Feiyan Ma

Weiyang College, Tsinghua University

February 23, 2026

1. Academic Performance
2. Research Experiences
3. Future Directions

Bios

- Major: Mathematics and Physics + Civil Engineering and Systems, Tsinghua University, 2021-2026 (expected).
- Overall GPA: 3.9/4.0, $\sim$ Top 15%.
- Research Assistant to Prof. Weichi Wu and Prof. Chenlei Leng, at Tsinghua (2024.05 - 2025.09). Topic: graphon estimation.
- Research Assistant to Prof. Ji Zhu and Prof. Gongjun Xu, at UMich (2025.03 - 2026.01). Topic: generative models for structured data.

Academic Performance

- **Undergraduate-Level Math Courses**: Probability Theory (1) (A+), Measures and Integrals (A), Abstract Algebra (A), Topology (A), Differential Geometry (A), Numerical Analysis (A), Advanced Topics in Linear Algebra (A-), Basic Functional Analysis (B+).
- **Graduate-Level Courses**: Advanced Mathematical Statistics I (A), Advanced Mathematical Statistics II (A-), Computational Probability (A), Statistical Analysis of Network Data (A), Probability (2) (B+).
- **Statistics Relevant Courses**: Reliability Data and Survival Analysis (A), Linear Regression Analysis (A-), Statistical Inference (A-), Financial Statistics (A+), Operation Research (A), Intro to Biostatistics (A), Introduction to Optimization Theory (A-), Topics in Logics (A).

1 Academic Performance

2 Research Experiences

Low-Rank Graphon Learning for Networks

Generative Model for Hypergraph Data with Hyperlink-wise Attributes

3 Future Directions

1. Academic Performance

2. Research Experiences
   - Low-Rank Graphon Learning for Networks
   - Generative Model for Hypergraph Data with Hyperlink-wise Attributes

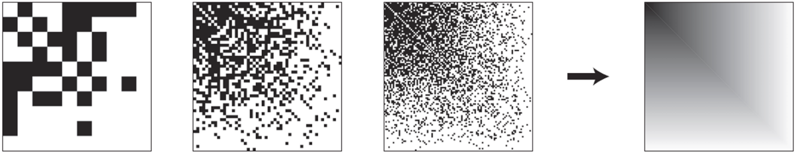
3. Future Directions

Low-Rank Graphon Learning for Networks

Low-Rank Graphon Learning for Networks: Supervised by Prof. Weichi Wu and Prof. Chenlei Leng at Tsinghua.
Co-first-authored paper accepted by **NeurIPS 2025**.

- We propose a novel approach that leverages a low-rank additive representation, yielding both a low-rank connection probability matrix and a low-rank graphon. Our method resolves identification issues and enables an efficient sequential algorithm based on subgraph counts and interpolation.
- We establish consistency and demonstrate strong empirical performance in terms of computational efficiency and estimation accuracy through simulations and data analysis.

Graphon (Graph Limits)



- We “lift” graphs of different sizes to analyze them in the same space by defining **graphon** (also called graph limit):
- A graphon is a symmetric measurable function $f : [0,1]^2 \to [0,1]$.
- Consider a random graph $\mathcal{G} = ([n], E)$ within the graphon model framework. For $i = 1, \ldots, n$, each node i is associated with an i.i.d. random variable $U_i \sim \text{Uniform}(0,1)$.
- The edges E_{ij} are independently drawn as $E_{ij} \sim \text{Bernoulli}(f(U_i, U_j))$ for $i < j$, and $E_{ii} = 0$.

Related Works

Our Goal: Estimate the **connection probability matrix** $P = \{P_{ij}\}_{1\leq i,j\leq n} \triangleq \{\mathbb{E}(E_{ij})\}_{1\leq i,j\leq n}$, and the **graphon function** $f(\cdot,\cdot)$ simultaneously, based on a single observed graph $G([n], E)$.

Graphon-based approaches:

- The approaches of Olhede & Wolfe (2014) relies on permutation maximization via a greedy algorithm, which is **computationally intensive**.
- Chan & Airoldi (2015) requires **strictly** monotonic marginals (doesn't allow SBMs), and the resulting P is generally not low-rank.

P-**based approaches** do not directly recover the graphon f, and all focus on mean squared error bounds:

- Chatterjee (2015), Zhang et al. (2017), Gao et al. (2016), $\cdots$

$\Rightarrow$ **Inconsistency** in low-rank P and underlying f usually exists.

Key Idea

- **Key idea**: Utilize a low-rank additive separable representation of f:

$$f(U_i, U_j) = \sum_{k=1}^{r} \lambda_k G_k(U_i) G_k(U_j),$$

where $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_r| > 0$, G_k is a measurable function with $\int_0^1 G_k^2(u)\, du = 1$ for $k = 1, \ldots, r$, and $\int_0^1 G_k(u) G_l(u)\, du = 0$ for $k \neq l$ (orthonormal conditions).

- This is a **truncated eigen-decomposition**, as suggested by the Hilbert-Schmidt theorem. It includes both the SBM and RDPG as special cases.
- The estimation of graphon f reduces to estimating $\{\lambda_k\}$'s and $\{G_k\}$'s.

- We develop an **efficient** sequential fitting algorithm using **subgraph counts**.

Algorithm 2 Estimation for $\{p_{ij}\}_{i,j=1}^n$ in Rank-r Model.

Require: The graph $\mathcal{G} = (V, E)$.

1: For $i = 1, \ldots, n$, compute $L_i^{(a)}, 1 \le a \le r$ and $C_i^{(a)}, 3 \le a \le r+2$ defined in (6) and (4).

2: Solve the system of equations

$$\begin{cases} y_k \ge 0, \text{ for } 1 \le k \le r, |\hat{\lambda}_1| > \cdots > |\hat{\lambda}_r|, \\ \sum_{k=1}^r \hat{\lambda}_k^a = \frac{1}{\prod_{j=0}^{a-1}(n-j)} \sum_{i=1}^n C_i^{(a)} \text{ for } 3 \le a \le r+2, \\ \sum_{k=1}^r \hat{\lambda}_k^a y_k^2 = \frac{1}{\prod_{j=0}^{a}(n-j)} \sum_{i=1}^n L_i^{(a)} \text{ for } 1 \le a \le r. \end{cases} \tag{8}$$

to obtain $(\hat{\lambda}_1, \cdots, \hat{\lambda}_r, y_1, \cdots, y_r)$.

3: For $i = 1, 2, \ldots, n$, compute the estimators $\hat{G}_1(U_i), \cdots, \hat{G}_r(U_i)$ from

$$\frac{1}{\prod_{j=1}^a (n-j)} L_i^{(a)} = \sum_{k=1}^r \hat{\lambda}_k^a y_k G_k(U_i) \text{ for } 1 \le a \le r. \tag{9}$$

4: Compute the standardized estimators $\tilde{G}_1(U_i), \cdots, \tilde{G}_r(U_i)$ from

$$\tilde{G}_k(U_i) = \hat{G}_k(U_i) \Big/ \sqrt{\sum_{i=1}^n \hat{G}_k^2(U_i)/n}. \tag{10}$$

5: For each pair (i, j), where $i \ne j$, estimate p_{ij} as $\hat{p}_{ij} = \left[1 \wedge \left(0 \vee (\sum_{k=1}^r \hat{\lambda}_k \tilde{G}_k(U_i) \tilde{G}_k(U_j))\right)\right]$. Set $\hat{p}_{ii} = 0$ for $i = 1, \ldots, n$.

6: Output $\{\hat{p}_{ij}\}_{i,j=1}^n$.

- With estimated discrete values of G_k's, under the assumption that G_1 is **monotonic**, we can **sort the nodes** to identify their latent variables $\{U_i\}$'s. Then recover G_k's and f by interpolation.

Main Theoretical Results

- We also establish the **error rate** of the proposed method regarding estimated connection probability matrix $\hat{P}$ and estimated graphon function $\hat{f}(\cdot,\cdot)$ under regular assumptions.

Theorem 3.6. *For $r \geq 2$, under Assumption 3.5, when n is sufficiently large, there exists an open set $U \subset \mathbb{R}^{2r}$ containing the point $(\lambda_1, \cdots, \lambda_r, \int_0^1 G_1(u)\,du, \cdots, \int_0^1 G_r(u)\,du)$ such that, with probability 1, the system of equations in (8) has a unique solution within this region. Moreover, for $\hat{\lambda}_k, 1 \leq k \leq r, \hat{p}_{ij}$, we have $\max_{1\leq k\leq r} |\hat{\lambda}_k - \lambda_k| = O_p(n^{-1/2})$, and $\sup_{i,j} |\hat{p}_{ij} - p_{ij}| = O_p(\sqrt{\log(n)/n})$.*

Theorem 3.9. *For $r \geq 2$, under Assumptions 3.5 and 3.7, the estimated graphon given by (11) satisfies*

$$\sup_{u,v\in[0,1]} |\hat{f}(u,v) - f(u,v)| \xrightarrow{a.s., L_2} 0, \text{and} = O_p(\sqrt{\log(n)/n}).$$

- Our result is based on the sup-norm, providing a stronger **uniform convergence** guarantee compared to point-wise or average error metrics. The achieved rate $\sqrt{\log(n)/n}$ matches that of Chan & Airoldi (2014).

Main Empirical Results

- We demonstrate its **computational efficiency** and **estimation accuracy** through extensive simulation studies.

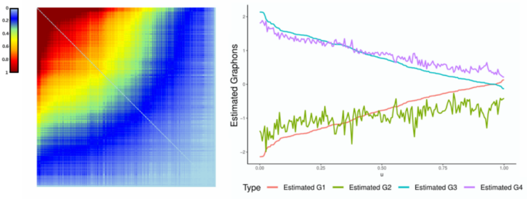


Figure 1: Learned $\hat{P}$ (Left) and learned graphon $\hat{f}$ (Right) for the primary school student interaction dataset [2].

- In **sparse** graphon cases, our method consistently outperforming all other approaches. This is expected, as it directly incorporates the sparsity parameter ρ_n during the equation solving procedure.

[2] http://www.sociopatterns.org/

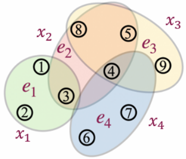
1. Academic Performance
2. Research Experiences
    - Low-Rank Graphon Learning for Networks
    - Generative Model for Hypergraph Data with Hyperlink-wise Attributes
3. Future Directions

Generative model for hypergraph with hyperlink-wise attributes

ReLaSH: Reconstructing Joint Latent Spaces for Efficient Generation of Synthetic Hypergraphs with Hyperlink Attributes: Supervised by Prof. Ji Zhu and Prof. Gongjun Xu at UMich.
First-authored paper accepted by **ICLR 2026**.

- We introduce ReLaSH (REconstructing joint LAtent Spaces for Hypergraphs with attributes), a general generative framework for producing realistic synthetic hypergraph data with hyperlink attributes via training a likelihood-based joint embedding model and reconstructing the joint latent space.
- ReLaSH explicitly accounts for the unique structure of hypergraphs and jointly models hyperlinks and their attributes. It also provides flexibility, efficiency, and interpretability relative to deep black-box architectures.
- We theoretically demonstrate consistency and generalizability of ReLaSH. Empirical results on a range of real-world datasets from diverse domains demonstrate its strong performance.

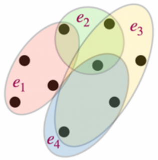
Hypergraph Network Data



e_1 $\{1,2,3\}$ $-$ x_1
e_2 $\{3,4,5,8\}$ $-$ x_2
e_3 $\{4,5,8,9\}$ $-$ x_3
e_4 $\{4,6,7\}$ $-$ x_4

- Hypergraph $\mathcal{H}(\mathcal{V}_n, \mathcal{E}_m, \mathcal{X}_m)$
- Nodes $\mathscr{V}_n = [n] = \{1, \ldots, n\}$.
- Hyperlinks $\mathscr{E}_m = \{e_1, \ldots, e_m\}$.
- Hyperlink-wise Attributes $\mathcal{X}_m = \{x_1, \ldots, x_m\}$

Hypergraph characteristics: sparsity of hyperlinks, degree heterogeneity of nodes, mixed data types of hyperlink attributes (and hyperlinks, since hyperlinks can be expressed as a binary vector).



hypergraph network

Examples:

Nodes	Relations (hyperlinks)	Additional information (hyperlink attributes)
medical symptoms	co-occurrence in patients' profiles	patients' demographics (e.g., BMI, age, …)
ingredients	co-occurrence in recipes	nutrition contents (e.g., calories, fat, …)
scholars	co-citation in journal papers	paper metadata (e.g., abstract keywords)
……		

Examples of generative modeling:

- Create synthetic medical records to share across medical centers while preserving patient privacy.
- Generate new recipes with predicted nutrition contents.

Goal: generate new hyperlinks and hyperlink-wise attributes on the same node set, and preserve structural properties of the original hypergraph.

Related Work

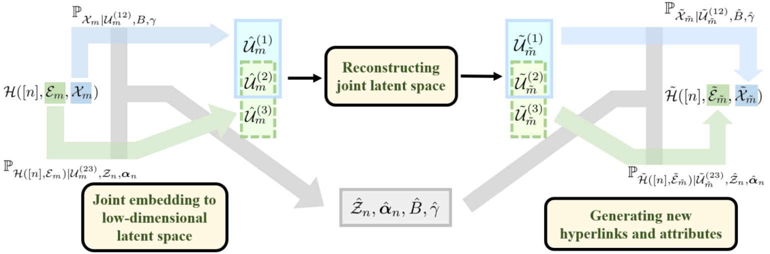
Existing research for tabular data generation: Generative models operating in discrete state spaces/mixed-type data have **slow** convergence in training and sampling [1, 6]. They do not account for the special **characteristics** of structured hypergraph data.

Existing research for graph generation:

- Existing graph generative models capture **pairwise relations** [9, 3].
- Representative learning on hypergraphs captures structural characteristics, but doesn't **extend to generative models**[7, 8].
- A thread of recent works on attributed hypergraphs address **different problems and objectives** compared to ours.[2, 4, 5]

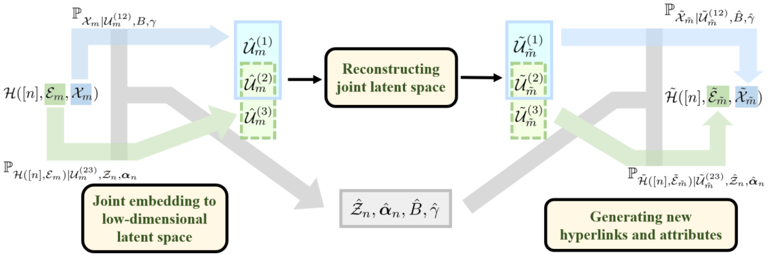
Key ideas: ReLaSH first maps the mixed-type data onto continuous spaces, then works on the continuous manifold.

Pipeline of ReLaSH



Step 1: Joint latent space embedding to low-dimensional latent space based on $\mathbb{P}_{\mathcal{H}([n],\{E\})|u^{(23)},\mathcal{Z}_n,\alpha_n}$ and $\mathbb{P}_{X|u^{(12)},B,\gamma}$.

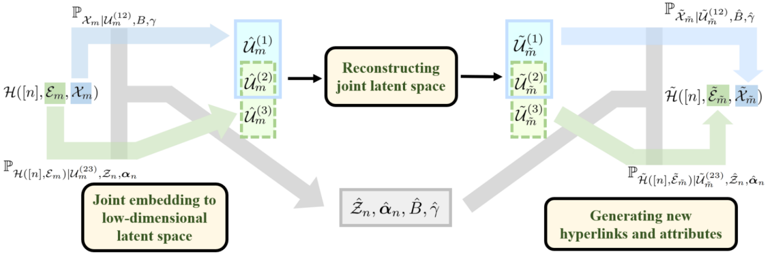
- Given a hypergraph dataset $\mathcal{H}([n],\mathcal{E}_m,\mathcal{X}_m)$, ReLaSH first embeds the hyperlinks and their attributes into a joint latent space by training a likelihood-based model.
- $\Rightarrow$ interpretable, preserving the characteristics of structured data, computationally efficient.



Step 2: Reconstructing joint latent space.

- Train a distribution-free generator (diffusion models[3]) **in the low-dim latent space** $\mathbb{R}^K$ to learn the distribution of $\hat{\mathcal{U}}_m$, which estimates $\{u_1, u_2, \cdots, u_m\}$, an empirical distribution of $\mathbb{P}_U$.
- Sample the new embeddings $\tilde{\mathcal{U}}_{\tilde{m}} = \{\tilde{u}_1, \cdots, \tilde{u}_{\tilde{m}}\}$ from the distribution-free generator.

[3] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. In International Conference on Learning Representations, 2021.



Step 3: Generating new hyperlinks and attributes.

- New hyperlinks $\tilde{e}_j$ and attributes $\tilde{x}_j$ are **decoded** via the trained likelihood-based model $\mathbb{P}_{\mathcal{H}([n],\{E\},\{X\})|\tilde{u},\hat{\mathcal{Z}}_n,\hat{\alpha}_n,\hat{B},\hat{\gamma}}$.

Key ideas: Generative models working on mixed-type data are slow in training and sampling. ReLaSH mitigates this by mapping the mixed-type data onto continuous spaces, and then works on the continuous manifold.

Main Theoretical Results

We theoretically demonstrate consistency and generalizability of ReLaSH by decomposing the overall error.

Theorem 2. *The KL-divergence between the true distribution $\mathbb{P}_{(E,X,U)}$ and the generated distribution $\mathbb{P}_{(\tilde{E},\tilde{X},\tilde{U})}$ admits the following decomposition:*

$$\mathrm{d}_{\mathrm{KL}}(\mathbb{P}_{(E,X,U)} \,\|\, \mathbb{P}_{(\tilde{E},\tilde{X},\tilde{U})}) = \Delta_{(\mathcal{Z}_n,B,\alpha,\gamma)\text{-}\textit{estimation}} + \Delta_{\mathbb{P}_U\text{-}\textit{estimation}} + \Delta_{\textit{latent-reconstruction}},$$

Depends on the likelihood model

Utilizes a discretization strategy to understand

Approximately diffusing independent samples in a low-dim continuous space

Under regular assumptions, the error rates can be computed.

Theorem 3. *Suppose that Assumptions 1 and 2 hold, and $\lambda \asymp \exp(\bar{\alpha}_{m,n})$, then as $(m,n) \to \infty$, the rate of estimation-related error satisfies*

$$\frac{1}{(n \vee p)} \Delta_{(\mathcal{Z}_n,B,\alpha,\gamma)\text{-}\textit{estimation}} = O_p\left(\frac{\log(m \vee n)}{\min\{m,n,p\}}\right).$$

Consequently, when $m \asymp n \asymp p$, we have $n^{-1}\Delta_{(\mathcal{Z}_n,B,\alpha,\gamma)\text{-}\textit{estimation}} = O_p(\log n/n)$, thus the error in final generative performance from estimating $\mathcal{Z}_n, B, \alpha, \gamma$ shrinks as fast as $\log n \cdot n^{-1}$.

Proposition 1 (Theorem 2 in Chen et al. (2022)). *Under Assumption 3, if $L \geq 1, h \leq 1$ and $T \geq 1$. We have $\Delta_{\textit{latent-reconstruction}} \lesssim (M_U + K)e^{-T} + T\varepsilon_0^2 + N^{-1}KT^2L^2$, where $K = k_1 + k_2 + k_3$. Then by choosing $T = \log((M_U + K)/\varepsilon_0^2)$ and $N = \Omega(KTL^2/\varepsilon_0^2)$, we have $\Delta_{\textit{latent-reconstruction}} = O(T\varepsilon_0^2)$.*

Main Empirical Results

Empirical results on a range of **real-world datasets from diverse domains** demonstrate the strong performance of ReLaSH, underscoring its broad utility and effectiveness in practical applications.

Personal Information		
Name: Jane Doe	Gender: ☐Male ☒Female	Religion: Catholic
Marital Status: ☐Single ☒Married ☐Divorced ☐Widowed ☐Separated ☐Life Partner		
Ethnicity: ☒ White ☐Black ☐Hispanic/Latino ☐ Asian ☐ American Indian/Alaska Native ☐ Other		
Lifetime: 86.19 yrs	Hospital Stay Time: 14d 19h	ICU Stay Time: 8d 6h

Representative Major Diseases	Other Diseases and Complications Record
Coronary Atherosclerosis Congestive Heart Failure Chronic Kidney Disease Intracerebral Hemorrhage Dementia	Hyperlipidemia; Hyperpotassemia; Pneumococcus infection; Atrial fibrillation; Primary cardiomyopathies; Long-term (current) use of anticoagulants; Chronic systolic heart failure; Abdominal aneurysm, ruptured; Embolism and thrombosis of iliac artery; Chronic obstructive asthma; Chronic airway obstruction; Noninfectious gastroenteritis and colitis; Hemorrhage of gastrointestinal tract; Acute kidney failure; Sinoatrial node dysfunction; Hematoma complicating a procedure; Personal history of malignant neoplasm of breast.
Total Diseases: 23	

Figure 2: An example of synthetic ICU medical record forms generated from ReLaSH, trained on a symptom co-occurrence hypergraph from Johnson et al. (2016), which includes 3,000 ICU patient profiles and 2,230 distinct disease and symptom codes. The disease combinations in this synthetic record reflect the characteristics of an aged, medically complex ICU patient, where the co-occurrence of symptoms often leads to the development of new syndromes.

We compare ReLaSH with 9 methods that can be used to produce synthetic hyperlinks with attributes: Gau-Diff, RealNVP, WGAN, VAE, ForestDiffusion, TabPFGen, CTAB-GAN, CTAB-GAN+, and CTGAN.

For the last five tabular data generation baselines, these methods do not scale to the patient-profile and co-citation generation tasks, so we just test them on a smaller recipe hypergraph dataset.

	$\Delta_{\mathcal{H}_v} \downarrow$	$\Delta_{\mathcal{X}_m} \downarrow$	$\Delta_{\mathcal{X}_v} \downarrow$	FED $\downarrow$	a-FED $\downarrow$
ReLaSH-$(5, 0, 2)$	1.978	2.236	0.894	0.293	0.356
ReLaSH$_c$-$(5, 0, 2)$	7.504	2.236	0.894	0.182	0.248
ReLaSH-$(5, 0, 6)$	2.129	2.174	**0.820**	**0.003**	**0.048**
ReLaSH$_c$-$(5, 0, 6)$	3.583	2.174	**0.820**	0.191	0.258
ReLaSH-$(5, 0, 16)$	2.355	1.533	1.112	0.766	0.847
ReLaSH$_c$-$(5, 0, 16)$	1.847	1.533	1.112	0.180	0.255
Gau-Diff	2.375	2.154	4.256	0.802	0.828
RealNVP	2.484	**1.146**	3.562	0.909	0.997
WGAN	2.208	21.428	1.351	0.907	0.928
VAE	21.587	9.883	5.180	11.553	10.285
CTGAN	2.519	28.799	4.983	0.847	0.865
ForestDiffusion	1.886	8.211	2.073	0.848	0.303
TabPFGen	**1.565**	1.915	1.205	0.297	0.884
CTAB-GAN	2.552	19.367	3.858	0.925	0.947
CTAB-GAN+	2.488	8.330	3.821	0.898	0.902

Table 3: Results for recipe generation. Scales of $\Delta_{\mathcal{H}_v}$, $\Delta_{\mathcal{X}_m}$, $\Delta_{\mathcal{X}_v}$, FED and a-FED are 10^{-3}, 10^{-2}, 10^{-2}, 10^{-1}, 10^{-1}, respectively.

Future Directions

- In addition to areas related to statistical network analysis, I am also open to explore broader topics in statistical machine learning and embrace new challenges.
- I have a strong foundation in mathematics and statistics, with extensive experience in both theoretical analysis and code implementation.
- My research background spans multiple areas, with tangible outcomes, enabling me to rapidly adapt to new research fields and make meaningful contributions during my Ph.D.

[1] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

[2] Anna Badalyan, Nicolò Ruggeri, and Caterina De Bacco. Structure and inference in hypergraphs with node attributes. *Nature Communications*, 15(1):7073, 2024.

[3] Xiaohui Chen, Jiaxing He, Xu Han, and Li-Ping Liu. Efficient and degree-guided graph generation via discrete diffusion modeling. *arXiv preprint arXiv:2305.04111*, 2023.

[4] Jaewan Chun, Seokbum Yoon, Minyoung Choe, Geon Lee, and Kijung Shin. Attributed hypergraph generation with realistic interplay between structure and attributes. *arXiv preprint arXiv:2509.21838*, 2025.

[5] Dorian Gailhard, Enzo Tartaglione, Lirida Naviner, and Jhony H Giraldo. Feature-aware hypergraph generation via next-scale prediction. *arXiv preprint arXiv:2506.01467*, 2025.

[6] Emiel Hoogeboom, Didrik Nielsen, Rianne van den Berg, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *arXiv preprint arXiv:2102.05379*, 2021.

[7] Jaehyeong Jo, Jinheon Baek, Seul Lee, Dongki Kim, Minki Kang, and Sung Ju Hwang. Edge representation learning with hypergraphs. *Advances in Neural Information Processing Systems*, 34:7534–7546, 2021.

[8] Shihao Wu, Junyi Yang, Gongjun Xu, and Ji Zhu. Denoising diffused embeddings: a generative approach for hypergraphs, 2025.

[9] Jiaxuan You, Rex Ying, Xiang Ren, William Hamilton, and Jure Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models. In *International Conference on Machine Learning*, pages 5708–5717. PMLR, 2018.

Thank you!