ECML–PKDD 2016 RIVA DEL GARDA

Data Mining and Knowledge Discovery

ClusPath: a temporal-driven clustering to infer typical evolution paths

20 September 2016

Marian-Andrei Rizoiu
Julien Velcin

Stéphane Bonnevay
Stéphane Lallich

Australian National University

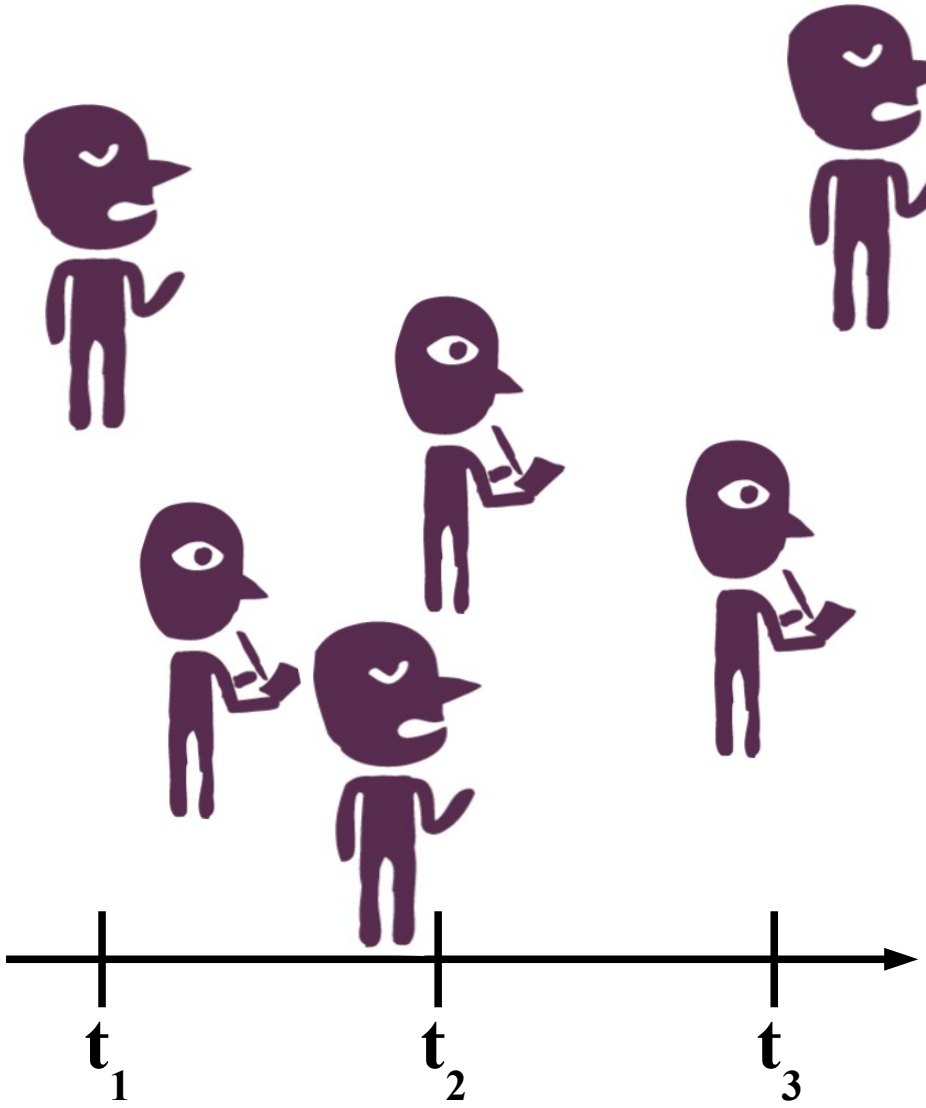
DATA 61

CSIRO

eric Entrepôts, Représentation & Ingénierie des Connaissances

Université Lumière Lyon 2
Université de Lyon

Dataset: descriptive features (x^d) for multiple entities (φ) at different moments of time (t)

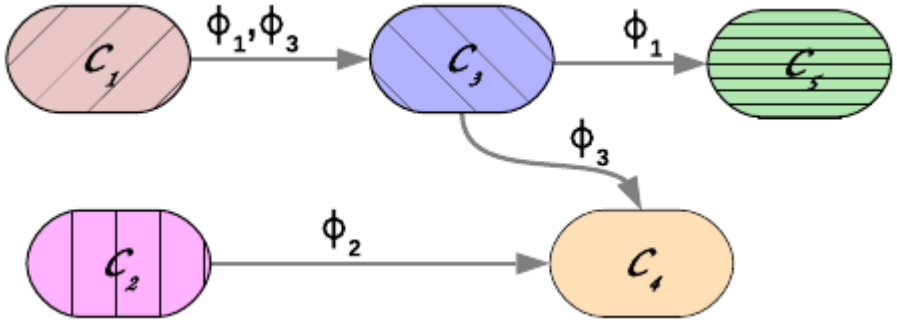


φ_1	t_1	x_1^d
	t_2	x_2^d
	t_3	x_3^d
φ_2	t_1	x_4^d
	t_2	x_5^d
	t_3	x_6^d

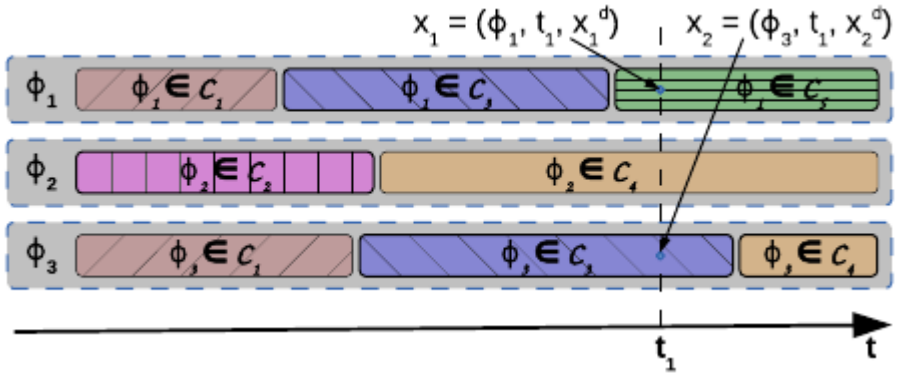
Goal:

- Detect typical evolution paths of individuals
- "Slow changing world" assumption
- Allow the relations between phases to emerge from the data

a) The evolution phases and their relational structure



b) Evolution paths → the trajectory of the entities through the different phases



Summary:

Proposed solution:

- A temporal-aware constrained clustering algorithm, resulted clusters serve as phases
- The relations between evolution phases inferred simultaneously with the partition

Objectives:

Obj. 1 Construct clusters which are coherent in the temporal and the descriptive space.

Obj. 2 Segment, as contiguously as possible, the series of observations for each entity.

Obj. 3 Present smooth passages between phases on evolution paths. i.e., changes should come in small increments.

Obj. 1 Construct clusters which are coherent in the temporal and the descriptive space.

- Use the temporal aware dissimilarity measure proposed in ***[Rizoiu et al '12]***

$$||x_i - x_j||_{TA} = 1 - \left(1 - \gamma_d \frac{||x_i^d - x_j^d||^2}{\Delta d_{max}^2}\right)\left(1 - \gamma_t \frac{||x_i^t - x_j^t||^2}{\Delta t_{max}^2}\right)$$

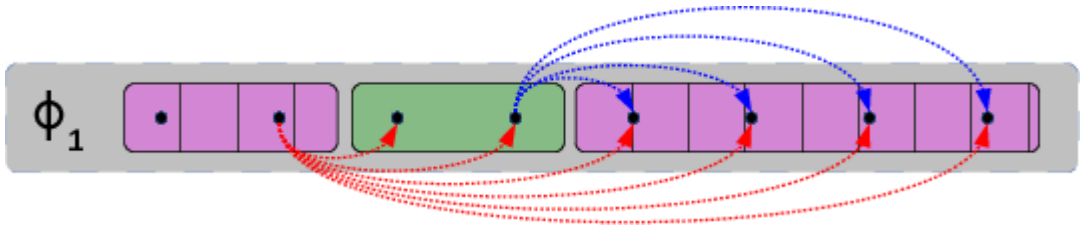
Properties:

- $||x_i - x_j||_{TA} \in [0,1], \forall\, x_i, x_j \in X$
- $||x_i - x_j||_{TA} = 0 \Leftrightarrow x_i^d = x_j^d \wedge x_i^t = x_j^t$
- $||x_i - x_j||_{TA} = 1 \Leftrightarrow ||x_i^d - x_j^d|| = \Delta x_{max} \vee |x_i^t - x_j^t| = \Delta t_{max}$

Obj. 2 Segment, as contiguously as possible, the series of observations for each entity.

- Use temporally-oriented soft must-link pair-wise constraints
- Extend the contiguity penalty function in ***[Rizoiu et al '12]***

$$w(x_i, x_k) = \beta * e^{-\frac{1}{2}\left(\frac{||x_i^t - x_k^t||}{\delta}\right)^2}\left(1 - a_{j,l}^2\right) \quad for\ x_i^\varphi = x_j^\varphi, x_i^t < x_k^t$$



Obj. 3 Present smooth passages between phases on evolution paths. i.e., changes should come in small increments.

- Evolution phases structured as an oriented graph

  $a_{p,q}$ strength of link between $\mathscr{C}_p$ and $\mathscr{C}_q$

- The strength of the link from $\mathscr{C}_p$ to $\mathscr{C}_q$ is proportional to the similarity of their prototypes

$$T_2 = \sum_{\mu_p \in \mathscr{M}} \sum_{\substack{\mu_q \in \mathscr{M} \\ p \neq q}} a_{p,q}^2 ||\mu_p - \mu_q||_{TA}.$$

- The strength of the link from $\mathscr{C}_p$ to $\mathscr{C}_q$ is dependent on the number of entities which present a transition from $\mathscr{C}_p$ to $\mathscr{C}_q$

$$T_3 = \sum_{\mu_p \in \mathscr{M}} \sum_{\substack{\mu_q \in \mathscr{M} \\ p \neq q}} a_{p,q}^2 inter_\phi^2(\mathscr{C}_p, \mathscr{C}_q). \quad inter_\phi(\mathscr{C}_p, \mathscr{C}_q) = 1 - \frac{|\{\phi_l \in \Phi | \mathscr{C}_p \xrightarrow{\phi_l} \mathscr{C}_q\}|}{|\Phi|}$$

The ClusPath algorithm:

Inspired from K-Means.

3 "ingredients": i) observations, ii) prototypes and iii) relations between clusters

Iterates 3 update phases:

- recompute prototypes
- assignments of observations to clusters
- recompute adjacency matrix

$$\mathcal{J} = \lambda_1 T_1 + \lambda_2 T_2 + \lambda_3 T_3$$

$$= \lambda_1 \sum_{\mu_p \in \mathcal{M}} \sum_{x_i \in \mathcal{C}_p} \left(||x_i - \mu_p||_{TA} + \sum_{\substack{x_k \in \mathcal{C}_q \\ q \neq p,\ x_i^\phi = x_k^\phi}}^{x_i^t < x_k^t} \beta * e^{-\frac{1}{2}\left(\frac{||x_i^t - x_k^t||}{\delta}\right)^2} \left(1 - a_{p,q}^2\right) \right)$$

$$+ \lambda_2 \sum_{\mu_p \in \mathcal{M}} \sum_{\substack{\mu_q \in \mathcal{M} \\ p \neq q}} a_{p,q}^2 ||\mu_p - \mu_q||_{TA} + \lambda_3 \sum_{\mu_p \in \mathcal{M}} \sum_{\substack{\mu_q \in \mathcal{M} \\ p \neq q}} a_{p,q}^2 inter_\phi^2(\mathcal{C}_p, \mathcal{C}_q),$$

Summary:

Four measures to evaluate a partition

- Descriptive and temporal coherence of clusters
  Classical Variance: MDvar and Tvar

- Contiguous segmentation **ShaP** *[Rizoiu et al '12]*

$$ShaP = \frac{1}{|\Phi|} \sum_{\phi \in \Phi} \sum_{i=1}^{k} \left[-p_\phi(\mathscr{C}_i) \log_2(p_\phi(\mathscr{C}_i)) \right] \left(1 + \frac{n_{ch} - n_{min}}{N - 1} \right), \; p_\phi(\mathscr{C}_i) = \sum_{\substack{x_j \in \mathscr{C}_i \\ x_j^\phi = \phi}} \frac{1}{N}$$

- Smooth passage of entities along an evolution path **SPass**

$$SPass = \sum_{\phi \in \Phi} \sum_{\mathscr{C}_i \xrightarrow{\phi} \mathscr{C}_j}^{i,j \in 1,\ldots,k} \frac{||\mu_i - \mu_j||_{TA}}{n_{ch}}$$

Parameter tuning heuristic: Use an evolutionary technique to approximate the Pareto front in the space of measures.

- Genome of individuals → the six parameters of ClusPath

$$\alpha, \beta, \delta, \lambda_1, \lambda_2, \lambda_3$$

- Approx. front in the space of the four measure of a partition

$$MDvar, Tvar, ShaP, SPass$$

- Elitist technique, crossover, mutation

- Choose parameters that output a balanced solution in the space of the four measures.

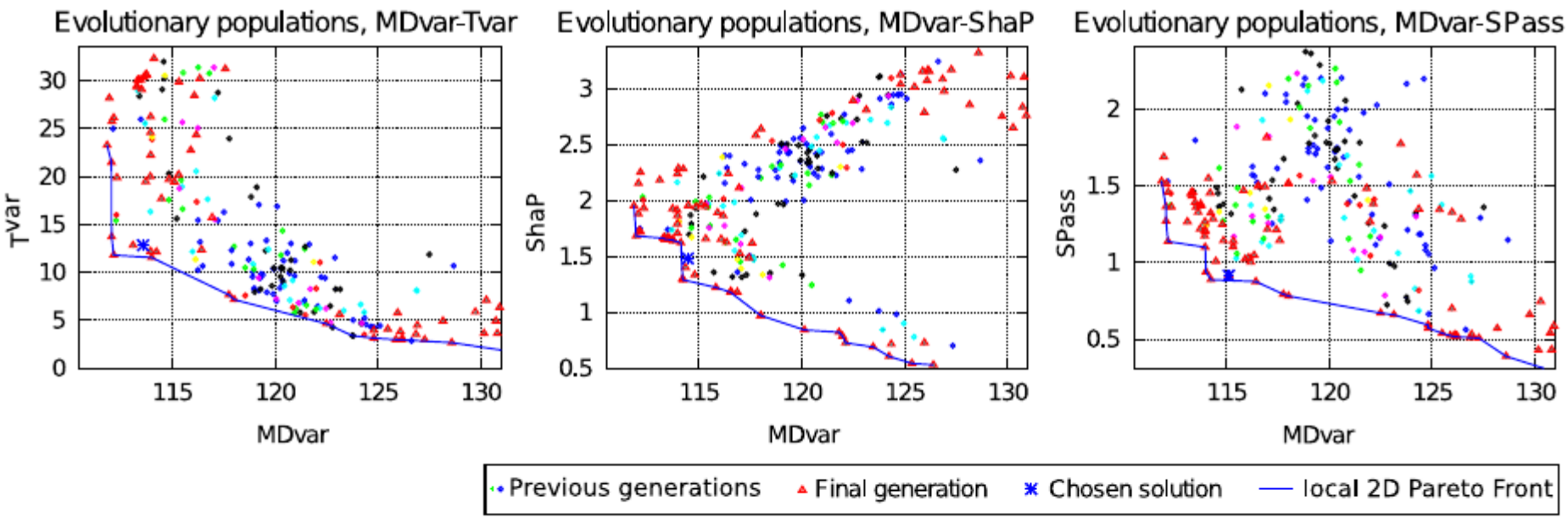
Summary:

Two datasets

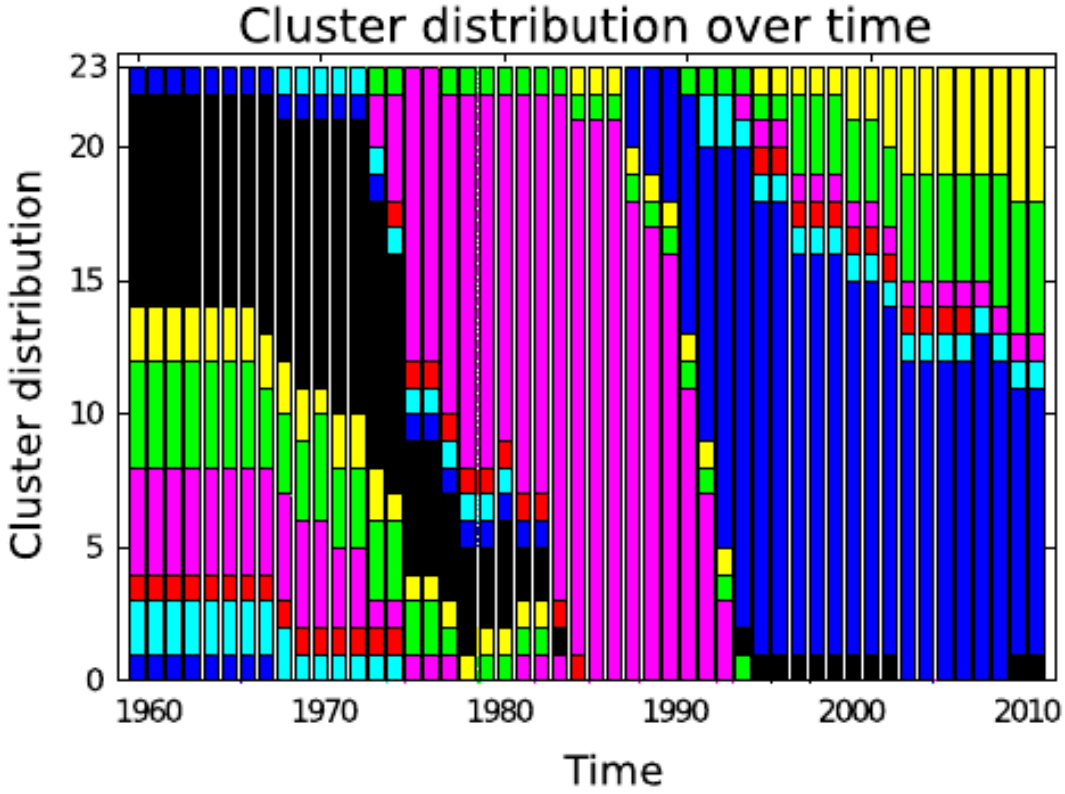
Compared Political Dataset I (CPDS) 23 countries, 60 years, 207 political, demographic, social and economic vars.

European Companies (EC) 836 companies, 5 years, 7 economic vars.

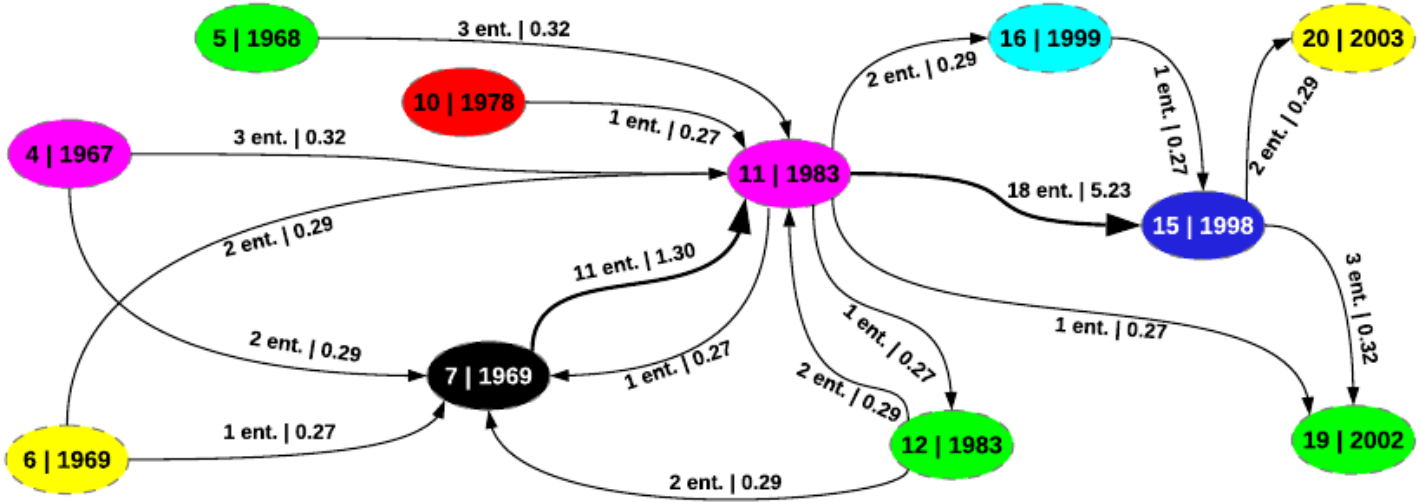
Setting parameters using the evolutionary technique



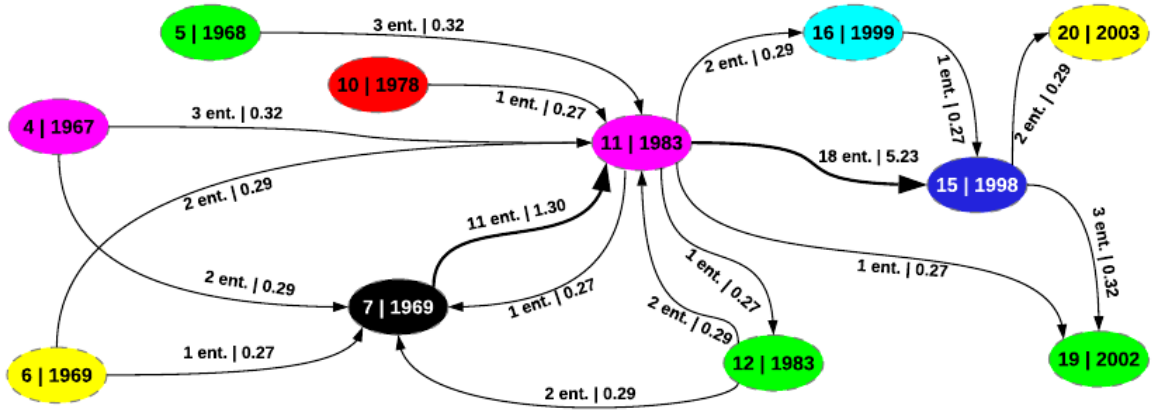
Results on CPDS

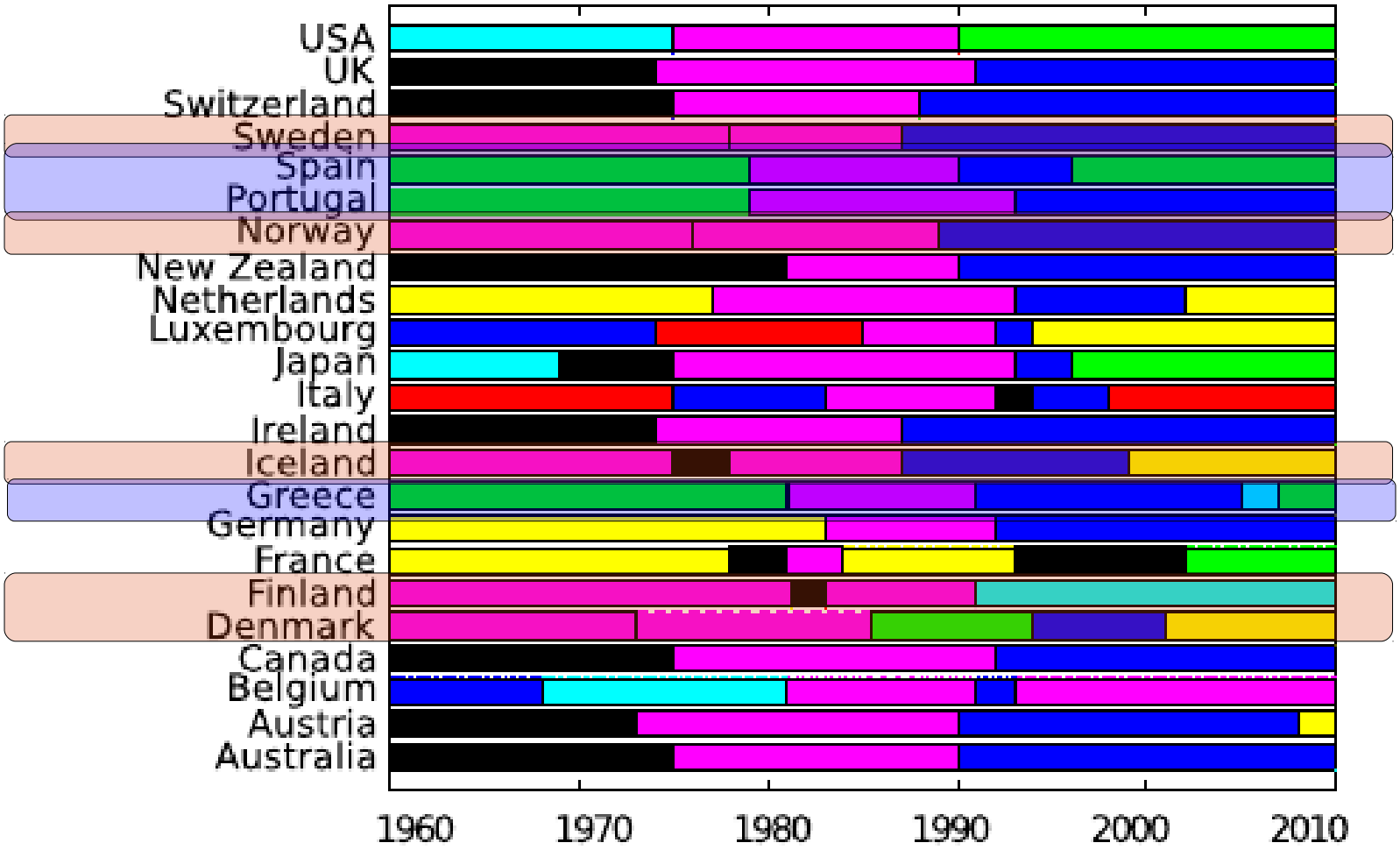


Evolution paths



Results on CPDS





Results on EC

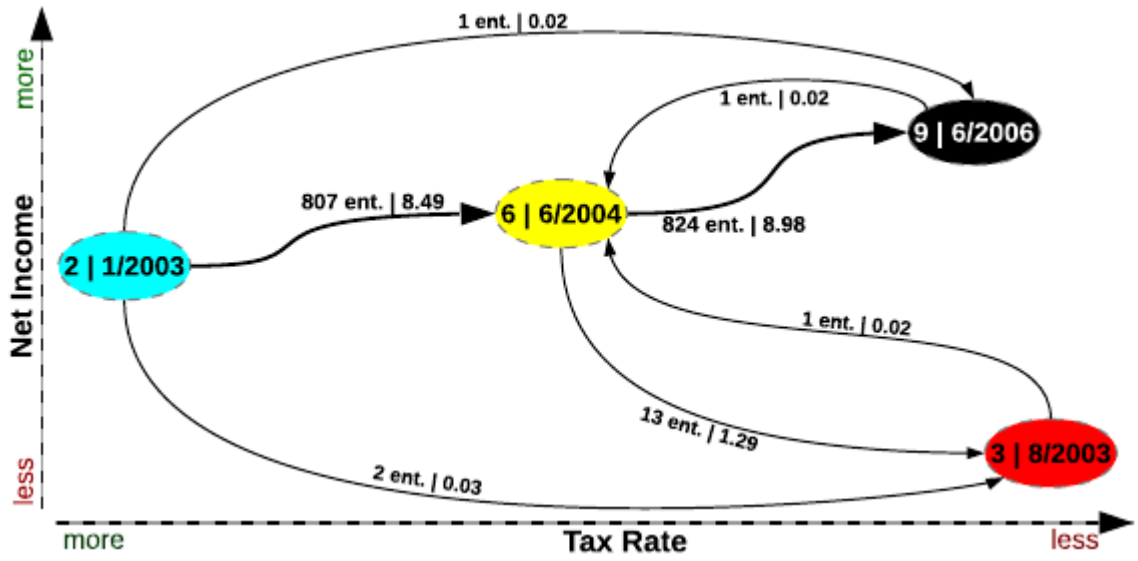


Table 1 Most common evolution phases in EC, described over the 7 dimensions of the dataset

Ph.	Time	FCFF	TotalDebt	Revenues	NetCapExp	EBITDA	TaxRate	NetIncome
$\mathscr{C}_2$	01/2003	−0.00	−0.01	−0.02	−0.00	−0.04	0.08	−0.09
$\mathscr{C}_3$	08/2003	−0.94	−0.06	−1.82	−0.67	−2.02	−0.07	−4.04
$\mathscr{C}_6$	06/2004	−0.01	−0.01	−0.02	−0.04	−0.02	−0.04	−0.04
$\mathscr{C}_9$	06/2006	0.05	0.01	0.07	0.04	0.07	−0.06	0.15

Quantitative evaluation

6 algorithms:

- K-Means *[MacQueen '67]*;
- Temporal-Driven K-Means *[Rizoiu et al '12]*;
- Constrained K-Means *[Rizoiu et al '12]*;
- tcK-Means *[Lin and Hauptmann '10];*
- TDCK-Means *[Rizoiu et al '12]*;
- ClusPath.

4 measures:

- MDvar
- Tvar
- ShaP
- SPass

ClusPath consistently obtains a better tradeoff between the four opposing measures.

Summary:

Thank you!

Conclusion:

- Studied the detection of typical evolutions paths, using a "slow changing word" assumption;
- The connections between evolution phases are inferred simultaneously with the clustering algorithm;
- An evolutionary technique to set parameters on unseen dataset, by searching a balanced tradeoff.

Perspectives:

- Automatic description of generated evolution phases (clusters);
- Use with other applications, e.g. in Computational Social Media (to find behavioral roles)