Discovering Interesting Patterns in Large Graph Cubes

2017 BigGraphs Workshop at IEEE BigData'17

Florian Demesmaeker, Consultant @EURA NOVA

EURA NOVA

UCL Université catholique de Louvain

Discovering Interesting Patterns in Large Graph Cubes

Florian Demesmaeker, Amine Ghrab,
Prof. Siegfried Nijssen & Sabri Skhiri

Agenda

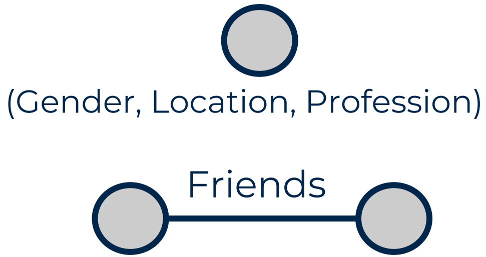
1. **Objectives**
2. **Interesting Itemset Mining Approach**
3. **Graph Cube Based Approach**
4. **Experiments**
5. **Conclusion**

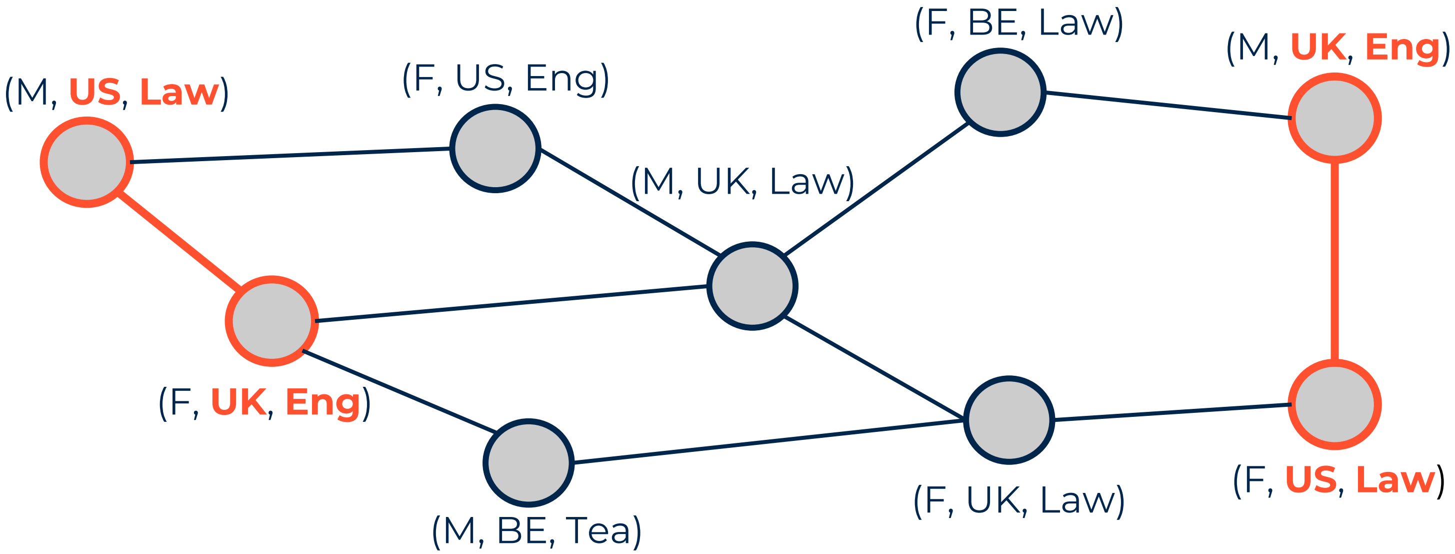
Agenda

1. **Objectives**
2. Interesting Itemset Mining Approach
3. Graph Cube Based Approach
4. Experiments
5. Conclusion

Objectives

- Exploratory data analysis
- Networks
- Surprising features associations

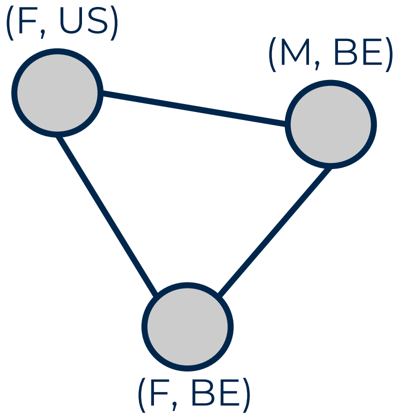




Agenda

1. Objectives
2. **Interesting Itemset Mining Approach**
3. Graph Cube Based Approach
4. Experiments
5. Conclusion

Pattern Mining with Itemsets



Edge	Transactions

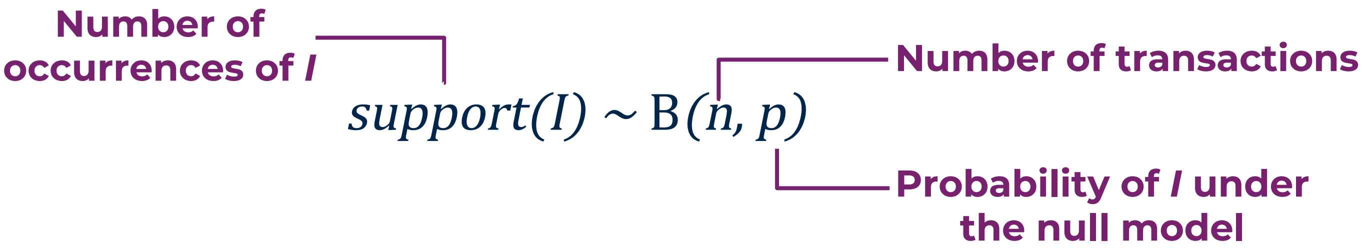
(F, US)-(M, BE)	{ (F,M), (US,BE) }
	{ (M,F), (BE,US) }

(F, BE)-(M, BE)	{ (F,M), (BE,BE) }
	{ (M,F), (BE,BE) }

(F, US)-(F, BE)	{ (F,F), (US,BE) }
	{ (F,F), (BE,US) }

Interesting Itemsets

- The independence between nodes as null model
- Under the null model, compute the probability of *support(I)*



Number of occurrences of *I*

$$support(I) \sim \mathrm{B}(n, p)$$

Number of transactions

Probability of *I* under the null model

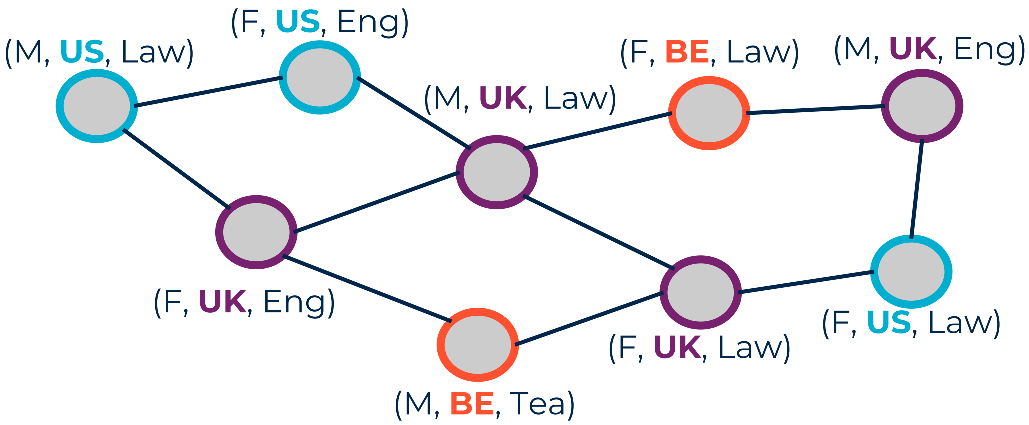
Arianna Gallo, Tijl De Bie, and Nello Cristianini. Mini: Mining informative non-redundant itemsets. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 438–445. Springer, 2007

Agenda

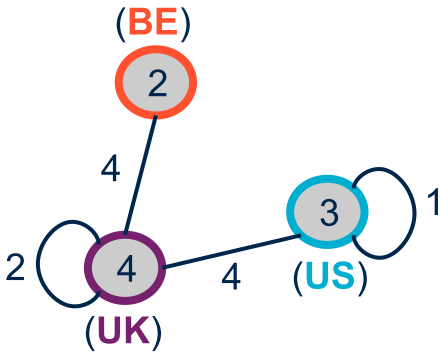
1. Objectives
2. Interesting Itemset Mining Approach
3. **Graph Cube Based Approach**
4. Experiments
5. Conclusion

The Search Space

The graph cube as a search structure



Original network



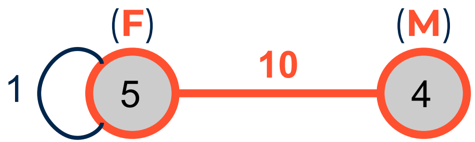
Aggregate network
w.r.t. (Location)

Peixiang Zhao, Xiaolei Li, Dong Xin, and Jiawei Han. Graph cube: on warehousing and olap multidimensional networks. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 853–864. ACM, 2011.

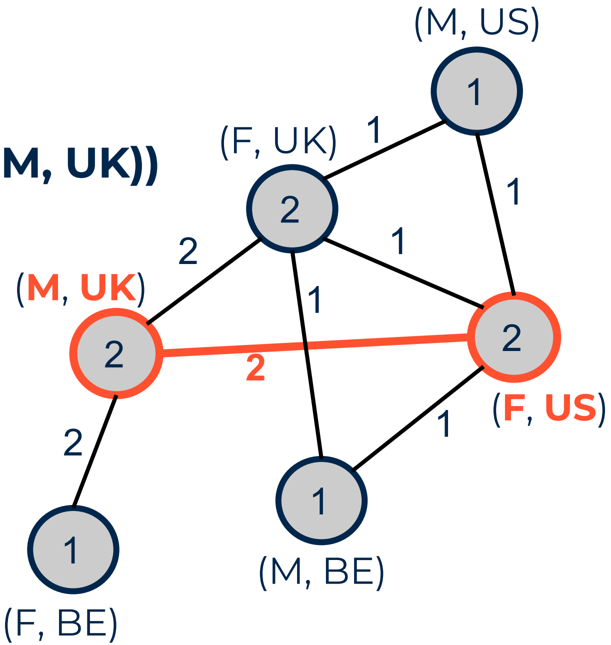
Interesting Patterns

The independence model

P’ = ((F), (M))



P = ((F, US), (M, UK))

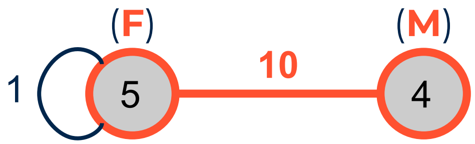


$$\Pr((\mathrm{F},\ \mathrm{US}),\ (\mathrm{M},\ \mathrm{UK}) \mid (\mathrm{F}),\ (\mathrm{M})) = \frac{support((\mathrm{F},\ \mathrm{US}))}{support((\mathrm{F}))} \times \frac{support((\mathrm{M},\ \mathrm{UK}))}{support((\mathrm{M}))}$$

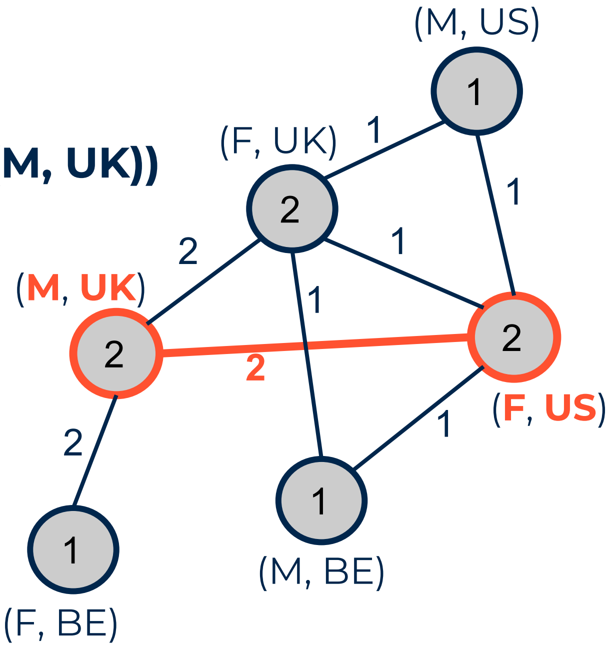
Interesting Patterns

The support probability

P’ = ((F), (M))



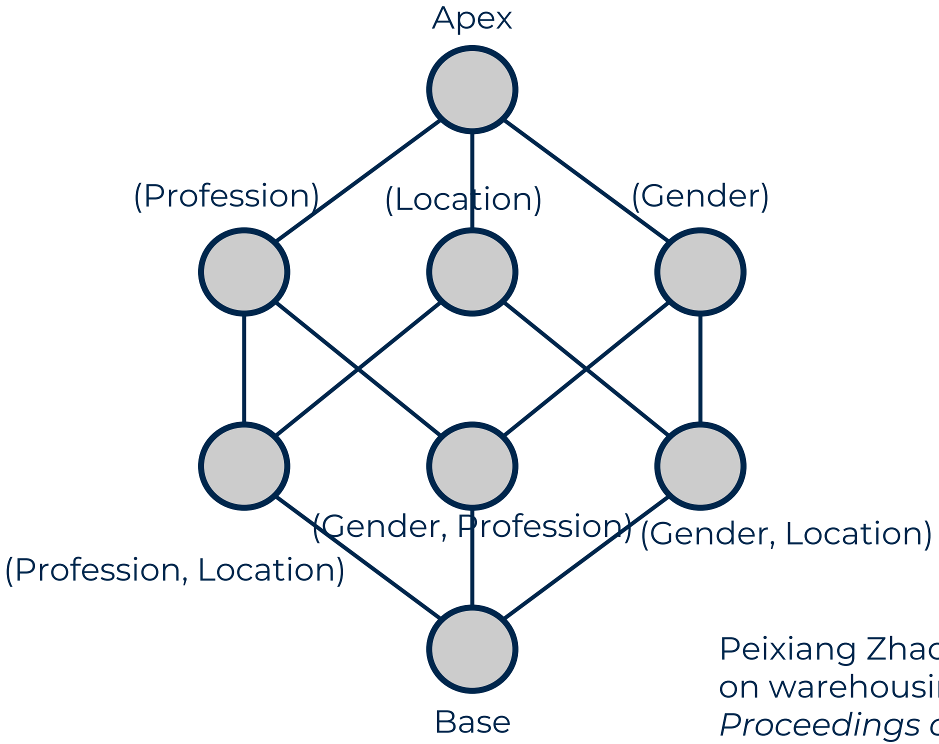
P = ((F, US), (M, UK))



$$\underbrace{support(P)}_{\text{Number of occurrences of P}} \sim \mathrm{B}(\overset{\text{supp(P’)}}{n}, \underset{\text{Pr(P | P’)}}{p})$$

The Graph Cube Lattice

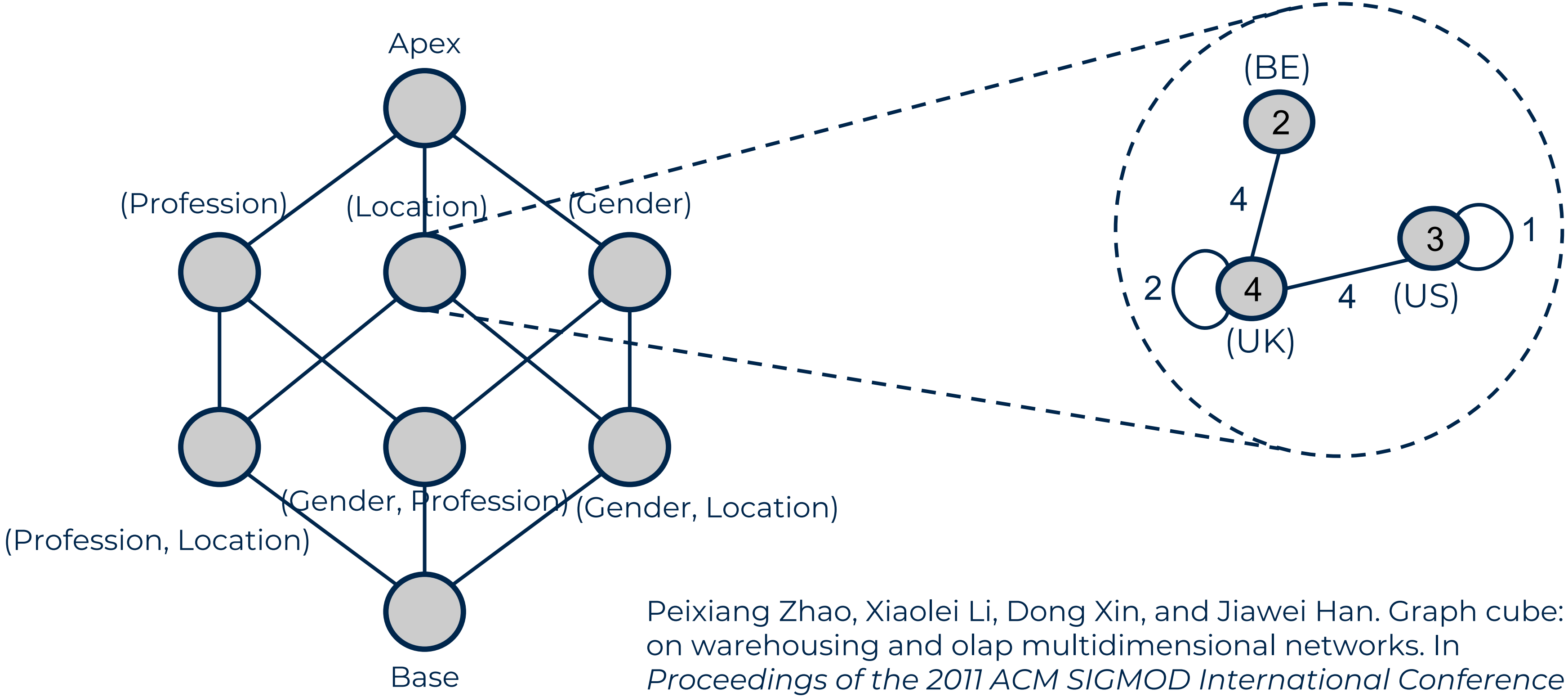
Represents the search space



Peixiang Zhao, Xiaolei Li, Dong Xin, and Jiawei Han. Graph cube: on warehousing and olap multidimensional networks. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 853–864. ACM, 2011.

The Graph Cube Lattice

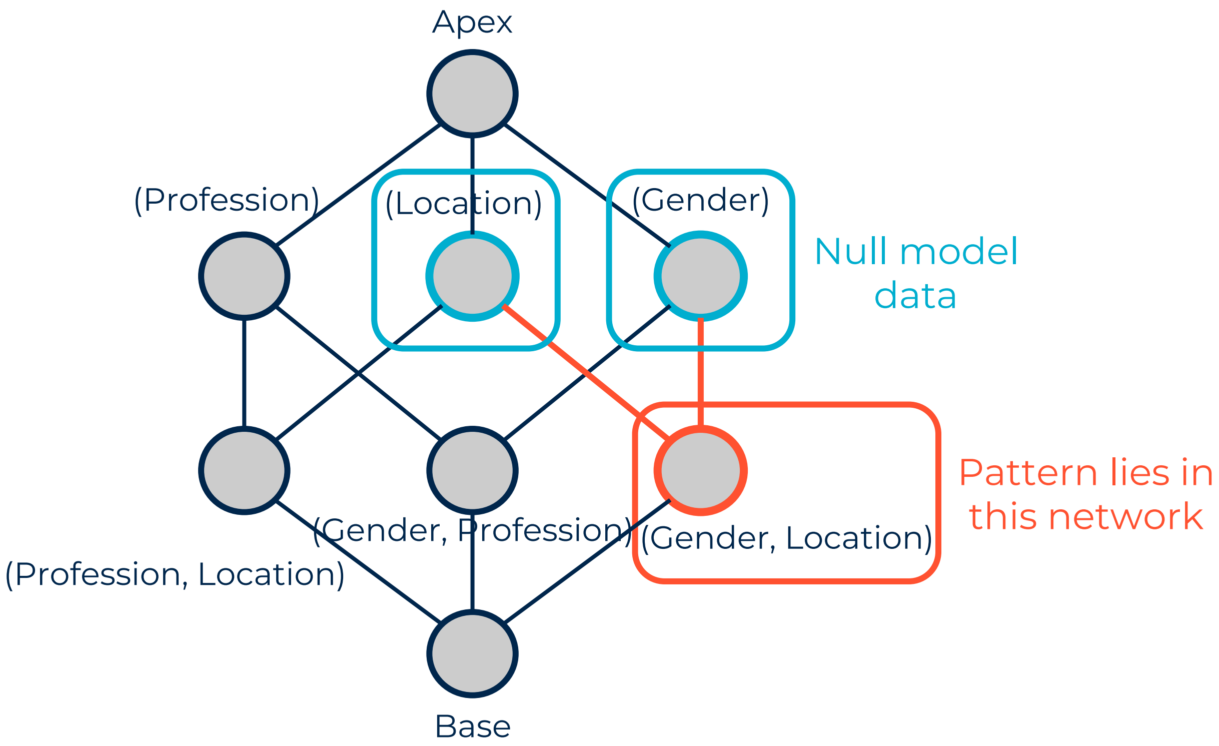
Represents the search space



Peixiang Zhao, Xiaolei Li, Dong Xin, and Jiawei Han. Graph cube: on warehousing and olap multidimensional networks. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 853–864. ACM, 2011.

Search

Uses the lattice



Agenda

1. Objectives
2. Interesting Itemset Mining Approach
3. Graph Cube Based Approach
4. **Experiments**
5. Conclusion

Experiments

- Extended MovieLens dataset: categorical data
- MINI: a greedy approach
- Graph cube mining: an exhaustive approach

Cuboid	Pattern	p-value
Critic	(Good)-(Very good)	3.67×10^{-194}
Critic	(Good)-(Very bad)	4.19×10^{-134}
Critic	(Very good)-(Very good)	4.19×10^{-130}
Country	(USA)-(USA)	5.32×10^{-117}
Critic	(Average)-(Very bad)	9.52×10^{-106}

Top 5 by graph cube miner

Cuboid	Pattern	p-value
Country, Critic	(USA, Good)-(USA, Good)	8.85×10^{-11}
Critic	(Good)-(Very good)	0.74
Critic	(Very good)-(Good)	0.74
Decade	(2000)-(2000)	0.74
Decade	(2000)-(1990)	0.74

Top 5 by MINI

Agenda

1. Objectives
2. Interesting Itemset Mining Approach
3. Graph Cube Based Approach
4. Experiments
5. **Conclusion**

Conclusion

- A data exploratory analysis technique in networks
  - Searching for surprising features associations
- A statistical approach
  - Compute the probability of observing a pattern property under the null model
  - An itemset mining approach
  - A graph cube based approach
- Perspectives
  - Perform experiments on synthetic data
  - Handle numerical attributes

Q/A time

Frequent Itemset Mining (FIM)

Transaction ID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke

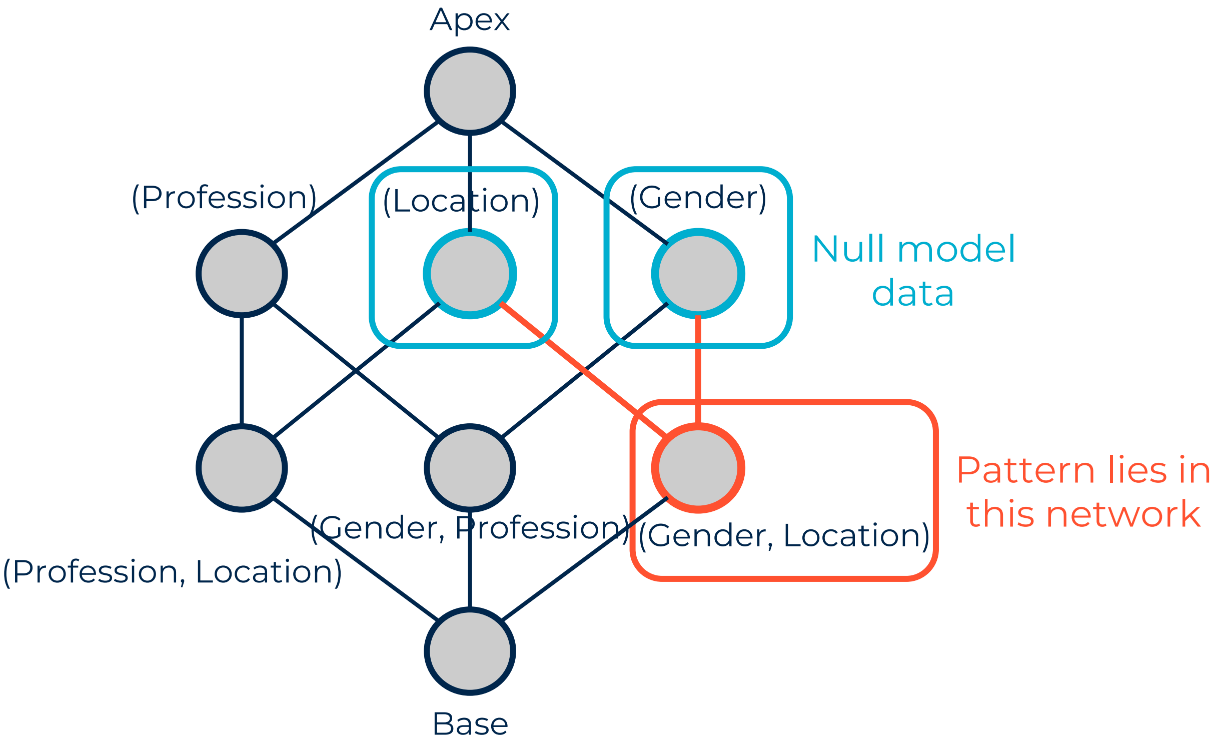
Size 1
Diaper
Milk
...

Size 2
Bread, Diaper
Diaper, Eggs
...

Size 3
Bread, Diaper, Coke
Diaper, Milk, Coke
...

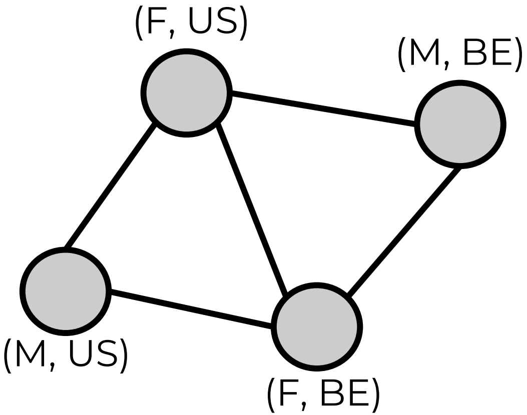
Size 4
Eggs, Coke, Beer, Milk
Bread, Milk, Diaper, Coke
...

Search



$$interest(P) = \max_{C' \in ancestors(C)} \Pr(support(P) \mid P')$$

Pattern Mining with Itemsets



Edge	Transactions

(F, US)-(M, BE)	{ (F,M), (US,BE) }
	{ (M,F), (BE,US) }

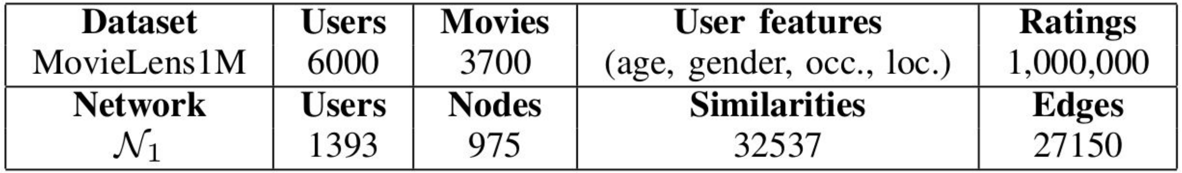
(F, BE)-(M, BE)	{ (F,M), (BE,BE) }
	{ (M,F), (BE,BE) }

(F, US)-(F, BE)	{ (F,F), (US,BE) }
	{ (F,F), (BE,US) }

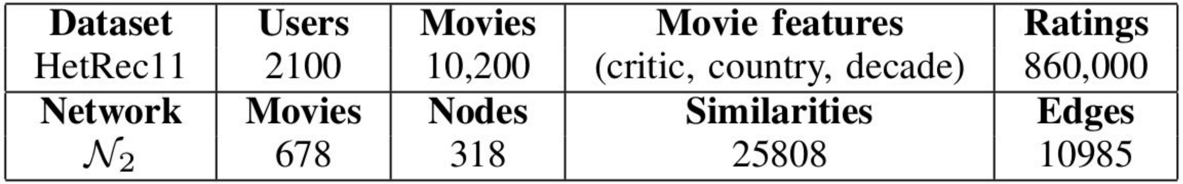
(F, US)-(M, US)	{ (F,M), (US,US) }
	{ (M,F), (US,US) }

(F, US)-(F, BE)	{ (F,F), (US,BE) }
	{ (F,F), (BE,US) }

Datasets statistics



Dataset	Users	Movies	User features	Ratings
MovieLens1M	6000	3700	(age, gender, occ., loc.)	1,000,000
Network	**Users**	**Nodes**	**Similarities**	**Edges**
$\mathcal{N}_1$	1393	975	32537	27150



Dataset	Users	Movies	Movie features	Ratings
HetRec11	2100	10,200	(critic, country, decade)	860,000
Network	**Movies**	**Nodes**	**Similarities**	**Edges**
$\mathcal{N}_2$	678	318	25808	10985

Search time

Cutoff	Movielens	Extended Movielens
$p < 1$	322 sec	3 sec
$p < 0.01$	32 sec	1.5 sec

Bibliography

- Arianna Gallo, Tijl De Bie, and Nello Cristianini. Mini: Mining informative non-redundant itemsets. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 438–445. Springer, 2007

- Peixiang Zhao, Xiaolei Li, Dong Xin, and Jiawei Han. Graph cube: on warehousing and olap multidimensional networks. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 853–864. ACM, 2011.

- https://grouplens.org/datasets/movielens/1m/

- I. Cantador, P. Brusilovsky, and T. Kuflik, “2nd workshop on information heterogeneity and fusion in recommender systems (hetrec 2011),” in Proceedings of the 5th ACM conference on Recommender systems, ser. RecSys 2011. New York, NY, USA: ACM, 2011.