Mining Skypatterns in Fuzzy Tensors

dupe hacks in the sky

Nicolas Nadisic, Aurélien Coussat, Loïc Cerf

September 19th, 2019 — ECMLPKDD 2019

Universidade Federal de Minas Gerais

Motivating example

1,198,292 spectators watching 92 Twitch channels during 19 weeks.

week	channel	user	time (hours)
10	90stardust	alelau18	1.35
9	zombiegrub	starfire1986	0.01
4	anosssc2	scalpl	0.04
4	anosssc2	vargase_lam	2.73
4	hazzardous1984	aarror	0.50
5	japanesports	zzz87012	0.04
5	rrb115	j30723	2.31

Objective: Find groups of users showing interest for groups of channels during certain weeks.

Motivating example

1,198,292 spectators watching 92 Twitch channels during 19 weeks.

week	channel	user	interest degree
10	90stardust	alelau18	0.874297
9	zombiegrub	starfire1986	0.00483032
4	anosssc2	scalpl	0.00552462
4	anosssc2	vargase_lam	0.999913
4	hazzardous1984	aarror	0.0645855
5	japanesports	zzz87012	0.00549996
5	rrb115	j30723	0.999171

Objective: Find groups of users showing interest for groups of channels during certain weeks.

What is a pattern?

	egg	ice cream	water	wine	yogurt
Alice	0.5	0.3	0	1	1
Bob	0.9	0.2	1	0	0
Carol	1	0.1	0	0.4	0.8
Dave	1	0	0.9	0.6	0

A fuzzy matrix.

- A subset of elements from every dimension.
- Noise-Tolerance: one ϵ_i per dimension.
- Ex $\epsilon_{\text{row}} = 0.8$, $\epsilon_{\text{col}} = 0.6$.
- Every ϵ_i upper-bounds the absolute amount of noise in every "slice" of the pattern that relates to the elements in its i^{th} dimension.

Problem

- The number of patterns exponentialy grows with the size of the dataset
- The computing time too

Problem

- The number of patterns exponentialy grows with the size of the dataset
- The computing time too

- We want only the *best* patterns
- We want to extract them fast

Measures

	egg	ice cream	water	wine	yogurt
Alice	0.5	0.3	0	1	1
Bob	0.9	0.2	1	0	0
Carol	1	0.1	0	0.4	0.8
Dave	1	0	0.9	0.6	0

- A measure is a function that take a pattern at input and returns a real number, the *score*.
- "The higher the better".
- Ex: frequency, area, growth-rate, utility...

Pattern $(\{\texttt{Alice}, \texttt{Carol}\}, \{\texttt{egg}, \texttt{wine}, \texttt{yogurt}\})$:
freq $= 2$, area $= 6$, growth rate $= +\infty$ if $C = \{\texttt{Alice}, \texttt{Carol}\}$.

Problem

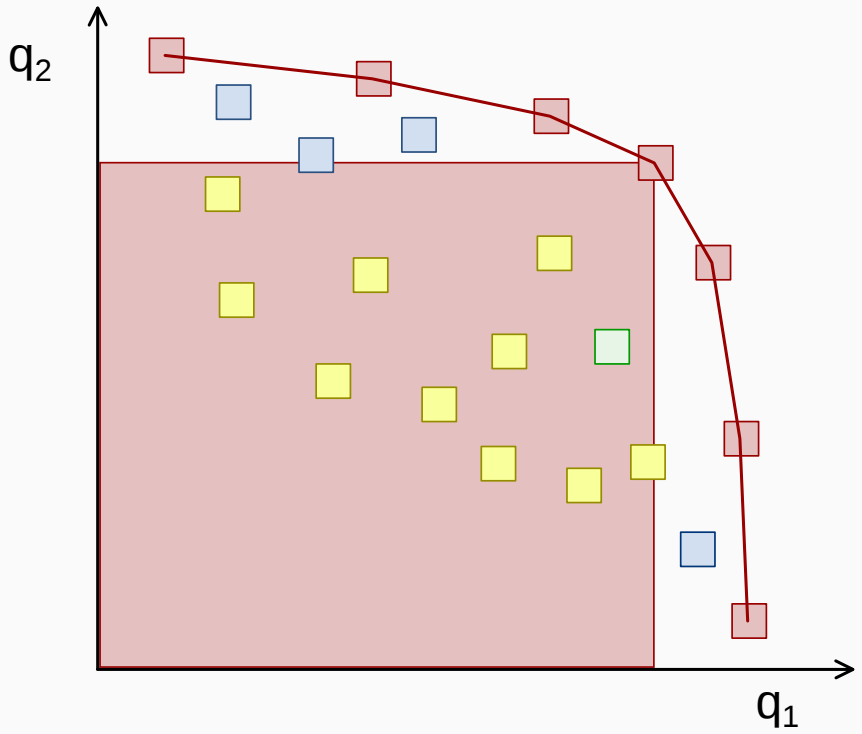
- Classic approaches use thresholds on measures.
- Hard to tune.
- We don't want patterns satisfying constraints, we want the *best* ones!

Problem

- Classic approaches use thresholds on measures.
- Hard to tune.
- We don't want patterns satisfying constraints, we want the *best* ones!

- How to define the *best* pattern with various measures?

Pareto domination



- A pattern dominates another if it scores at least as well on all measures, and strictly better on one measure.
- Pareto-optimal = Non-dominated.
- Pareto optimal set = Skyline (Börzsöny *et al.*, 2001).

Dominating patterns

	egg	ice cream	water	wine	yogurt
Alice	0.5	0.3	0	1	1
Bob	0.9	0.2	1	0	0
Carol	1	0.1	0	0.4	0.8
Dave	1	0	0.9	0.6	0

The blue pattern dominates the red one:

	freq	area	gr
$(\{\texttt{Alice}, \texttt{Carol}\}, \{\texttt{egg}, \texttt{wine}, \texttt{yogurt}\})$	2	6	$+\infty$
$(\{\texttt{Bob}, \texttt{Dave}\}, \{\texttt{egg}, \texttt{water}\})$	2	4	0

Also, the blue pattern is non-dominated: it is a skypattern.

Problem

Problem statement

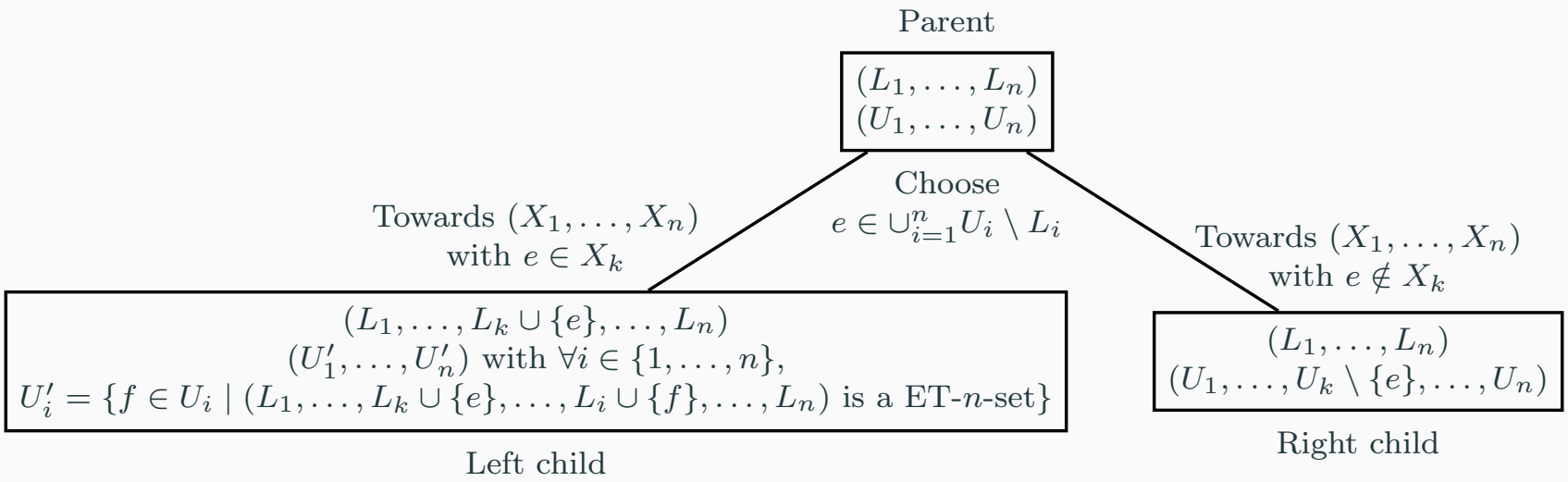
Given a fuzzy tensor, noise-tolerance thresholds, and a set of measures to optimize, find all the patterns on the skyline: the skypatterns.

Problem

Problem statement
Given a fuzzy tensor, noise-tolerance thresholds, and a set of measures to optimize, find all the patterns on the skyline: the skypatterns.

- How to make the mining fast?
- How to prune the pattern space?

`multidupehack` (Cerf and Meira Jr., ICDE'14)



Pattern space traversal by `multidupehack`

- The pattern space is span by an enumeration tree.
- A node defines a subspace of patterns, with a lower bound L and an upper bound U.
- Initialy, $L = \emptyset$ and $U =$ all elements from all dimensions
- When $L = U$, we have a leaf: a potential pattern
- Exact algorithm!

Lower and upper bounds, optimism

L

	egg	ice cream	water	wine	yogurt
Alice	0.5	0.3	0	1	1
Bob	0.9	0.2	1	0	0
Carol	1	0.1	0	0.4	0.8
Dave	1	0	0.9	0.6	0

U

	egg	ice cream	water	wine	yogurt
Alice	0.5	0.3	0	1	1
Bob	0.9	0.2	1	0	0
Carol	1	0.1	0	0.4	0.8
Dave	1	0	0.9	0.6	0

- To prune a subtree, we need an optimistic evaluation of measures

Pruning a branch

L

	egg	ice cream	water	wine	yogurt
Alice	0.5	0.3	0	1	1
Bob	0.9	0.2	1	0	0
Carol	1	0.1	0	0.4	0.8
Dave	1	0	0.9	0.6	0

U

	egg	ice cream	water	wine	yogurt
Alice	0.5	0.3	0	1	1
Bob	0.9	0.2	1	0	0
Carol	1	0.1	0	0.4	0.8
Dave	1	0	0.9	0.6	0

(L, U): freq $= 2$, area $= 6$, gr $= \frac{1}{2}$

Pruning a branch

L

	egg	ice cream	water	wine	yogurt
Alice	0.5	0.3	0	1	1
Bob	0.9	0.2	1	0	0
Carol	1	0.1	0	0.4	0.8
Dave	1	0	0.9	0.6	0

U

	egg	ice cream	water	wine	yogurt
Alice	0.5	0.3	0	1	1
Bob	0.9	0.2	1	0	0
Carol	1	0.1	0	0.4	0.8
Dave	1	0	0.9	0.6	0

$A \subset Skyline$

	egg	ice cream	water	wine	yogurt
Alice	0.5	0.3	0	1	1
Bob	0.9	0.2	1	0	0
Carol	1	0.1	0	0.4	0.8
Dave	1	0	0.9	0.6	0

(L, U): freq $= 2$, area $= 6$, gr $= \frac{1}{2}$ and A: freq $= 2$, area $= 6$, gr $= +\infty$: safe pruning!

Piecewise (anti-)monotonicity

How to evaluate measures on (L, U)?

Example (Area)
$(X_1, X_2) \mapsto |X_1 \times X_2| \Leftrightarrow (L_1, L_2, U_1, U_2) \mapsto |U_1 \times U_2|$

Example (Growth-rate to a subset $C \subseteq D_1$)
$$(X_1, X_2) \mapsto \frac{|D_1 \setminus C| \times |X_1 \cap C|}{|C| \times |X_1 \setminus C|} \Leftrightarrow (L_1, L_2, U_1, U_2) \mapsto \frac{|D_1 \setminus C| \times |U_1 \cap C|}{|C| \times |L_1 \setminus C|}$$

Skypatterns: state of the art

- Only manage 0/1 matrices.

Skypatterns: state of the art

- Only manage 0/1 matrices.

- "Mining dominant patterns in the sky", Soulet *et al.*, 2011.
- Generalization of the skyline query to itemset mining.
- Motivation: optimize various measures, no need to define thresholds.
- `Aetheris` algorithm: post-process approach.

Skypatterns: state of the art

- Only manage 0/1 matrices.

- "Mining dominant patterns in the sky", Soulet *et al.*, 2011.
- Generalization of the skyline query to itemset mining.
- Motivation: optimize various measures, no need to define thresholds.
- `Aetheris` algorithm: post-process approach.

- "Dominance programming for itemset mining",
  Negrevergne *et al.*, 2013: `DP` algorithm.
- "Mining (Soft-) Skypatterns Using Dynamic CSP",
  Ugarte *et al.*, 2015: `CP+SKY` algorithm.

Comparative study: Experimental protocol

- Comparison with `Aetheris`, `DP`, `CP+SKY`.
- 23 UCI datasets (0/1 matrices).
- 2 sets of measures:
  - frequency and area.
  - frequency, area, and growth-rate over all classes.
- All algorithms output the same skypatterns.
- Timeout of 2 hours.

Comparative study

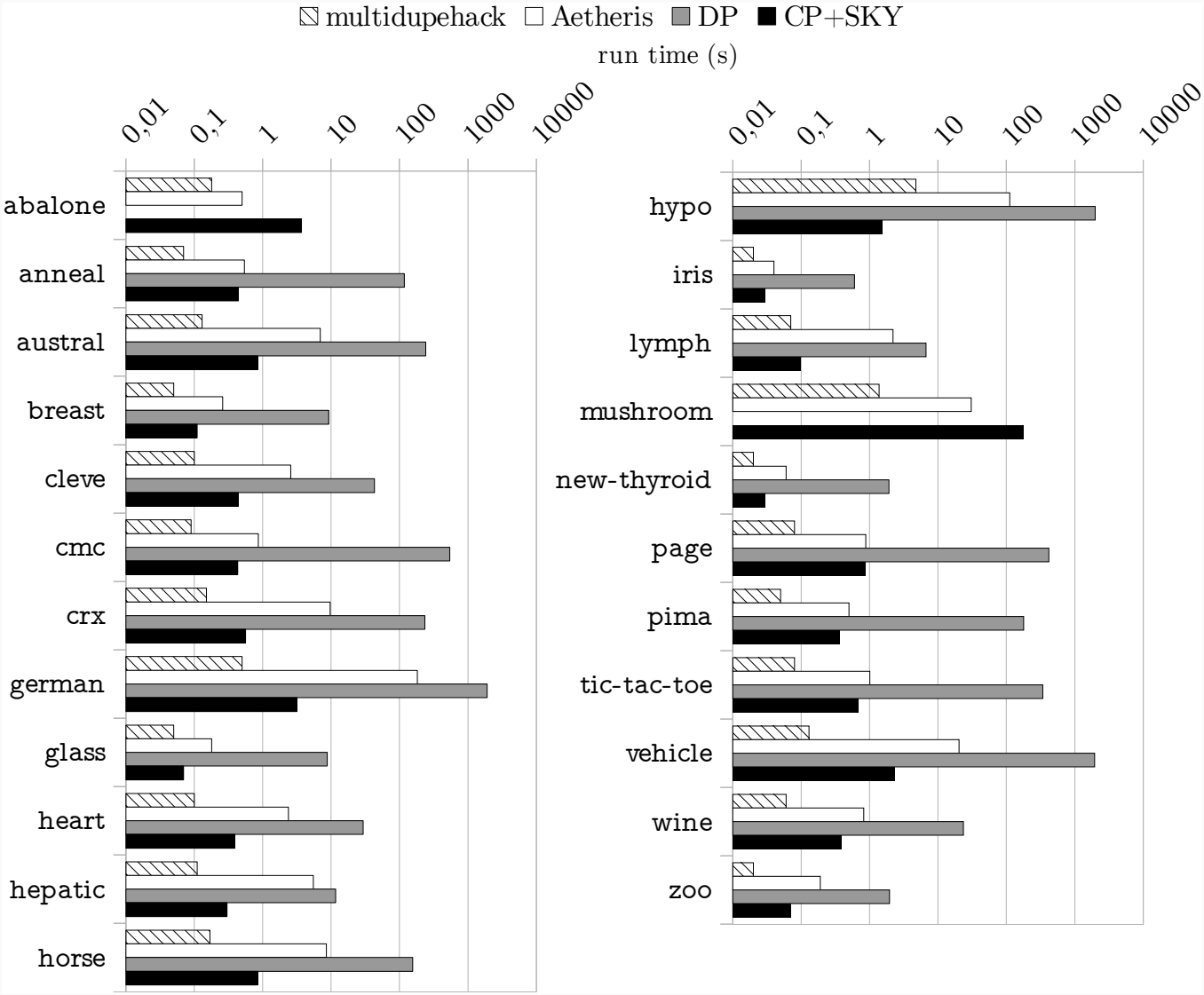


Figure 1: Run times to extract skypatterns maximizing frequency and area

Comparative study

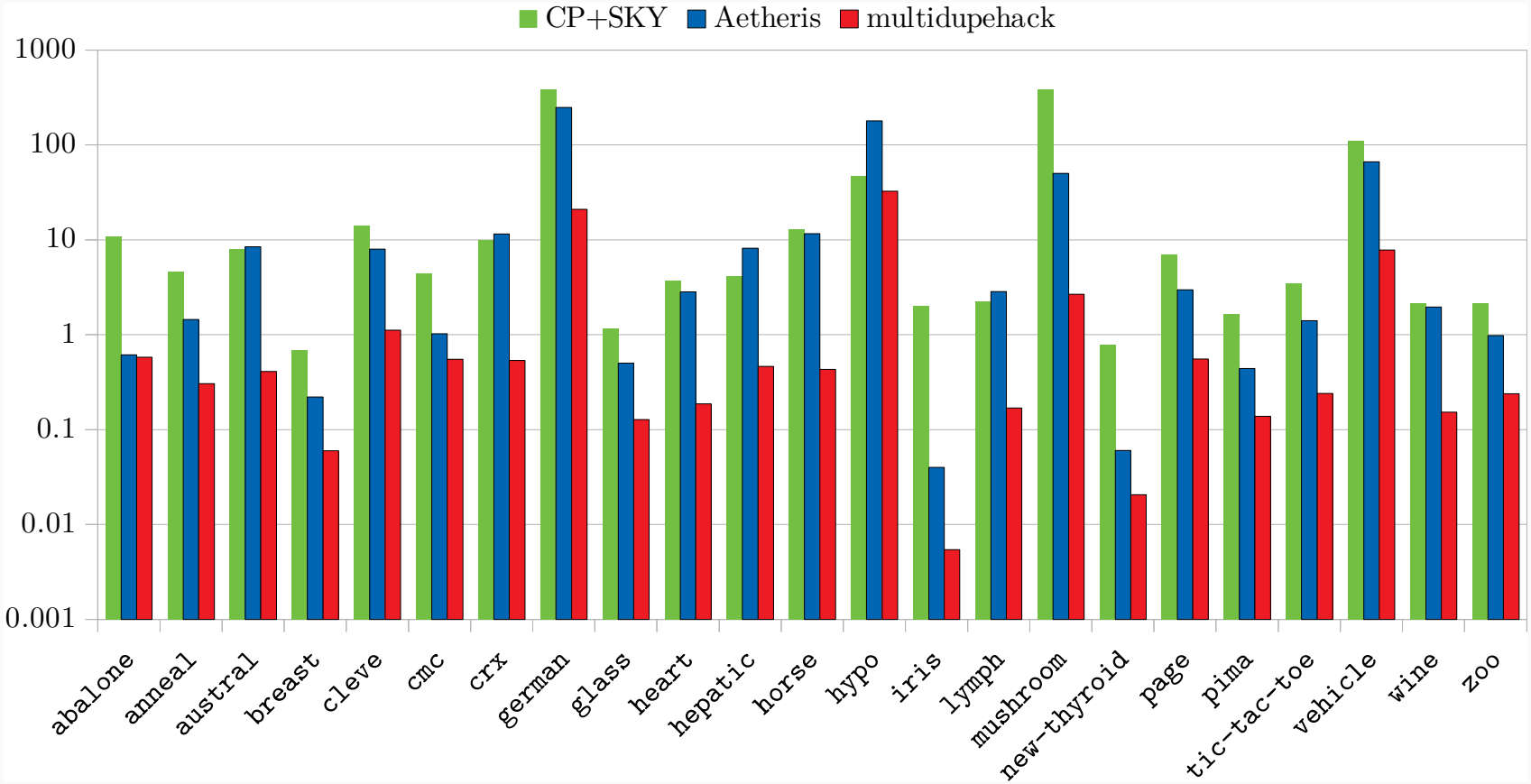


Figure 2: Run times to extract skypatterns maximizing frequency, area and growth-rate

Comparative study: Results

For the set frequency-area:

- `multidupehack` > `CP+SKY` > `Aetheris` > `DP`.
- On average, `multidupehack` is 11 × faster than `CP+SKY` and 42 × faster than `Aetheris`.

For the set frequency-area-growth-rate:

- `multidupehack` > `Aetheris` > `CP+SKY`.
- On average, `multidupehack` is 10 × faster than `Aetheris` and 36 × faster than `CP+SKY`.

The Twitch dataset

- Twitch: video games streaming website
- 3-way fuzzy tensor
- Connection and disconnection times of 1,198,292 spectators, watching 92 channels related to Starcraft 2, during 19 weeks.

Example of data:

```
10 90stardust alelau18 0.874297
9 zombiegrub starfire1986 0.00483032
4 anosssc2 scalpl 0.00552462
4 anosssc2 vargase_lam 0.999913
4 hazzardous1984 aarror 0.0645855
5 japanesports zzz87012 0.00549996
5 rrb115 j30723 0.999171
```

A complex piecewise (anti-)monotone measure

Definition (The slope measure)

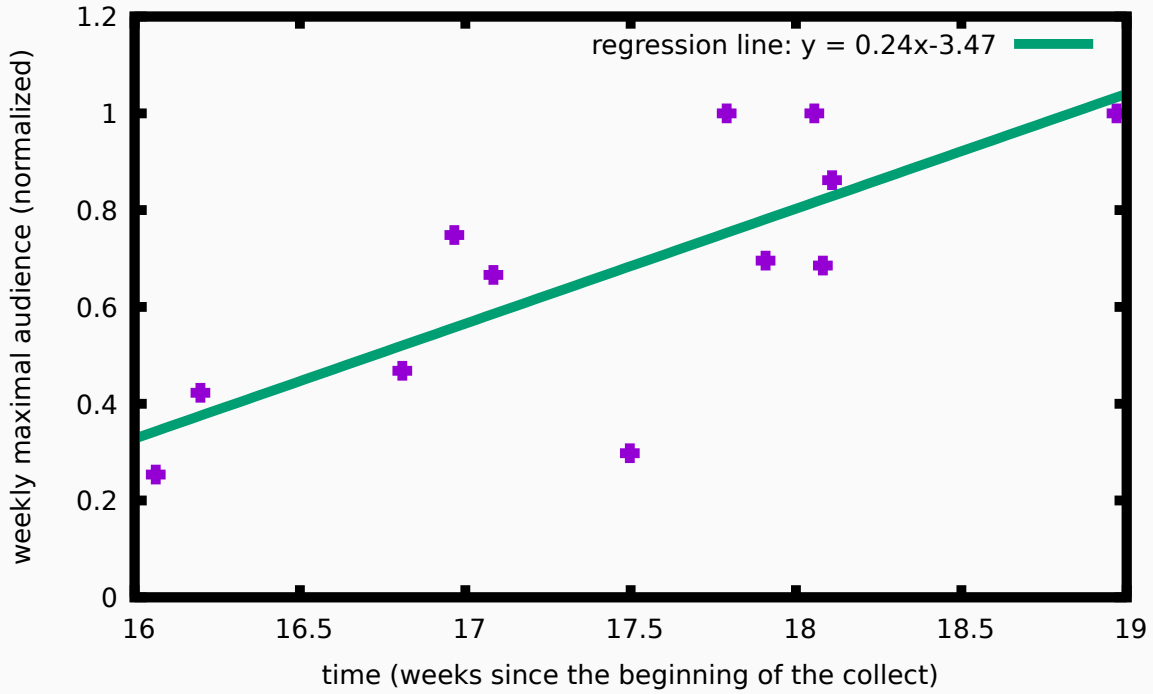
Given $I \subseteq \{1, \ldots, n\}$ and two external-data-access functions x and y over I, the *slope* is the following measure:

$$(X_1, \ldots, X_n) \mapsto \frac{\sum_{t \in \prod_{i \in I} X_i} x(t) \sum_{t \in \prod_{i \in I} X_i} y(t) - \left|\prod_{i \in I} X_i\right| \sum_{t \in \prod_{i \in I} X_i} x(t)y(t)}{\left(\sum_{t \in \prod_{i \in I} X_i} x(t)\right)^2 - \left|\prod_{i \in I} X_i\right| \sum_{t \in \prod_{i \in I} X_i} x(t)^2} .$$

Yes, it's piecewise (anti-)monotone !

Slope example

	week 17	week 18	week 19
egstephano	(17.0, 0.75)	(17.1, 0.67)	(18.1, 0.86)
esltv_sc2	(16.8, 0.47)	(17.5, 0.30)	(19.0, 1.00)
liquidtlo	(16.1, 0.25)	(17.8, 1.00)	(18.1, 0.69)
sc2proleague	(16.2, 0.42)	(17.9, 0.70)	(18.1, 1.00)



Protocol

- Maximize number of channels, number of weeks, and the *slope* of the regression line of the points associated with every pair (channel, week).
- Semantically, we look for big groups of channels whose number of viewers grows during a big group of weeks.

The Twitch dataset: Results

- 114 skypatterns extracted in 7 minutes 34 seconds

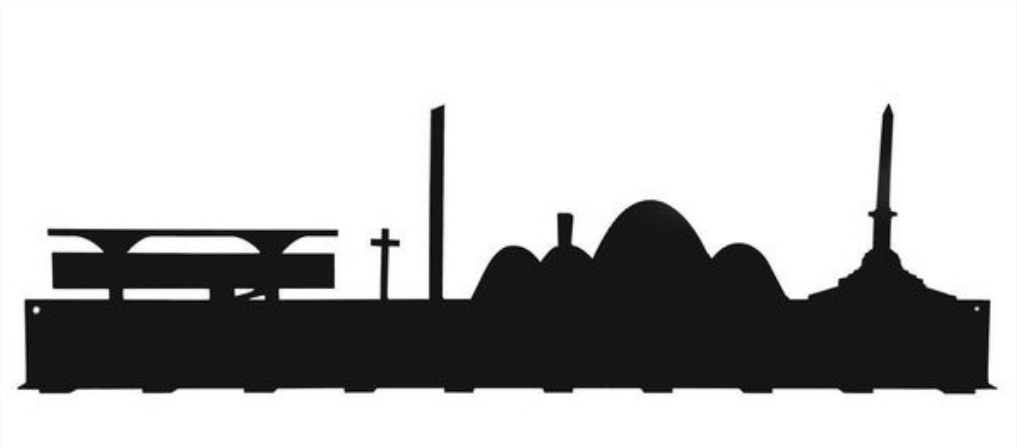
Example:

$$(\{\texttt{balosaar}, \texttt{darrenlorduk}, \texttt{japi1}, \texttt{lumberdavid}, \texttt{yoctz22}, \texttt{91kitsune91}\},$$
$$\{\texttt{egjd}, \texttt{egstephano}, \texttt{esltv\_sc2}, \texttt{sc2proleague}\},$$
$$\{\texttt{week17}, \texttt{week18}, \texttt{week19}\})$$

Conclusion

Our contributions:

- Generalization of skypattern mining, from 0/1 matrices to fuzzy tensors.
- Design of a generic algorithm, implementation by extending `multidupehack`.
  - Manage a large class of measures (piecewise (anti-)monotone).
  - Efficient, outperform existing proposals.
  - Able to mine relevant patterns in real-life situations.



`gitlab.com/nnadisic/skypatterns-uncertain-tensors`