networks domain, we developed the a novel link prediction model by taking the rich input of hypergraphs, which effectively enhances traditional link prediction techniques by incorporating node significance through centrality scores. Our empirical results, evaluated across several real-world datasets, demonstrate that hypercentrality-based models, particularly those utilizing hyperdegree and hyperclustering coefficients, consistently outperform existing link prediction methods in terms of both F1-score and AUPR. This suggests that hypercentrality measures provide a more accurate and nuanced approach for predicting link formation in hyper-networks, especially in networks characterized by dense connectivity or strong clustering.

As part of future work, we aim to extend our investigation beyond the hyperdegree $\mathbb{HC}_d$, hyperclustering coefficient $\mathbb{HC}_{cc}$, hyperbetweenness $\mathbb{HC}_b$, and hypercloseness $\mathbb{HC}_{cl}$ measures used in this study. Additionally, instead of considering all non-adjacent node pairs, we plan to examine the effect of restricting the prediction to node pairs within a hop distance of 2, 3, and beyond. An important direction for future work is to extend the current model, which is focused on non-adjacent node pairs of size 2, to include larger hyperedge sizes, such as non-adjacent node triplets (size 3), quadruplets (size 4), and beyond. This extension would enable the model to capture and analyze more complex group interactions,

enhancing its applicability to hyper-networks with diverse structural properties and larger hyperedges. Additionally, we aim to incorporate node and edge attributes to further enhance the model's predictive capabilities. These extensions will allow for a more comprehensive understanding of link prediction in hyper-networks, further refining our model's effectiveness in real-world applications.

Deep learning approaches have significantly advanced the field of link prediction in hypergraphs by effectively modeling and leveraging the intricate relationships present in complex networked systems. Neural Hypergraph Link Prediction (NHP) [31], Heterogeneous Hypergraph Representation Learning (HHRL) [32] represent notable advancements in this domain. We intend to extend our research in this direction as part of our future work.

REFERENCES

[1] R. Albert and A. Barabási, "Statistical mechanics of complex networks," *Rev. Modern Phys.*, vol. 74, no. 1, pp. 47–97, Jan. 2002.

[2] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proc. 12th Int. Conf. Inf. Knowl. Manage.*, Nov. 2003, pp. 556–559.

[3] A. Vazquez, "Complex hypergraphs," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 107, no. 2, Feb. 2023, Art. no. 024316.

[4] H. Wu, Y. Yan, and M. K. Ng, "Hypergraph collaborative network on vertices and hyperedges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3245–3258, Mar. 2023.

[5] S. Feng, "Hypergraph models of biological networks to identify genes critical to pathogenic viral response," *BMC Bioinf.*, vol. 22, no. 1, p. 287, Dec. 2021.

[6] C. Chen and Y.-Y. Liu, "A survey on hyperlink prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 15034–15050, Nov. 2024.

[7] E. Nasiri, K. Berahmand, and Y. Li, "A new link prediction in multiplex networks using topologically biased random walks," *Chaos, Solitons Fractals*, vol. 151, Oct. 2021, Art. no. 111230.

[8] K. Berahmand, S. Bahadori, M. N. Abadeh, Y. Li, and Y. Xu, "SDAC-DA: Semi-supervised deep attributed clustering using dual autoencoder," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 11, pp. 6989–7002, Nov. 2024.

[9] R. Sheikhpour, K. Berahmand, M. Mohammadi, and H. Khosravi, "Sparse feature selection using hypergraph Laplacian-based semi-supervised discriminant analysis," *Pattern Recognit.*, vol. 157, Jan. 2025, Art. no. 110882.

[10] M. Shen, C. Wang, Q.-G. Wang, Y. Sun, and G. Zong, "Synchronization of fractional reaction-diffusion complex networks with unknown couplings," *IEEE Trans. Netw. Sci. Eng.*, vol. 11, no. 5, pp. 4503–4512, Sep. 2024.

[11] L. Yao, L. Wang, L. Pan, and K. Yao, "Link prediction based on common-neighbors for dynamic social network," *Proc. Comput. Sci.*, vol. 83, pp. 82–89, Jan. 2016.

[12] A. Saxena and S. Iyengar, "Centrality measures in complex networks: A survey," 2020, *arXiv:2011.07190*.

[13] Y. V. Nandini, T. J. Lakshmi, M. K. Enduri, and H. Sharma, "Link prediction in complex networks using average centrality-based similarity score," *Entropy*, vol. 26, no. 6, p. 433, May 2024.

[14] E. Estrada and J. A. Rodriguez-Velazquez, "Complex networks as hypergraphs," 2005, *arXiv:physics/0505137*.

[15] T. Kumar, K. Darwin, S. Parthasarathy, and B. Ravindran, "HPRA: Hyperedge prediction using resource allocation," in *Proc. 12th ACM Conf. Web Sci.*, Jul. 2020, pp. 135–143.

[16] Y. V. Nandini, T. J. Lakshmi, M. K. Enduri, H. Sharma, and M. W. Ahmad, "Extending graph-based LP techniques for enhanced insights into complex hypergraph networks," *IEEE Access*, vol. 12, pp. 51208–51222, 2024.

[17] L. C. Freeman, "Centrality in social networks: Conceptual clarification," in *Social Network: Critical Concepts in Sociology*, vol. 1. London, U.K.: Routledge, 2002, pp. 238–263.

[18] M. Á. Serrano and M. Boguñá, "Clustering in complex networks. I. General formalism," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 74, no. 5, Nov. 2006, Art. no. 056114.

[19] G. Costantini and M. Perugini, "Generalization of clustering coefficients to signed correlation networks," *PLoS ONE*, vol. 9, no. 2, Feb. 2014, Art. no. e88669.

[20] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, p. 35, Mar. 1977.

[21] M. Newman, *Networks*. London, U.K.: Oxford Univ. Press, 2018.

[22] S. Roy and B. Ravindran, "Measuring network centrality using hypergraphs," in *Proc. 2nd ACM IKDD Conf. Data Sci.*, Mar. 2015, pp. 59–68.

[23] D. Li, Z. Xu, S. Li, and X. Sun, "Link prediction in social networks based on hypergraph," in *Proc. 22nd Int. Conf. World Wide Web*, May 2013, pp. 41–42.

[24] İ. Tuğal and Z. Pala, "Centrality with entropy in hypergraphs based on similarity measures," *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, vol. 14, no. 3, pp. 407–419, 2023.

[25] S. G. Aksoy, C. Joslyn, C. O. Marrero, B. Praggastis, and E. Purvine, "Hypernetwork science via high-order hypergraph walks," *EPJ Data Sci.*, vol. 9, no. 1, p. 16, Dec. 2020.

[26] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, "Simplicial closure and higher-order link prediction," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 48, pp. E11221–E11230, Nov. 2018.

[27] R. Mastrandrea, J. Fournet, and A. Barrat, "Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys," *PLoS ONE*, vol. 10, no. 9, Sep. 2015, Art. no. e0136497, doi: 10.1371/journal.pone.0136497.

[28] X. Jiao, S. Wan, Q. Liu, Y. Bi, Y.-L. Lee, E. Xu, D. Hao, and T. Zhou, "Comparing discriminating abilities of evaluation metrics in link prediction," *J. Phys., Complex.*, vol. 5, no. 2, Jun. 2024, Art. no. 025014.

[29] Y. Yang, R. N. Lichtenwalter, and N. V. Chawla, "Evaluating link prediction methods," *Knowl. Inf. Syst.*, vol. 45, no. 3, pp. 751–782, Dec. 2015.

[30] K. Boyd, K. H. Eng, and C. D. Page, "Area under the precision-recall curve: Point estimates and confidence intervals," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Prague, Czech Republic. Cham, Switzerland: Springer, 2013, pp. 451–466.

[31] N. Yadati, V. Nitin, M. Nimishakavi, P. Yadav, A. Louis, and P. Talukdar, "NHP: Neural hypergraph link prediction," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 1705–1714.

[32] Z. Zhao, K. Yang, and J. Guo, "Heterogeneous hypergraph representation learning for link prediction," *Eur. Phys. J. B*, vol. 97, no. 10, pp. 1–9, Oct. 2024.

Y. V. NANDINI received the B.Tech. degree in computer science and engineering from JNTUK, Kakinada, Andhra Pradesh, India, in 2014, and the M.Tech. degree in computer science and engineering from Andhra University, Visakhapatnam, Andhra Pradesh, India, in 2016. She is currently pursuing the Ph.D. degree in computer science and engineering with SRM University-AP, Amaravati, India. Her research interests include recommender systems, graph analytics, and link prediction.

T. JAYA LAKSHMI (Member, IEEE) received the Ph.D. degree from the School of Computer and Information Sciences, University of Hyderabad, India, in 2019, focusing on "LP in heterogeneous social networks." She is currently a Lecturer with the Department of Computing, Sheffield Hallam University, Sheffield, U.K. She is also an active reviewer of prominent international journals and brings over 22 years of teaching experience. Her research interests include graph mining, recommender systems, natural language processing, and security analytics.

MURALI KRISHNA ENDURI (Member, IEEE) received the M.Sc. degree in mathematics from Acharya Nagarjuna University, Andhra Pradesh, India, in 2008, the M.Tech. degree in systems analysis and computer applications from the National Institute of Technology Karnataka, India, in 2011, and the Ph.D. degree in computer science and engineering from Indian Institute of Technology Gandhinagar, Gujarat, India, in 2018. He conducted postdoctoral research with Indian Institute of Technology Madras, Tamil Nadu, India, in 2018. He is currently an Assistant Professor with the Department of Computer Science and Engineering, SRM University-AP, Amaravati, India. His research interests include recommender systems, algorithms, complexity theory, and complex networks. His work includes developing algorithms for graph isomorphism in specific graph classes, evaluating article diversity, and modeling disease spread predictions.

MOHD ZAIRUL MAZWAN JILANI is currently a Senior Lecturer in computer science with Sheffield Hallam University. His research interests include computer science is machine learning and evolutionary algorithms, which are applied to solving complex problems in health informatics. His Ph.D. research area mainly focused on heuristic search, classification, and clustering that applied to the glaucoma dataset. Prior to joining academia, he had nine years of professional experience working in financial institutions in Malaysia. He hold various positions in these experiences, including an IT auditor and an IT project auditor for the banking system in Malaysia. Therefore, he has global experience of interactions and applications with the computing industry.