Logical Aspects of Spatial Databases

Part I: First-order geometric and topological queries

Jan Van den Bussche
Hasselt University

Spatial data

Q: what is a spatial dataset?

A: a set $S \subset \mathbb{R}^n$

equivalently, an n-ary relation S over $\mathbb{R}$ (Cartesian coordinates)

$\mathbb{R} = (\mathbb{R}, 0, 1, +, \cdot, <)$

Use first-order logic to express properties of spatial datasets

E.g. $$\exists a \exists b \forall x \forall y (S(x,y) \rightarrow y = a \cdot x + b)$$

$(\mathbb{R}, S) \models \phi$ is abbreviated $S \models \phi$

Geometric properties

Let G be a group of transformations of $\mathbb{R}^n$

- similarities (Euclidean geometry)
- affinities (affine geometry)
- continuous transformations (topology)
- ...

Property ϕ is called G-*geometric* if it is invariant under G:

$$\forall S\,\forall g \in G : S \models \phi \Leftrightarrow g(S) \models \phi$$

- "S lies on a circle" is Euclidean, not affine
- "S lies on a straight line" is affine, not topological
- "S has dimension two" is topological

Capturing the G-geometric first-order properties

Easy when G is *first-order parameterisable:*

- injection $p : G \to \mathbb{R}^\ell$
- $\{(p(g), \bar{x}, \bar{y}) \mid g \in G \text{ and } \bar{y} = g(\bar{x})\}$ is first-order definable in $\mathbb{R}$

E.g. affinities in $\mathbb{R}^2$ are tuples (a, b, c, d, e, f) such that

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} \neq 0 \text{ and } \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$$

A first-order property is G-geometric

$\Updownarrow$

expressible by a sentence of the form

$$\phi \wedge \forall p(g) \in p(G)[\phi(S) \leftrightarrow \phi(g(S))]$$

with ϕ arbitrary sentence over $(\mathbb{R}, S)$.

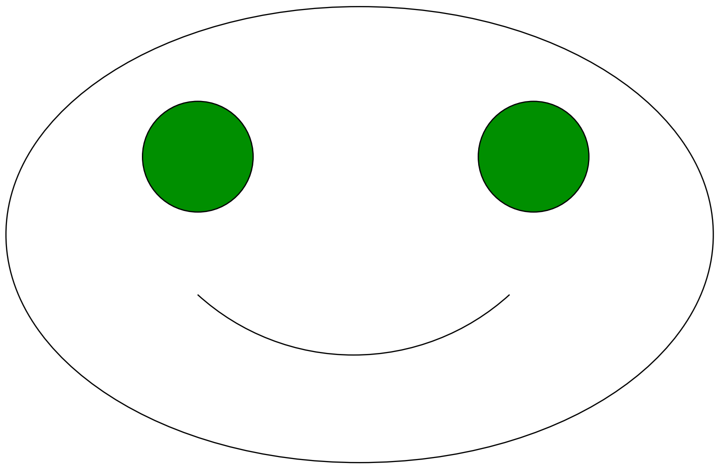
Topological properties

Invariant under continuous transformations (isotopies)

Not first-order parameterisable

We can capture them on the class of datasets in $\mathbb{R}^2$ that are

- *semi-algebraic:* definable in $\mathbb{R}$
- *closed* in the topological sense



$$x^2/25 + y^2/16 = 1 \\ \vee\, x^2 + 4x + y^2 - 2y \leqslant -4 \vee x^2 - 4x + y^2 - 2y \leqslant -4 \\ \vee\, (x^2 + y^2 - 2y = 8 \wedge y \leqslant -1)$$

We call such sets "plain"

Which topological properties of plain sets are FO?

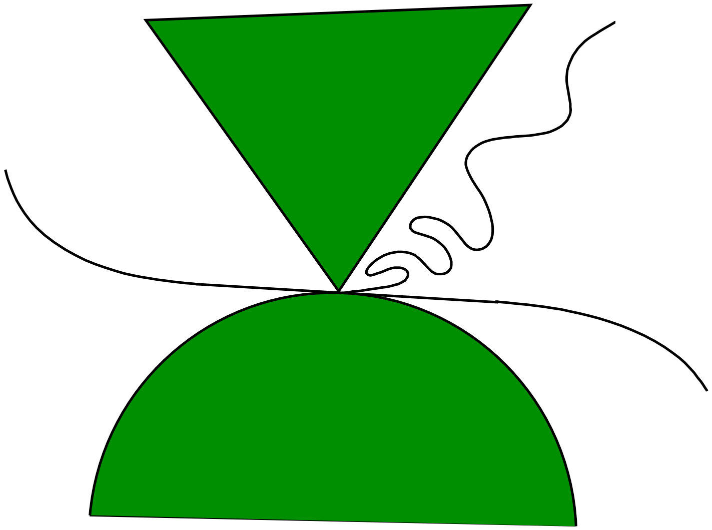
FO-expressible:

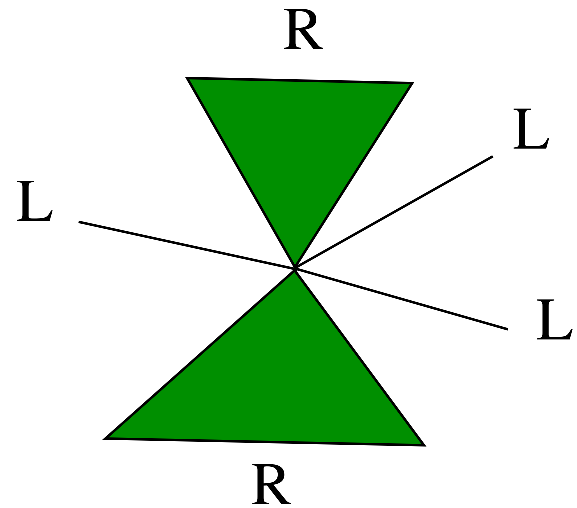
- "The dimension is 0 (1, 2)"
- "There is a point where three lines intersect"
- "There is a point where two 2-dim regions touch"

Not FO-expressible:

- "There is a point where an even number of lines intersect"
- "The number of points where two 2-dim regions touch is even"
- "The set is topologically connected"

Cones





Around each point on the boundary we see a circular list of L's and R's, called the *cone*

- points with cone (LL) or (R), or interior points: *regular*

- others: *singular* (finitely many)

W.l.o.g. we can focus on the singular points

Cone Logic

Atomic formulas: $|e| \geqslant n$
with e a star-free regular expression over $\Sigma = \{L, R\}$

Meaning: there are at least n points whose cone satisfies e

A CL-sentence is a boolean combination of atomic formulas.

E.g. "The dimension is 0":

$$|L\Sigma^*| = 0 \wedge |R\Sigma^*| = 0$$

E.g. "There is a point where three lines intersect":

$$|LLLLLL| \geqslant 1$$

E.g. "There is a point where two regions touch":

$$|RR| \geqslant 1$$

The first-order topological properties of plain sets are precisely those expressible in CL

[Benedikt, Kuijpers, Löding, VdB, Wilke]

Proof

1. Topological elementary equivalence

2. Flower datasets

3. Finite structures over the reals, collapse theorems

4. Coding flower datasets by finite structures

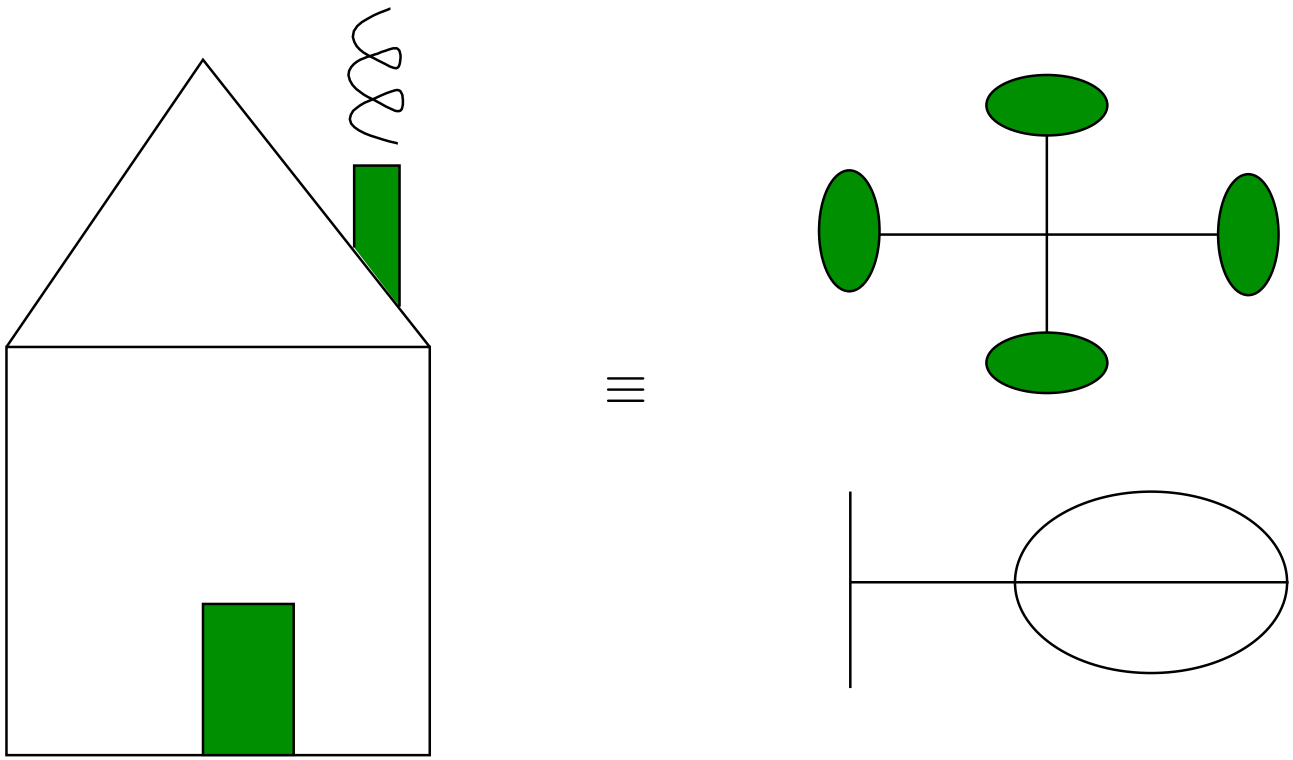
5. Translating sentences about datasets into sentences about codes

6. Invariance arguments over codes

Topological elementary equivalence

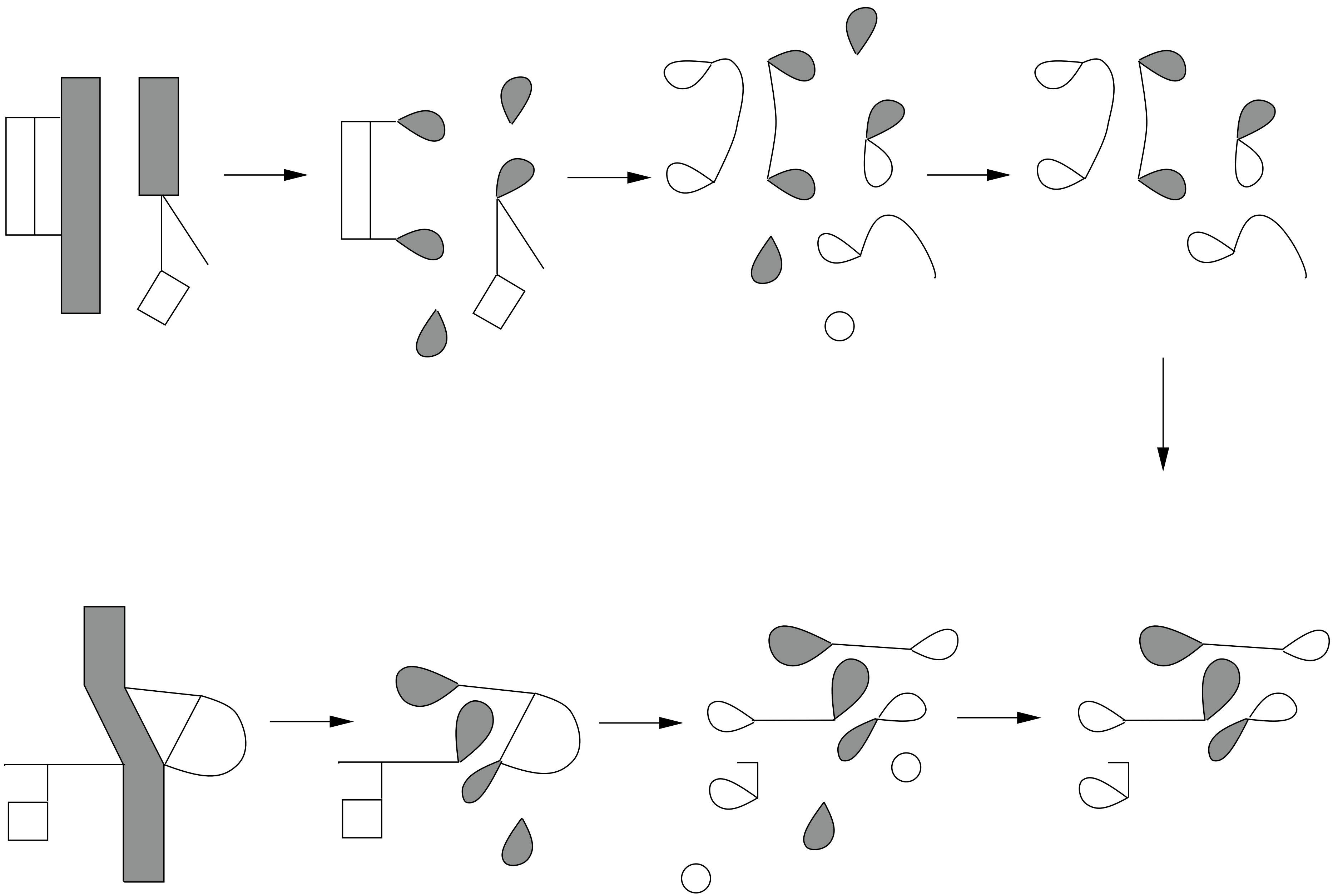
For plain sets A and B, write $A \equiv B$ if indistinguishable by topological first-order sentences

$A \equiv B \Leftrightarrow A$ and B have precisely the same cones, with the same multiplicities



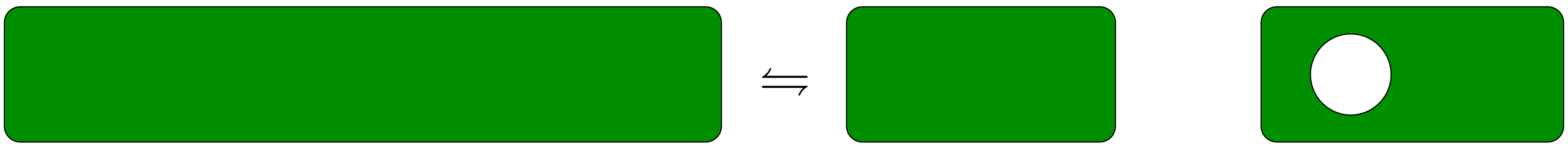
[Kuijpers, Paredaens, VdB]

TEE proof: transformation into *flower normal form*



Transformation rules

E.g. "cut and paste":

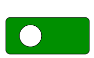


Show that this is indistinguishable by topological first-order sentences, using a reduction from queries on *finite structures over the reals*

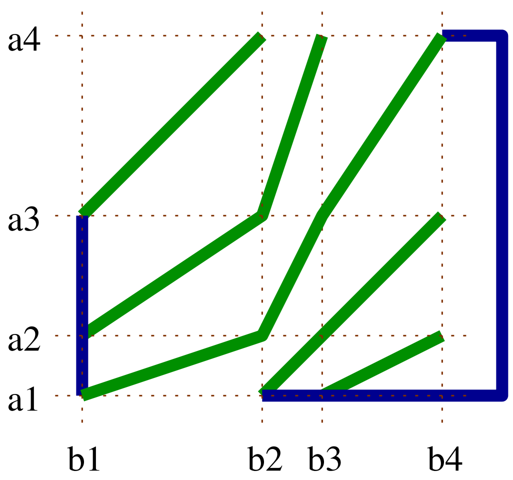
These are structures of the form $(\mathbb{R}, R_1, \ldots, R_k)$ with R_i finite

E.g. *Majority:* given finite unary relations R_1 and R_2, is $\#R_1 \geqslant \#R_2$?

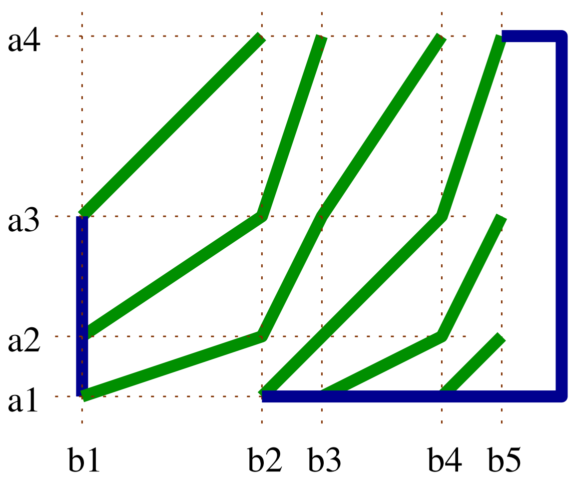
Write an FO-formula $\psi(x, y)$ such that for each finite structure $D = (\mathbb{R}, R_1, R_2)$:

$\psi(D) \cong$  if $\#R_1 \geqslant \#R_2$;
$\psi(D) \cong$   if $\#R_1 < \#R_2$.

$R_1 = \{a_1, a_2, a_3, a_4\},\ R_2 = \{b_1, b_2, b_3, b_4\}$:



$R_1 = \{a_1, a_2, a_3, a_4\},\ R_2 = \{b_1, b_2, b_3, b_4, b_5\}$:



Collapse theorems

Natural–active collapse:

Every first-order query on finite structures over the reals is already expressible by a sentence in which all quantifiers are relativised to the finite relations.

Generic collapse:

Every first-order query on finite structures over the reals (in the language $(0, 1, +, \cdot, <, R_1, \ldots, R_k)$) that is *order-generic* (invariant under all monotone permutations of $\mathbb{R}$) is already expressible by a sentence in the language $(<, R_1, \ldots, R_k)$.

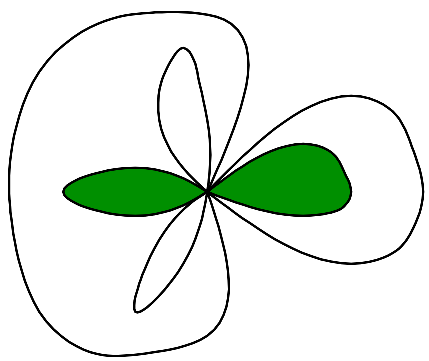
So, order-generic first-order sentences view finite structures over the reals just as abstract, ordered, finite structures.

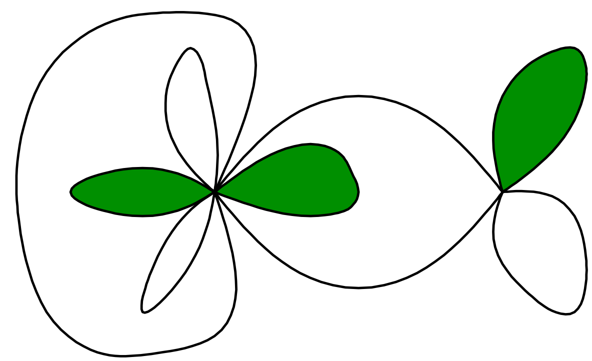
[Benedikt, Libkin, et al.]

Flower datasets

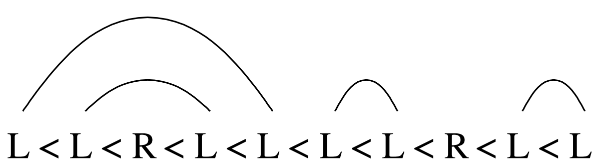
A normal form for datasets (as far as topological FO is concerned)

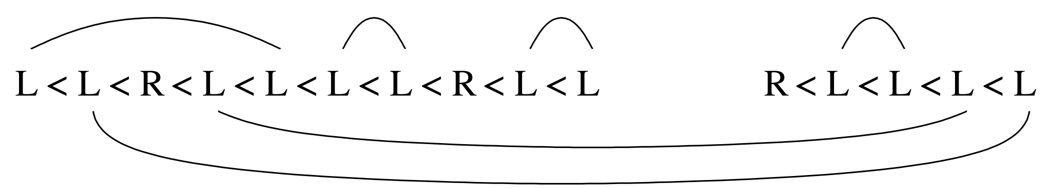
Disjoint union of single or paired *flowers*





Represent by abstract finite structure called *code:*
disjoint union of single or paired *cycles*





These are (possible paired) word structures equipped with a planar matching on the L's

Translation argument

Drawing Lemma: We can write an FO-formula $\delta(x, y)$ such that for any code C embedded in the reals, $\delta(C)$ is a flower dataset that is a drawing of C.

$\Rightarrow$ Translate a topological sentence ϕ about flower datasets into a sentence $\psi := \phi \circ \delta$ about codes, called *implementation* of ϕ

Using collapse theorems, we may assume ψ sees only an ordered version of the abstract code. But this ordering $<$ is not the $\prec$ of the word structures!

W.l.o.g. assume that $<$ agrees with $\prec$, so all $<$ does is shuffle the separate cycles in some order

ψ is *invariant* under the way this shuffling is done

$\Rightarrow$ Show that $<$-invariant FO on ordered codes collapses to FO on codes

Planar-matching-invariant FO on word structures

Word structures over finite alphabet Σ, additionally equipped with a planar matching G

Main Invariance Lemma: G-invariant FO collapses to FO on the class of word structures with a planar matching

Cf. logical characterisation of context-free languages
[Lautemann, Schwentick, Thérien]

Main Invariance Lemma can be adapted to cycles and cycle pairs

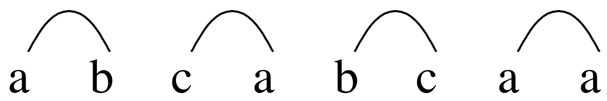
Implementations of topological FO-sentences are indeed G-invariant

- "Push down" invariance to individual cycles and cycle pairs
- Get rid of pairings by rearrangement argument (TEE)
- Use equivalence of FO and star-free regular expressions

$\Rightarrow$ Cone-Logic Theorem is proved.

Proof of Invariance Lemma

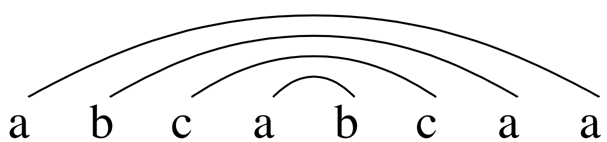
A *chain matching* can be simulated using *alternating markers*:



a b’ c a’ b c’ a a’

Can translate FO over chain matchings to FO over marked words

A *parenthetical matching* can be simulated using *folding*:



a b c a
a a c b

Can translate FO over parenthetical to FO over folded words

Both translations imply that set W of words is surely regular, and can have only very limited kind of *counters*

Final argument shows that $W = W' \cap (\Sigma\Sigma)^*$ with W' counter-free regular $\Rightarrow$ first-order

Corollary: topological collapse

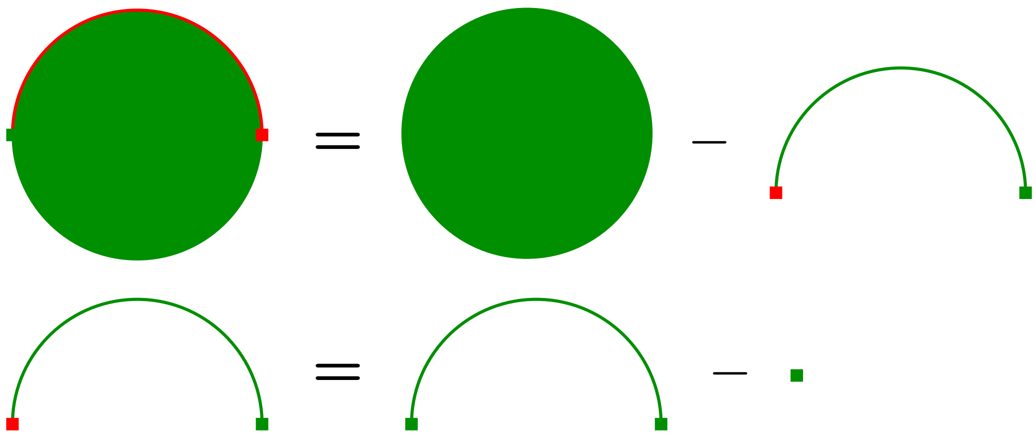
CL can already be expressed in FO over $(\mathbb{R}, S)$ using only $<$ and S

$\Rightarrow$ Every topological first-order property of plain sets is already expressible by a sentence using only $<$ and S

Open problems

What about non-closed sets?

We can always decompose a set in $\mathbb{R}^n$ in $n+1$ closed sets:



$\Rightarrow$ What about ensembles of closed sets?

[Grohe & Segoufin]

No problem for FO-parameterisable geometric queries

More open problems

What about $\mathbb{R}^3$ and higher?

And, what about non-semialgebraic sets?

E.g. "Every point in the set has cone (LL)":

- FO
- topological over semialgebraic sets
- not topological over all sets

Example of a topological property that is FO over semialgebraic sets but not over all sets?

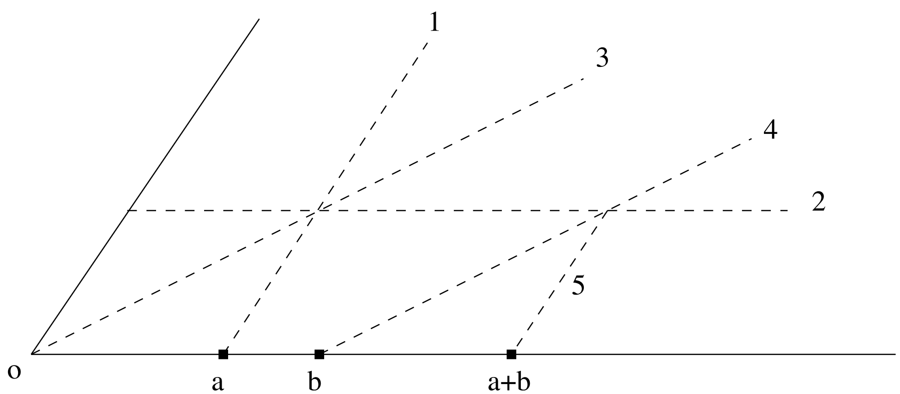
Point-based logics

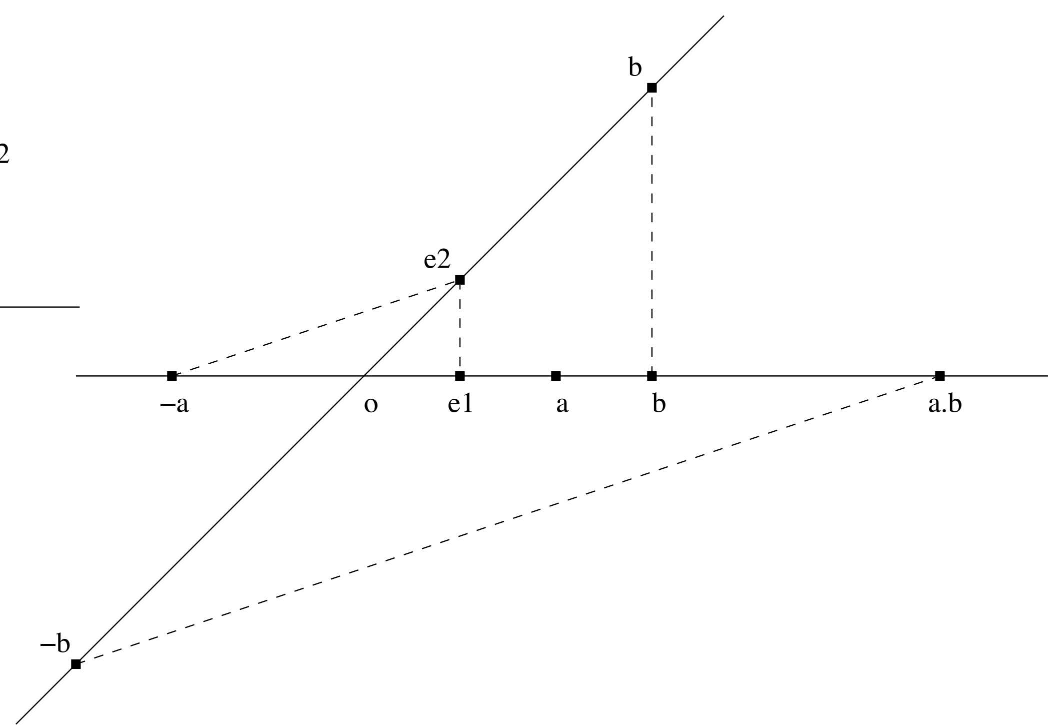
FO over $\mathbb{R}$ is a coordinate-based logic

Cone Logic is a point-based logic

Can we find point-based logics for other kinds of geometric queries?

[Tarski:] Geometric constructions of addition and multiplication are FO-expressible using a single ternary predicate β ("between")





Affine queries

View $S \subset \mathbb{R}^2$ as a *unary* relation over the structure $(\mathbb{R}^2, \beta)$

Denote $\mathrm{FO}(\mathbb{R}, 0, 1, +, \cdot, <, S^{(2)})$ by $\mathrm{FO}(\mathbb{R})$

Denote $\mathrm{FO}(\mathbb{R}^2, \beta, S^{(1)})$ by $\mathrm{FO}(\beta)$

Call a triple (o, e_1, e_2) of non-collinear points a *basis*

For each $\mathrm{FO}(\mathbb{R})$-sentence ϕ there exists an $\mathrm{FO}(\beta)$-formula $\psi(o, e_1, e_2)$ such that for every dataset S and

$$\text{for every basis } (o, e_1, e_2) : S \models \psi(o, e_1, e_2) \Leftrightarrow \alpha(S) \models \phi$$

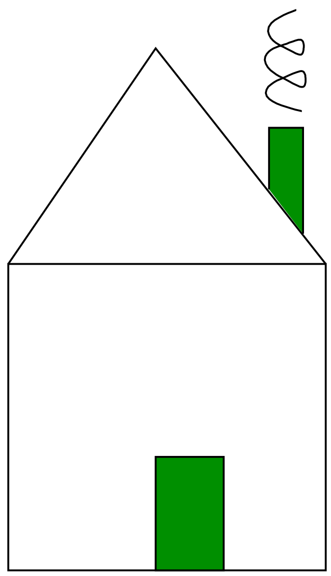
where

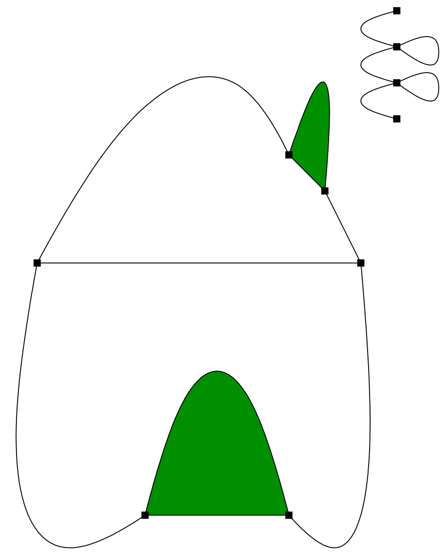
$$\alpha \text{ is the unique affinity} : (o, e_1, e_2) \mapsto ((0,0), (1,0), (0,1))$$

For each affine $\mathrm{FO}(\mathbb{R})$-sentence ϕ there exists an equivalent $\mathrm{FO}(\beta)$-sentence ψ (and vice versa).

Plane graphs

The topology of a semialgebraic set in the plane can be represented by a finite structure





Every topological first-order sentence about semialgebraic sets in the plane, using only $<$ and S, can be translated to a first-order sentence about the corresponding plane graphs.

[Segoufin & Vianu]

By topological collapse, we know that (for a single plain set) the restriction to only $<$ is harmless

References (http://alpha.uhasselt.be/~vdbuss)

- M. Gyssens, J. Van den Bussche, D. Van Gucht. Complete geometric query languages. *JCSS* 58(1):54–68, 1999.
- M. Benedikt, B. Kuijpers, C. Löding, J. Van den Bussche, T. Wilke. A characterization of first-order topological properties of planar spatial data. *JACM*, to appear.
- B. Kuijpers, J. Paredaens, J. Van den Bussche. On topological elementary equivalence of closes semi-algebraic sets in the real plane. *JSL* 65(4):1530–1555, 2000.
- S. Grumbach, J. Su. Queries with arithmetical constraints. *TCS* 173(1):151–181, 1997.
- L. Libkin. *Elements of Finite Model Theory*. Springer, 2004.
- M. Grohe, L. Segoufin. On first-order topological queries. *TOCL* 3(3):336–358, 2002.
- L. Segoufin, V. Vianu. Querying spatial databases via topological invariants. *JCSS* 61(2):270–301, 2000.