Kronecker Generative Models for Power-Law Patterns in Real-World Hypergraphs

Minyoung Choe[1], Jihoon Ko[1], Taehyung Kwon[1], Kijung Shin[1], and Christos Faloutsos[2]

[1]{minyoung.choe, jihoonko, Taehyung.kwon, kijungs}@kaist.ac.kr, [2]christos@andrew.cmu.edu

Code and Data: https://github.com/young917/HyRec

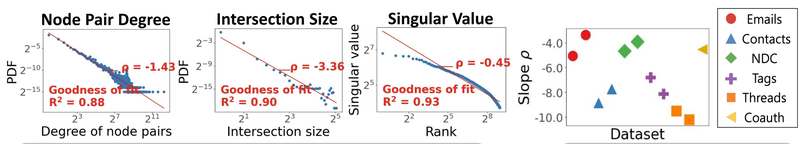
Motivation & Summary

- **RQ1. What Patterns Exist in Real-world Hypergraphs?**
  - **Why It Matters:** Recurring structural patterns show how group interactions typically form and evolve, and they define realism in hypergraphs.
  - **Our Answer:** We discover eight power-law properties across eleven real-world hypergraphs.
- **RQ2. What Mechanisms Underlie These Patterns?**
  - **Why It Matters:** Understanding what derives these patterns is a key to building realistic and generalizable generators.
  - **Our Answer:** We propose HyRec, a hypergraph generator based on self-similar Kronecker structure.
- **RQ3. How Can We Efficiently Fit HyRec to Real Data?**
  - **Why It Matters:** A well-fitted generator enables extrapolation, anonymization, data augmentation, and summarization.
  - **Our Answer:** We develop SingFit, an efficient algorithm that scales to large real-world hypergraphs.
- **Extensive Experiments**
  - **Accuracy:** HyRec effectively fits and extrapolates real-world hypergraphs with minimal input parameters.
  - **Efficiency:** HyRec achieves near-linear-time fitting and linear-time generation with the support of SingFit.

Discoveries: Patterns in real-world hypergraphs

- **Studied 11 Real-World Hypergraphs from Six Domains:**
  - Emails, Contacts, Drugs (NDC), Tags, Threads, and Co-authorship
- **Identified Three Key Discoveries (D1-D3):**

Discovery (D1): Node pair degrees, intersection sizes, and singular values exhibit power-law distributions



Strong linear regression fitting on log-log scale (R^2 scores near 1) confirms power-law behavior

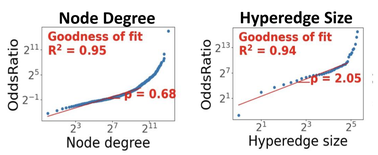
Similar slopes within the same domains

Discovery (D2) : Node degrees and hyperedge sizes follow log-logistic distributions rather than perfect power laws

x **follows log-logistic**

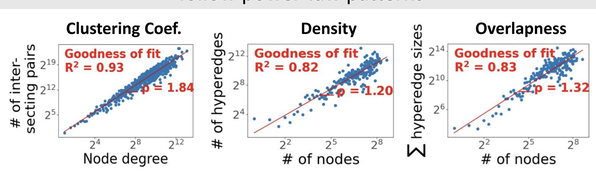
$\Leftrightarrow$

OddsRatio$(x) := \frac{CDF(x)}{1-CDF(x)}$ **follows power-law**



Linear regression on a log-log scale of odds ratios yields high R^2 scores, confirming strong log-logistic characteristics

Discovery (D3) : Clustering coefficients, density, and overlapness in egonets follow power-law patterns

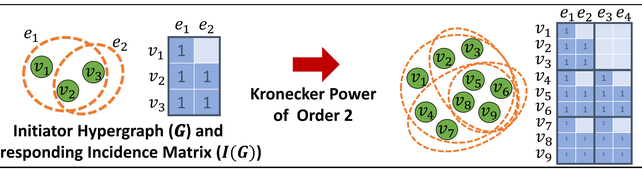


High R^2 values in log-log regression confirm power-law behavior

HyRec: A Tractable and Realistic Hypergraph Generator

- Inspired by observed **power-law patterns** in real-world hypergraphs and the **Kronecker graph model,** we present **HyRec,** which generates hypergraphs using the Kronecker product of incidence matrices.

Given an **initiator hypergraph** $\boldsymbol{G}$ and **order** $\boldsymbol{K}$, **HyRec($\boldsymbol{G}$, $\boldsymbol{K}$)** generates a hypergraph with an incidence matrix $\boldsymbol{I}(\boldsymbol{G})^{[K]}$, representing the K-th Kronecker power of $\boldsymbol{I}(\boldsymbol{G})$



- **Theoretical Characteristics of HyRec**

[**Structural Patterns**] **Theorem 1 and 2**. HyRec($\boldsymbol{G}$, $\boldsymbol{K}$) generates hypergraphs where key statistics follow **multinomial distributions***: (1) degrees, (2) hyperedge sizes, (3) pair degrees, (4) intersection sizes, and (5) singular values

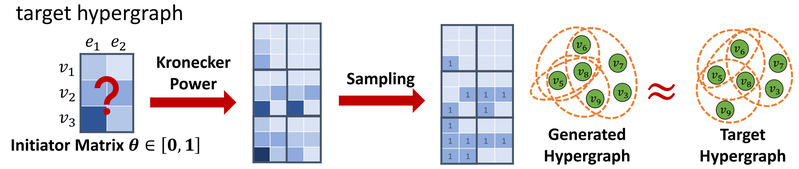
** **Multinomials** closely resemble power-law and log-logistic distributions found in real-world hypergraphs*

[**Evolutionary Patterns**] **Theorem 3–5**. HyRec($\boldsymbol{G}$, $\boldsymbol{K}$) models hypergraph growth by simulating changes in (1) density and (2) (effective) diameter as the Kronecker power exponent K increases, reflecting the evolution of hypergraphs over time.

- This **tractability** enables easier analysis and better insight into HyRec's behavior.

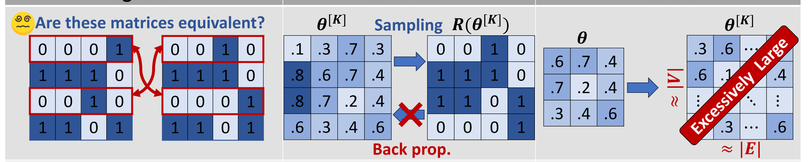
SingFit: An Efficient Algorithm for Fitting HyRec to Real-World Hypergraphs

- **Goal:** Find the initiator matrix ($\boldsymbol{\theta} \in [0,1]^{N \times M}$) that best captures key properties of the target hypergraph

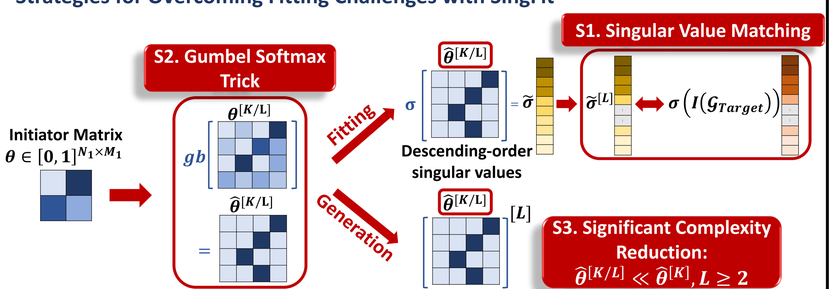


- **Challenges in Fitting HyRec**

C1. Computational Cost of Alignment	C2. Non-differentiability of Generation	C3. Density of Probability Matrix
Are these matrices equivalent?	Sampling $R(\theta^{[K]})$	$\theta^{[K]}$
Permutation-equivalent incidence matrices require **exhaustive** search for alignment.	The sampling process is **non-differentiable**, impeding gradient-based optimization.	**Full node-hyperedge** processing causes **heavy computation** and **memory use.**

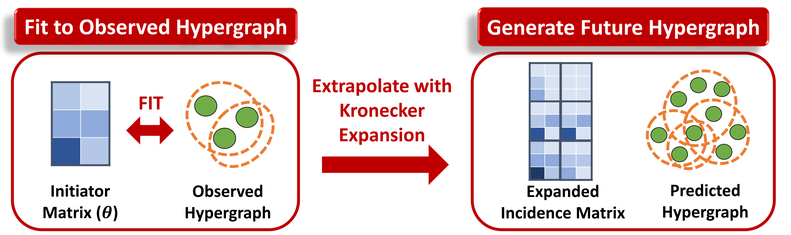


- **Strategies for Overcoming Fitting Challenges with SingFit**



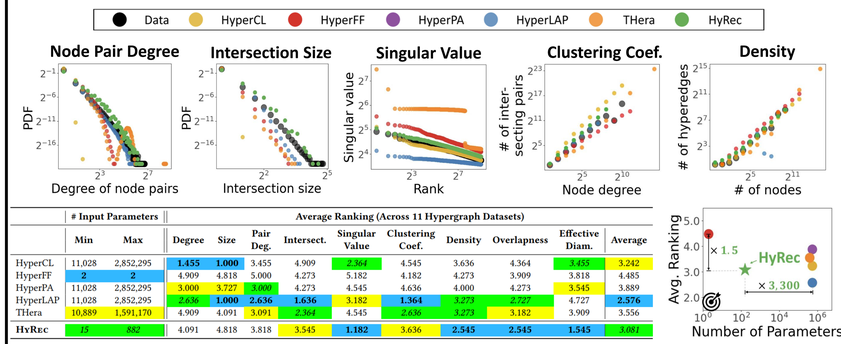
- **S1. Singular Value Matching:** Uses permutation-invariant **singular values**
- **S2. Differentiable Sampling:** Applies **gumbel softmax trick**
- **S3. Kronecker-based Acceleration:** Reduces cost via **Kronecker-based decomposition**

Application: Hypergraph Extrapolation using HyRec



Experimental Results

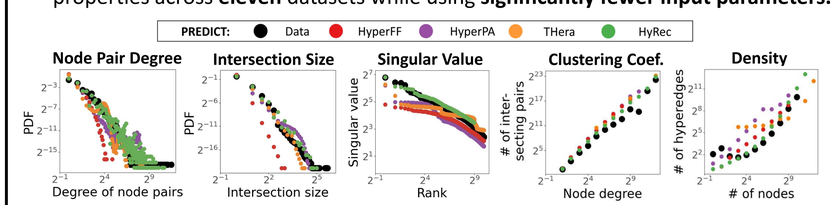
- **Fitting to Real-World Hypergraphs**
  - **HyRec effectively captures real-world hypergraph patterns with minimal input parameters**, outperforming in matching all **nine** properties across **eleven** datasets.



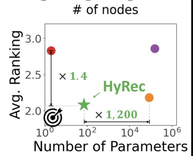
	# Input Parameters		Average Ranking (Across 11 Hypergraph Datasets)								
	Min	Max	Degree	Size	Pair Deg.	Intersect.	Singular Value	Clustering Coef.	Density	Overlapness	Effective Diam.	Average
HyperCL	11,028	2,852,295	1.455	1.000	3.455	4.909	2.364	4.545	3.636	4.364	3.435	3.242
HyperFF	2	2	4.909	4.818	5.000	4.273	5.182	4.182	4.273	3.909	3.818	4.485
HyperPA	11,028	2,852,295	3.000	3.727	1.000	4.273	4.545	4.636	4.000	4.273	3.545	3.889
HyperLAP	11,028	2,852,295	2.636	1.000	2.636	1.636	3.182	1.364	3.273	2.727	4.727	2.576
THera	10,889	1,591,170	4.909	4.091	3.091	2.364	4.545	2.636	3.273	3.182	3.909	3.556
HyRec	15	882	4.091	4.818	3.818	3.545	1.182	3.636	2.545	2.545	1.545	3.081

** The **best**, **second-best**, and **third-best** performance are highlighted in **blue**, **green**, and **yellow**, respectively*

- **Extrapolating Real-World Hypergraphs**
  - **Step 1.** Fit HyRec using only the hyperedges observed before 50% of the nodes appear
  - **Step 2.** Expand via Kronecker power and compare it to the full original hypergraph
  - **HyRec extrapolates well,** accurately predicting the evolution of **nine** hypergraph properties across **eleven** datasets while using **significantly fewer input parameters.**



	# Input Parameters		Average Ranking (Across 11 Hypergraph Datasets)								
	Min	Max	Degree	Size	Pair Deg.	Intersect.	Singular Value	Clustering Coef.	Density	Overlapness	Effective Diam.	Average
HyperFF	2	2	2.818	2.909	3.273	2.818	2.909	1.818	2.909	2.909	3.091	2.828
HyperPA	466	1,167,675	2.636	1.818	3.273	3.182	3.636	2.818	3.000	3.091	2.273	2.859
THera	349	537,114	3.000	2.273	1.636	1.909	1.636	2.909	2.000	2.182	2.091	2.182
HyRec	12	396	1.545	2.545	1.818	2.091	1.818	2.455	2.091	1.818	2.545	2.081



- **Efficiency in Fitting and Generating Large Hypergraphs**
  - **SingFit achieves near-linear scalability in both fitting and generation**, ensuring efficiency for large-scale hypergraphs.
  - **HyRec achieves the second-fastest generation time** while requiring significantly fewer input parameters than the fastest method.