## Dirty Dataset $D$

| | Name | Gender | Depart. | Age | Working Years |
|---|---|---|---|---|---|
| $t_1$ | Nancy | F | Sales | 24 | 2 |
| $t_2$ | Amy | | Finance | 42 | 10 |
| $t_3$ | Lei | M | | 35 | |
| $t_4$ | Lily | F | Market | | 8 |
| $t_5$ | James | M | Sales | 38 | 2 |
| $t_6$ | Tom | M | | 36 | 6 |

## A Coreset $C$

| | Name | Gender | Depart. | Age | Working Years | |
|---|---|---|---|---|---|---|
| $c_1$ | Nancy | F | Sales | 24 | 2 | $t_1$ |
| $c_2$ | Lei | M | | 35 | | $t_3$ |
| $c_3$ | Lily | F | Market | | 8 | $t_4$ |

$$\gamma(1) = 1 \qquad \gamma(2) = 3 \qquad \gamma(3) = 4$$

(a) An example of dataset with missing values and a coreset

### $1^{st}$ possible world

| | | | | |
|---|---|---|---|---|
| $t_1$ | F | Sales | 24 | 2 |
| $t_2$ | F | Finance | 42 | 10 |
| $t_3$ | M | Sales | 35 | 1 |
| $t_4$ | F | Market | 45 | 8 |
| $t_5$ | M | Sales | 38 | 2 |
| $t_6$ | M | Market | 36 | 6 |

### $2^{nd}$ possible world

| | | | | |
|---|---|---|---|---|
| F | Sales | 24 | 2 |
| F | Finance | 42 | 10 |
| M | Market | 35 | 10 |
| F | Market | 45 | 8 |
| M | Sales | 38 | 2 |
| M | Market | 36 | 6 |

### $32{,}000^{th}$ possible world

| | | | | |
|---|---|---|---|---|
| F | Sales | 24 | 2 |
| M | Finance | 42 | 10 |
| M | Finance | 35 | 5 |
| F | Market | 25 | 8 |
| M | Sales | 38 | 2 |
| M | Sales | 36 | 6 |

. . .



$p_1 S_1 \qquad\qquad p_2 S_2$
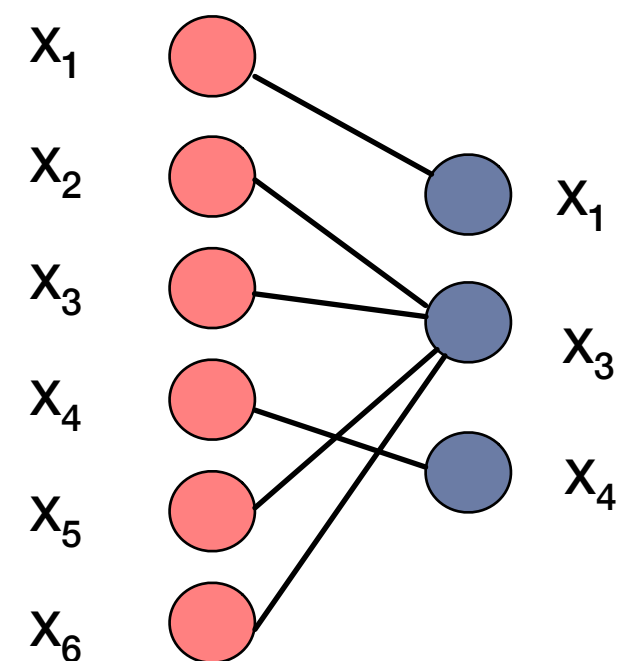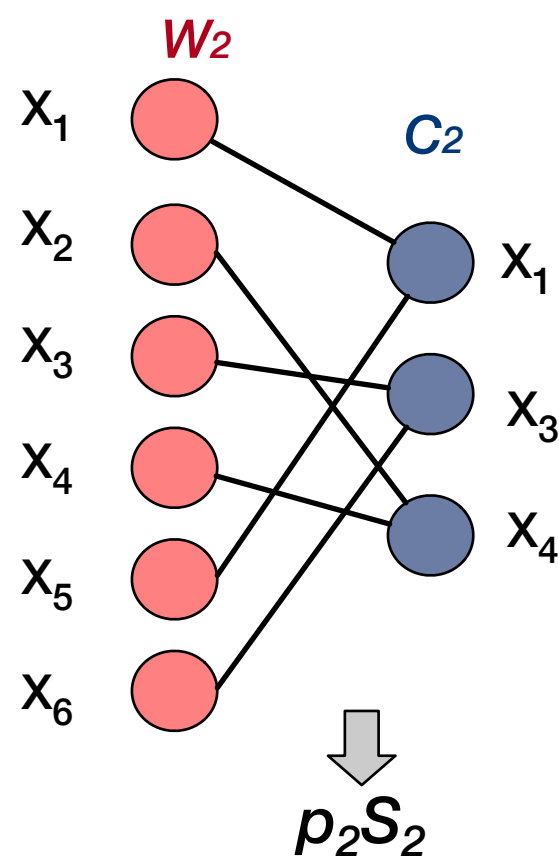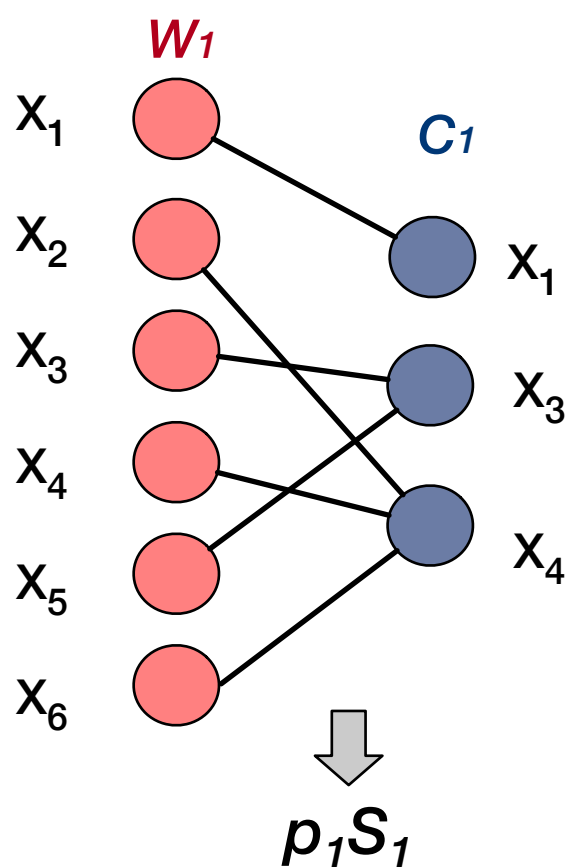
(b) An example of the possible worlds