ClavaDDPM: Multi-relational Data Synthesis with Cluster-guided Diffusion Models

Wei Pang

UNIVERSITY OF WATERLOO

Single-Table Synthesis

Individual ID	Married	Age	Language
1	F	1	English
2	F	2	English
3	F	10	Chinese
4	T	25	English
5	F	35	English
6	F	7	French
7	F	9	French
8	F	14	English
9	F	29	English
10	T	78	Chinese

Single-Table Synthesis

Individual ID	Married	Age	Language
1	F	1	English
2	F	2	English
3	F	10	Chinese
4	T	25	English
5	F	35	English
6	F	7	French
7	F	9	French
8	F	14	English
9	F	29	English
10	T	78	Chinese

- Model the correlation between columns.
- Each row is i.i.d.

Multi-table Synthesis

Individual ID	Married	Age	Language	Household ID
1	F	1	English	1
2	F	2	English	1
3	F	10	Chinese	2
4	T	25	English	1
5	F	35	English	3
6	F	7	French	4
7	F	9	French	4
8	F	14	English	1
9	F	29	English	3
10	T	78	Chinese	2

Household ID	Country	Size
1	US	4
2	China	2
3	Canada	2
4	Canada	2

Multi-table Synthesis

Individual ID	Married	Age	Language	Household ID
1	F	1	English	1
2	F	2	English	1
3	F	10	Chinese	2
4	T	25	English	1
5	F	35	English	3
6	F	7	French	4
7	F	9	French	4
8	F	14	English	1
9	F	29	English	3
10	T	78	Chinese	2

A ***US*** household with ***4*** individuals

Household ID	Country	Size
1	US	4
2	China	2
3	Canada	2
4	Canada	2

UNIVERSITY OF WATERLOO

Multi-table Synthesis

Individual ID	Married	Age	Language	Household ID
1	F	1	English	1
2	F	2	English	1
3	F	10	Chinese	2
4	T	25	English	1
5	F	35	English	3
6	F	7	French	4
7	F	9	French	4
8	F	14	English	1
9	F	29	English	3
10	T	78	Chinese	2

A **Chinese** household with ***2*** individuals

Household ID	Country	Size
1	US	4
2	China	2
3	Canada	2
4	Canada	2

Multi-table Synthesis

Individual ID	Married	Age	Language	Household ID
1	F	1	English	1
2	F	2	English	1
3	F	10	Chinese	2
4	T	25	English	1
5	F	35	English	3
6	F	7	French	4
7	F	9	French	4
8	F	14	English	1
9	F	29	English	3
10	T	78	Chinese	2

A **Canadian** household with **2** individuals

Household ID	Country	Size
1	US	4
2	China	2
3	Canada	2
4	Canada	2

Multi-table Synthesis

Individual ID	Married	Age	Language	Household ID
1	F	1	English	1
2	F	2	English	1
3	F	10	Chinese	2
4	T	25	English	1
5	F	35	English	3
6	F	7	French	4
7	F	9	French	4
8	F	14	English	1
9	F	29	English	3
10	T	78	Chinese	2

A **Canadian** household with ***4*** individuals

Household ID	Country	Size
1	US	4
2	China	2
3	Canada	2
4	Canada	2

Multi-table Synthesis

Individual ID	Married	Age	Language	Household ID
1	F	1	English	1
2	F	2	English	1
3	F	10	Chinese	2
4	T	25	English	1
5	F	35	English	3
6	F	7	French	4
7	F	9	French	4
8	F	14	English	1
9	F	29	English	3
10	T	78	Chinese	2

Household ID	Country	Size
1	US	4
2	China	2
3	Canada	2
4	Canada	2

- **Inter-column** correlation still exists.

Multi-table Synthesis

Individual ID	Married	Age	Language	Household ID
1	F	1	English	1
2	F	2	English	1
3	F	10	Chinese	2
4	T	25	English	1
5	F	35	English	3
6	F	7	French	4
7	F	9	French	4
8	F	14	English	1
9	F	29	English	3
10	T	78	Chinese	2

Household ID	Country	Size
1	US	4
2	China	2
3	Canada	2
4	Canada	2

- **Inter-column** correlation still exists.
- **Inter-table** columns can also be correlated.

Individual's language is strongly correlated with **Household**'s country!

Multi-table Synthesis

Individual ID	Married	Age	Language	Household ID
1	F	1	English	1
2	F	2	English	1
3	F	10	Chinese	2
4	T	25	English	1
5	F	35	English	3
6	F	7	French	4
7	F	9	French	4
8	F	14	English	1
9	F	29	English	3
10	T	78	Chinese	2

Household ID	Country	Size
1	US	4
2	China	2
3	Canada	2
4	Canada	2

- **Inter-column** correlation still exists.
- **Inter-table** columns can also be correlated.
- Child table **rows are no longer i.i.d.**, but dependent on parent table.

Individuals within the same **Household** tend to speak the same language!

Multi-table Synthesis

Individual ID	Married	Age	Language	Household ID
1	F	1	English	1
2	F	2	English	1
3	F	10	Chinese	2
4	T	25	English	1
5	F	35	English	3
6	F	7	French	4
7	F	9	French	4
8	F	14	English	1
9	F	29	English	3
10	T	78	Chinese	2

Household ID	Country	Size
1	US	4
2	China	2
3	Canada	2
4	Canada	2

- **Inter-column** correlation still exists.
- **Inter-table** columns can also be correlated.
- Child table **rows are no longer i.i.d.**, but dependent on parent table.
- The **size of a group** referring to the same parent is correlated with parent table.

The number of **Individuals** within the same **Household** is also dependent on parent table!

Multi-table Synthesis: Motivation

To address these:

- **Inter-column** correlation still exists.
- **Inter-table** columns can also be correlated.
- Child table **rows are no longer i.i.d.**, but dependent on parent table.
- The **size of a group** referring to the same parent is correlated with parent table.

Multi-table Synthesis: Motivation

To address these:

- **Inter-column** correlation still exists.
- **Inter-table** columns can also be correlated.
- Child table **rows are no longer i.i.d.**, but dependent on parent table.
- The **size of a group** referring to the same parent is correlated with parent table.

We aim to design a model that:

- Maintains **single-table quality**.
- Captures **inter-table correlations**.
- Models **row-wise correlations**.
- Models **group size distributions**.

Multi-relational Data

Multi-relational database:

$$\mathcal{R} = (R_1, \dots, R_m)$$

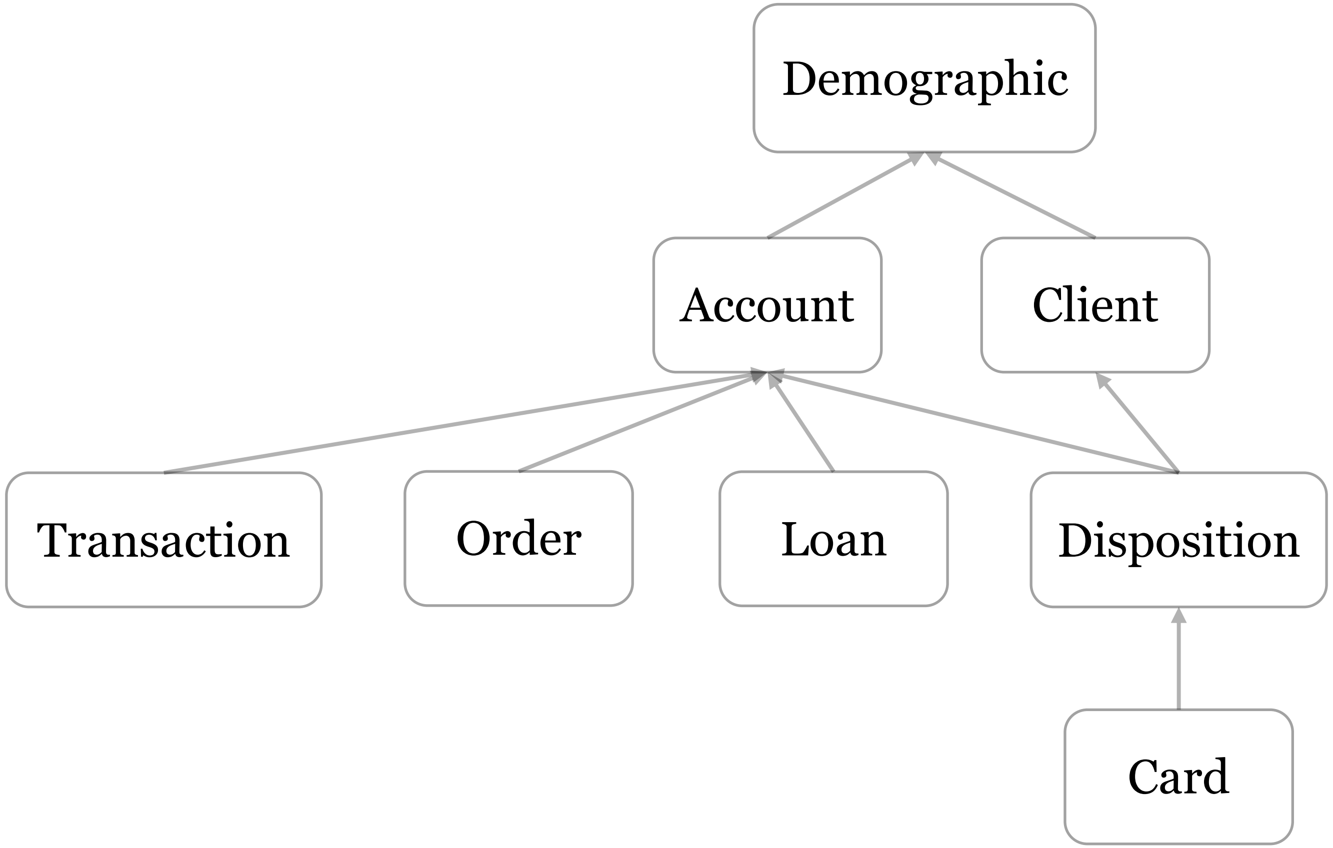
Multi-relational database with foreign key constraints (DAG):

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

$$\mathcal{E} = \{(R_i \to R_j) | i, j \in \{1, \dots, m\}, i \neq j, R_i \text{ refers to } R_j\}$$

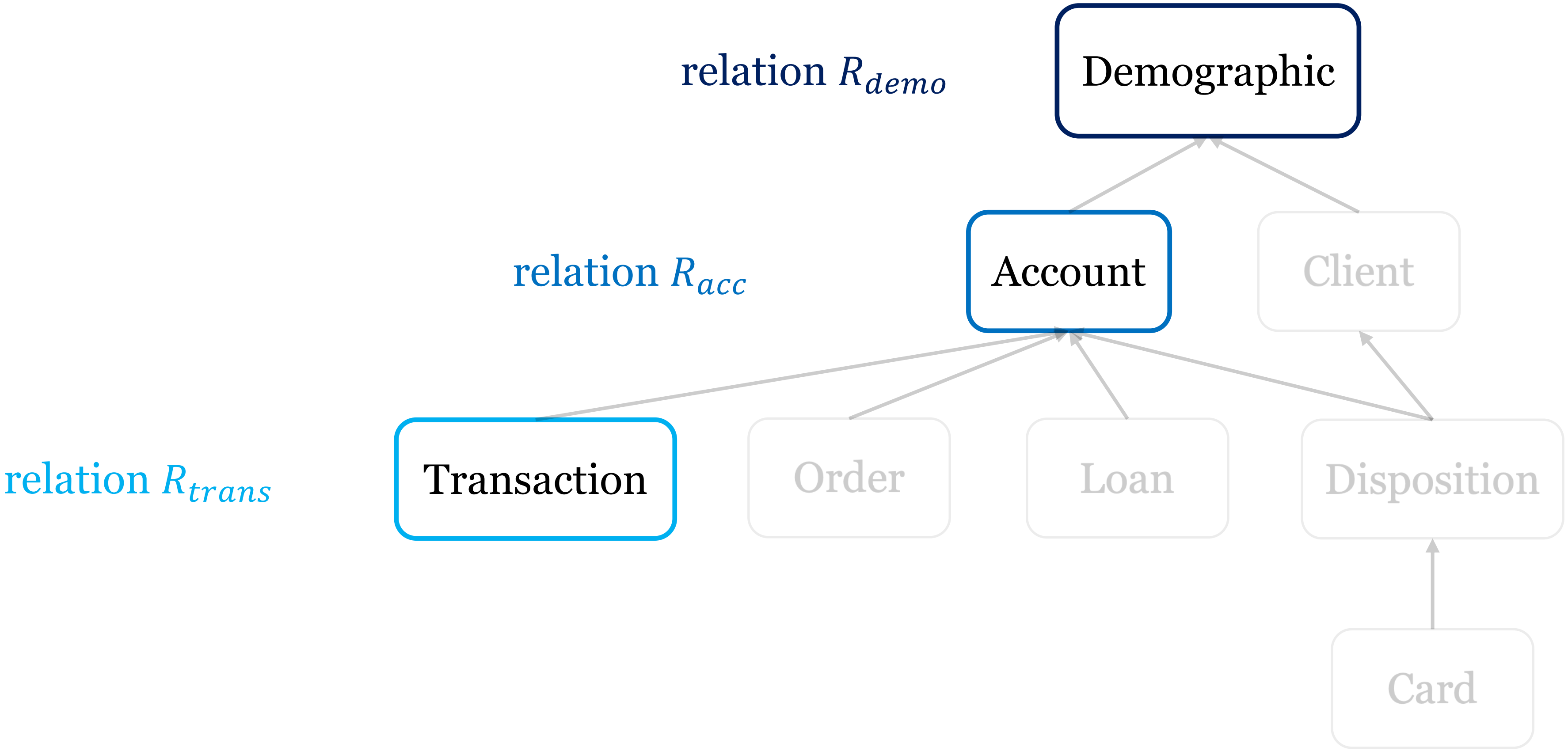
We also call $(R_i \to R_j)$ a **parent-child** relationship.

Real-world Data

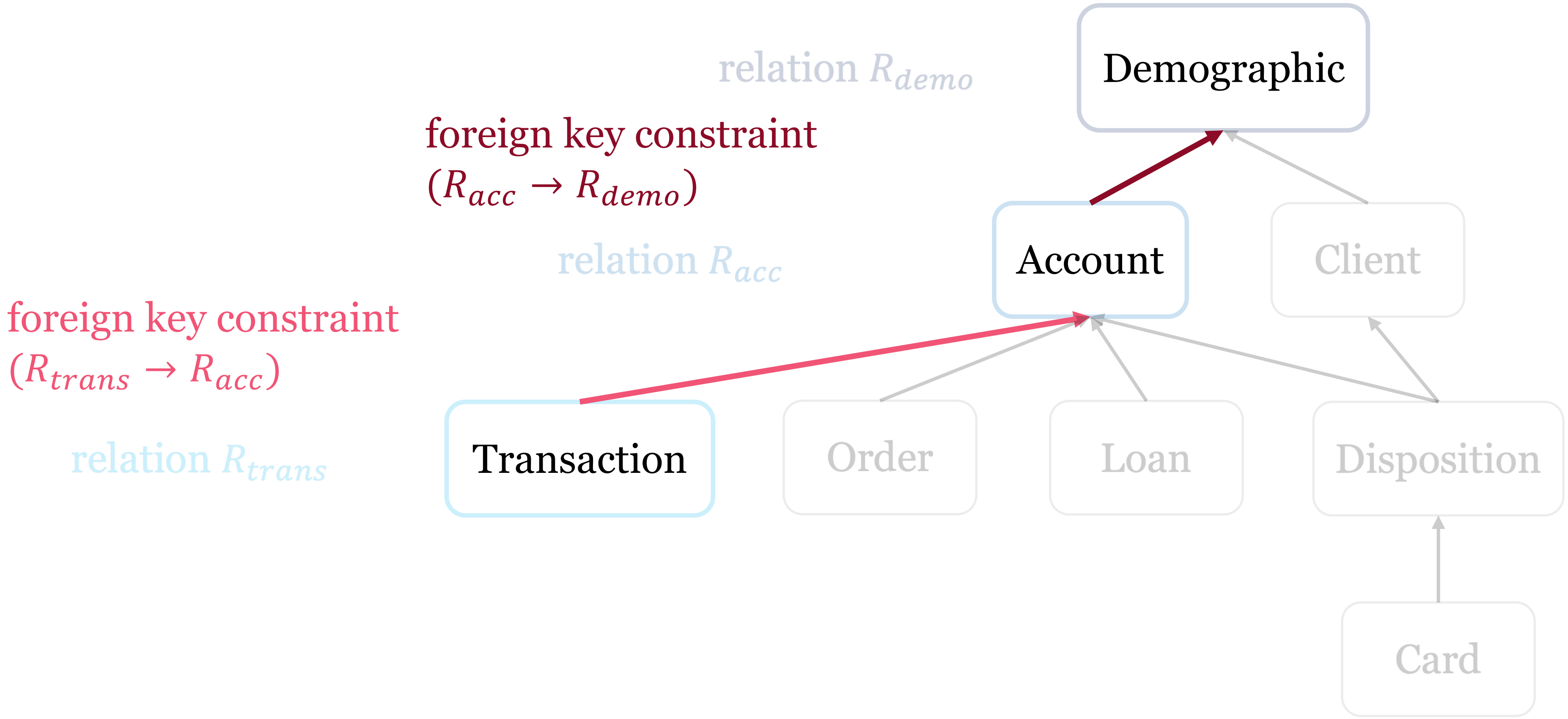


UNIVERSITY OF WATERLOO

Real-world Data



Real-world Data



Real-world Data

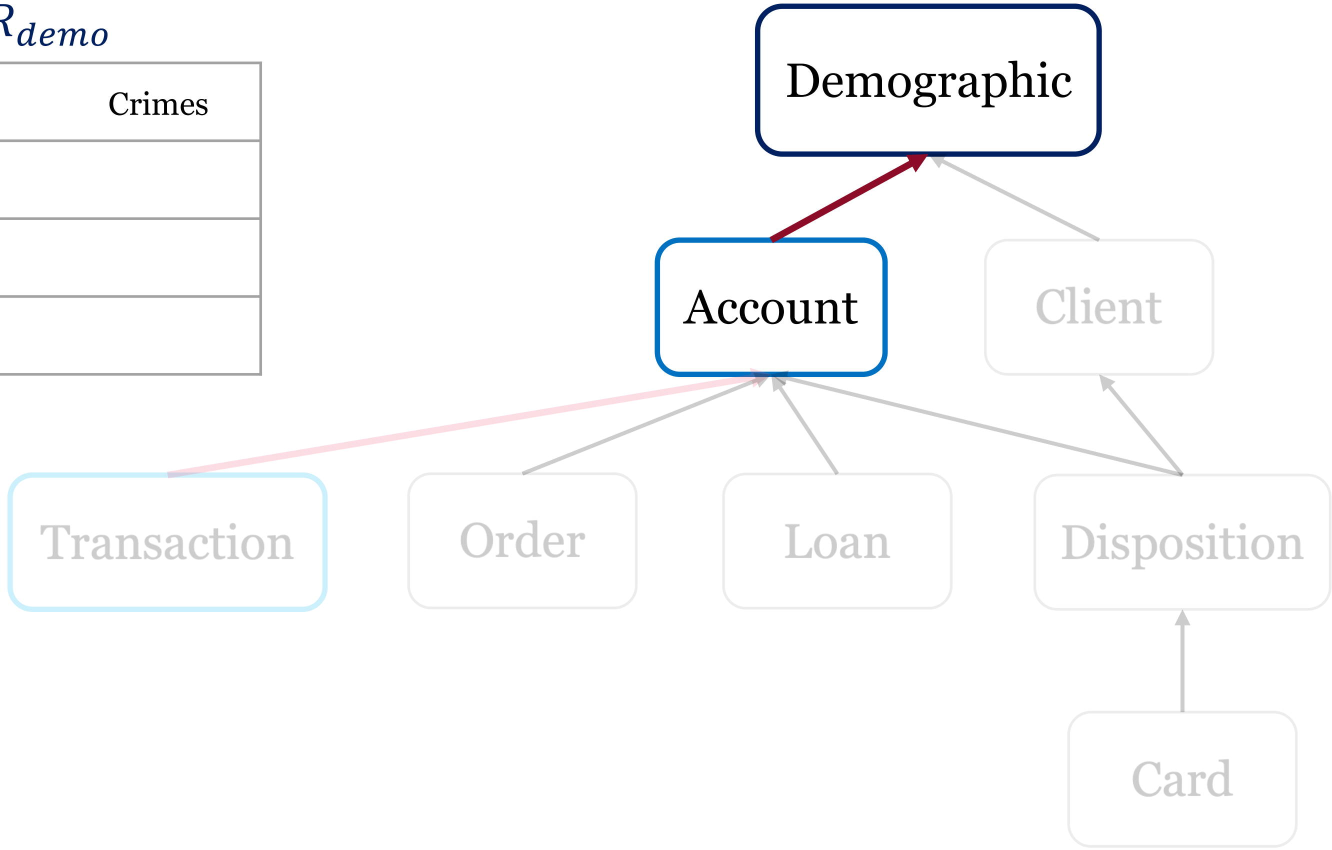
relation R_{demo}

Demo ID	Name	...	Crimes
1			
2			
3			

foreign key constraint ($R_{acc} \rightarrow R_{demo}$)

relation R_{acc}

Acc ID	Demo ID	Date	...	Freq
1	2			
2	3			
3	1			
4	1			



A foreign key group with size 2.

Real-world Data

relation R_{acc}

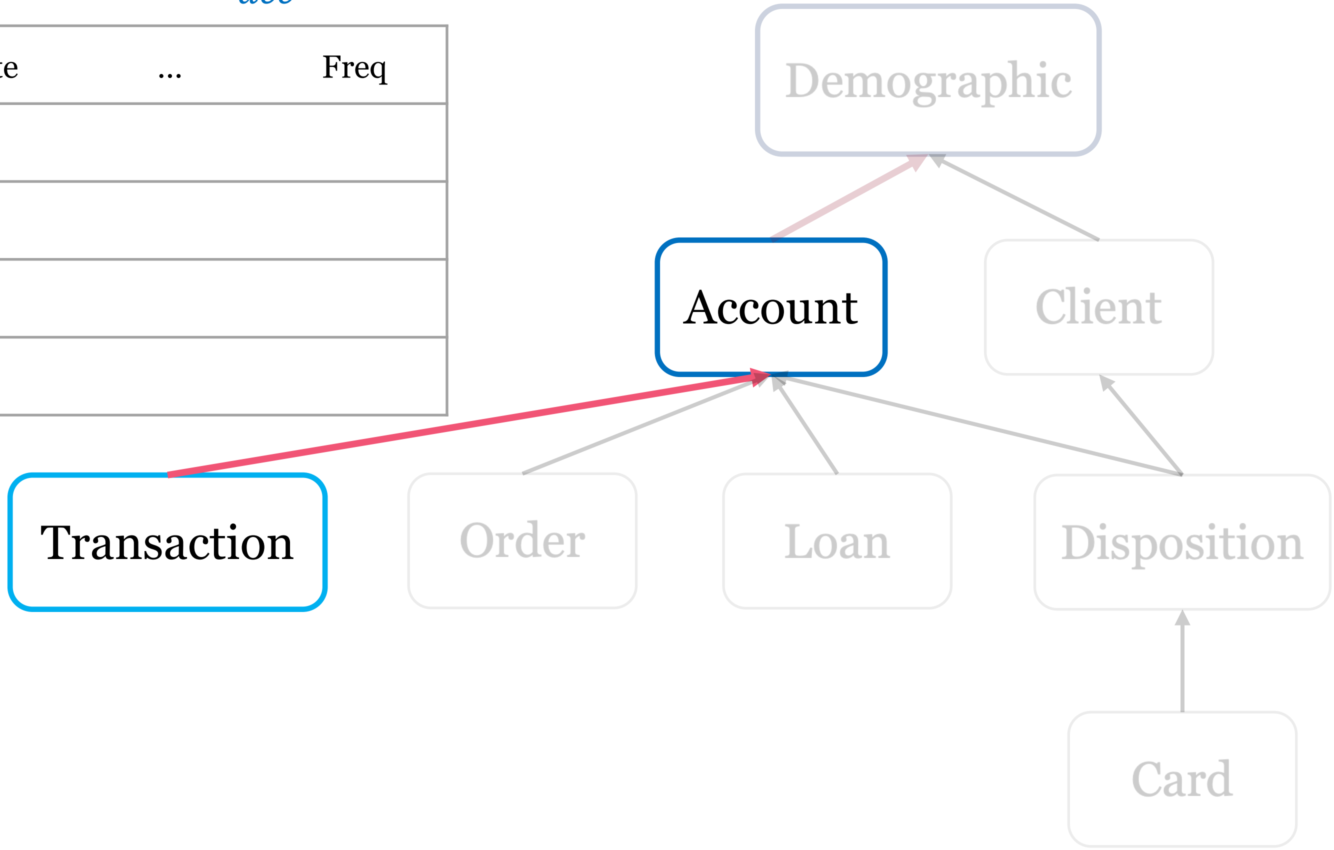
Acc ID	Demo ID	Date	...	Freq
1	2			
2	3			
3	1			
4	1			

foreign key constraint
$(R_{trans} \rightarrow R_{acc})$

relation R_{trans}

Trans ID	Acc ID	Amount	...	Type
1	4			
2	4			
3	4			
4	3			
5	1			

A foreign key group with size 3.



Multi-table Synthesis

- Follows the same assumption on **categorical** and **numerical** values.
- Assumptions:
  - Different columns are correlated.
  - Different tables are correlated. (parent-child relationships)
  - Rows are not i.i.d. due to foreign key constraints.
- Desiderata:
  - **Inter-column** correlations within the same table.
  - **Intra-group** correlations within the same foreign key group.
  - **Inter-table** correlations.
  - **Group size** distributions.

ClavaDDPM: Gaussian Diffusion as Backbone

$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}\right)$ Gaussian transition **forward** process

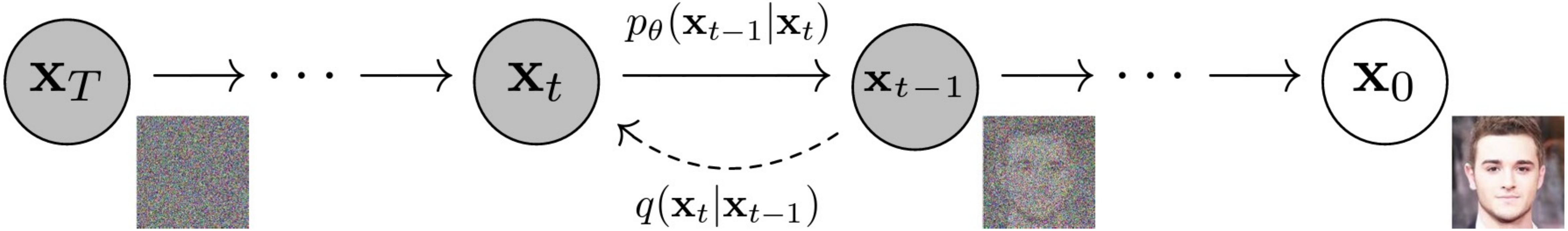
$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)\right)$ Learnable parameterized **reverse** process with a Gaussian form

$$\log\left(p_{\theta,\varphi}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{y})\right) \approx \log\left(p(\boldsymbol{z})\right) + C$$

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{g}, \boldsymbol{\Sigma})$$

Classifier-guided sampling

$$\boldsymbol{g} = \nabla_{\boldsymbol{x}_{t-1}} \log\left(p_\varphi(\boldsymbol{y}|\boldsymbol{x}_t)|_{\boldsymbol{x}_{t-1}=\boldsymbol{\mu}}\right)$$



UNIVERSITY OF WATERLOO

ClavaDDPM: Two Tables

Child ID	Parent ID	X
1	2	x_1
2	2	x_2
3	1	x_3
4	3	x_4
5	3	x_5
6	3	x_6
7	4	x_7
8	4	x_8
9	5	x_9
10	5	x_{10}

Parent ID	Y
1	y_1
2	y_2
3	y_3
4	y_4
5	y_5

ClavaDDPM: Two Tables

Child ID	Parent ID	X
1	2	x_1
2	2	x_2
3	1	x_3
4	3	x_4
5	3	x_5
6	3	x_6
7	4	x_7
8	4	x_8
9	5	x_9
10	5	x_{10}

Foreign key group g_1

Parent ID	Y
1	y_1
2	y_2
3	y_3
4	y_4
5	y_5

ClavaDDPM: Two Tables

Child ID	Parent ID	X
1	2	x_1
2	2	x_2
3	1	x_3
4	3	x_4
5	3	x_5
6	3	x_6
7	4	x_7
8	4	x_8
9	5	x_9
10	5	x_{10}

Foreign key group g_2

Parent ID	Y
1	y_1
2	y_2
3	y_3
4	y_4
5	y_5

ClavaDDPM: Two Tables

Child ID	Parent ID	X
1	2	x_1
2	2	x_2
3	1	x_3
4	3	x_4
5	3	x_5
6	3	x_6
7	4	x_7
8	4	x_8
9	5	x_9
10	5	x_{10}

Foreign key group g_3

Parent ID	Y
1	y_1
2	y_2
3	y_3
4	y_4
5	y_5

ClavaDDPM: Two Tables

Child ID	Parent ID	X
1	2	x_1
2	2	x_2
3	1	x_3
4	3	x_4
5	3	x_5
6	3	x_6
7	4	x_7
8	4	x_8
9	5	x_9
10	5	x_{10}

Foreign key group g_4

Parent ID	Y
1	y_1
2	y_2
3	y_3
4	y_4
5	y_5

ClavaDDPM: Two Tables

Child ID	Parent ID	X
1	2	x_1
2	2	x_2
3	1	x_3
4	3	x_4
5	3	x_5
6	3	x_6
7	4	x_7
8	4	x_8
9	5	x_9
10	5	x_{10}

Foreign key group g_5

Parent ID	Y
1	y_1
2	y_2
3	y_3
4	y_4
5	y_5

ClavaDDPM: Two Tables

Child ID	Parent ID	X
1	2	g_2
2	2	
3	1	g_1
4	3	g_3
5	3	
6	3	
7	4	g_4
8	4	
9	5	g_5
10	5	

Instead of modeling $\boldsymbol{x}$ directly, we model foreign key groups $\boldsymbol{g}$.

Parent ID	Y
1	y_1
2	y_2
3	y_3
4	y_4
5	y_5

ClavaDDPM: Modelling

Assumptions

- Each parent row $\boldsymbol{y}$ is i.i.d.
- The child row distribution $\boldsymbol{x}$ is and only is constrained by its parent y.
  - Child table X is formed by a collection of foreign key groups $X = \{g_1, \dots, g_{|y|}\}$.
  - Each foreign key group g_j is formed by a collection of rows $g_j = \{x_j^i | i = 1, \dots, |g_j|\}$, which corresponds to parent row y_j.

UNIVERSITY OF WATERLOO

ClavaDDPM: Modelling

Idea

- Model parent table distribution $p(y)$.
- Model conditional foreign key group distribution $p(g|y)$.

Difficulties

- Parent table space Y can be sparse and badly shaped.
- Vectors y can be high-dimensional.

Modelling the full conditional distribution $p(g|y)$ can be **costly** and leads to **bad performance**.

ClavaDDPM: Cluster as Latents

- Instead of learning the full conditional distribution $p(g|y)$ directly:
  - We quantize $(\boldsymbol{g}, \boldsymbol{y})$ into codebook $\boldsymbol{c}$. We call this *relation-aware clustering*.
  - Use c as a proxy for modelling foreign key group distributions.

$$p(g_j, y_j) = \sum_c p(g_j|c)p(y, c)$$

- Gaussian Mixture Models (GMM) clustering.

ClavaDDPM: Cluster as Latents

Child ID	Parent ID	X
1	2	g_2
2	2	
3	1	g_1
4	3	g_3
5	3	
6	3	
7	4	g_4
8	4	
9	5	g_5
10	5	

JOIN

Parent ID	Y
1	y_1
2	y_2
3	y_3
4	y_4
5	y_5

⇒

Child ID	Parent ID	X	Y
1	2	g_2	y_2
2	2		
3	1	g_1	y_1
4	3	g_3	y_3
5	3		
6	3		
7	4	g_4	y_4
8	4		
9	5	g_5	y_5
10	5		

ClavaDDPM: Cluster as Latents

Child ID	Parent ID	X	Y
1	2	g_2	y_2
2	2		
3	1	g_1	y_1
4	3		
5	3	g_3	y_3
6	3		
7	4	g_4	y_4
8	4		
9	5	g_5	y_5
10	5		

GMM $\Rightarrow$

Child ID	Parent ID	X	Y	C
1	2	g_2	y_2	c_2
2	2			
3	1	g_1	y_1	c_1
4	3			
5	3	g_3	y_3	c_3
6	3			
7	4	g_4	y_4	c_2
8	4			
9	5	g_5	y_5	c_3
10	5			

Same cluster indicates similar parent and children, serving as a quantization.

ClavaDDPM: Cluster as Latents

Child ID	Parent ID	X	Y	C
1	2	g_2	y_2	c_2
2	2			
3	1	g_1	y_1	c_1
4	3	g_3	y_3	c_3
5	3			
6	3			
7	4	g_4	y_4	c_2
8	4			
9	5	g_5	y_5	c_3
10	5			

Keep parents $\Longrightarrow$

Augmented parent table

Parent ID	Y	C
2	y_2	c_2
1	y_1	c_1
3	y_3	c_3
4	y_4	c_2
5	y_5	c_3

ClavaDDPM: Cluster as Latents

Original parent table

Parent ID	Y
1	y_1
2	y_2
3	y_3
4	y_4
5	y_5

Augmentation $\Longrightarrow$

Augmented parent table

Parent ID	Y	C
2	y_2	c_2
1	y_1	c_1
3	y_3	c_3
4	y_4	c_2
5	y_5	c_3

ClavaDDPM: Cluster as Latents

Child ID	Parent ID	X	Y
1	2	g_2	y_2
2	2		
3	1	g_1	y_1
4	3	g_3	y_3
5	3		
6	3		
7	4	g_4	y_4
8	4		
9	5	g_5	y_5
10	5		

GMM $\Rightarrow$

Child ID	Parent ID	X	Y	C
1	2	g_2	y_2	c_2
2	2			
3	1	g_1	y_1	c_1
4	3	g_3	y_3	c_3
5	3			
6	3			
7	4	g_4	y_4	c_2
8	4			
9	5	g_5	y_5	c_3
10	5			

Sampled from $\boldsymbol{g}_2|\boldsymbol{c}_2$

ClavaDDPM: Cluster as Latents

Child ID	Parent ID	X	Y
1	2	g_2	y_2
2	2		
3	1	g_1	y_1
4	3	g_3	y_3
5	3		
6	3		
7	4	g_4	y_4
8	4		
9	5	g_5	y_5
10	5		

GMM

Child ID	Parent ID	X	Y	C
1	2	g_2	y_2	c_2
2	2			
3	1	$\boxed{g_1}$	y_1	c_1
4	3	g_3	y_3	c_3
5	3			
6	3			
7	4	g_4	y_4	c_2
8	4			
9	5	g_5	y_5	c_3
10	5			

Sampled from $\boldsymbol{g}_1|\boldsymbol{c}_1$

ClavaDDPM: Cluster as Latents

Child ID	Parent ID	X	Y
1	2	g_2	y_2
2	2		
3	1	g_1	y_1
4	3	g_3	y_3
5	3		
6	3		
7	4	g_4	y_4
8	4		
9	5	g_5	y_5
10	5		

GMM ⇒

Child ID	Parent ID	X	Y	C
1	2	g_2	y_2	c_2
2	2			
3	1	g_1	y_1	c_1
4	3	$\boxed{g_3}$	y_3	c_3
5	3			
6	3			
7	4	g_4	y_4	c_2
8	4			
9	5	g_5	y_5	c_3
10	5			

Sampled from $\boldsymbol{g}_3|\boldsymbol{c}_3$

ClavaDDPM: Cluster as Latents

Child ID	Parent ID	X	Y
1	2	g_2	y_2
2	2		
3	1	g_1	y_1
4	3	g_3	y_3
5	3		
6	3		
7	4	g_4	y_4
8	4		
9	5	g_5	y_5
10	5		

$\xrightarrow{\text{GMM}}$

Child ID	Parent ID	X	Y	C
1	2	g_2	y_2	c_2
2	2			
3	1	g_1	y_1	c_1
4	3	g_3	y_3	c_3
5	3			
6	3			
7	4	g_4	y_4	c_2
8	4			
9	5	g_5	y_5	c_3
10	5			

Sampled from $\boldsymbol{g}_4|\boldsymbol{c}_2$

ClavaDDPM: Cluster as Latents

<table>
<tr><th>Child ID</th><th>Parent ID</th><th>X</th><th>Y</th></tr>
<tr><td>1</td><td>2</td><td rowspan="2">g_2</td><td rowspan="2">y_2</td></tr>
<tr><td>2</td><td>2</td></tr>
<tr><td>3</td><td>1</td><td>g_1</td><td>y_1</td></tr>
<tr><td>4</td><td>3</td><td rowspan="3">g_3</td><td rowspan="3">y_3</td></tr>
<tr><td>5</td><td>3</td></tr>
<tr><td>6</td><td>3</td></tr>
<tr><td>7</td><td>4</td><td rowspan="2">g_4</td><td rowspan="2">y_4</td></tr>
<tr><td>8</td><td>4</td></tr>
<tr><td>9</td><td>5</td><td rowspan="2">g_5</td><td rowspan="2">y_5</td></tr>
<tr><td>10</td><td>5</td></tr>
</table>

$\xrightarrow{\text{GMM}}$

<table>
<tr><th>Child ID</th><th>Parent ID</th><th>X</th><th>Y</th><th>C</th></tr>
<tr><td>1</td><td>2</td><td rowspan="2">g_2</td><td rowspan="2">y_2</td><td rowspan="2">c_2</td></tr>
<tr><td>2</td><td>2</td></tr>
<tr><td>3</td><td>1</td><td>g_1</td><td>y_1</td><td>c_1</td></tr>
<tr><td>4</td><td>3</td><td rowspan="3">g_3</td><td rowspan="3">y_3</td><td rowspan="3">c_3</td></tr>
<tr><td>5</td><td>3</td></tr>
<tr><td>6</td><td>3</td></tr>
<tr><td>7</td><td>4</td><td rowspan="2">g_4</td><td rowspan="2">y_4</td><td rowspan="2">c_2</td></tr>
<tr><td>8</td><td>4</td></tr>
<tr><td>9</td><td>5</td><td rowspan="2">$\boxed{g_5}$</td><td rowspan="2">y_5</td><td rowspan="2">c_3</td></tr>
<tr><td>10</td><td>5</td></tr>
</table>

Sampled from $\boldsymbol{g}_5|\boldsymbol{c}_3$

ClavaDDPM: Cluster as Latents

Child ID	Parent ID	X	Y
1	2	g_2	y_2
2	2		
3	1	g_1	y_1
4	3	g_3	y_3
5	3		
6	3		
7	4	g_4	y_4
8	4		
9	5	g_5	y_5
10	5		

GMM $\Rightarrow$

Child ID	Parent ID	X	Y	C
1	2	g_2	y_2	c_2
2	2			
3	1	g_1	y_1	c_1
4	3	$\boxed{g_3}$	y_3	c_3
5	3			
6	3			
7	4	g_4	y_4	c_2
8	4			
9	5	g_5	y_5	c_3
10	5			

How many rows does g_3 contain?

ClavaDDPM: Group Size

- Model group size $s = |g|$.
- Two-step generation:
  - Sample group size s.
  - Sample s rows in foreign key group g.

$$p(g_j|c) = p(s_j|c) \prod_{i=1}^{s_j} p(x_j^i|c)$$

UNIVERSITY OF WATERLOO

ClavaDDPM: Group Size

Sample $p(c, y)$ $\boxed{c}$

ClavaDDPM: Group Size

Sample $p(c, y)$ $\boxed{c} \xrightarrow{\text{Sample } p(s|c)}$

ClavaDDPM: Group Size

Sample $p(c, y)$

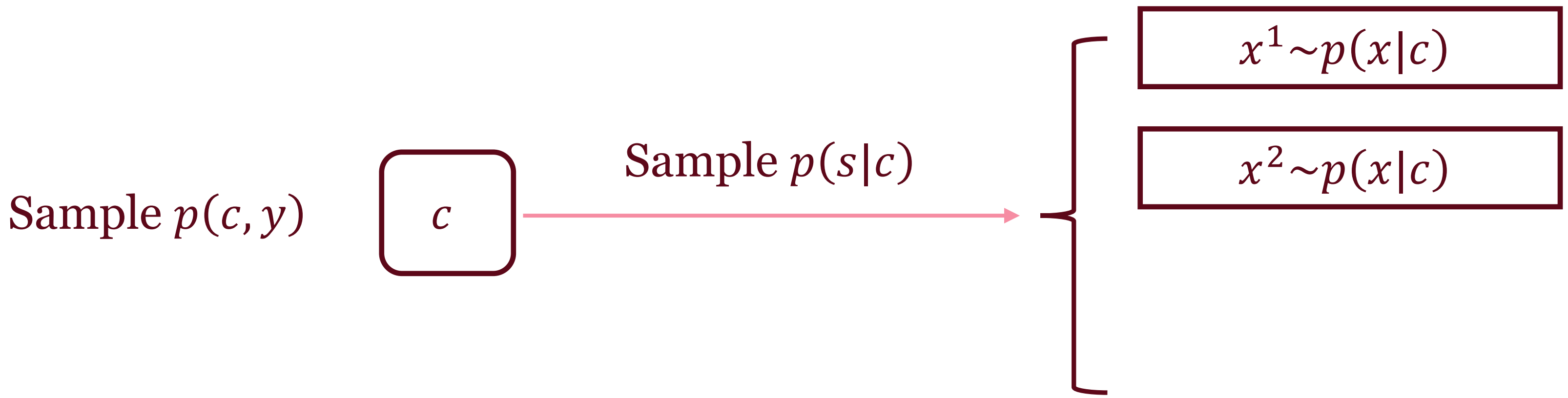
c

Sample $p(s|c)$

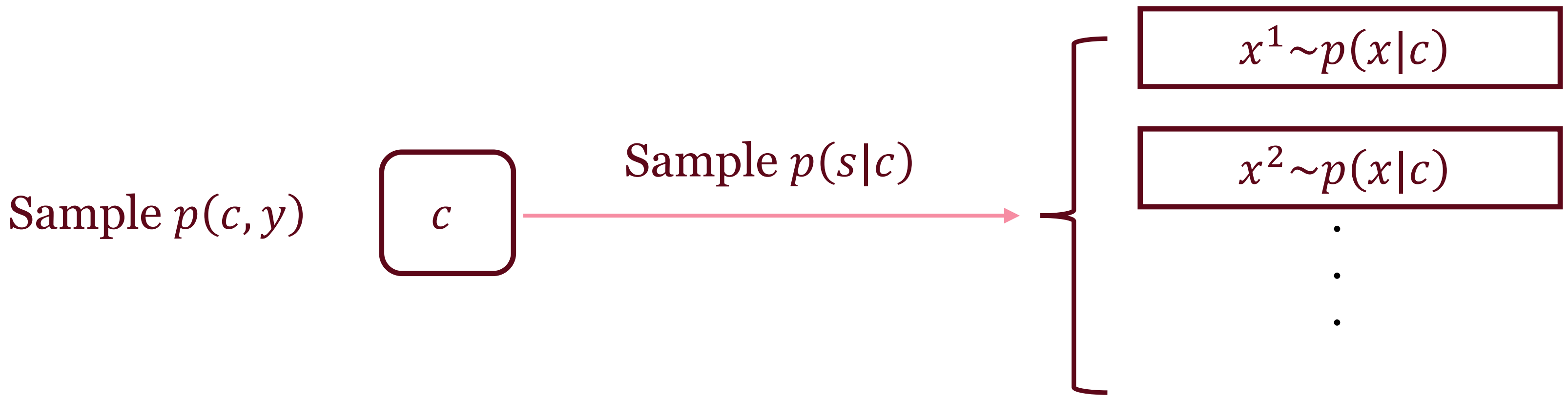
$x^1 \sim p(x|c)$

UNIVERSITY OF WATERLOO

ClavaDDPM: Group Size

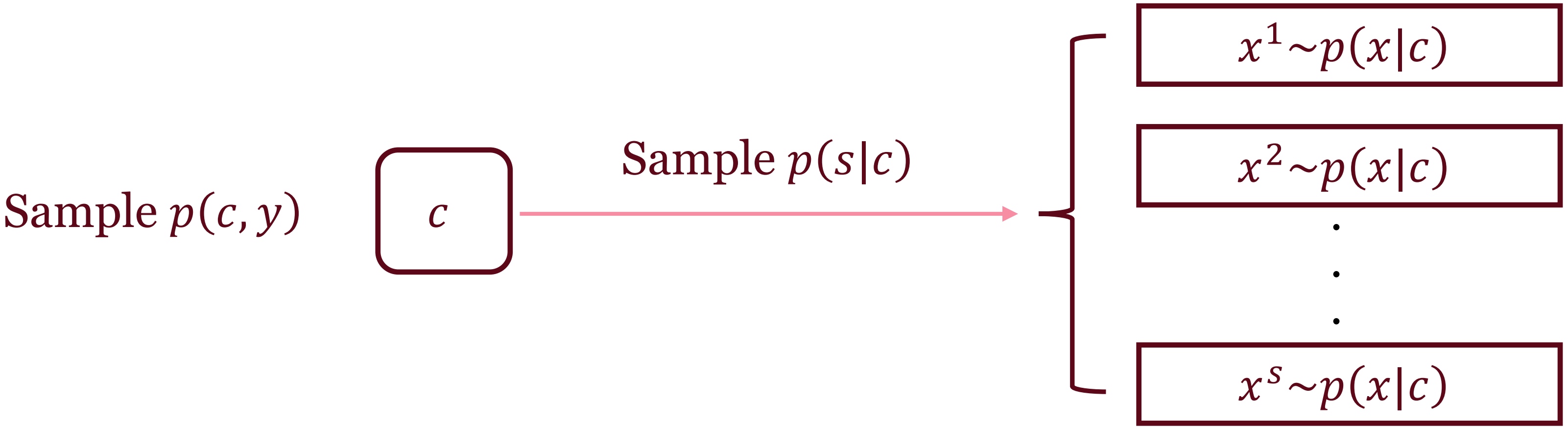


ClavaDDPM: Group Size



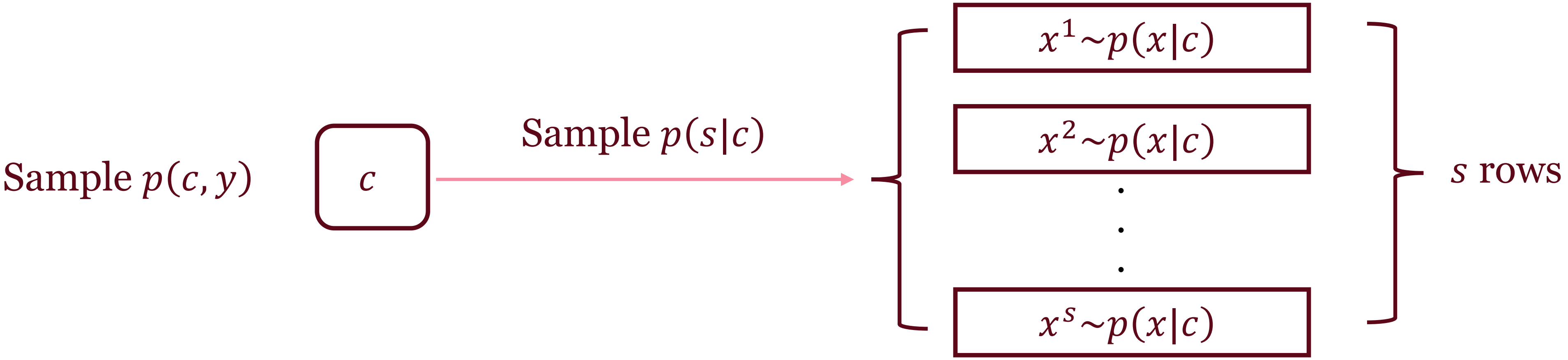
UNIVERSITY OF WATERLOO

ClavaDDPM: Group Size

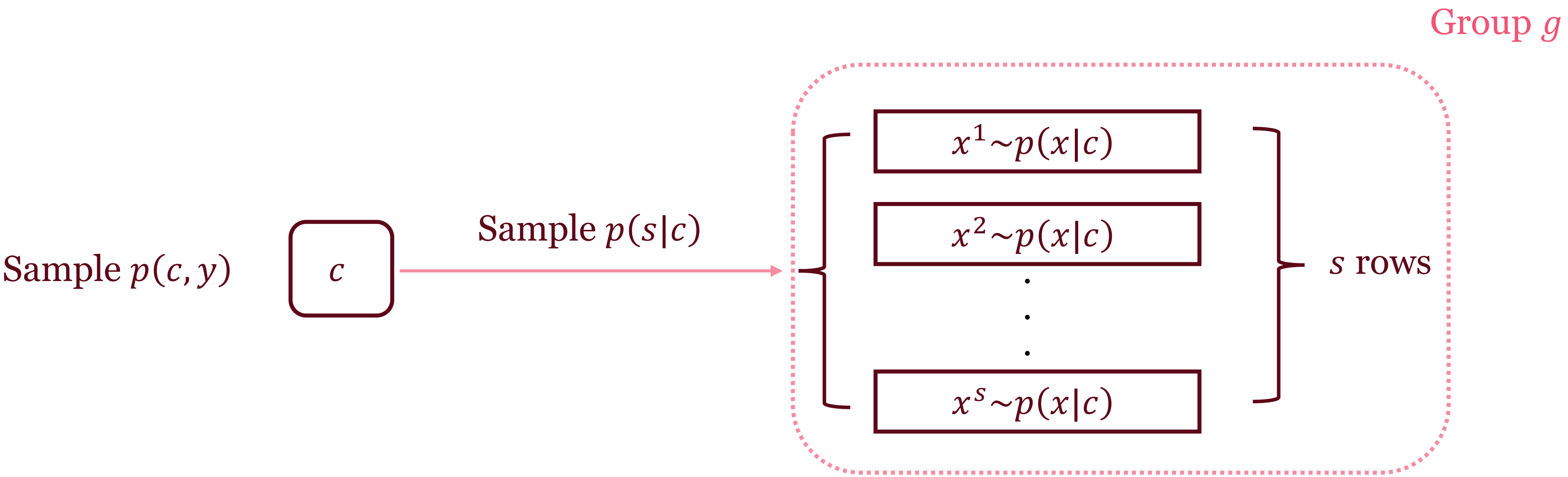


UNIVERSITY OF WATERLOO

ClavaDDPM: Group Size



ClavaDDPM: Group Size



ClavaDDPM: Wrapped Up

- Parent table R_1, data denoted Y.
- Child table R_2, data denoted X.
- Cluster latent c, group size s.

$$p(X,Y) \approx \prod_{j=1}^{|R_2|} \sum_{c} p(y_j, c)\, p(s_j|c) \prod_{i=1}^{s_j} p(x_j^i|c)$$

UNIVERSITY OF WATERLOO

ClavaDDPM: Wrapped Up

- Parent table R_1, data denoted Y.
- Child table R_2, data denoted X.
- Cluster latent c, group size s.

$$p(X,Y) \approx \prod_{j=1}^{|R_2|} \sum_{c} p(y_j, c) p(s_j | c) \prod_{i=1}^{s_j} p(x_j^i | c)$$

Diffusion model for augmented parent table

ClavaDDPM: Wrapped Up

- Parent table R_1, data denoted Y.
- Child table R_2, data denoted X.
- Cluster latent c, group size s.

$$p(X,Y) \approx \prod_{j=1}^{|R_2|} \sum_{c} p(y_j, c)\, \underbrace{p(s_j|c)}_{\text{Frequency estimation}} \prod_{i=1}^{s_j} p(x_j^i|c)$$

UNIVERSITY OF WATERLOO

ClavaDDPM: Wrapped Up

- Parent table R_1, data denoted Y.
- Child table R_2, data denoted X.
- Cluster latent c, group size s.

$$p(X,Y) \approx \prod_{j=1}^{|R_2|} \sum_{c} p(y_j, c) p(s_j|c) \prod_{i=1}^{s_j} p(x_j^i|c)$$

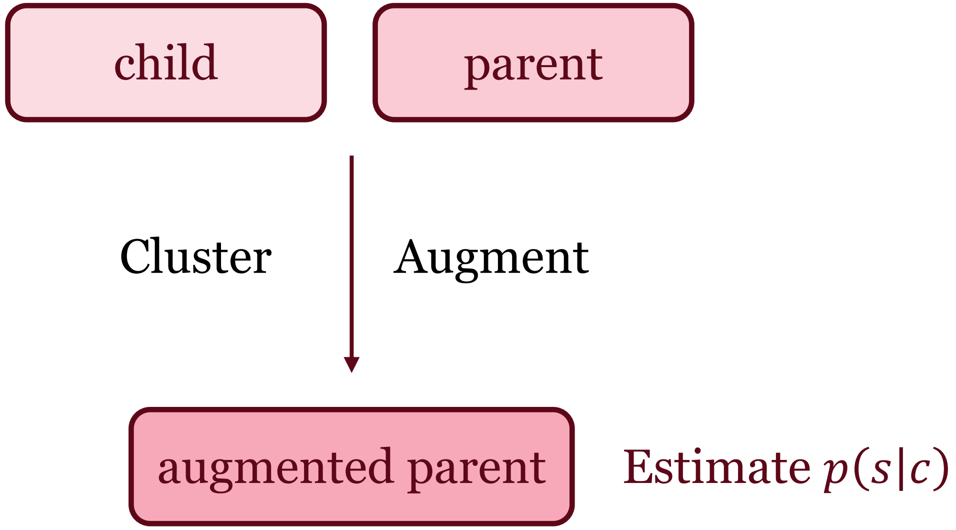
Classifier guided sampling using child diffusion model $p(x)$ and classifier $p(c|x)$

ClavaDDPM: Two Tables Training

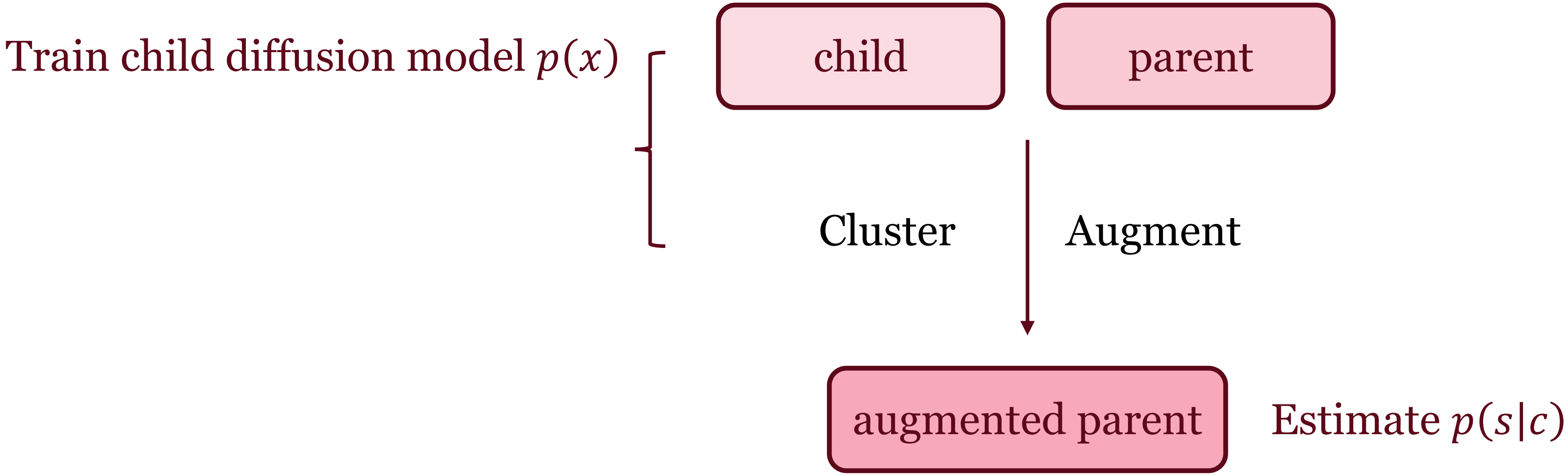
child

parent

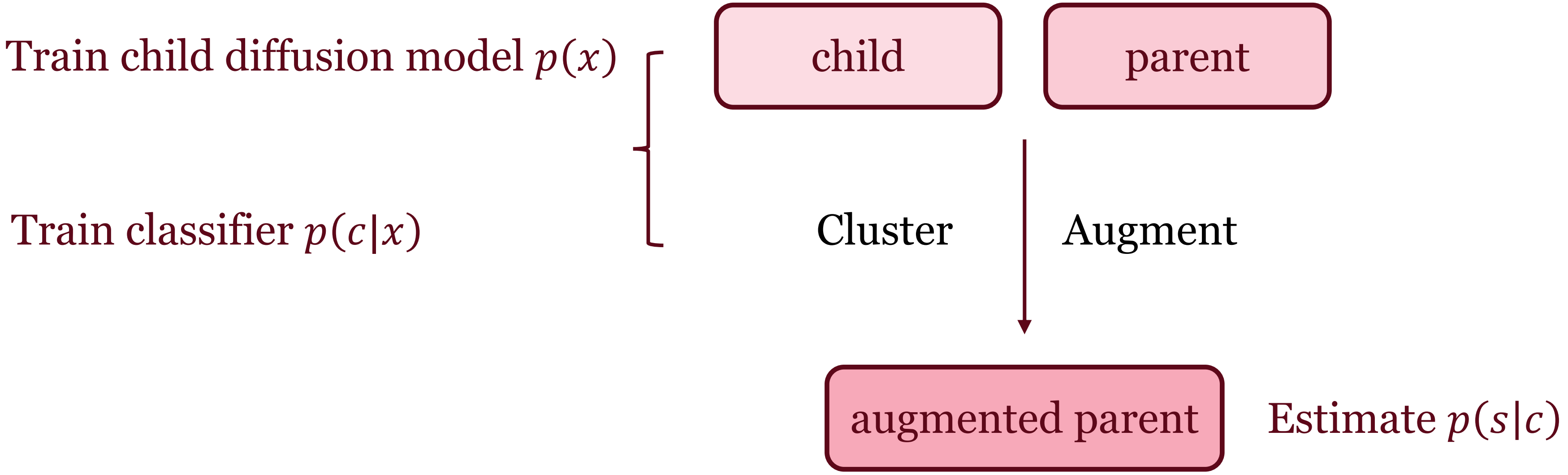
ClavaDDPM: Two Tables Training



ClavaDDPM: Two Tables Training



ClavaDDPM: Two Tables Training



ClavaDDPM: Two Tables Training

Train child diffusion model $p(x)$

Train classifier $p(c|x)$

child

parent

Cluster

Augment

augmented parent

Estimate $p(s|c)$

Train diffusion model $p(y, c)$ on augmented parent

ClavaDDPM: Two Tables Training

Train child diffusion model $p(x)$

Train classifier $p(c|x)$

child

parent

Cluster

Augment

augmented parent

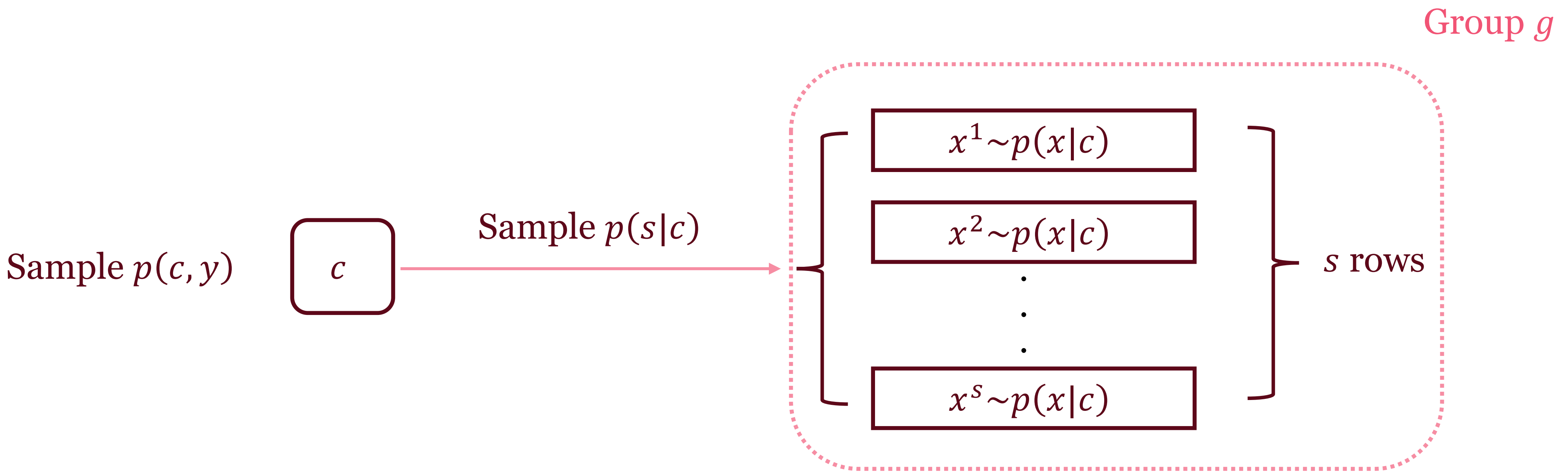
Estimate $p(s|c)$

Train diffusion model $p(y, c)$ on augmented parent

Note: the **parent** augmentation depends on **child**.

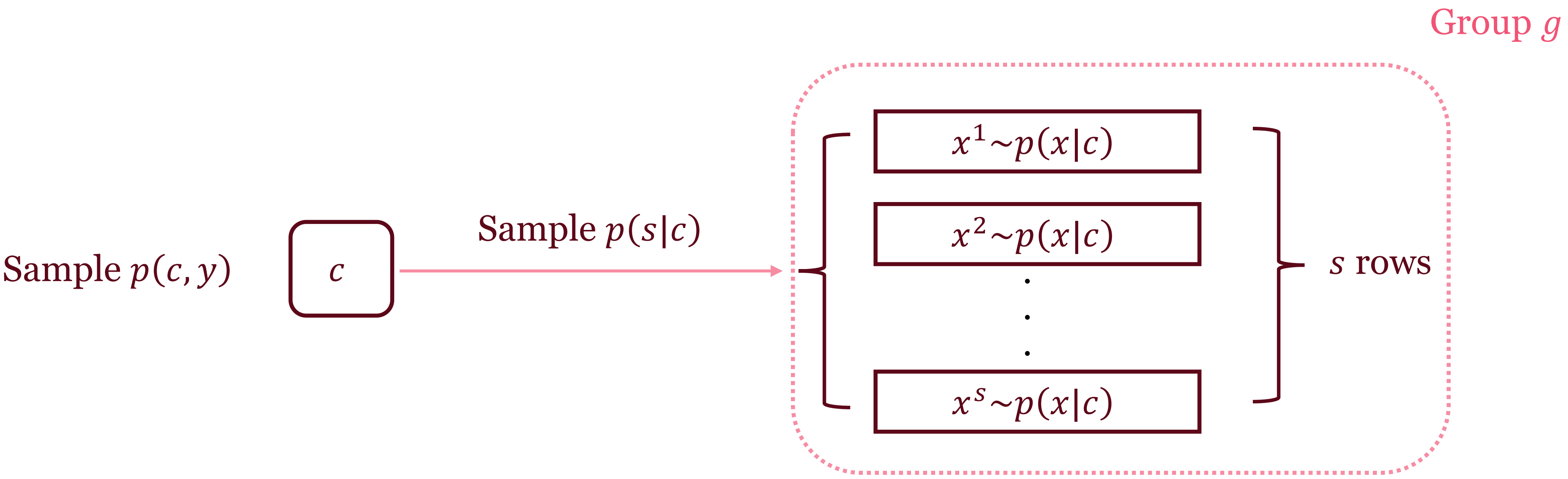
UNIVERSITY OF WATERLOO

ClavaDDPM: Two Tables Sampling



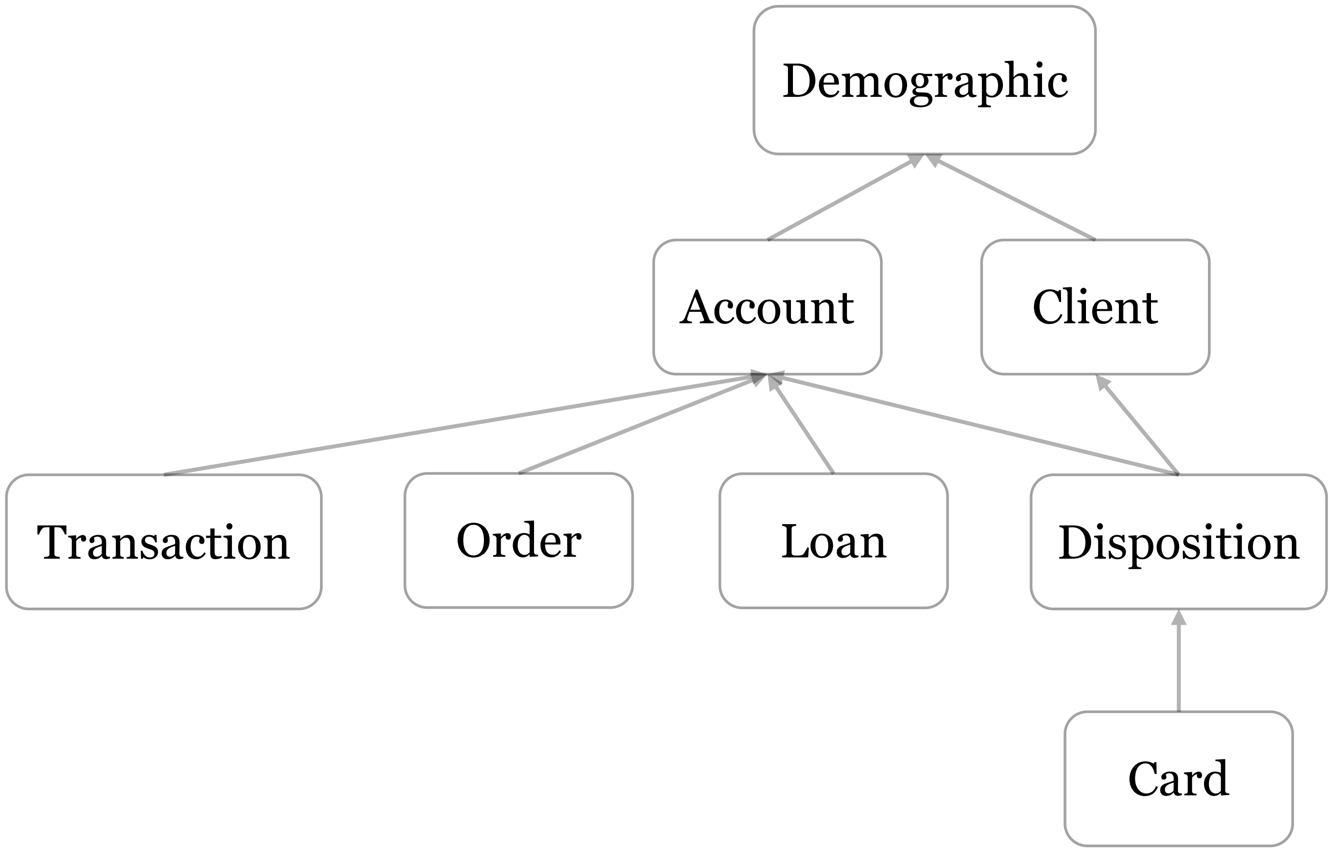
UNIVERSITY OF WATERLOO

ClavaDDPM: Two Tables Sampling



Note: the **child** sampling depends on **parent**.

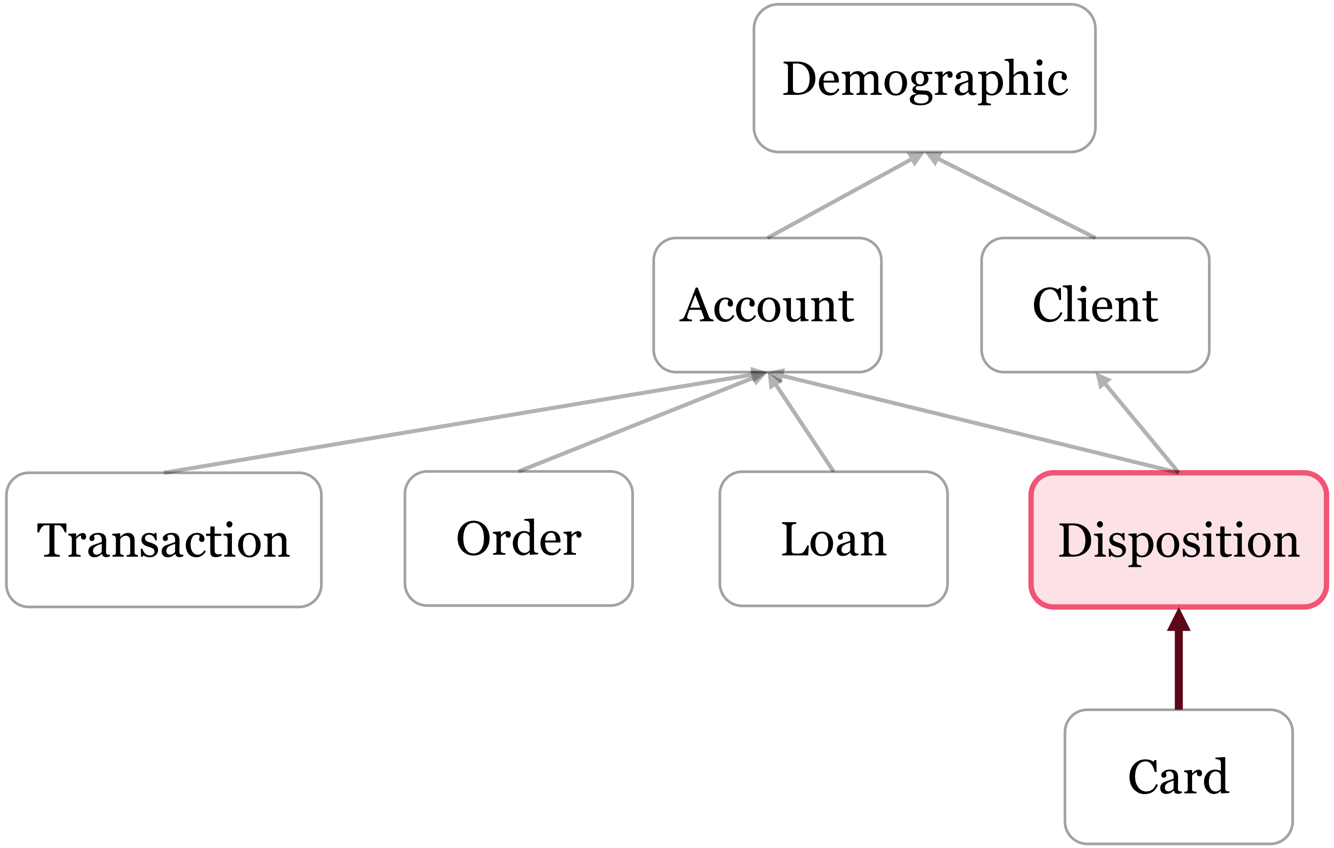
Extension to More: Training



Extension to More: Training

Cluster, augment, and train

- Parent: Disposition
- Child: Card

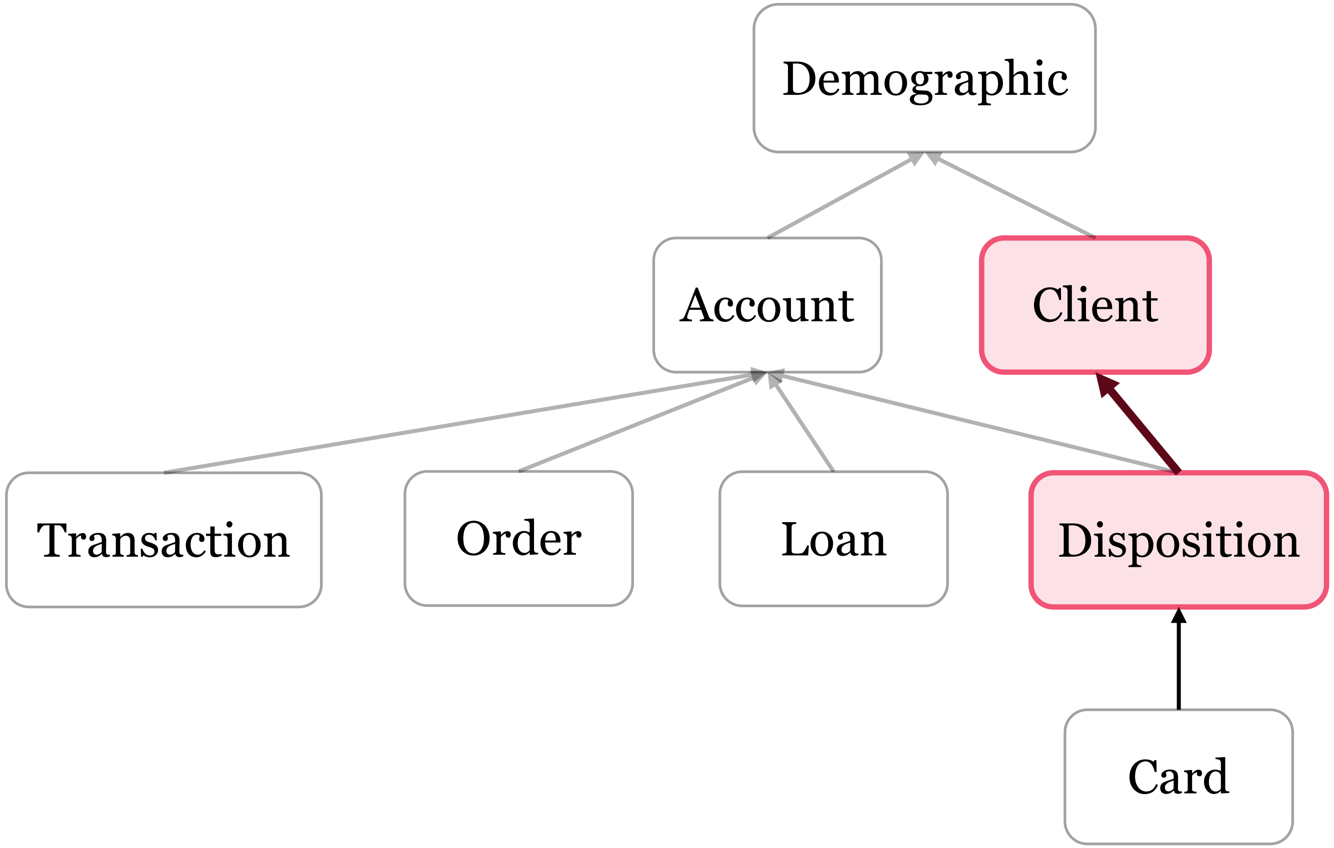


UNIVERSITY OF WATERLOO

Extension to More: Training

Cluster, augment, and train

- Parent: Client
- Child: **augmented** Disposition

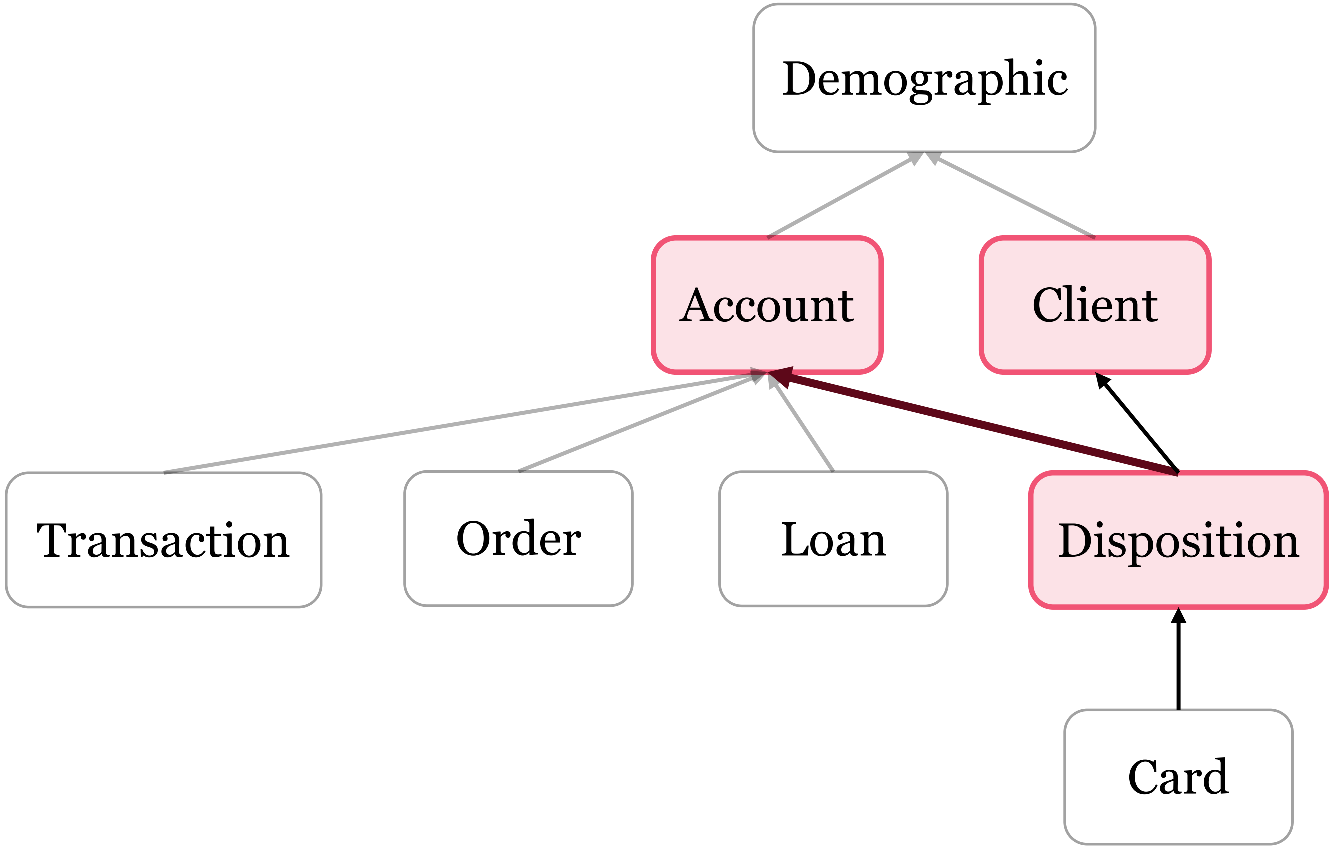


UNIVERSITY OF WATERLOO

Extension to More: Training

Cluster, augment, and train

- Parent: Account
- Child: **augmented** Disposition

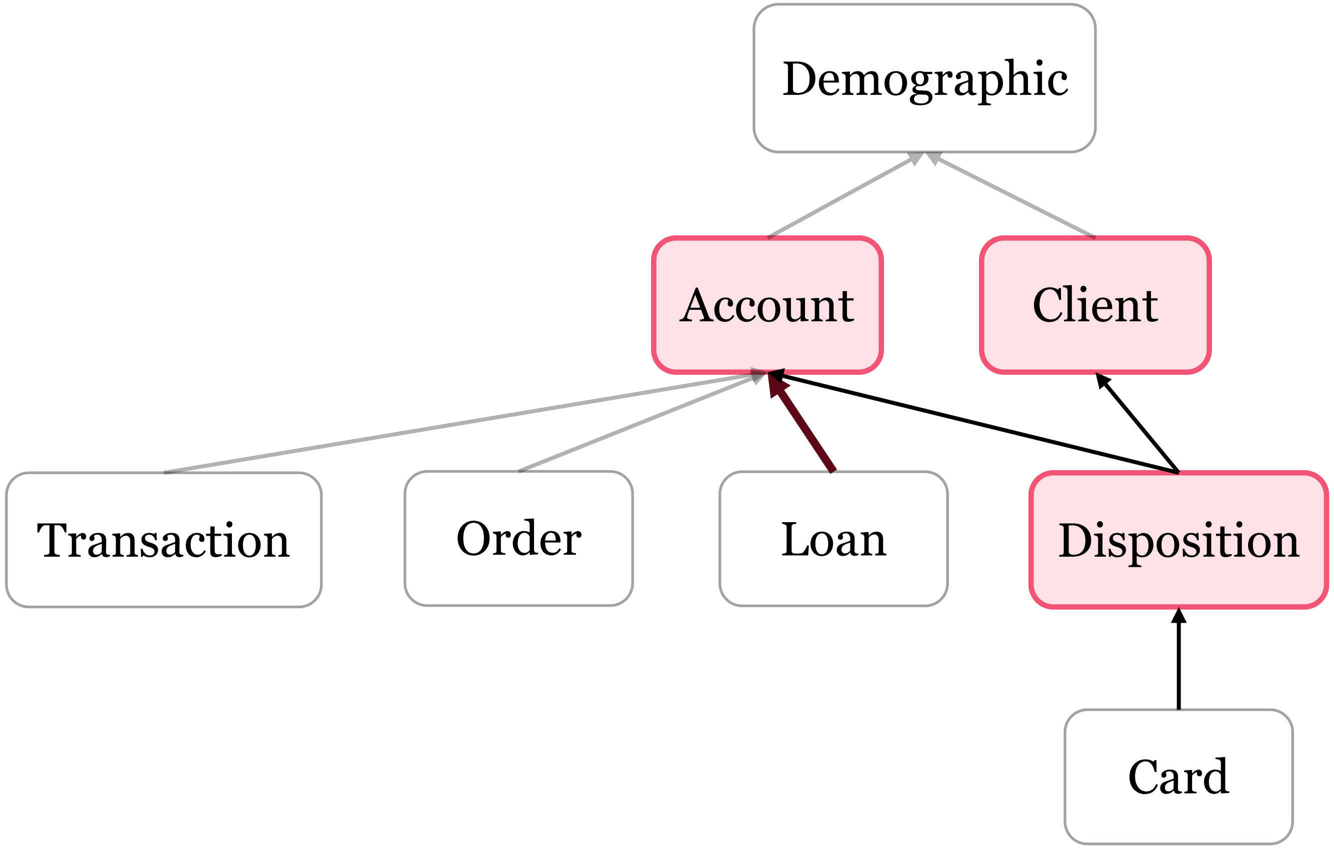


UNIVERSITY OF WATERLOO

Extension to More: Training

Cluster, augment, and train

- Parent: **augmented** Account
- Child: Loan

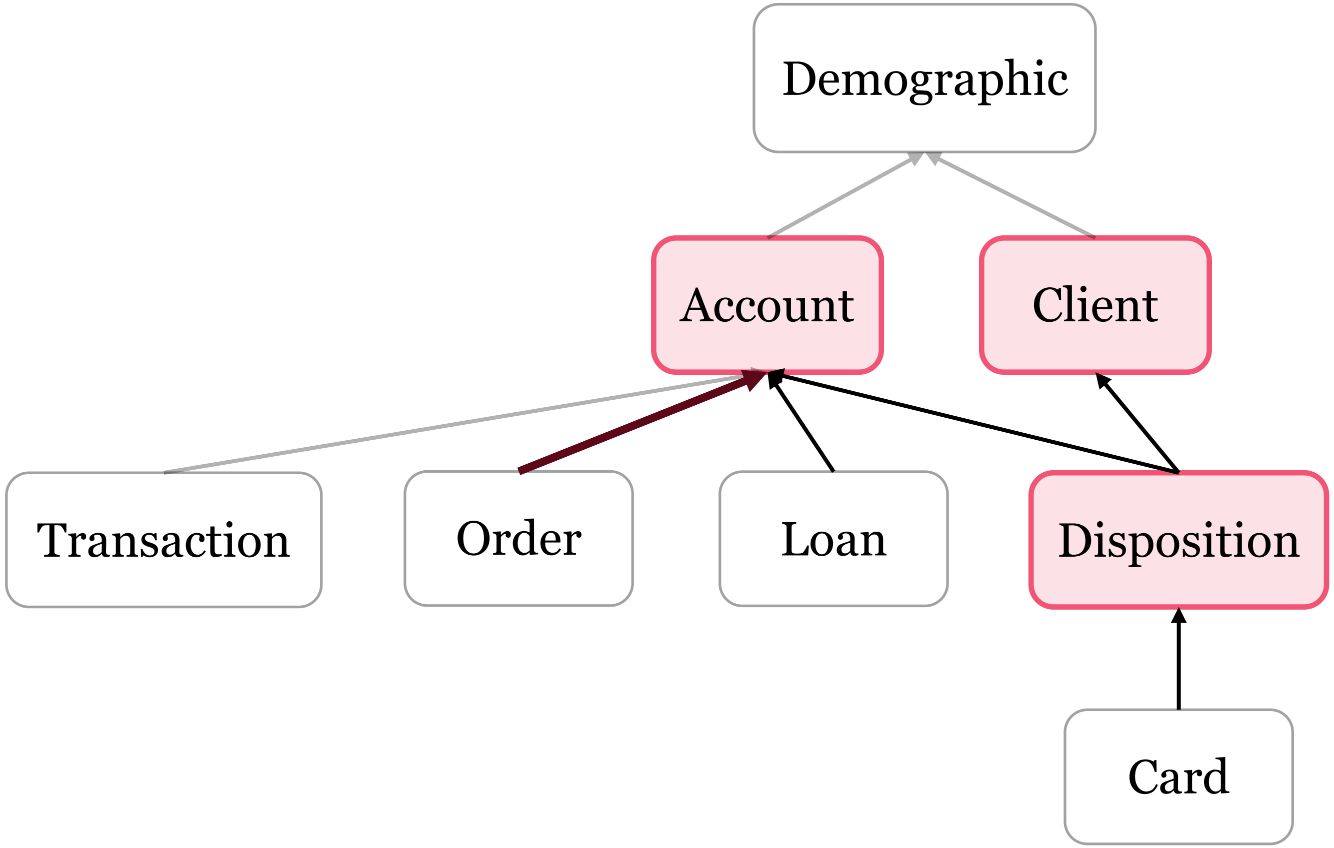


UNIVERSITY OF WATERLOO

Extension to More: Training

Cluster, augment, and train

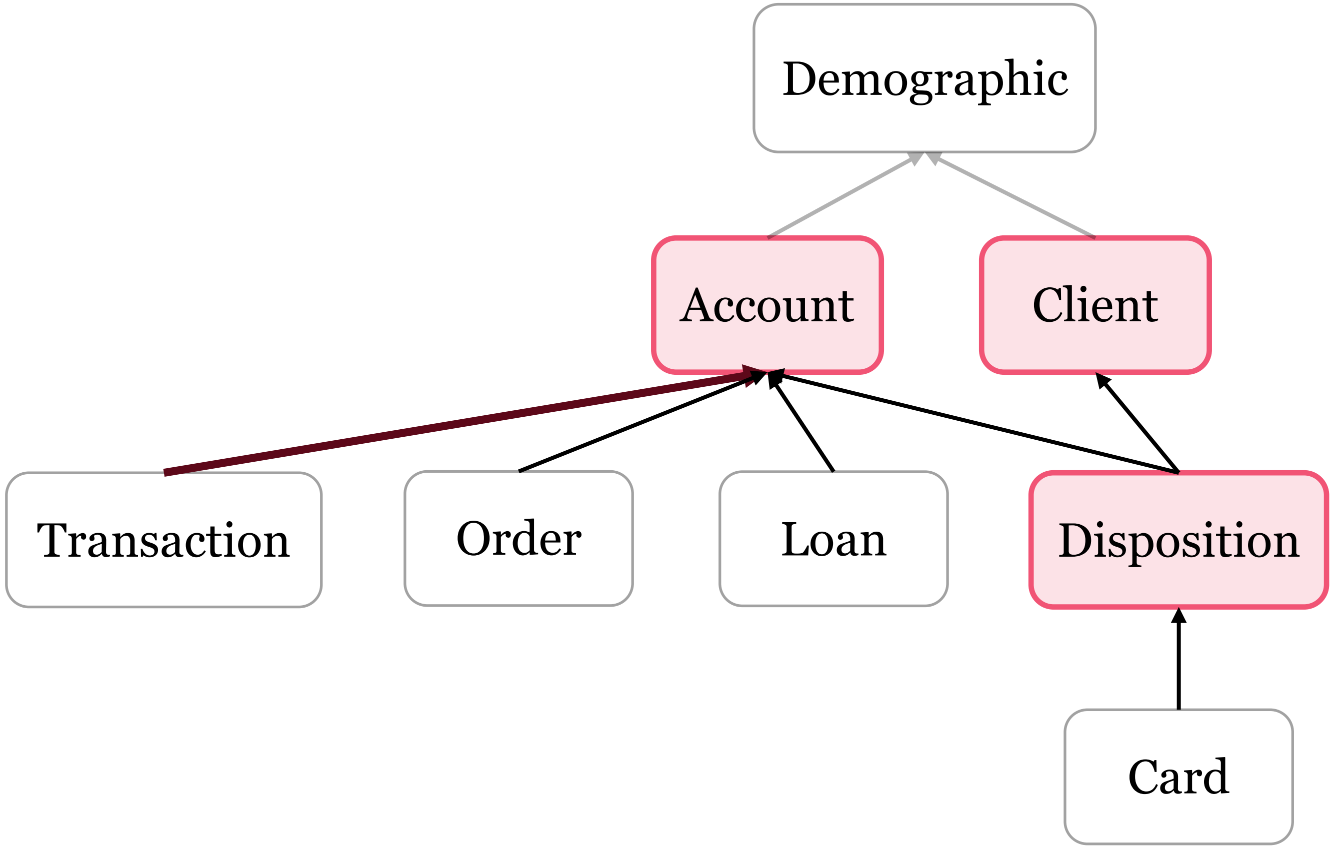
- Parent: **augmented** Account
- Child: Order



Extension to More: Training

Cluster, augment, and train

- Parent: **augmented** Account
- Child: Transaction

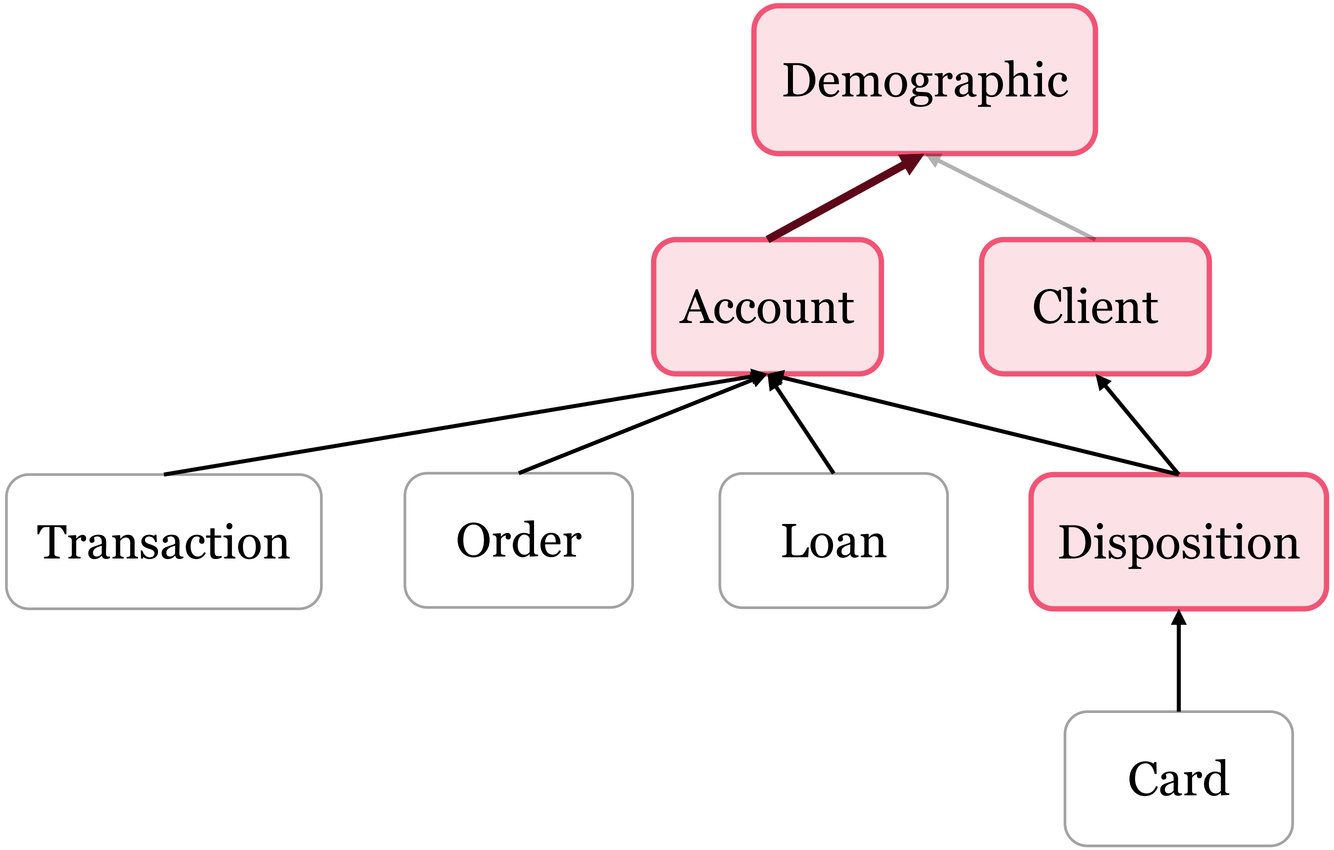


UNIVERSITY OF WATERLOO

Extension to More: Training

Cluster, augment, and train

- Parent: Demographic
- Child: **augmented** Account

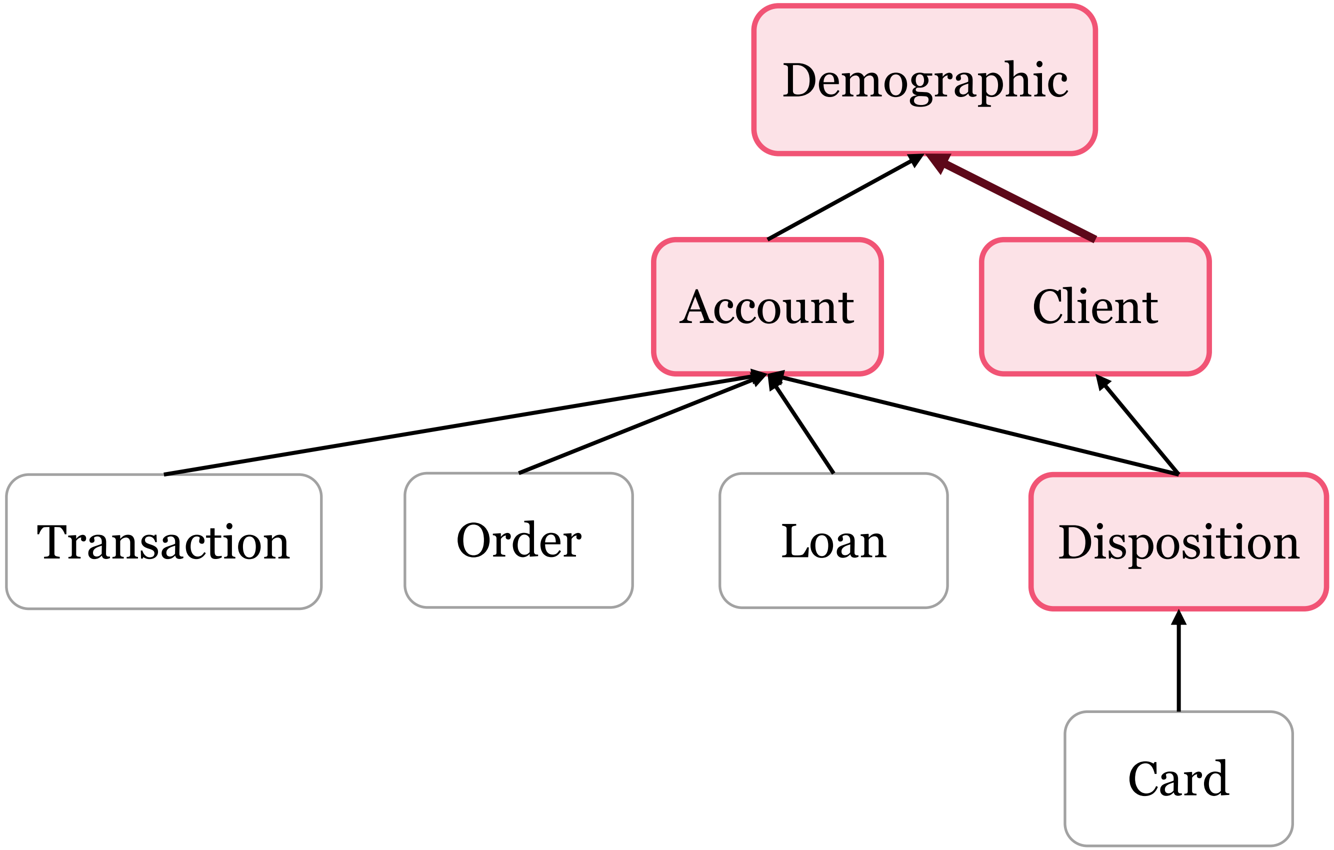


UNIVERSITY OF WATERLOO

Extension to More: Training

Cluster, augment, and train

- Parent: **augmented** Demographic
- Child: **augmented** Client



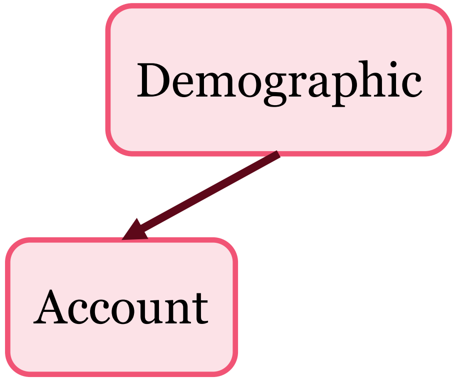
Extension to More: Synthesis

Synthesize **augmented** Demographic

Demographic

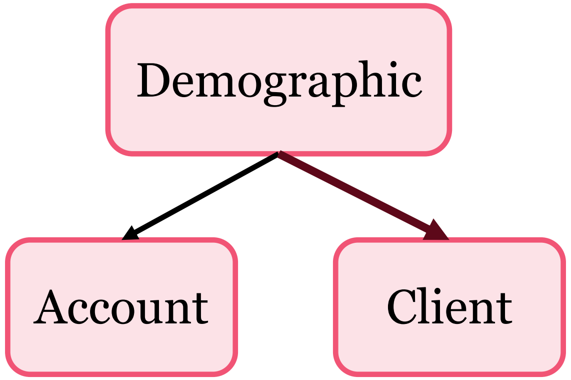
Extension to More: Synthesis

Conditioned on **augmented** Demographic
Synthesize **augmented** Demographic



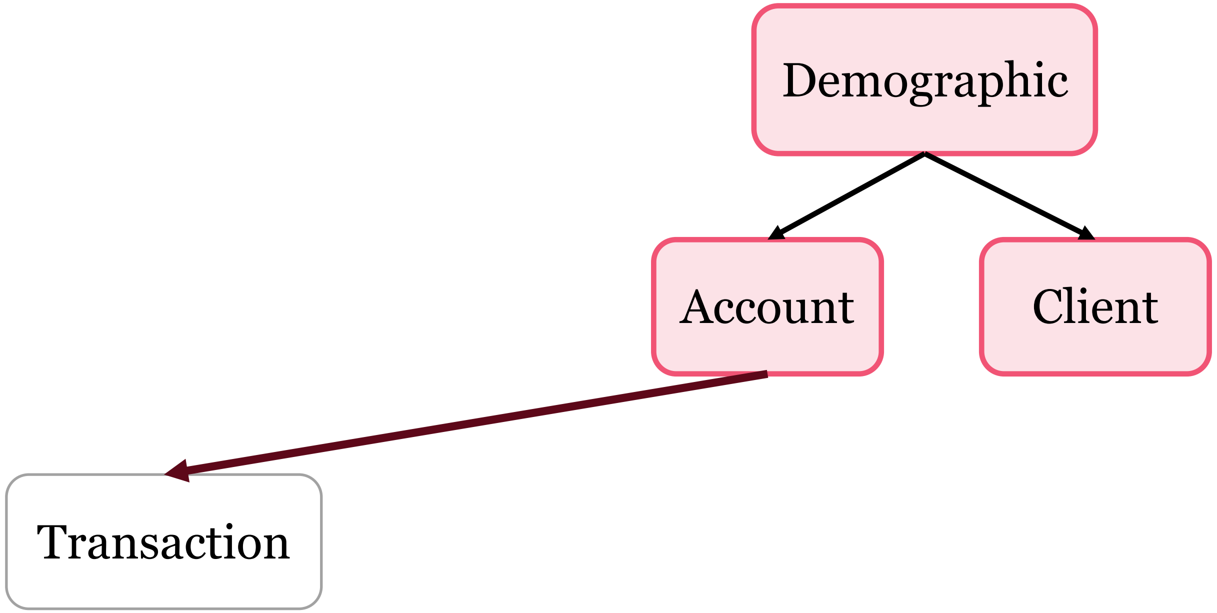
Extension to More: Synthesis

Conditioned on **augmented** Demographic
Synthesize **augmented** Client



Extension to More: Synthesis

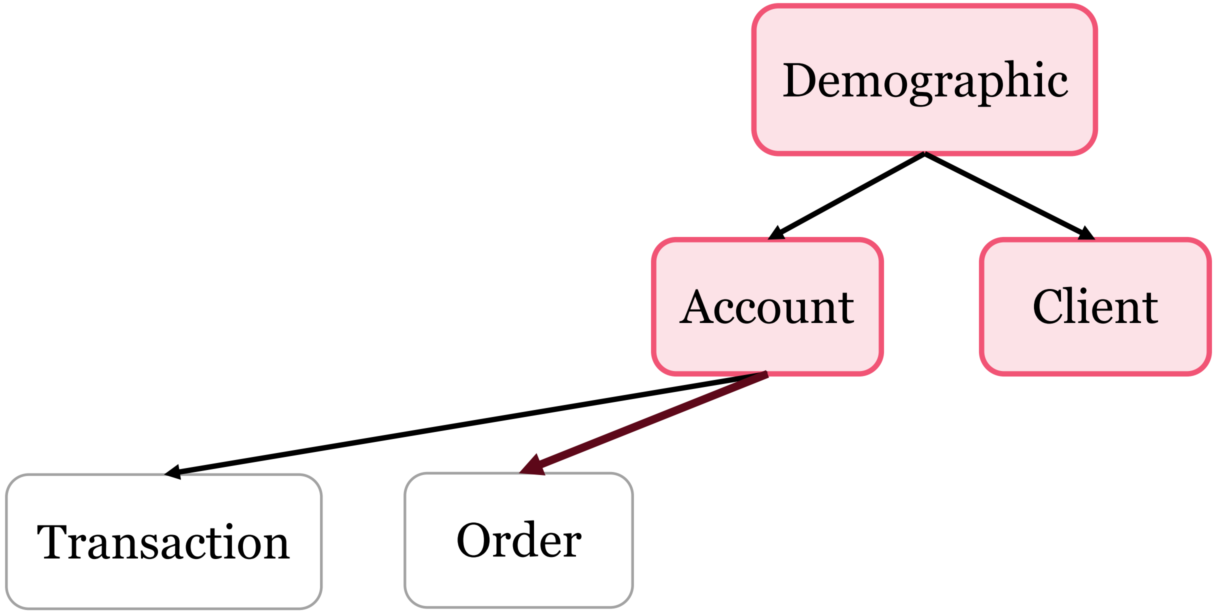
Conditioned on **augmented** Account
Synthesize Transaction



UNIVERSITY OF WATERLOO

Extension to More: Synthesis

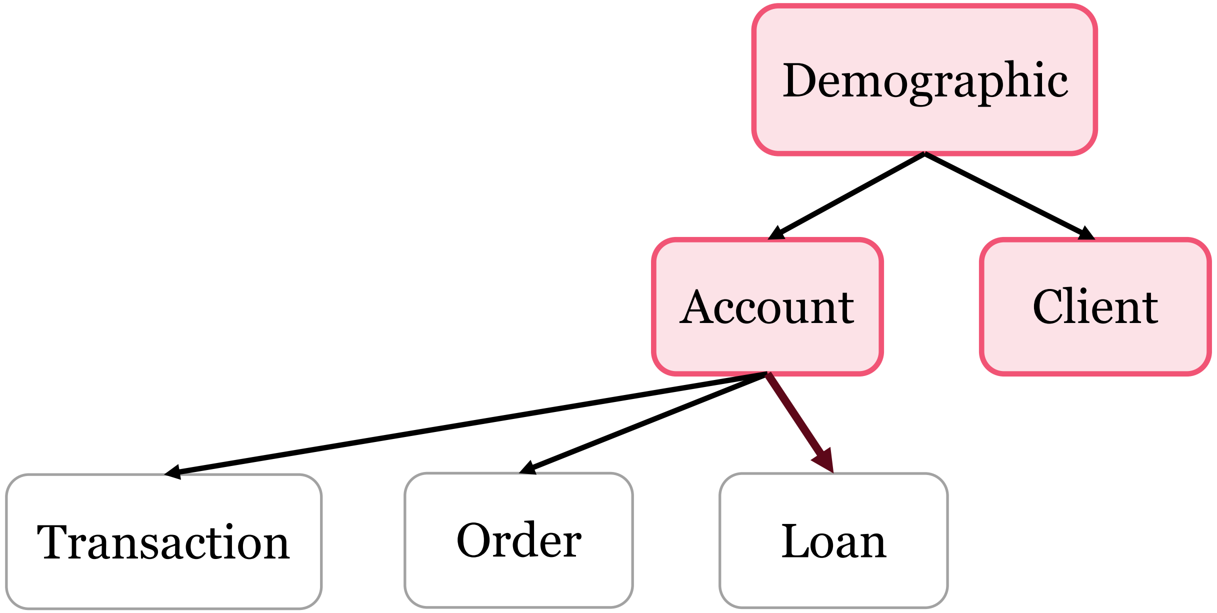
Conditioned on **augmented** Account
Synthesize Order



UNIVERSITY OF WATERLOO

Extension to More: Synthesis

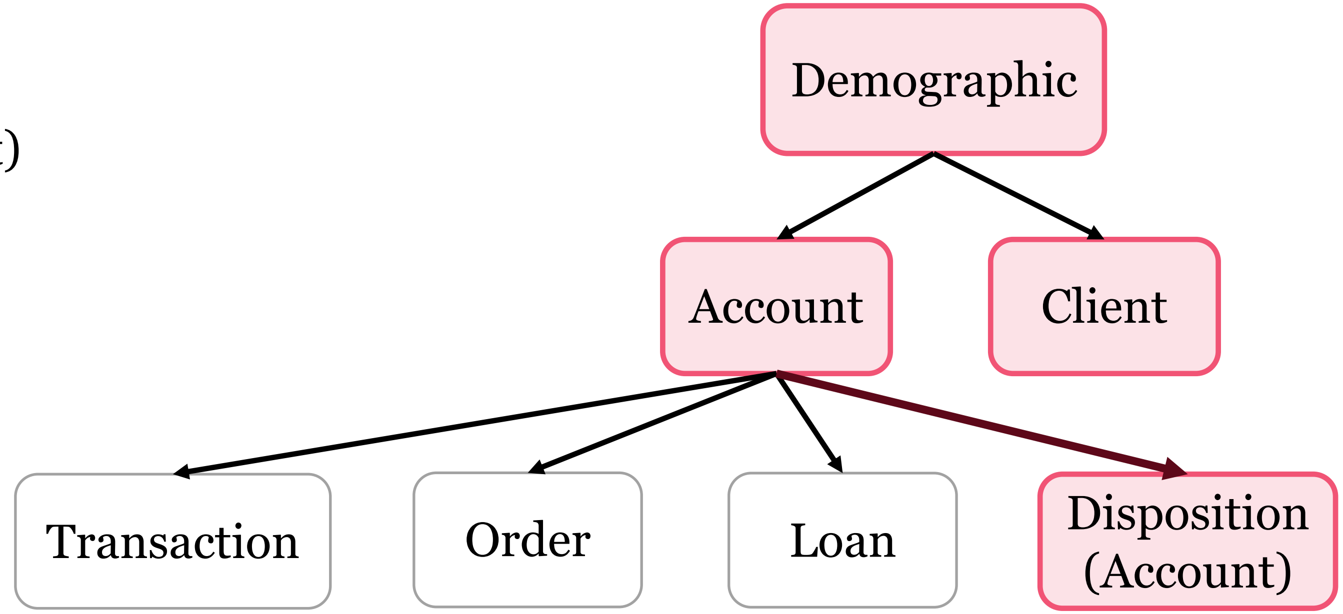
Conditioned on **augmented** Account
Synthesize Loan



UNIVERSITY OF WATERLOO

Extension to More: Synthesis

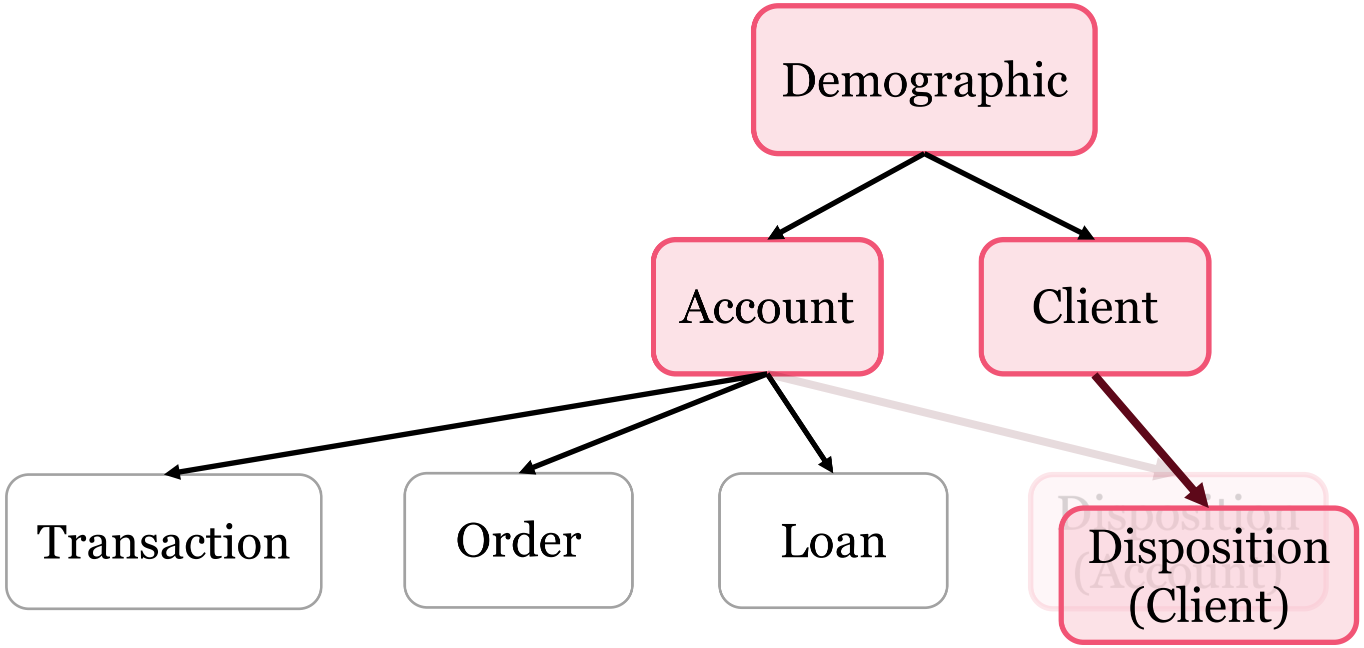
Conditioned on **augmented** Account
Synthesize **augmented** Disposition (Account)



UNIVERSITY OF WATERLOO

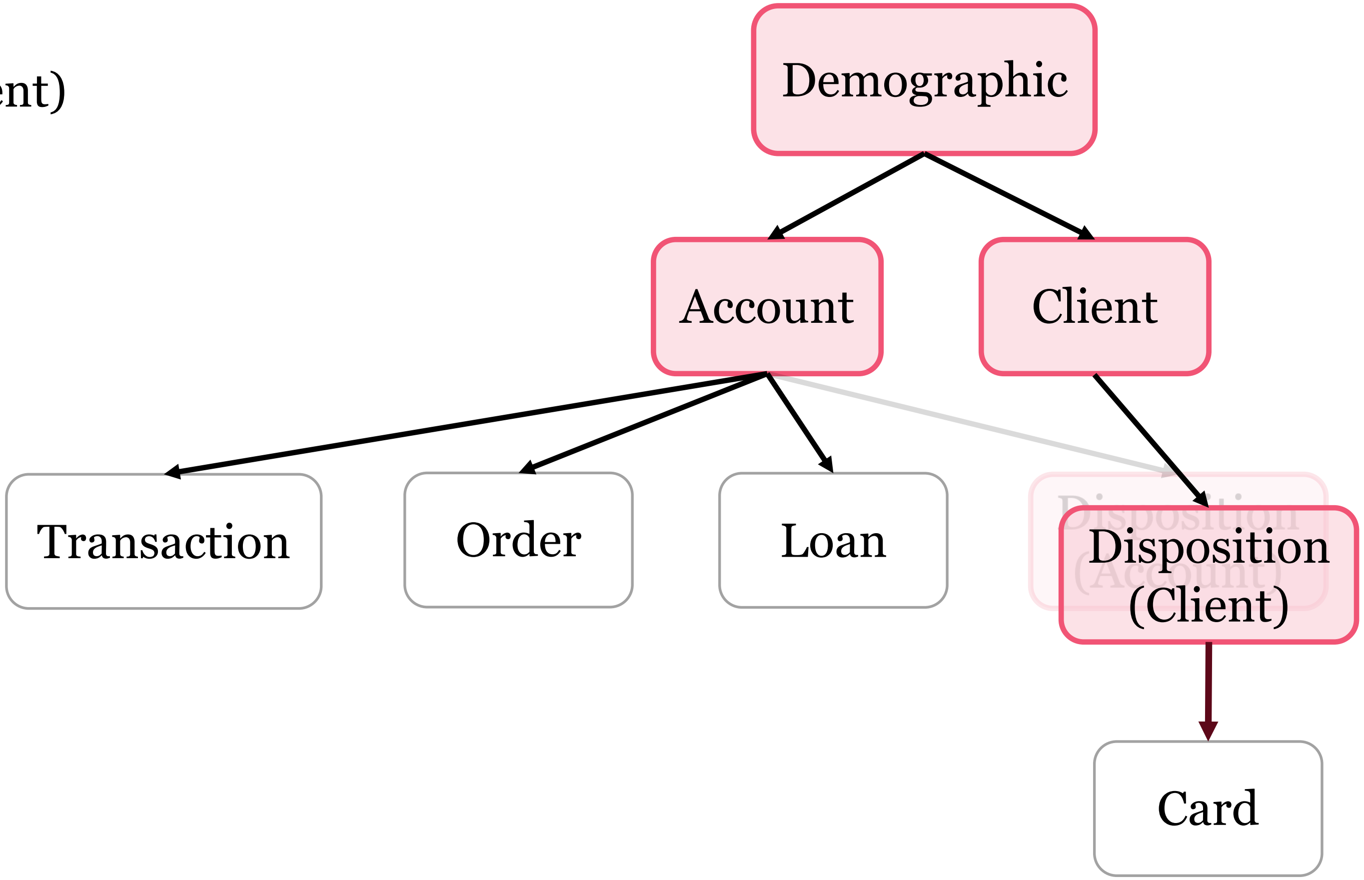
Extension to More: Synthesis

Conditioned on **augmented** Client
Synthesize **augmented** Disposition (Client)



Extension to More: Synthesis

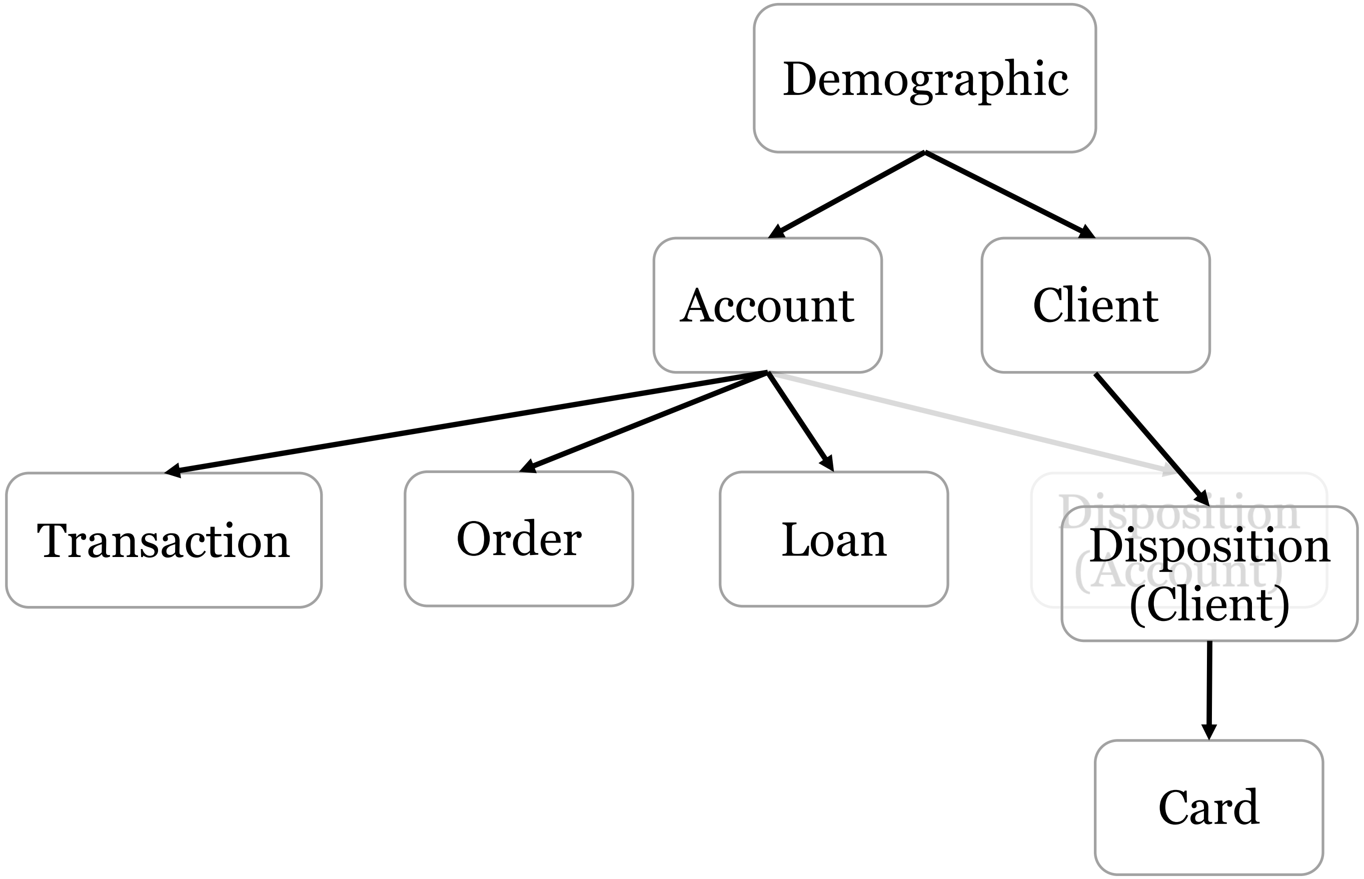
Conditioned on **augmented** Disposition (Client)
Synthesize Card



UNIVERSITY OF WATERLOO

Extension to More: Synthesis

Remove augmented columns



UNIVERSITY OF WATERLOO

Extension to More: Multi-parent Dilemma

Disposition (Client)

Disp ID	Client ID	X^c
1	2	x_1^c
2	2	x_2^c
3	1	x_3^c
4	3	x_4^c
5	3	x_5^c
6	3	x_6^c
7	4	x_7^c
8	4	x_8^c

Disposition (Account)

Disp ID	Account ID	X^a
1	2	x_1^a
2	1	x_2^a
3	3	x_3^a
4	5	x_4^a
5	5	x_5^a
6	2	x_6^a
7	2	x_7^a
8	1	x_8^a
9	3	x_9^a

Extension to More: Multi-parent Dilemma

Disposition (Client)

Disp ID	Client ID	X^c
1	2	x_1^c
2	2	x_2^c
3	1	x_3^c
4	3	x_4^c
5	3	x_5^c
6	3	x_6^c
7	4	x_7^c
8	4	x_8^c

Disposition (Account)

X^a	Disp ID	Account ID
x_1^a	1	2
x_2^a	2	1
x_3^a	3	3
x_4^a	4	5
x_5^a	5	5
x_6^a	6	2
x_7^a	7	2
x_8^a	8	1
x_9^a	9	3

Extension to More: Matching

Disposition (Client)

Disp ID	Client ID	X^c
1	2	x_1^c
2	2	x_2^c
3	1	x_3^c
4	3	x_4^c
5	3	x_5^c
6	3	x_6^c
7	4	x_7^c
8	4	x_8^c

Disposition (Account)

X^a	Disp ID	Account ID
x_1^a	1	2
x_2^a	2	1
x_3^a	3	3
x_4^a	4	5
x_5^a	5	5
x_6^a	6	2
x_7^a	7	2
x_8^a	8	1
x_9^a	9	3

$\Rightarrow$

Disposition

Disp ID	Client ID	Account ID	X
1	2		
2	2		
3	1		
4	3		
5	3		
6	3		
7	4		
8	4		

Extension to More: Matching

Disposition (Client)

Disp ID	Client ID	X^c
1	2	x_1^c
2	2	x_2^c
3	1	x_3^c
4	3	x_4^c
5	3	x_5^c
6	3	x_6^c
7	4	x_7^c
8	4	x_8^c

Disposition (Account)

X^a	Disp ID	Account ID
x_1^a	1	2
x_2^a	2	1
x_3^a	3	3
x_4^a	4	5
x_5^a	5	5
x_6^a	6	2
x_7^a	7	2
x_8^a	8	1
x_9^a	9	3

$\Rightarrow$

Disposition

Disp ID	Client ID	Account ID	X
1	2	2	(x_1^c, x_1^a)
2	2		
3	1		
4	3		
5	3		
6	3		
7	4		
8	4		

Extension to More: Matching

Disposition (Client)

Disp ID	Client ID	X^c
1	2	x_1^c
2	2	x_2^c
3	1	x_3^c
4	3	x_4^c
5	3	x_5^c
6	3	x_6^c
7	4	x_7^c
8	4	x_8^c

Disposition (Account)

X^a	Disp ID	Account ID
x_1^a	1	2
x_2^a	2	1
x_3^a	3	3
x_4^a	4	5
x_5^a	5	5
x_6^a	6	2
x_7^a	7	2
x_8^a	8	1
x_9^a	9	3

Disposition

Disp ID	Client ID	Account ID	X
1	2	2	(x_1^c, x_1^a)
2	2	3	(x_2^c, x_3^a)
3	1		
4	3		
5	3		
6	3		
7	4		
8	4		

Extension to More: Matching

Disposition (Client)

Disp ID	Client ID	X^c
1	2	x_1^c
2	2	x_2^c
3	1	x_3^c
4	3	x_4^c
5	3	x_5^c
6	3	x_6^c
7	4	x_7^c
8	4	x_8^c

Disposition (Account)

X^a	Disp ID	Account ID
x_1^a	1	2
x_2^a	2	1
x_3^a	3	3
x_4^a	4	5
x_5^a	5	5
x_6^a	6	2
x_7^a	7	2
x_8^a	8	1
x_9^a	9	3

$\Rightarrow$

Disposition

Disp ID	Client ID	Account ID	X
1	2	2	(x_1^c, x_1^a)
2	2	3	(x_2^c, x_3^a)
3	1	5	(x_3^c, x_4^a)
4	3		
5	3		
6	3		
7	4		
8	4		

UNIVERSITY OF WATERLOO

Extension to More: Matching

Disposition (Client)

Disp ID	Client ID	X^c
1	2	x_1^c
2	2	x_2^c
3	1	x_3^c
4	3	x_4^c
5	3	x_5^c
6	3	x_6^c
7	4	x_7^c
8	4	x_8^c

Disposition (Account)

X^a	Disp ID	Account ID
x_1^a	1	2
x_2^a	2	1
x_3^a	3	3
x_4^a	4	5
x_5^a	5	5
x_6^a	6	2
x_7^a	7	2
x_8^a	8	1
x_9^a	9	3

Disposition

Disp ID	Client ID	Account ID	X
1	2	2	(x_1^c, x_1^a)
2	2	3	(x_2^c, x_3^a)
3	1	5	(x_3^c, x_4^a)
4	3	2	(x_4^c, x_7^a)
5	3		
6	3		
7	4		
8	4		

Extension to More: Matching

Disposition (Client)

Disp ID	Client ID	X^c
1	2	x_1^c
2	2	x_2^c
3	1	x_3^c
4	3	x_4^c
5	3	x_5^c
6	3	x_6^c
7	4	x_7^c
8	4	x_8^c

Disposition (Account)

X^a	Disp ID	Account ID
x_1^a	1	2
x_2^a	2	1
x_3^a	3	3
x_4^a	4	5
x_5^a	5	5
x_6^a	6	2
x_7^a	7	2
x_8^a	8	1
x_9^a	9	3

$\Rightarrow$

Disposition

Disp ID	Client ID	Account ID	X
1	2	2	(x_1^c, x_1^a)
2	2	3	(x_2^c, x_3^a)
3	1	5	(x_3^c, x_4^a)
4	3	2	(x_4^c, x_7^a)
5	3	1	(x_5^c, x_2^a)
6	3		
7	4		
8	4		

Extension to More: Matching

Disposition (Client)

Disp ID	Client ID	X^c
1	2	x_1^c
2	2	x_2^c
3	1	x_3^c
4	3	x_4^c
5	3	x_5^c
6	3	x_6^c
7	4	x_7^c
8	4	x_8^c

Disposition (Account)

X^a	Disp ID	Account ID
x_1^a	1	2
x_2^a	2	1
x_3^a	3	3
x_4^a	4	5
x_5^a	5	5
x_6^a	6	2
x_7^a	7	2
x_8^a	8	1
x_9^a	9	3

$\Rightarrow$

Disposition

Disp ID	Client ID	Account ID	X
1	2	2	(x_1^c, x_1^a)
2	2	3	(x_2^c, x_3^a)
3	1	5	(x_3^c, x_4^a)
4	3	2	(x_4^c, x_7^a)
5	3	1	(x_5^c, x_2^a)
6	3	5	(x_6^c, x_5^a)
7	4		
8	4		

Extension to More: Matching

Disposition (Client)

Disp ID	Client ID	X^c
1	2	x_1^c
2	2	x_2^c
3	1	x_3^c
4	3	x_4^c
5	3	x_5^c
6	3	x_6^c
7	4	x_7^c
8	4	x_8^c

Disposition (Account)

X^a	Disp ID	Account ID
x_1^a	1	2
x_2^a	2	1
x_3^a	3	3
x_4^a	4	5
x_5^a	5	5
x_6^a	6	2
x_7^a	7	2
x_8^a	8	1
x_9^a	9	3

$\Rightarrow$

Disposition

Disp ID	Client ID	Account ID	X
1	2	2	(x_1^c, x_1^a)
2	2	3	(x_2^c, x_3^a)
3	1	5	(x_3^c, x_4^a)
4	3	2	(x_4^c, x_7^a)
5	3	1	(x_5^c, x_2^a)
6	3	5	(x_6^c, x_5^a)
7	4	1	(x_7^c, x_8^a)
8	4		

Extension to More: Matching

Disposition (Client)

Disp ID	Client ID	X^c
1	2	x_1^c
2	2	x_2^c
3	1	x_3^c
4	3	x_4^c
5	3	x_5^c
6	3	x_6^c
7	4	x_7^c
8	4	x_8^c

Disposition (Account)

X^a	Disp ID	Account ID
x_1^a	1	2
x_2^a	2	1
x_3^a	3	3
x_4^a	4	5
x_5^a	5	5
x_6^a	6	2
x_7^a	7	2
x_8^a	8	1
x_9^a	9	3

$\Rightarrow$

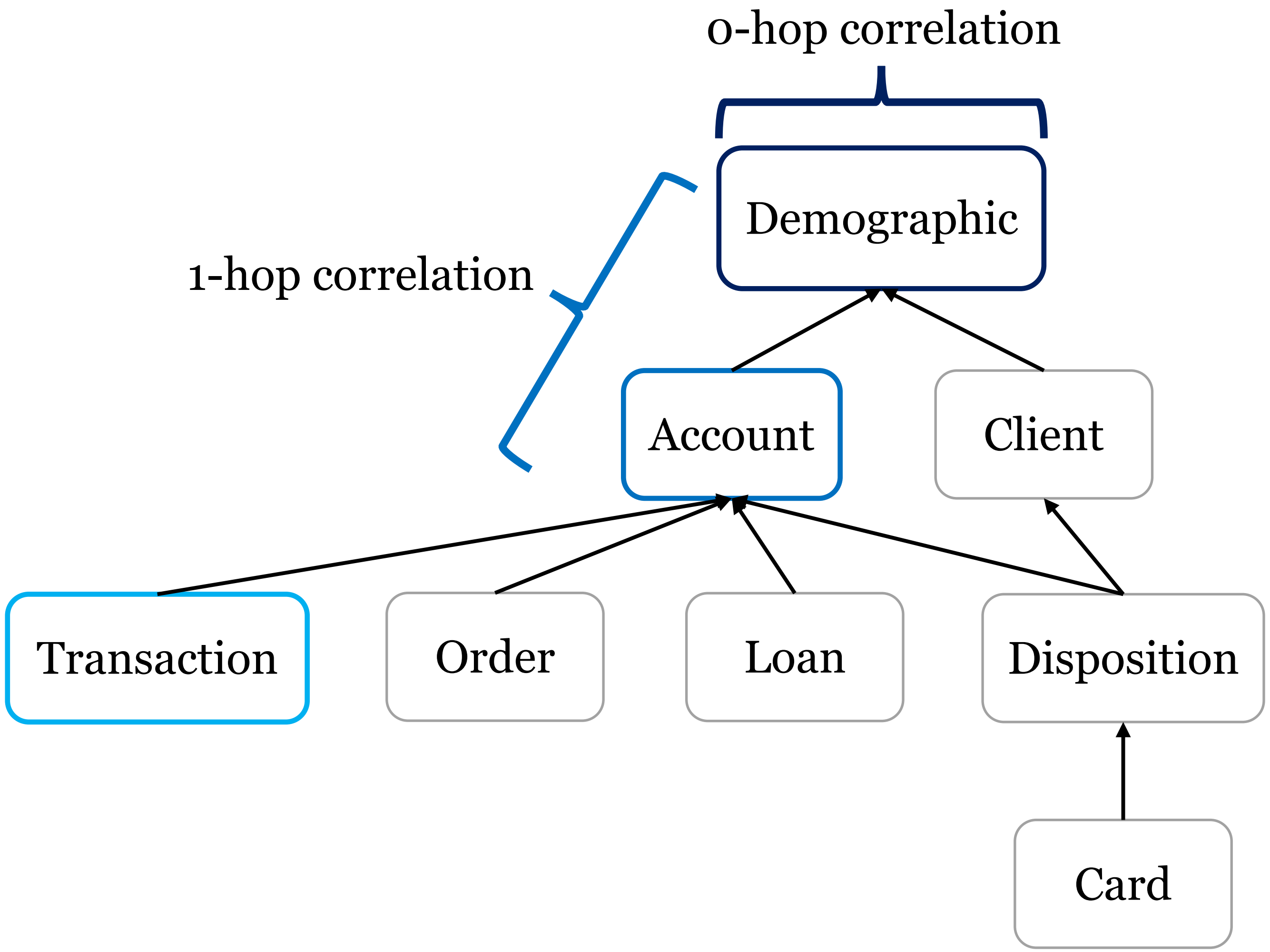
Disposition

Disp ID	Client ID	Account ID	X
1	2	2	(x_1^c, x_1^a)
2	2	3	(x_2^c, x_3^a)
3	1	5	(x_3^c, x_4^a)
4	3	2	(x_4^c, x_7^a)
5	3	1	(x_5^c, x_2^a)
6	3	5	(x_6^c, x_5^a)
7	4	1	(x_7^c, x_8^a)
8	4	2	(x_8^c, x_6^a)

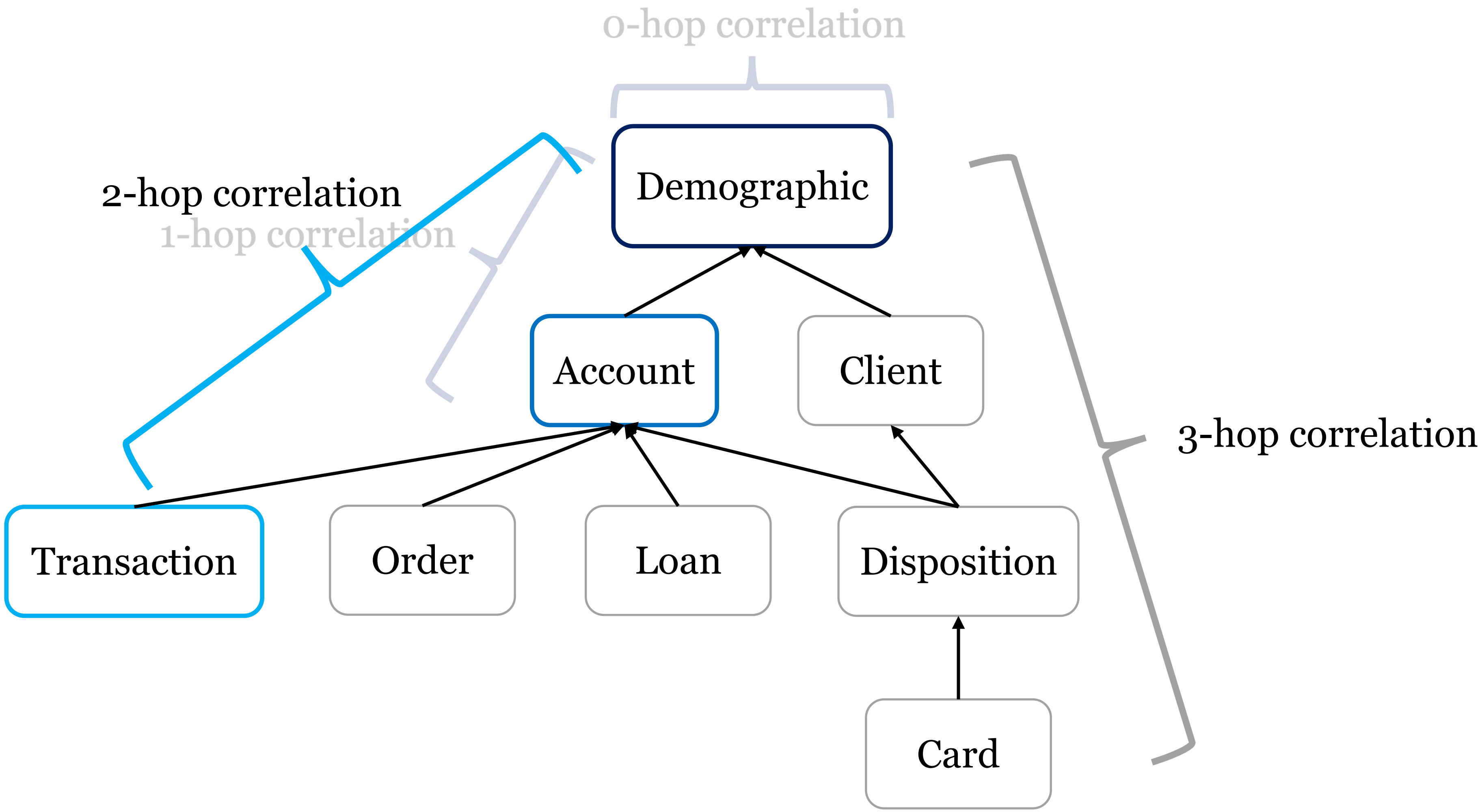
Evaluation: Metrics

- Kolmogorov-Sirnov Test (KST): measures the distance between two continuous distributions.
- Total Variation Distance (TVD): measures the distance between two discrete distributions.
- Pearson Correlation Coefficient: measures the correlation between two continuous distributions.
- Contingency Similarity: measures the distance between two discrete joint distributions.

Evaluation: Long-range Dependency



Evaluation: Long-range Dependency



Evaluation: Datasets

	# Tables	# Foreign Key Constraints	Depth	Total # of Attributes	# Rows in Largest Table
California	2	1	2	15	1,690,642
Instacart 05	6	6	3	12	1,616,315
Berka	8	8	4	41	1,056,320
Movie Lens	7	6	2	14	996,159
CCS	5	4	2	11	383,282

Evaluation: Baselines

- SDV HMA Synthesizer
- PrivLava $\varepsilon = 50$
- Single Table (ST): each table is synthesized independently.
- Denorm (D): synthesizes the joint table, then split into separate tables.
- Single table synthesis backbones:
  - CTGAN
  - TabDDPM
  - ClavaDDPM

Evaluation: Results

Berka	PrivLava	SDV	ST-CTGAN	ST-TabDDPM	ST-ClavaDDPM	D-CTGAN	D-TabDDPM	D-ClavaDDPM	ClavaDDPM
Cardinality	DNC	DNC	96.08 ±0.18	68.29±0.00	97.06±0.80	97.72±0.29	97.71±0.00	96.06±1.15	96.92±0.71
1-way			79.78±0.75	76.41±2.21	94.58±0.01	83.00±0.65	80.09±0.68	83.28±0.97	94.29±0.44
0-hop			74.24±0.32	72.80±1.23	91.72±0.23	76.04±0.34	74.82±0.49	72.12±0.73	91.49±0.82
1-hop			66.59±0.54	54.01±2.35	81.77±1.19	75.25±0.55	61.99±2.10	55.77±2.80	86.86±2.74
2-hop			75.83±1.07	59.88±1.39	78.09±0.53	72.40±0.43	63.94±1.33	57.68±1.67	89.25±2.27
3-hop			72.58±0.86	55.29±1.58	75.56±0.34	71.74±0.69	62.67±2.26	55.59±1.48	87.27±1.92
AVG 2-way			73.22±0.45	61.74±1.57	82.33±0.40	73.94±0.37	66.29±1.30	60.93±1.49	89.21±1.95

Evaluation: Results

End-to-end	PrivLava	SDV	ST-CTGAN	ST-TabDDPM	ST-ClavaDDPM	D-CTGAN	D-TabDDPM	D-ClavaDDPM	ClavaDDPM
California									
CARDINALITY	99.90 ±0.03	71.45 ±0.00	99.93 ±0.02	99.94 ±0.00	99.89 ±0.04	99.90 ±0.07	99.94 ±0.00	99.87 ±0.02	99.19 ±0.29
1-WAY	99.71 ±0.02	72.32 ±0.00	91.59 ±0.50	83.27 ±0.07	99.51 ±0.04	91.22 ±0.07	93.10 ±0.84	94.99 ±0.02	98.77 ±0.02
0-HOP	98.49 ±0.05	50.23 ±0.00	87.67 ±0.63	79.27 ±0.08	98.69 ±0.08	86.58 ±0.44	91.12 ±1.35	94.17 ±0.01	97.65 ±0.05
1-HOP	97.46 ±0.12	54.89 ±0.00	84.82 ±0.61	78.44 ±0.04	92.96 ±0.05	82.72 ±0.30	84.43 ±1.80	87.24 ±0.10	95.16 ±0.39
AVG 2-WAY	97.97 ±0.09	52.56 ±0.00	86.25 ±0.60	78.85 ±0.06	95.83 ±0.07	84.65 ±0.35	87.78 ±1.57	90.71 ±0.04	96.41 ±0.20
Instacart 05									
CARDINALITY			95.78 ±0.96		94.73 ±0.14	93.81 ±0.39		94.98 ±0.84	95.30 ±0.79
1-WAY			79.85 ±0.96		89.30 ±0.00	69.07 ±0.57		71.83 ±0.32	89.84 ±0.29
0-HOP	DNC	DNC	78.27 ±0.28	TLE	99.70 ±0.00	84.85 ±0.44	TLE	88.74 ±0.00	99.62 ±0.04
1-HOP			62.48 ±0.16		66.93 ±0.07	60.26 ±0.38		62.58 ±0.05	76.42 ±0.39
2-HOP			24.82 ±8.02		16.22 ±13.41	0.00 ±0.00		0.00 ±0.00	39.29 ±3.38
AVG 2-WAY			60.05 ±1.40		66.66 ±2.37	56.19 ±0.33		58.52 ±0.03	76.02 ±0.78
Berka									
CARDINALITY			96.08 ±0.18	68.29 ±0.00	97.06 ±0.80	97.72 ±0.29	97.71 ±0.00	96.06 ±1.15	96.92 ±0.71
1-WAY			79.78 ±0.75	76.41 ±2.21	94.58 ±0.01	83.00 ±0.65	80.09 ±0.68	83.28 ±0.97	94.29 ±0.44
0-HOP			74.24 ±0.32	72.80 ±1.23	91.72 ±0.23	76.04 ±0.34	74.82 ±0.49	72.12 ±0.73	91.49 ±0.82
1-HOP	DNC	DNC	66.59 ±0.54	54.01 ±2.35	81.77 ±1.19	75.25 ±0.55	61.99 ±2.10	55.77 ±2.80	86.86 ±2.74
2-HOP			75.83 ±1.07	59.88 ±1.39	78.09 ±0.53	72.40 ±0.43	63.94 ±1.33	57.68 ±1.67	89.25 ±2.27
3-HOP			72.58 ±0.86	55.29 ±1.58	75.56 ±0.34	71.74 ±0.69	62.67 ±2.26	55.59 ±1.48	87.27 ±1.92
AVG 2-WAY			73.22 ±0.45	61.74 ±1.57	82.33 ±0.40	73.94 ±0.37	66.29 ±1.30	60.93 ±1.49	89.21 ±1.95
Movie Lens									
CARDINALITY			98.91 ±0.06		98.99 ±0.16	98.70 ±0.40		98.87 ±0.26	99.07 ±0.18
1-WAY			86.58 ±0.80		99.19 ±0.00	68.38 ±0.36		78.03 ±0.17	99.34 ±0.10
0-HOP	DNC	DNC	72.80 ±0.86	TLE	98.56 ±0.01	31.96 ±0.32	TLE	57.33 ±0.10	98.69 ±0.15
1-HOP			74.86 ±0.63		92.72 ±0.09	58.00 ±0.05		77.45 ±1.93	96.19 ±0.11
AVG 2-WAY			74.10 ±0.62		94.87 ±0.06	48.45 ±0.09		70.07 ±1.19	97.11 ±0.02
CCS									
CARDINALITY		74.36 ±8.40	99.00 ±0.53	93.70 ±0.00	99.37 ±0.16	26.98 ±0.05	26.97 ±0.00	26.70 ±0.20	99.25 ±0.16
1-WAY		69.04 ±4.38	82.21 ±0.32	82.72 ±0.06	95.20 ±0.00	73.68 ±0.35	79.28 ±0.10	79.29 ±0.13	92.37 ±2.30
0-HOP	DNC	94.84 ±1.00	87.02 ±0.18	88.10 ±0.07	98.96 ±0.00	81.70 ±0.33	87.15 ±0.16	86.60 ±0.14	98.47 ±0.79
1-HOP		21.74 ±9.62	49.84 ±2.30	47.11 ±0.06	51.62 ±0.22	56.86 ±0.66	61.53 ±1.50	57.77 ±0.69	83.15 ±4.22
AVG 2-WAY		41.68 ±6.73	59.98 ±1.72	58.29 ±0.06	64.53 ±0.16	63.64 ±0.57	68.51 ±1.11	65.64 ±0.50	87.33 ±3.12

UNIVERSITY OF WATERLOO

Thank you!

UNIVERSITY OF WATERLOO

YOU+WATERLOO

Our greatest impact happens together.