University of BRISTOL

Making Sense of (Multi-)Relational Data

Part IV: Exploration by Descriptive Modelling – Fully Relational Local Approaches

Jefrey Lijffijt

Eirini Spyropoulou

Tijl De Bie

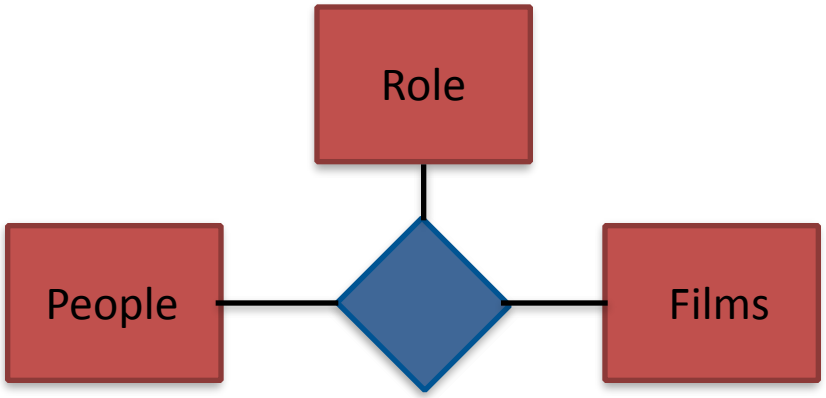
Fully relational local algorithms

- N-set mining
- RMiner \& variants
- Constraint programming for closed relational sets
- Uncovering the plot

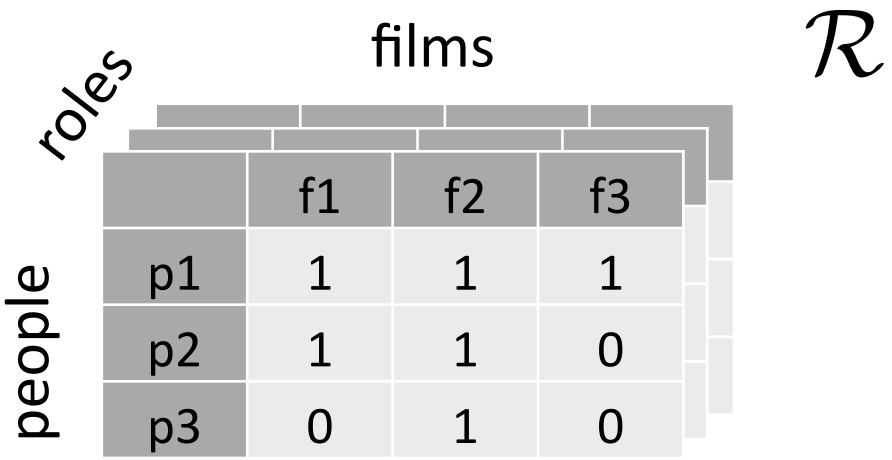
N-set mining

University of BRISTOL

N-set mining (Cerf et al., Trans. Knowl. Discov. Data, 2009



Pattern Syntax, N-sets



University of BRISTOL

N-sets

$\mathcal{R}$

films

people

	f1	f2	f3
p1	1	1	1
p2	1	1	0
p3	1	0	0

actor

	f1	f2	f3
p1	1	1	1
p2	1	1	0
p3	0	1	0

director

	f1	f2	f3
p1	0	1	1
p2	0	0	0
p3	0	0	1

producer

Sets of entities that are complete with respect to the relationship and maximal

University of BRISTOL

N-sets

$\mathcal{R}$

films

people

	f1	**f2**	**f3**
p1	1	1	1
p2	1	1	0
p3	1	0	0

actor

	f1	**f2**	**f3**
p1	1	1	1
p2	1	1	0
p3	0	1	0

director

	f1	**f2**	**f3**
p1	0	1	1
p2	0	0	0
p3	0	0	1

producer

N-set: {p1, f2, f3, actor, director, producer}

University of BRISTOL

N-sets

$\mathcal{R}$

films

people

	f1	**f2**	**f3**
p1	1	1	1
p2	1	1	0
p3	1	0	0

actor

	f1	**f2**	**f3**
p1	1	1	1
p2	1	1	0
p3	0	1	0

director

	f1	**f2**	**f3**
p1	0	1	1
p2	0	0	0
p3	0	0	1

producer

N-set: {p1, f2, f3, actor, director, producer}

{p1, f2, actor}$\in \mathcal{R}$ {p1, f2, director}$\in \mathcal{R}$

{p1, f3, actor}$\in \mathcal{R}$ {p1, f3, director}$\in \mathcal{R}$

University of BRISTOL

Interestingness of N-sets

- No interestingness measure defined.
- Constraints on the number of entities per entity type help to focus on a smaller pattern set.

University of BRISTOL

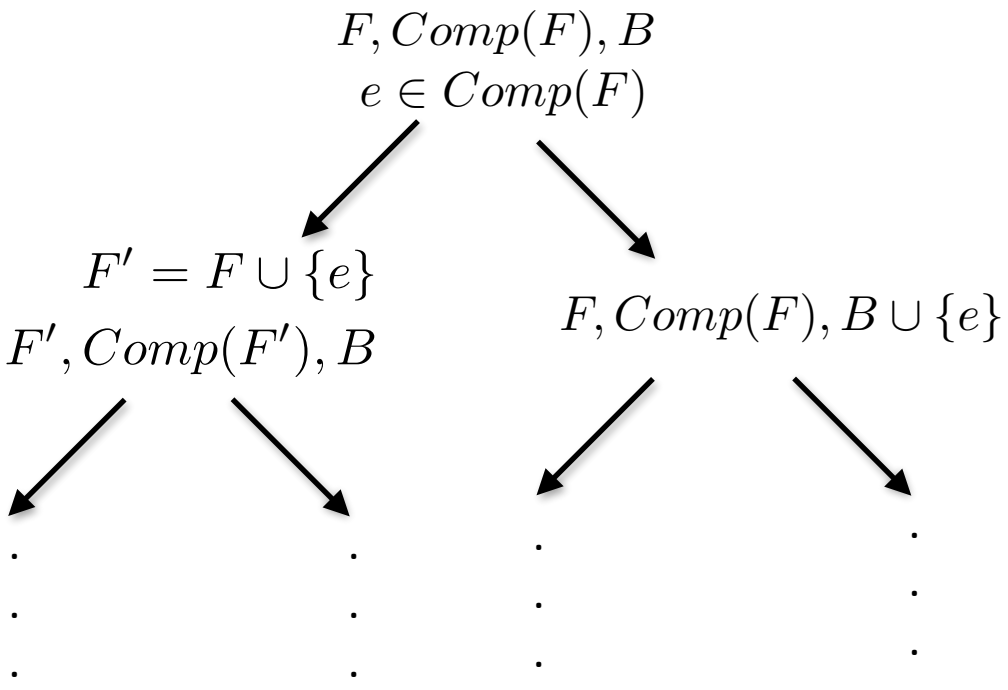
N-set Mining Algorithm - DataPeeler

- Exhaustive Search
- Divide and conquer enumeration strategy
- Use completeness to reduce the search space

DataPeeler

F : current solution

$Comp(F) : \forall e \in E, F \cup \{e\}$ is complete

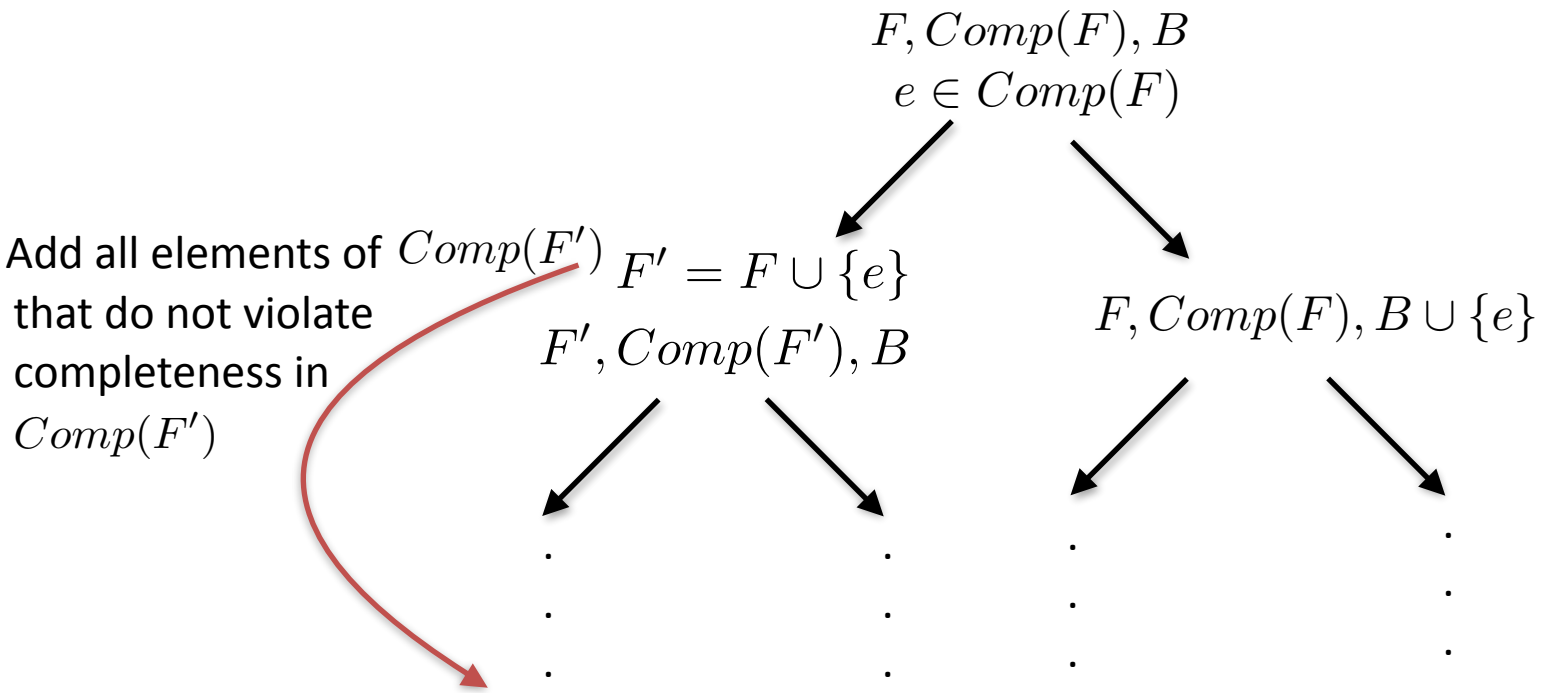


University of BRISTOL

DataPeeler

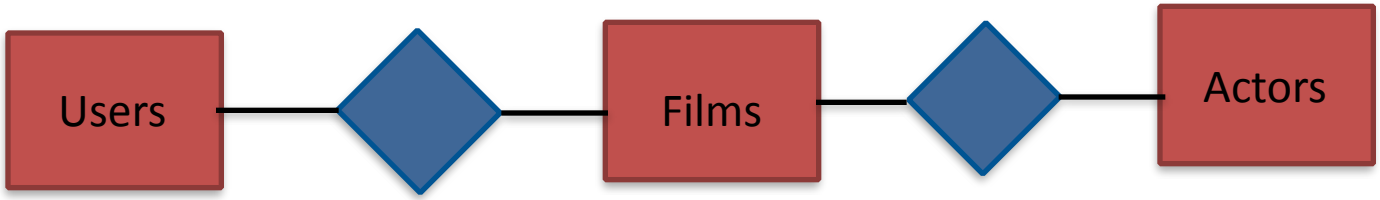
F : current solution

$Comp(F)$: $\forall e \in E, F \cup \{e\}$ is complete



RMiner & variants

RMiner - E.S. et al.,DMKD, 2014



University of BRISTOL

RMiner - Pattern Syntax (MCCSs)

films

users

	f1	f2	f3
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	a2	a3
f1	1	0	1
f2	1	0	1
f3	0	1	0

University of BRISTOL

MCCSs (Maximal Complete Connected Subsets)

films

users

	f1	f2	f3
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	a2	a3
f1	1	0	1
f2	1	0	1
f3	0	1	0

University of BRISTOL

MCCSs (Maximal Complete Connected Subsets)

Completeness

$$\mathcal{R} = R_{users,films} \cup R_{films,actors}$$

films

users

	f1	f2	f3
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	a2	a3
f1	1	0	1
f2	1	0	1
f3	0	1	0

University of BRISTOL

MCCSs (Maximal Complete Connected Subsets)

Complete Subset: {u1, f1, f2, a1}

films

users

	f1	**f2**	f3
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	a2	a3
f1	1	0	1
f2	1	0	1
f3	0	1	0

University of BRISTOL

MCCSs (Maximal Complete Connected Subsets)

Complete Subset: {u1, f1, f2, a1}

$\{u1,f1\} \in \mathcal{R}$
$\{u1,f2\} \in \mathcal{R}$
$\{f1,a1\} \in \mathcal{R}$
$\{f2,a1\} \in \mathcal{R}$

users \ films

	f1	**f2**	f3
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

films \ actors

	a1	a2	a3
f1	1	0	1
f2	1	0	1
f3	0	1	0

University of BRISTOL

RMiner - Pattern Syntax (MCCSs)

Connectedness

films

users

	f1	**f2**	f3
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	a2	a3
f1	1	0	1
f2	1	0	1
f3	0	1	0

RMiner - Pattern Syntax (MCCSs)

Complete Connected Subset (CCS): {u1, f1, f2, a1}

films

users

	f1	**f2**	f3
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	a2	a3
f1	1	0	1
f2	1	0	1
f3	0	1	0

University of BRISTOL

RMiner - Pattern Syntax (MCCSs)

Complete **not connected** Subset: {u1, a2}

films

users

	f1	f2	f3
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	**a2**	a3
f1	1	0	1
f2	1	0	1
f3	0	1	0

RMiner - Pattern Syntax (MCCSs)

Complete Connected Subset (CCS): {u1, f1, f2, a1}

films

users

	f1	**f2**	f3
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	a2	a3
f1	1	0	1
f2	1	0	1
f3	0	1	0

University of BRISTOL

RMiner - Pattern Syntax (MCCSs)

Maximality

films

users

	f1	**f2**	f3
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	a2	a3
f1	1	0	1
f2	1	0	1
f3	0	1	0

University of BRISTOL

RMiner - Pattern Syntax (MCCSs)

Maximal Complete Connected Subset (MCCS): {u1, u2, f1, f2, a1, a3}

films

users

	f1	**f2**	f3
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	a2	**a3**
f1	1	0	1
f2	1	0	1
f3	0	1	0

University of BRISTOL

films

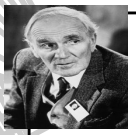
users

43603
30977
46603
24231
58087
48395
48357
33141
45478
38373

genres

Action
Thriller
Adventure

actors



Desmond Llewelyn

Lois Maxwell

RMiner - Interestingness

$$\text{Interestingness}(\gamma) = \frac{\text{Self Information}(\gamma)}{\text{Description Length}(\gamma)}.$$

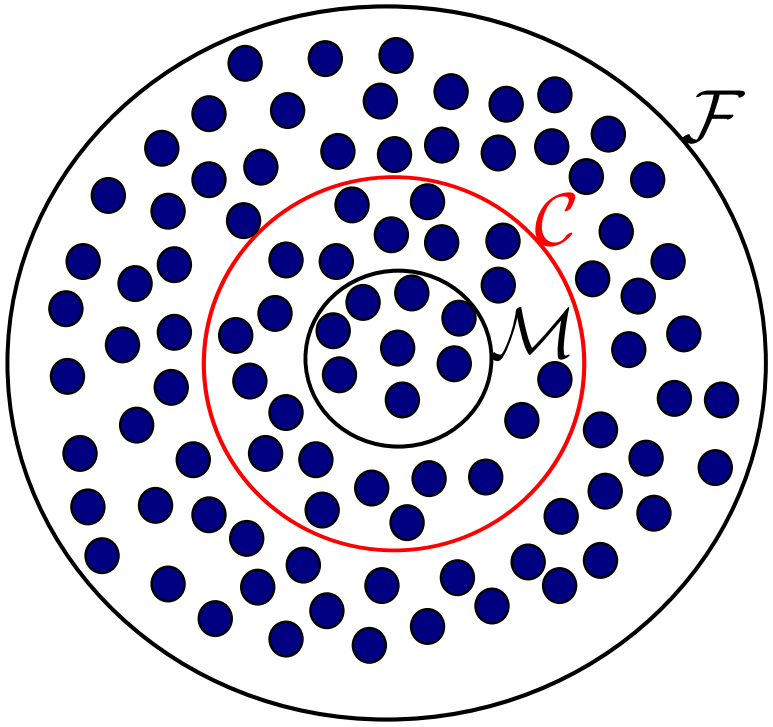
University of BRISTOL

Interestingness - Background model

- Maximum Entropy distribution on the data
- Constraints on the expected number of “1s” in every row and every column to be equal to the actual number
- The probability between entities that exist in many relationship instances is going to be high

University of BRISTOL

RMiner - Algorithm



$\mathcal{F}$: the set of CCSs

$\mathcal{M}$: the set of MCCSs

$\mathcal{C}$: the set of closed CCSs

$$\mathcal{M} \subseteq \mathcal{C} \subseteq \mathcal{F}$$

Closed CCSs

- Fixpoints of a closure operator

Mapping $\rho : \mathcal{F} \to \mathcal{F}$

Extensivity: $F \subseteq \rho(F)$ for all $F \in \mathcal{F}$

Monotonicity: $\rho(F) \subseteq \rho(F')$ for all $F, F' \in \mathcal{F}$ with $F \subseteq F'$

Idempotence: $\rho(\rho(F)) = \rho(F)$ for all $F \in \mathcal{F}$

RMiner - Algorithm

- Exhaustive search
- Based on divide and conquer algorithmic framework of Boley et al.

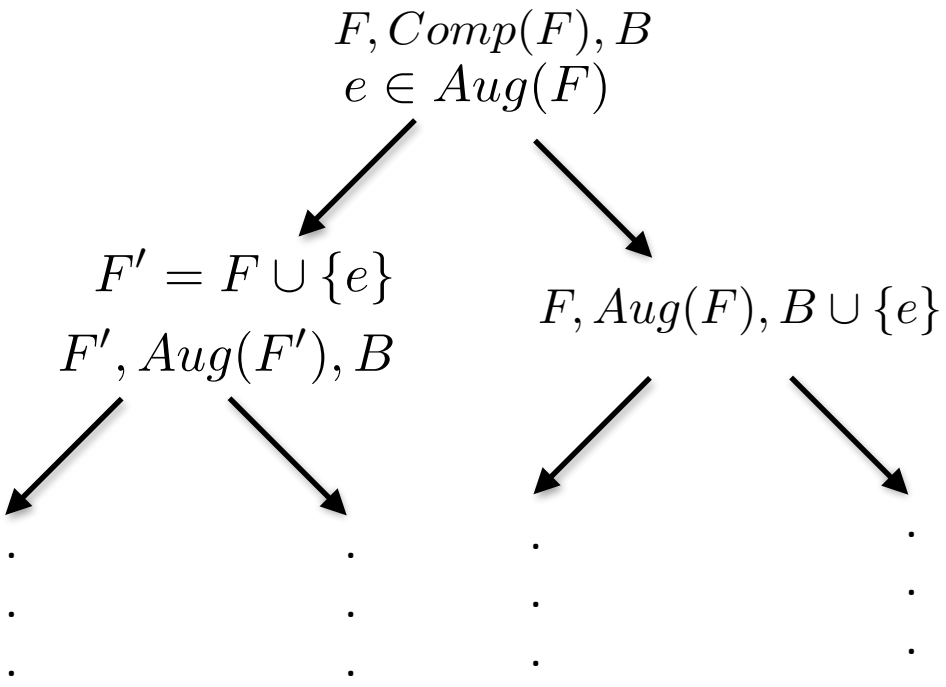
University of BRISTOL

RMiner - Algorithm

F : current solution

$Comp(F) : \forall e \in E, F \cup \{e\}$ is complete

$Aug(F) : \forall e \in E, F \cup \{e\}$ is complete and connected



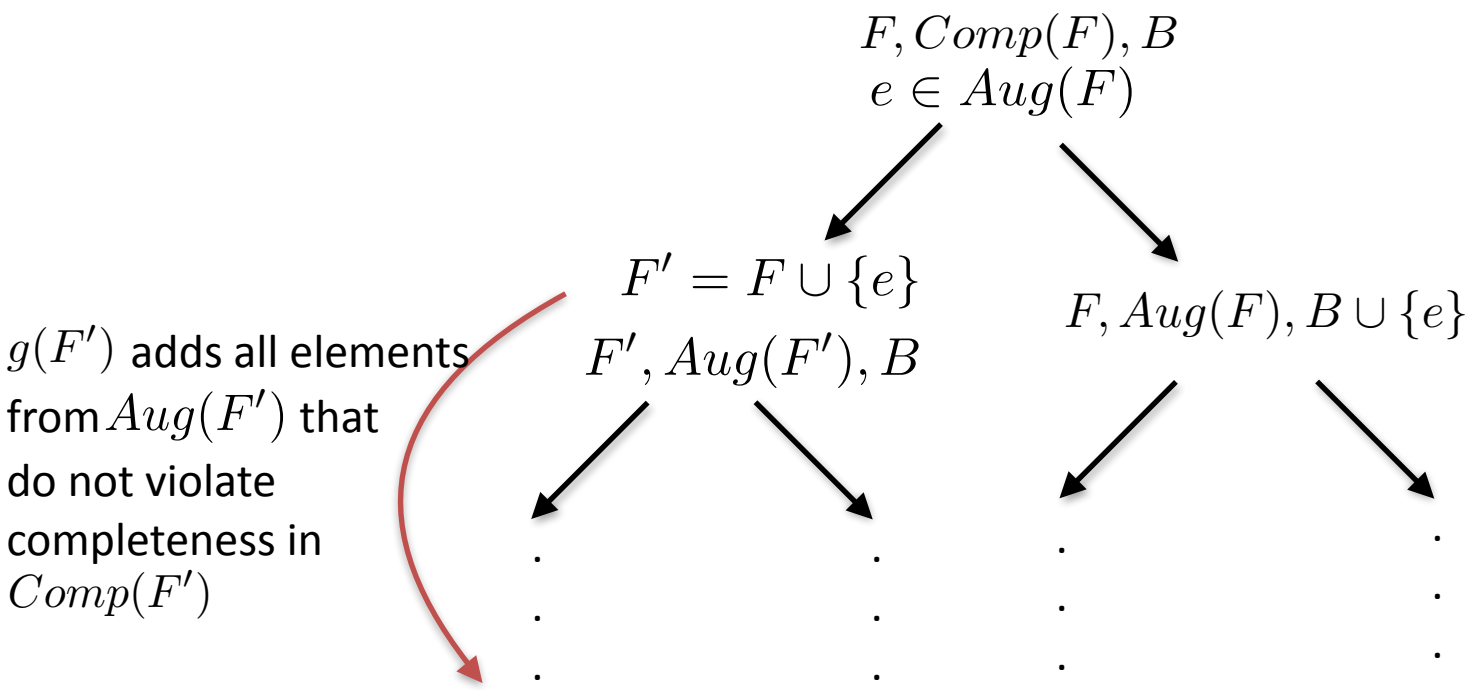
University of BRISTOL

RMiner - Algorithm

F : current solution

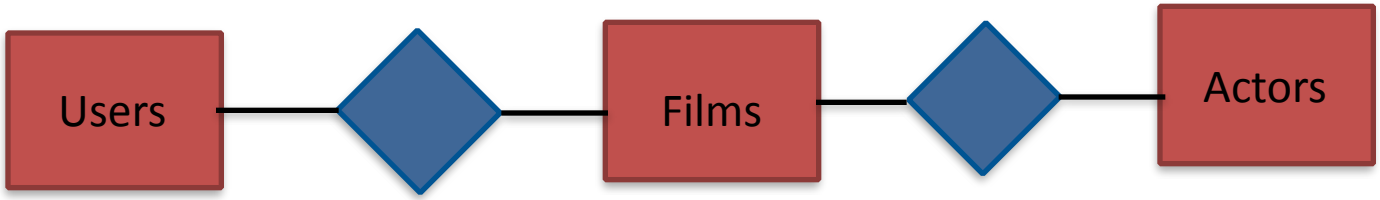
$Comp(F) : \forall e \in E, F \cup \{e\}$ is complete

$Aug(F) : \forall e \in E, F \cup \{e\}$ is complete and connected



A-RMiner

A-RMiner (E.S. et al., DSAA, 2014)



University of BRISTOL

A-RMiner Pattern Syntax (a-CCSs)

- Unions of MCCSs that are maximal extensions of a closed CCS.

University of BRISTOL

A-RMiner Pattern Syntax (a-CCSs)

films

users

	f1	f2	f3
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	a2	a3
f1	1	0	1
f2	1	0	1
f3	0	1	0

University of BRISTOL

A-RMiner Pattern Syntax (a-CCSs)

M1 = {u1, u2, f1, f2, a1, a3}

films

users

	f1	**f2**	f3
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	a2	**a3**
f1	1	0	1
f2	1	0	1
f3	0	1	0

University of BRISTOL

A-RMiner Pattern Syntax (a-CCSs)

M2 = {u3, u4, f2, f3}

films

users

	f1	**f2**	**f3**
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	a2	a3
f1	1	0	1
f2	1	0	1
f3	0	1	0

University of BRISTOL

A-RMiner Pattern Syntax (a-CCSs)

a-CCS: M1 U M2 = {u1, u2, u3, u4, f1, f2, f3, a1, a3}

films

users

	f1	**f2**	**f3**
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	a2	a3
f1	1	0	1
f2	1	0	1
f3	0	1	0

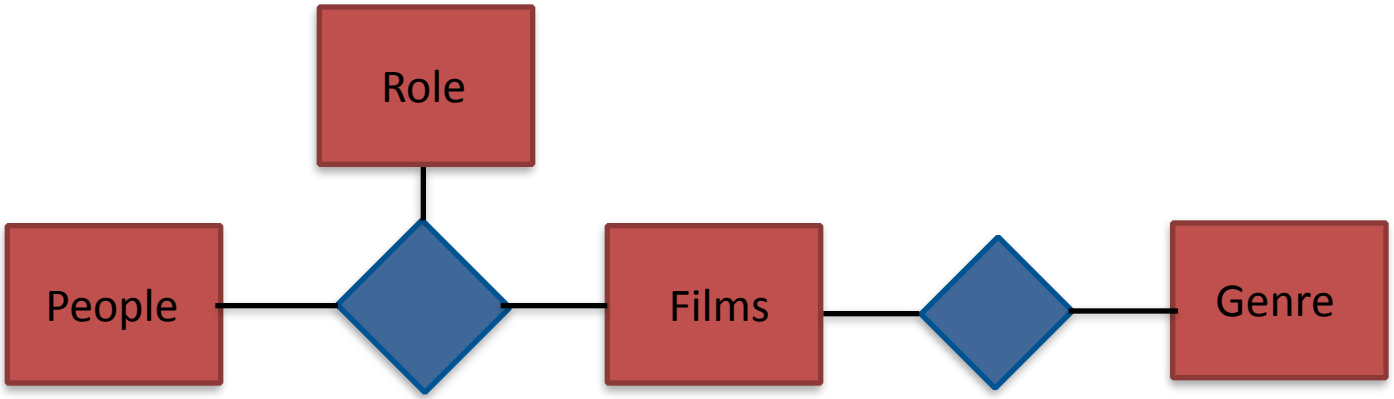
University of BRISTOL

A-RMiner - Algorithm

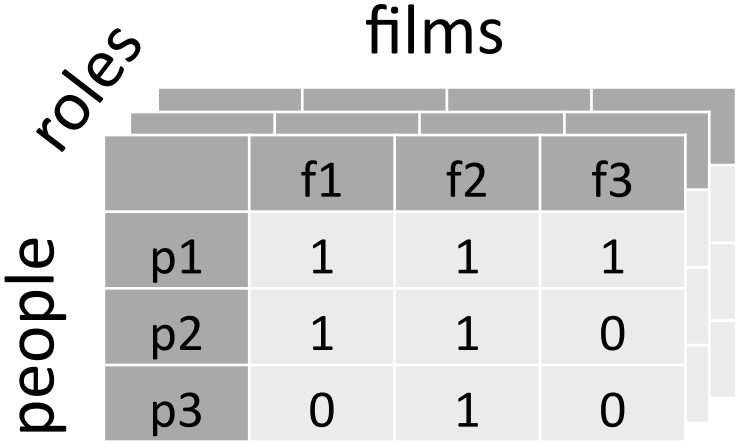
- Same algorithmic framework as RMiner
- Finds approximate patterns at no extra computational cost

University of BRISTOL

N-RMiner (E.S. et al., Discovery Science 2013)



N-RMiner Pattern Syntax (NMCCSs)



University of BRISTOL

NMCCSs

films

people

	f1	f2	f3
p1	1	1	1
p2	1	1	0
p3	1	0	0

actor

	f1	f2	f3
p1	1	1	1
p2	1	1	0
p3	0	1	0

director

	f1	f2	f3
p1	0	1	1
p2	0	0	0
p3	0	0	1

producer

University of BRISTOL

NMCCSs

films

people

	f1	f2	f3
p1	1	1	1
p2	1	1	0
p3	1	0	0

actor

	f1	f2	f3
p1	1	1	1
p2	1	1	0
p3	0	1	0

director

	f1	f2	f3
p1	0	1	1
p2	0	0	0
p3	0	0	1

producer

Completeness: Similar to N-sets with but allowing a subset of the entity types to be in the pattern.

NMCCSs

films

people

	f1	**f2**	f3
p1	1	1	1
p2	1	1	0
p3	1	0	0

actor

	f1	**f2**	f3
p1	1	1	1
p2	1	1	0
p3	0	1	0

director

	f1	f2	f3
p1	0	1	1
p2	0	0	0
p3	0	0	1

producer

N-MCCS: {f1, f2, p1, p2, actor, director}

University of BRISTOL

NMCCSs

films

people

	f1	**f2**	**f3**
p1	1	1	1
p2	1	1	0
p3	1	0	0

actor

	f1	**f2**	**f3**
p1	1	1	1
p2	1	1	0
p3	0	1	0

director

	f1	**f2**	**f3**
p1	0	1	1
p2	0	0	0
p3	0	0	1

producer

N-MCCS: {p1, f2, f3, actor, director, producer}

University of BRISTOL

N-RMiner - Algorithm

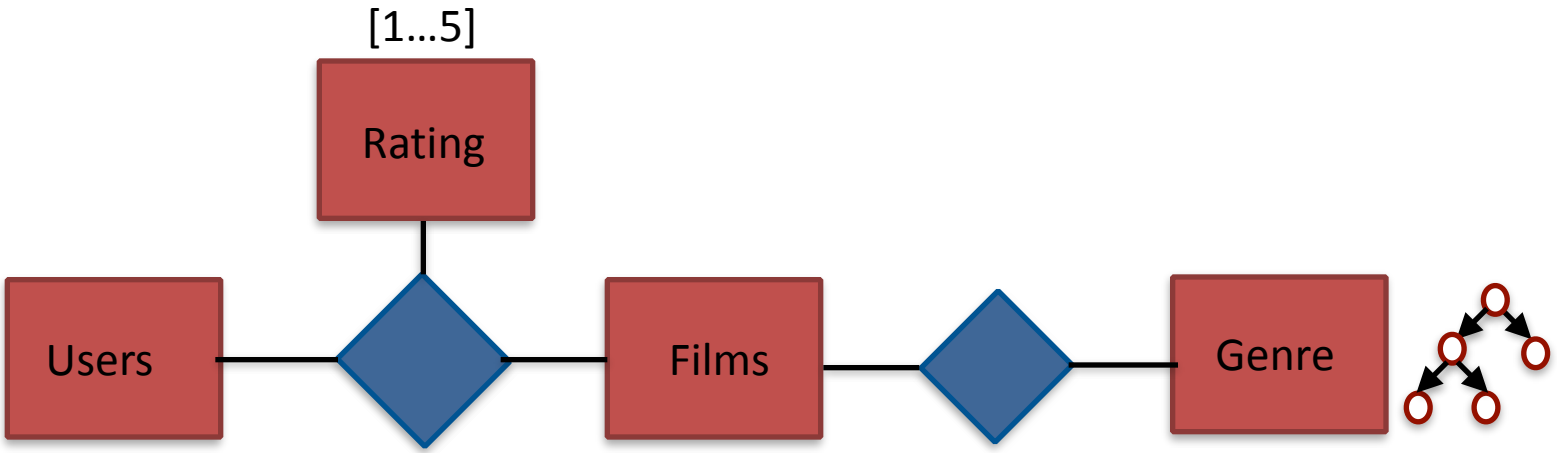
- Same enumeration algorithm
- Different in the way completeness is checked

N-RMiner - Interestingness

- Similar to RMiner

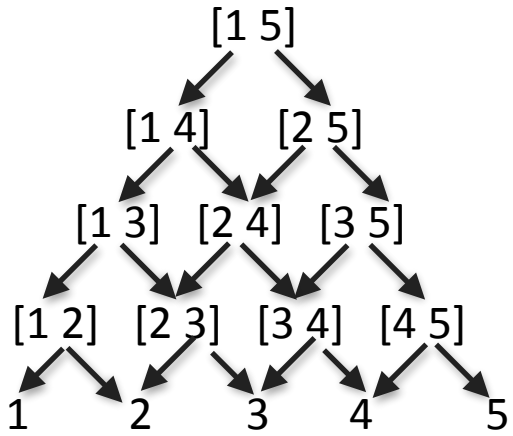
University of BRISTOL

P-N-RMiner (Lijffijt et al., DSAA, 2015)



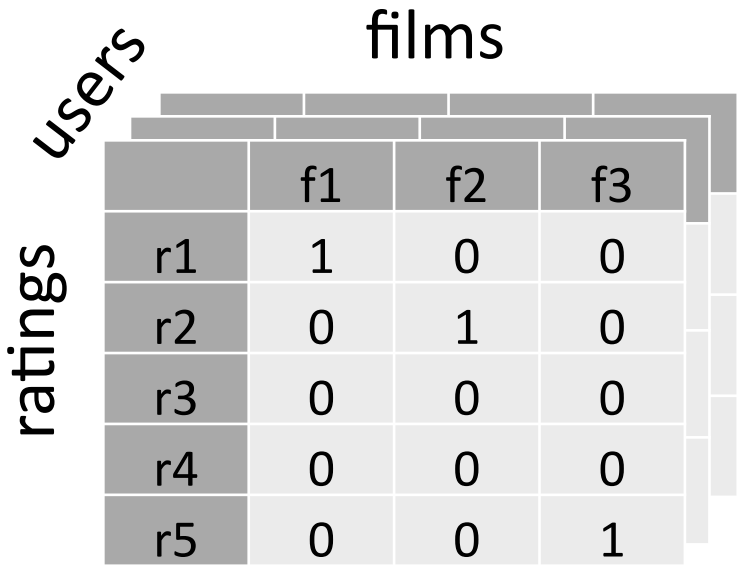
P-N-RMiner, Pattern Syntax

- Represents numerical attributes as partial orders



- It can handle both numerical and hierarchical attributes in the same way

P-N-RMiner, Pattern Syntax (MCCPSs)



	f1	f2	f3
r1	1	0	0
r2	0	1	0
r3	0	0	0
r4	0	0	0
r5	0	0	1

genres

films

	comedy	history	drama
f1	0	1	1
f2	0	1	1
f3	1	0	0

University of BRISTOL

P-N-RMiner, Pattern Syntax (MCCPSs)

users

films

ratings

	f1	f2	f3
r1	1	0	0
r2	0	1	0
r3	0	0	0
r4	0	0	0
r5	0	0	1

genres

films

	comedy	history	drama
f1	0	1	1
f2	0	1	1
f3	1	0	0

MCCPS: {user1, user2, [1 2], f1, f2, history, drama}

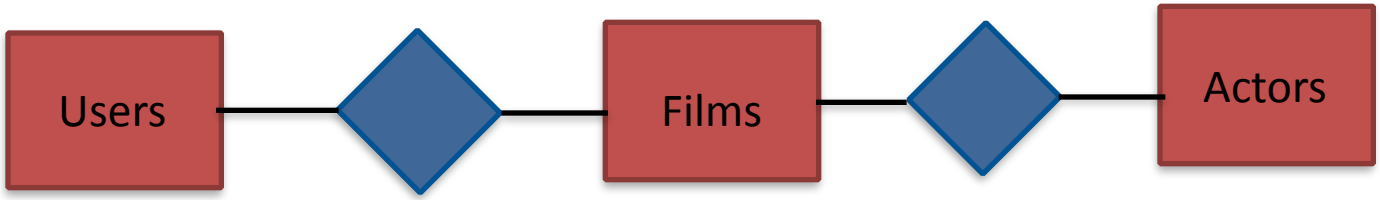
P-N-RMiner, Algorithm

- Same enumeration framework as RMiner
- Uses the partial order to reduce the search space

Constraint programming for closed relational sets

University of BRISTOL

Constraint programming for closed relational sets (Nijssen et al., ICDM Workshops 2011)



University of BRISTOL

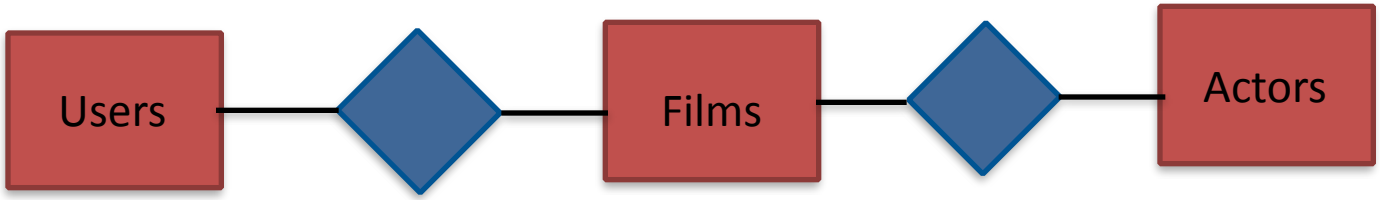
Constraint programming for closed relational sets (Nijssen et al., ICDM Workshops 2011)

- Constraint programming approach
- Defines both pattern syntax and pattern quality constraints
- Pattern syntax is the same as that of MCCSs
- Pattern quality constraints are on the minimum number of entities per entity type

Uncovering the plot

University of BRISTOL

Uncovering the plot (Wu et al. DMKD, 2014)



University of BRISTOL

Uncovering the plot (Wu et al. DMKD, 2014)

- Approximate multi-relational patterns similar to a-CCSs.
- Defined as chains of bi-clusters such that they overlap on one entity type.
- Works only for binary relationships.

Uncovering the plot - Bi-cluster Chains

films

users

	f1	**f2**	f3
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	a2	a3
f1	1	0	1
f2	1	0	1
f3	0	1	0

Uncovering the plot - Bi-cluster Chains

films

users

	f1	**f2**	f3
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	a2	a3
f1	1	0	1
f2	1	0	1
f3	0	1	0

University of BRISTOL

Uncovering the plot - Bi-cluster Chains

films

users

	f1	**f2**	f3
u1	1	1	0
u2	1	1	0
u3	0	1	1
u4	0	1	1

actors

films

	a1	a2	a3
f1	1	0	1
f2	1	0	1
f3	0	1	0

{u1, u2, f1, f2, a1}

Uncovering the plot - Algorithm

- Based on mining all bi-clusters on every relationship first.
- Then greedily combining them to form a chain.

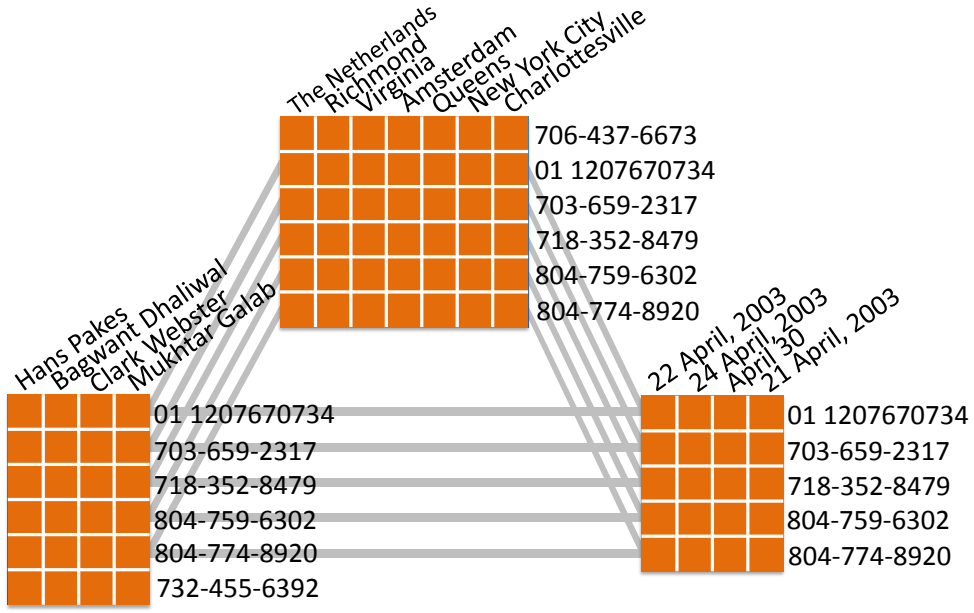
University of BRISTOL

Uncovering the plot - Pattern interestingness

- Information content with respect to MaxEnt background model
- Different prior knowledge to RMiner and variants
    - Defined over the join of all relationships
    - Has the form of constraints over the area of a tile.

University of BRISTOL

Uncovering the plot - Pattern example



University of BRISTOL

References

1. L. Cerf, J. Besson, C. Robardet, and J.-F. Boulicaut. Closed patterns meet n-ary relations. ACM Trans. Knowl. Discov. Data, 3(1):3:1–3:36, 2009.

2. E. Spyropoulou, T. De Bie and M. Boley. Interesting Pattern Mining in Multi-Relational Data. Data Mining & Knowledge Discovery, 28(3), pp.808-849, 2014.

3. E. Spyropoulou and T. De Bie. Approximate Multi-Relational Patterns. Proceedings of the International Conference on Data Science and Advanced Analytics (DSAA), 2014.

4. E. Spyropoulou, T. De Bie and M. Boley. Mining Interesting Patterns in Multi-Relational Data with N-ary Relationships. Proceedings of the International Conference on Discovery Science, pp. 217-232, 2013.

5. S. Nijssen, A. Jimenez and T Guns. Constraint-based pattern mining in multi-relational databases. Proceedings of Data Mining Workshops (ICDMW), 2011.

6. Uncovering the Plot: Detecting Surprising Coalitions of Entities in Multi-Relational Schemas. Data Mining and Knowledge Discovery vol.28(5), pp 1398-1428, Springer, 2014.

7. J. Lijffijt, E. Spyropoulou and T. De Bie. P-N-RMiner: A Generic Framework for Mining Interesting Structured Relational Patterns. Proceedings of the International Conference on Data Science and Advanced Analytics (DSAA),