Topological RNA structures over one and two backbones

Christian M. Reidys

[1]University of Southern Denmark

2013

Outline

Collaborators

- Andersen, J.E.
- Penner, R.C.
- Waterman, M.
- Fenix W. Huang (folding, shapes)
- Thomas J.X. Li (bijection for genus 1-unicellular maps)
- Hillary S. Han (bijection for bicellular maps)

• Andersen, J.E., Huang, W.D., Penner, R.C. and Reidys, C.M., 2012, *Topology of RNA-RNA interaction structures*, JCB, Volume 19, Number 7, pp. 928â943.
• Andersen, J.E., Penner, R.C., Reidys, C.M. and Waterman, M.S., 2012, *Topological classification and enumeration of RNA structures by genus*, J. Math. Bio.,, DOI 10.1007/s00285-012-0594-x.
• Li, T.J.X. and Reidys, C.M., 2011, *The genus filtration of γ-structures*, Math Biosci. 2013 Jan; 241(1):24-33. doi: 10.1016/j.mbs.2012.09.006. Epub 2012 Sep 26.
• Reidys, C.M., Huang, W.D., Andersen, J.E., Penner, R.C., Stadler, P.F. and Nebel, M.E., 2011, *Topology and prediction of RNA pseudoknots*, Bioinformatics, **(27)**, 8, 1076-1085.

RNA structures as planar graphs and diagrams

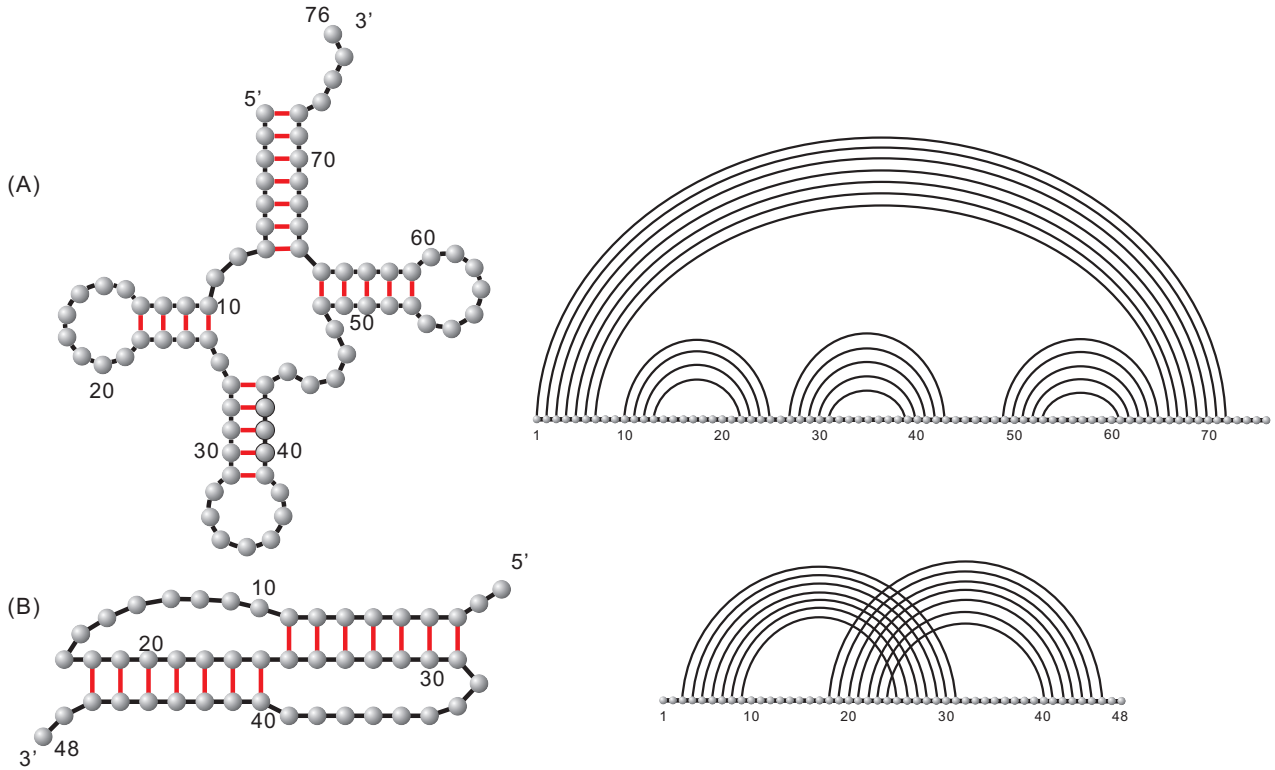
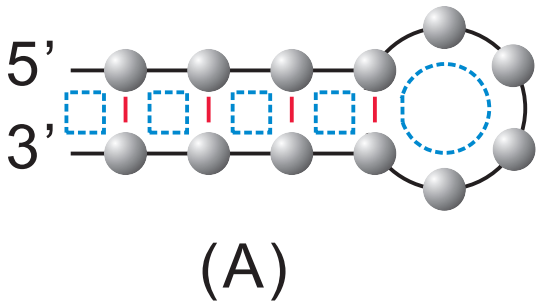


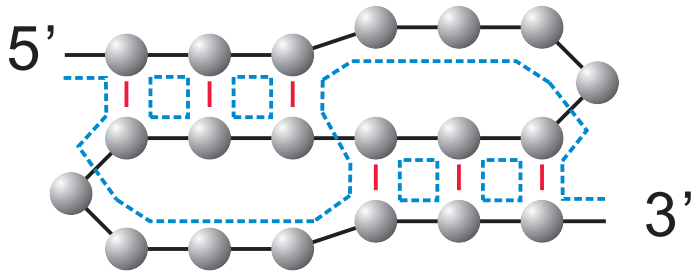
Figure: RNA structures as planar graphs and diagrams.

More than graphs...

- a secondary structure can be decomposed into “loops”,
- to specify a loop requires some kind of “orientation” i.e. how to “turn” around a vertex,
- its energy is loop-based depends on base pairs, bases and loop-type,
- pseudoknot structures have also a loop-decomposition
- energy is loop-based depends on base pairs, bases and loop-type or even simpler, when flat penalties of crossings are applied.



(A)



(B)

Loop-based folding: RNA pseudoknot

`Gfold`:

- Input: one sequence of length n.
- Type of structure: 1-structure, i.e., irreducible component has genus 1.
- Time complexity: $O(n^6)$.
- Space complexity: $O(n^4)$.
- Run time: 20 hours for 300nt, maximum 300nt.

`Sdufold`:

- Input: one or two sequence of length n.
- Type of structure: 1-structure, i.e., irreducible component has genus 1.
- Time complexity: $O(n^5)$ for shape (H) and (K), $O(n^6)$ for shape (L) and (M).
- Space complexity: $O(n^4)$.
- Run time: 30 min for 300nt, maximum 500nt.

Loop-based folding: RNA-RNA interaction structure

`rip2`:

- Input: two sequence of length n and m with $n \geq m$.
- Time complexity: $O(n^6)$.
- Space complexity: $O(n^4)$.

`Sdufold`:

- Input: two or two sequence of length n and m with $n \geq m$.
- Time complexity: $O((n+m)^5)$ for non-crossing hybrid, $O(n)^6$ for crossing hybrids.
- Space complexity: $O(n^4)$.

More than graphs: “fat”graphs

A **graph** consists of a set of half-edges, H, its **vertices** are **subsets** of half-edges and its **edges** are disjoint **pairs** of half-edges. A **fatgraph** consists of a set of half-edges, H its **vertices** are **cycles** of half-edges and its **edges** are disjoint **pairs** of half-edges. Thus, a fatgraph is given by (H, σ, α), where σ is the vertex-permutation and α a fixed-point free involution.

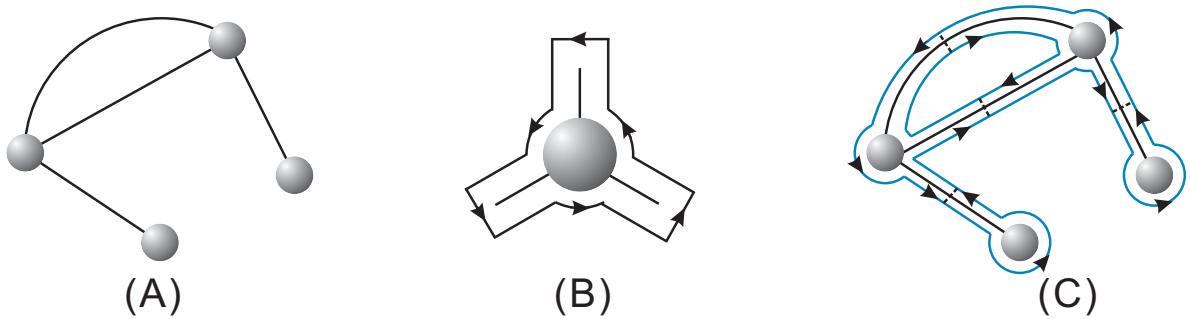


Figure: (A) a graph with 4 vertexes and 4 edges, (B) fattening of a vertex, (C) a fatgraph derived from (A). Any fatgraph induces a topological surface.

Fatgraphs in the computer

A fatgraph having n edges can be presented by

- the vertex permutation σ and the involution α, representing the arcs,
- we can consider the permutation $\gamma = \alpha \circ \sigma$, whose cycles are called **boundary components**.

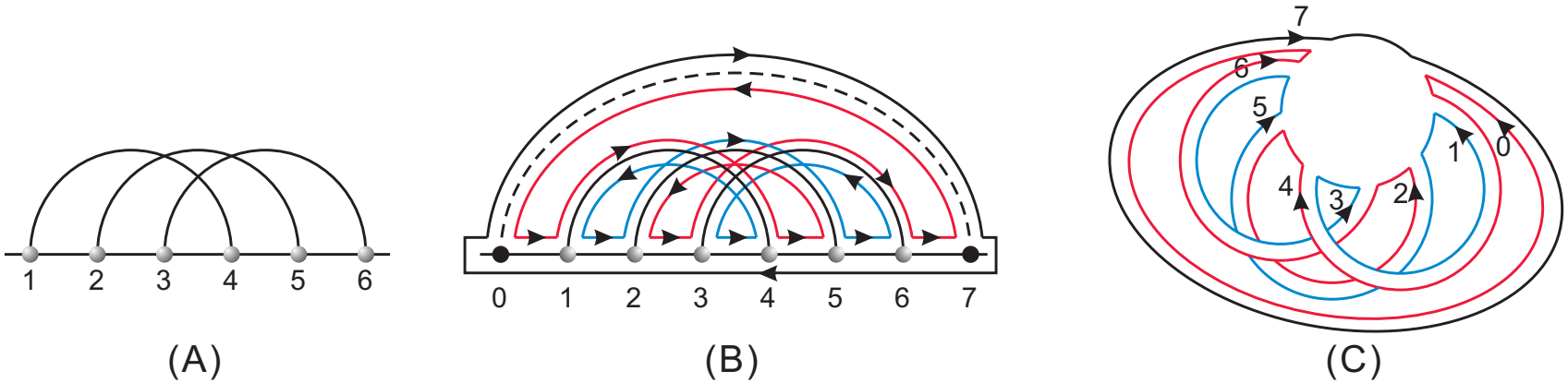
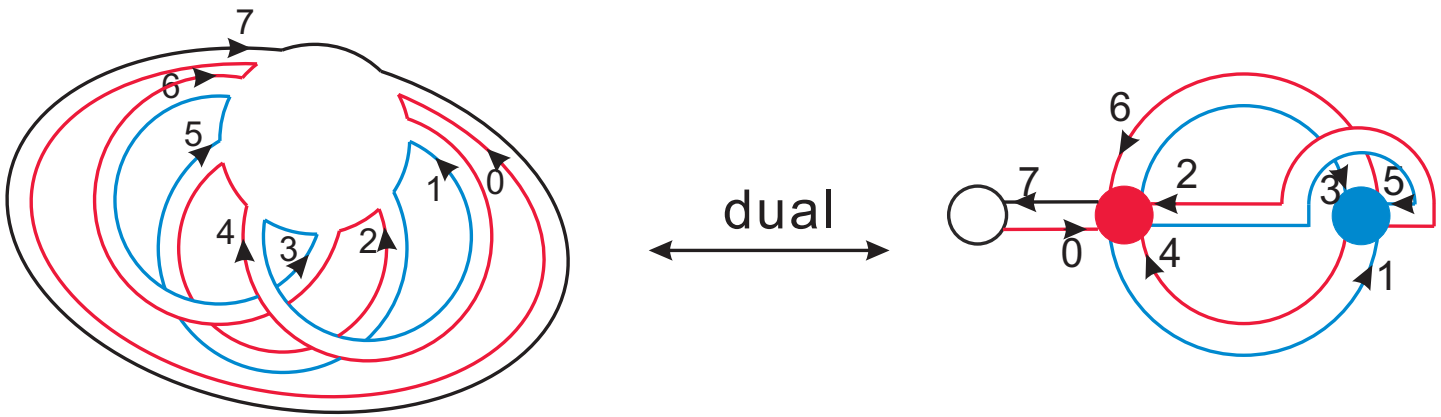


Figure: (A) A diagram, (B) a fatgraph of (A) augmented by an additional "rainbow" arc $(0, 7)$. (C) collapsing the backbone. Here $\gamma = \alpha \circ \sigma = (0, 4, 2, 6)(1, 5, 3)(7)$ has two cycles.

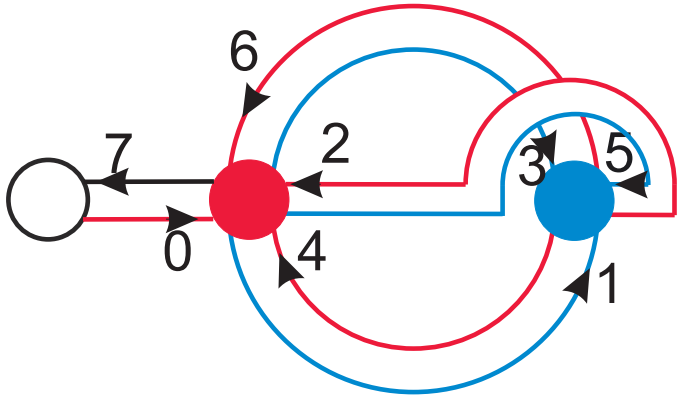
The Poincaré dual

Poincaré dual: Mapping a fatgraph (σ, α) to $(\alpha \circ \sigma, \alpha)$.



Unicellular maps

A fatgraph with **one** boundary component is called a *unicellular map*. A *planted* unicellular map contains an additional vertex of degree one serving as its distinguished root.



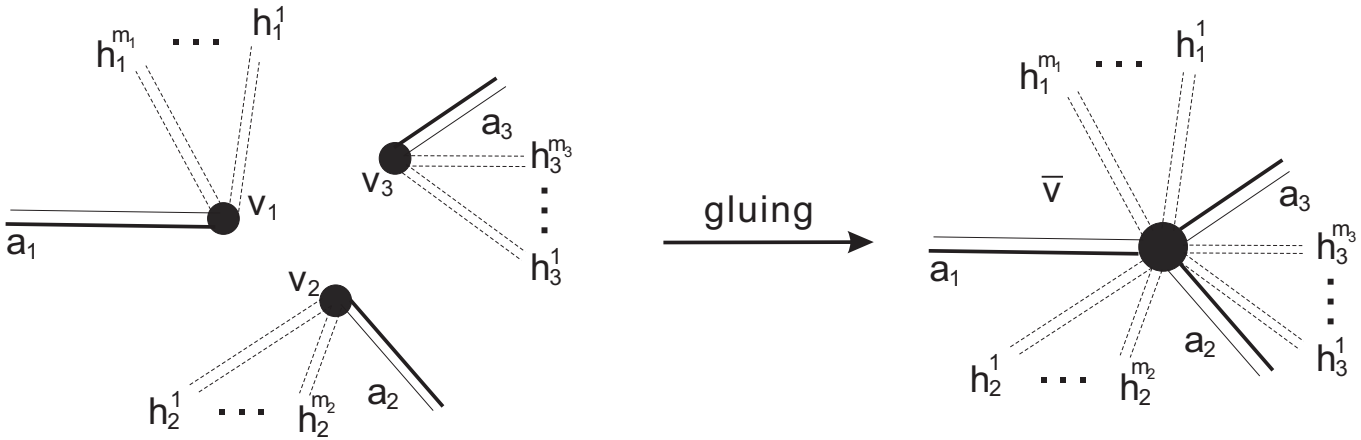
Two orders

- The tour of the unique boundary component. We write $a_1 <_\gamma a_2$ if a_1 appears before a_2 in this tour.
- The order of the half-edges induced by the vertex-cycle. We call $a_1 <_\sigma a_2$ if a_1 appears before a_2 counterclockwise in the vertex.
- suppose three half-edges $a_1 <_\gamma a_2 <_\gamma a_3$ are contained in one vertex v. Then a_1, a_2, a_3 are **intertwined** iff $a_1 <_\sigma a_3 <_\sigma a_2$ holds.

Gluing

Lemma

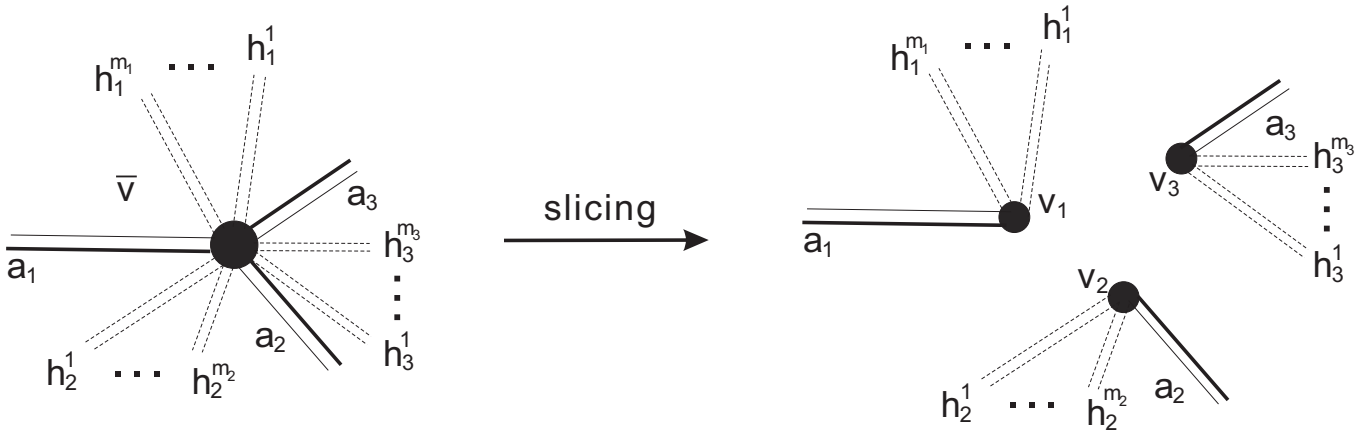
Chapuy (2011) *Given* $\mathfrak{m}_g$ *with* $\gamma = (w_A, a_1, w_B, a_2, w_C, a_3, w_D)$. *Gluing three half-edges* $a_1 <_\gamma a_2 <_\gamma a_3$ *belonging to three distinct vertices* $v_i = (a_i, h_i^1, \ldots, h_i^{m_i})$, $i = 1, 2, 3$, *into* $\overline{v} = (a_1, h_2^1, \ldots, h_2^{m_2}, a_2, h_3^1, \ldots, h_3^{m_3}, a_3, h_1^1, \ldots, h_1^{m_1})$ *generates* $\mathfrak{m}_{g+1}$, *with* $\overline{\gamma} = (w_A, a_1, w_C, a_3, w_B, a_2, w_D)$. a_1, a_2, a_3 *are intertwined in* $\overline{v}$.



Slicing

Lemma

Chapuy (2011) *Consider* $\mathfrak{m}_{g+1}$*, with* $\overline{\gamma} = (w_A, a_1, w_C, a_3, w_B, a_2, w_D)$ *and* $\overline{v} = (a_1, h_2^1, \ldots, h_2^{m_2}, a_2, h_3^1, \ldots, h_3^{m_3}, a_3, h_1^1, \ldots, h_1^{m_1})$*, in which* a_1, a_2, a_3 *are intertwined. Then slicing* $\overline{v}$ *into* $v_i = (a_i, h_i^1, \ldots, h_i^{m_i})$ *produces* $\mathfrak{m}_g$ *with* $\gamma = (w_A, a_1, w_B, a_2, w_C, a_3, w_D)$.



Trisection types

Definition

Let $\mathcal{D}_{g+1}$ be the set of unicellular maps $\overline{\mathfrak{m}}$ with a labeled trisection τ in $V(\tau)$. Suppose

$$V(\tau) = (a_1, h_2^1, \ldots, h_2^{m_2}, a_2, h_3^1, \ldots, h_3^{m_3}, a_3, h_1^1, \ldots, h_1^{m_1}),$$

where $a_1 = \min V(\tau)$, $a_3 = \sigma(\tau)$ and a_2 is the smallest half-edge that preceeds a_3 in $V(\tau)$ and that is greater than a_3. Then a_1, a_2, a_3 are interwined in $V(\tau)$ and τ is of type I, iff $a_3 = \min v_3$ after splicing and of type II, otherwise.

Bijection for trisections of type I

Theorem

Let $\mathcal{V}_g^1(n)$ denotes the set of unicellular maps with three labeled vertices. We have the bijection

$$\Gamma_1 : \mathcal{V}_g^1(n) \to \mathcal{D}_{g+1}^1(n), \quad \Gamma(\mathfrak{m}, v_1, v_2, v_3) = (\overline{\mathfrak{m}}, \tau),$$

where $\mathcal{D}_{g+1}^1$ is the set of unicellular maps with labeled trisection of type I.

Bijection for trisection of type II

Theorem

Let $\mathcal{V}_g^2(n)$ denote the set of unicellular maps with two labeled vertices and a trisection. Then we have the bijection

$$\Gamma_2 \colon \mathcal{V}_g^2(n) \to \mathcal{D}_{g+1}^2(n), \quad \Gamma_2(\mathfrak{m}, v_1, v_2, \tau) = (\overline{\mathfrak{m}}, \tau),$$

where $v_1, v_2, V(\tau)$ are $\mathfrak{m}$-vertices such that

$$\min v_1 <_{\mathfrak{m}} \min v_2 <_{\mathfrak{m}} \min V(\tau).$$

Note that $\mathcal{D}_{g+1}^2(n)$ is the set of unicellular maps with labeled trisection τ whose subsequent half-edge a_3 is not minimal in $v_3 = (a_3, h_1^3, \ldots, h_{m_3}^3)$ after slicing.

Trisections

Lemma

Chapuy (2011) *A planted unicellular map having n edges and genus g contains* $n+1$ *down-steps,* $n+1$ *up-steps and* $2g$ *trisections.*

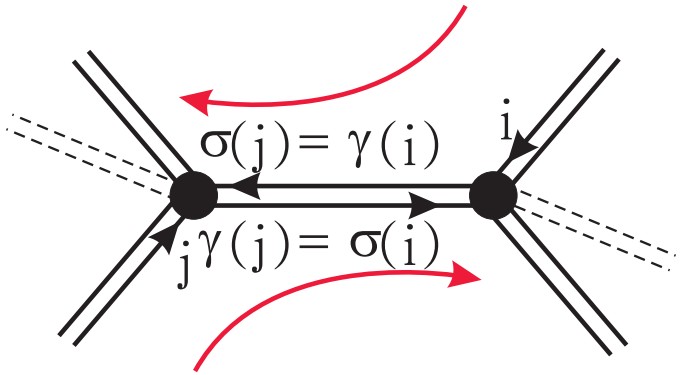


Figure: The tour visits i before $\sigma(i)$ if and only if it visits $\sigma(j)$ before j, and vice versa.

An enumerative corollary

Corollary

Let $U_g(n)$ denote the set of unicellular maps of genus g having n edges. Then $u_g(n) = |U_g(n)|$ satisfies

$$2g \cdot u_g(n) = \binom{n+3-2g}{3} u_{g-1}(n) + \binom{n+5-2g}{3} u_{g-2}(n) + \cdots + \binom{n+1}{2g+1} u_0(n).$$

Genus 1: towards a “real” bijection

Here we have just type I trisections and hence

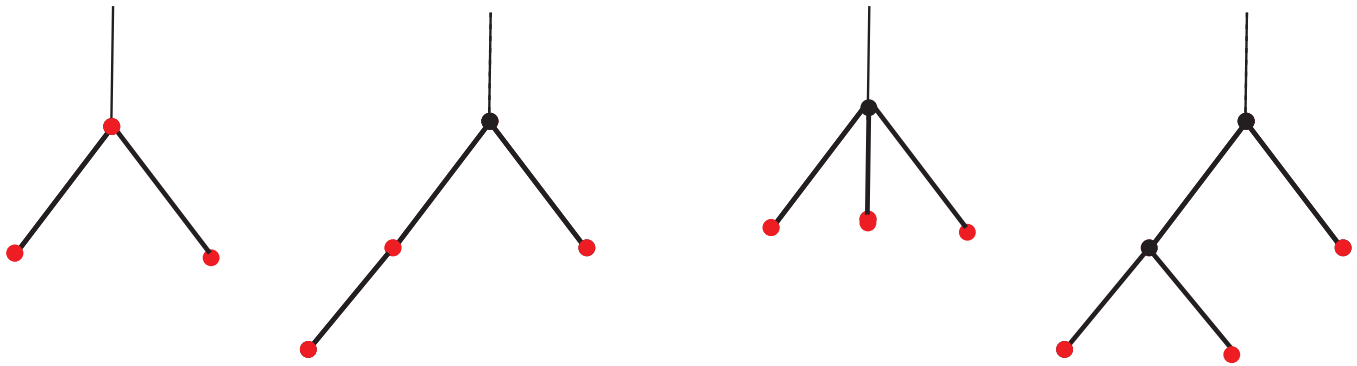
$$2\varepsilon_1(n) = \binom{n+1}{3}\varepsilon_0(n).$$

Is there a “canonical” choice–and if so what trees are being generated?

Genus 1: towards a “real” bijection cont.

Definition

Three different vertices v_1, v_2 and v_3 in a planar tree, where $\min(v_1) < \min(v_2) < \min(v_3)$, are *maximal* if $LCA(v_2, v_3)$ is an ancestor of $LCA(v_1, v_2)$.



Genus 1: its the maximal trees

Lemma

There exists a bijection from the set of planted plane trees with n edges and three maximum (minimum) distinguished vertices to the set of planted unicellular maps of genus 1 with n edges and the distinguished maximum (minimum) trisection.

Genus 1: maximality is “nice”

We next generalize a bijection of planted plane trees leading to

$$(n+1)\varepsilon_0(n) = 2(2(n-1)+1)\varepsilon_0(n-1)$$

to plane trees with k distinguished non-root vertices, $\mathcal{T}^{(k)}(n)$, where $t^{(k)}(n) = |\mathcal{T}^{(k)}(n)|$. Thus $t^{(k)}(n) = \binom{n+1}{k}\varepsilon_0(n)$.

Lemma

We have

$$(n+1-k)t^{(k)}(n) = 2(2n-1)t^{(k)}(n-1)$$

via a bijection that respects maximality.

Genus 1 cont.

Theorem

(Li & Reidys, 2012) *There is a bijection between the set of planted unicellular maps of genus* 1 *with n edges and the set of rooted planar trees with n edges and three maximal distinguished vertices. Furthermore we have the recursion*

$$(n-2)\varepsilon_1(n) = 2(2n-1)\varepsilon_1(n-1).$$

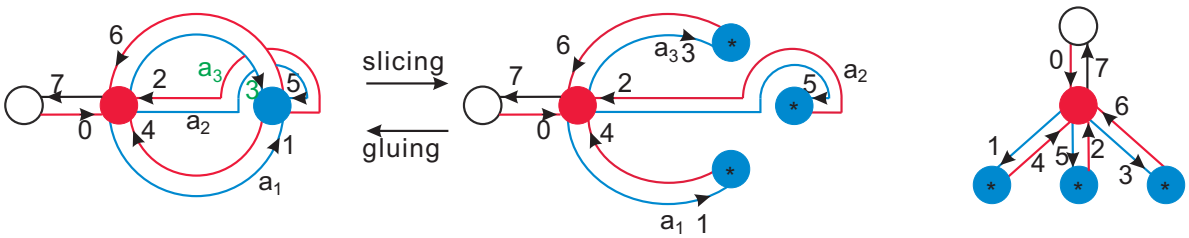
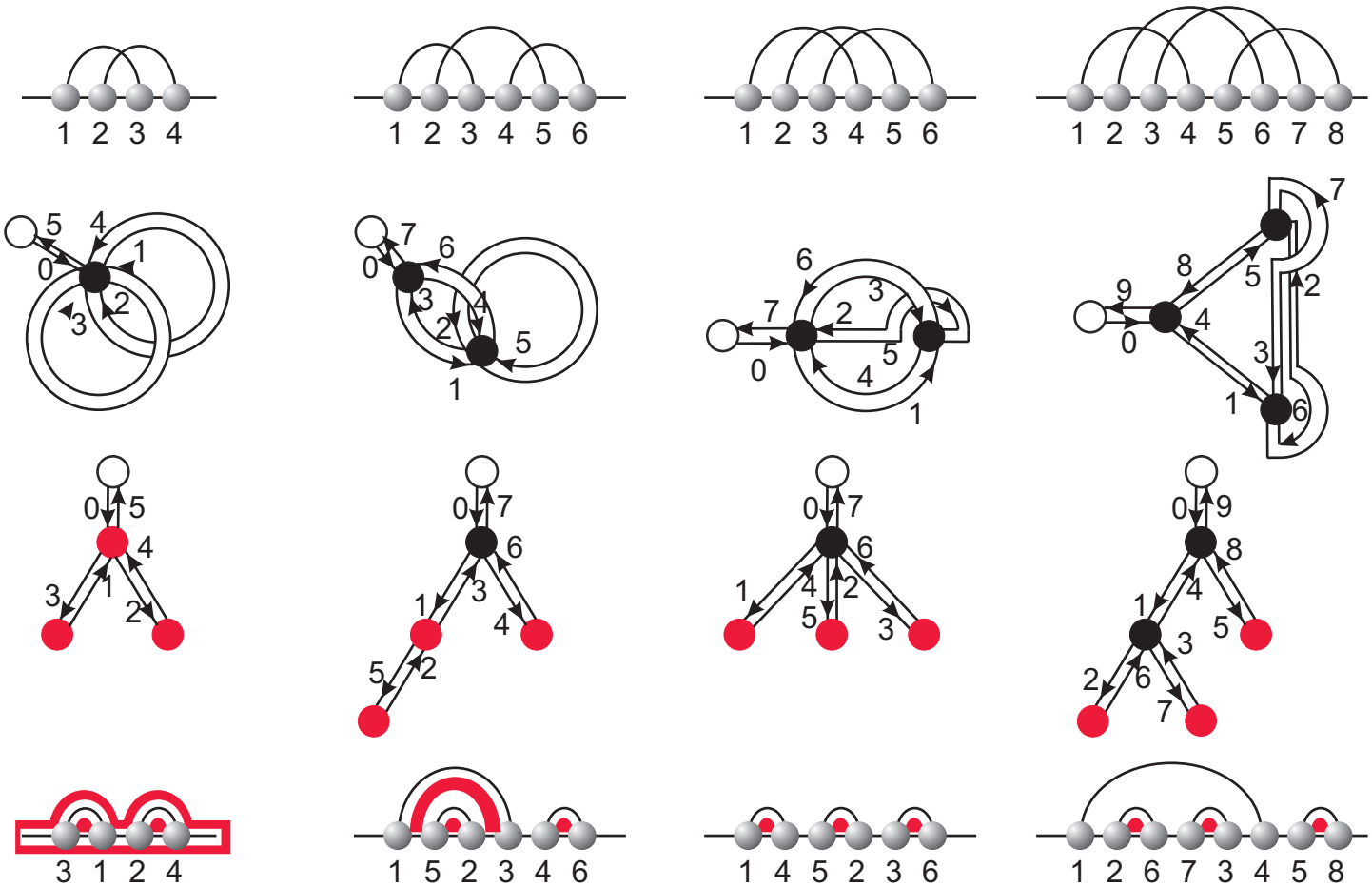


Figure: A unicellular map of genus 1 is sliced into a planar tree. Half-edge 3 is the maximal trisection.

An overview of the genus 1 case



Shapes

A shape is a diagram that

- contains no isolated vertices and 1-arcs,
- no “parallel” arcs i.e. each stack consists of a single arc.

Lemma

There exist only finitely many shapes of fixed genus g. Let S be a shape of genus g having n arcs and $m_g^S(n)$ is the planted unicellular map associated to S. Then each $m_g^S(n)$-vertex contains at least three half-edges.

Irreducible shapes

An irreducible shape is a shape that can not decomposed w.r.t. concatenation and nesting.

Lemma

There are only four irreducible shapes of genus 1.

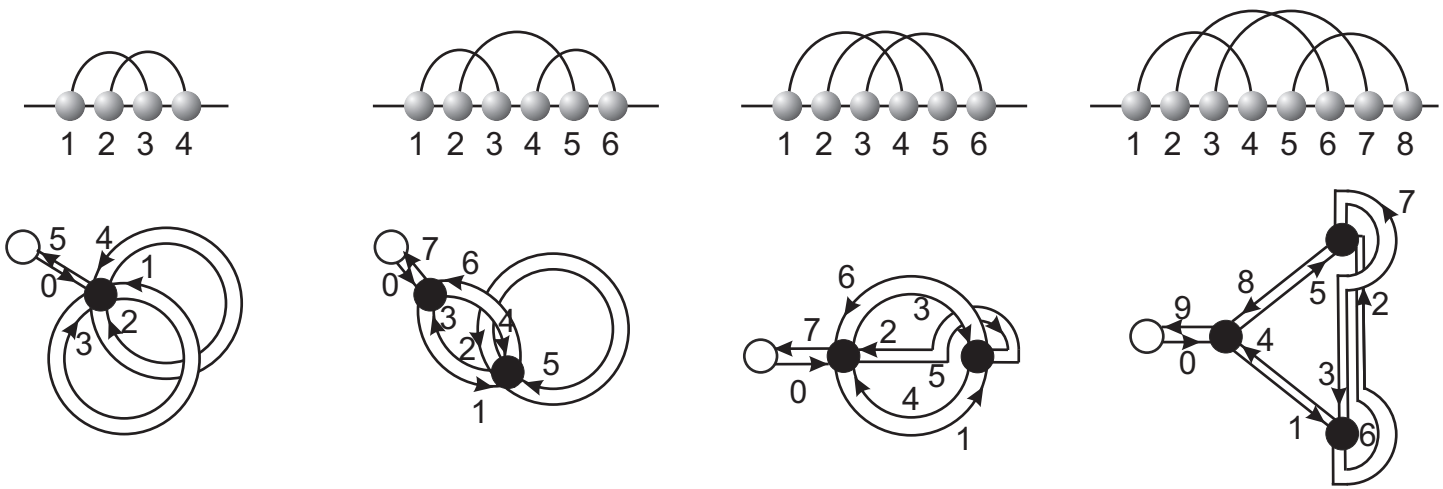


Figure: The four irreducible shapes of genus 1 and their corresponding unicellular maps.

Labeled shapes

Definition

Let $U_g^S(n,m)$ denote the set of unicellular maps of genus g having n edges with m labeled vertices such that each **unlabeled** vertex contains at least three half-edges.

We shall consider vertices derived from slicings as labeled. Thus, slicing a shape having n edges produces an element of $U_g^S(n,m)$ for some g. Vice versa, any unlabeled vertex obtained by gluing contains always at least three half-edges.

Shapes with labeled vertices: a recursion

Let $u_g^S(n, m_1, m_2)$ denote the set of shapes containing $m_1 + m_2$ labeled vertices partitioned into m_1 blue and m_2 red vertices. Then

$$u_g^S(n, m_1, m_2) = \binom{m_1 + m_2}{m_1} u_g^S(n, m_1 + m_2),$$

Lemma

$$\begin{aligned} 2g \cdot u_g^S(n, m) &= \sum_{k=1}^{g} \binom{m + 2k + 1}{2k + 1} u_{g-k}^S(n, m + 2k + 1) \\ &\quad + \sum_{k=1}^{g} \binom{m + 2k}{2k + 1} u_{g-k}^S(n, m + 2k). \end{aligned}$$

Shapes with labeled vertices: reduction to trees

Lemma

$$u_g^S(n,0) = \sum_{t=1}^{p} a_g(t) u_0^S(n, 2g+t),$$

where $p = g$ *when* $n \geq 3g$ *and* $p = n - 2g + 1$ *when* $n < 3g$.

$$a_g(t) = \sum_{\substack{0=g_0<g_1<\cdots<g_r=g \\ 0=t_0=t_1\leq t_2\leq\cdots\leq t_r=r-t}} \prod_{i=1}^{r} \frac{1}{2g_i} \binom{2g+t-(2g_{i-1}+(i-1))+t_i}{2(g_i-g_{i-1})+1}$$

where $(t_i)_{r-t}$ *is a sequence s.t.* $t_0 = t_1 = 0$, $t_r = r - t$ *and* $t_i - t_{i-1} = 0$ *or* 1, $\forall 1 \leq i \leq r$. $(2g+t)$ *is the number of labeld vertices of* $u_0^S(n, 2g+t)$, $1 \leq t \leq g$.

Rewriting $u_0^S(n, 2g + t)$: Rémy's bijection

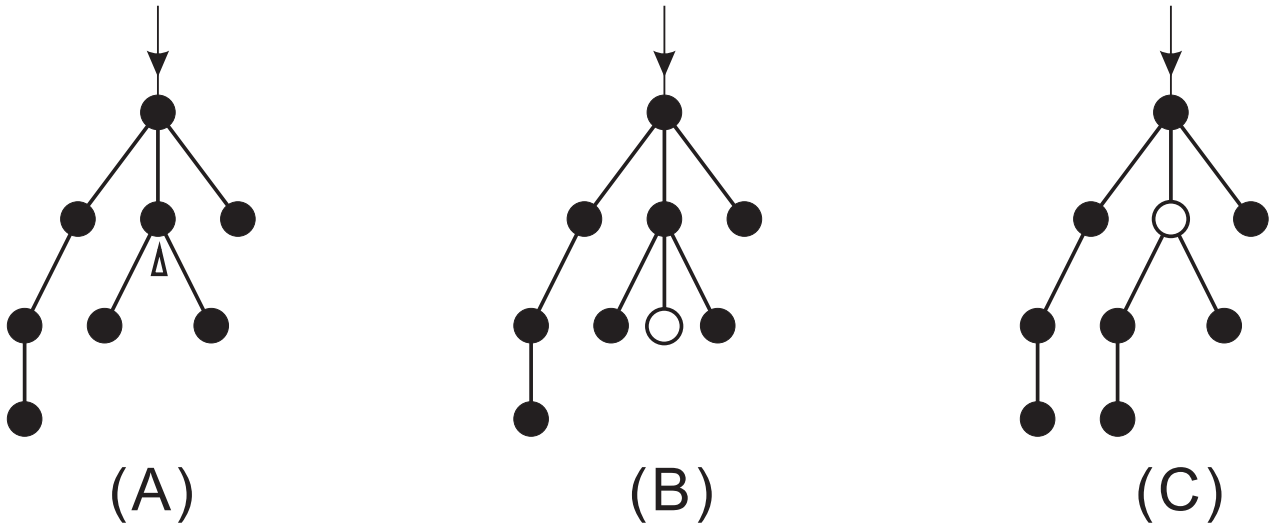


Figure: Rémy's procedure gives two ways of obtaining a planar tree with n edges and a labeled vertex from a planar tree with $n-1$ edges with a labeled sector. (A) to (B) inserting a labeled vertex as a leave to the labeled sector. (A) to (C) replacing the vertex containing the sector by the labeled vertex, and carrying the subtree on the left of the sector as the left most subtree. In this case, the labeled vertex is not a leave.

Shape preserving sectors

Lemma

Let $u_0^S(n,m) \in U_0^S(n,m)$, then there are $2m-n-2$ shape preserving sectors in $u_0^S(n,m)$. Rémy's edge-insertion produces here an element $u_0^S(n+1,m) \in U_0^S(n+1,m)$.

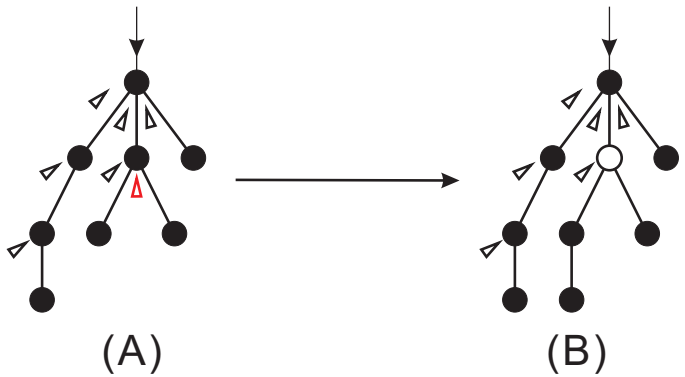


Figure: Illustration of Rémy's procedure inserting a non-leave vertex to a tree. The marked sectors are shape preserving. One insertion of a vertex that has at least 3 half-edges removes one shape preserving sector.

Computing $u_0^S(n,m)$

Corollary

In $u_0^S(m-1,m)$ *we have* $2m-(m-1)-2 = m-1$ *shape preserved sectors. Inserting* $n-(m-1)$ *unlabeled vertices we have*

$$u_0^S(n,m) = \binom{m-1}{n-m+1} u_0^S(m-1,m) = \binom{m-1}{n-m+1} Cat(m-1),$$

where $Cat(n)$ *is the n-th Catalan number given by* $\frac{1}{n+1}\binom{2n}{n}$.

The shape polynomial

Theorem

The shape generating function is given by

$$S_g(z) = \sum_{t=1}^{g} \kappa_g(t)[z(1+z)]^{2g+t-1},$$

where $\kappa_g(t)$ *is the coefficient given by*

$$\kappa_g(t) = a_g(t)Cat(2g+t-1).$$

Shapes and $\kappa_g(t)$

Observing the shape polynomial, for fixed genus g we have

$$u_g^S(n) = \binom{2g}{n-2g}\kappa_g(1) + \binom{2g+1}{n-2g-1}\kappa_g(2) + \cdots + \binom{n}{0}\kappa_g(i),$$

$\forall n \leq 3g$, and

$$u_g^S(n) = \binom{2g}{n-2g}\kappa_g(1) + \binom{2g+1}{n-2g-1}\kappa_g(2) + \cdots + \binom{n}{n-3g}\kappa_g(g),$$

$\forall n > 3g$.

The analogue for unicellular maps

Corollary

Let $U_g(n)$ be the set of unicellular map of genus g having n edges. Then we have

$$u_g(n) = \sum_{t=1}^{p} \kappa_g(t) \frac{\prod_{i=1}^{n-(2g+t-1)} 2(2(n-i)+1)}{(n-(2g+t-1))!}.$$

Here $p = g$ when $n \geq 3g$ and $p = n - 2g + 1$ when $n < 3g$.

Planted bicellular maps

Definition

A planted bicellular map b having n edges is a triple $b = (L, \beta, \tau)$, where L is a set of cardinality $(2n + 4)$,

$$L = \{1_R, 1, \ldots, m, m_R, (m+1)_R, m+1, \ldots, 2n, 2n_R\},$$

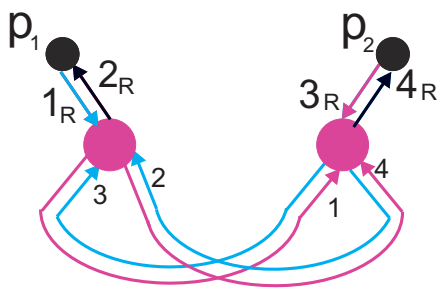
where $1 < m < 2n - 1$. β is a fixed-point free involution containing the cycles $(1_R, m_R)$ and $((m+1)_R, 2n_R)$. $\beta \circ \tau = \omega_1 \circ \omega_2$, where

$$\begin{aligned} \omega_1 &= (1_R, 1, 2, \ldots, m, m_R), \\ \omega_2 &= ((m+1)_R, m+1, m+2, \ldots, 2n, 2n_R) \end{aligned}$$

and there exists some half-edge $x \in \omega_1$, such that $\beta(x) \in \omega_2$

Planted bicellular maps cont.

- the cycles $\beta \setminus \{(1_R, m_R) \cup ((m+1)_R, 2n_R)\}$ are the edges; $\tau \setminus \{(m)_R, (2n_R)\}$ are the vertices of b,
- ω_1 and ω_2 are the faces,
- the cycles $p_1 = (m_R)$ and $p_2 = (2n_R)$ are the two plants.



(a)

$L = [1_R, 1, 2, 2_R, 3_R, 3, 4, 4_R]$

$\beta = (1_R, 2_R)(1, 3)(2, 4)(3_R, 4_R)$

$\tau = (1_R, 3, 2)(3_R, 1, 4)(2_R)(4_R)$

$\omega = \omega_1\omega_2 = [1_R, 1, 2, 2_R][3_R, 3, 4, 4_R]$

(b)

Figure: A bicellular map with 2 edges, 2 vertices, and genus 0 as: (a) ribbon graph; (b) pair of permutations.

Diagram, fat graph, planted bicellular map

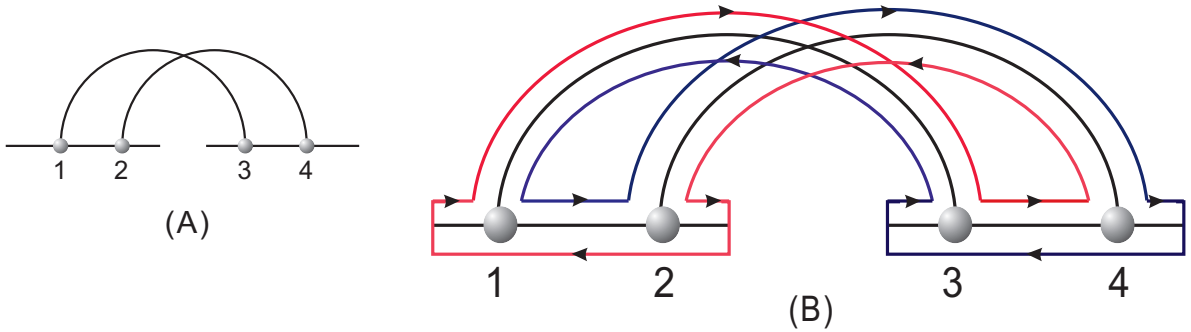


Figure: A diagram with 2 arcs over two backbones and its fattening (B). $\omega = \beta \circ \tau = (1_R, 3, 2)(3_R, 1, 4)(2_R)(4_R)$ are its two boundary components (red) (blue).

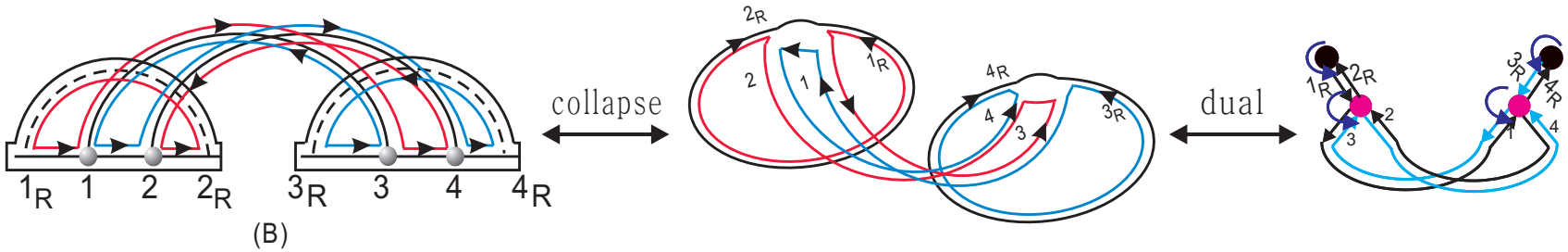


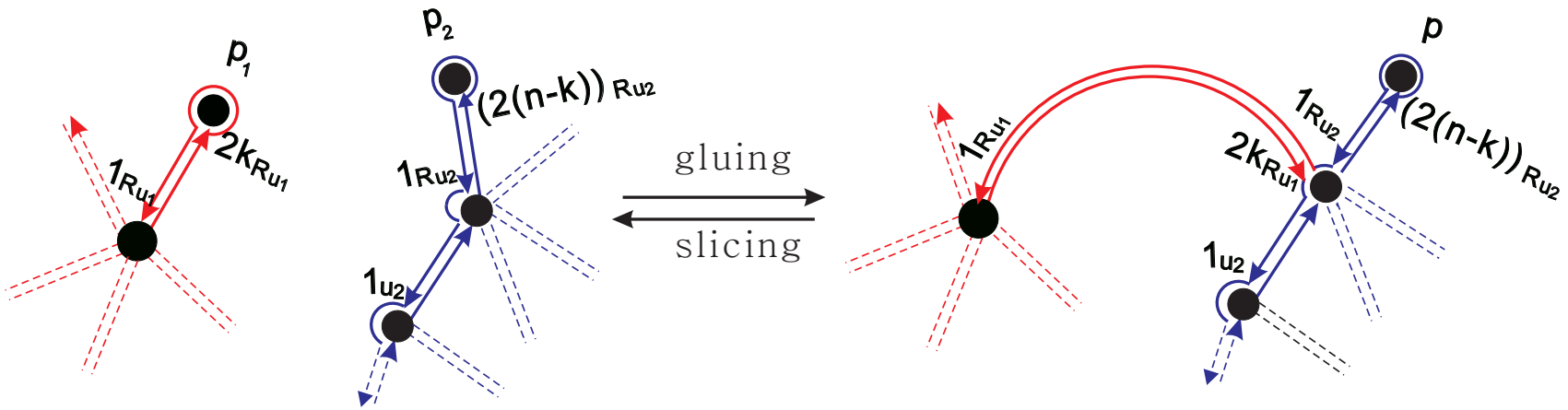
Figure: A fattened diagram over two backbones, backbone collapse

The first lemma

Lemma

There is a bijection

$$\theta : \dot{\bigcup}_{0\leq g_1\leq g+1,\ 0\leq j\leq n} \left(U_{g_1,j} \times U_{g+1-g_1,n-j}\right) \longrightarrow U^{III}_{g+1,n+1}.$$



The second lemma

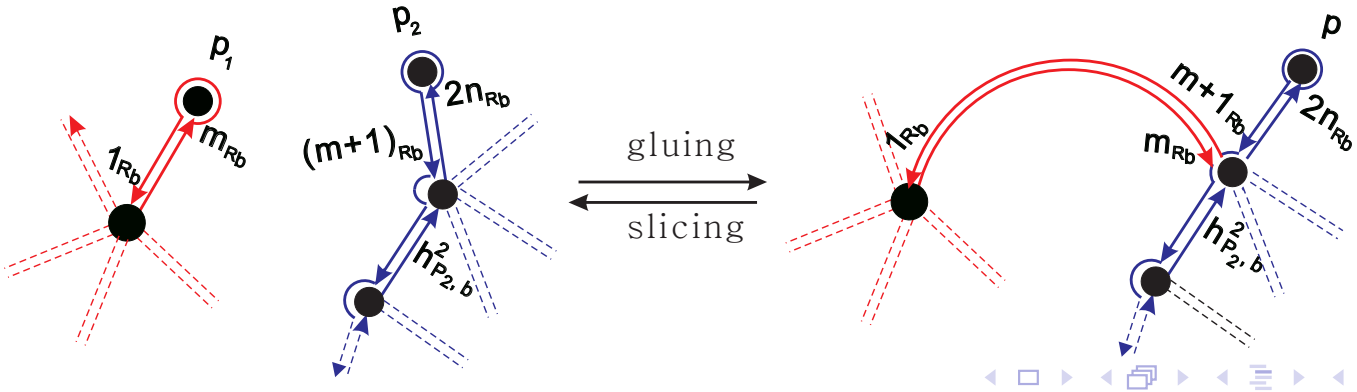
Lemma

There exists a bijection

$$\eta : B^I_{g,n} \dot\cup B^{II}_{g,n} \longrightarrow U^I_{g+1,n+1} \dot\cup U^{II}_{g+1,n+1},$$

and η induces by restriction the two bijections

$$\eta_I : B^I_{g,n} \longrightarrow U^I_{g+1,n+1} \quad \text{and} \quad \eta_{II} : B^{II}_{g,n} \longrightarrow U^{II}_{g+1,n+1}.$$



The case $B^{II}_{g,n+1}$

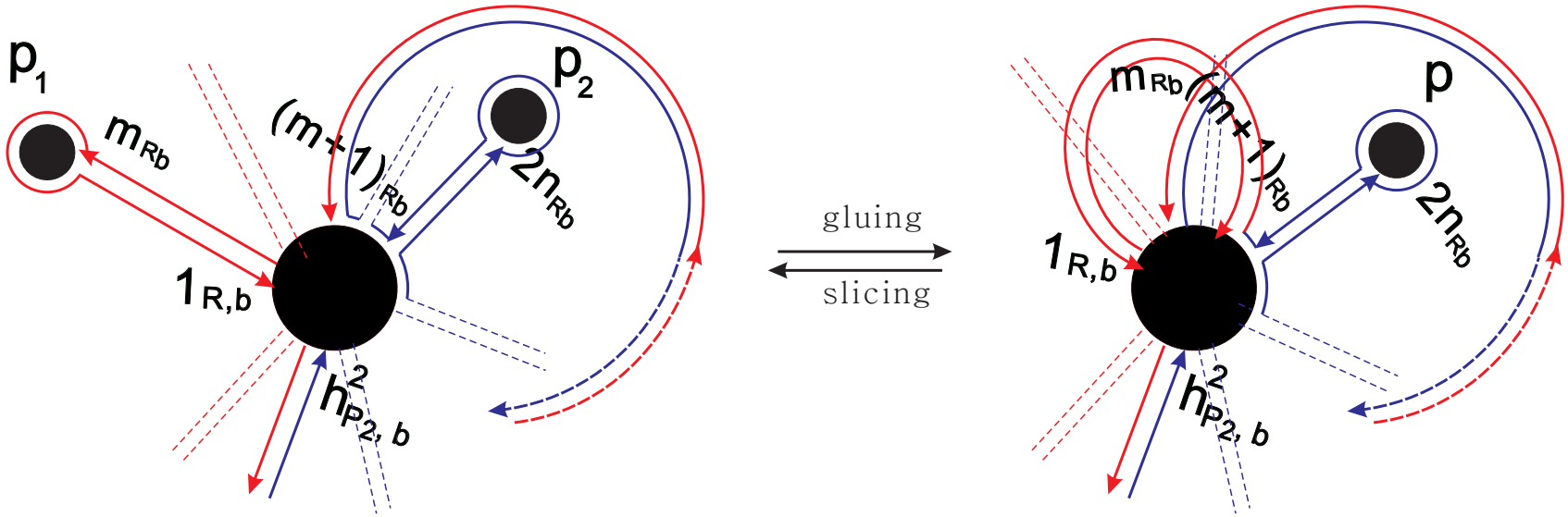


Figure: Lemma 23: gluing and slicing, the case $B^{II}_{g,n+1}$.

The bijection

Theorem

Let $U_{g,n}$ and $B_{g,n}$ denote the sets of unicellular and bicellular maps containing n edges and genus g. Then there is a bijection

$$\beta : \dot{\bigcup}_{0\leq g_1\leq g+1,\ 0\leq j\leq n} \left(U_{g_1,j} \times U_{g+1-g_1,n-j}\right) \dot{\cup} B_{g,n} \longrightarrow U_{g+1,n+1}. \tag{1}$$

An enumerative corollary

Corollary

The generating function of unicellular and bicellular maps, $\mathbf{C}_g(z)$ *and* $\mathbf{C}_g^{[2]}(z)$ *satisfy the following equation*

$$\sum_{g_1=0}^{g+1} \mathbf{C}_{g_1}(z)\mathbf{C}_{g+1-g_1}(z) + \mathbf{C}_g^{[2]}(z) = \mathbf{C}_{g+1}(z)/z \qquad (2)$$