ECDA 2013 - Luxembourg

Mining preferential datasets in MCDA

Alexandru-Liviu Olteanu [1,2] Raymond Bisdorff [1]

1. Université du Luxembourg

2. Institut Télécom, Télécom Bretagne
Université Européenne de Bretagne

10th of July 2013

Contents

1. **Data mining and clustering**

2. **Multiple criteria decision aid**

3. **Clustering in MCDA**

4. **Case Study: U.S. Toxic Chemicals Release Practices**

5. **Conclusions and Perspectives**

Data mining and clustering

Data mining and clustering

Data
- many forms;

  (measurements, observations, dynamics of processes, text, images, etc.)
- large quantities [Gantz and Reinsel 2011];

  $\approx 10^{21}$ bytes (100 TB for each person on the planet)

Data mining

- process that **extracts information** from a data set and **transforms** it into an **understandable structure** for further use;

Data mining and clustering

Data

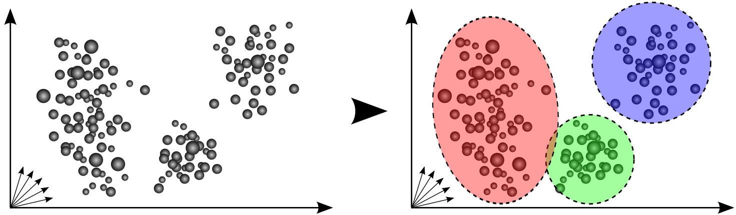
- many forms;

  (measurements, observations, dynamics of processes, text, images, etc.)
- large quantities [Gantz and Reinsel 2011];

  ≈ 10^{21} bytes (100 TB for each person on the planet)

Data mining

- process that **extracts information** from a data set and **transforms** it into an **understandable structure** for further use;



Multiple criteria decision aid

Multiple Criteria Decision Aid

- aims at modelling the **preferences** of decision-makers;
- **aids** them in reaching certain **decisions**;

Objects	Attributes			
	Price	Acceleration	Safety	⋯
Car 1	18,342	30.7s	good	⋯
Car 2	15,335	30.2s	medium	⋯
Car 3	16,973	29s	v.good	⋯
⋮	⋮	⋮	⋮	

Multiple criteria decision aid

Multiple Criteria Decision Aid

- aims at modelling the **preferences** of decision-makers;
- **aids** them in reaching certain **decisions**;

Alternatives	Criteria			
	Price ↓	Acceleration ↓	Safety ↑	⋯
Car 1	**18,342**	**30.7s**	good	⋯
Car 2	**15,335**	30.2s	**medium**	⋯
Car 3	16,973	**29s**	**v.good**	⋯
⋮	⋮	⋮	⋮	

Modelling preferences

Value functions

- aggregate all the criteria into a **score**;
- $(x_i, x_j, x_k, \ldots) \rightarrow U(x)$;
- **trade-offs** between criteria;

Outranking relations

- x **outranks** y iff:

1) x is **at least as good as** y on a **weighted majority** of criteria;
2) x is **not much worse** than y on any criterion;
$\rightarrow \quad x\,\mathrm{S}\,y$

- similar to **voting**;

Multiple criteria decision aid

Modelling preferences

Value functions

- aggregate all the criteria into a **score**;
- $(x_i, x_j, x_k, \dots) \rightarrow U(x)$;
- **trade-offs** between criteria;

Outranking relations

- x **outranks** y iff:

1) x is **at least as good as** y on a **weighted majority** of criteria;
2) x is **not much worse** than y on any criterion;
$\rightarrow \quad x\,S\,y$

- similar to **voting**;

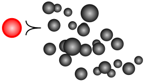
Preferential situations

$U(x) = U(y)$	Indifference (I)	$x\,S\,y \wedge y\,S\,x$
$U(x) > U(y)$	Strict preference (P)	$x\,S\,y \wedge y\,\not S\,x$
$U(x) \geqslant U(y)$	Weak preference (Q)	$x\,S\,y$
	Incomparability (R)	$x\,\not S\,y \wedge y\,\not S\,x$

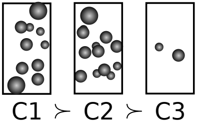
Multiple criteria decision aid

Decision problems

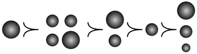
Choice



Sorting



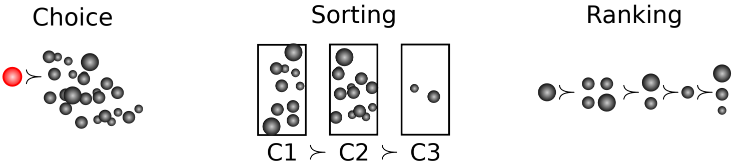
Ranking



s

Multiple criteria decision aid

Decision problems



Clustering in MCDA

Existing approaches:

- that use similarity measures:
- that use **preferential information**:

Formally defined using preferential relations in [Meyer, Olteanu 2013]s

Multiple criteria decision aid

Data mining		MCDA
• objects + attributes • **similarity**	▶	• alternatives + criteria • **indifference**, strict preference incomparability
• clustering - formally defined using similarity measures	▶	• clustering - formally defined using preferential measures [Meyer, Olteanu, 2013]
• problem size - easily > 10^6	▶	• problem size - rarely > 100

Clustering in MCDA

Clustering in MCDA

Clustering in Data mining

• process that groups objects that are **similar** and separates those that are **dissimilar**;

- objects that **cannot be distinguished** are **similar**;

Clustering in MCDA

Clustering in Data mining

- process that groups objects that are **similar** and separates those that are **dissimilar**;

- objects that **cannot be distinguished** are **similar**;

  BUT

- alternatives that **cannot be distinguished** are **indifferent**.

Clustering in MCDA

Clustering in Data mining

• process that groups objects that are **similar** and separates those that are **dissimilar**;

- objects that **cannot be distinguished** are **similar**;

  BUT

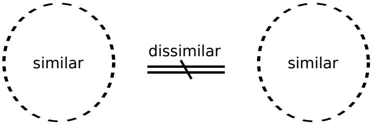
- alternatives that **cannot be distinguished** are **indifferent**.

Clustering in MCDA

• process that groups alternatives that are **indifferent** and separates those that are **not indifferent**;

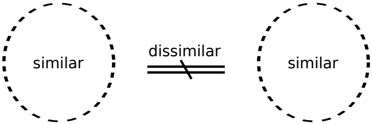
Clustering in MCDA

Classical clustering

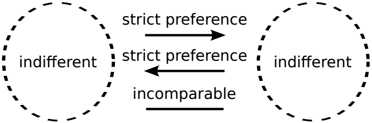


Clustering in MCDA

Classical clustering

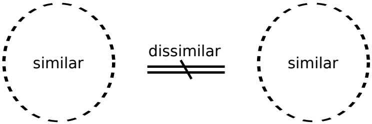


Non-relational clustering

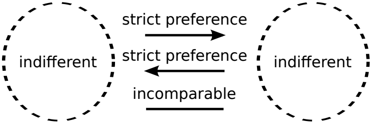


Clustering in MCDA

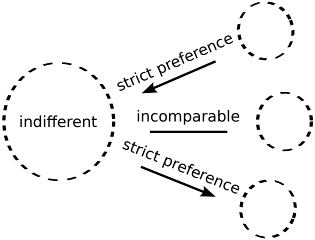
Classical clustering



Non-relational clustering

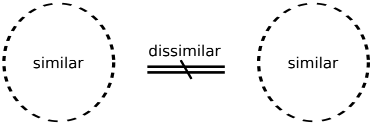


Relational clustering

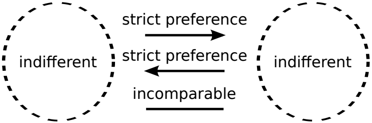


Clustering in MCDA

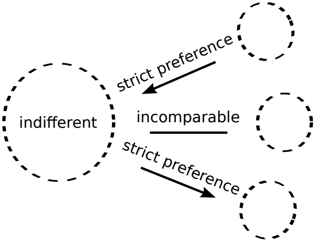
Classical clustering



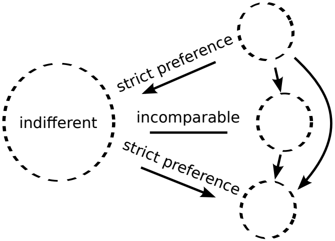
Non-relational clustering



Relational clustering



Ordered clustering



Case Study: U.S. Toxic Chemicals Release Practices

Case Study: U.S. Toxic Chemicals Release Practices

The data

- Toxic Chemical Release Practices of facilities in the U.S.;
- $> 53,000$ facilities reporting over 25 years;
- **selected** data from 2010 ($\sim 22,000$ reports);

- chemical **toxicity** information;
- reports containing the **release amounts** of a chemical;
- reports containing the **mitigated amounts** of a chemical;

Case Study: U.S. Toxic Chemicals Release Practices

The data

- Toxic Chemical Release Practices of facilities in the U.S.;
- $> 53,000$ facilities reporting over 25 years;
- **selected** data from 2010 ($\sim 22,000$ reports);

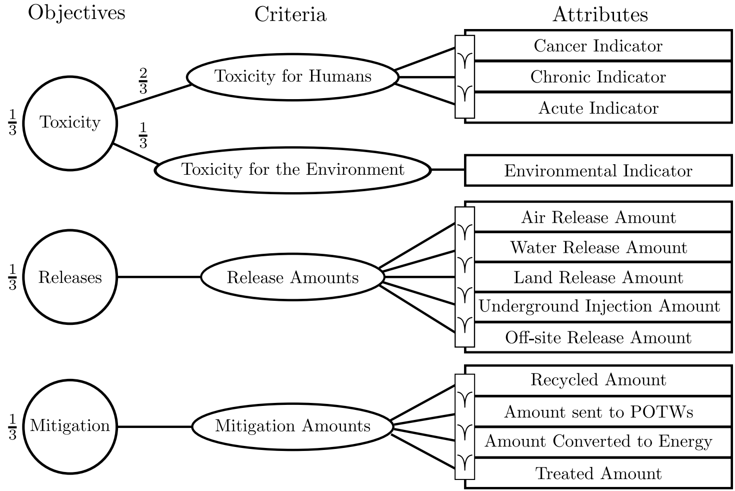
- chemical **toxicity** information;
- reports containing the **release amounts** of a chemical;
- reports containing the **mitigated amounts** of a chemical;

The problem

- **classifying** these practices w.r.t. their quality **without knowing** the classes a priori;
- **PREFERENCES**: handling of **less toxic** chemicals, **fewer releases** and **better mitigation** procedures;

Case Study: U.S. Toxic Chemicals Release Practices

Structuring the problem



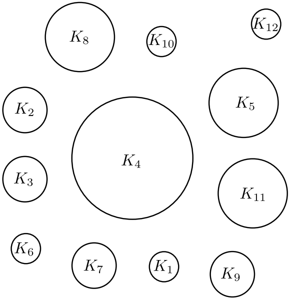
- Fictive decision-maker: bipolar-valued outranking relation [Bisdorff 2012];

Case Study: U.S. Toxic Chemicals Release Practices

Non-relational clustering

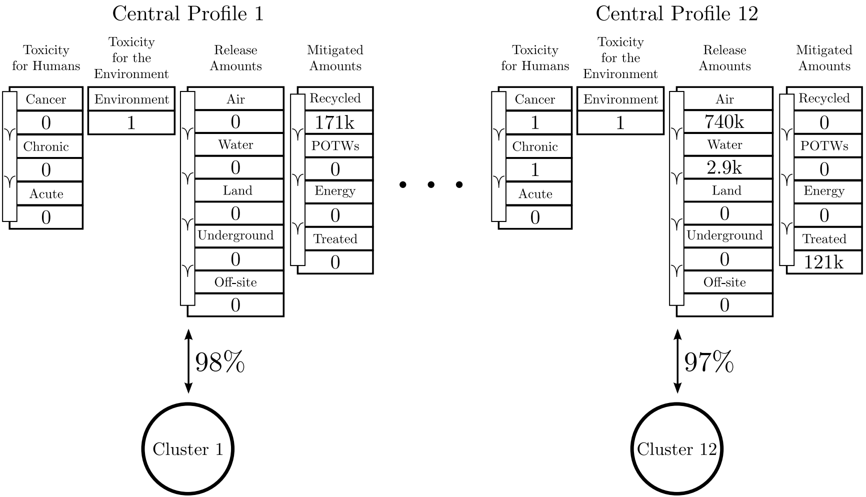
- used algorithms from [Meyer, Olteanu 2013]
- selected one result to illustrate (12 clusters);

Fitness (%)	
$\mathbf{f^*_{NR}}$	**80.0**
f_{NR}	62.7
f^{min}_{NR}	0.0
f^*_{R}	64.0
f_{R}	55.2
f^{min}_{R}	0.0



Cluster sizes	
K_1	296
K_2	**1,429**
K_3	**1,632**
K_4	**6,237**
K_5	**2,973**
K_6	167
K_7	**1,316**
K_8	**2,615**
K_9	**1,688**
K_{10}	540
K_{11}	**3,518**
K_{12}	356

Case Study: U.S. Toxic Chemicals Release Practices



Central Profile 1

Toxicity for Humans	Toxicity for the Environment	Release Amounts	Mitigated Amounts
Cancer	Environment	Air	Recycled
0	1	0	171k
Chronic		Water	POTWs
0		0	0
Acute		Land	Energy
0		0	0
		Underground	Treated
		0	0
		Off-site	
		0	

98%

Cluster 1

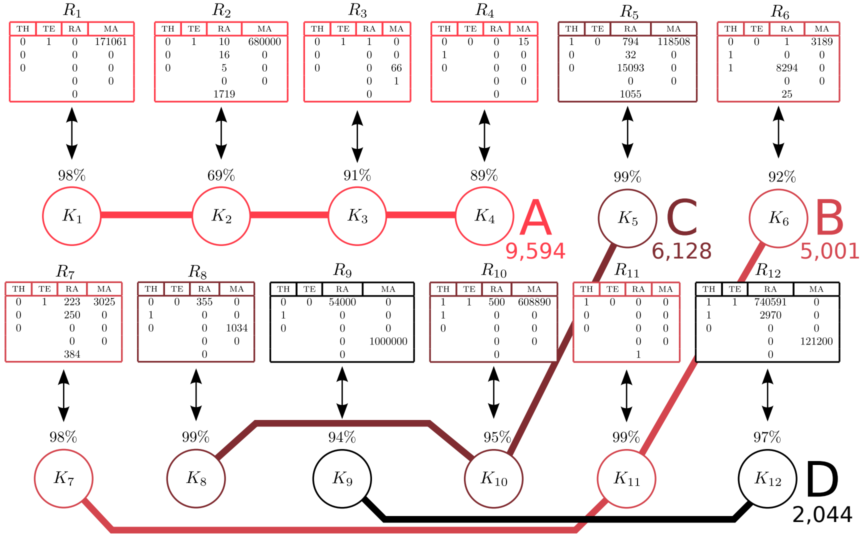
Central Profile 12

Toxicity for Humans	Toxicity for the Environment	Release Amounts	Mitigated Amounts
Cancer	Environment	Air	Recycled
1	1	740k	0
Chronic		Water	POTWs
1		2.9k	0
Acute		Land	Energy
0		0	0
		Underground	Treated
		0	121k
		Off-site	
		0	

97%

Cluster 12

Case Study: U.S. Toxic Chemicals Release Practices



R_1

TH	TE	RA	MA
0	1	0	171061
0		0	0
0		0	0
		0	0
		0	

R_2

TH	TE	RA	MA
0	1	10	680000
0		16	0
0		5	0
		0	0
		1719	

R_3

TH	TE	RA	MA
0	1	1	0
0		0	0
0		0	66
		0	1
		0	

R_4

TH	TE	RA	MA
0	0	0	15
1		0	0
0		0	0
		0	0
		0	

R_5

TH	TE	RA	MA
1	0	794	118508
0		32	0
0		15093	0
		0	0
		1055	

R_6

TH	TE	RA	MA
0	0	1	3189
1		0	0
1		8294	0
		0	0
		25	

R_7

TH	TE	RA	MA
0	1	223	3025
0		250	0
0		0	0
		0	0
		384	

R_8

TH	TE	RA	MA
0	0	355	0
1		0	0
0		0	1034
		0	0
		0	

R_9

TH	TE	RA	MA
0	0	54000	0
1		0	0
0		0	0
		0	1000000
		0	

R_{10}

TH	TE	RA	MA
1	1	500	608890
1		0	0
0		0	0
		0	0
		0	

R_{11}

TH	TE	RA	MA
1	0	0	0
0		0	0
0		0	0
		0	0
		1	

R_{12}

TH	TE	RA	MA
1	1	740591	0
1		2970	0
0		0	0
		0	121200
		0	

Conclusions and Perspectives

Conclusions and Perspectives

Conclusions:

- highlighted clustering using preferential information;
- illustrated an application of clustering in MCDA;

Perspectives:

- further explore clustering in MCDA (different structures);
- methodology for using clustering when eliciting the parameters of a preference model;
- combining similarity-based and indifference-based clustering (2 layers).

Mining preferential datasets in MCDA

1. **Data mining and clustering**
2. **Multiple criteria decision aid**
3. **Clustering in MCDA**
4. **Case Study: U.S. Toxic Chemicals Release Practices**
5. **Conclusions and Perspectives**

Annex

Illustrative example:

F	i	j	k
w	1	1	1
x	GOOD	MEDIUM	BAD
y	BAD	MEDIUM	GOOD

Annex

Illustrative example:

F	i	j	k
w	1	1	1
x	GOOD	MEDIUM	BAD
y	BAD	MEDIUM	GOOD

Similarity:

$x_i \neq y_i$ $x_j = y_j$ $x_k \neq y_k$ → x,y - **dissimilar**;

Indifference:

$$\left.\begin{array}{l} x_i \succcurlyeq y_i \;\; x_j \succcurlyeq y_j \;\; x_k \not\succcurlyeq y_k \rightarrow \text{x outranks y} \\ y_i \not\succcurlyeq x_i \;\; y_j \succcurlyeq x_j \;\; y_k \succcurlyeq x_k \rightarrow \text{y outranks x} \end{array}\right\} \rightarrow \text{x,y - } \textbf{indifferent}.$$

Comparative analysis

Measures:

- **similarity** measures from the Manhattan distance (S_{L_1}), the Euclidian distance (S_{L_2}), from [Bisdorff,Meyer,Olteanu 2011] (S_{THR}) and **indifference** measure from the outranking relation in [Bisdorff,Meyer,Roubens 2007] ($I_{\tilde{S}}$);

Comparative analysis

Measures:

- **similarity** measures from the Manhattan distance (S_{L_1}), the Euclidian distance (S_{L_2}), from [Bisdorff,Meyer,Olteanu 2011] (S_{THR}) and **indifference** measure from the outranking relation in [Bisdorff,Meyer,Roubens 2007] ($I_{\tilde{S}}$);

Experiment:

- all feasible alternatives on a fixed number of criteria with fixed number of values;
- compared similarity and indifferent measures for all pairs of alternatives;

Comparative analysis

Measures:

- **similarity** measures from the Manhattan distance (S_{L_1}), the Euclidian distance (S_{L_2}), from [Bisdorff,Meyer,Olteanu 2011] (S_{THR}) and **indifference** measure from the outranking relation in [Bisdorff,Meyer,Roubens 2007] ($I_{\tilde{S}}$);

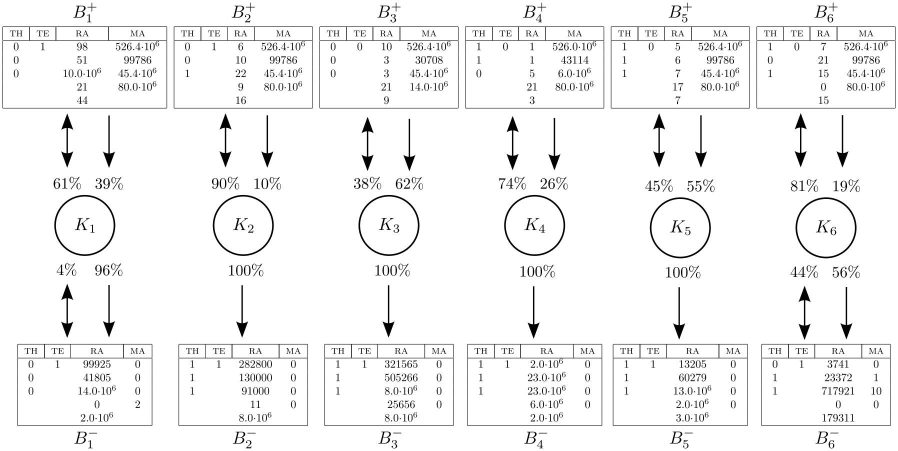
Experiment:

- all feasible alternatives on a fixed number of criteria with fixed number of values;
- compared similarity and indifferent measures for all pairs of alternatives;

Results:

- in at least 25% cases **dissimilar** alternatives were in fact **indifferent**;
- **significant differences** between similarity and indifference.

Annex



B_1^+

TH	TE	RA	MA
0	1	98	$526.4\cdot10^6$
0		51	99786
0		$10.0\cdot10^6$	$45.4\cdot10^6$
		21	$80.0\cdot10^6$
		44	

61% 39%

K_1

4% 96%

TH	TE	RA	MA
0	1	99925	0
0		41805	0
0		$14.0\cdot10^6$	0
		0	2
		$2.0\cdot10^6$	

B_1^-

B_2^+

TH	TE	RA	MA
0	1	6	$526.4\cdot10^6$
0		10	99786
1		22	$45.4\cdot10^6$
		9	$80.0\cdot10^6$
		16	

90% 10%

K_2

100%

TH	TE	RA	MA
1	1	282800	0
1		130000	0
1		91000	0
		11	0
		$8.0\cdot10^6$	

B_2^-

B_3^+

TH	TE	RA	MA
0	0	10	$526.4\cdot10^6$
0		3	30708
0		3	$45.4\cdot10^6$
		21	$14.0\cdot10^6$
		9	

38% 62%

K_3

100%

TH	TE	RA	MA
1	1	321565	0
1		505266	0
1		$8.0\cdot10^6$	0
		25656	0
		$8.0\cdot10^6$	

B_3^-

B_4^+

TH	TE	RA	MA
1	0	1	$526.0\cdot10^6$
1		1	43114
0		5	$6.0\cdot10^6$
		21	$80.0\cdot10^6$
		3	

74% 26%

K_4

100%

TH	TE	RA	MA
1	1	$2.0\cdot10^6$	0
1		$23.0\cdot10^6$	0
1		$23.0\cdot10^6$	0
		$6.0\cdot10^6$	0
		$2.0\cdot10^6$	

B_4^-

B_5^+

TH	TE	RA	MA
1	0	5	$526.4\cdot10^6$
1		6	99786
1		7	$45.4\cdot10^6$
		17	$80.0\cdot10^6$
		7	

45% 55%

K_5

100%

TH	TE	RA	MA
1	1	13205	0
1		60279	0
1		$13.0\cdot10^6$	0
		$2.0\cdot10^6$	0
		$3.0\cdot10^6$	

B_5^-

B_6^+

TH	TE	RA	MA
1	0	7	$526.4\cdot10^6$
0		21	99786
1		15	$45.4\cdot10^6$
		0	$80.0\cdot10^6$
		15	

81% 19%

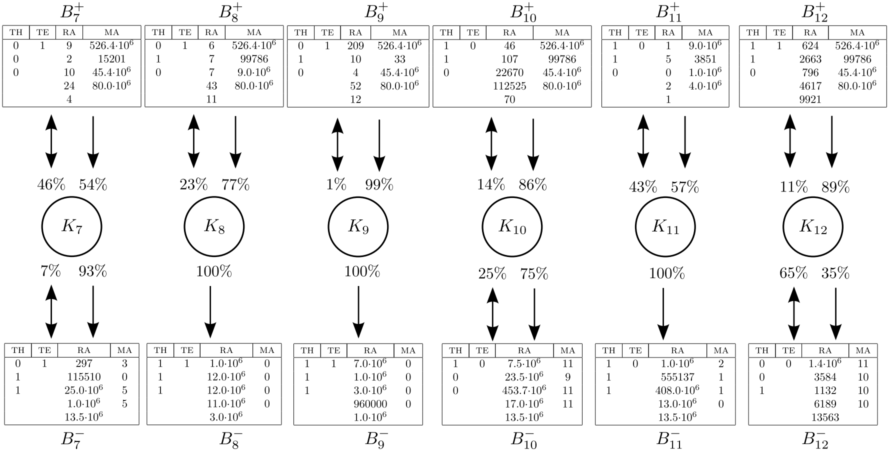
K_6

44% 56%

TH	TE	RA	MA
0	1	3741	0
1		23372	1
1		717921	10
		0	0
		179311	

B_6^-

Annex



B_7^+

TH	TE	RA	MA
0	1	9	$526.4{\cdot}10^6$
0		2	15201
0		10	$45.4{\cdot}10^6$
		24	$80.0{\cdot}10^6$
		4	

46% 54%

K_7

7% 93%

TH	TE	RA	MA
0	1	297	3
1		115510	0
1		$25.0{\cdot}10^6$	5
		$1.0{\cdot}10^6$	5
		$13.5{\cdot}10^6$	

B_7^-

B_8^+

TH	TE	RA	MA
0	1	6	$526.4{\cdot}10^6$
1		7	99786
0		7	$9.0{\cdot}10^6$
		43	$80.0{\cdot}10^6$
		11	

23% 77%

K_8

100%

TH	TE	RA	MA
1	1	$1.0{\cdot}10^6$	0
1		$12.0{\cdot}10^6$	0
1		$12.0{\cdot}10^6$	0
		$11.0{\cdot}10^6$	0
		$3.0{\cdot}10^6$	

B_8^-

B_9^+

TH	TE	RA	MA
0	1	209	$526.4{\cdot}10^6$
1		10	33
0		4	$45.4{\cdot}10^6$
		52	$80.0{\cdot}10^6$
		12	

1% 99%

K_9

100%

TH	TE	RA	MA
1	1	$7.0{\cdot}10^6$	0
1		$1.0{\cdot}10^6$	0
1		$3.0{\cdot}10^6$	0
		960000	0
		$1.0{\cdot}10^6$	

B_9^-

B_{10}^+

TH	TE	RA	MA
1	0	46	$526.4{\cdot}10^6$
1		107	99786
0		22670	$45.4{\cdot}10^6$
		112525	$80.0{\cdot}10^6$
		70	

14% 86%

K_{10}

25% 75%

TH	TE	RA	MA
1	0	$7.5{\cdot}10^6$	11
0		$23.5{\cdot}10^6$	9
0		$453.7{\cdot}10^6$	11
		$17.0{\cdot}10^6$	11
		$13.5{\cdot}10^6$	

B_{10}^-

B_{11}^+

TH	TE	RA	MA
1	0	1	$9.0{\cdot}10^6$
1		5	3851
0		0	$1.0{\cdot}10^6$
		2	$4.0{\cdot}10^6$
		1	

43% 57%

K_{11}

100%

TH	TE	RA	MA
1	0	$1.0{\cdot}10^6$	2
1		555137	1
1		$408.0{\cdot}10^6$	1
		$13.0{\cdot}10^6$	0
		$13.5{\cdot}10^6$	

B_{11}^-

B_{12}^+

TH	TE	RA	MA
1	1	624	$526.4{\cdot}10^6$
1		2663	99786
0		796	$45.4{\cdot}10^6$
		4617	$80.0{\cdot}10^6$
		9921	

11% 89%

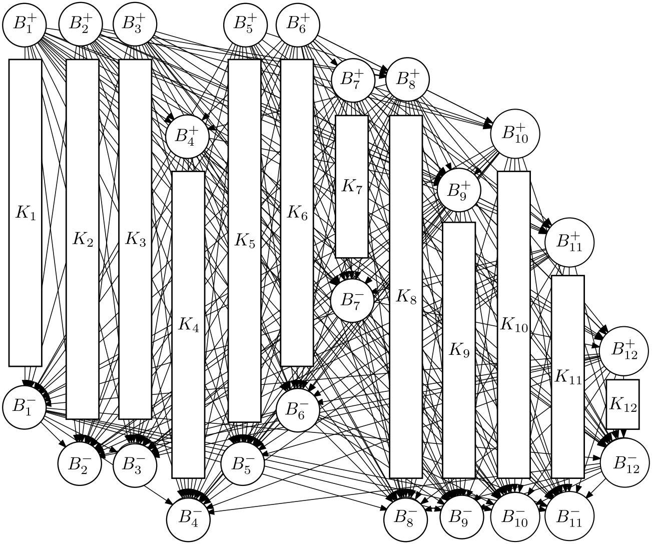
K_{12}

65% 35%

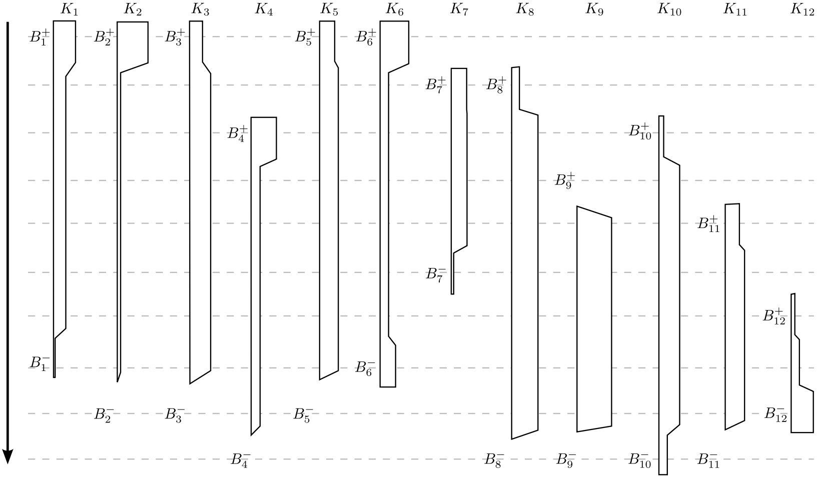
TH	TE	RA	MA
0	0	$1.4{\cdot}10^6$	11
0		3584	10
1		1132	10
		6189	10
		13563	

B_{12}^-

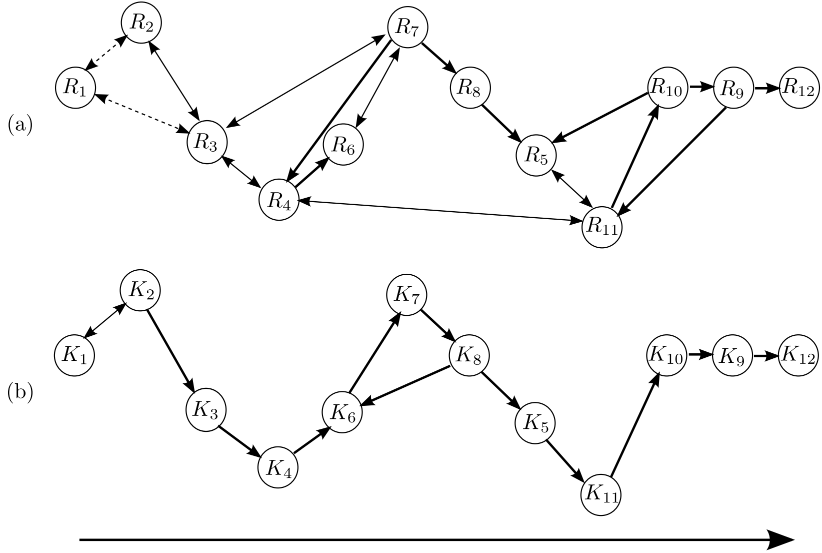
Annex


B_1^+
B_2^+
B_3^+
B_5^+
B_6^+
B_7^+
B_8^+
B_4^+
B_{10}^+
B_9^+
B_{11}^+
B_{12}^+
K_1
K_2
K_3
K_4
K_5
K_6
K_7
K_8
K_9
K_{10}
K_{11}
K_{12}
B_7^-
B_1^-
B_6^-
B_2^-
B_3^-
B_5^-
B_{12}^-
B_4^-
B_8^-
B_9^-
B_{10}^-
B_{11}^-

Annex


K_1
K_2
K_3
K_4
K_5
K_6
K_7
K_8
K_9
K_{10}
K_{11}
K_{12}
B_1^+
B_2^+
B_3^+
B_5^+
B_6^+
B_7^+
B_8^+
B_4^+
B_{10}^+
B_9^+
B_{11}^+
B_7^-
B_{12}^+
B_1^-
B_6^-
B_2^-
B_3^-
B_5^-
B_{12}^-
B_4^-
B_8^-
B_9^-
B_{10}^-
B_{11}^-

Annex


(a)
R_1
R_2
R_3
R_4
R_5
R_6
R_7
R_8
R_9
R_{10}
R_{11}
R_{12}
(b)
K_1
K_2
K_3
K_4
K_5
K_6
K_7
K_8
K_9
K_{10}
K_{11}
K_{12}