Mining Dynamic and Augmented Graphs

A Constraint-Based Pattern Mining View

Marc Plantevit

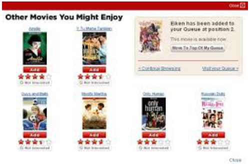
MEET THE INDUSTRY DAY,
UNIVERSITY-INDUSTRY WORKSHOP ON SYSTEMS BIOLOGY

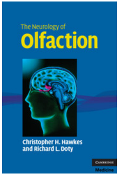
Data Mining and Mining (DM2L) Research Group
LIRIS UMR5205

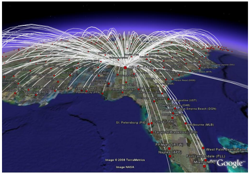
Data: a new “natural ressource”

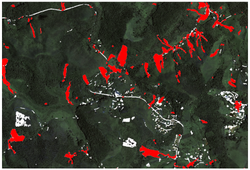
MEET THE INDUSTRY DAY
NOVEMBER 19TH 2014 · Villeurbanne, France

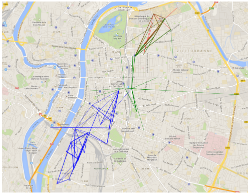
Potential increase of our knowledge



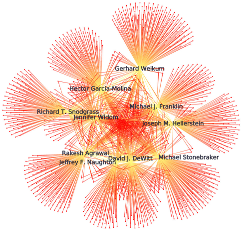


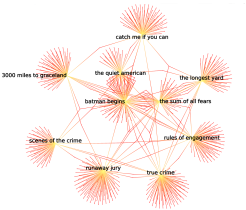


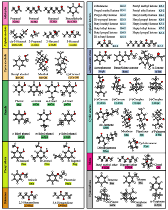


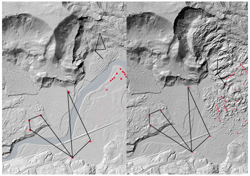


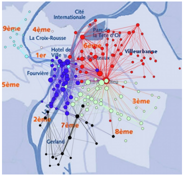
Viewed as augmented graphs











- Graphs are dynamic with attributes associated to vertices and/or edges.
- Generic techniques to understand the underlying mechanisms.

Mining augmented graphs

Network data brings several questions:

- Working with network data is messy
  - Not just "wiring diagrams" but also dynamics and data (features, attributes) on nodes and edges
- Computational challenges
  - Large scale network data
- Algorithmic models as vocabulary for expressing complex scientific questions
  - Social science, physics, biology, neuroscience

☛ Understanding how network structure and node attribute values relate and influence each other.

- **A constraint-based pattern mining view**

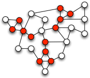
Constraint-based pattern mining view

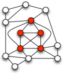
A (local) pattern φ describes a sub-group of the data $\mathcal{D}$

- observed several times
- or characterized by specific properties

The pattern shape is fixed: $\varphi \in \mathcal{L}$

- whose cardinality is exponential in the size of the data or infinite





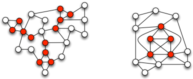
Constraint-based pattern mining view

A (local) pattern φ describes a sub-group of the data $\mathcal{D}$

- observed several times
- or characterized by specific properties

The pattern shape is fixed: $\varphi \in \mathcal{L}$

- whose cardinality is exponential in the size of the data or infinite



The constraints

$\mathcal{C}$ evaluates the adequacy of the pattern to the data

$$\mathcal{C}(\varphi, \mathcal{D}) \rightarrow \text{Boolean}$$

To express the interest of the end-user

- Taking into account the domain knowledge
- objective interest, statistical assessment

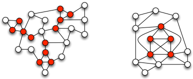
Constraint-based pattern mining view

A (local) pattern φ describes a subgroup of the data $\mathcal{D}$

- observed several times
- or characterized by specific properties

The pattern shape is fixed: $\varphi \in \mathcal{L}$

- whose cardinality is exponential in the size of the data or infinite



The constraints

$\mathcal{C}$ evaluates the adequacy of the pattern to the data

$$\mathcal{C}(\varphi, \mathcal{D}) \rightarrow \text{Boolean}$$

To express the interest of the end-user

- Taking into account the domain knowledge
- objective interest, statistical assessment

Pattern mining task: Find all interesting subgroups

$$Th(\mathcal{L}, \mathcal{D}, \mathcal{C}) = \{\varphi \in \mathcal{L} \mid \mathcal{C}(\varphi, \mathcal{D}) \text{ is true } \}$$

$Th(\mathcal{L}, \mathcal{D}, \mathcal{C})$ is an inductive query.

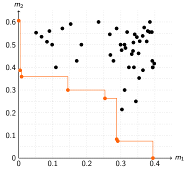
Fully taking into account user preferences

:-(A constraint ≡ some (too many) thresholds to set !!!

- A well-known issue in data mining that limits the full use of this paradigm

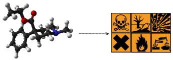
Let's see the constraints as preferences !

- Computing only the patterns that maximize the user preferences
- [Soulet et al., ICDM 2011]



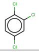
⇒ Skyline Analysis

to compute only the (sky)patterns that are pareto-dominant w.r.t. to the user's preferences.

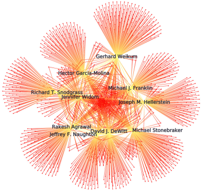


Case Study: Discovering Toxicophores

- Skypatterns are useful to discover toxicophores
- background knowledge can easily be integrated, adding aromaticity and density measures



Some inductive queries for augmented graphs

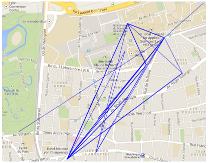


- What are the node attributes that strongly co-vary with the graph structure?
  - Co-authors that published at ICDE with a high degree and a low clustering coefficient.

  ✎ [Prado et al., IEEE TKDE 2013]

- What are the sub-graphs whose node attributes evolve similarly?
  - Airports whose arrival delays increased over the three weeks following Katrina hurricane

  ✎ [Desmier et al., ECMLPKDD 2013]



- For a given population, what is the most related subgraphs (i.e., behavior)? For a given subgraph, which is the most related subpopulation?
  - People born after 1979 are over represented on the campus.

MEET THE INDUSTRY DAY
NOVEMBER 19TH 2014 · Villeurbanne, France

Talk Outline

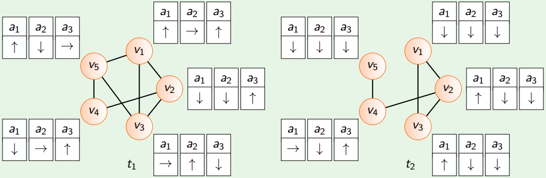
MEET THE INDUSTRY DAY

Dynamic Attributed Graphs

A dynamic attributed graph $\mathcal{G} = (\mathcal{V}, \mathcal{T}, \mathcal{A})$ is a sequence over $\mathcal{T}$ of attributed graphs $G_t = (\mathcal{V}, E_t, A_t)$, where:

- $\mathcal{V}$ is a set of vertices that is fixed throughout the time,
- $E_t \in \mathcal{V} \times \mathcal{V}$ is a set of edges at time t,
- A_t is a vector of numerical values for the attributes of $\mathcal{A}$ that depends on t.

Example

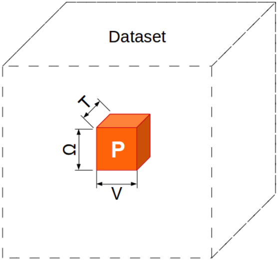


MEET THE INDUSTRY DAY

Co-evolution Pattern

Given $\mathcal{G} = (\mathcal{V}, \mathcal{T}, \mathcal{A})$, a co-evolution pattern is a triplet $P = (V, T, \Omega)$ s.t.:

- $V \subseteq \mathcal{V}$ is a subset of the vertices of the graph.
- $T \subset \mathcal{T}$ is a subset of not necessarily consecutive timestamps.
- Ω is a set of signed attributes, i.e., $\Omega \subseteq A \times S$ with $A \subseteq \mathcal{A}$ and $S = \{+, -\}$ meaning respectively a $\{increasing, decreasing\}$ trend.

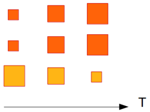


Predicates

A co-evolution pattern must satisfy two types of constraints:

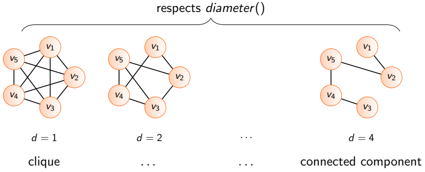
Constraint on the evolution:

- Makes sure attribute values co-evolve
- We propose δ-**strictEvol**.
- $\forall v \in V$, $\forall t \in T$ and $\forall a^s \in \Omega$ then $\delta\text{-}trend(v,t,a) = s$



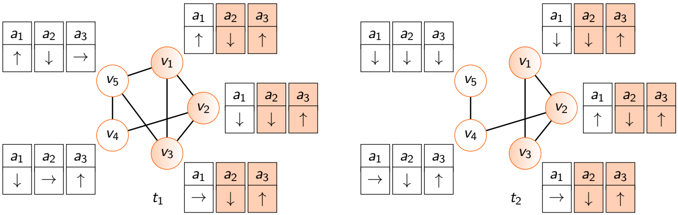
Constraint on the graph structure:

- Makes sure vertices are related through the graph structure.
- We propose **diameter**.
- $\Delta\text{-}diameter(V, T, \Omega) = true \Leftrightarrow \forall t \in T \; diam_{G_t(V)} \leq \Delta$

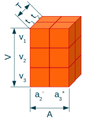


Example

$$P = \{(v_1, v_2, v_3)(t_1, t_2)(a_2^-, a_3^+)\}$$



- 1-Diameter(P) is true,
- 0-strictEvol(P) is true.

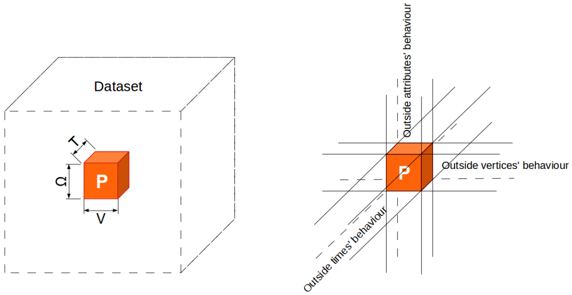


MEET THE INDUSTRY DAY NOVEMBER 19TH 2014

Density Measures

Intuition

Discard patterns that depict a behaviour supported by many other elements of the graph. We propose : **vertex specificity**, **temporal dynamic** and **trend relevancy**.

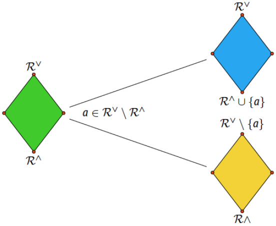


MEET THE INDUSTRY DAY NOVEMBER 19TH 2014 Villeurbanne, France

Algorithm

How to use the properties of the constraints to reduce the search space?

- Binary enumeration of the search space.
- Using the properties of the constraints to reduce the search space
  - Monotone, anti-monotone, piecewise (anti-)monotone, etc.
- Constraints are fully or partially pushed:
  - to prune the search space (i.e., stop the enumeration of a node),
  - to propagate among the candidates.



✎[Cerf et al, ACM TKDD 2009]

☛Our algorithms aim to be complete but other heuristic search can be used in a straightforward way (e.g., beam-search) to be more scalable


Top temporal_dynamic trend dynamic sub-graph (in red)

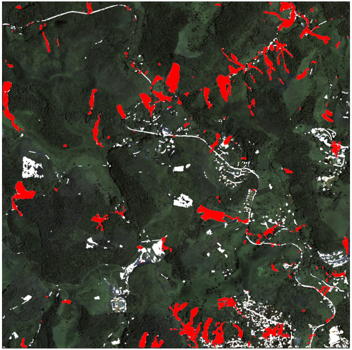
- 71 airports whose arrival delays increase over 3 weeks.
- *temporal_dynamic* = 0, which means that arrival delays never increased in these airports during another week.
- The hurricane strongly influenced the domestic flight organization.

	$\|V\|$	$\|T\|$	$\|A\|$	density
Katrina	280	8	8	5×10^{-2}

Top trend_relevancy (Yellow)

- 5 airports whose number of departures and arrivals increased over the three weeks following Katrina hurricane.
- *trend_relevancy* value equal to 0.81
- Substitutions flights were provided from these airports during this period.
- This behavior is rather rare in the rest of the graph

Brazil landslides

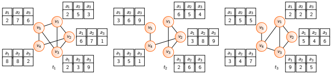


	$\|V\|$	$\|T\|$	$\|A\|$	density
Brazil landslide	10521	2	9	0.00057

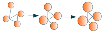
Discovering lanslides

- Taking into account expert knowledge, focus on the patterns that involve NDVI^{+}.
- Regions involved in the patterns: true landslides (red) and other phenomena (white).
- Compare to previous work, much less patterns to characterize the same phenomena (4821 patterns vs millions).

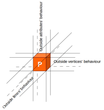
Overview of our proposal



Co-evolution patterns



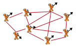
Interestingness Measures

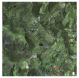


(Desmier et al., ECML/PKDD 2013)

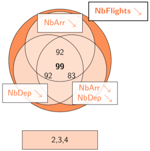
Experimental results

DBLP US flights Brazil landslides

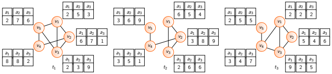




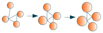
- Some obvious patterns are discarded ...
- ... but some patterns need to be generalized



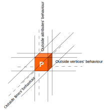
Overview of our proposal



Co-evolution patterns



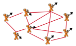
Interestingness Measures



(Desmier et al., ECML/PKDD 2013)

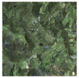
Experimental results

DBLP

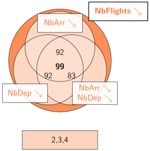


US flights

Brazil landslides



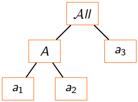
- Some obvious patterns are discarded ...
- ... but some patterns need to be generalized



Hierarchical co-evolution patterns

Take benefits from a hierarchy over the vertex attributes to :

- return a more concise collection of patterns;
- discover new hidden patterns;



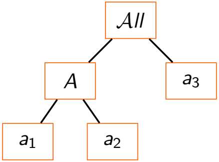
MEET THE INDUSTRY DAY
NOVEMBER 19TH 2014 · Villeurbanne, France

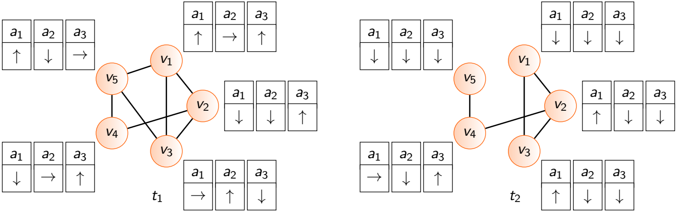
Talk Outline

Hierarchy

A hierarchy $\mathcal{H}$ on $\mathcal{A}$ is a tree where:

- the edges are a relation is_a,
- the node $\mathcal{A}ll$ is the root of the tree,
- the leaves are attributes of $\mathcal{A}$,
- dom($\mathcal{H}$) is all the nodes except the root.



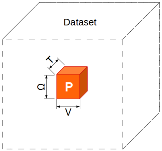


MEET THE INDUSTRY DAY NOVEMBER 19TH 2014 Villeurbanne, France

Hierarchical co-evolution Patterns

Given $\mathcal{G} = (\mathcal{V}, \mathcal{T}, \mathcal{A})$ and $\mathcal{H}$, a hierarchical co-evolution pattern is a triplet $P = (V, T, \Omega)$ s.t.:

- $V \subseteq \mathcal{V}$ is a subset of the vertices of the graph.
- $T \subset \mathcal{T}$ is a subset of not necessarily consecutive timestamps.
- Ω is a set of signed attributes, i.e., $\Omega \subseteq A \times S$ with $A \subseteq dom(\mathcal{H})$ and $S = \{+, -\}$ meaning respectively a $\{increasing, decreasing\}$ trend.

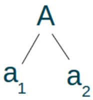


It must respect the following constraints:

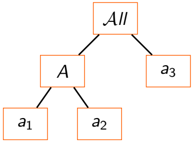
1. Constraint on the evolution.
2. Constraint on the graph structure.

Evolution Constraint

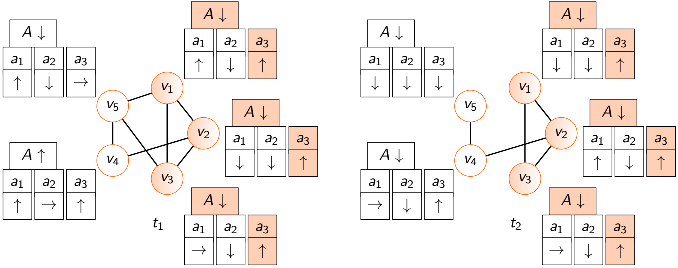
For an attribute A, its evolution is computed from the evolution of the leaves it covers.



Example



$$P = \{(v_1, v_2, v_3)(t_1, t_2)(A^-, a_3^+)\}$$

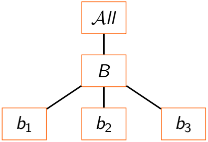


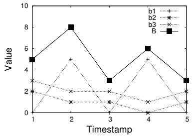
- 1-Diameter(P) is true,
- 0-strictEvolHierarchical(P) is true.

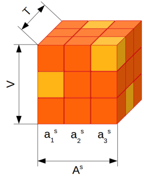
Purity of the pattern

Is the pattern described with the good level of granularity?

Purity computes the proportion of valid triplet (v, t, a^s) with regard to the number of possible triplets.



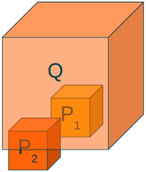




$$purity(P) = \frac{\sum_{v \in V} \sum_{t \in T} \sum_{a^s \in leaf(\Omega)} \delta_{a^s(v,t)}}{|V| \times |T| \times |leaf(\Omega)|}$$

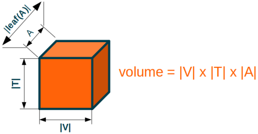
Use of hierarchies does not impact other measures/constraints

✔ **Maximality:**

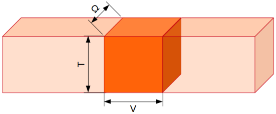


✔ **Size measures:**

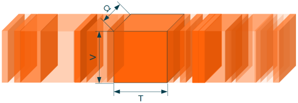
- $|leaf(A)| \geq min_A$,



✔ **Vertex specificity:**



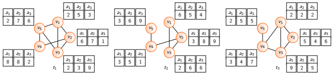
✔ **Temporal dynamicity:**



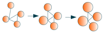
✘ **No trend relevancy** with hierarchies.

- What level of hierarchy do we consider?
- What about attributes discarded because of a too small purity gain?

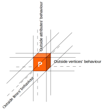
Overview



Co-evolution patterns



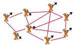
Interestingness Measures

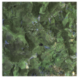


(Desmier et al., ECML/PKDD 2013)

Experimental results

DBLP US flights Brazil landslides

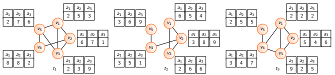




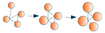
- Some obvious patterns are discarded ...
- ... but some patterns need to be generalized ✓
- [Desmier et al, IDA 2014]
- Difficulties to set parameters.

MEET THE INDUSTRY DAY
NOVEMBER 19TH 2014 · Villeurbanne, France

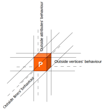
Overview



Co-evolution patterns



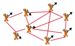
Interestingness Measures



(Desmier et al., ECML/PKDD 2013)

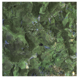
Experimental results

DBLP



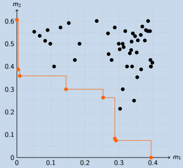
US flights

Brazil landslides



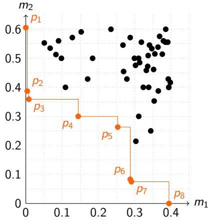
- Some obvious patterns are discarded ...
- ... but some patterns need to be generalized ✓
- [Desmier et al, IDA 2014]
- Difficulties to set parameters.

⇒ Skyline Analysis



Skyline analysis

The skyline operator returns all the skypatterns:



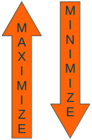
$$sky(\mathcal{P}, M) = \{P \in \mathcal{P} | \nexists Q \in \mathcal{P} \text{ s.t. } Q \succ_M P\}$$

$Q \succ_M P$ iff:

- Q is better (i.e., more preferred) than P in at least one measure,
- Q is not worse than P on every other measure.

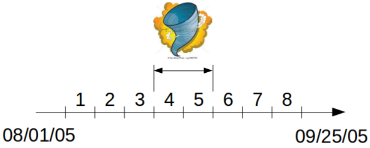
We propose to discover skypatterns considering a multidimensional space composed with a subset of the measures:

- sizeV, sizeT, sizeA
- volume
- purity

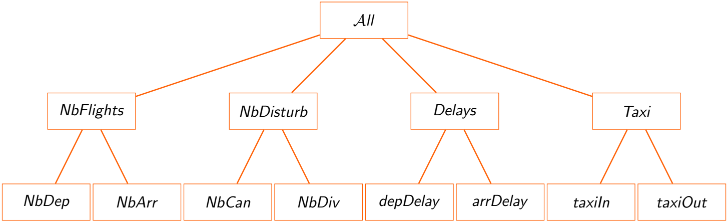


- vertexSpecificity
- temporalDynamic

US flights datasets



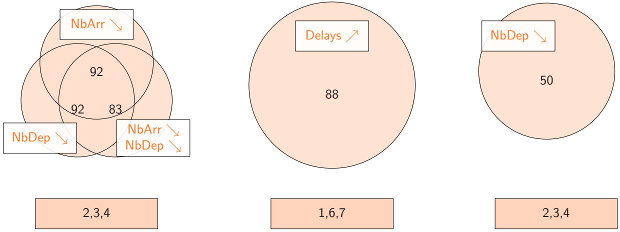
- Vertices: 280 airports.
- Times: 8 weeks around the Katrina hurricane.
- Attributes: number of departure/arrival/cancelled/deviated flights, departure/arrival delays and ground times.

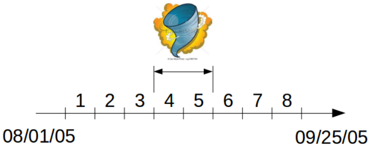


RITA "On-Time Performance" database.
(http://www.transtats.bts.gov)

Hierarchy impact

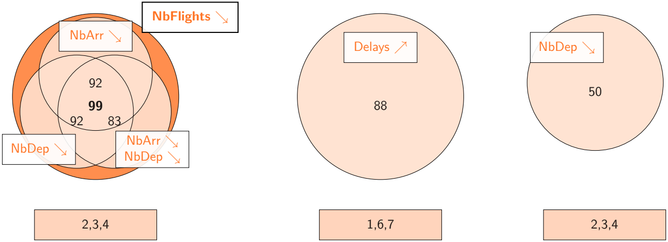
- 2 experiments with and without a hierarchy,
- Thresholds: min_V=40, min_T=min_A=ϑ=1, ψ=0.9, κ=0.2, τ=0.4.

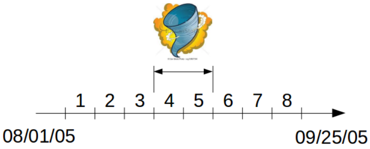




Hierarchy impact

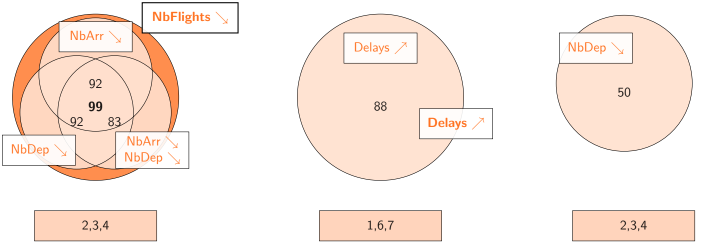
- 2 experiments with and without a hierarchy,
- Thresholds: min_V=40, min_T=min_A=ϑ=1, ψ=0.9, κ=0.2, τ=0.4.

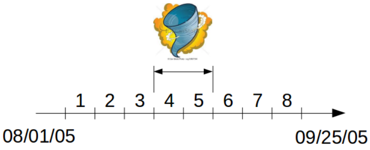




Hierarchy impact

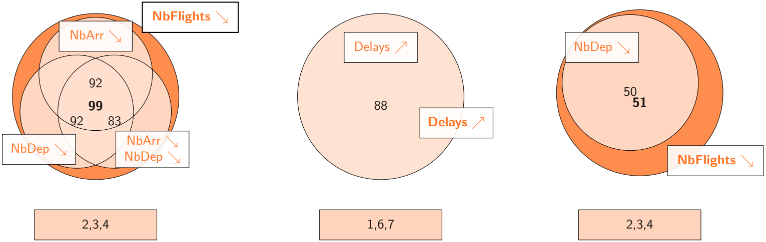
- 2 experiments with and without a hierarchy,
- Thresholds: min_V=40, min_T=min_A=ϑ=1, ψ=0.9, κ=0.2, τ=0.4.

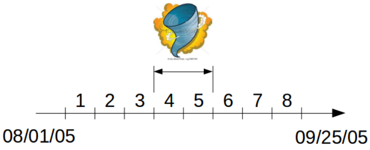




Hierarchy impact

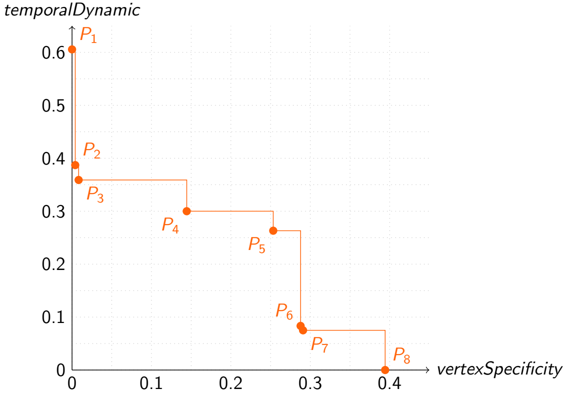
- 2 experiments with and without a hierarchy,
- Thresholds: min_V=40, min_T=min_A=ϑ=1, ψ=0.9, κ=0.2, τ=0.4.



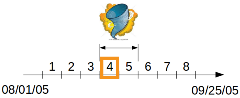


Qualitative experiments: Using skyline analysis

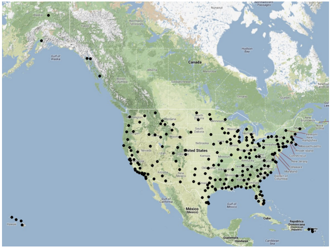
- $\vartheta = min_V = 5$, $min_T = min_A = 1$, ψ=0.9
- Skyline dimensions: *VS*, *TD*

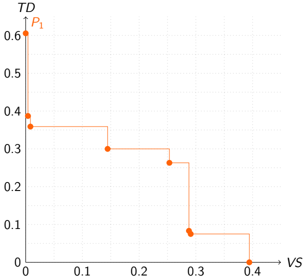


Qualitative experiments: Using skyline analysis



	$\|V\|$	T	A	purity	VS	TD
P_1	213	4	nbFlights$^-$	0.96	0	0.61





- This behavior is not followed by another node (airport) at this timestamp.

MEET THE INDUSTRY DAY
NOVEMBER 19TH 2014 · Villeurbanne, France

Talk Outline

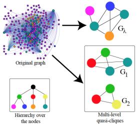
1. Co-evolution patterns in dynamic attributed graphs
2. Extensions to hierarchies and skyline analysis
3. Conclusion

(dynamic) Augmented graphs:

- A powerful mathematical abstraction that makes possible to depict many phenomena
- We have to define a large variety of inductive queries:
  - to focus on the evolution (of the attributes, the graph structure),
  - to take into account the intrinsic richness of the edges and the nodes.
  - [Pitarch et al, ASONAM 2014]: triggering attributes.

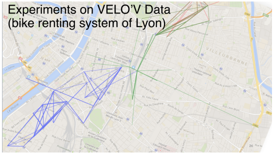
Multi-level graphs

- find all dense multi-level graphs
- hypothesis elicitation (rare diseases), clustering



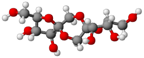
Contextualized trajectories

- Find subgraphs that are specific to a subpopulation
- recommendation, link prediction.



3D graphs

- Are there some 3D configurations specific to a class?
- hypothesis elicitation (olfaction)



MEET THE INDUSTRY DAY NOVEMBER 19TH 2014 · Villeurbanne, France

Skyline analysis to support more interaction

Skypattern mining is particularly well suited to interactive research:

- it proposes a *reduced collection* of patterns to the data expert which can quickly analyze it.

- Integration of the user feedbacks to make to foster iterative and interactive process.
  - refining the dominance relation;
  - computing the cube of all possible measures;
  - the skypattern cube exploration will provide a better understanding of the impact of the measures on the problem at hand;
  - Removing some uninteresting skypatterns and recompute the local changes;

A challenging issue, especially with augmented graphs!

Thank you for your attention.