Joint Relational Database Generation via Graph-Conditional Diffusion Models

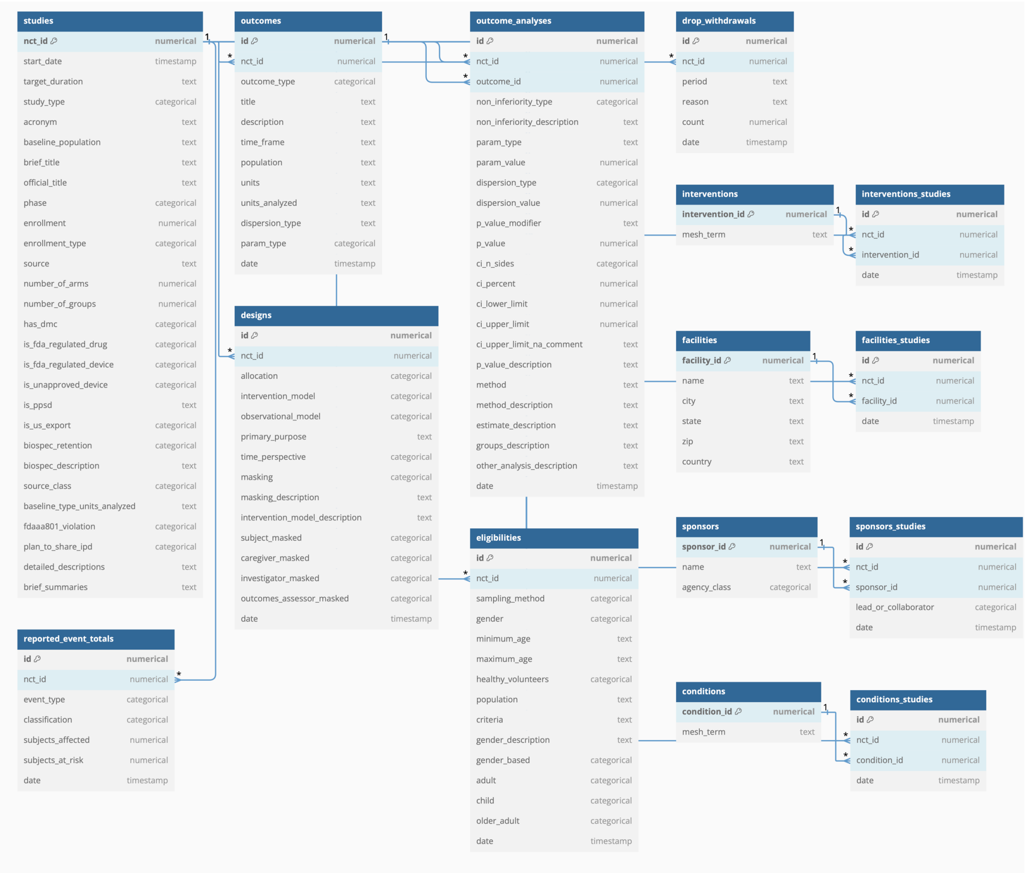
Mohamed Amine Ketata, David Lüdke, Leo Schwinn, Stephan Günnemann

School of Computation, Information and Technology & Munich Data Science Institute
Technical University of Munich, Germany

December 2025

Relational Databases (RDBs) ...

- ... organize data into ***multiple***, ***interlinked*** tables.
- ... are the most widely used data management system: they store over ***70%*** of the world's structured data [1].



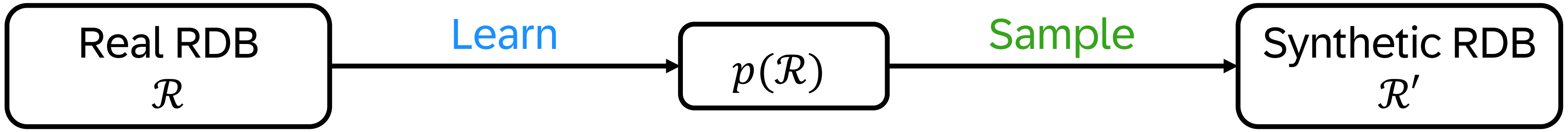
Source: https://relbench.stanford.edu/datasets/rel-trial/

[1] DB-Engines. DBMS popularity broken down by database model, 2023. Available: https: //db-engines.com/en/ranking_categories.

Synthetic RDB Generation

Applications:

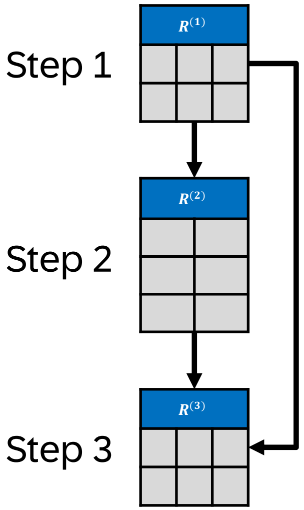
- Privacy-preserving data sharing.
- Data augmentation for machine learning models.
- Synthetic data for training foundation models.
- ...



How to model $p(\mathcal{R})$?

Autoregressive (prior works)

$$p(\mathcal{R}) = \prod_{i=1}^{m} p(R^{(i)} | \{R^{(j)} | j < i\})$$

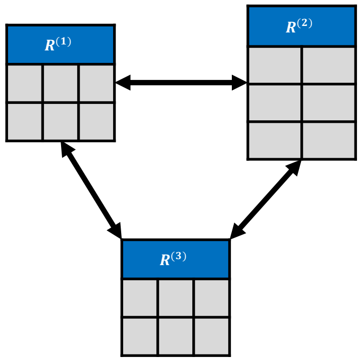


– Fixed table order

– Sequential generation

Joint (Ours)

$$p(\mathcal{R}) = p(R^{(1)}, \ldots, R^{(m)})$$



+ No table order

+ Parallel generation

RDBs as Heterogeneous Graphs

Customers

ID	Age	Income	City
1	27	33K	A
2	56	60K	C
3	32	40K	B

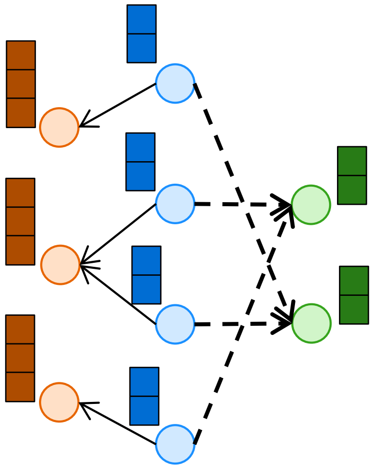
Products

ID	Brand	Category
1	X	Food
2	Y	Clothing

Transactions

ID	Customer ID	Product ID	Price	Time
1	1	2	30.0	09:35
2	2	1	15.7	12:45
3	2	2	35.4	15:34
4	3	1	17.5	19:03

RDB $\mathcal{R}$



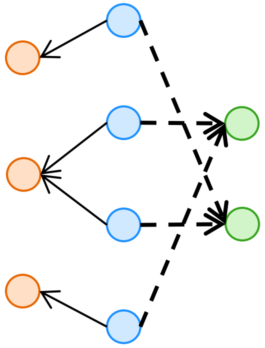
Graph $\mathcal{G}$ with nodes $\mathcal{V}$, edges $\mathcal{E}$, and features $\mathcal{X}$

$$p(\mathcal{R}) = p(\mathcal{G}) = p(\mathcal{V}, \mathcal{E}, \mathcal{X})$$

Method: RDB generation as two-step graph generation procedure

$$p(\mathcal{R}) = p(\mathcal{G}) = p(\mathcal{V}, \mathcal{E})p(\mathcal{X} | \mathcal{V}, \mathcal{E})$$

Step 1: $p(\mathcal{V}, \mathcal{E})$



Customers

ID	Age	Income	City
1	??	??	??
2	??	??	??
3	??	??	??

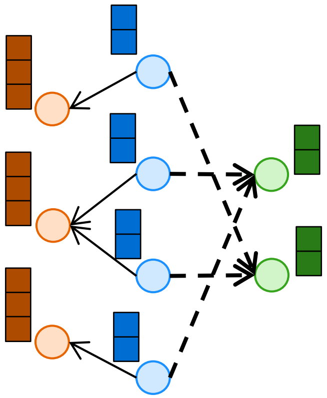
Products

ID	Brand	Category
1	??	??
2	??	??

Transactions

ID	Customer ID	Product ID	Price	Time
1	1	2	??	??
2	2	1	??	??
3	2	2	??	??
4	3	1	??	??

Step 2: $p(\mathcal{X} | \mathcal{V}, \mathcal{E})$



Customers

ID	Age	Income	City
1	27	33K	A
2	56	60K	C
3	32	40K	B

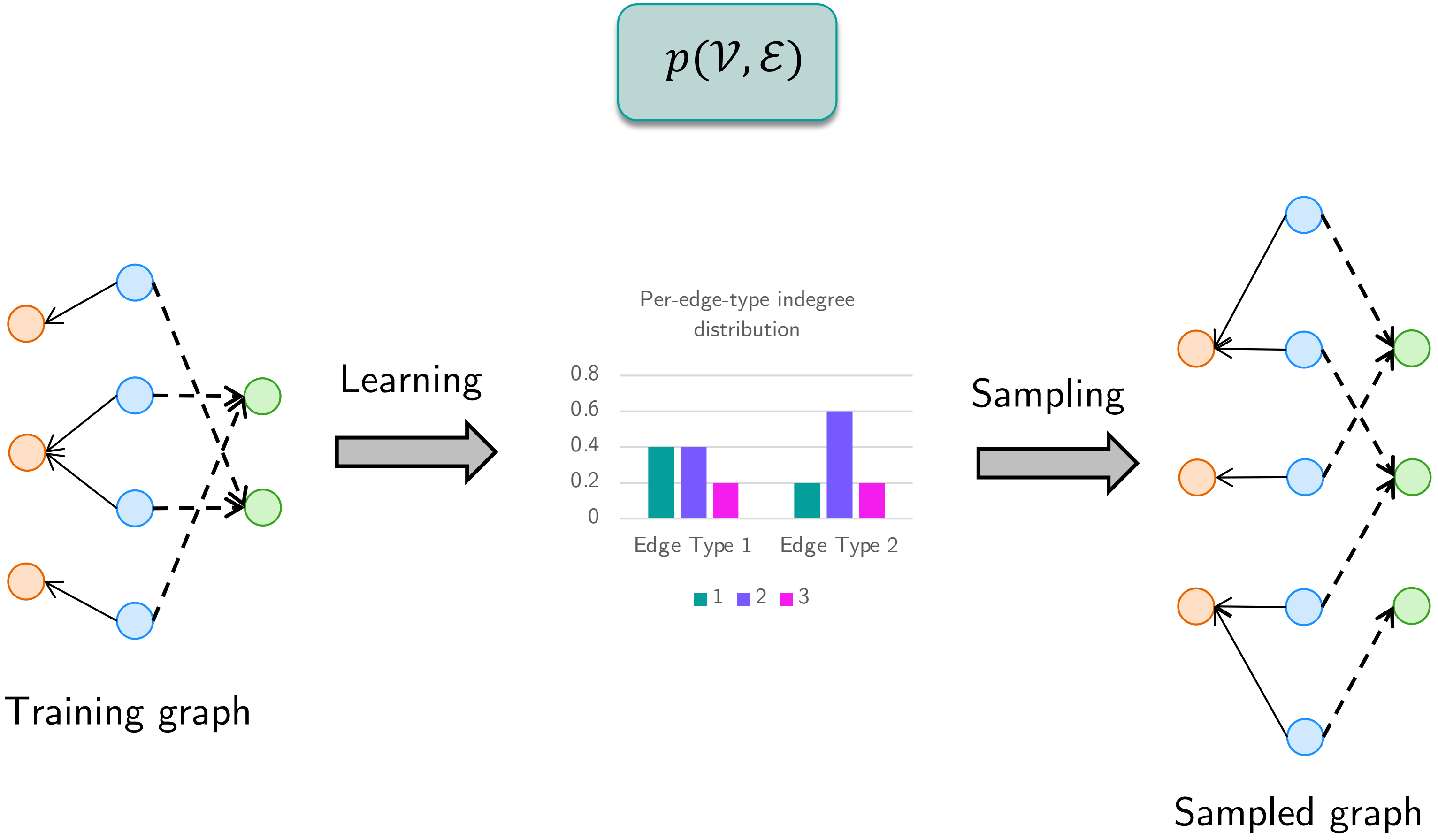
Products

ID	Brand	Category
1	X	Food
2	Y	Clothing

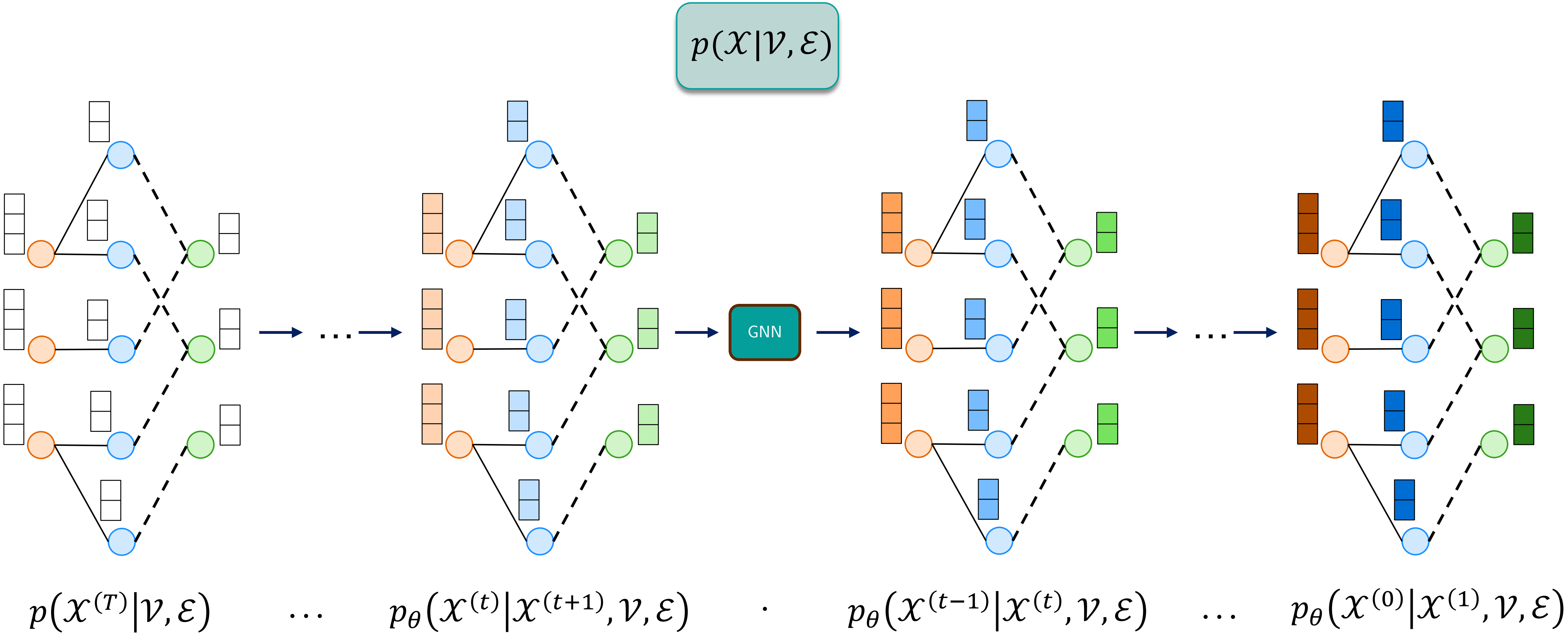
Transactions

ID	Customer ID	Product ID	Price	Time
1	1	2	30.0	09:35
2	2	1	15.7	12:45
3	2	2	35.4	15:34
4	3	1	17.5	19:03

Node Degree-Preserving Random Graph Generation

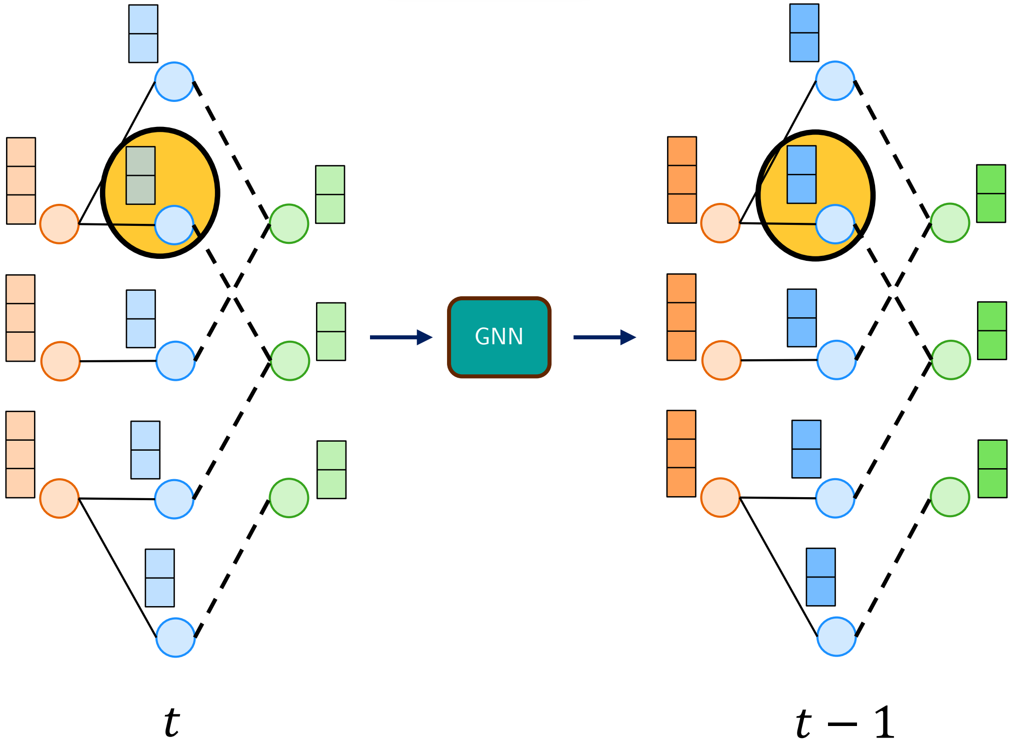


Graph-Conditional Relational Diffusion Model (GRDM)



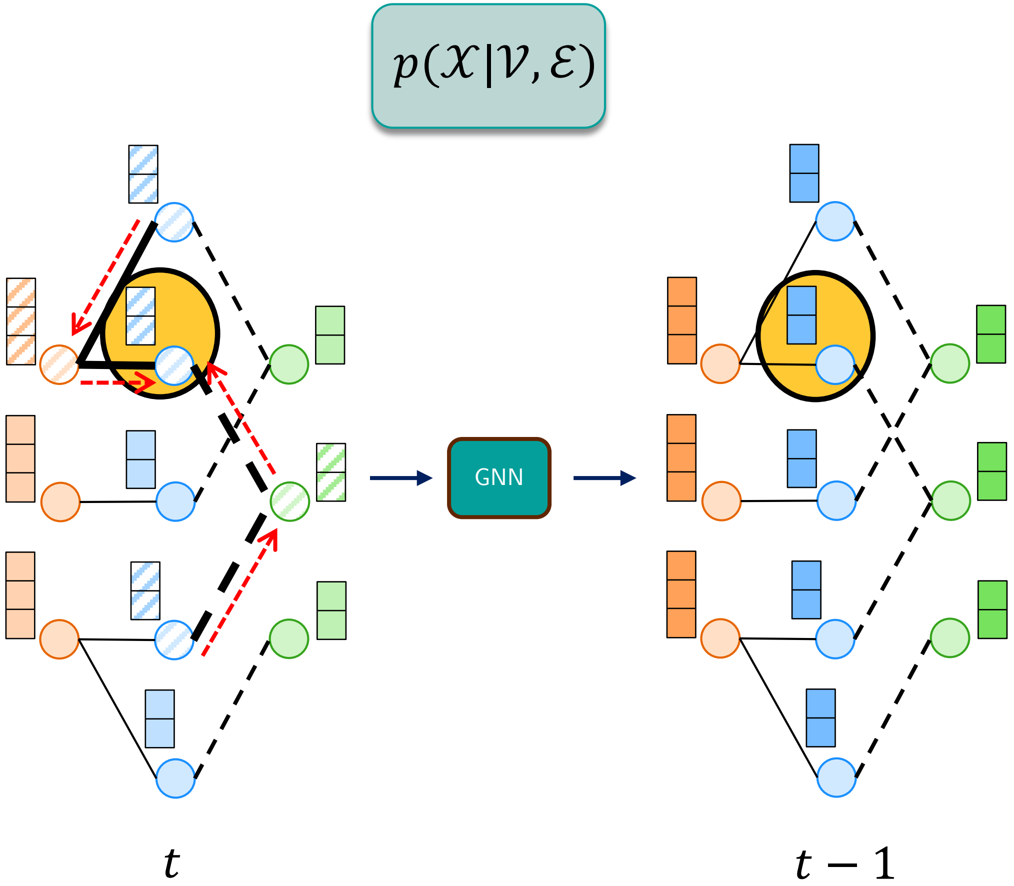
Graph-Conditional Relational Diffusion Model (GRDM)

$$p(\mathcal{X} | \mathcal{V}, \mathcal{E})$$



$$p_\theta\left(\mathcal{X}^{(t-1)} \middle| \mathcal{X}^{(t)}, \mathcal{V}, \mathcal{E}\right) = \prod_{v \in \mathcal{V}} p_\theta\left(x_v^{(t-1)} \middle| \mathcal{G}_{v,K}^{(t)}\right)$$

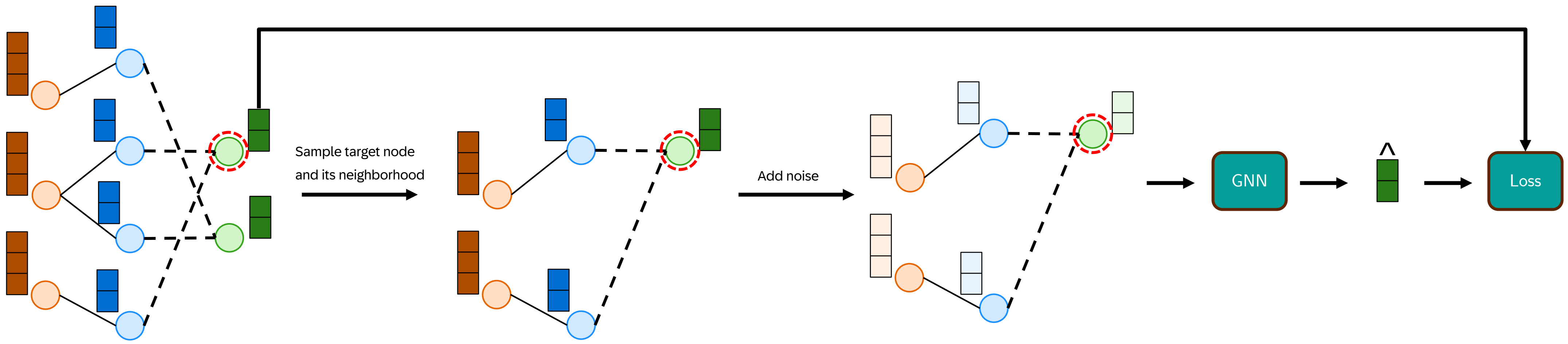
Graph-Conditional Relational Diffusion Model (GRDM)



$$p_\theta(\mathcal{X}^{(t-1)}|\mathcal{X}^{(t)},\mathcal{V},\mathcal{E}) = \prod_{v\in\mathcal{V}} p_\theta(x_v^{(t-1)}|\mathcal{G}_{v,K}^{(t)})$$

Message passing

Training of GRDM



$$L(\theta) = \mathbb{E}_{v\sim\mathcal{V},t\sim\{1,\ldots,T\},\epsilon\sim\mathcal{N}(0,I)}\left[\left\|\epsilon_v - \epsilon_\theta\left(\mathcal{G}_{v,K}^{(t)}, t\right)\right\|^2\right]$$

Evaluation of Synthetic RDB quality

Metrics:

- **Cardinality**: compares graph structure.
- **Column Shapes**: compare column-wise marginal distributions.
- **Intra-Table Trends**: compare correlations of column pairs from the same table.
- **Inter-Table Trends**: compare correlations of column pairs from different tables (1-hop, 2-hop, etc.)

	SDV	SingleTable	Denorm	ClavaDDPM	GRDM (Ours)	Improvement (%)
Berka						
Cardinality		97.06 ± 0.80	96.06 ± 1.15	96.75 ± 0.26	99.70 ± 0.07	+2.72
Column Shapes		94.58 ± 0.01	83.28 ± 0.97	94.60 ± 0.41	96.90 ± 0.07	+2.43
Intra-Table Trends	> 5 tables	91.72 ± 0.23	72.12 ± 0.73	90.53 ± 1.93	98.21 ± 0.05	+7.08
Inter-Table Trends (1-hop)		81.77 ± 1.19	55.77 ± 2.80	83.79 ± 4.21	92.94 ± 0.06	+10.92
Inter-Table Trends (2-hop)		78.09 ± 0.53	57.68 ± 1.67	85.87 ± 2.72	96.04 ± 0.20	+11.84
Inter-Table Trends (3-hop)		75.56 ± 0.34	55.59 ± 1.48	80.98 ± 3.12	93.13 ± 0.47	+15.00
Instacart 05						
Cardinality		94.73 ± 0.14	94.98 ± 0.84	94.91 ± 1.50	99.96 ± 0.01	+5.24
Column Shapes		89.30 ± 0.00	71.83 ± 0.32	90.18 ± 0.43	98.87 ± 0.06	+9.64
Intra-Table Trends	> 5 tables	99.70 ± 0.00	88.74 ± 0.00	99.68 ± 0.02	98.33 ± 0.04	-1.37
Inter-Table Trends (1-hop)		66.93 ± 0.07	62.58 ± 0.05	75.84 ± 0.36	92.03 ± 0.30	+21.35
Inter-Table Trends (2-hop)		16.22 ± 13.41	0.00 ± 0.00	14.40 ± 20.37	96.17 ± 0.15	+492.91
RelBench-F1						
Cardinality		94.94 ± 1.06	93.39 ± 1.56	94.04 ± 2.33	98.26 ± 0.12	+3.50
Column Shapes		95.93 ± 0.16	89.25 ± 0.28	95.19 ± 0.75	97.28 ± 0.29	+1.41
Intra-Table Trends	> 5 tables	95.80 ± 0.29	90.30 ± 0.56	94.71 ± 0.63	96.74 ± 0.36	+0.98
Inter-Table Trends (1-hop)		79.61 ± 0.65	65.60 ± 0.45	88.19 ± 0.27	93.74 ± 0.50	+6.29
Inter-Table Trends (2-hop)		74.10 ± 2.71	62.21 ± 0.38	83.17 ± 1.39	96.81 ± 0.28	+16.40
Movie Lens						
Cardinality		98.99 ± 0.16	98.87 ± 0.26	98.79 ± 0.13	98.80 ± 0.36	-0.19
Column Shapes	> 5 tables	99.19 ± 0.00	78.03 ± 0.17	99.11 ± 0.09	98.22 ± 0.05	-0.98
Intra-Table Trends		98.56 ± 0.01	57.33 ± 0.10	98.52 ± 0.05	96.24 ± 0.24	-2.35
Inter-Table Trends (1-hop)		92.72 ± 0.09	77.45 ± 1.93	92.11 ± 2.12	95.60 ± 0.17	+3.11
CCS						
Cardinality	74.36 ± 8.40	99.37 ± 0.16	26.70 ± 0.20	98.96 ± 0.79	99.79 ± 0.03	+0.42
Column Shapes	69.04 ± 4.38	95.20 ± 0.00	79.29 ± 0.13	92.64 ± 3.93	97.03 ± 0.37	+1.92
Intra-Table Trends	94.84 ± 1.00	98.96 ± 0.00	86.60 ± 0.14	97.75 ± 1.70	93.60 ± 2.61	-5.42
Inter-Table Trends (1-hop)	21.74 ± 9.62	51.62 ± 0.22	57.77 ± 0.69	72.65 ± 8.10	85.54 ± 2.91	+17.74
California						
Cardinality	71.45 ± 0.00	99.89 ± 0.04	99.87 ± 0.02	98.99 ± 0.69	99.96 ± 0.01	+0.07
Column Shapes	72.32 ± 0.00	99.51 ± 0.04	94.99 ± 0.02	98.76 ± 0.27	99.15 ± 0.02	-0.36
Intra-Table Trends	50.23 ± 0.00	98.69 ± 0.08	94.17 ± 0.01	97.65 ± 0.39	98.00 ± 0.01	-0.70
Inter-Table Trends (1-hop)	54.89 ± 0.00	92.96 ± 0.05	87.24 ± 0.10	95.34 ± 0.48	97.68 ± 0.01	+2.45

GRDM consistently outperforms all baselines on Inter-Table Trends across all databases.

Ablation Study

We conduct an ablation study to understand the impact of the number of hops K and the joint modeling aspect.

	$K = 2$ (Default)	$K = 1$	$K = 1$, Sequential
Berka			
Cardinality	99.70 ± 0.07	99.70 ± 0.07	99.70 ± 0.07
Column Shapes	96.90 ± 0.07	97.18 ± 0.03	38.97 ± 3.65
Intra-Table Trends	98.21 ± 0.05	98.47 ± 0.07	52.10 ± 2.98
Inter-Table Trends (1-hop)	92.94 ± 0.06	93.21 ± 0.04	28.53 ± 3.84
Inter-Table Trends (2-hop)	96.04 ± 0.20	95.69 ± 0.19	52.01 ± 2.55
Inter-Table Trends (3-hop)	93.13 ± 0.47	92.84 ± 0.14	49.45 ± 4.14
Instacart 05			
Cardinality	99.96 ± 0.01	99.96 ± 0.01	99.96 ± 0.01
Column Shapes	98.87 ± 0.06	98.40 ± 0.01	67.51 ± 3.93
Intra-Table Trends	98.33 ± 0.04	97.53 ± 0.06	89.33 ± 2.27
Inter-Table Trends (1-hop)	92.03 ± 0.30	88.83 ± 0.70	63.32 ± 2.01
Inter-Table Trends (2-hop)	96.17 ± 0.15	93.30 ± 0.08	13.41 ± 3.64
RelBench-F1			
Cardinality	98.26 ± 0.12	98.26 ± 0.12	98.26 ± 0.12
Column Shapes	97.28 ± 0.29	96.76 ± 0.06	94.83 ± 0.09
Intra-Table Trends	96.74 ± 0.36	95.92 ± 0.04	95.01 ± 0.75
Inter-Table Trends (1-hop)	93.74 ± 0.50	93.33 ± 0.17	90.82 ± 0.37
Inter-Table Trends (2-hop)	96.81 ± 0.28	96.33 ± 0.08	87.73 ± 0.86

Thank You!

If you have any questions, send us an email at:

a.Ketata@tum.de