The genus of regular languages and other ideas from low-dimensional topology

Florian Deloup

Institut de Mathématiques de Toulouse, France

June 21, 2016

Joint work with Guillaume Bonfante.

Joint work with Guillaume Bonfante.

1) The genus of regular languages, 2012. Math. Str. Computer Sc., 2016.

2) The decidability of language genus computation, 2016. Available on ArXiv.

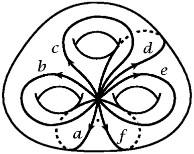
What I won't talk about in this talk

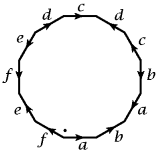
Topology $\Longrightarrow$ Languages (as tool to study topology): *languages as topological invariants*

What I won't talk about in this talk

Topology $\Longrightarrow$ Languages (as tool to study topology): *languages as topological invariants*

- Fundamental group of a topological space, languages (Poincaré, 1895, "Analysis situs" paper, also Riemann and Klein)

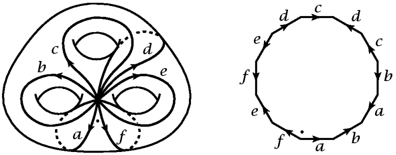




What I won't talk about in this talk

Topology $\Longrightarrow$ Languages (as tool to study topology): *languages as topological invariants*

- Fundamental group of a topological space, languages (Poincaré, 1895, "Analysis situs" paper, also Riemann and Klein)



- Knots: encoding Reidemeister moves (1927) yields language(s). Particular cases: quandles, Wirtinger presentation of the fundamental group of the complement of a knot.

What I will talk about in this talk

Languages $\Longrightarrow$ Topology (as a tool to study languages): *topology as a language invariant*

This talk: language invariants from *low-dimensional topology*.

"Moore's Law"

Moore's "Law" (1960s)

The number of transistors in a dense integrated circuit doubles every two years.

"Moore's Law"

Moore's "Law" (1960s)

The number of transistors in a dense integrated circuit doubles every two years.

Correction to Moore's "Law" (2005)

Moore's Law has to end.

"Moore's Law"

Moore's "Law" (1960s)

The number of transistors in a dense integrated circuit doubles every two years.

Correction to Moore's "Law" (2005)

Moore's Law has to end.

Reason invoked: physical limit of matter processing.

"Moore's Law"

Moore's "Law" (1960s)

The number of transistors in a dense integrated circuit doubles every two years.

Correction to Moore's "Law" (2005)

Moore's Law has to end.

Reason invoked: physical limit of matter processing.
Shape and space organization become central $\Longrightarrow$
Low-dimensional topology $\Longrightarrow$ Invariants of Languages from low-dimensional topology

Regular languages

Set-up:

- the class $\mathrm{Reg}_{\mathscr{A}}$ of regular languages on a finite alphabet $\mathscr{A}$.
- the class $\mathrm{DFA}_{\mathscr{A}}$ of deterministic finite automata on $\mathscr{A}$.

Regular languages

Set-up:

- the class $\mathrm{Reg}_{\mathscr{A}}$ of regular languages on a finite alphabet $\mathscr{A}$.
- the class $\mathrm{DFA}_{\mathscr{A}}$ of deterministic finite automata on $\mathscr{A}$.

Working-out definition: a *regular language L* on alphabet $\mathscr{A}$ is a subset of $\mathscr{A}^*$, starting from a subset of $\mathscr{A}$ and recursively computed by a finite number of the familiar 3 operations:

Regular languages

Set-up:
- the class $\mathrm{Reg}_{\mathscr{A}}$ of regular languages on a finite alphabet $\mathscr{A}$.
- the class $\mathrm{DFA}_{\mathscr{A}}$ of deterministic finite automata on $\mathscr{A}$.

Working-out definition: a *regular language* L on alphabet $\mathscr{A}$ is a subset of $\mathscr{A}^*$, starting from a subset of $\mathscr{A}$ and recursively computed by a finite number of the familiar 3 operations:

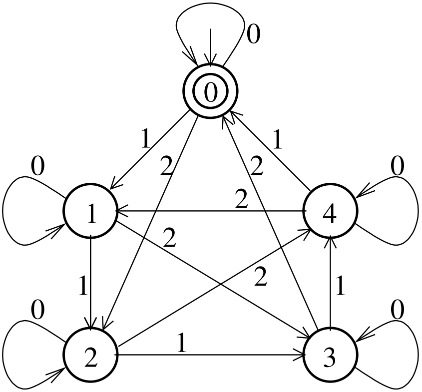
- Union of two languages:
$(L, L') \mapsto L \cup L' = \{w \in A^* \mid w \in L, \mathrm{or}\ w \in L'\}$.
- Composition of two languages:
$(L, L') \mapsto LL' = \{ww' \mid w \in L, w' \in L'\}$.
- Star operation: $L \mapsto L^* = \bigcup_{n \geq 0} L^n$

Automata

An *automaton* is a decorated directed (multi)graph.

Automata

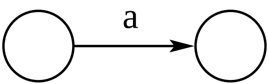
An *automaton* is a decorated directed (multi)graph.



Automata

Decoration:

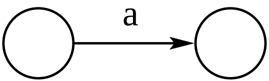
- label each directed edge (transition) by a letter of the alphabet $\mathscr{A}$.



Automata

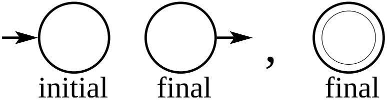
Decoration:

- label each directed edge (transition) by a letter of the alphabet $\mathscr{A}$.



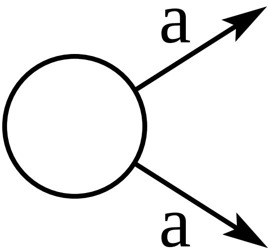
- distinguish special states: one initial state, one subset of final states.

Pictorial convention for initial and final states:



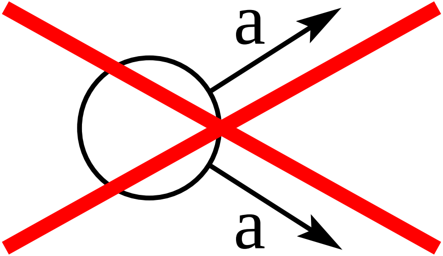
Deterministic automaton

The automaton is *deterministic* if there is at most one transition labelled by a given letter.



Deterministic automaton

The automaton is *deterministic* if there is at most one transition labelled by a given letter.



Kleene

Let $\mathtt{A} \in \mathrm{DFA}$. The language $L(\mathtt{A})$ computed by $\mathtt{A}$ is the set of all words $w \in \mathscr{A}^*$ read from (the sequence of labels of) a path starting at the initial state and ending at some final state of $\mathtt{A}$.

Theorem

The assignment $\mathtt{A} \mapsto L(\mathtt{A})$ *defines a surjective map* $\mathrm{DFA}_{\mathscr{A}} \to \mathrm{Reg}_{\mathscr{A}}$.

In the words of the topologists

Challenge: define “quantum invariants” of $L = L(\mathtt{A})$ (beyond the size of L), locally computable from a picture of any automaton $\mathtt{A}$ computing L. The computation from two equivalent automata should give the same invariant.

Why is it a challenge ? Nonlocal nature of the computation: two automata can be nonlocally equivalent.

The simplest invariant of language.

The simplest invariant of language.

Definition

The *size* $|L|$ of a language L is the smallest number of states required to produce a deterministic automaton $\mathtt{A}$ computing L:

$$|L| = \min\{|\mathtt{A}| \mid \mathtt{A} \in \mathrm{DFA},\ L(\mathtt{A}) = L\}.$$

The simplest invariant of language.

Definition

The *size* $|L|$ of a language L is the smallest number of states required to produce a deterministic automaton $\mathtt{A}$ computing L:

$$|L| = \min\{|\mathtt{A}| \mid \mathtt{A} \in \mathrm{DFA},\ L(\mathtt{A}) = L\}.$$

Theorem (Myhill-Nerode, 1950s)

Let L be a regular language. There is a unique automaton $\mathtt{A} \in \mathrm{DFA}$ such that $L(\mathtt{A}) = L$ with number $|A|$ of states equal to $|L|$.

The simplest invariant of language.

Definition

The *size* $|L|$ of a language L is the smallest number of states required to produce a deterministic automaton $\mathtt{A}$ computing L:

$$|L| = \min\{|\mathtt{A}| \mid \mathtt{A} \in \mathrm{DFA},\ L(\mathtt{A}) = L\}.$$

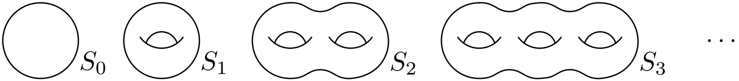
Theorem (Myhill-Nerode, 1950s)

Let L be a regular language. There is a unique automaton $\mathtt{A} \in \mathrm{DFA}$ such that $L(\mathtt{A}) = L$ with number $|A|$ of states equal to $|L|$.

Such an automaton is called the *minimal* automaton of L.

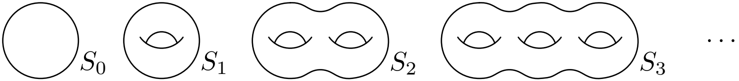
Classification of closed oriented surfaces

Theorem (first stated ˜1850, proved ˜1920): The topological type of a closed oriented surface Σ is determined by one natural number $g(\Sigma) \in \mathbb{N}$.



Classification of closed oriented surfaces

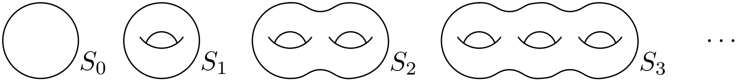
Theorem (first stated ˜1850, proved ˜1920): The topological type of a closed oriented surface Σ is determined by one natural number $g(\Sigma) \in \mathbb{N}$.



The genus is the number of "handles" required to produce the surface Σ from the sphere.

Classification of closed oriented surfaces

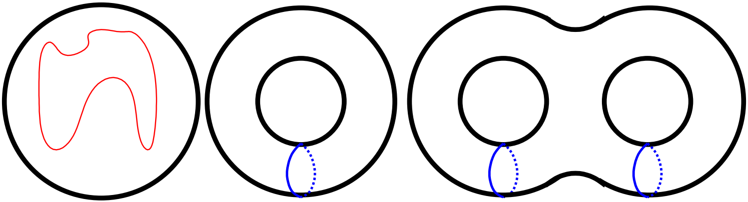
Theorem (first stated ~1850, proved ~1920): The topological type of a closed oriented surface Σ is determined by one natural number $g(\Sigma) \in \mathbb{N}$.



The genus is the number of "handles" required to produce the surface Σ from the sphere.

The genus $g(\Sigma)$ of Σ is the maximal number of mutually disjoint simple closed curves $C_1, \dots, C_g$ such that the complement $\Sigma - (C_1 \cup \dots \cup C_g)$ remains connected.

Examples



genus = 0 genus = 1 genus = 2 . . . . . .

Embedding an automaton into a closed oriented surface

An embedding of a graph is essentially a "drawing of the graph without crossings of the edges".

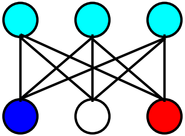
Embedding an automaton into a closed oriented surface

An embedding of a graph is essentially a "drawing of the graph without crossings of the edges".

Definition

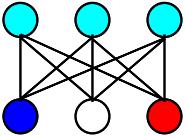
An embedding of a graph $G = (E, V)$ into a closed oriented surface Σ is a map $\varphi : (E, V) \to \Sigma$ sending injectively vertices to points, sending edges to simple arcs in Σ such that $\varphi(\partial e) = \partial\varphi(e)$ for any edge $e \in E$, $\varphi(e) \cap \varphi(e') = \varphi(\partial e) \cap \varphi(\partial e')$ for any pair $e, e' \in E$.

Example. The "Utility Graph" (complete bipartite $K_{3,3}$)

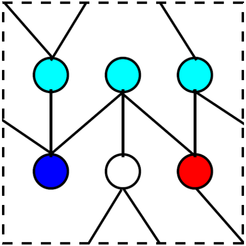


is not embeddable in the sphere (Kuratowski).

Example. The "Utility Graph" (complete bipartite $K_{3,3}$)



is not embeddable in the sphere (Kuratowski). However, $K_{3,3}$ embeds into a torus (genus 1).



Genus of a regular language

Definition

Let L be a regular language. The *genus* $g(L)$ is defined as

$$g(L) = \min\{g(\mathtt{A}) \mid \mathtt{A} \in \mathrm{DFA},\ L(\mathtt{A}) = L\}.$$

If $g(L) = 0$, then L is said to be *planar*.

Genus of a regular language

Definition

Let L be a regular language. The *genus* $g(L)$ is defined as

$$g(L) = \min\{g(\mathtt{A}) \mid \mathtt{A} \in \mathrm{DFA},\ L(\mathtt{A}) = L\}.$$

If $g(L) = 0$, then L is said to be *planar*.

Remark: the definition makes sense because any graph embeds into some closed oriented surface.

Recall: the simplest invariant of a regular language L is its *size*

$$|L| = \min\{|\mathtt{A}| \mid \mathtt{A} \in \mathrm{DFA}, L(\mathtt{A}) = L\}.$$

Recall: the simplest invariant of a regular language L is its *size*

$$|L| = \min\{|\mathtt{A}| \mid \mathtt{A} \in \mathrm{DFA}, L(\mathtt{A}) = L\}.$$

Question: relation between the genus and the size of a language ?

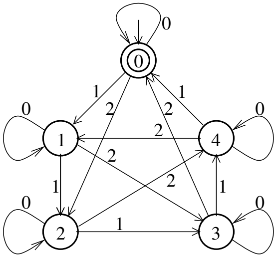
Basic observation: the automaton $\mathtt{A}$ for which a minimal embedding (with minimal genus) is realized *may not be* the minimal automaton.

Basic observation: the automaton A for which a minimal embedding (with minimal genus) is realized *may not be* the minimal automaton.

Alphabet: $\mathscr{A} = \{0, 1, 2\}$

Morphism: $\varphi : \mathscr{A}^* \to \mathbb{Z}/5\mathbb{Z}$ defined by $\varphi(aw) = \varphi(a) + \varphi(w)$ for any $a \in \mathscr{A}$, $w \in \mathscr{A}^*$.

Language: $L = \{w \in \mathscr{A}^* \mid \varphi(w) = 0 \mod 5\}$

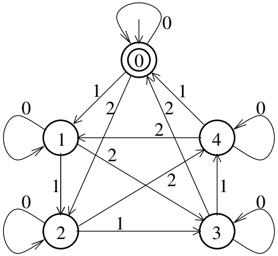


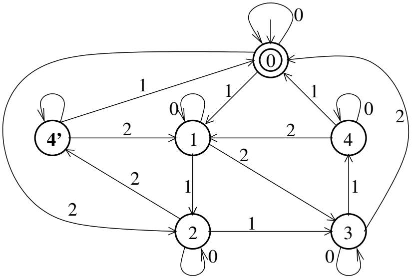
Basic observation: the automaton A for which a minimal embedding (with minimal genus) is realized *may not be* the minimal automaton.

Alphabet: $\mathscr{A} = \{0, 1, 2\}$

Morphism: $\varphi : \mathscr{A}^* \to \mathbb{Z}/5\mathbb{Z}$ defined by $\varphi(aw) = \varphi(a) + \varphi(w)$ for any $a \in \mathscr{A}$, $w \in \mathscr{A}^*$.

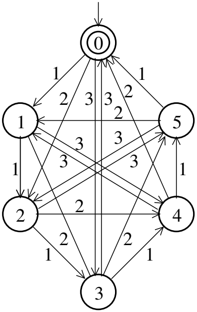
Language: $L = \{w \in \mathscr{A}^* \mid \varphi(w) = 0 \mod 5\}$



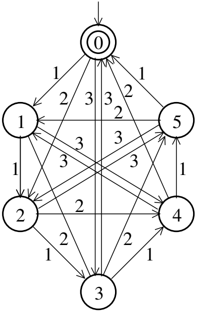


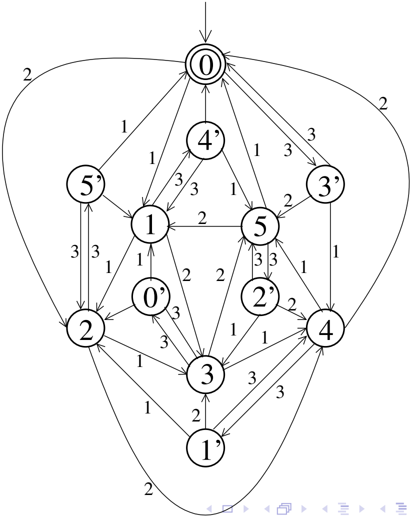
Another example.

Another example.



Another example.





Topological size

Definition

The *topological size* of a language L is

$$|L|_{\mathrm{top}} = \min\{|\mathtt{A}| \mid L(\mathtt{A}) = L, g(\mathtt{A}) = g(L)\}.$$

Topological size

Definition

The *topological size* of a language L is

$$|L|_{\mathrm{top}} = \min\{|\mathtt{A}| \mid L(\mathtt{A}) = L, g(\mathtt{A}) = g(L)\}.$$

By definition: $|L|_{\mathrm{top}} \geq |L|$.

Topological size

Definition

The *topological size* of a language L is

$$|L|_{\mathrm{top}} = \min\{|\mathtt{A}| \mid L(\mathtt{A}) = L, g(\mathtt{A}) = g(L)\}.$$

By definition: $|L|_{\mathrm{top}} \geq |L|$.

The topological size $|L|_{\mathrm{top}}$ is regarded as "the cost" you are willing to pay for the simplest topological embedding of the representing automaton of L.

Question

Is there a universal bound $|L|_{\mathrm{top}} \leq f(|L|)$ for some explicit function f ?

Question

Is there a universal bound $|L|_{\mathrm{top}} \leq f(|L|)$ for some explicit function f ?

If such a function exists, it has to be at least exponential.

Theorem (2015)

There is a family of planar regular languages $(L_n)_{n\geq 1}$ such that for some $K > 2$, $|L_n|_{\mathrm{top}} = O(K^{|L_n|})$.

Book and Chandra (1978) raised the question of whether the planarity of a language is decidable.

Book and Chandra (1978) raised the question of whether the planarity of a language is decidable.
One may generalize the question and ask whether the following is true.

Book and Chandra (1978) raised the question of whether the planarity of a language is decidable.
One may generalize the question and ask whether the following is true.

Conjecture

The genus of a regular language is computable.

Book and Chandra (1978) raised the question of whether the planarity of a language is decidable.
One may generalize the question and ask whether the following is true.

Conjecture

The genus of a regular language is computable.

Partial positive answer:

Theorem (2012, 2015)

If the language has "no short cycles", the conjecture is true.

"No short cycles".

Definition

A language has no cycles of length less than k if the underlying graph of its minimal automaton has no cycles of length less than k.

"No short cycles".

Definition

A language has no cycles of length less than k if the underlying graph of its minimal automaton has no cycles of length less than k.

Cycle = simple cycle = closed path without repeated edge (no matter its orientation), regardless of the orientation of the original graph.

"No short cycles".

Definition

A language has no cycles of length less than k if the underlying graph of its minimal automaton has no cycles of length less than k.

Cycle = simple cycle = closed path without repeated edge (no matter its orientation), regardless of the orientation of the original graph.

Theorem (2012)

Let L be a language on m letters. Assume that $m \geq 4$ and that L has no cycles of length ≤ 2. Then

$$1 + \frac{m-3}{6}|L| \leq g(L) \leq 1 + \frac{m-1}{2}|L|.$$

Hierarchies of languages

Remark Every language on one letter is planar (exercise).

Hierarchies of languages

Remark Every language on one letter is planar (exercise).

Book and Chandra (1978) construct an example of a language on two letters which is nonplanar from a minimal deterministic automaton with 35 states.

Hierarchies of languages

Remark Every language on one letter is planar (exercise).

Book and Chandra (1978) construct an example of a language on two letters which is nonplanar from a minimal deterministic automaton with 35 states.

Theorem (2012, 2015)

Let A be an alphabet of at most 2 letters. There exists a family of languages $(L_n)_{n\in\mathbb{N}}$ on alphabet A such that $g(L_n) = n$.

Hierarchies of languages

Remark Every language on one letter is planar (exercise).

Book and Chandra (1978) construct an example of a language on two letters which is nonplanar from a minimal deterministic automaton with 35 states.

Theorem (2012, 2015)

Let A be an alphabet of at most 2 letters. There exists a family of languages $(L_n)_{n\in\mathbb{N}}$ on alphabet A such that $g(L_n) = n$.

Remark. The minimal example of nonplanar language on two letters has 30 states.

Hierarchies of languages

Remark Every language on one letter is planar (exercise).

Book and Chandra (1978) construct an example of a language on two letters which is nonplanar from a minimal deterministic automaton with 35 states.

Theorem (2012, 2015)

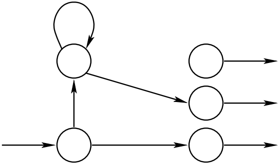
Let A be an alphabet of at most 2 letters. There exists a family of languages $(L_n)_{n\in\mathbb{N}}$ on alphabet A such that $g(L_n) = n$.

Remark. The minimal example of nonplanar language on two letters has 30 states. Conjecture. 30 is optimal.

The following definition is the "directed version" of Fellows' graph emulator (in connection with the planar finite cover conjecture in the 1980s).

Definition

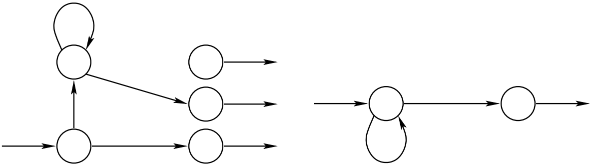
Let $G = (E, V)$ be a directed graph. A *directed emulator* of G is a graph $\tilde{G} = (\tilde{E}, \tilde{V})$ such that there is a surjective simplicial map $\varphi : \tilde{G} \to G$ sending *surjectively outgoing edges* of each vertex $\tilde{v} \in \tilde{V}$ onto outgoing edges of the *image vertex* $\varphi(\tilde{v})$.



The following definition is the "directed version" of Fellows' graph emulator (in connection with the planar finite cover conjecture in the 1980s).

Definition

Let $G = (E, V)$ be a directed graph. A *directed emulator* of G is a graph $\tilde{G} = (\tilde{E}, \tilde{V})$ such that there is a surjective simplicial map $\varphi : \tilde{G} \to G$ sending *surjectively outgoing edges* of each vertex $\tilde{v} \in \tilde{V}$ onto outgoing edges of the *image vertex* $\varphi(\tilde{v})$.



Idea: the directed emulator map mimicks the canonical projection map between an automaton and its minimal automaton.

Idea: the directed emulator map mimicks the canonical projection map between an automaton and its minimal automaton.

Theorem

A language L has genus $\leq g$ iff (the underlying directed graph of) its minimal automaton $\mathtt{A}_{\min}$ has a directed emulator of genus $\leq g$.

Idea: the directed emulator map mimicks the canonical projection map between an automaton and its minimal automaton.

Theorem

A language L has genus $\leq g$ iff (the underlying directed graph of) its minimal automaton $\mathtt{A}_{\min}$ has a directed emulator of genus $\leq g$.

This leads to a "directed minor" approach to the computation of the genus of a language

Idea: the directed emulator map mimicks the canonical projection map between an automaton and its minimal automaton.

Theorem

A language L has genus $\leq g$ iff (the underlying directed graph of) its minimal automaton $\mathtt{A}_{\min}$ has a directed emulator of genus $\leq g$.

This leads to a "directed minor" approach to the computation of the genus of a language as an analogy to the Robertson-Seymour theorem for graphs.

Idea: the directed emulator map mimicks the canonical projection map between an automaton and its minimal automaton.

Theorem

A language L has genus $\leq g$ iff (the underlying directed graph of) its minimal automaton $\mathtt{A}_{\min}$ has a directed emulator of genus $\leq g$.

This leads to a "directed minor" approach to the computation of the genus of a language as an analogy to the Robertson-Seymour theorem for graphs.

More on this: come to Denis Kuperberg's talk.

The genus of regular languages is only the tip of the iceberg. Many other invariants inspired from low-dimensional topology and graph theory admit nontrivial reincarnations in the study of languages.

The genus of regular languages is only the tip of the iceberg. Many other invariants inspired from low-dimensional topology and graph theory admit nontrivial reincarnations in the study of languages.
Chromaticity: no

The genus of regular languages is only the tip of the iceberg. Many other invariants inspired from low-dimensional topology and graph theory admit nontrivial reincarnations in the study of languages.
Chromaticity: no
Star height: has a topological refinement. Existence of a hierarchy. Computability: unknown.

The genus of regular languages is only the tip of the iceberg. Many other invariants inspired from low-dimensional topology and graph theory admit nontrivial reincarnations in the study of languages.
Chromaticity: no
Star height: has a topological refinement. Existence of a hierarchy. Computability: unknown.
Graph width, *cohomology theories based on graphs*...