Families of dendrograms

Patrick Erik Bradley

1. Introduction

- Dendrograms are ultrametric spaces

- Ultrametricity is pervasive (F. Murtagh)

- p-adic geometry = natural environment for ultrametricity

2. Some p-adic geometry

2.1. Eucledian vs. p-adic geometry

Euclidean geometry.

- modelled on the real numbers $\mathbb{R}$

2. Some p-adic geometry

2.1. Eucledian vs. p-adic geometry

Euclidean geometry.

- modelled on the real numbers $\mathbb{R}$
- decimal system

$$\sum_{\nu=m}^{\infty} a_\nu 10^{-\nu}, \quad a_\nu \in \{0,\ldots,9\}$$

2. Some p-adic geometry

2.1. Eucledian vs. p-adic geometry

Euclidean geometry.

- modelled on the real numbers $\mathbb{R}$
- decimal system

$$\sum_{\nu=m}^{\infty} a_\nu 10^{-\nu}, \quad a_\nu \in \{0, \dots, 9\}$$

- $\mathbb{R}$ = completion of rationals $\mathbb{Q}$ w.r.t. absolute norm:

$$|x| = \begin{cases} x, & x \geq 0 \\ -x, & x < 0 \end{cases}$$

2. Some p-adic geometry

2.1. Eucledian vs. p-adic geometry

p-adic geometry.

Let p be a prime number.

2. Some p-adic geometry

2.1. Eucledian vs. p-adic geometry

p-adic geometry.

Let p be a prime number.

- p-adic norm on $\mathbb{Q}$: $|x|_p = \begin{cases} p^{-\nu_p(x)}, & x \neq 0 \\ 0 & \text{otherwise} \end{cases}$ $x = \frac{a}{b}$, $a, b \in \mathbb{Z}$,

  $\nu_p(x) = \nu_p(a) - \nu_p(b)$.

  $\nu_p(n) =$ multiplicity with which p divides integer n:

  $n = p^{\nu_p(n)} \cdot u, \quad p \not| \, u.$

2. Some p-adic geometry

2.1. Eucledian vs. p-adic geometry

p-adic geometry.

Let p be a prime number.

- p-adic norm on $\mathbb{Q}$: $|x|_p = \begin{cases} p^{-\nu_p(x)}, & x \neq 0 \\ 0 & \text{otherwise} \end{cases}$ $x = \frac{a}{b}$, $a, b \in \mathbb{Z}$,

  $\nu_p(x) = \nu_p(a) - \nu_p(b)$.

  $\nu_p(n) =$ multiplicity with which p divides integer n:

  $n = p^{\nu_p(n)} \cdot u, \quad p \not| u$.

- $\mathbb{Q}_p =$ completion w.r.t. $|\cdot|_p$

$$\Rightarrow x = \sum_{\nu=m}^{\infty} a_\nu p^\nu, \quad a_\nu \in \{0, \ldots, p-1\}$$

  p-adic expansion of p-adic numbers.

Remark. $|\cdot|_p$ is an ultrametric:

$$|x|_p \geq 0, \quad \text{and} \quad |x|_p = 0 \Leftrightarrow x = 0 \tag{1}$$

$$|x \cdot y|_p = |x|_p \cdot |y|_p \tag{2}$$

$$|x + y|_p \leq \max\{|x|_p, |y|_p\} \tag{3}$$

The last property is the *ultrametric triangle inequality*.

2.2. The Bruhat-Tits tree for $\mathbb{Q}_p$

Unit disc: $\mathbb{D} = \{x \in \mathbb{Q}_p \mid |x|_p \leq 1\} = \mathbb{Z}_p$ is a ring.

Arbitrary closed disc: $B_{p^{-r}}(a) = \{x \in \mathbb{Q}_p \mid |x - a|_p \leq p^{-r}\}$

2.2. The Bruhat-Tits tree for $\mathbb{Q}_p$

Unit disc: $\mathbb{D} = \{x \in \mathbb{Q}_p \mid |x|_p \leq 1\} = \mathbb{Z}_p$ is a ring.

Arbitrary closed disc: $B_{p^{-r}}(a) = \{x \in \mathbb{Q}_p \mid |x - a|_p \leq p^{-r}\}$

The max. strict subdiscs of $\mathbb{D}$: $B_{p^{-1}}(0), B_{p^{-1}}(1), \ldots, B_{p^{-1}}(p-1)$ are a partition of $\mathbb{D}$.
Reason: $\mathbb{Z}_p/p\mathbb{Z}_p \cong \mathbb{F}_p = \mathbb{Z}/p\mathbb{Z} = \{0, \ldots, p-1\}$, and residue classes are discs.

2.2. The Bruhat-Tits tree for $\mathbb{Q}_p$

Unit disc: $\mathbb{D} = \{x \in \mathbb{Q}_p \mid |x|_p \leq 1\} = \mathbb{Z}_p$ is a ring.

Arbitrary closed disc: $B_{p^{-r}}(a) = \{x \in \mathbb{Q}_p \mid |x - a|_p \leq p^{-r}\}$

The max. strict subdiscs of $\mathbb{D}$: $B_{p^{-1}}(0), B_{p^{-1}}(1), \ldots, B_{p^{-1}}(p-1)$ are a partition of $\mathbb{D}$.
Reason: $\mathbb{Z}_p/p\mathbb{Z}_p \cong \mathbb{F}_p = \mathbb{Z}/p\mathbb{Z} = \{0, \ldots, p-1\}$, and residue classes are discs.

One minimal disc strictly containing $\mathbb{D}$: $B_p(0) = \{x \in \mathbb{Q}_p \mid |x|_p \leq p\}$.

Define a graph $\mathscr{T}_{\mathbb{Q}_p}$.

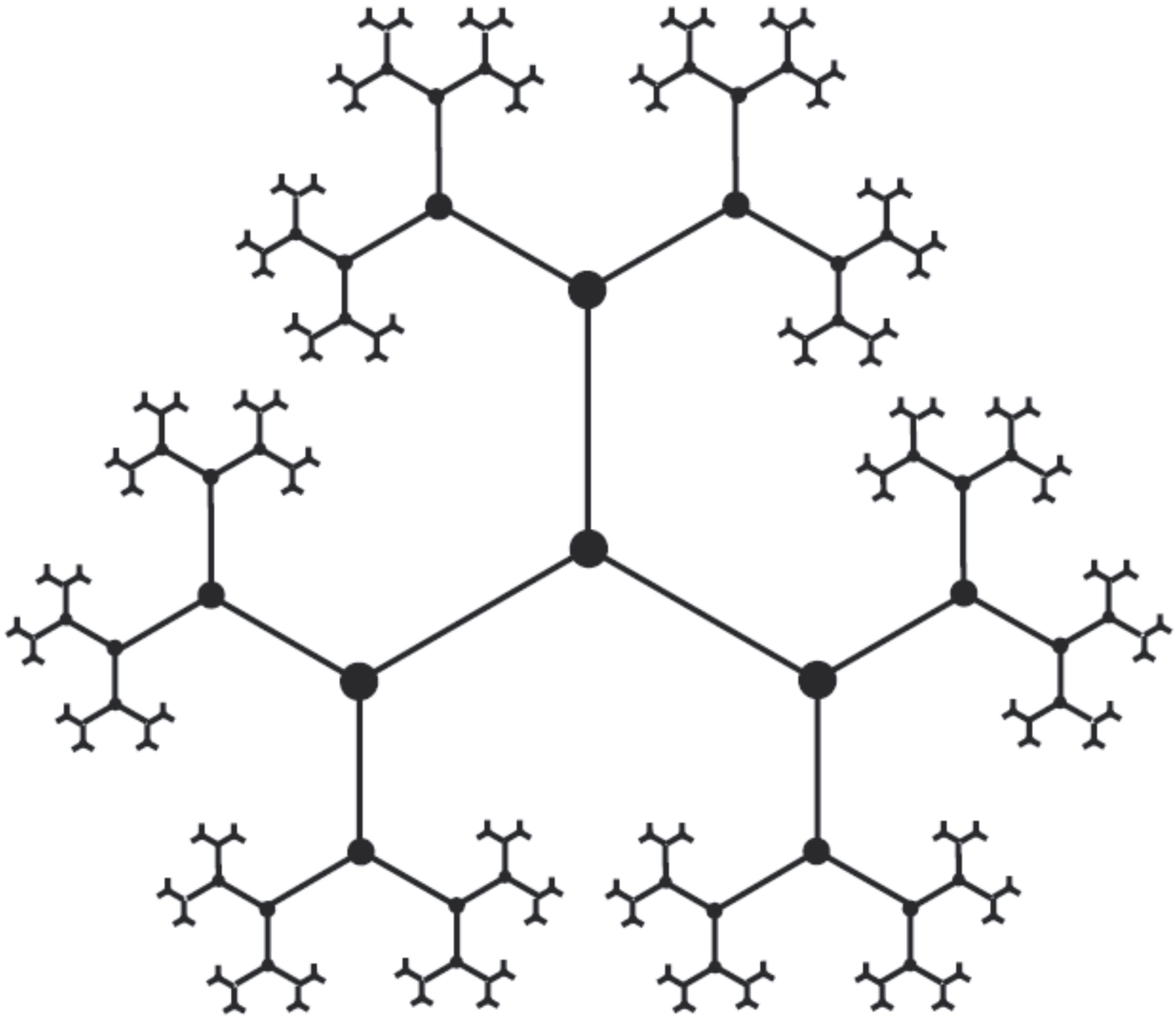
Vertices: the p-adic discs.
Edges: strict inclusions $B_1 \subseteq B_2$ not allowing intermediate discs.

Define a graph $\mathscr{T}_{\mathbb{Q}_p}$.

Vertices: the p-adic discs.
Edges: strict inclusions $B_1 \subseteq B_2$ not allowing intermediate discs.

$\Rightarrow$ $\mathscr{T}_{\mathbb{Q}_p}$ is a $p+1$-regular tree,

the *Bruhat-Tits tree* for $\mathbb{Q}_p$.



The Bruhat-Tits tree for $\mathbb{Q}_2$

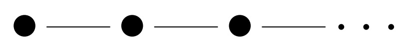
Strictly descending chain of discs

$$B_1 \supseteq B_2 \supseteq \ldots$$

converges to

$$\bigcap_n B_n = x \in \mathbb{Q}_p,$$

and corresponds to a halfline in $\mathscr{T}_{\mathbb{Q}_p}$



Strictly descending chain of discs

$$B_1 \supsetneq B_2 \supsetneq \dots$$

converges to

$$\bigcap_n B_n = x \in \mathbb{Q}_p,$$

and corresponds to a halfline in $\mathscr{T}_{\mathbb{Q}_p}$

$$\bullet \text{---} \bullet \text{---} \bullet \text{---} \cdots$$

An *end* of $\mathscr{T}_{\mathbb{Q}_p}$ is given by halflines with all but finitely many vertices in common

Strictly descending chain of discs

$$B_1 \supsetneq B_2 \supsetneq \ldots$$

converges to

$$\bigcap_n B_n = x \in \mathbb{Q}_p,$$

and corresponds to a halfline in $\mathscr{T}_{\mathbb{Q}_p}$

$$\bullet \text{---} \bullet \text{---} \bullet \text{---} \cdots$$

An *end* of $\mathscr{T}_{\mathbb{Q}_p}$ is given by halflines with all but finitely many vertices in common

One extra end:

$$B_1 \subsetneq B_2 \subsetneq \ldots$$

corresponds to the point ∞ on the p-adic projective line

$$\mathbb{P}^1(\mathbb{Q}_p) = \mathbb{Q}_p \cup \{\infty\}.$$

Result. Ends of $\mathscr{T}_{\mathbb{Q}_p} \cong \mathbb{P}^1(\mathbb{Q}_p)$.

2.3. Berkovich Topology

Problem. $(\mathbb{Q}_p, |\cdot|_p)$ is totally disconnected.

2.3. Berkovich Topology

Problem. $(\mathbb{Q}_p, |\cdot|_p)$ is totally disconnected.

Popular way out. Define new points:

2.3. Berkovich Topology

Problem. $(\mathbb{Q}_p, |\cdot|_p)$ is totally disconnected.

Popular way out. Define new points:

corresponding to sequences of p-adic discs

$$B_1 \supseteq B_2 \supseteq \dots$$

of arbitrary real radii s.t. $B = \bigcap B_n$ is one of:

1. $x \in \mathbb{Q}_p$,
2. a closed p-adic disc with radius $r \in |\mathbb{Q}_p|_p$,
3. a closed p-adic disc with radius $r \notin |\mathbb{Q}_p|_p$,
4. empty.

2.3. Berkovich Topology

Problem. $(\mathbb{Q}_p, |\cdot|_p)$ is totally disconnected.

Popular way out. Define new points:

corresponding to sequences of p-adic discs

$$B_1 \supseteq B_2 \supseteq \dots$$

of arbitrary real radii s.t. $B = \bigcap B_n$ is one of:

1. $x \in \mathbb{Q}_p$,
2. a closed p-adic disc with radius $r \in |\mathbb{Q}_p|_p$,
3. a closed p-adic disc with radius $r \notin |\mathbb{Q}_p|_p$,
4. empty.

Type 1: *classical points*, types 2–4: *generic points*.

Theorem (Berkovich)

1. The new $\mathbb{P}^1$ is compact, hausdorff, arc-wise connected.

2. $\mathscr{T}_{\mathbb{Q}_p} \subseteq \mathbb{P}^1$ is a retraction of $\mathbb{P}^1 \setminus \mathbb{P}^1(\mathbb{Q}_p)$.

3. Ends of $\mathscr{T}_{\mathbb{Q}_p} = \{$classical points of $\mathbb{P}^1\}$.

3. p-adic dendrograms

Take a finite set $X \subseteq \mathbb{P}^1(\mathbb{Q}_p)$.
$\rightsquigarrow X$ corresponds to a choice of ends in $\mathscr{T}_{\mathbb{Q}_p}$.

Definition. The smallest subtree $\mathscr{D}(X)$ of $\mathscr{T}_{\mathbb{Q}_p}$ whose ends are X is called the *p-adic dendrogram* for X.

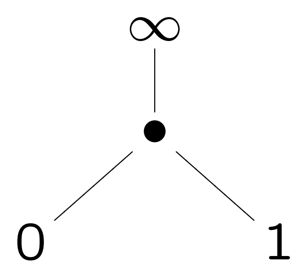
3. p-adic dendrograms

Take a finite set $X \subseteq \mathbb{P}^1(\mathbb{Q}_p)$.
$\rightsquigarrow X$ corresponds to a choice of ends in $\mathscr{T}_{\mathbb{Q}_p}$.

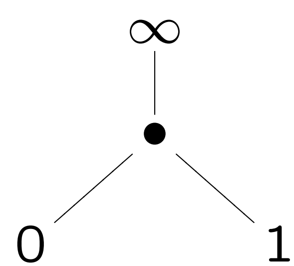
Definition. The smallest subtree $\mathscr{D}(X)$ of $\mathscr{T}_{\mathbb{Q}_p}$ whose ends are X is called the *p-adic dendrogram* for X.

Idea came from study of p-adic symmetries (G. Cornelissen, F. Kato, 2000)

Examples. 1. $X = \{0, 1, \infty\} \rightsquigarrow \mathscr{D}(X)$ consists of one vertex $v(0, 1, \infty)$ and three ends:



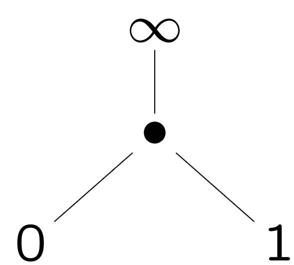
Examples. 1. $X = \{0, 1, \infty\} \rightsquigarrow \mathscr{D}(X)$ consists of one vertex $v(0, 1, \infty)$ and three ends:



2. $\{0, 1, \infty\} \subseteq X \subseteq \mathbb{N} \cup \{\infty\}$

$\rightsquigarrow \mathscr{D}(X)$ is a rooted tree with root $v(0, 1, \infty)$.

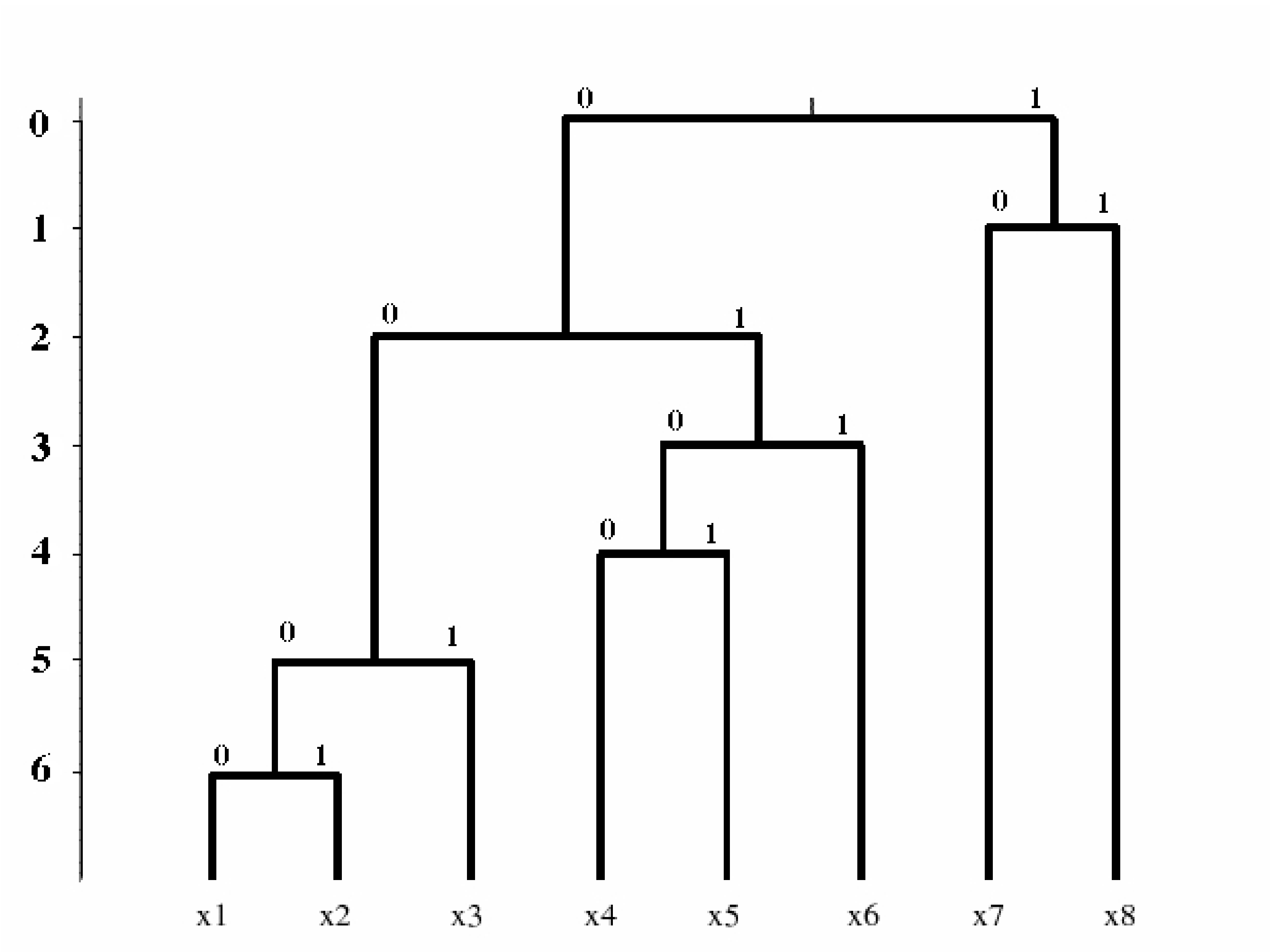
Examples. 1. $X = \{0, 1, \infty\} \rightsquigarrow \mathscr{D}(X)$ consists of one vertex $v(0, 1, \infty)$ and three ends:

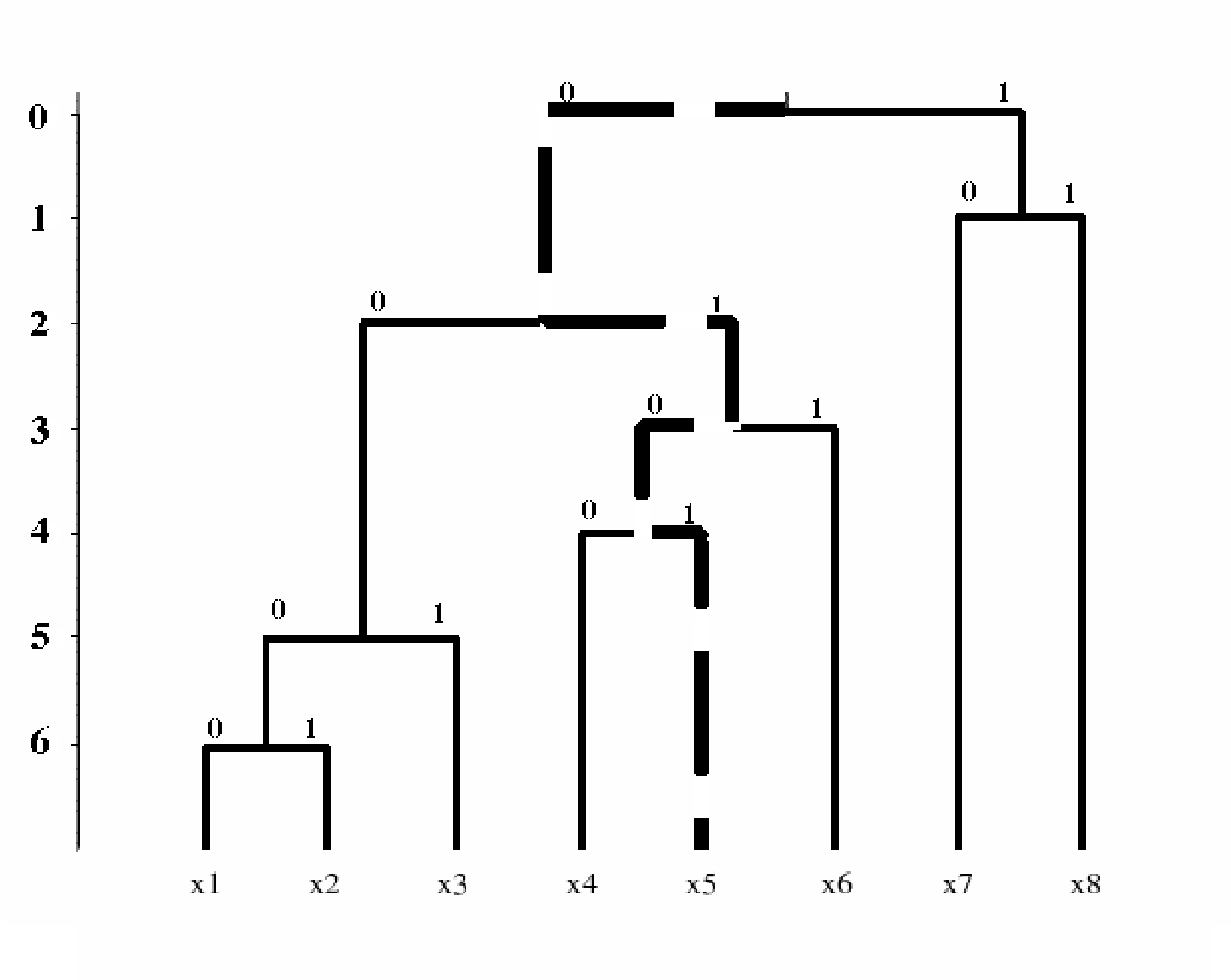


2. $\{0, 1, \infty\} \subseteq X \subseteq \mathbb{N} \cup \{\infty\}$

$\rightsquigarrow \mathscr{D}(X)$ is a rooted tree with root $v(0, 1, \infty)$.

3. F. Murtagh's p-adic dendrogram:


0
1
2
3
4
5
6
0
1
0
1
0
1
0
1
0
1
0
1
0
1
x1
x2
x3
x4
x5
x6
x7
x8



$$x_5 = 0 \cdot 2^0 + 1 \cdot 2^2 + 0 \cdot 2^3 + 1 \cdot 2^4.$$

Remark. All binary dendrograms are 2-adic dendrograms.

Arbitrary dendrograms. Either take larger prime p.

or: use a little algebra!

Extension fields of $\mathbb{Q}_p$.

Facts. 1. $\mathbb{Q}_p$ has arbitrarily large finite extension fields $K \supseteq \mathbb{Q}_p$.

2. $|\cdot|_p$ extends uniquely to a norm $|\cdot|_K$ on $K \rightsquigarrow (K, |\cdot|_K)$ is a complete field, called p-*adic number field*.

Extension fields of $\mathbb{Q}_p$.

Facts. 1. $\mathbb{Q}_p$ has arbitrarily large finite extension fields $K \supseteq \mathbb{Q}_p$.

2. $|\cdot|_p$ extends uniquely to a norm $|\cdot|_K$ on $K \rightsquigarrow (K, |\cdot|_K)$ is a complete field, called p-*adic number field*.

3. The *integers* of K are a ring $\mathcal{O}_K = \{x \in K \mid |x|_K \leq 1\}$.

4. There is a *uniformiser* $\pi \in \mathcal{O}_K$ s.t. $\mathcal{O}_K/\pi\mathcal{O}_K$ is a finite field with $q = p^f$ elements.

Extension fields of $\mathbb{Q}_p$.

Facts. 1. $\mathbb{Q}_p$ has arbitrarily large finite extension fields $K \supseteq \mathbb{Q}_p$.

2. $|\cdot|_p$ extends uniquely to a norm $|\cdot|_K$ on $K \rightsquigarrow (K, |\cdot|_K)$ is a complete field, called *p-adic number field*.

3. The *integers* of K are a ring $\mathcal{O}_K = \{x \in K \mid |x|_K \leq 1\}$.

4. There is a *uniformiser* $\pi \in \mathcal{O}_K$ s.t. $\mathcal{O}_K/\pi\mathcal{O}_K$ is a finite field with $q = p^f$ elements.

5. The Bruhat-Tits tree $\mathscr{T}_K$ is a $q+1$-regular tree.

Consequence. $p = 2$ suffices.

Let in $\mathscr{D}$ the maximal number of children vertices be $n \geq 2$,
$\rightsquigarrow$ take K large enough such that $2^f \geq n$.

By number theory, such K exist.

(Take K *unramified*, i.e. $\dim_{\mathbb{Q}_2} K = f$.)

Consequence. $p = 2$ suffices.

Let in $\mathscr{D}$ the maximal number of children vertices be $n \geq 2$,
$\rightsquigarrow$ take K large enough such that $2^f \geq n$.

By number theory, such K exist.

(Take K *unramified*, i.e. $\dim_{\mathbb{Q}_2} K = f$.)

$\rightsquigarrow$ From now on, "pretend" that $K = \mathbb{Q}_p$.

4. The space of dendrograms

Take $S = \{x_1, \ldots, x_n\} \subseteq \mathbb{P}^1(\mathbb{Q}_p)$ s.t. $x_1 = 0$, $x_2 = 1$, $x_3 = \infty$.

$\sigma(X) = \mathscr{D}(S)$ is the *skeleton* of $X = \mathbb{P}^1 \setminus S$.

Call X an n*-pointed projective line* and S the *markings*.

4. The space of dendrograms

Take $S = \{x_1, \ldots, x_n\} \subseteq \mathbb{P}^1(\mathbb{Q}_p)$ s.t. $x_1 = 0$, $x_2 = 1$, $x_3 = \infty$.

$\sigma(X) = \mathscr{D}(S)$ is the *skeleton* of $X = \mathbb{P}^1 \setminus S$.

Call X an n*-pointed projective line* and S the *markings*.

$\mathfrak{M}_n := \{X = \mathbb{P}^1 \setminus S \mid S = \{x_1 = 0, x_2 = 1, x_3 = \infty, \ldots, x_n\},\ \#S = n\}$

$\mathfrak{D}_n := \{\sigma(X) \mid X \in \mathfrak{M}_{n+1}\}$, the *space of dendrograms for n data*.

Facts. 1. $\dim \mathfrak{M}_n = n - 3$.

2. $\mathfrak{D}_{n-1} \subseteq \mathfrak{M}_n$ is a real polyhedral complex of dimension $n - 3$.

3. Maximal cells of $\mathfrak{D}_{n-1}$ consist of the binary dendrograms.

4. Moving inside cell $\leftrightarrow$ varying lengths of bounded edges.

5. Passing to neighbouring cells $\leftrightarrow$ contracting edges.

Facts. 1. $\dim \mathfrak{M}_n = n-3$.

2. $\mathfrak{D}_{n-1} \subseteq \mathfrak{M}_n$ is a real polyhedral complex of dimension $n-3$.

3. Maximal cells of $\mathfrak{D}_{n-1}$ and consist of the binary dendrograms.

4. Moving inside cell $\leftrightarrow$ varying lengths of bounded edges.

5. Passing to neighbouring cells $\leftrightarrow$ contracting edges.

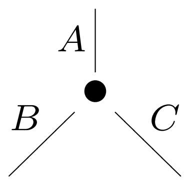
Remark. $n-3 = \#$ "free markings" which move within $\mathbb{P}^1$ without colliding.

Examples. 1. $\mathfrak{M}_3 = \mathfrak{D}_2 = \{pt\}$.

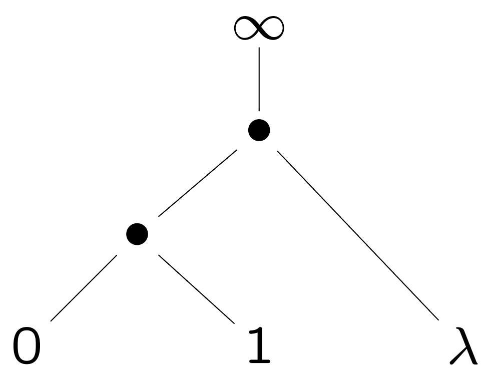
2. $\mathfrak{M}_4$ has one free marking $\lambda \in \mathbb{P}^1(\mathbb{Q}_p) \setminus \{0, 1, \infty\}$
$\rightsquigarrow \mathfrak{M}_4 = \mathbb{P}^1 \setminus \{0, 1, \infty\}$.

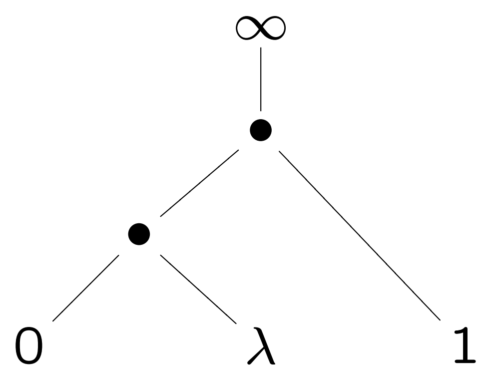
Examples. 1. $\mathfrak{M}_3 = \mathfrak{D}_2 = \{pt\}$.

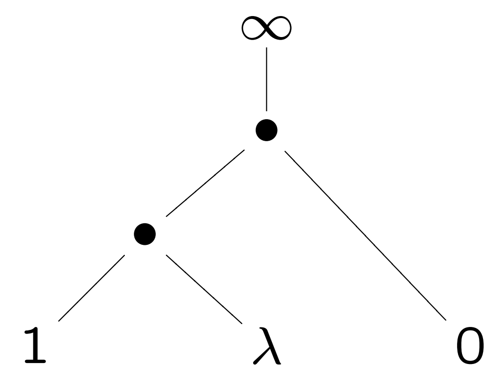
2. $\mathfrak{M}_4$ has one free marking $\lambda \in \mathbb{P}^1(\mathbb{Q}_p) \setminus \{0, 1, \infty\}$
$\rightsquigarrow \mathfrak{M}_4 = \mathbb{P}^1 \setminus \{0, 1, \infty\}$.

$\rightsquigarrow \mathfrak{D}_3 =$ 

, where

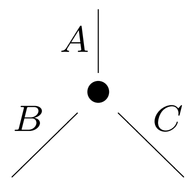
A: 

B: 

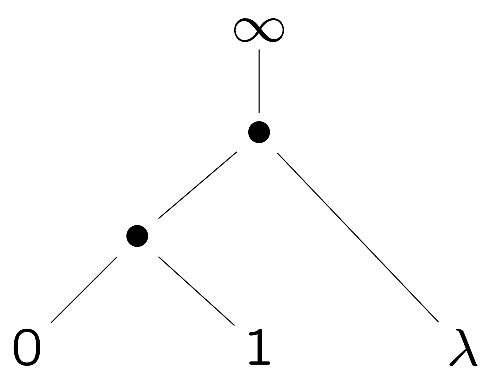
C: 

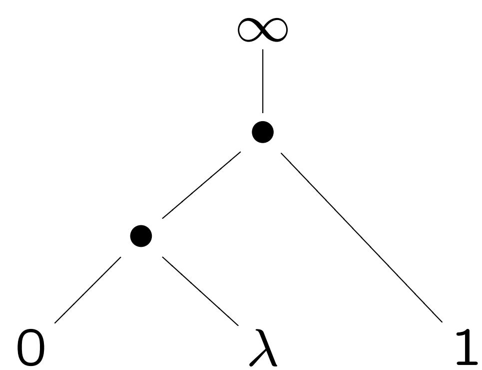
Examples. 1. $\mathfrak{M}_3 = \mathfrak{D}_2 = \{pt\}$.

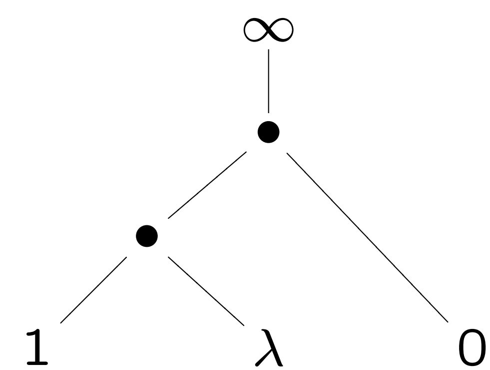
2. $\mathfrak{M}_4$ has one free marking $\lambda \in \mathbb{P}^1(\mathbb{Q}_p) \setminus \{0, 1, \infty\}$
$\rightsquigarrow \mathfrak{M}_4 = \mathbb{P}^1 \setminus \{0, 1, \infty\}$.

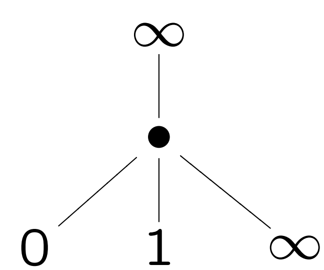
$\rightsquigarrow \mathfrak{D}_3 =$ 

, where

A: 

B: 

C: 

and the unique vertex in $\mathfrak{D}_3$ corresponds to 

Definition. A *family of dendrograms with n data over a space Y* is a map $Y \to \mathfrak{D}_n$ from some p-adic space Y.

Example. $Y = \{y_1, \ldots, y_T\}$. Interpret $t \in \{1, \ldots, T\}$ as time.

$\rightsquigarrow$ A family $Y \to \mathfrak{D}_n$ is a time series of n non-colliding particles.

Definition. A *family of dendrograms with* n *data over a space* Y is a map $Y \to \mathfrak{D}_n$ from some p-adic space Y.

Example. $Y = \{y_1, \ldots, y_T\}$. Interpret $t \in \{1, \ldots, T\}$ as time.

$\rightsquigarrow$ A family $Y \to \mathfrak{D}_n$ is a time series of n non-colliding particles.

Particles with collisions: $\rightsquigarrow$ compactify $\mathfrak{M}_n$.

But this is another story . . .

5. Distributions on dendrograms

Given dendrogram $\mathscr{D}(S)$ for data $S = \{x_1, \dots, x_n\}$.

Want to insert further datum $x \notin S$ into classification scheme.

5. Distributions on dendrograms

Given dendrogram $\mathscr{D}(S)$ for data $S = \{x_1, \ldots, x_n\}$.

Want to insert further datum $x \notin S$ into classification scheme.

Idea: *Classifier* $\rightsquigarrow$ Assign probabilities to vertices of $\mathscr{D}(S)$ depending on x.

5. Distributions on dendrograms

Given dendrogram $\mathscr{D}(S)$ for data $S = \{x_1, \ldots, x_n\}$.

Want to insert further datum $x \notin S$ into classification scheme.

Idea: *Classifier* $\rightsquigarrow$ Assign probabilities to vertices of $\mathscr{D}(S)$ depending on x.

Result: A family of dendrograms for $S \cup \{x\}$ with variable x
$+$ a probability distribution.

Definition. A *universal p-adic classifier $\mathcal{C}$ for n given points* is a probability distribution on $\mathfrak{M}_{n+1}$.

Borel σ-algebra on $\mathfrak{M}_{n+1}$ generated by the Berkovich open sets.

Definition. A *universal p-adic classifier $\mathcal{C}$ for n given points* is a probability distribution on $\mathfrak{M}_{n+1}$.

Borel σ-algebra on $\mathfrak{M}_{n+1}$ generated by the Berkovich open sets.

Let $X \in \mathfrak{M}_{n+1}$ with skeleton $\sigma(X) \in \mathfrak{D}_n$

$\rightsquigarrow$ $\mathcal{C}$ induces a measure on $X = \mathbb{P}^1 \setminus S$

$\rightsquigarrow$ probability distribution on $\sigma(X) = \mathscr{D}(S)$ (renormalisation)

Definition. A *universal p-adic classifier $\mathcal{C}$ for n given points* is a probability distribution on $\mathfrak{M}_{n+1}$.

Borel σ-algebra on $\mathfrak{M}_{n+1}$ generated by the Berkovich open sets.

Let $X \in \mathfrak{M}_{n+1}$ with skeleton $\sigma(X) \in \mathfrak{D}_n$

$\rightsquigarrow$ $\mathcal{C}$ induces a measure on $X = \mathbb{P}^1 \setminus S$

$\rightsquigarrow$ probability distribution on $\sigma(X)\mathscr{D}(S)$ (renormalisation)

The like for families of dendrograms.

6. Hidden vertices

Definition. A vertex v in a dendrogram $\mathscr{D}$ is *hidden*, if all edges emanating from v are bounded.

I.e. the cluster corresponding to v is non-trivially composed of non-singleton subclusters.

Theorem. Let $\mathscr{D} \in \mathfrak{D}_n$. Then

$$v^h \leq \frac{n+1}{4} - b_0^h + 1 \tag{4}$$

$$b_0^h \leq \frac{n-4}{3} \tag{5}$$

and the bound in (5) is sharp.

$v^h :=$ # of hidden vertices in $\mathscr{D}$

$\mathscr{D}^h :=$ the subforest of $\mathscr{D}$ spanned by all hidden vertices

$b_0^h :=$ # connected components of $\mathscr{D}^h$,
measures the internal structure of $\mathscr{D}$.

Proof. Inductive pasting of trees. □

7. Conclusions

- Geometric foundation towards p-adic data encoding.
- Encoding = embedding dendrogram into Bruhat-Tits tree.
- Embedding uniquely determined by p-adic data representation.

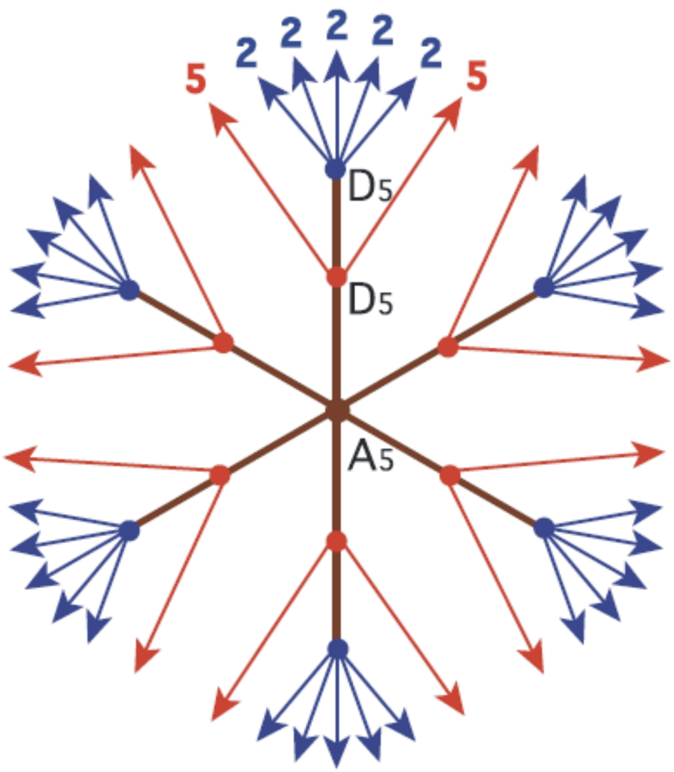
7. Conclusions

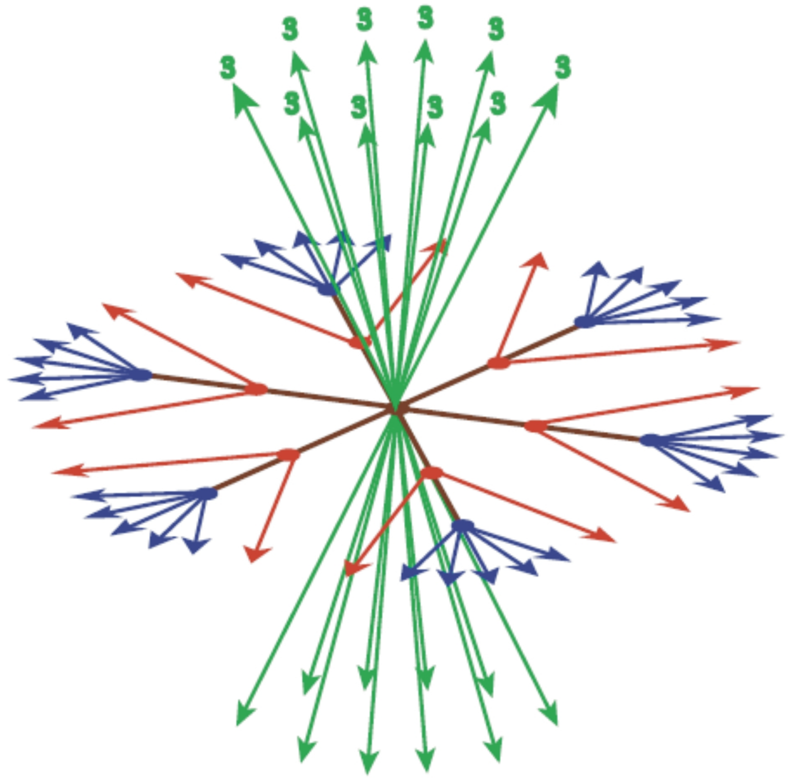
- Geometric foundation towards p-adic data encoding.
- Encoding = embedding dendrogram into Bruhat-Tits tree.
- Embedding uniquely determined by p-adic data representation.
- $\mathfrak{D}_n \subseteq \mathfrak{M}_{n+1}$
- Moving particles $\leftrightarrow$ family of dendrograms.
- Classifiers via measures on $\mathfrak{M}_n$.

7. Conclusions

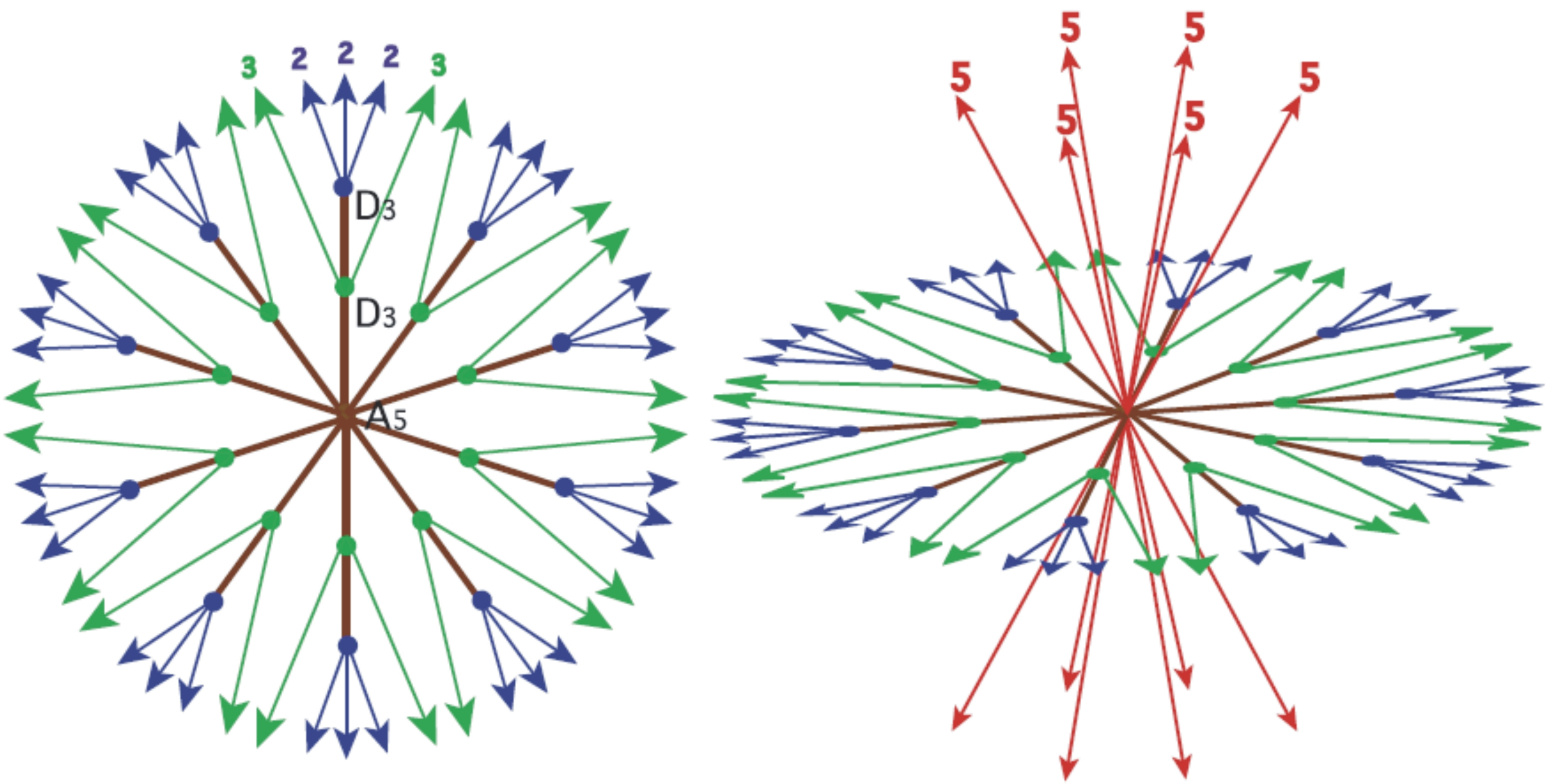
- Geometric foundation towards p-adic data encoding.
- Encoding = embedding dendrogram into Bruhat-Tits tree.
- Embedding uniquely determined by p-adic data representation.
- $\mathfrak{D}_n \subseteq \mathfrak{M}_{n+1}$
- Moving particles $\leftrightarrow$ family of dendrograms.
- Classifiers via measures on $\mathfrak{M}_n$.
- Bounds for # hidden vertices and components.

8. Epilogue: symmetric dendrograms

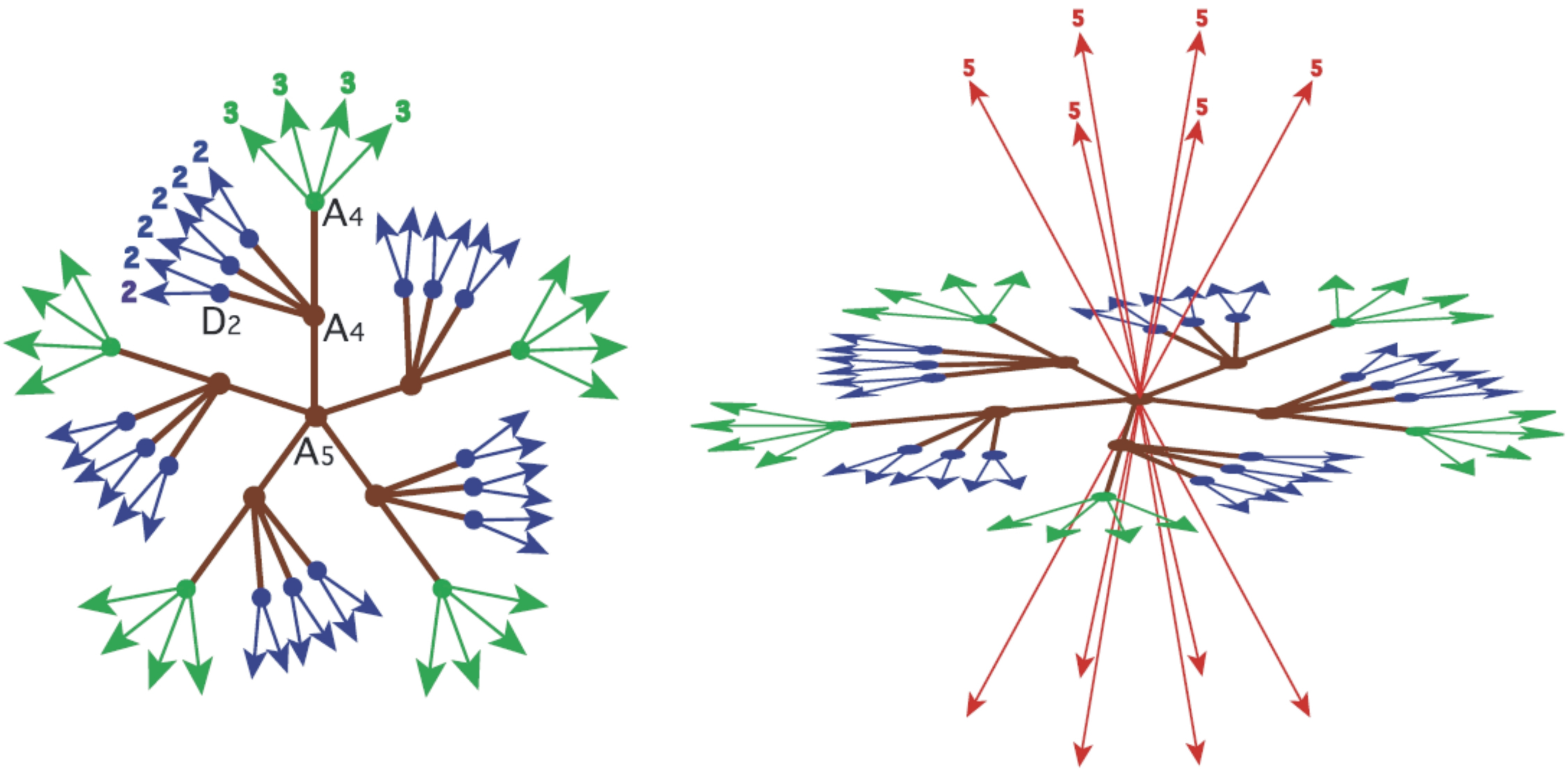




5-adic icosahedron (G. Cornelissen & F. Kato)



3-adic icosahedron (G. Cornelissen & F. Kato)



2-adic icosahedron (G. Cornelissen & F. Kato)