Higher-Order Systems Analysis applications to the knowledge graph analysis

Liubov Tupikina and collaborators

Abstract

Hypergraph theory has been established in recent years as the primary framework to capture higher-order interactions in complex systems. However, due to computational limitations, many standard techniques rely on lower-order projections of hypergraphs that can be treated with conventional network theory methods [2,4]. A common example of such an order-reducing procedure is the replacement of higher-order hyperedges by cliques of pair-wise edges. This is equivalent to performing trace operations (summing over indices) on the adjacency tensor – containing the full information of a hypergraph – until a single 2-index matrix is obtained. Such matrices form associative algebras, which are mathematically well-understood and afford powerful analytical tools such as spectral theory. Despite these advantages, the information loss of order-reduction procedures may obscure patterns emerging from irreducible higher-order interactions. Besides the information loss while treating the projected structures, there are also computational issues during working with large hypegraphs [Evans paper on innovation hypergraphs], or processing the dynamically updating graphs [kappalanguage.org link] applying higher order theory for hypergraphs analysis, would be really beneficial.

Specifically applied to the knowledge and innovation space analysis, one has shown that a focus on lower-level, dyadic interactions among knowledge components is not coping with enormous volumes of prior scientific and technological knowledge, which requires different lenses for seeing significant moves and comprehending the state of scientific revolutions.

To address these challenges we propose a novel approach to the mathematical and computational treatment of hypergraphs based on rewriting systems and higher-arity algebra [3,6]. In some cases of rewriting systems we can extract some invariants (non-local, counts of motifs, local patterns). Even though it is difficult to say something concrete about general rewriting systems for hypergraphs, one can perform computational frameworks for hypergraphs structures.

1 Methods

Here we construct hypergraphs from the datasets of scientific articles as well as from patents. The hypergraph $H(V,E)$ construction from the scientific articles dataset is generalized from the graph construction explained in [Singh et al. 2023, Muscitto et al.2021].

1.1 Weighted hypergraph

We construct a weighted hypergraph given the set of sets of articles using similar construction as for the hypergeometric p-value. We assign a weight to each hyperedge of a a hypergraph, which corresponds to the hypergeometric p-value assigned to the overlapping sets.

Based on the weighted hypergraph we can generate the projection to the simple weighted graph. There are several ways of doing such projection. Edge paths and distances (not cognitive ones) were defined in Distances in Higher-Order Networks and the Metric Structure of Hypergraphs [Vasilieva et al.].

1.2 Paths in weighted hypergraphs

We first will need to specify the notion of path for weighted hypergraphs, and how we estimate the shortest path for such hypergraphs.

Definition 1. Hypergraph path in $H=(V,E)$ between two distinct vertices i,j is a sequence $i,e_1,...,e_k,j$, where A_i, have the following properties: (i) $k>1$ (the number of edges on the path) is a positive integer; (ii) $i,...,j$ are distinct vertices.

1.3 Distances and metrics in hypergraphs

One of the main problems we are interested to investigate is the representation of hypergraph structures by lower-order structures. This problem can be formulated and formalized as follows.

Problem 1. Given a hypergraph H and its projection P into the graph G, what should be the set of properties of projection P of a hypergraph such that metrics M introduced on the hypergraph H keeps its properties as projected metrics M' on a graph G. By metrics M for a hypergraph H we mean a function defined on a set of nodes of a hypergraph H:

$$M(xi, xj) = f(x_i, x_j, n(x_i), x(x_j)),$$

where x_i and x_j are elements from the set of nodes X of hypergraph $H(X, E)$ and graph G such that it fullfills the triangle inequality, function f is positively defined, $n(x_j)$ is defined as the neighboring . COMMENT: this can be special in the case for the graphs.
By metrics M' for a graph H we mean a function defined on a set of nodes of a hypergraph H: M(xi,xj), where xi and xj are elements from the set of nodes of hypergraph H and graph G. This problem can be formalised into a problem of finding the correspondence between all sets of hypergraphs on a set of nodes X and sets of graphs with metrics on the same set of nodes X.

Problem 2 (variation of Problem 1). Given a hypergraph H and its projection P into the lower-arity hypergraph H', what should be the set of properties of projection P of a hypergraph such that metrics M introduced on the hypergraph H keeps its properties as projected metrics M' on a hypergraph H' of a specific arity k.

1.4 Cognitive distance for hypergraphs

We generalize the concept of cognitive distance for hypergraph H using the notion of the generalized path on a hypergraph $H(V, E)$ between two nodes $i, j \in V$.
There are several results, which connect path on projected associated hypergraph and associated graph $G(V, E)$ [Dhamarujan 2014].
Definition 2. We first remind the definition of the cognitive distance definition for the graph G. Cognitive distance $C(i, j)$ between field tags i and j as the weighted shortest path $C(i, j) = \sum_e 1/W_e$, where e are the edges on the shortest path.
Definition 3. Cognitive distance on the hypergraph is then defined by the analogy. Cognitive distance $C(i, j)$ between field tags i and j as the weighted shortest path

$$C(i, j) = \sum_e 1/W_e,$$

where e are the hyper-edges on the shortest hyper-path of H according to the Definition 1.

1.5 Null models for hypergraphs

In order to see the significance of the results of the estimated measures, we need to compare it with the null model. For this we need to identify rewiring of edges and nodes for the initial hypergraph.

2 Rewriting systems

There are currently some issues with motifs analysis such as the following: 1) considered motifs are of simple topology and of small cardinality, such as intersections of maximum three hyperedges [Hypergraph Motifs: Concepts, Algorithms, and Discoveries Lee et al.]. 2) distributions of h-motifs' (hyperedges combinations) instances precisely characterize local structural patterns of real-world hypergraphs, yet are not often linked to potential topological analysis of hypergraphs H.
Here we link the h-motifs structure to the hypergraphs to motifs of three types for a hypergraph H, illustrated as follows [Figure 1]: 1. BMP related motif, 2. generalization of BMP with n number of components, 3. lollipop generalization of the motif, 4. blades motifs and their generalizations.
Patterns counting for patterns on 4 nodes: In fact, the number of possible non-isomorphic patterns of higher-order interactions involving 4 nodes is 171, a number that makes all the relabelings storable in memory.

3 Results

3.0.1 Cognitive distance for hypergraphs

We test whether the distribution of the cognitive distance (and not just the first moment of it, average) are giving different results for H than for $G = P(H)$.
The distributions for the hypergraph H cognitive distance C_{ij} distribution with the first moment $\overline{(C_{ij})}$ which may give the same value as for the graph $G = P(H)$, yet the values of the distribution may be different:

$$\overline{(C_{ij})}_H vs. \overline{(C_{ij})}_G,$$

where $\overline{C}$ denotes the averaging of the distribution of C_{ij}. For the weighted graph and hypergraph cases, the correspondence between $\overline{(C_{ij})}_H$ and $\overline{(C_{ij})}_G$ is more intricate, since the weighted projection of the hypergraph is less strictly defined.

Here we aim to generalize and study the emergence of scientific novelty and scientific landscape using the rewriting systems.

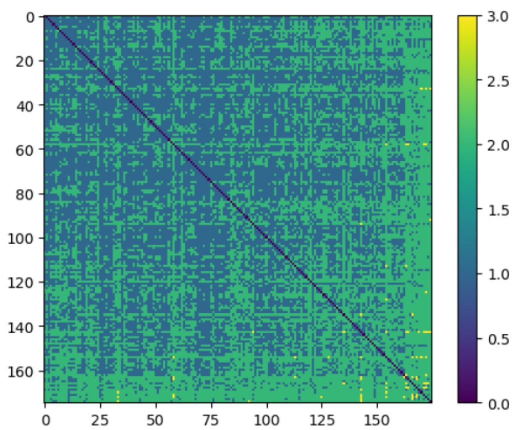

Figure 1: The distributions estimated from the graph and higher-order structure representation of the data. Graph is constructed using the clique projection method. The generalised cognitive distance measure is estimated for the hypergraph and for the graph (as expected, we get the same result, when we use unweighted version of the hypergraph).

4 Further steps

We are interested to extend and explore the possible metrics on the knowledge hypergraphs to characterize the emergence of new scientific fields. The cognitive distance measure can be extended by the vector centrality measure for hypergraphs [Kovalenko et al. 2022] as well as other measures, which could characterize the ternary and higher order structures of knowledge landscape, not captured by network centralities (in progress).
We also would like to explore the algebras for hypergraphs defined for the extended centrality measures.
SciBERT (autoencoder) presumably embeds data of arxiv data to the higher order .

[1] F. Battiston, G. Cencetti, et al. I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, G. Petri, Networks beyond pairwise interactions: Structure and dynamics, Phys. Rep., 874 (2020)
[2] A. Bretto, Hypergraphs theory: An Introduction, Springer, 2192-4732 (2013)
[3] C. Zapata-Carratala, et al., Hypermatrix Algebra and Irreducible Arity in Higher-Order systems, Adv. Compl. Sys. (2023)
[4] T. Carletti, F. Battiston, G. Cencetti, D. Fanelli, Random walks on hypergraphs, Phys. Rev. E 101, 022308 (2020)
[5] C. Singh, L. Tupikina, F. Lecuyer, et al. M. Starnini, M. Santolini, Charting mobility patterns in the scientific knowledge landscape, EPJ (in rev.), arxiv.org.2302.13054 (2023)
[6] N. Behr, V. Danos, I. Garnier, Stochastic mechanics of graph rewriting, Proceedings of the 31st Annual ACM/IEEE (2016)
[7] E. K. Gnang, Y. Filmus, On the Bhattacharya-Mesner rank of third order hypermatrices, Lin. Alg. and Appl., 588 (2020)
[8] A. Iranmanesh et al. The combinatorial and algebraic structure of the hypergroup associated to a hypergraph Article in Journal of Multiple-value

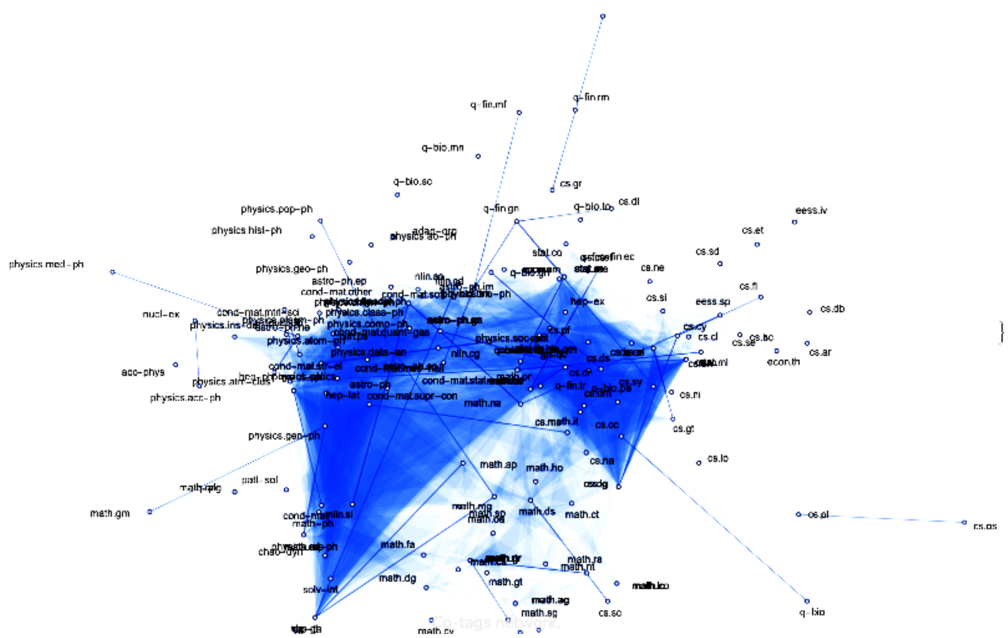

Figure 2: The illustration of the knowledge hypergraph reconstructed from the arxiv data.