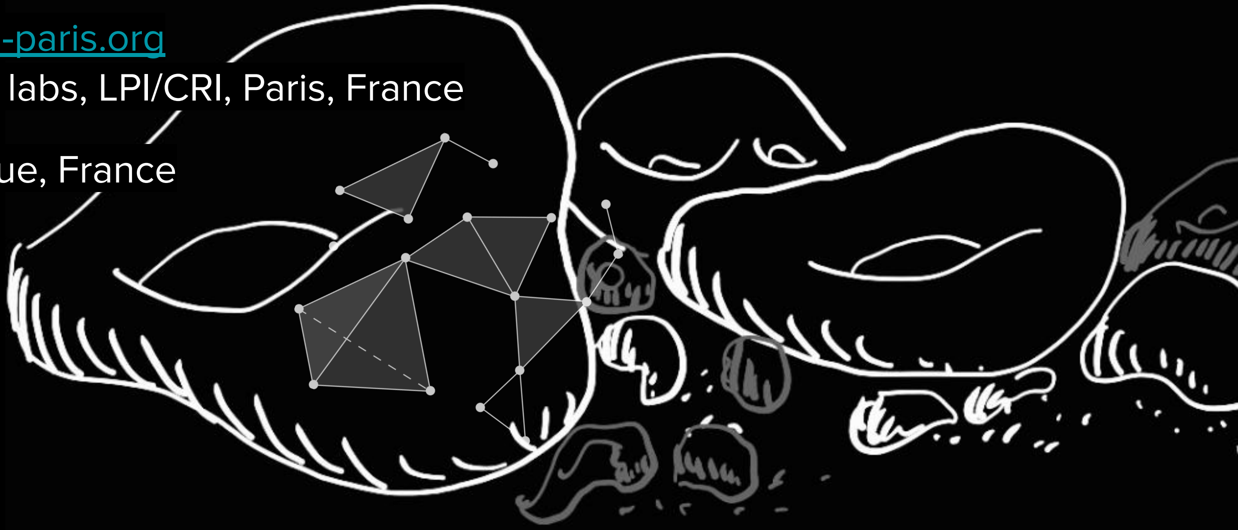
Dissecting embedding method: learning higher-order structures from data

Wolfram Foundation, Wolfram Seminar Community talk

Liubov Tupikina liubov.tupikina@cri-paris.org
ITMO, Snt. Petersburg, Russia, Bell labs, LPI/CRI, Paris, France

Hritika Kathuria, Ecole Polytechnique, France



$'$ — отрицание: p' — не p;

$\wedge$ — конъюнкция: $p_1 \wedge p_2$ — p_1 и p_2;

$\vee$ — дизъюнкция: $p_1 \vee p_2$ — p_1 или p_2;

$\rightarrow$ — импликация: $p_1 \rightarrow p_2$ — если p_1, то p_2;

$\leftrightarrow$ — эквивалентность: $p_1 \leftrightarrow p_2$ — p_1 эквивалентно p_2,

$\bigwedge_a$ — квантор общности: $\bigwedge_a \Pi(a)$ — все a удовлетворяют функции-высказыванию $\Pi(a)$;

$\bigvee_a$ — квантор существования: $\bigvee_a \Pi(a)$ — существует a, удовлетворяющее функции-высказыванию $\Pi(a)$;

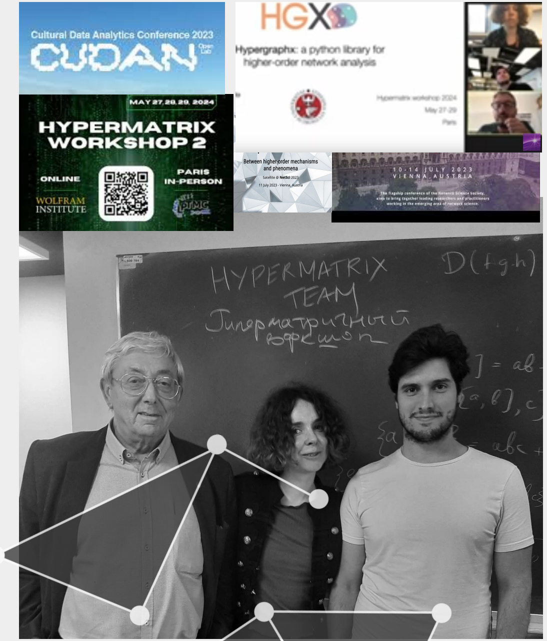
$\bigwedge_{\Pi_1(a)}$ — ограниченный квантор общности: $\bigwedge_{\Pi_1(a)} \Pi_2(a)$ — все a, удовлетворяющие функции-высказыванию $\Pi_1(a)$, удовлетворяют функции-высказыванию $\Pi_2(a)$ или

$$\bigwedge_{\Pi_1(a)} \Pi_2(a) \leftrightarrow \bigwedge_a (\Pi_1(a) \rightarrow \Pi_2(a));$$

$\bigvee_{\Pi_1(a)}$ — ограниченный квантор существования: $\bigvee_{\Pi_1(a)} \Pi_2(a)$ — среди a, удовлетворяющих функции-высказыванию $\Pi_1(a)$, существует такое, которое удовлетворяет функции-высказыванию $\Pi_2(a)$, или

$$\bigvee_{\Pi_1(a)} \Pi_2(a) \leftrightarrow \bigvee_a (\Pi_1(a) \wedge \Pi_2(a)).$$

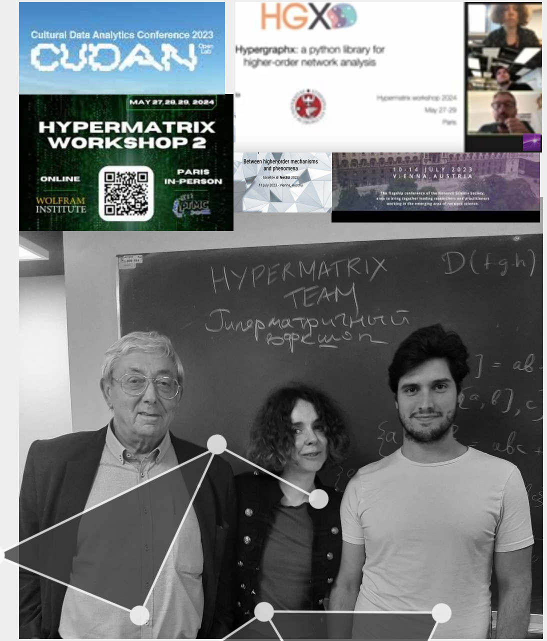
- **Main question:** how to encode embeddings from data using higher-orders structures?



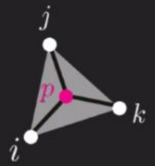
- **Main question**: how to encode embeddings from data using n-ary hypergraph structures?

- **Collaborations**: Zapata (Wolfram Institute, SEMF), Santolini (LPI), Kathuria (Bell labs), Kerner (LPTMC), Bocalleti team (MIPT)

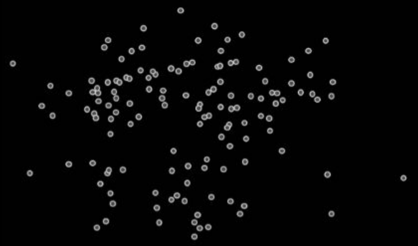
- **Events** [CUDAN lab conference, Estonia] 2023
  [Hypermatrix workshop 2024] (Wolfram Institute, Paris, France)
  NetSci Conference, Canada 2024
  [presentation]
  Embed-days (ENS, MIPT, hybrid) 2025



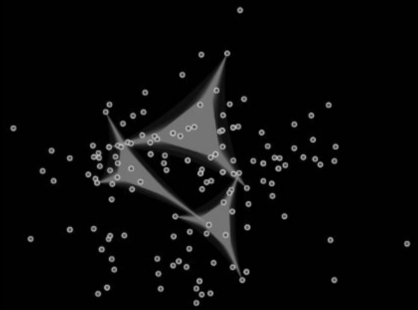
Relation to other Wolfram Institute projects

Hypergraph Motif	Adjacency Condition	Hypermatrix Operation	Higher-Arity Algebra
		$B_{ij} = \sum_{p=1}^{N} A_{ip} \cdot A_{pj}$	binary associative (ordinary matrix multiplication)
		$B_{ijk} = \sum_{p=1}^{N} A_{ip} \cdot A_{jp} \cdot A_{kp}$	ternary non-closed (triple index contraction)
		$B_{ijk} = \sum_{p=1}^{N} A_{ijp} \cdot A_{ipk} \cdot A_{pjk}$	ternary non-associative (Bhattacharya-Mesner algebra)

Relation to Wolfram Institute projects

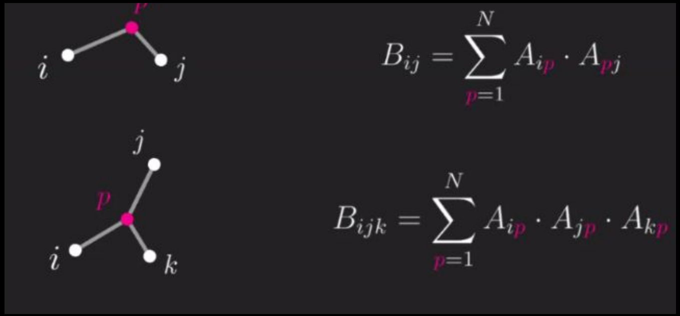


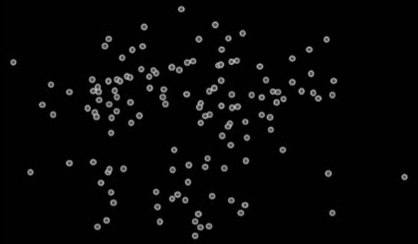
`OnePassRewriting[BladesRule, Hypdatacut]`



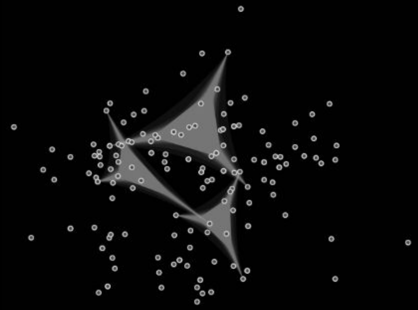
```
BMIN=Hypergraph[{1,2,3,4},Hyperedges[{1, 2, 4}, {1, 4, 3}, {4, 2, 3}]];
BMOUT=Hypergraph[{1,2,3,4},Hyperedges[{1,2,3}]];
BMRule=HypergraphRule[BMIN,BMOUT]
```

Zapata et al. An Invitation to Higher Arity Science, Complex Systems 32(4) 2024 Zapata, LT (2024) NetSci Conference, Canada 2024 youtube presentation



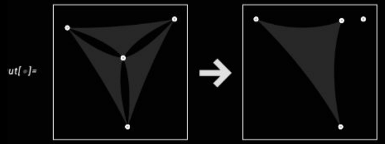


```
OnePassRewriting[BladesRule, Hypdatacut]
```

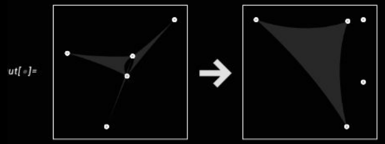


```
BMIN=Hypergraph[{1,2,3,4},Hyperedges[{1, 2, 4}, {1, 4, 3}, {4, 2, 3}]];
BMOUT=Hypergraph[{1,2,3,4},Hyperedges[{1,2,3}]];
BMRule=HypergraphRule[BMIN,BMOUT]
```

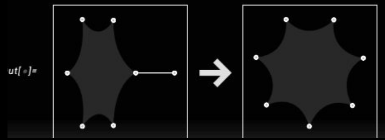
```
BMRule=HypergraphRule[BMIN,BMOUT]
```



```
In[∘]:= BladesIN=Hypergraph[{1,2,3,4,5},Hyperedges[{1, 2, 3}, {2, 3, 4}, {2, 3, 5}]];
BladesOUT=Hypergraph[{1,2,3,4,5},Hyperedges[{1,4,5}]];
BladesRule=HypergraphRule[BladesIN,BladesOUT]
```

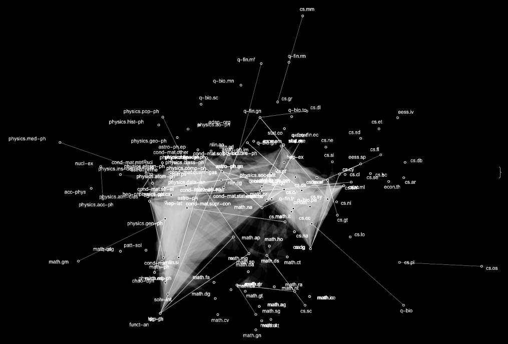


```
In[∘]:= LollipopIN=Hypergraph[{1,2,3,4,5,6,7},Hyperedges[{1, 2, 3,4,5,6}, {6,7}]];
LollipopOUT=Hypergraph[{1,2,3,4,5,6,7},Hyperedges[{1, 2, 3,4,5,6,7}]];
LollipopRule=HypergraphRule[LollipopIN,LollipopOUT]
```



Outline

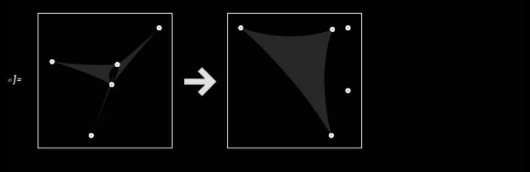
Part 1:
Formalisation of the problem of the embedding of data into low dimensional spaces



Outline

Part 1:
Formalisation of the problem of the embedding of data into low dimensional spaces

Part 2:
Development of theory for higher-order structures and its adaptation to analysis of embeddings

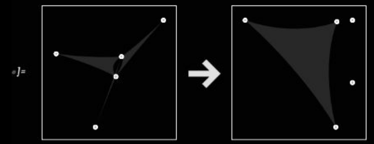


Part 3:
Applications to science of science data

Take-home message

Part 1:
Formalisation of the problem of the embedding of data into low dimensional spaces

Part 2:
Development of the higher-order structures and its adaptation to analysis of embeddings

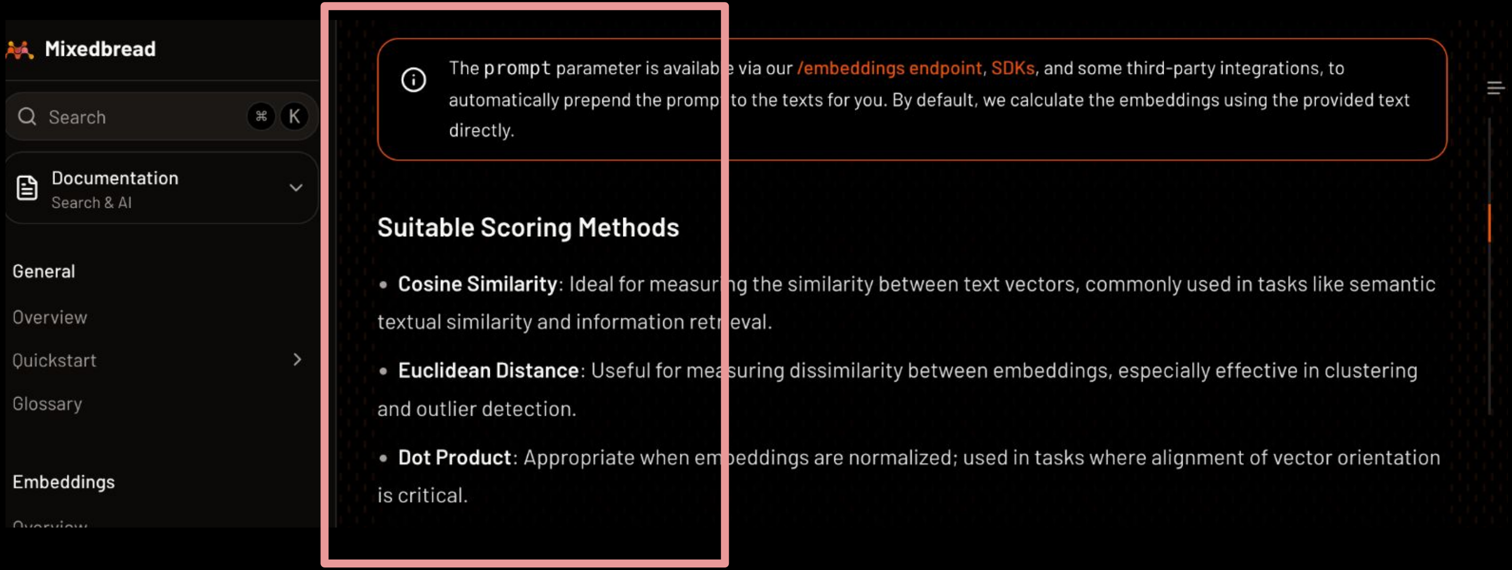


Part 3:
Applications to science of science data

Problem of embeddings theory

Part 1:
Formalisation of the problem of the embedding of data into low dimensional spaces
...

Why some of the embedding techniques are not giving us good results?



Welearn project, LPI

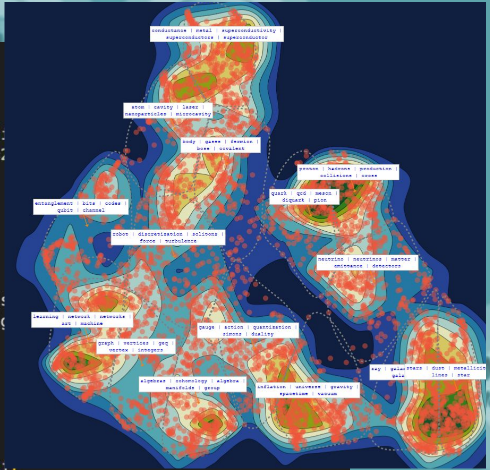
Why some of the embedding methods do not work as well as theory promises?

[colab]

[github]

title	categories	abstract
sparsity-certifying graph decompositions	math.co cs.cg	we describe a new algorithm, the (k, ℓ)-pe...
a limit relation for entropy and channel capac...	quant-ph cs.it math.it	in a quantum mechanic[illegible] feldmann...
intelligent location of simultaneously active ...	cs.ne cs.ai	the intelligent acoustic emission locator is d...
intelligent location of simultaneously active ...	cs.ne cs.ai	part i describes an intelligent acoustic emiss...
on-line viterbi algorithm and its relationship...	cs.ds	in this paper, we introduce the on-line viterb...

```
def jump_distance_vector(trajectory):
    j_dist=[]
    for i in range(len(trajectory)-1):
        p1,p2=trajectory[i],trajectory[i
        eucl = np.sqrt(((p2[0]-p1[0])**2
        j_dist.append(eucl)

    return j_dist

def z_scoreD(dist_vec):

    u,sig = np.mean(dist_vec),np.std(dis
    zscore = list(map(lambda x:(x-u)/sig

    return zscore

    nce_from0_vector(trajectory):

    i in range(1,len(trajectory)):
        p0,p1=trajectory[0],trajectory[i]
        eucl = np.sqrt(((p1[0]-p0[0])**2) + ((p1[1]-p0[1])**2))
        dist.append(eucl)

    return dist

def jump_duration(days):
    time=[]
    for i in range(len(days)-1):
        d0,d1=days[i],days[i+1]
        time.append(d1-d0)
```

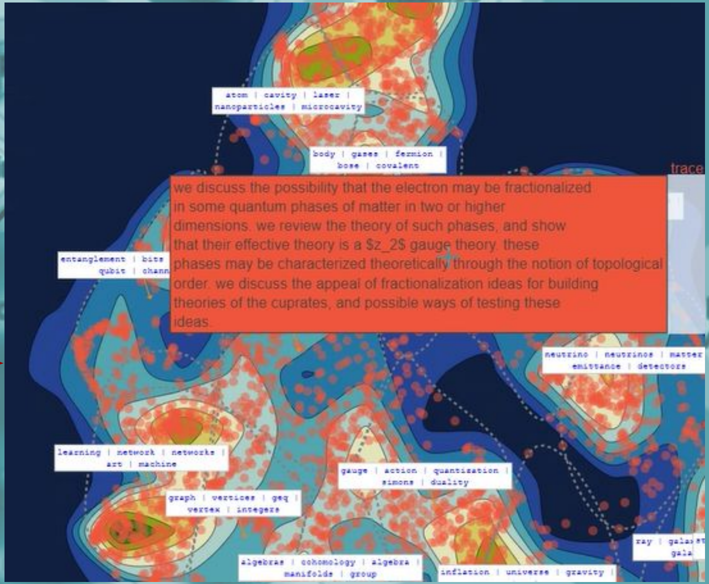


Why some of the embedding methods do not work as well as theory promises?

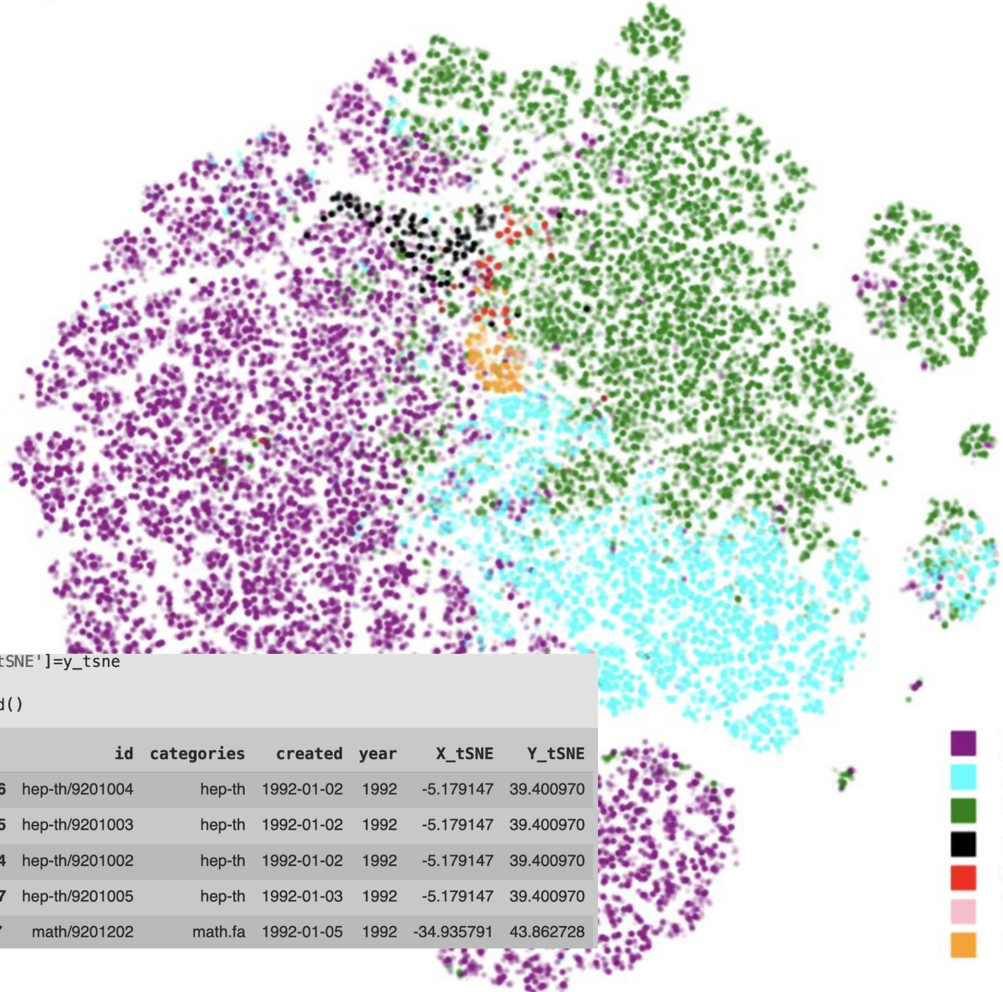
[colab]

[github]

title	categories	abstract
sparsity-certifying graph decompositions	math.co cs.cg	we describe a new algorithm, the (k, ℓ)-pe...
a limit relation for entropy and channel capac...	quant-ph cs.it math.it	in a quantum mechanica... feldmann...
intelligent location of simultaneously active ...	cs.ne cs.ai	the intelligent acoustic emission locator is d...
intelligent location of simultaneously active ...	cs.ne cs.ai	part i describes an intelligent acoustic emiss...
on-line viterbi algorithm and its relationship...	cs.ds	in this paper, we introduce the on-line viterb...



a

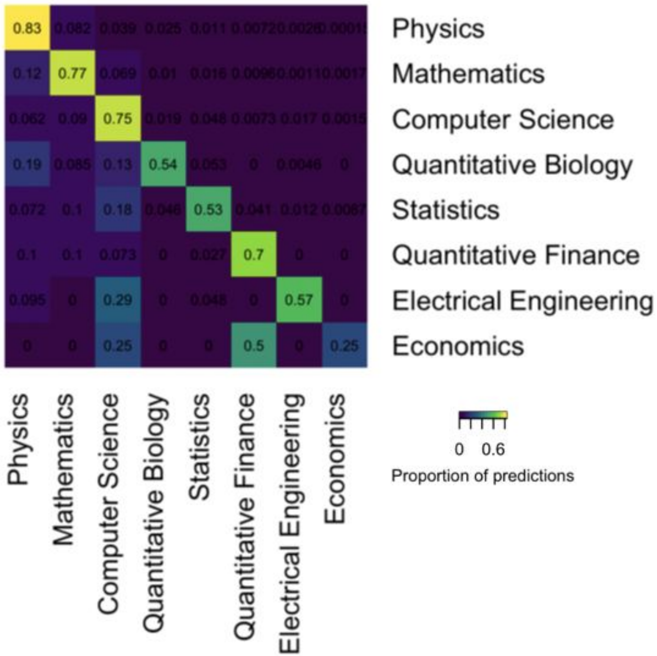


```
DF['Y_tSNE']=y_tsne

DF.head()
```

	id	categories	created	year	X_tSNE	Y_tSNE
1251136	hep-th/9201004	hep-th	1992-01-02	1992	-5.179147	39.400970
1251135	hep-th/9201003	hep-th	1992-01-02	1992	-5.179147	39.400970
1251134	hep-th/9201002	hep-th	1992-01-02	1992	-5.179147	39.400970
1251137	hep-th/9201005	hep-th	1992-01-03	1992	-5.179147	39.400970
526877	math/9201202	math.fa	1992-01-05	1992	-34.935791	43.862728

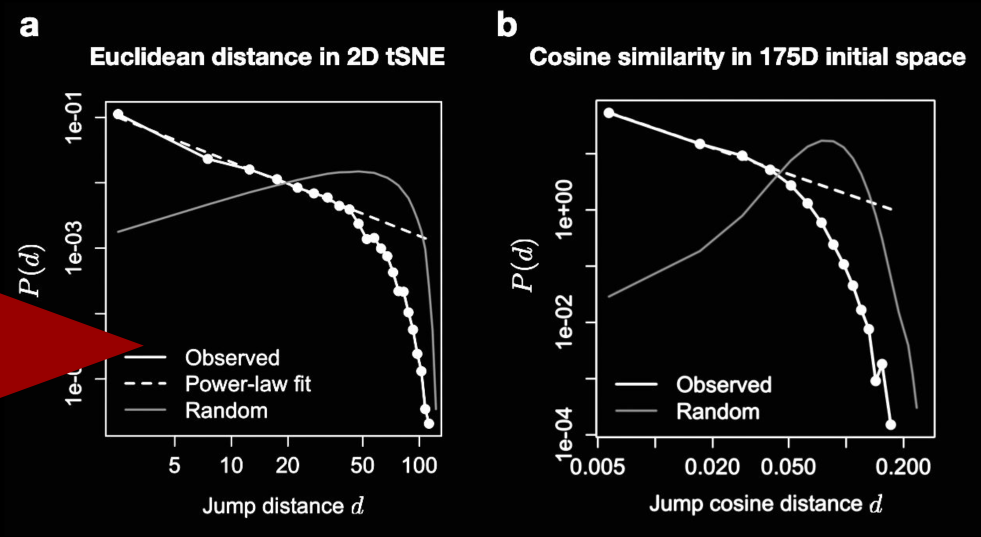
Physics
Mathematics
Computer Science
Quantitative Biology
Statistics
Economics
Quantitative Finance

b



Singh, [...], **Santolini+** PLOS ONE 2022
Singh, [...], **Santolini+** EPJ (in rev) 2023

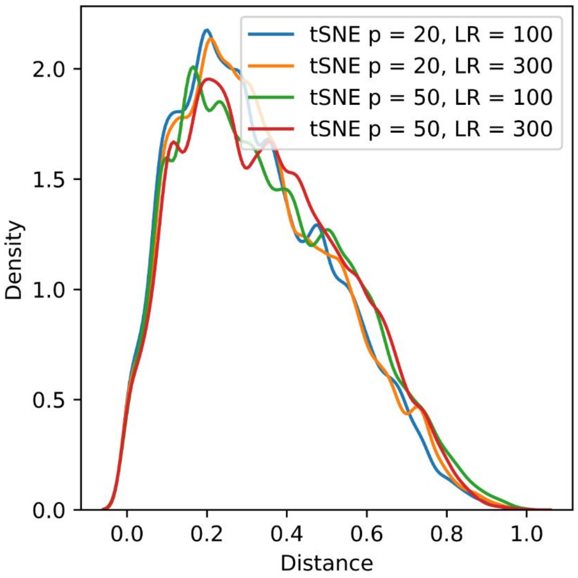
	topic_id	topic_name	size	percent
0	bt-6	conductance | metal | superconductivity | supe...	462	9.24
1	bt-5	learning | network | networks | art | machine	408	8.16
2	bt-1	ray | galaxies | redshift | galaxy | radio	394	7.88
3	bt-7	inflation | universe | gravity | spacetime | v...	375	7.50
4	bt-12	stars | dust | metallicity | lines | star	373	7.46



Singh, L. Tupikina, M. Starnini, M. Santolini, Charting mobility patterns in the scientific knowledge landscape, EPJ D.S. 13,12 (2024)
M. Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure (2022)
Kobak et al. The art of using t-SNE for single-cell (2019)

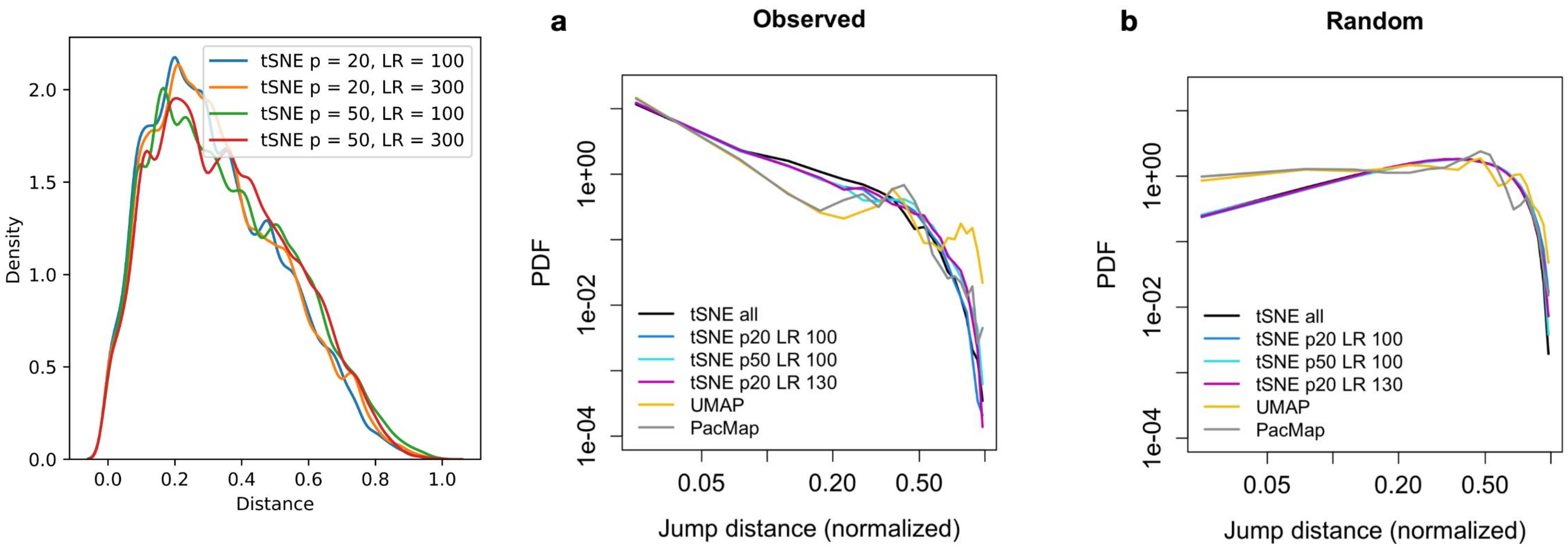
How to characterize accuracy of embeddings?

First method: compare jump distance distributions from embeddings.



How to characterize embedding methods?

First method: compare jump distance distributions from embeddings, **yet this considers only pairwise data relations.**



Singh et al. EPJ (2024), Vasilieva, LT et al. (in progr)

Challenges for explainable embedding analysis

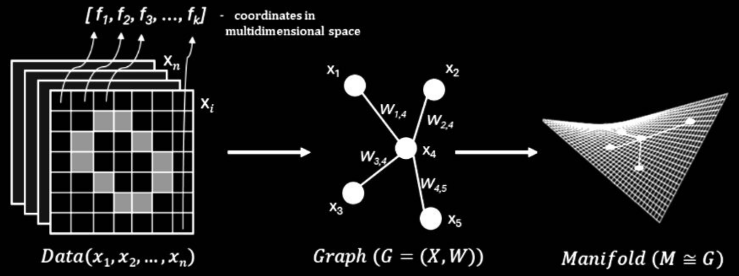
1. Manifold learning assumes that manifold properties are controlled mainly by the learning parameters, not data properties.

2. Initial feature space is assumed with Euclidean metric, and is transferred to a manifold.

3. Distance metric learning algorithms capture data relations from manifold structures.

Melas-Kyriazi, Mathematical Foundations of Manifold Learning (2020)

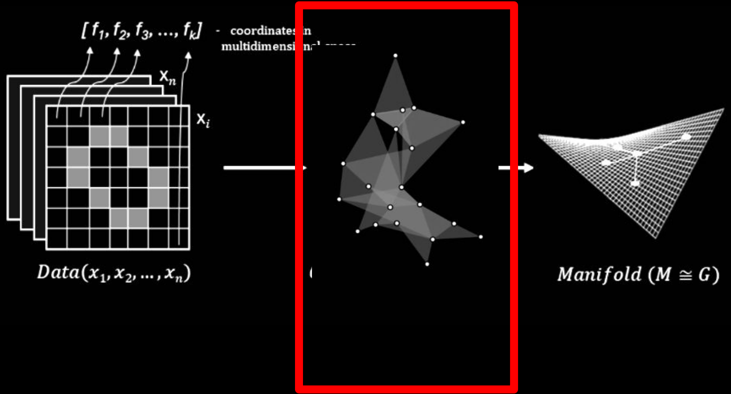
Proposed pathway to overcome challenges in manifold learning and explainable embeddings



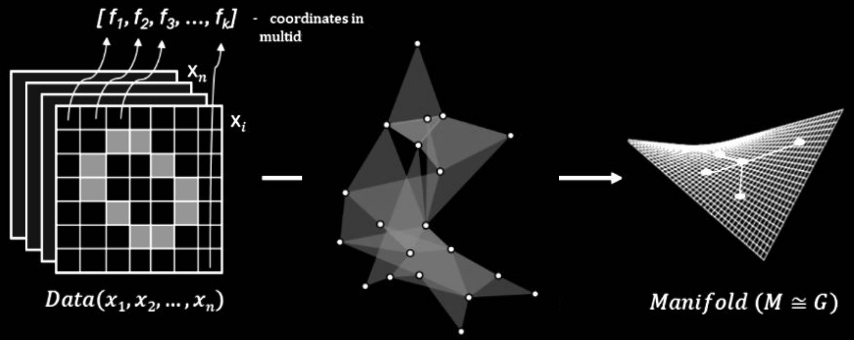
Theorem 2.1.1 (No Free Lunch Theorem). *No learning algorithm achieves uniform universal consistency. That is, for all* $\varepsilon > 0$*:*

$$\lim_{n\to\infty} \sup_{\rho} \mathbb{P}\left(\mathcal{E}(\hat{f}) - \inf_{f\in\mathcal{F}} \mathcal{E}(f) > \varepsilon\right) = \infty$$

Proposed method to overcome challenges in manifold learning and explainable embeddings



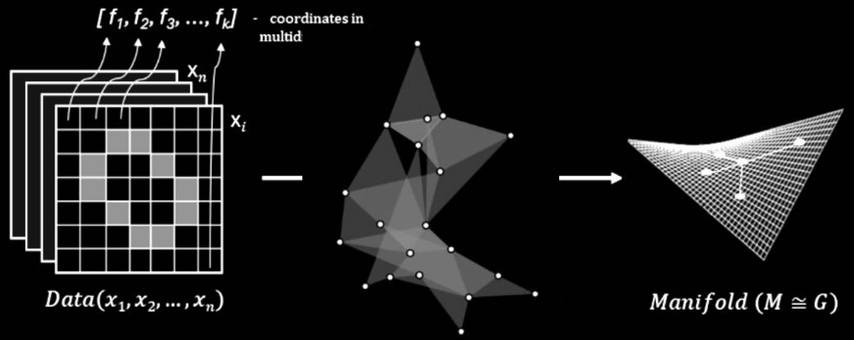
Proposed method



- Hypergraph H instead of G structure

- Use combinatorial approach instead of topological space analysis, which would require choice of (L2) metrics
- Hypergraph measures for the latent space representation e.g. motifs analysis

Proposed method for dissecting embedding methods



- Hypergraph H instead of G structure
- Use combinatorial approach instead of topological space analysis, which would require choice of (L2) metrics
-
- Hypergraph measures for the latent space representation e.g. motifs analysis

Main evidence:
Neural networks, LLMS can be successively trained in lower-dimensional subspaces (extracted curated manifolds).
It is not necessary to update millions of parameters on small fine-tuning datasets (Razhigaev et al. 2024 ACL, de Dampierre et al. arxiv 2024)

SOTA methods for reconstruction of lower-dimensional embedding space

Phase 1: Construct a graph representing the data

Phase 2: Embed this graph into a lower dimensional space

Method for reconstruction of knowledge space

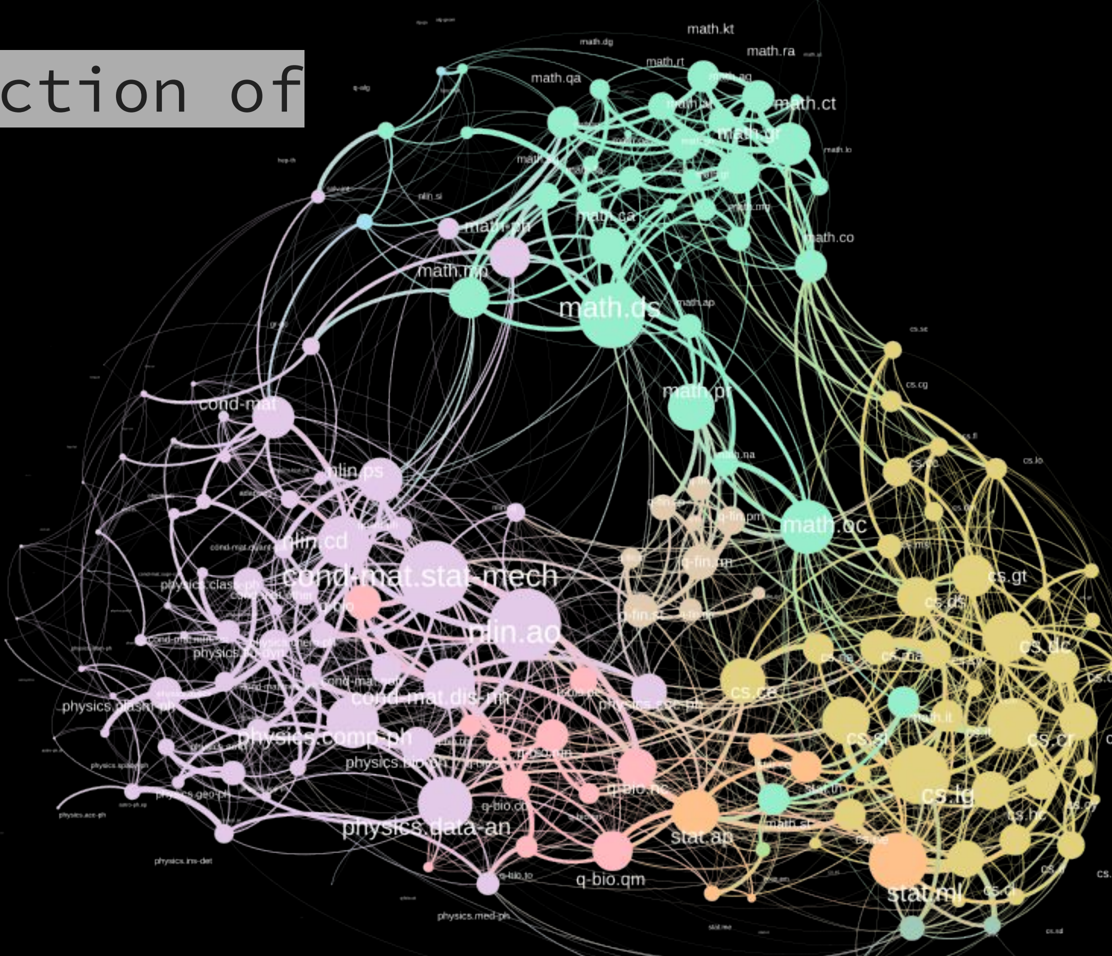
Phase 1: Construct a graph representing data

Phase 2: Embed this graph into a lower dimensional space

$$p_{ij} = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Edge Weight = $-log_{10}(p_{ij})$

N - total articles, K - articles in field i, n - articles in field j, k - common articles bw i, j



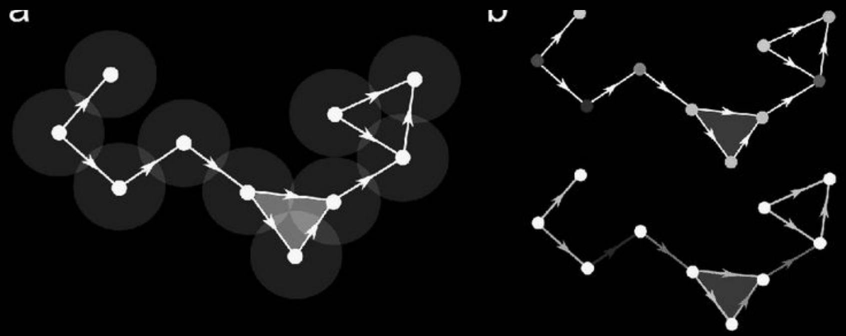
Co-tags network, Singh et al. Netsci 2022, Plos One 2023
UMAP McInnes et al. 2020

Example of UMAP method

Phase 1: Construct a graph representing data

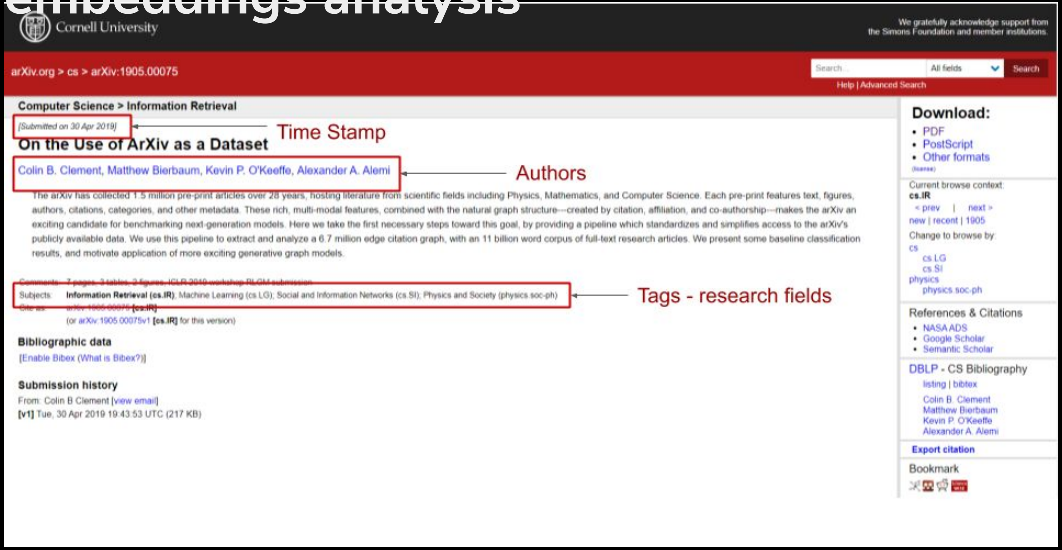
Phase 2: Embed this graph into a lower dimensional space

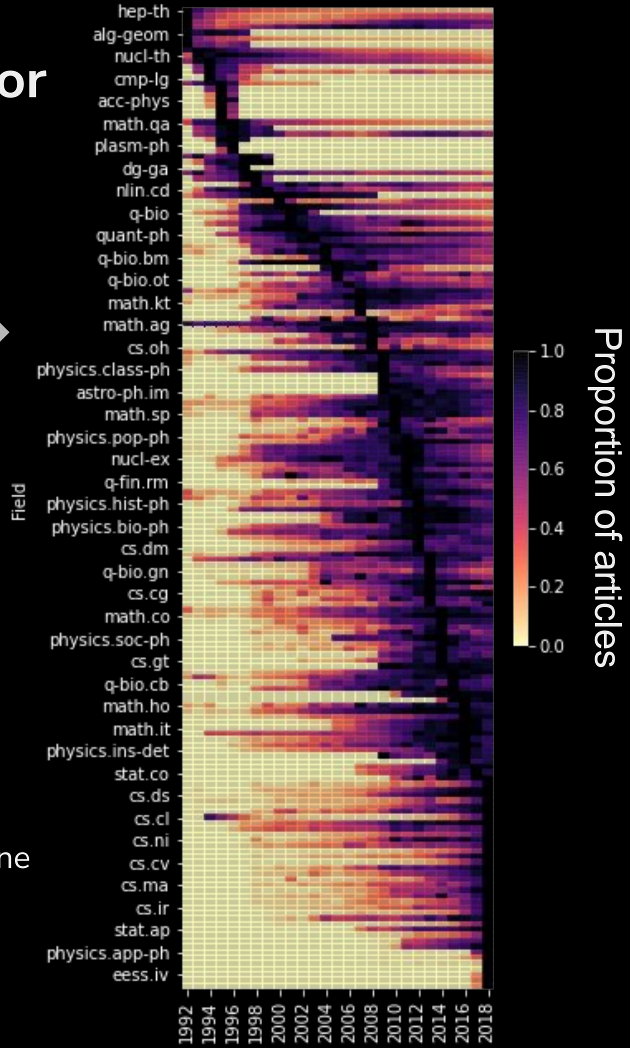
Theorem 1 (Nerve). Let $U = \{U_i\}_i \in I$ be a cover of topological space X. If, for all $\sigma \in I : \cap_{i \in \sigma} U_i$ is either contractible or empty, then $N(U)$ is homotopically equivalent to X.



$$C((A,\mu),(A,\nu)) \triangleq \sum_{a \in A} \left(\mu(a) \log \left(\frac{\mu(a)}{\nu(a)} \right) + (1-\mu(a)) \log \left(\frac{1-\mu(a)}{1-\nu(a)} \right) \right).$$

Data structures for testing higher-order methods for embeddings analysis





Clement et al. **On the Use of ArXiv as a Dataset**. *arXiv:1905.00075* (2019),
Singh (...) Tupikina, Santolini **Quantifying the rise and fall of scientific fields**", Plos One (2022)

Representative (lower-dimensional) ‘knowledge’ space

$$R_g^k = \sqrt{\frac{\sum_{i=1}^{k} M_i (R_i - R_{cm}^k)^2}{\sum_{i=1}^{k} M_i}}$$

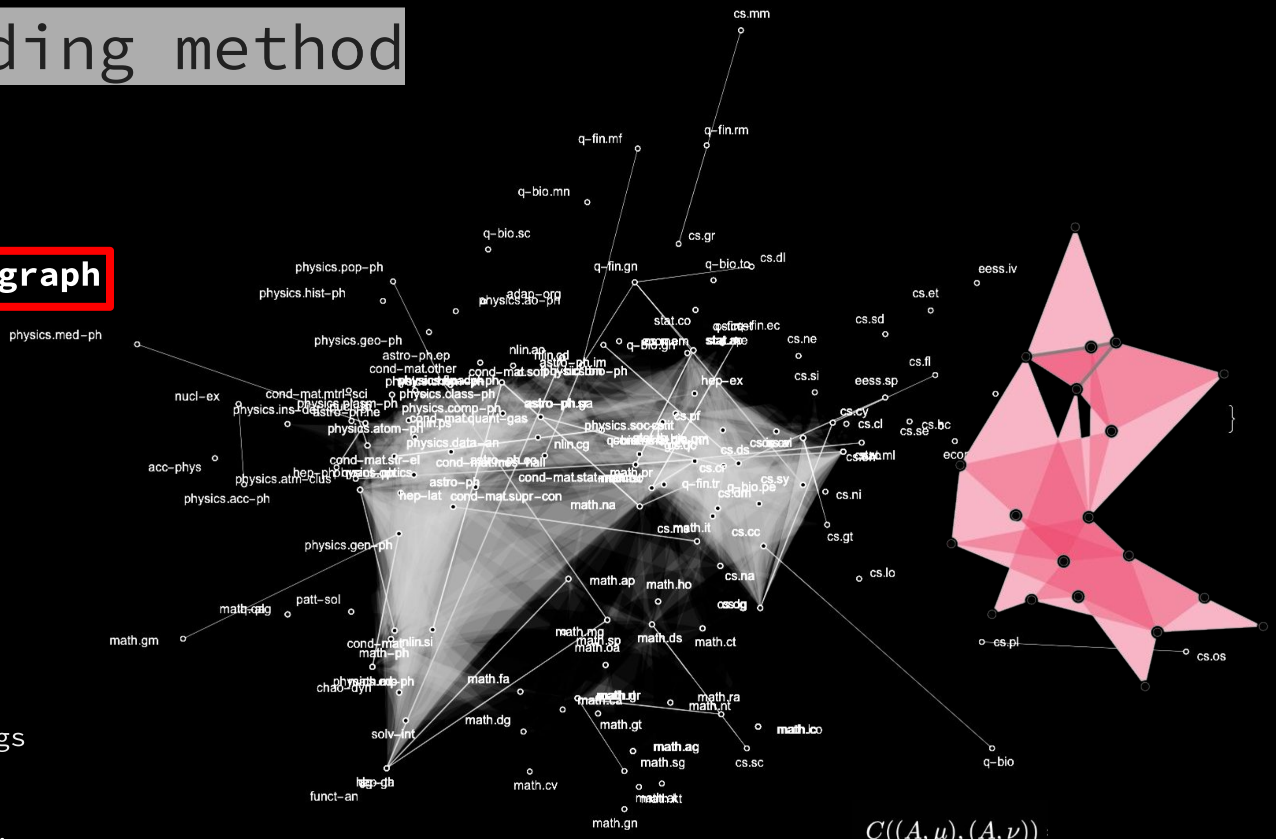
Dichotomy of human trajectories, radius of gyration (Pappalardo et al. 2015)

Proposed embedding method

Phase 1: Construct $\boxed{\textbf{hypergraph}}$ representing data

Phase 2. Embed the hypergraph/apply motifs, rewriting analysis

$$p_{ij} \rightarrow p_{ijk}$$

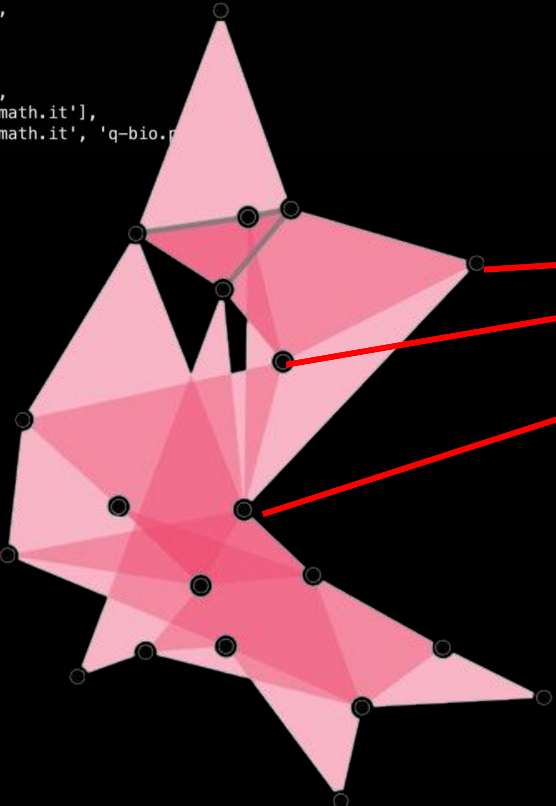


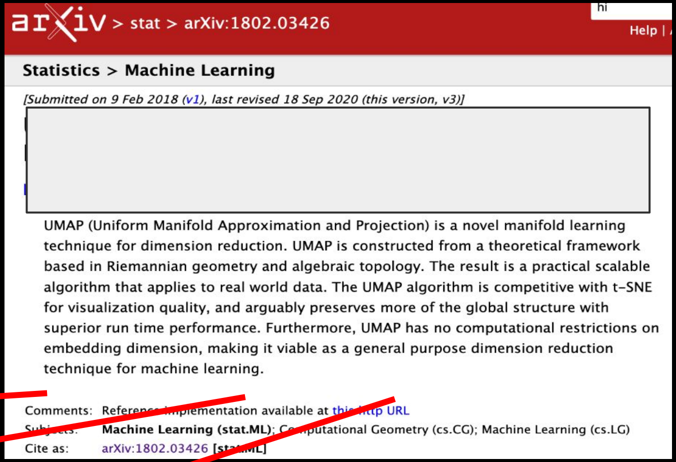
$$C((A,\mu),(A,\nu))$$

Learning hypergeometric embeddings
[Zhou et al. NIPS (2006)]()
Learning distances in weighted hypergraphs, cognitive distances, Vasilieva, LT, et al. (in prep.)

Encoding data with higher-order relations

```
['cs.ds'],
['quant-ph', 'cs.it', 'math.it'],
['cs.ne', 'cs.ai'],
['cs.ne', 'cs.ai'],
['cs.ce', 'cond-mat.stat-mech', 'cs.ms', 'cs.na', 'physics.data-an'],
['cs.it', 'math.it'],
['cs.cc'],
['cs.cc'],
['cs.cc'],
['cs.it', 'math.it'],
['cs.it', 'cs.cc', 'math.it'],
['cs.it', 'cs.ai', 'math.it', 'q-bio.
```





arXiv > stat > arXiv:1802.03426

Help |

Statistics > Machine Learning

[Submitted on 9 Feb 2018 (v1), last revised 18 Sep 2020 (this version, v3)]

UMAP (Uniform Manifold Approximation and Projection) is a novel manifold learning technique for dimension reduction. UMAP is constructed from a theoretical framework based in Riemannian geometry and algebraic topology. The result is a practical scalable algorithm that applies to real world data. The UMAP algorithm is competitive with t-SNE for visualization quality, and arguably preserves more of the global structure with superior run time performance. Furthermore, UMAP has no computational restrictions on embedding dimension, making it viable as a general purpose dimension reduction technique for machine learning.

Comments: Reference implementation available at this http URL

Subjects: **Machine Learning (stat.ML)**; Computational Geometry (cs.CG); Machine Learning (cs.LG)

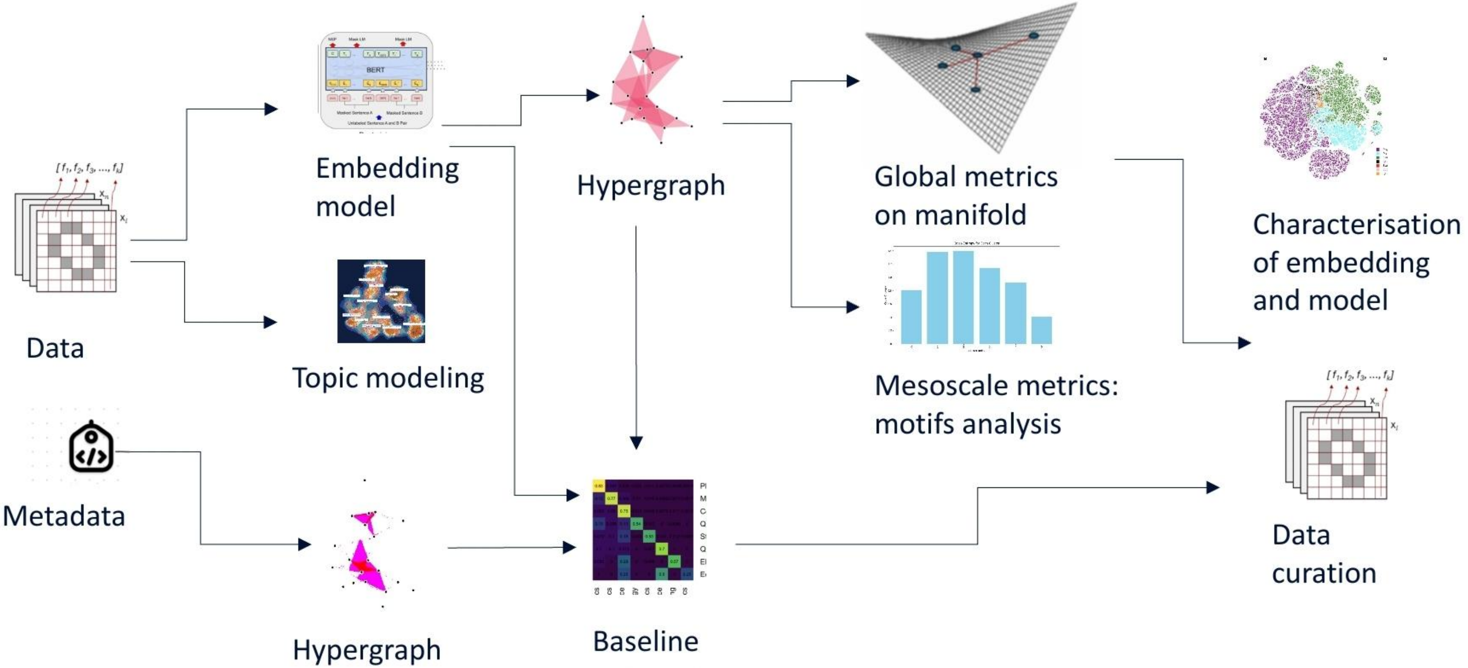
Cite as: arXiv:1802.03426 [stat.ML]

Hypergraph of co-tags is generalisation of co-tags network.

Data representations:
Nodes: fields
Hyperedges: papers

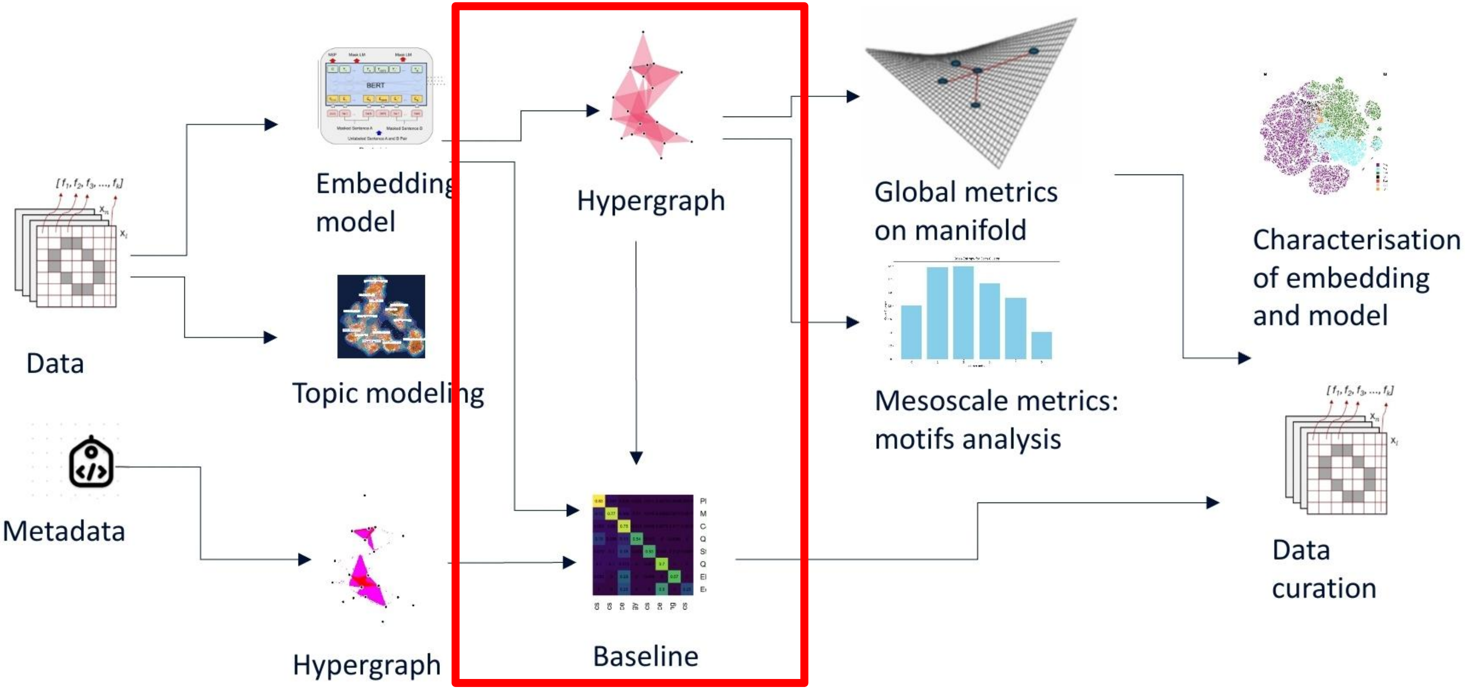
Nodes: articles
Hyperedges: relations between them

Encoding data with higher-order relations



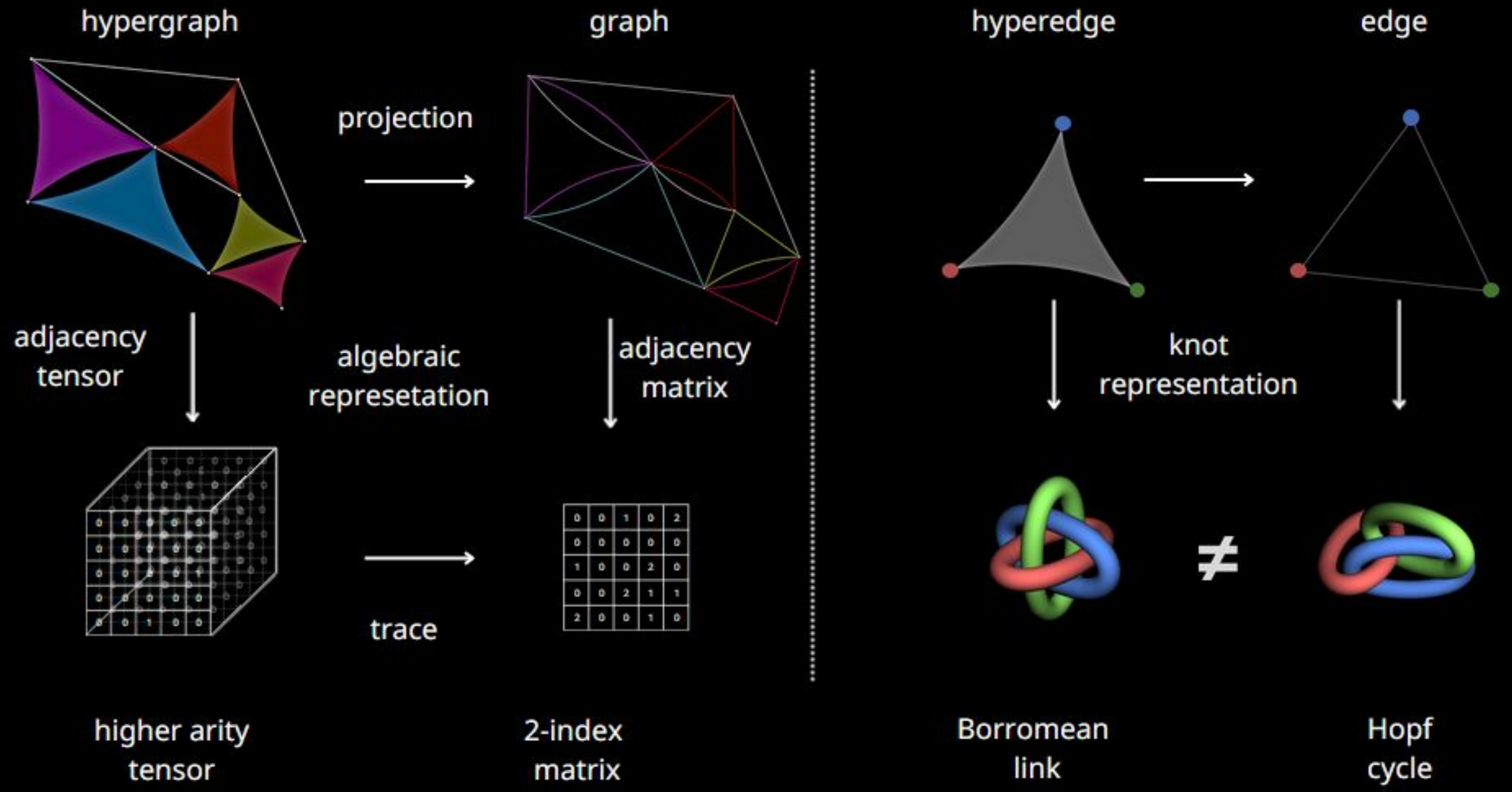
C. Zapata-Carratala, LT, Hypermatrix Algebra and Irreducible Arity in Higher-Order systems, NetSci (2024), LT, Kathuria, CNA Proceedings, (2024)

Hypergraphs structures to encode and compare embedding methods



C. Zapata-Carratala, LT, Hypermatrix Algebra and Irreducible Arity in Higher-Order systems, NetSci (2024),
LT, H. Kathuria, CNA Proceedings, (2024)

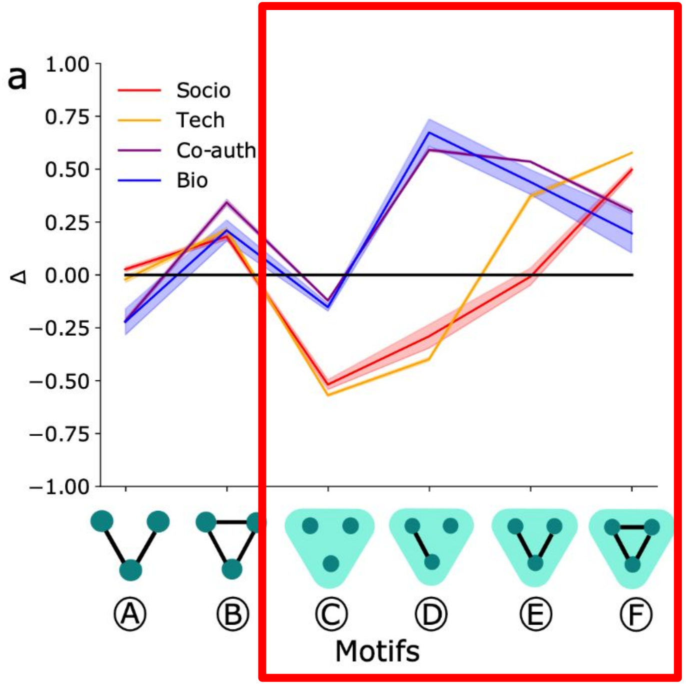
Higher-order structures require different theories to be developed



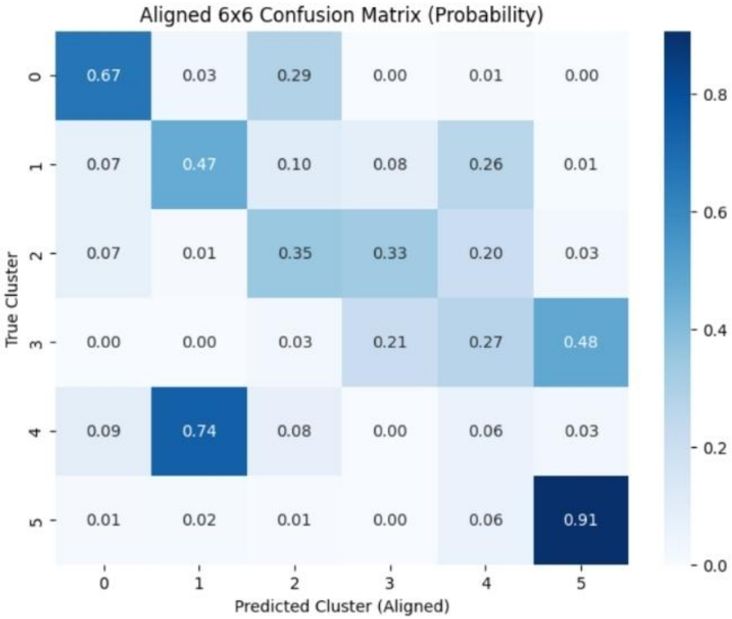
Applications of hypergraph motifs for dissection of embeddings

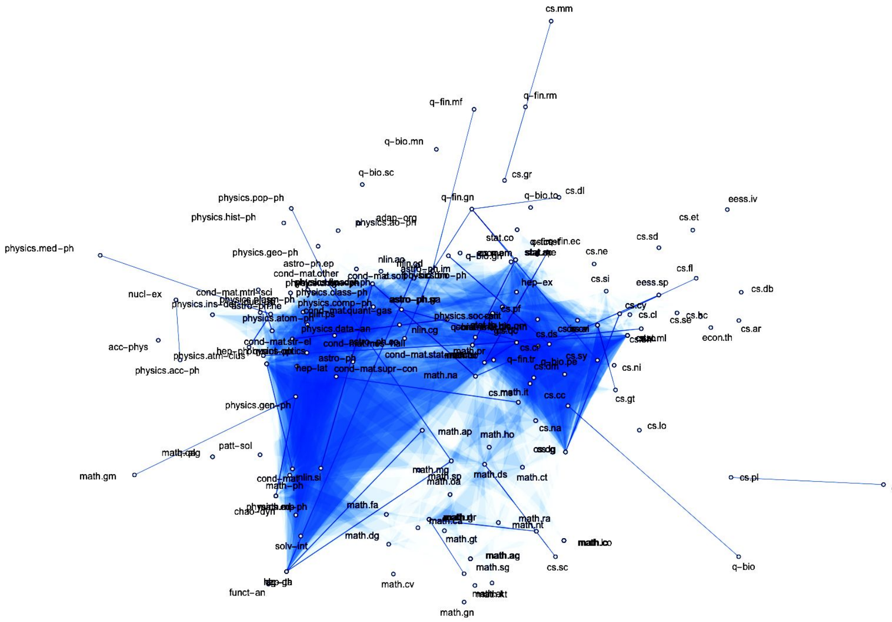
SOTA algorithms treat systems with higher order interactions as binary systems using linear algebraic mathematical structures [Lotito, Battiston et al. 2021]

Higher order interactions are neglected due to the lack of applicability of higher-order mathematical structures (non-associative algebras, n-ary algebras) [Kerner, Abramov et al. 2000, Zapata et al. 2022, Suchkevich, Wagner 1955]

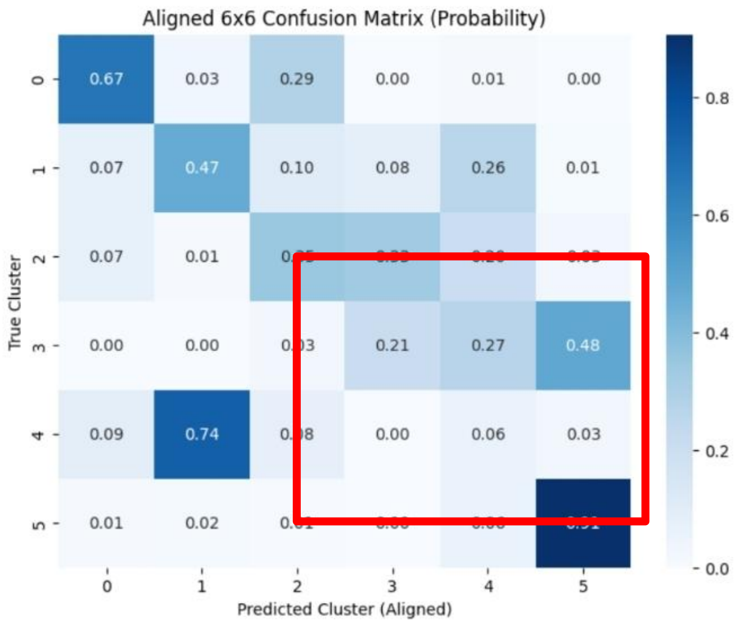


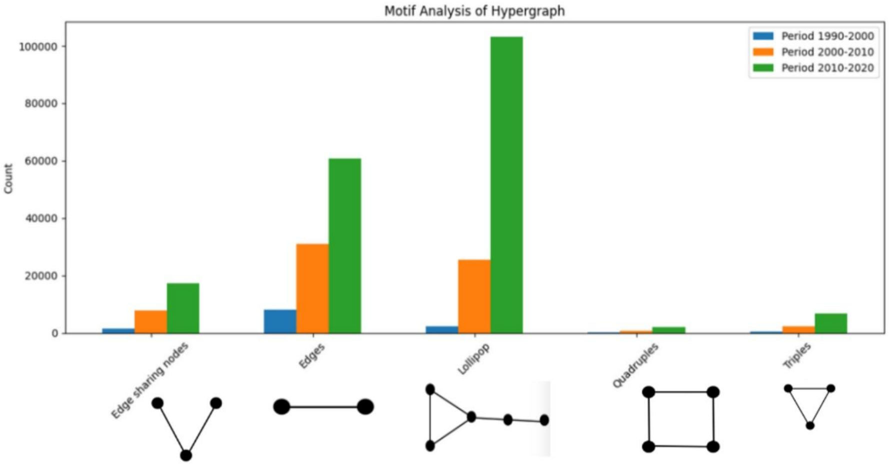
Applications of hypergraph motifs for dissection of embeddings





Applications of hypergraph motifs for dissection of embeddings





C. Zapata-Carratala, LT, Hypermatrix Algebra and Irreducible Arity in Higher-Order systems, NetSci presentation (2024), LT, Kathuria Dissecting embedding methods using higher order structures formalistsm, CNA Proceedings accept. (2024)

Outlook on the hypergraphs rewriting

Computational complexity of hypergraphs operations working with embeddings representation

https://hal.science/hal-04730183

https://github.com/Liyubov/hypergraph_distances

Dissecting embedding method: learning higher-order structures from data en fr

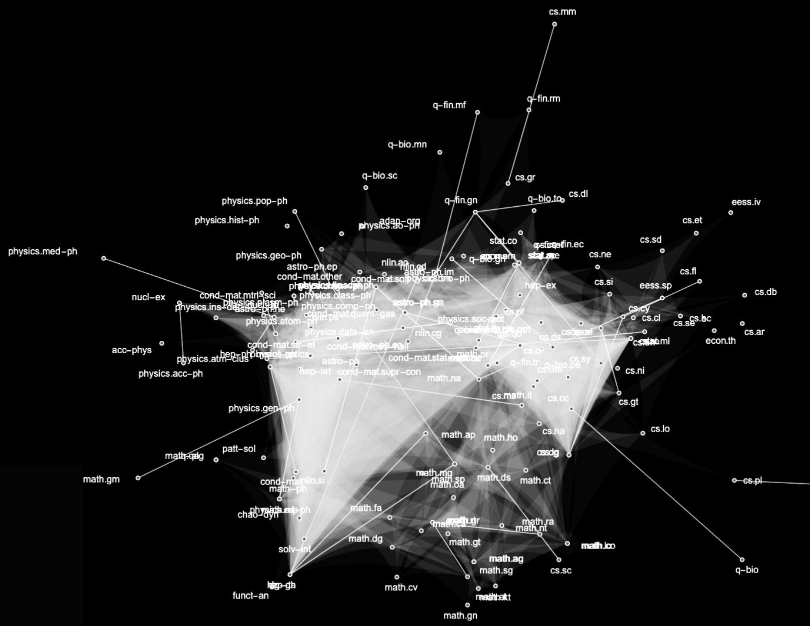
Liubov Tupikina (1, 2, 3) , Kathuria Hritika (3)

Show details

1 Nokia Bell Labs [Paris-Saclay]
2 UPD5 - Université Paris Descartes - Paris 5
3 LPI - Learning Planet Institute [Paris]

Abstract en

Keywords en



Conclusions & remarks

Hypergraphs and algebraic structures encoding hypergraphs are alternative methods to study embedding methods (github)

Comparison of latent space of models using combinatorial approaches, yet they are computationally expensive (MIPT project github)

Hypergraphs isomorphism, heaps and associativity theory (Wolfram Institute, ПОМИ Санкт-Петербург)

liubov.tupikina@cri-paris.org

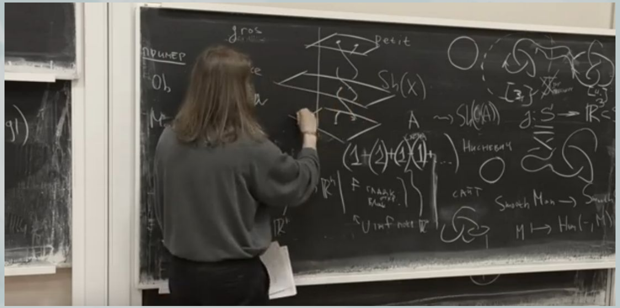
Ongoing seminars and workshops

Hypermatrix workshop TBD 2025

Course in MIPT on embeddings, hypergraphs (with Raigorodsky) 2025

Colloquium Embed-days
(Vaccarino, Petri, Rudin, ENS, Sony labs)
hybrid 2025

Network seminar BrAINs series (ITMO, LPI)
2024/2025



liubov.tupikina@cri-paris.org

Further directions: hypergraphs isomorphism problem

Main definitions

Def.1. Let us for simplicity consider an undirected finite graph G without multiple edges [4]. For defining the graph algebra $A(G)$ we define the canonical form of a graph as follows. We first need to associate every vertex of the graph and build the

class.

Def.3 A semiheap is a set S together with a ternary operation $[,,] : SxSxS \to S$ such that for all $a, b, c, d, e \in S$, $[[a, b, c], d, e] = [a, [d, c, b], e] = [a, b, [c, d, e]]$.

Def.4 Exceptional Bernstein algebra over a field K associated with a graph $G(V, E)$ where V is set of nodes: $U = \oplus kv, k \in K$, $Z = \oplus_{(a,b)\in E} kz_a^b$, $A(G) = U \oplus Z$.

Def.5 Exceptional Bernstein algebra for K associated with l-uniform hypergraph H then is defined using Grishkov extension. Given $v_i \in V(H)$ we can then first define auxiliary constructions $U = \oplus_{v\in V} kv, k \in F$,

$$Z_3 = \oplus kz_{a,b,c}, v_a, v_b, v_c \in V(H),$$

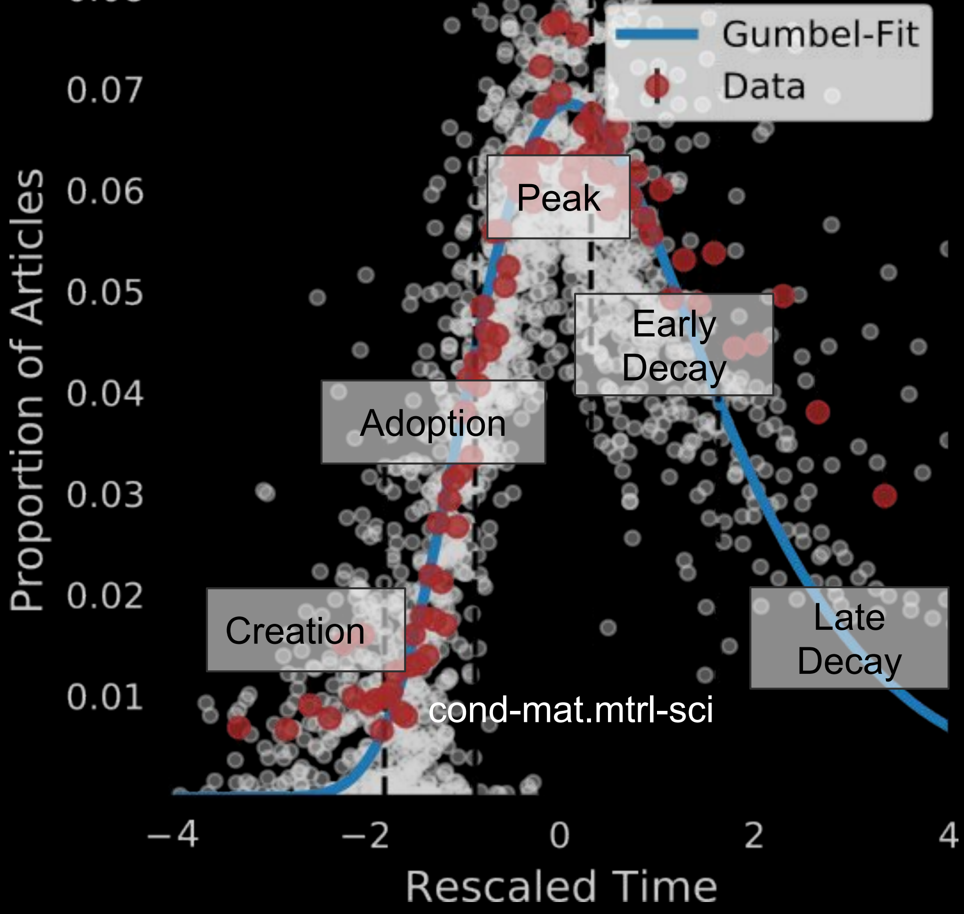
where H contains only ternary edges, F is some field, z_3 is the operator acting on the space defined for the set of vertices. Then $A(H)$ algebra is simply $A(H) = U \oplus V$.

Def.6 Projection of the hypergraph H to graph G is made using hyperedges mapping: l-order hyperedge is mapped to the $(l-1)$-order hyperedge.

Prop.1. Hypergraph algebra $A(H)$ is defined through the characteristic vectors of its vertices (on the concistency of Def.2 and Def.5).

Prop.2. Hypergraph algebra $A(H)$ fully defines the structure of the hypergraph H: given two hypergraph algebras equivalent, corresponding hypergraphs are isomorphic.

Rise and Fall of Research Fields

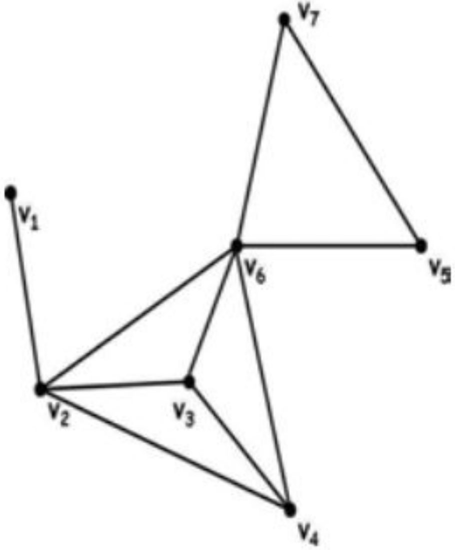


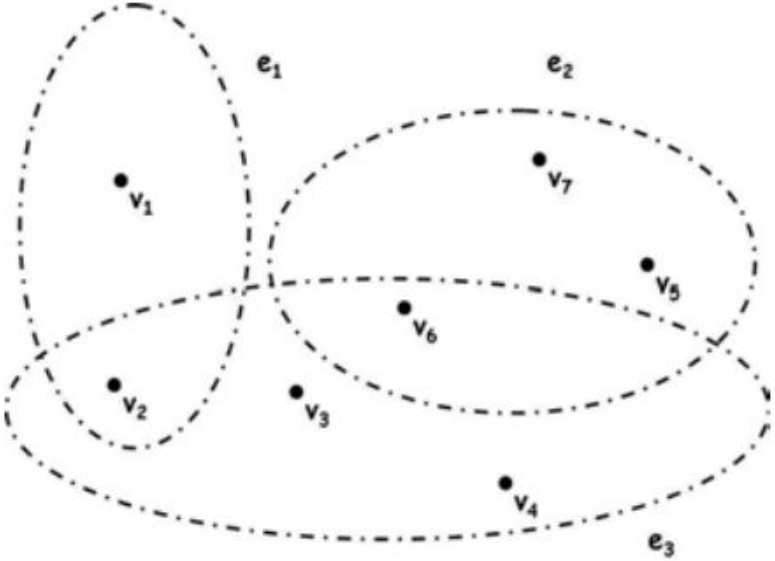
$$G = \frac{1}{\beta} e^{\frac{-(x-\alpha)}{\beta}} e^{-e^{\frac{-(x-\alpha)}{\beta}}}$$

$$t' = \frac{t-\alpha}{\beta}$$

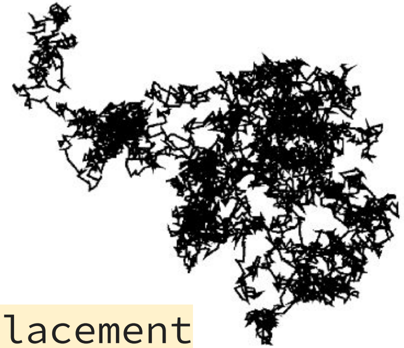
Field stages are defined at 2.5%, 16%,50% and 84% of the fit curve (blue). These number are borrowed from the **diffusion of innovation** literature

	e_1	e_2	e_3
v_1	1	0	0
v_2	1	0	1
v_3	0	0	1
v_4	0	0	1
v_5	0	1	0
v_6	0	1	1
v_7	0	1	0
v_2	1	0	1





Methods to study trajectories

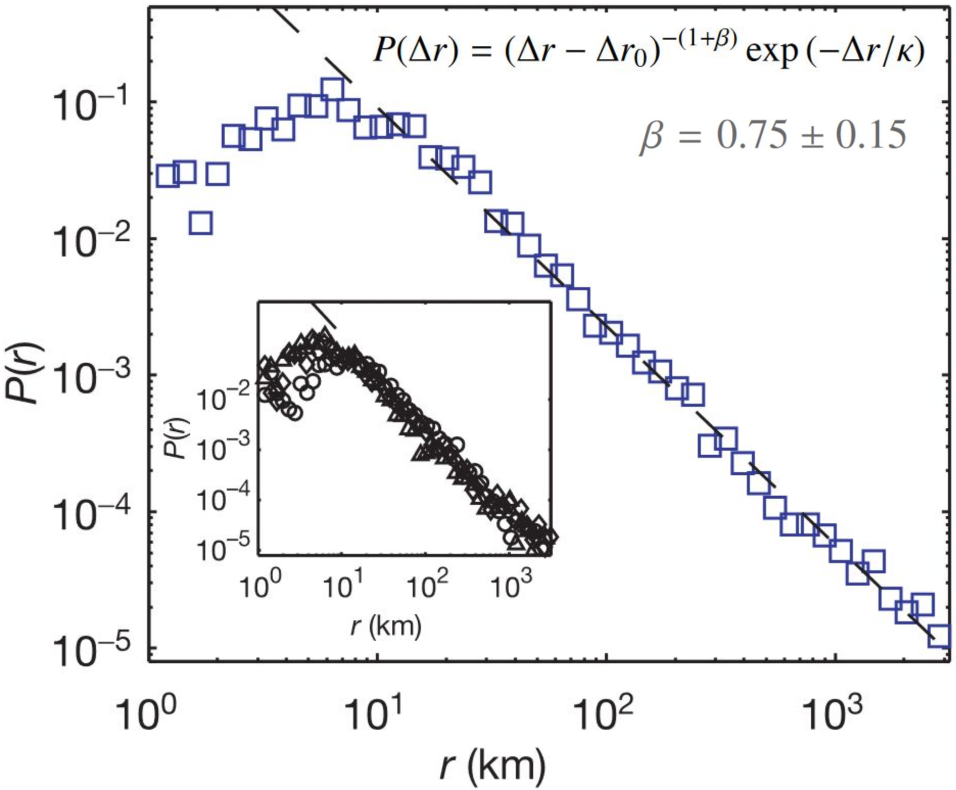


Mean squared displacement

$$\mathrm{MSD}(t) = \langle (\mathbf{r}(t) - \mathbf{r}_0)^2 \rangle \equiv \langle \Delta \mathbf{r}(t)^2 \rangle$$

$$\langle \Delta \mathbf{r}(t)^2 \rangle \sim t^{\nu}$$

Jump distance distribution



Klafter, Sokolov 'First steps in random walks' (2010), Barbosa et al. 'Human mobility: Models and applications' (2017), github.com/liyubov/Trajectory_analysis