C3MM : Clique-Closure based Hyperlink Prediction

Govind Sharma[*] , **Prasanna Patil** and **M. Narasimha Murty**
Indian Institute of Science, Bengaluru, India
{govinds, patilk, mnm}@iisc.ac.in

Abstract

Usual networks lossily (if not incorrectly) represent higher-order relations, which calls for complex structures such as hypergraphs to be used instead. Akin to the link prediction problem in graphs, we deal with hyperlink (higher-order link) prediction in hypergraphs. With a handful of solutions in the literature that seem to have merely scratched the surface, we provide improvements for the same. Motivated by observations in recent literature, we first formulate a "clique-closure" hypothesis (*viz.*, hyperlinks are more likely to be formed from near-cliques rather than from non-cliques), test it on real hypergraphs, and then exploit it for our very problem. In the process, we generalize hyperlink prediction on two fronts: (1) from small-sized to arbitrary-sized hyperlinks, and (2) from a couple of domains to a handful. We perform experiments (both the hypothesis-test as well as the hyperlink prediction) on multiple real datasets, report results, and provide both quantitative and qualitative arguments favouring better performances *w.r.t.* the state-of-the-art.

1 Introduction

Relations in nature, more often than not, exist between a *set* of entities rather than a *pair* thereof. For a lossless representation, complex structures such as *hypergraphs* [Berge, 1984] are used, wherein *hyperlinks* or *hyperedges* are used to represent higher-order relations. *Hyperlink prediction* refers to predicting future/missing hyperlinks in a given hypergraph [Xu *et al.*, 2013; Zhang *et al.*, 2018; Benson *et al.*, 2018]. We draw inspirations from recent literature and the existing state-of-the-art, *i.e.*, Coordinated Matrix Minimization (CMM) [Zhang *et al.*, 2018], to solve the problem of predicting arbitrary-sized hyperlinks in networks. We first formulate a ***C**lique-**C**losure **H**ypothesis* (**CCH**), which can be summarized as follows: ***Hyperlinks in a network are more likely to be formed from closures of cliques (and near-cliques) rather than those of non-cliques***. In simpler terms, we hypothesize that for a given hyperlink, prior to its first occurrence, its incident nodes should have had more interactions than a set of arbitrary number of nodes usually does. *I.e.*, we expect every hyperlink to have *evolved gradually*, rather than having spontaneously "sprung-up". Consider the example shown in Figure 1, where it could be noted how at any $t = t_i$, smaller hyperlinks from the past (dotted) combine together to form larger ones (solid) in the present.

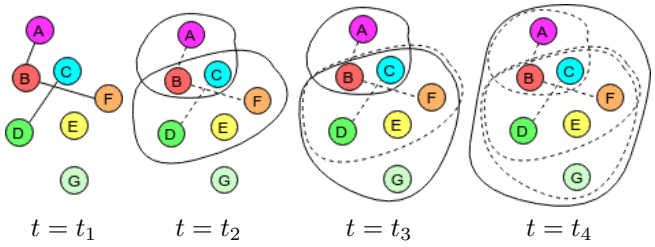

Figure 1: A toy example: For $i = 1, 2, 3, 4$, solid hyperlinks formed exactly at $t = t_i$, are eventually shown as dotted connections. The clique-closure hypothesis (CCH) we propose says that for a current hyperlink (solid) it is highly likely that its nodes had been densely connected via past connections (dotted).

We first test CCH on real datasets, and then use it for hyperlink prediction via a method we term *"**C**lique-**C**losure based **C**oordinated **M**atrix **M**inimization"* (**C3MM**). We ingest CCH into the objective function of CMM to get C3MM, and then solve it in a similar fashion. Choosing datasets from different domains, we note significant improvements over CMM. Major improvements come from the fact that C3MM gives a chance to those hyperlinks that could explain existing relations.

1.1 Our Contributions

1. We formulate and test a **clique-closure hypothesis (CCH) for hypergraph network evolution**. As a result, we provide novel insights into hyperlink evolution.
2. We provide a **hyperlink prediction algorithm C3MM that significantly improves upon CMM**.
3. We **extend hyperlink prediction** for hyperlinks of arbitrary **size** and to multiple **domains**.

[*]Contact Author

Symbol	Definition
V	Set of *vertices*
$F \subseteq 2^V$	Set of *hyperlinks / hyperedges*
$T : F \to \mathbb{R}$	Hyperlink *timestamp*
$H = (V, F)$	*Non-temporal Hypergraph*
$H = (V, F, T)$	*Temporal Hypergraph*
$F_{<t} := \{f \in F \mid T(f) < t\}$	Hyperlinks observed *before* time t
$H_{<t} = (V, F_{<t}, T)$	Hypergraph observed *before* time t
$H_{-f} = (V, F \setminus \{f\}, T)$	Hypergraph *punctured w.r.t.* (or *without*) f
$G = (V, E)$	Undirected *graph* with set E of edges over V
$E_f = \{\{u, v\} \mid u, v \in f\}$	Edges induced by a hyperedge f.
$\eta(F) = \cup_{f \in F} E_f$	Edges projected by hyperedges in F.
$\eta(H) = (V, \eta(F))$	*Clique-expansion* of H[1]
$d(G) = \frac{2 \cdot \lvert E \rvert}{\lvert V \rvert \cdot (\lvert V \rvert - 1)}$	Density of graph G
$G\lvert_f$	Subgraph of G *w.r.t.* nodes $f \subseteq V$

Table 1: Notations used in the article.

2 Preliminaries and Notation

We define basic notations in Table 1 for easy look-up. Hyperlink prediction, just like link prediction, could be formulated as a binary classification problem where *positive* and *negative* classes correspond to *hyperlink* and *non-hyperlink* respectively. For our experiments, we first partition the set of hyperlinks into two parts, namely observed and unobserved hyperlinks (F_{obs} and F_{unobs}). For temporal hypergraphs, the set of hyperlinks is partitioned chronologically whereas for non-temporal hypergraphs, it is done randomly.

Next, we pick non-hyperlinks through under-sampling of the negative class due to extreme class imbalance (*i.e.*, $O(2^{|V|})$ possible non-hyperlinks), and denote the sampled negative class as $\hat{F}_{\text{unobs}}$. Finally, we state the **hyperlink prediction problem** as follows: *Given a set of observed hyperlinks F_{obs}, find scores $s : F_{unobs} \cup \hat{F}_{unobs} \to [0, 1]$, mapping potential hyperlinks to their hyperlink-formation probabilities.* Later in this article, we also refer to F_{unobs} as ΔF, $\hat{F}_{unobs}$ as $\Delta\hat{F}$, and F_{obs} as plainly F.

3 The Clique Closure Hypothesis

Occurrence of a hyperlink marks the collaboration among multiple entities via a single common event. It is intuitive that subsets of these entities would have interacted in some form in the past, rather than the hyperlink getting formed spontaneously. In formal terms, in a temporal hypergraph $H = (V, F, T)$, corresponding to a hyperlink $f \in F$ formed at a given time $T(f)$, we could expect to find some hyperlinks $f' \in F_{<T(f)}$ that overlap densely with subsets of nodes incident on f. Since if that were not true, there is not much explanation — at least not any using the hypergraph topology — as to why the relation f is formed in the first place. In the projected graph $\eta(H_{<T(f)})$, this translates as densely connected subgraphs (near-cliques) or sometimes even cliques. In simple words, CCH states that with high probability, nodes of a hyperlink were part of dense subgraphs before they formed hyperlinks.

We formally define CCH in this section, but before doing so, we need to keep certain concepts well-defined, since their equivalents do not exist in the literature. We define **hypergraph density** of a hypergraph H as $d(H) := d(\eta(H))$, the density of its clique-expanded graph. Similarly, **subgraph density** $d(f, H)$ of any set of nodes $f \subseteq V$ (where f need not be a hyperlink) is defined as $d(f, H) := d(\eta(H)|_f)$. Note that $f \in F \implies d(f, H) = 1$. We define a slight modification of this notion for temporal and non-temporal hypergraphs, *viz.*, **pre-hyperlink density** $d_{pre}(f, H) := d(f, H_{<T(f)})$ and **punctured hyperlink density** $d_{punc}(f, H) := d(f, H_{-f})$ respectively. The prefixes *punctured-* and *pre-* here refer to the fact that density is calculated on the hypergraph that existed *without* and *before* hyperlink f respectively. A higher pre-hyperlink density for a hyperlink would mean it evolved from near-cliques. Moreover, a hyperlink f evolving from cliques would have $d_{pre}(f, H) = 1$, and those having an underlying clique structure would have $d_{punc}(f, H) = 1$. In other words, d_{punc} is used as an alternative for d_{pre} in a non-temporal hypergraph, where the concept of *evolution* (i.e. order of hyperedge discovery is irrelevant) does not exist.

Let the **clique-fraction** $cf(H)$ of hypergraph H be defined as $cf(H) := |\{f \in F : d_x(f, H) = 1\}| \,/\, |F|$, *the fraction of hyperlinks that formed from cliques*, where d_x denotes d_{punc} and d_{pre} for temporal and non-temporal hypergraphs respectively. Since $cf(H)$ is expected to be too low for non-temporal datasets, we also define a constant **minimum-clique-fraction** cf_{min} and fix it to be $cf_{min} = 0.05$, which is a little more than the maximum hypergraph density among all non-temporal hypergraphs (ref Table 3). Finally, we define **cliqueness** of a hyperlink $f \in F$ as follows:

$$\chi(f, H) := d_x(f, H) \cdot \max(cf_{min}, cf(H)) \qquad (1)$$

where d_x denotes d_{punc} and d_{pre} for temporal and non-temporal hypergraphs respectively. We are now ready with a well-defined measure — **cliqueness** — to capture the notion of how dense is the region from which a given hyperlink is formed. Cliqueness captures both clique-fraction, as well as density, thereby catering to both clique- as well as near-clique-structure of a given hyperlink.

Hypothesis 1 (CCH: Clique Closure Hypothesis) *Given a hypergraph $H = (V, F)$ (or (V, F, T)), with significance $\alpha = 0.1$, the null and alternate hypotheses for CCH are defined as follows for a hyperlink $f \in F$:*

$$\mathbb{H}'_0 : \chi(f, H) \leq \mathbb{E}[d|H], \quad \mathbb{H}'_1 : \chi(f, H) > \mathbb{E}[d|H], \qquad (2)$$

where $\mathbb{E}[d|H] := \frac{1}{|2^V|} \cdot \sum_{f' \in 2^V} d(f', H)$, the mean subgraph density over all subsets of V.

In order to simplify CCH, and to make it more deterministic, we have the following result in place.

Theorem 1 *Mean subgraph density of H over all subsets of V is equal to its hypergraph-density. In other words, $\mathbb{E}[d|H] = d(H)$.*

Hypothesis 2 (CCH restated)

$$\mathbb{H}_0 : \chi(f, H) \leq d(H), \quad \mathbb{H}_1 : \chi(f, H) > d(H), \qquad (3)$$

where $d(H)$ and $\chi(f, H)$ denote density of hypergraph H and cliqueness of hyperlink f therein.

[1] As defined by [Agarwal *et al.*, 2006]

Algorithm 1 An algorithm to test CCH on a temporal hypergraph $H = (V, F, T)$. Each hyperlink $f = \{v_1, \ldots, v_{|f|}\} \in F$ is evaluated *w.r.t.* connections $\eta(H_{<T(f)})$ in its past based on how densely $\{v_1, \ldots, v_{|f|}\}$ are connected.

Input: Temporal hypergraph, $H = (V, F, T)$
Output: p-value of $\mathbb{H}_0$ for H

1: $F_{>2} \leftarrow \{f \in F : |f| > 2\}$
2: $d_H \leftarrow \dfrac{2 * |\eta(F)|}{|V| * (|V| - 1)}$
3: $N_c \leftarrow 0$
4: $D \leftarrow \{\}$
5: **for** $f \in F_{>2}$ **do**
6: $\quad E_f \leftarrow \{\{u, v\} \mid \forall\, u, v \in f\}$
7: $\quad t \leftarrow T(f)$
8: $\quad F_{<t} \leftarrow T^{-1}([0, t))$
9: $\quad E_{<t} \leftarrow \eta(F_{<t})$
10: $\quad D[f] \leftarrow \dfrac{|E_f \cap E_{<t}|}{|E_f|}$
11: $\quad$ **if** $(f, E_f \cap E_{<t})$ is a clique **then**
12: $\quad\quad D[f] == 1$
13: $\quad$ **else**
14: $\quad\quad N_c \leftarrow N_c + 1$
15: $\quad$ **end if**
16: **end for**
17: $cf \leftarrow \max(cf_{min}, N_c/|F_{>2}|)$
18: $N_{CCH} \leftarrow 0$
19: **for** $f \in F_{>2}$ **do**
20: $\quad \chi_f \leftarrow D[f] * cf$
21: $\quad$ **if** $\chi_f \leq d_H$ **then**
22: $\quad\quad N_{CCH} \leftarrow N_{CCH} + 1$
23: $\quad$ **end if**
24: **end for**
25: $p \leftarrow N_{CCH}/|F_{>2}|$
26: **return** p

We test CCH on a given temporal hypergraph H using Algorithm 1 and report results in Table 3.

Applying CCH to a non-temporal hypergraph would be futile, since there's no concept of *evolution* per se defined for it. For instance, reactions (hyperlinks) in a metabolite hypergraph [Zhang *et al.*, 2018] cannot be arranged in a chronological order. However, we attempt to test CCH for such networks using a proxy mechanism, in that we set $d_x = d_{punc}$ in eq. 1 while calculating cliqueness $\chi(f, H)$. Making few changes to lines 7–10 in Algorithm 1, so as to calculate $D[f] \leftarrow d_{punc}(f, H)$ for each hyperlink f, we could find p-values for a non-temporal hypergraph as well. The idea is to validate whether for a hyperedge, its incident nodes are well-connected even without its presence. Finally, we present the results in Table 3 for both temporal and non-temporal datasets.

In the literature, Benson et al. [Benson *et al.*, 2018], who restrict themselves to 3- and 4-sized hyperlinks only, refer (although implicitly) to a similar phenomenon, wherein they argue that a clique (an open simplex) eventually forms a hyperlink (a closed simplex). The results of evaluating CCH on various datasets are tabulated and discussed in Section 7.1.

As would be discussed later, **temporal hypergraphs strongly satisfy CCH**, *i.e.*, it is evident that **most hyperlinks were cliques or near-cliques (densely connected) in the projected graph before they become hyperlinks.**

4 C3MM: CCH based Hyperlink Prediction

We exploit this unique characteristic of clique-closure to predict hyperlinks. The approach is similar to *Coordinated Matrix Minimization* (CMM) by Zhang et al. [Zhang *et al.*, 2018]. We call our method *Clique-Closure based CMM* (C3MM). It is formed as follows.

For a given hypergraph $H = (V, F)$, define $S \in \{0, 1\}^{|V| \times |F|}$ to be its *incidence matrix*. And let $\Delta F \subseteq \mathcal{P}(V) \setminus F$ represent the hyperlinks that are *missing* from (or *yet to occur* in) H. Clearly, $F \cap \Delta F = \emptyset$. Let $\Delta S \in \{0, 1\}^{|V| \times |\Delta F|}$ be the incidence matrix corresponding to ΔF. Let $H' := (V, F \uplus \Delta F)$ be the completed hypergraph, whose incidence matrix $S' \in \{0, 1\}^{|V| \times (|F| + |\Delta F|)}$ could be represented as follows[2]:

$$S' := [S;\ \Delta S]. \tag{4}$$

Adjacency matrix for projected graph $\eta(H)$ is defined as $A := \eta(S) := SS^T \in \mathbb{R}^{|V| \times |V|}$. Similarly, $A' := S'S'^T$ refers to the adjacency matrix of $\eta(H')$, for which we have:

$$\begin{aligned} A' = S'S'^T &= [S; \Delta S][S; \Delta S]^T \\ &= SS^T + \Delta S \Delta S^T \\ &= A + \Delta A, \end{aligned} \tag{5}$$

where ΔA refers to the links (edges) in adjacency space that get projected by missing hyperlinks ΔF represented by ΔS.

Let $F_{univ} = \{f_1, f_2, \ldots, f_{|F_{univ}|}\}$ represent the set of *universal hyperlinks* (or candidate hyperlinks), forming our test set. Of F_{univ}, ΔF corresponds to true hyperlinks (the positive class); the remaining hyperlinks, $F_{univ} \setminus \Delta F$, called as *non-hyperlinks*, can be represented by $\Delta \hat{F}$. and let $\Delta \hat{S} \in \{0, 1\}^{|V| \times |\Delta \hat{F}}$ be the corresponding incidence matrix.

Let $U = [\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_{|F_{univ}|}] \in \mathbb{R}^{|V| \times |F_{univ}|}$ to be the incidence matrix for the of set of candidate hyperlinks F_{univ}. Once the adjacency matrix ΔA of the missing links is predicted, the next step would be to pick those hyperlinks from F_{univ}, that best *explain* A and ΔA. For a given diagonal matrix $\Lambda_U = diag(\lambda_1, \lambda_2, \ldots, \lambda_{|F_{univ}|}) \in \{0, 1\}^{|F_{univ}| \times |F_{univ}|}$, the product incidence matrix $U\Lambda_U$ would "select" exactly those columns $\mathbf{u}_i$ from U, for which $\lambda_i = 1$. The corresponding adjacency matrix is then $U\Lambda_U(U\Lambda_U)^T = U\Lambda_U^2 U^T = U\Lambda_U U^T$. Hence Λ_U functions as **hyperlink selector** or predictor.

For the purpose of link prediction, any feasible link prediction method can be used. Here, we use a Common Neighbor (CN) [Newman, 2001; Liben-Nowell and Kleinberg, 2003] based link prediction technique. We first complete the adjacency matrix using the CN score, and then achieve its low rank approximation via Symmetric NMF [Kuang *et al.*,

[2]Here, $[A; B]$ denote column-wise concatenation of matrices A and B having the same number of rows.

2012]. The matrix $A_{CN} = A^2 - diag(A^2)$ captures the common neighbor information of the projected graph $\eta(H)$ of H. To predict missing links ΔA we first approximate $A + A_{CN}$ with a low-rank matrix $W \in \mathbb{R}^{|V| \times k}$, where $k < |V|$, such that $A + A_{CN} \approx WW^T$.

$$\min_{W \in \mathbb{R}_+^{|V| \times k}} \left\| A + A_{CN} - WW^T \right\|_F^2. \tag{6}$$

The representation capability of W is low and hence such approximation ends up removing noisy links which might have been introduced due to A_{CN}. Thus we define the predicted links as $\Delta A := WW^T - A$.

Next step is to predict the missing hyperlinks ΔF from the predicted missing links ΔA. Since U contains missing hyperlinks as well as non-hyperlinks, the diagonal matrix Λ_U should be such that the hyperlinks selected by $U\Lambda_U$ correspond to the links in ΔA when they are projected on graph. This can be obtained by optimizing Λ_U w.r.t following objective function:

$$\min_{\Lambda_U \in diag\left(\{0,1\}^{|F_{univ}|}\right)} \left\| \Delta A - U\Lambda_U U^T \right\|_F^2 \tag{7}$$

This is where we bring CCH into the picture. According to CCH, links in A also play a major role in formation of future hyperlinks. Hence, the predicted hyperlinks in ΔF should not only explain missing links of ΔA but also existing links in A (through clique and near-clique closure). However, as we already know, links in A are formed by F and hence can always be explained by S. Then hyperlinks can be predicted by optimizing Λ_U *w.r.t* following objective function:

$$\min_{\substack{\Lambda_U \in diag\left(\{0,1\}^{|F_{univ}|}\right) \\ \Lambda_S \in diag\left(\{0,1\}^{|F|}\right)}} \left\| A - S\Lambda_S S^T - U\Lambda_U U^T \right\|_F^2 + \|\Lambda_S\|_1. \tag{8}$$

The L1-penalty imposed on Λ_S is important to avoid a trivial solution where $\Lambda_S = I$ and $\Lambda_U = 0$.

Satisfying the objective functions specified in eqs. (8) and (7) leads to a joint optimization problem for hyperlink prediction, which we formulate next. We note that since the problems in eqs. (7 – 8) fall into the integer programming paradigm, and since such problems are NP-complete, we relax the domains of Λ_U and Λ_S to the unit interval $[0, 1]$ instead of $\{0, 1\}$. Hence our final problem boils down to the following:

$$\begin{aligned} \min_{\substack{\Lambda_U \in diag\left([0,1]^{|F_{univ}|}\right) \\ \Lambda_S \in diag\left([0,1]^{|F|}\right)}} & \left\| A - S\Lambda_S S^T - U\Lambda_U U^T \right\|_F^2 \\ & + \left\| \Delta A - U\Lambda_U U^T \right\|_F^2 + \|\Lambda_S\|_1 . \end{aligned} \tag{9}$$

In summary, **we have exploited our clique-closure hypothesis by explicitly forcing the objective function to consider cliques and near-cliques from the projected graph of the observed hyperlinks, as well as new information ΔA simultaneously, and predict hyperlinks that explain them both.**

Alternating Minimization

Finding an optimal solution to the problem (9) can be done by minimizing it alternatively — first for W, and then for Λ_S and Λ_U.

This leads to an alternating optimization where we first predict missing links with the help of W obtained as per eq (6) and then predict missing hyperlinks by solving the optimization problem in eq (9). At the end of each iteration we update A with the new predicted links by adding $U\Lambda_U U^T$. Overall C3MM predicts hyperlinks by performing following steps alternatively:

Step 1: For fixed Λ_U from Step 2 below (or by fixing it to be a random matrix for the first iteration), solve for W:

$$\min_{W \in [0,\infty)^{|V| \times |V|}} \left\| A + A_{CN} + U\Lambda_U U^T - WW^T \right\|_F^2 \tag{10}$$

Step 2: Defining $\Delta A := WW^T - A$ for W fixed from Step 1 above, find the optimal Λ_U according to Eq. (9).

Since both Step 1 and Step 2 are convex optimization problems, we solve them by alternatively minimizing them for matrices Λ_U, Λ_S and W, and finally use Λ_U, that denotes the newly predicted hyperlinks.

5 Related Work

The hyperlink prediction problem focuses on predicting unknown/unseen interactions between a set of nodes, whose analogue in usual networks is the link prediction problem. Here, we given a brief overview of the related work in both link- as well as hyperlink-prediction.

Although research in hyperlink prediction has been limited, its literature is convincing enough to vouch for its importance. Ever since the near-seminal works by [Agarwal *et al.*, 2006] and [Zhou *et al.*, 2006] that unite the fields of hypergraphs and machine learning, there has been four major works focusing on hyperlink prediction. Xu et al. [Xu *et al.*, 2013] and both works by Zhang et al. [Zhang *et al.*, 2018; Zhang *et al.*, 2016] deal with specific domains, *viz.*, email and metabolite networks respectively. Benson et al. [Benson *et al.*, 2018], on the other hand, bring a multitude of domains to the table (see Section 6 for more details). While Zhang et al. [Zhang *et al.*, 2018] introduce a matrix completion based solution called Coordinated Matrix Minimization (CMM) that works well for a *non-temporal* network of metabolites, Benson et al. [Benson *et al.*, 2018] restrict the problem to that of predicting the closure of a 3-4 sized open simplex, which is a problem temporal in nature.

Researchers have previously worked on the task of predicting links in heterogeneous and bipartite networks as well [Kunegis *et al.*, 2010; Yu *et al.*, 2014], however, their relevance to the current work is limited since hyperlink prediction parallels neither to link prediction on such networks, nor their one-mode projections.

6 Experimental Setup

We test our algorithm (C3MM[3]) on both structural as well as temporal link prediction problems, and report results on diverse datasets, using a few baselines to compare against. But

[3]Code available at `https://github.com/govindjsk/c3mm`

Dataset	Temporal?	$\lvert V\rvert$	$\lvert F\rvert$
contact-ps	Yes	242	11,161
contact-hs	Yes	327	6,700
MAG-G	Yes	1,876	9,471
tags-ms	Yes	862	9,098
iJO1366	No	1,805	2,583
iAF1260b	No	1,668	2,388
iAF692	No	628	690
iHN637	No	698	785
iIT341	No	485	554
iAB_RBC_283	No	342	469

Table 2: Temporal and non-temporal datasets that we use in our experiments.

before that, we test our hypothesis (CCH) on these datasets, and elucidate that it holds statistically significantly for most of them. Let us describe the datasets we have used.

6.1 Datasets

We have performed our experiments on altogether *ten* datasets, of which *four* are *temporal* hypergraphs and we use the *six non-temporal metabolite* hypergraphs from Zhang et al. [Zhang *et al.*, 2018].

We provide a brief introductory overview of these datasets below. For more information, we suggest the reader to refer to Benson et al. [Benson *et al.*, 2018] for an extensive analysis of the four (and more) *temporal* datasets, and to Zhang et al. [Zhang *et al.*, 2018] for the six *metabolites* datasets. Summary is in Table 2.

1. *contact-ps* and *contact-hs*: These are *contact* networks, wherein nodes are primary/high school students and a hyperlink between them represents those observed to be within a close proximity to each other over a period of three days.
2. *MAG-G*: This refers to a *co-authorship* network, wherein nodes are authors and a hyperlink between them denotes a set of authors who have exclusively co-authored at least one paper. Since this is one of the biggest datasets we have, we reduce it using the same *core*-based filtering technique as Liben-Nowell et al. [Liben-Nowell and Kleinberg, 2003].
3. *tags-ms*: It is a hypergraph where nodes refer to tags given to question-answer threads on Math StackExchange, and a set of all tags associated a thread forms a hyperlink. We take only one year's (most recent) data for tags-ms.
4. *Metabolites*: This is a group of six datasets of *metabolic reactions*, where nodes are metabolites (reactants and products) of a metabolic reaction, which represents a hyperlink. They have been named *iJO1366*, *iAF1260b*, *iAF692*, *iHN637*, *iIT341*, and *iAB_RBC_283*.

6.2 Baselines

Coordinated Matrix Minimization (CMM) [Zhang *et al.*, 2018] as well as baseline algorithms mentioned in their paper form the baselines for our experiments. More specifically, we use the following methods as our baselines: Bayesian Sets (BS) [Ghahramani and Heller, 2006], Spectral Hypergraph Clustering (SHC) [Zhou *et al.*, 2006], Factorization Machines (FM) [Rendle, 2012], Katz [Katz, 1953], and Hyper Common Neighbors (CN) [Zhang *et al.*, 2018; Liben-Nowell and Kleinberg, 2007]. For more information we refer to Zhang et al. [Zhang *et al.*, 2018] or the respective references therein.

To evaluate the performance of hyperlink prediction algorithms, we make use of the area under ROC curves (AUC) metric.

Negative Sampling

Owing to extreme class imbalance (of ratio $\mathcal{O}\left(2^{|V|}/|F|\right)$) between non-hyperlinks and hyperlinks in hyperlink prediction, the issue of negative sampling (*i.e.*, undersampling the negative/dominant class) becomes serious, lest biased results get reported. The literature on fair evaluation of link prediction algorithms [Garcia Gasulla *et al.*, 2015; Lichtenwalter *et al.*, 2010; Lichtnwalter and Chawla, 2012; Yang *et al.*, 2015] highlights the role of negative sampling in evaluating solutions, which extends to hyperlink prediction as well. We extend Lichtenwalter et al.'s [Lichtenwalter *et al.*, 2010] geodesic-distance $\ell = 2$ (*i.e.*, one-hop neighbor) based negative-sampling technique to hypergraphs as follows [Patil *et al.*, 2020].

We sample the negative class (*i.e.*, non-hyperlinks) by randomly picking a hyperlink $f \in F$ (of size, say s) and replacing its *lowest degree* node (say $v_0 \in f$) with a *common-neighbor* (say $v' \notin f$) of *all* other $s-1$ nodes in $f_{-v_0} := f \setminus \{v_0\}$, such that $f' := (f_{-v_0}) \cup \{v'\} \notin F$. We repeat this process to extract multiple non-hyperlinks $f' \notin F$ corresponding to a hyperlink f', and stop when we have an enough (defined as a factor $p \in \mathbb{R}$ of the positive class size) number of them. This method is a simple extension of the usual one-hop ($\ell = 2$) negative sampling performed in link prediction [Lichtenwalter *et al.*, 2010].

Data Preparation

We sample 15 times as many non-hyperlinks as there are hyperlinks in the unobserved hypergraph for all of the temporal hypergraphs. For the non-temporal hypergraphs (*i.e.*, the Metabolites datasets), Zhang et al. [Zhang *et al.*, 2018] already refer to a manually curated negative class (or non-hyperlinks)[4]; hence there is no need to generate any negative samples. We fix the size of latent dimension for symmetric NMF in (6) to be $k = 30$ for the all the datasets, just as Zhang et al. [Zhang *et al.*, 2018] do as a default choice for CMM.

7 Results and Discussion

7.1 CCH Hypothesis Testing

We test our hypothesis (CCH) on a total of ten datasets, four of which are temporal, while remaining are non-temporal (Table 3). On temporal datasets, we test CCH using Algorithm 1, whereas for non-temporal ones the variation as mentioned in Section 3 is used. More specifically, for each of our datasets,

[4]Owing to the knowledge domain experts have about "impossible" metabolic reactions.

Dataset	$d(H)$	$cf(H)$	p-value	Result ($\alpha = 0.1$)
contact-ps	0.285	0.92	**0.001**	Rejects $\mathbb{H}_0$
contact-hs	0.109	0.91	**0.000**	Rejects $\mathbb{H}_0$
MAG-G	0.014	0.27	**0.059**	Rejects $\mathbb{H}_0$
tags-ms	0.028	0.52	**0.021**	Rejects $\mathbb{H}_0$
iJO1366	0.009	0.07	0.102	Fails to reject $\mathbb{H}_0$
iAF1260b	0.008	0.07	**0.033**	Rejects $\mathbb{H}_0$
iAF692	0.027	0.08	0.571	Fails to reject $\mathbb{H}_0$
iHN637	0.028	0.03	0.658	Fails to reject $\mathbb{H}_0$
iIT341	0.034	0.04	0.813	Fails to reject $\mathbb{H}_0$
iAB_RBC_283	0.030	0.04	0.591	Fails to reject $\mathbb{H}_0$

Table 3: CCH Test on temporal and non-temporal datasets with significance level $\alpha = 0.1$. All temporal datasets reject $\mathbb{H}_0$ (*i.e.*, follow CCH) with significance $\alpha = 0.1$, and all but one (*iAF1260b*) non-temporal datasets fail to reject $\mathbb{H}_0$ (*i.e.*, don't follow CCH).

we report the values of $d(H)$, $cf(H)$, and also p-values of $\mathbb{H}_0$ (Hypothesis 2) over all hyperlinks $f \in F$.

The first set of results (first four rows of Table 3) show that all temporal hypergraphs satisfy the hypothesis by a decent margin, in that $p < \alpha$. One can infer that in these settings, it is highly required for a group of nodes to have had dense lower-order interactions before the group evolves into a hyperlink. Also, as the hyperlink size increases, so does its mean pre-hyperedge density. It is therefore observed that three or four authors can relatively easily group together to collaborate on a common work, than bigger groups.

The second set of results (bottom six rows of Table 3) clearly show that metabolite datasets, which are non-temporal in nature, show little-to-no support for the hypothesis. Not much could be commented on the relative comparison between datasets since they are all equally low, wherein $cf(H)$ lies in the range of 3–8% and p-value much higher as compared to its temporal counterparts.

In summary, temporal datasets satisfy CCH with high confidences, while non-temporal ones fail miserably. The results we report in the bottom part of Table 3 are certain summaries of the static analysis of metabolite networks, which is not bound to follow a particular pattern, at least not the pattern we expect it to (namely, CCH).

7.2 Hyperlink Prediction

We present the results for hyperlink prediction on the four temporal datasets in Table 4. Table 4 reports mean AUC scores for C3MM versus CMM and its other baselines.

In all the temporal datasets, C3MM performs better than the other baselines, of which in particular, CMM (an approach that is similar to C3MM) has much lower AUC scores. This supports the argument that our hypothesis (CCH) has helped identify hyperlinks that the earlier formulation did not. Of the datasets, MAG-G and tags-math-sx have the highest scores, since they are bigger datasets and have formed over a longer time range than the other ones. Of the other baselines, we have BS (Bayesian Sets) that has a decent AUC for all datasets, except for tags-ms, and SHC seems to be the third best baseline.

One dataset that has a relatively higher p-value for CCH and despite this fact C3MM performing well is MAG-G,

Dataset	C3MM	CMM	BS	SHC	FM	Katz	CN
contact-ps	**0.590**	0.455	0.580	0.563	0.497	0.324	0.413
contact-hs	**0.629**	0.382	0.624	0.537	0.490	0.308	0.391
MAG-G	**0.639**	0.380	0.637	0.626	0.262	0.274	0.350
tags-ms	**0.638**	0.476	0.590	0.549	0.374	0.374	0.430

Table 4: AUC scores of hyperlink prediction on temporal datasets. In all cases, C3MM outperforms CMM, our main baseline. The role of CCH in helping to identify hyperlinks better is hence evident.

where we see most (73%) of the hyperlinks forming from non-cliques. This is possibly due to MAG-G being a co-authorship network where one would anticipate future collaboration among authors who have worked together in the past in some form. The higher p-value could be attributed to the fact that we take the hypergraph snapshot of recent 7 years, which ends up ignoring meaningful connections of the past.

At the same time, performance of C3MM drops for most of the non-temporal metabolite datasets. The only dataset which shows better performance for C3MM is *iAF1260b* while for the rest of the datasets performance drop is anywhere between 12% to 1%. Also it should be noted that *iAF1260b* is the only non-temporal dataset that satisfies CCH hypothesis as seen in Table 3 while the other datasets don't as shown by Table 3. This shows that C3MM is a better algorithm for hyperlink prediction when the CCH hypothesis is strongly supported by dataset.

8 Conclusion and Future Work

Hyperlink prediction is a difficult task to perform, at least more difficult than what link prediction is. This is so both due to the number of possible hyperlinks in a given hypergraph (which is exponential in the number of nodes), as well as lack of multi-way heuristic scores. We set out to improve upon the current state-of-the-art (CMM) by introducing a clique-closure hypothesis into its objective function, ultimately forming C3MM. It is clear from the results on the hypothesis tests that we succeed in validating that it is cliques and co-cliques that close to form hyperlinks, instead of they being formed by co-cliques or disconnected structures. Embedding the hypothesis into the objective function leads to it significantly hunting down more hyperlinks which were missed by CMM. Another conclusion we draw is that hyperlink prediction on temporal and non-temporal datasets works differently, in that the latter predicts the future, and the former, the missing hyperlinks. While CMM works well on non-temporal datasets, C3MM better predicts future links.

In an extension, we would like to extract more concepts from baselines such as Bayesian Sets and Spectral Hypergraph Clustering and incorporate them into C3MM, to hopefully improve further. Also, we would aim to work with more variety of networks: heterogeneous hypergraphs, directed hypergraphs, and weighted hypergraphs.

Acknowledgements

Prasanna Patil was supported by a fellowship grant from the Centre for Networked Intelligence (a Cisco CSR initiative) of the Indian Institute of Science, Bengaluru.

References

[Agarwal *et al.*, 2006] Sameer Agarwal, Kristin Branson, and Serge Belongie. Higher order learning with graphs. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 17–24, New York, NY, USA, 2006. ACM.

[Benson *et al.*, 2018] Austin R Benson, Rediet Abebe, Michael T Schaub, Ali Jadbabaie, and Jon Kleinberg. Simplicial closure and higher-order link prediction. *arXiv preprint arXiv:1802.06916*, 2018.

[Berge, 1984] Claude Berge. *Hypergraphs: combinatorics of finite sets*, volume 45. Elsevier, 1984.

[Garcia Gasulla *et al.*, 2015] Dario Garcia Gasulla, Claudio Ulises Cortés García, Eduard Ayguadé Parra, and Jesús José Labarta Mancho. Evaluating link prediction on large graphs. In *Artificial Intelligence Research and Development: Proceedings of the 18th International Conference of the Catalan Association for Artificial Intelligence*, pages 90–99. IOS Press, 2015.

[Ghahramani and Heller, 2006] Zoubin Ghahramani and Katherine A Heller. Bayesian sets. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 435–442. MIT Press, 2006.

[Katz, 1953] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, Mar 1953.

[Kuang *et al.*, 2012] Da Kuang, Chris Ding, and Haesun Park. Symmetric nonnegative matrix factorization for graph clustering. In *Proceedings of the 2012 SIAM international conference on data mining*, pages 106–117. SIAM, 2012.

[Kunegis *et al.*, 2010] Jérôme Kunegis, Ernesto William De Luca, and Sahin Albayrak. The link prediction problem in bipartite networks. *CoRR*, abs/1006.5367, 2010.

[Liben-Nowell and Kleinberg, 2003] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 556–559, New York, NY, USA, 2003. ACM.

[Liben-Nowell and Kleinberg, 2007] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[Lichtenwalter *et al.*, 2010] Ryan N Lichtenwalter, Jake T Lussier, and Nitesh V Chawla. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–252. ACM, 2010.

[Lichtnwalter and Chawla, 2012] Ryan Lichtnwalter and Nitesh V Chawla. Link prediction: fair and effective evaluation. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 376–383. IEEE Computer Society, 2012.

[Newman, 2001] Mark EJ Newman. Clustering and preferential attachment in growing networks. *Physical review E*, 64(2):025102, 2001.

[Patil *et al.*, 2020] Prasanna Patil, Govind Sharma, and M Narasimha Murty. Negative sampling for hyperlink prediction in networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 607–619. Springer, 2020.

[Rendle, 2012] Steffen Rendle. Factorization machines with libfm. *ACM Trans. Intell. Syst. Technol.*, 3(3):57:1–57:22, May 2012.

[Xu *et al.*, 2013] Ye Xu, Dan Rockmore, and Adam M Kleinbaum. Hyperlink prediction in hypernetworks using latent social features. In *International Conference on Discovery Science*, pages 324–339. Springer, 2013.

[Yang *et al.*, 2015] Yang Yang, Ryan N Lichtenwalter, and Nitesh V Chawla. Evaluating link prediction methods. *Knowledge and Information Systems*, 45(3):751–782, 2015.

[Yu *et al.*, 2014] Philip S. Yu, Bing Liu, and Ouri Wolfson. Link prediction across heterogeneous social networks: A survey. 2014.

[Zhang *et al.*, 2016] Muhan Zhang, Zhicheng Cui, Tolutola Oyetunde, Yinjie Tang, and Yixin Chen. Recovering metabolic networks using a novel hyperlink prediction method. *arXiv preprint arXiv:1610.06941*, 2016.

[Zhang *et al.*, 2018] Muhan Zhang, Zhicheng Cui, Shali Jiang, and Yixin Chen. Beyond link prediction: Predicting hyperlinks in adjacency space. AAAI, 2018.

[Zhou *et al.*, 2006] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 1601–1608, Cambridge, MA, USA, 2006. MIT Press.