Towards a Geometry of Syntax

Matilde Marcolli

MoL 2019: Mathematics of Language
University of Toronto, 2019

A Mathematical Physicist's adventures in Linguistics

1. Alexander Port, Taelin Karidi, Matilde Marcolli, *Topological Analysis of Syntactic Structures*, arXiv:1903.05181
2. Andrew Ortegaray, Robert C. Berwick, Matilde Marcolli, *Heat Kernel analysis of Syntactic Structures*, arXiv:1803.09832
3. Kevin Shu, Andrew Ortegaray, Robert C. Berwick, Matilde Marcolli, *Phylogenetics of Indo-European Language families via an Algebro-Geometric Analysis of their Syntactic Structures*, arXiv:1712.01719
4. Kevin Shu, Matilde Marcolli, *Syntactic Structures and Code Parameters*, Math. Comput. Sci. 11 (2017) N.1, 79-90
5. Karthik Siva, Jim Tao, Matilde Marcolli, *Spin Glass Models of Syntax and Language Evolution*, Linguistic Analysis, Vol.41 (2017) N.3-4, 559-608.
6. Alexander Port, Iulia Gheorghita, Daniel Guth, John M.Clark, Crystal Liang, Shival Dasu, Matilde Marcolli, *Persistent Topology of Syntax*, Math. Comput. Sci. 12 (2018) N.1, 33-50

Syntax and Syntactic Parameters

• one of the key ideas of modern Generative Linguistics: Principles and Parameters (Chomsky, 1981)

- *parameters*: binary variables (on/off switches) that distinguish languages in terms of syntactic structures

• this idea is very appealing for a mathematician: at the level of syntax a language can be described by a set of coordinates given by binary variables

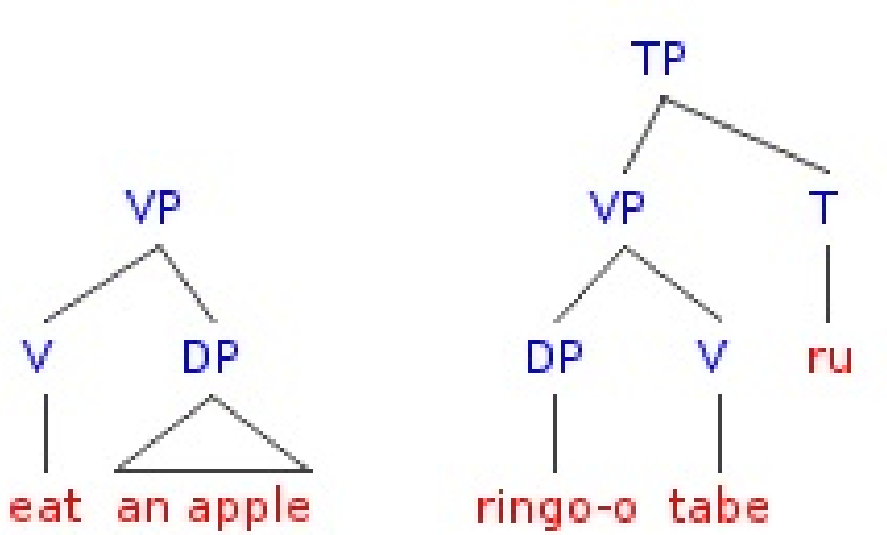
Geometric questions: can relations be modelled geometrically as a "manifold of syntax" inside a large binary space of arbitrary parameter configurations? What kind of manifold? Can it be cut out by algebraic equations (algebraic variety over $\mathbb{F}_2$)? etc.

Binary variables

- Example of parameter: head-directionality

(head-initial versus head-final)

English is head-initial, Japanese is head-final



VP= verb phrase, TP= tense phrase, DP= determiner phrase

- Other examples of parameters:
  - *Subject-side*
  - *Pro-drop*
  - *Null-subject*

Main Problems

- there is no complete classification of syntactic parameters
- Interdependencies between different syntactic parameters are poorly understood: what is a good independent set of variables, a good set of coordinates?
- syntactic parameters are dynamical: they change historically over the course of language change and evolution
- collecting reliable data is hard! (there are thousands of world languages and analyzing them at the level of syntax is much more difficult for linguists than collecting lexical data; few ancient languages have enough written texts)

Databases of syntactic structures of world languages

1. Syntactic Structures of World Languages (SSWL) `http://sswl.railsplayground.net/`
2. TerraLing `http://www.terraling.com/`
3. World Atlas of Language Structures (WALS) `http://wals.info/`
4. LanGeLin set of data from Longobardi–Guardiano, Lingua 119 (2009) 1679-1706; more recent update with more extensive database 2017.
5. new set of data announced by Longobardi's LanGeLin collaboration: should be available this year

• First Step: data analysis of syntax of world languages with various mathematical tools (persistent topology, phylogenetic algebraic geometry, coding theory, etc.)

Phylogenetic reconstruction in Linguistics

* Kevin Shu, Andrew Ortegaray, Robert Berwick, Matilde Marcolli, *Phylogenetics of Indo-European Language families via an Algebro-Geometric Analysis of their Syntactic Structures*, arXiv:1712.01719

- Can one reconstruct phylogenetic trees computationally using only information on the modern languages?
- Can one reconstruct phylogenetic trees using syntactic parameters data? (Syntax is more stable than lexicon, slower changes, rare borrowing...)
- Long standing open problems: for example the question of the early Indo-European tree
- Linguistics has studied in depth how languages change over time (Philology, Historical Linguistics) usually via lexical and morphological analysis
- Goal: understand the historical relatedness of different languages, subdivisions into families and sub-families, phylogenetic trees of language families

Phylogenetic Algebraic Geometry

Several available methods of computational phylogenetic reconstruction. The one that performs best is *Phylogenetic Algebraic Geometry*

- L. Pachter, B. Sturmfels, *Algebraic Statistics for Computational Biology*, Cambridge University Press, 2005
- L. Pachter, B. Sturmfels, *The Mathematics of Phylogenomics*, SIAM Review, Vol.49 (2007) N.1, 3–31
- E. Allman, J. Rhodes, *Phylogenetic ideals and varieties for general Markov models*, Adv. Appl. Math. Vol.40 (2008) 127–148
- M. Casanellas, J. Fernández–Sánchez, *Performance of a new invariants method on homogeneous and nonhomogeneous quartet trees*, Mol. Biol. Evol. 24 (2007) N.1, 288–293
- N. Eriksson, *Using invariants for phylogenetic tree construction*, in "Emerging applications of algebraic geometry", pp. 89–108, IMA Vol. Math. Appl., 149, Springer, 2009

General Idea of Phylogenetic Algebraic Geometry

- Markov process on a binary rooted tree (gen. Jukes-Cantor model)
- probability distribution at the root $(\pi, 1-\pi)$

(frequency of 0/1 for parameters at root vertex) and transition matrices along edges M^e bistochastic

$$M^e = \begin{pmatrix} 1-p_e & p_e \\ p_e & 1-p_e \end{pmatrix}$$

- observed distribution at the n leaves polynomial function

$$p_{i_1,\dots,i_n} = \Phi(\pi, M^e) = \sum_{w_v \in \{0,1\}} \pi_{w_{v_r}} \prod_e M^e_{w_{s(e)}, w_{t(e)}}$$

with sum over "histories" consistent with data at leaves

- polynomial map that assigns

$$\Phi : \mathbb{C}^{4n-5} \to \mathbb{C}^{2^n}, \quad \Phi(\pi, M^e) = p_{i_1,\dots,i_n}$$

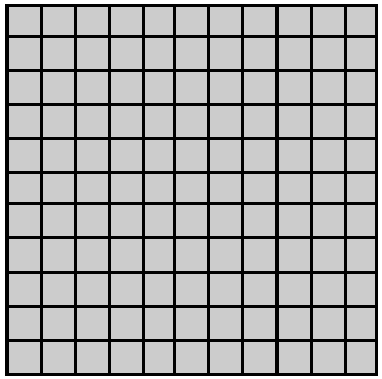
defines an *algebraic variety*

$$V_T = \overline{\Phi(\mathbb{C}^{4n-5})} \subset \mathbb{C}^{2^n}$$

What kinds of algebraic varieties occur in these models?

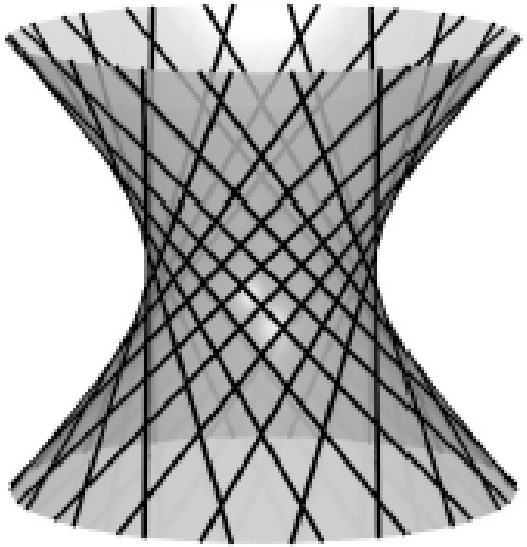
- Toric varieties (including Segre varieties and Veronese varieties)
- Determinantal varieties: the tree structure imposes rank constraints on matrices built starting from observed probabilities at the leaves
- Focus on following cases:
  1. Segre embeddings
  2. Secant varieties
- in our application to syntax we will encounter other varieties (defined as intersections of determinantal varieties in an ambient projective space) but we will use rough estimations of distances based on these simpler pieces

Segre embeddings



$\mathbb{P}^1 \times \mathbb{P}^1$

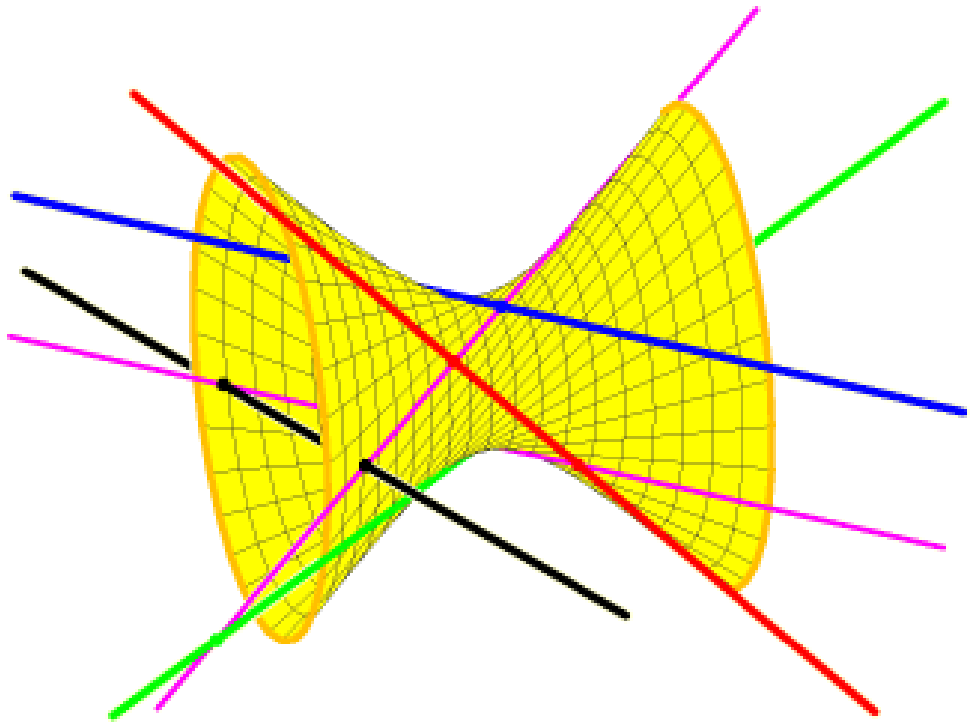
f



$X \subset \mathbb{P}^3$

$\mathbb{P}^1 \times \mathbb{P}^1 \hookrightarrow \mathbb{P}^3$ with $((x_0 : x_1), (y_0 : y_1)) \mapsto (x_0y_0 : x_0y_1 : x_1y_0 : x_1y_1)$

Secant varieties



variety of cords, closure (Zariski) of union of all secant lines of a variety V

Main Toolbox: Phylogenetic Invariants

• Allman–Rhodes theorem: ideal $\mathcal{I}_T$ defining V_T generated by all 3×3 minors of all *edge flattenings* of tensor $P = (p_{i_1,\ldots,i_n})$: $2^r \times 2^{n-r}$-matrix $Flat_{e,T}(P)$

$$Flat_{e,T}(P)(u, v) = P(u_1, \ldots, u_r, v_1, \ldots, v_{n-r})$$

where edge e removal separates boundary distribution into 2^r variable and 2^{n-r} variables

• phylogenetic invariants $\phi_T(P)$: 3×3 minors evaluated at boundary distribution $P = (p_{i_1,\ldots,i_n})$ given by data

- candidate trees T test by phylogenetic invariants
  - if T is the correct tree the phylogenetic invariants ϕ_T vanish when evaluated on the observed boundary distribution P (obtained from the data)

$$\phi_T(P) = 0$$

  - usually some noise in the data, so compare trees by how closely satisfied is the vanishing condition
  - closeness in some norm: ℓ^∞-norm or ℓ^1-norm

$$\|\phi_T(P)\|_{\ell^\infty} = \max_{M \in 3\times 3\text{-minors of } Flat_{e,T}(P)} |\det(M)|$$

$$\|\phi_T(P)\|_{\ell^1} = \sum_{M \in 3\times 3\text{-minors of } Flat_{e,T}(P)} |\det(M)|$$

  - ℓ^∞-norm is a weaker invariant of the ℓ^1-norm: loses information about the ϕ_T

Main Toolbox: Euclidean Distance

- Euclidean distance of the point P from the variety V_T (in ambient affine space)
- Eckrat–Young formula: for a determinantal variety

$$\mathcal{D}_r(n,m) = \{n \times m \text{ matrices of rank } \leq r\}$$

$$\text{dist}(M, \mathcal{D}_r(n,m)) = (\sum_{i=r+1}^{n} \sigma_i^2)^{1/2}$$

with σ_i singular values of the $n \times m$ flattenings M

Estimates of distance

- Euclidean distance of the point P from certain intersections $V_k \cap W$
- in general $\mathrm{dist}(P, V_1) < \mathrm{dist}(P, V_2)$ does not imply $\mathrm{dist}(P, V_1 \cap W) < \mathrm{dist}(P, V_2 \cap W)$
- assume established that $P \in W$
- then conditional case: if know that $P \in W$, then minimizing $\mathrm{dist}(P, V_k)$ suffices
- in more general cases, can use separate distances $\mathrm{dist}(P, V)$ and $\mathrm{dist}(P, W)$ as estimates from below of $\mathrm{dist}(P, V \cap W) \geq \max\{\mathrm{dist}(P, V), \mathrm{dist}(P, W)\}$ if easier to compute: if large can be used to rule out candidate trees

Procedure

- set of languages $\mathcal{L} = \{\ell_1, \ldots, \ell_n\}$ (selected subfamily)
- set of syntactic parameters mapped for all: π_i, $i = 1, \ldots, N$
- gives vectors $\pi_i = (\pi_i(\ell_j)) \in \mathbb{F}_2^n$
- compute frequencies

$$P = \{p_{i_1,\ldots,i_n} = \frac{N_{i_1,\ldots,i_n}}{N}\}$$

with $N_{i_1,\ldots,i_n}$ = number of occurrences of binary string $(i_1, \ldots, i_n) \in \mathbb{F}_2^n$ among the $\{\pi_i\}_{i=1}^N$

Test procedure to check reliability of data compared to what known from historical linguistics

- Produce a set of candidate trees (eg PHYLIP)
- Given a *candidate tree* T, compute all 3×3 minors of each flattening matrix $Flat_{e,T}(P)$, for each edge
- evaluate ℓ^∞ and ℓ^1 norm of $\phi_T(P)$ over all 3×3 minors of flattening matrices
- obtain estimates of Euclidean distance of P to V_T (or part of V_T that distinguishes candidate trees)
- select best fit tree on the basis of these tests
- compare with what known from historical linguistics

First Example: Germanic Languages

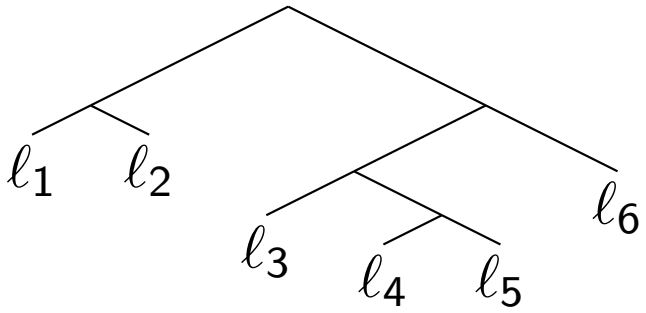
• small set of languages: ℓ_1 =Dutch, ℓ_2 =German, ℓ_3 =English, ℓ_4 =Faroese, ℓ_5 =Icelandic, ℓ_6 =Swedish

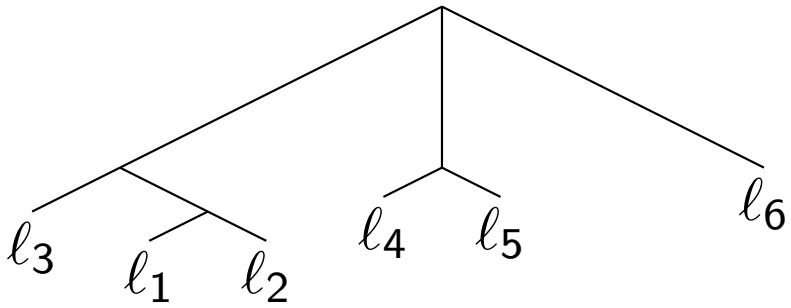
• candidate trees produced by PHYLIP on SSWL data

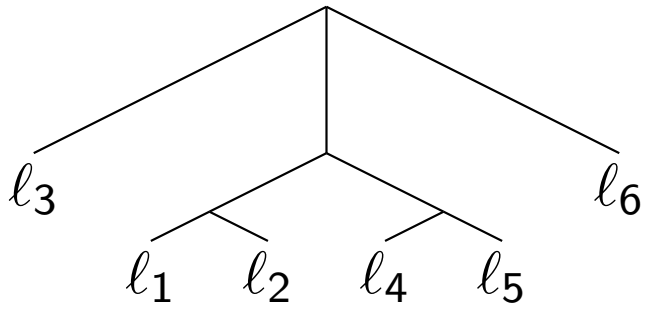
$$\begin{aligned} \texttt{pars1} &= ((\ell_1, \ell_2), (\ell_3, (\ell_4, \ell_5)), \ell_6) \\ \texttt{pars2} &= ((\ell_3, (\ell_1, \ell_2)), (\ell_4, \ell_5), \ell_6) \\ \texttt{pars3} &= (\ell_3, ((\ell_1, \ell_2), (\ell_4, \ell_5)), \ell_6) \end{aligned}$$

• compute flattenings for each of these trees (after resolving trivalent ambiguities into binary trees)

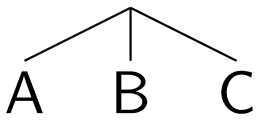
- `pars1`, `pars2`, and `pars3` trees

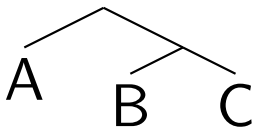


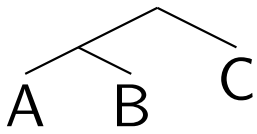


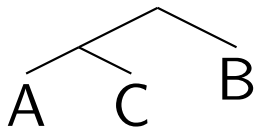


- resolve non-binary trees

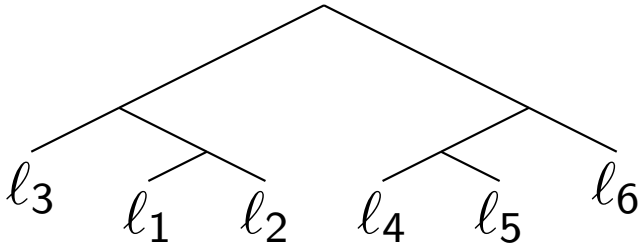


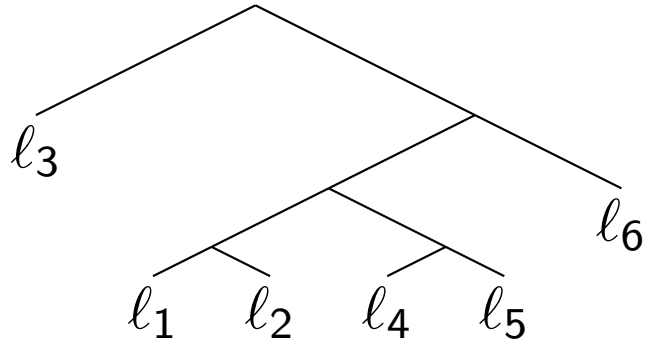






• up to shifts in the position of the root vertex binary trees for `pars2` and `pars3`





• position of the root vertex not determined by this algorithm, only tree topology: need to use additional information to locate it

• note that all these candidate trees agree on the proximity of ℓ_1 and ℓ_2 (Dutch and German) and of ℓ_4 and ℓ_5 (Faroese and Icelandic) ...conditional case

- Flattenings:
  - `pars1`:

$$\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P) \quad 4\times 16\text{matrix}$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P) \quad 8\times 8\text{matrix}$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3,\ell_6\}\cup\{\ell_4,\ell_5\}}(P) \quad 16\times 4\text{matrix}$$

  - `pars2`:

$$\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P) \quad 4\times 16\text{matrix}$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P) \quad 8\times 8\text{matrix}$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3,\ell_6\}\cup\{\ell_4,\ell_5\}}(P) \quad 16\times 4\text{matrix}$$

  - `pars3`:

$$\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P) \quad 4\times 16\text{matrix}$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3,\ell_6\}\cup\{\ell_4,\ell_5\}}(P) \quad 16\times 4\text{matrix}$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P) \quad 16\times 4\text{matrix}$$

- $\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P)$ and $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3,\ell_6\}\cup\{\ell_4,\ell_5\}}(P)$ contribute to all candidate trees, do not discriminate between them

- conditional problem: assuming that P lies on the varieties cut out by the phylogenetic invariants of $\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P)$ and $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3,\ell_6\}\cup\{\ell_4,\ell_5\}}(P)$ select the most likely candidate that makes the remaining condition satisfied

- left with simpler setting:
  - $F_1 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P)$ for `pars1`
  - $F_2 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P)$ for `pars2`
  - $F_3 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P)$ for `pars3`

- single flattening: phylogenetic ideal generated by its 3×3 minors

• the geometry involved consists of classical algebro-geometric spaces:

- `pars1`: secant variety $\mathrm{Sec}(\mathcal{S}(8,8))$ of Segre variety $\mathcal{S}(8,8) = \mathbb{P}^7 \times \mathbb{P}^7$ embedded in $\mathbb{P}^{63}$ via Segre embedding $u_{i_1,\ldots,i_6} = x_{i_1,i_2,i_6} y_{i_3,i_4,i_5}$
- `pars2`: $\mathrm{Sec}(\mathcal{S}(8,8))$, with $\mathcal{S}(8,8)$ embedded in $\mathbb{P}^{63}$ via $u_{i_1,\ldots,i_6} = x_{i_1,i_2,i_3} y_{i_4,i_5,i_6}$.
- `pars3`: secant variety $\mathrm{Sec}(\mathcal{S}(16,4))$ of Segre variety $\mathcal{S}(16,4) = \mathbb{P}^{15} \times \mathbb{P}^3$ embedded in $\mathbb{P}^{63}$ via Segre embedding $u_{i_1,\ldots,i_6} = x_{i_1,i_2,i_4,i_5} y_{i_3,i_6}$.

Boundary distribution

- 90 SSWL parameters are completely mapped for these languages
- for each binary string $(i_1, \ldots, i_6)$ count occurrences as values of some syntactic parameter on the languages $\ell_1, \ldots, \ell_6$
- frequency matrix:

$$
\begin{array}{lll}
n_{110111} = 3 & n_{000011} = 1 & n_{000010} = 4 \\
n_{000000} = 40 & n_{110000} = 2 & n_{001110} = 1 \\
n_{000100} = 2 & n_{111111} = 22 & n_{111110} = 1 \\
n_{000110} = 1 & n_{111101} = 3 & n_{100000} = 2 \\
n_{010000} = 1 & n_{111001} = 2 & n_{110110} = 1 \\
n_{010111} = 1 & n_{001000} = 2 & n_{000111} = 1
\end{array}
$$

$n_{i_1,\ldots,i_6} = 0$ otherwise; frequencies $p_{i_1,\ldots,i_6} = n_{i_1,\ldots,i_6}/90$

- from this compute the flattening matrices

Example:

flattening matrix $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P)$

$$\begin{pmatrix}
\frac{4}{9} & \frac{1}{45} & \frac{1}{45} & 0 & \frac{2}{45} & \frac{1}{90} & 0 & \frac{1}{90} \\
\frac{1}{90} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{45} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{45} & 0 & 0 & 0 & 0 & \frac{1}{90} & 0 & \frac{1}{90} \\
0 & 0 & 0 & 0 & \frac{1}{90} & \frac{1}{90} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{1}{90} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{45} & \frac{1}{30} & 0 & \frac{1}{30} & 0 & \frac{11}{45}
\end{pmatrix}$$

ℓ_1 =Dutch, ℓ_2 =German, ℓ_3 =English, ℓ_4 =Faroese, ℓ_5 =Icelandic, ℓ_6 =Swedish

Phylogenetic invariants favor the tree binary tree pars2:

- for the tree pars1

$$\|\phi_{T_1}(P)\|_{\ell^\infty} = \max_{\phi \in 3\times 3 \text{ minors of } F_1} |\phi(P)| = \frac{22}{18225}$$

$$\|\phi_{T_1}(P)\|_{\ell^1} = \sum_{\phi \in 3\times 3 \text{ minors of } F_1} |\phi(P)| = \frac{3707}{364500}$$

- for the tree pars2

$$\|\phi_{T_2}(P)\|_{\ell^\infty} = \max_{\phi \in 3\times 3 \text{ minors of } F_2} |\phi(P)| = \frac{419}{364500}$$

$$\|\phi_{T_2}(P)\|_{\ell^1} = \sum_{\phi \in 3\times 3 \text{ minors of } F_2} |\phi(P)| = \frac{2719}{364500}$$

- for the tree pars3

$$\|\phi_{T_3}(P)\|_{\ell^\infty} = \max_{\phi \in 3\times 3 \text{ minors of } F_3} |\phi(P)| = \frac{22}{18225}$$

$$\|\phi_{T_3}(P)\|_{\ell^1} = \sum_{\phi \in 3\times 3 \text{ minors of } F_3} |\phi(P)| = \frac{949}{91125}$$

Euclidean distance

• varieties defined by the 3×3-minors of the three flattening matrices:

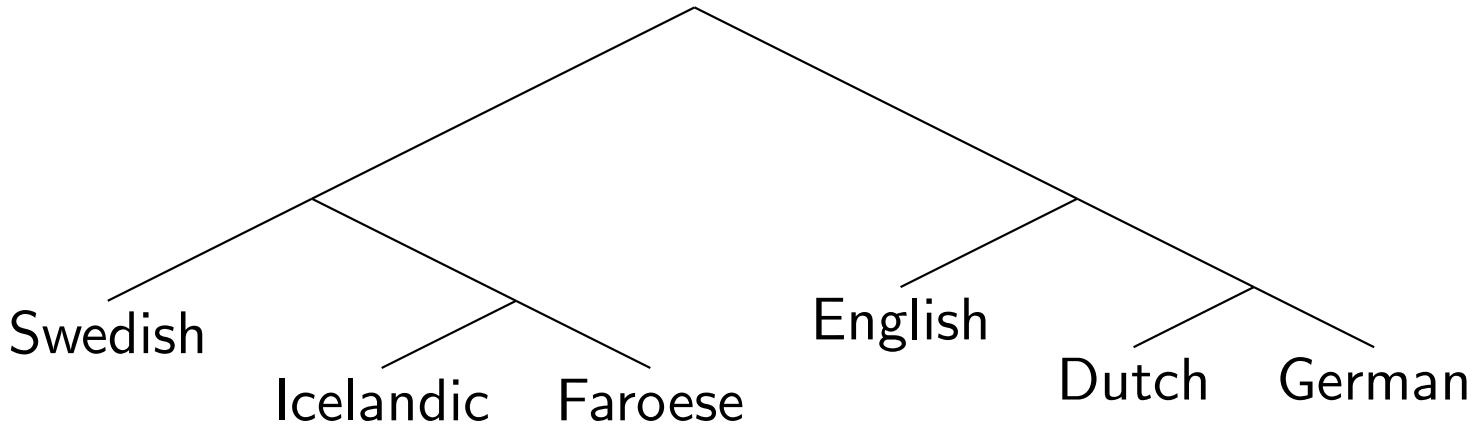
- $\mathcal{D}_2(8,8) = \mathrm{Sec}(\mathcal{S}(8,8))$: 28-dimensional determinantal variety of all 8×8 matrices of rank at most two
- $\mathcal{D}_2(16,4) = \mathrm{Sec}(\mathcal{S}(16,4))$: 36-dimensional determinantal variety of all 16×4 matrices of rank at most two

• phylogenetic algebraic variety of a candidate tree: intersection with remaining equations coming from the 3×3 minors of the other common flattenings (intersections of three different determinantal varieties inside a common ambient space $\mathbb{A}^{2^6}$)
• conditional case, assuming P on the common varieties, evaluate distance from the remaining one

- Euclidean distance of $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P)$ from $\mathcal{D}_2(8,8)$
- Euclidean distance of $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P)$ from $\mathcal{D}_2(8,8)$
- Euclidean distance of the point $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P)$ from $\mathcal{D}_2(16,4)$.

Result: Eckart-Young theorem Euclidean distance favors the tree `pars2`

- correctly identifies the West Germanic/North Germanic split



- other PHYLIP candidate trees misplaced it

Other examples: Romance languages, Slavic languages

- same method; comparative use of SSWL and Longobardi data
- placement of ancient languages and root vertex
- estimates of Euclidean distance

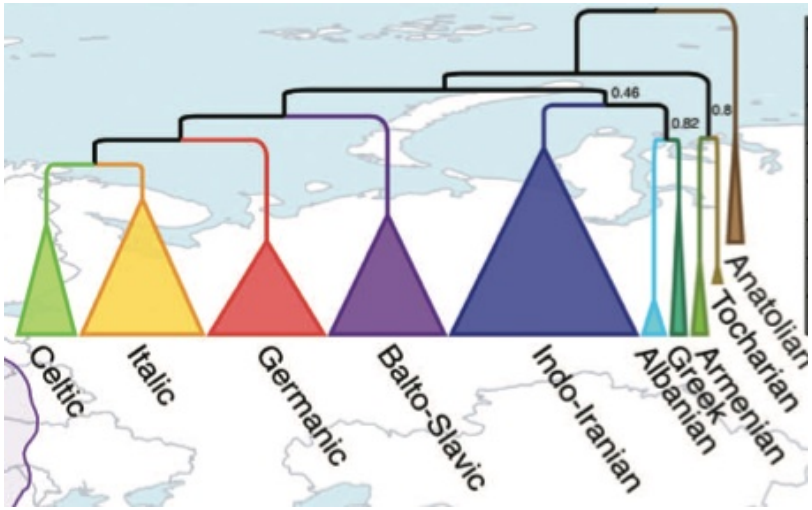
Early Indo-European tree: can one use this method to say something about the early branched of the Indo–European tree?

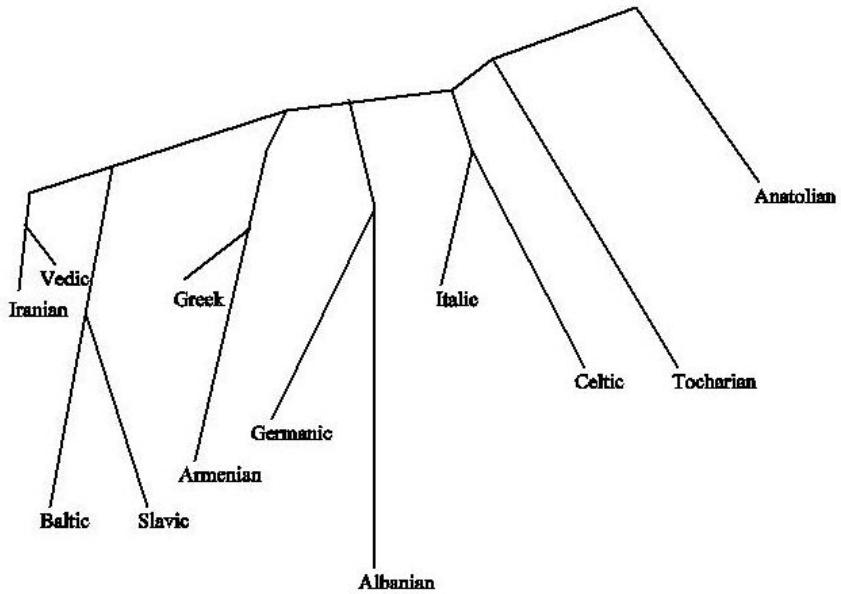
• Examples of questions about early branches of the tree of Indo-European languages:

- The relative positions of the Greco-Armenian subtrees;
- The position of Albanian in the tree;
- The relative positions of these languages with respect to the Anatolian-Tocharian subtrees.

• Try a comparison, based on SSWL data, between tree of Gray and Atkinson (Nature, 2003) and tree via morphological analysis (Ringe, Warnow, Taylor, 2002)

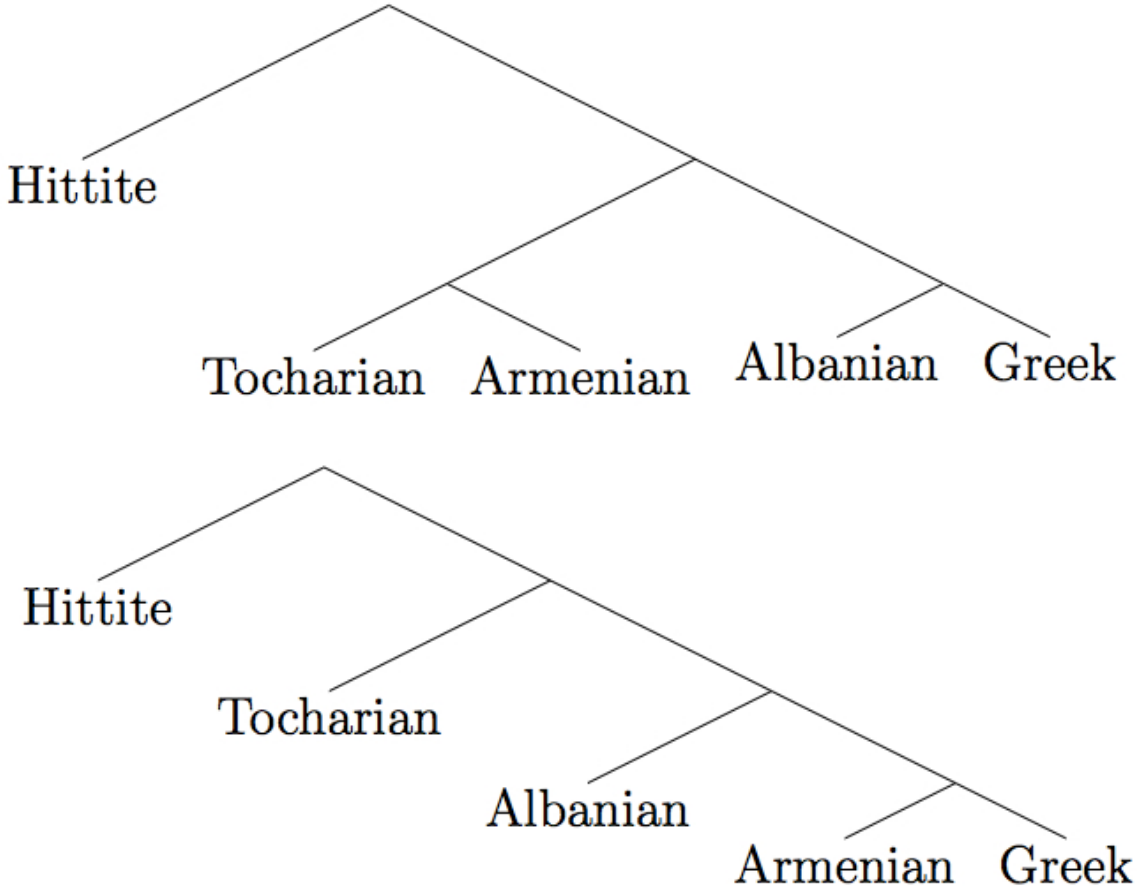
• A. Perelysvaig, M.W. Lewis, *The Indo-European controversy: facts and fallacies in Historical Linguistics*, Cambridge University Press, 2015.





The Atkinson–Gray early Indo-European tree and the Ringe–Warnow–Taylor tree

Focus on a smaller part of the tree: relative position of these languages



Can detect the difference from syntactic parameters? Using Phylogenetic Algebraic Geometry of Syntactic Parameters?

• Problem: SSWL data for Hittite, Tocharian, Albanian, Armenian, and Greek have a small number of parameters that is completely mapped for all these languages (and these parameters largely agree); Hittite and Tocharian not mapped in Longobardi's data.

• only 22 of the SSWL parameters are completely mapped for all of these languages

$$p_{00000} = 4/11, \quad p_{11111} = 3/11, \quad p_{11101} = 2/11,$$
$$p_{11011} = 1/22, \quad p_{10111} = 1/11, \quad p_{01000} = 1/22$$

with $p_{i_1,\ldots,i_5} = 0$ for all the remaining binary vectors in $\{0,1\}^5$.

First Case: flattening matrices

$$\begin{pmatrix} \frac{4}{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{11} \\ 0 & 0 & 0 & \frac{1}{22} & 0 & \frac{2}{11} & 0 & \frac{3}{11} \end{pmatrix}$$

$$\begin{pmatrix} \frac{4}{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{11} \\ 0 & 0 & 0 & \frac{1}{22} \\ 0 & \frac{2}{11} & 0 & \frac{3}{11} \end{pmatrix}$$

Second Case: flattening matrices

$$\begin{pmatrix} \frac{4}{11} & 0 & \frac{1}{22} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{22} \\ 0 & 0 & 0 & \frac{2}{11} & 0 & \frac{1}{11} & 0 & \frac{3}{11} \end{pmatrix}$$

$$\begin{pmatrix} \frac{4}{11} & 0 & \frac{1}{22} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{22} \\ 0 & 0 & 0 & \frac{2}{11} \\ 0 & \frac{1}{11} & 0 & \frac{3}{11} \end{pmatrix}$$

Phylogenetic Invariants

1. For the Gray-Atkins tree T_1:

$$\|\Phi_{T_1}(P)\|_{\ell^\infty} = \max_{\substack{\phi \in 3\times 3 \text{minors} \\ \text{of flattenings of } T_1}} |\phi(P)| = \frac{8}{1331}$$

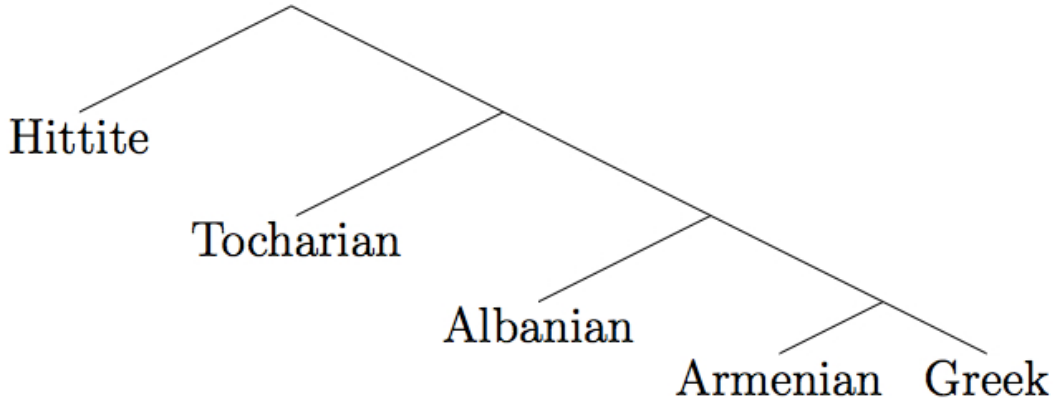
$$\|\Phi_{T_1}(P)\|_{\ell^1} = \sum_{\substack{\phi \in 3\times 3 \text{minors} \\ \text{of flattenings of } T_1}} |\phi(P)| = \frac{61}{2662}$$

2. For the Ringe–Warnow–Taylor tree T_2:

$$\|\Phi_{T_1}(P)\|_{\ell^\infty} = \max_{\substack{\phi \in 3\times 3 \text{minors} \\ \text{of flattenings of } T_1}} |\phi(P)| = \frac{8}{1331}$$

$$\|\Phi_{T_1}(P)\|_{\ell^1} = \sum_{\substack{\phi \in 3\times 3 \text{minors} \\ \text{of flattenings of } T_1}} |\phi(P)| = \frac{18}{1331}$$

• the ℓ^∞ norm does not distinguish the two trees while the ℓ^1 norm prefers the Ringe–Warnow–Taylor tree T_2
• the SSWL data favor the Ringe–Warnow–Taylor tree over the Atkinson–Gray tree, *but the SSWL data is problematic!* ...need better syntactic data on these languages (especially Hittite and Tocharian that are very poorly mapped in databases)



More ongoing work on phylogenetic trees:

* Sitanshu Gakkhar, Matilde Marcolli, *Metric Distortion, Graph Expansion, and Syntactic Structures*, in preparation, 2019

Question: can one get new historical linguistic information about language families, not in the form of phylogenetic trees?

Topological Analysis of Syntactic Structures

* Alexander Port, Taelin Karidi, Matilde Marcolli, *Topological Analysis of Syntactic Structures*, arXiv:1903.05181
* Alexander Port, Iulia Gheorghita, Daniel Guth, John M.Clark, Crystal Liang, Shival Dasu, Matilde Marcolli, *Persistent Topology of Syntax*, Math. Comput. Sci. 12 (2018) N.1, 33-50

Some Questions:

- Is topology different for different language families?
- Persistent connected components H_0: can construct a tree, how does it correlate to phylogenetic trees?
- Persistent first homology H_1: loops can detect historical cross-family influences at the syntactic level? when homeoplasy phenomena?
- Are there non-trivial higher homology groups like H_2? What meaning do they have?

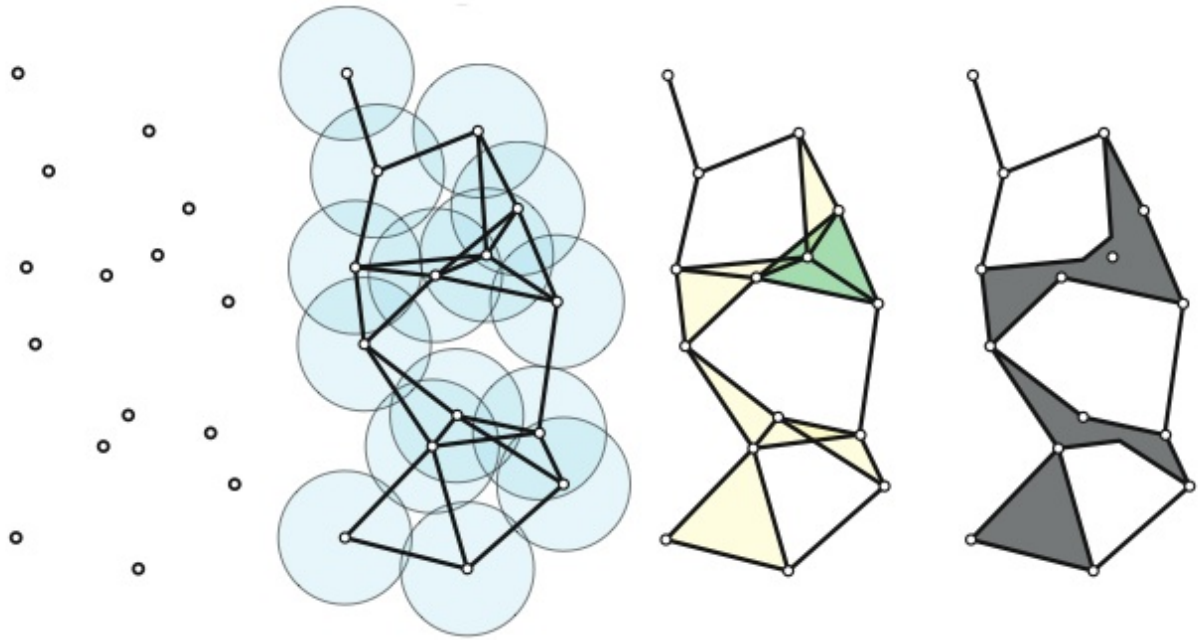
Some general references on Persistent Topology

- G. Carlsson, *Topology and data*, Bull. Amer. Math. Soc. (N.S.) 46 (2009) N.2, 255-308.
- H. Edelsbrunner, J. Harer, *Computational topology*, American Mathematical Society, 2010.
- A.J. Zomorodian, *Topology for computing*, Cambridge University Press, 2005
- R. Ghrist, *Elementary Applied Topology*, CreateSpace, 2014.
- J. Boissonnat, F. Chazal, M. Yvinec, *Geometric and topological inference*, Cambridge University Press, 2018.

Persistent Homology

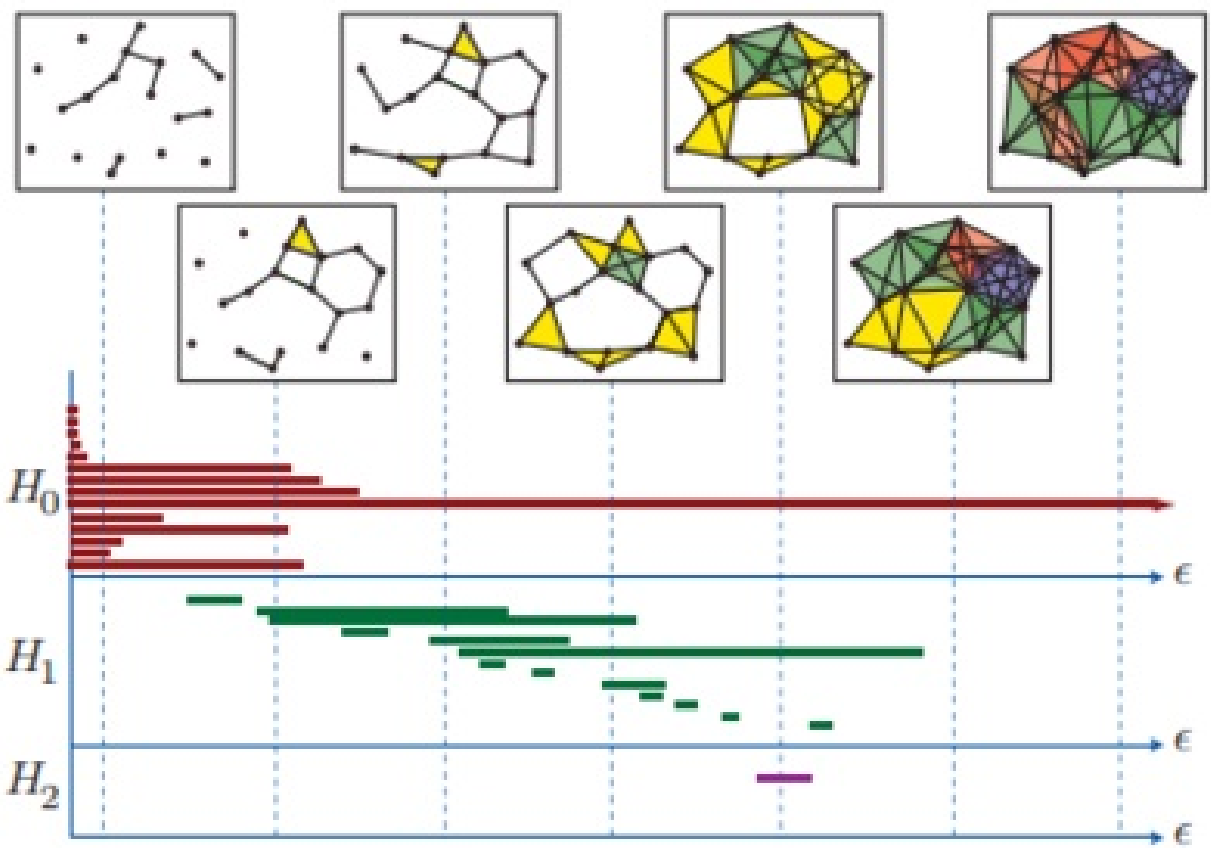
Vietoris–Rips complexes

- set $X = \{x_\alpha\}$ of points in Euclidean space $\mathbb{E}^N$, distance $d(x, y) = \|x - y\| = (\sum_{j=1}^{N}(x_j - y_j)^2)^{1/2}$
- Vietoris-Rips complex $R(X, \epsilon)$ of scale ϵ over field $\mathbb{K}$:
- $R_n(X, \epsilon)$ is $\mathbb{K}$-vector space spanned by all unordered $(n+1)$-tuples of points $\{x_{\alpha_0}, x_{\alpha_1}, \ldots, x_{\alpha_n}\}$ in X where all pairs have distances $d(x_{\alpha_i}, x_{\alpha_j}) \leq \epsilon$



Barcode Diagrams

- inclusion maps $R(X, \epsilon_1) \hookrightarrow R(X, \epsilon_2)$ for $\epsilon_1 < \epsilon_2$ induce maps in homology by functoriality $H_n(X, \epsilon_1) \to H_n(X, \epsilon_2)$
- barcode diagrams: births and deaths of persistent generators



Persistent Components Tree

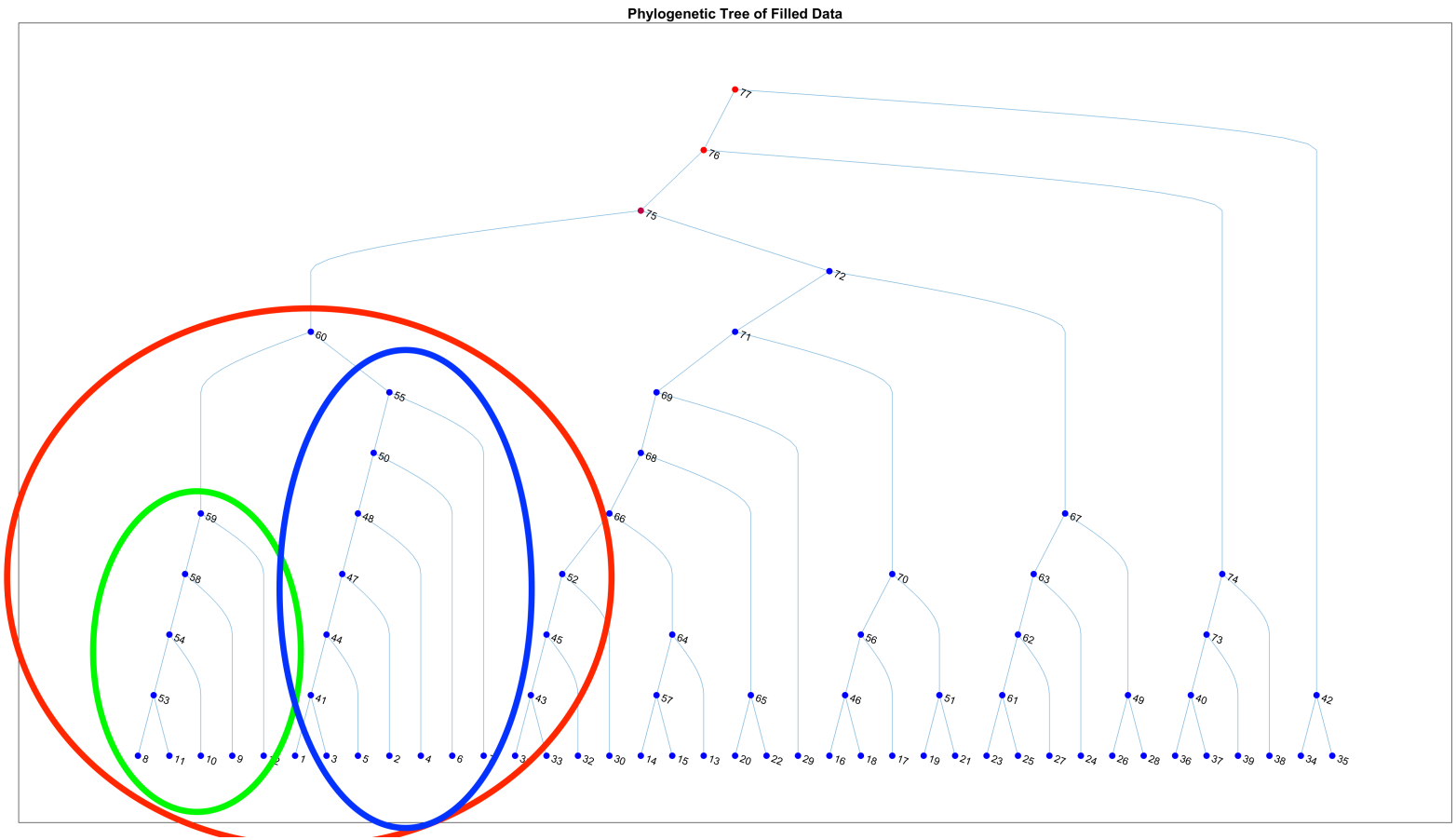
- for very small ϵ each point a singleton component
- for very large ϵ all points have joined into the same persistent connected component
- in between the components join in a certain order as function of ϵ that reflects the barcode diagram
- construct a tree that follows the merging of connected components as ϵ grows

Observations

- generally persistent component clustering has closer correlation to phylogenetic trees of historical relatedness for LanGeLin than for SSWL data
- closer look at some subfamilies reveal misplacements
- misplacements within smaller subfamilies also affected by changing the PCA variance level (PCA needed because persistent homology is computationally heavy)
- not a problem of the data: phylogenetic algebraic geometry method (applied to same data) gives correct historical tree

Some examples

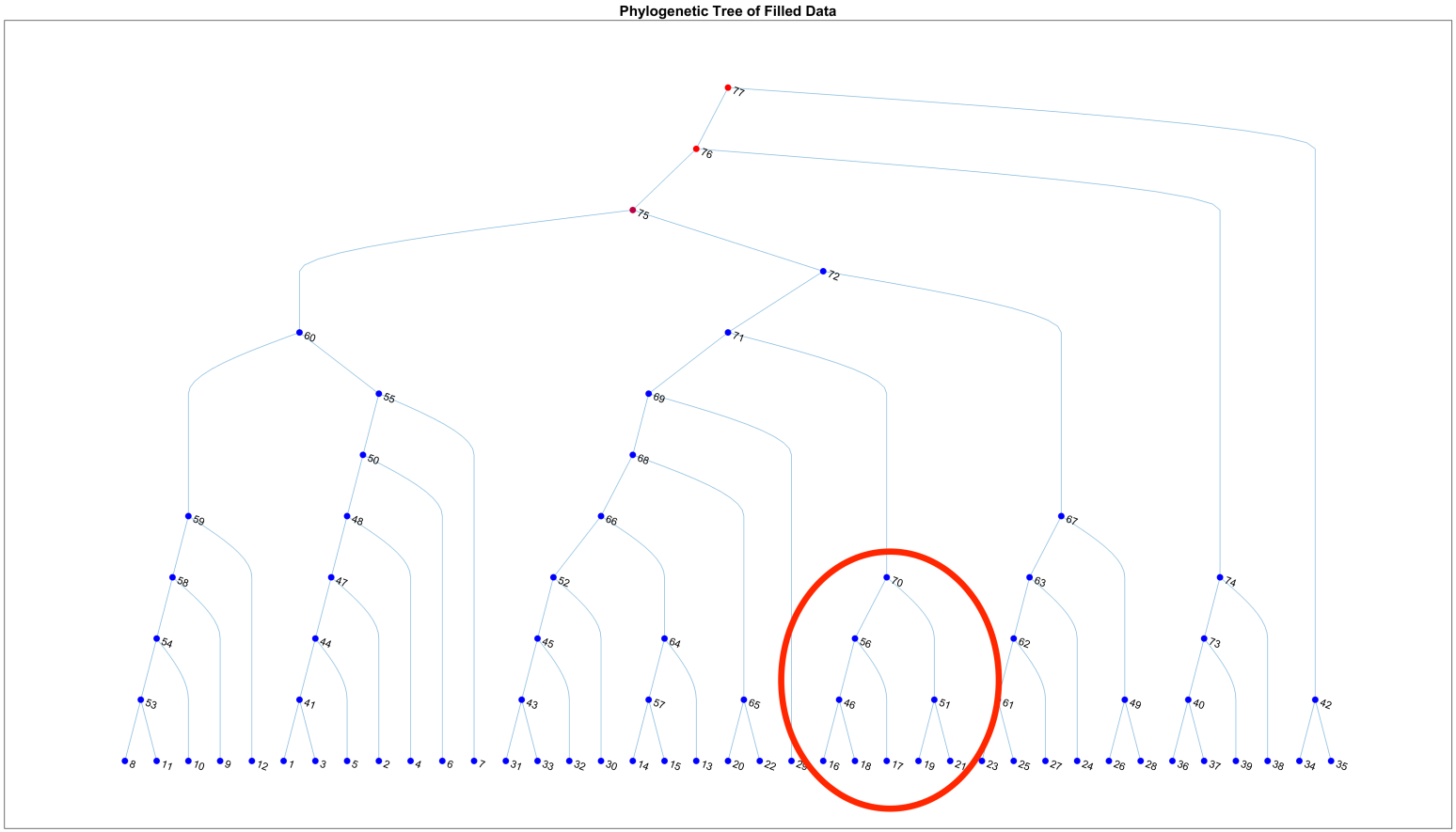
LanGeLin data:



subcluster: modern Romance languages: Italian, Spanish, French, Portuguese, Romanian; subcluster: Romance Southern Italian dialects: Ragusa, Mussomeli, Aidone, Southern Calabrese, Salentino, Northern Calabrese, Campano

Example

LanGeLin data:



Hellenic languages: Salento Greek, Calabrian Greek A, Calabrian Greek B, Modern Greek, Cypriot Greek

Example: Clustering and the Greek-Italian Microvariations

- different clustering in SSWL and LanGeLin: Hellenic languages
  - in SSWL certain Hellenic languages (Cappadocian Greek, Modern Greek) remain singletons for a long range of radii and join other clusters very late in the persistence scale
  - in LanGeLin the Hellenic languages join into clusters earlier in the persistence diagram
- LanGeLin data include a range of Southern-Italian dialect that are either Romance or Hellenic (Salento Greek, Calabrian Greek A, Calabrian Greek B)

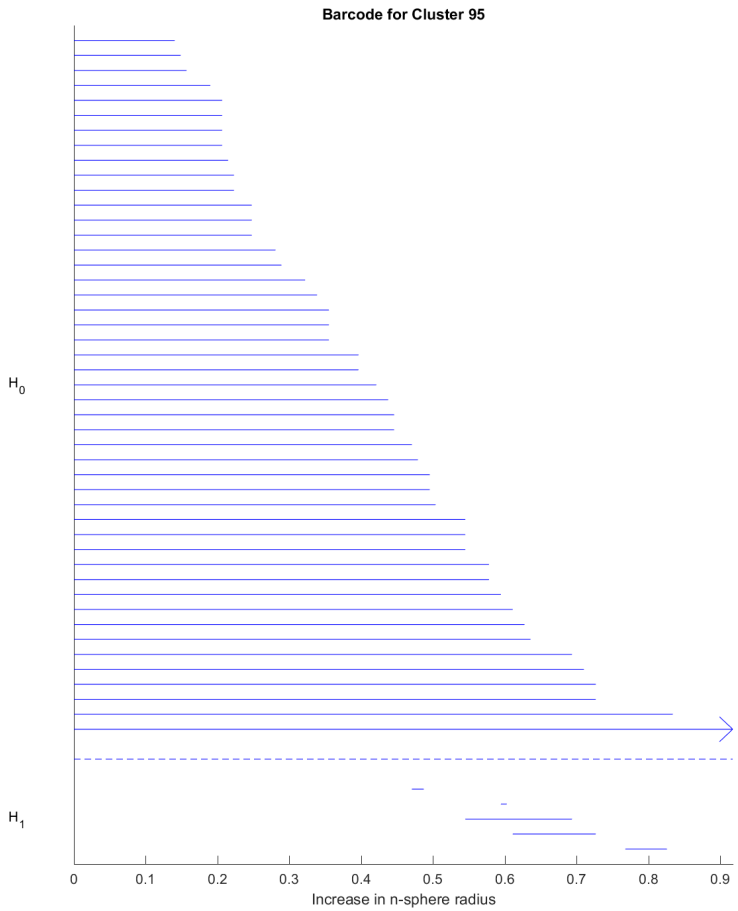
• Microvariations: languages either genealogically very closely related or in distinct genealogical groups but in close geographic proximity and interaction

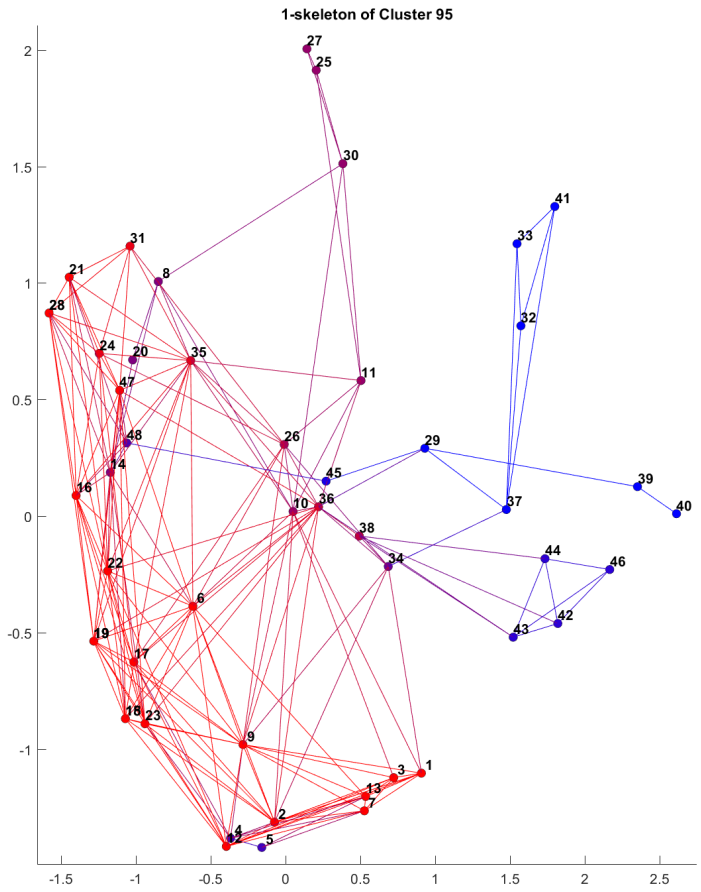
• These Italian-Greek Microvariations studied at length in

- C. Guardiano, D. Michelioudakis, A. Ceolin, M. Irimia, G. Longobardi, N. Radkevich, I. Sitaridou, G. Silvestri, *South by Southeast. A Syntactic Approach to Greek and Romance Microvariation*, L'Italia Dialettale, Vol. 77 (2016) 95–166.

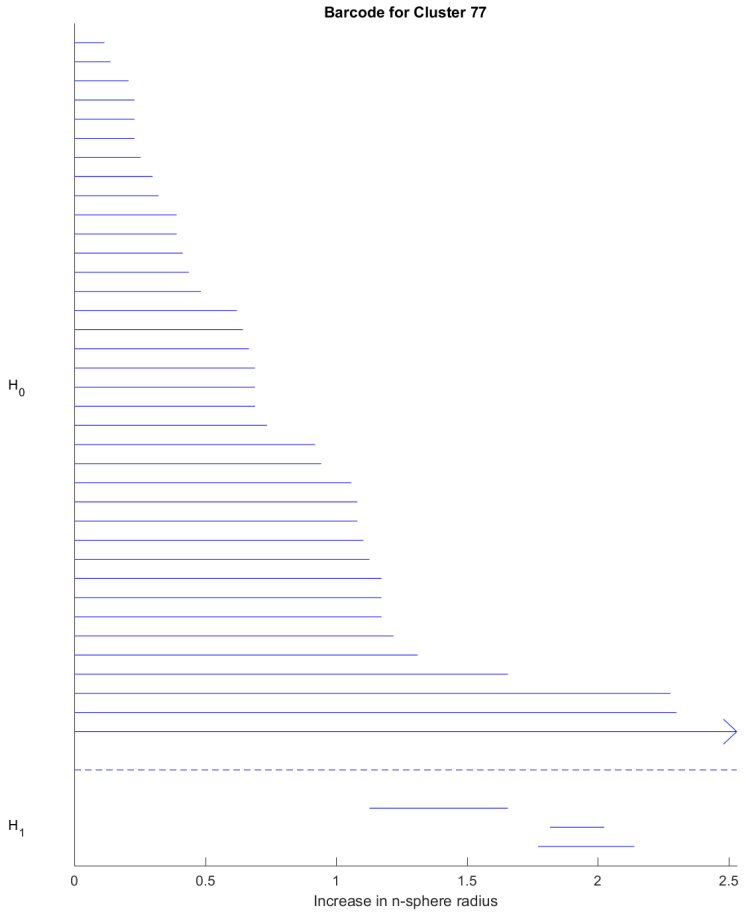
Persistent First Homology of Language Families

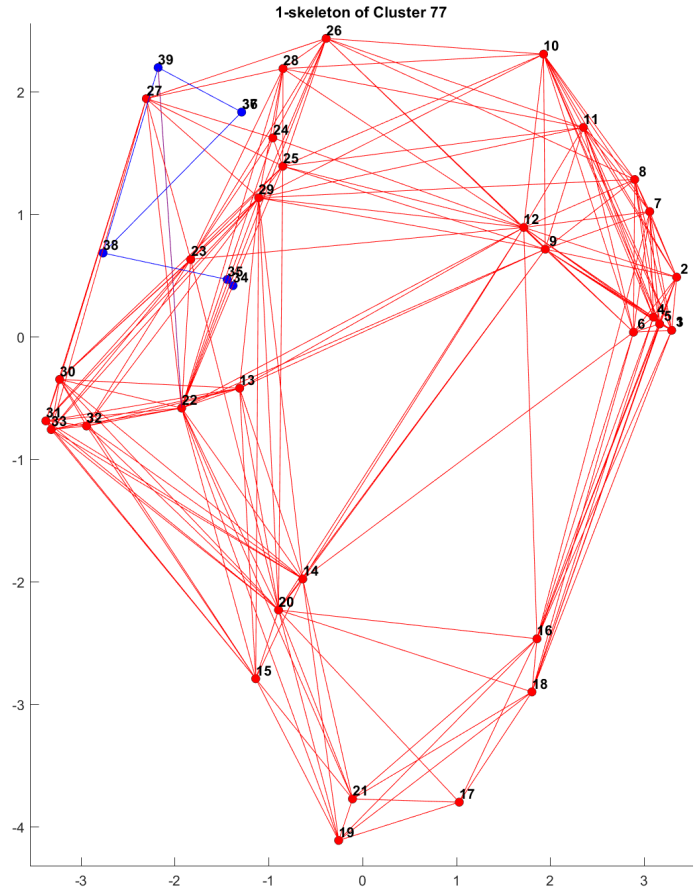
- SSWL data: Indo-European + Ural and Altaic languages





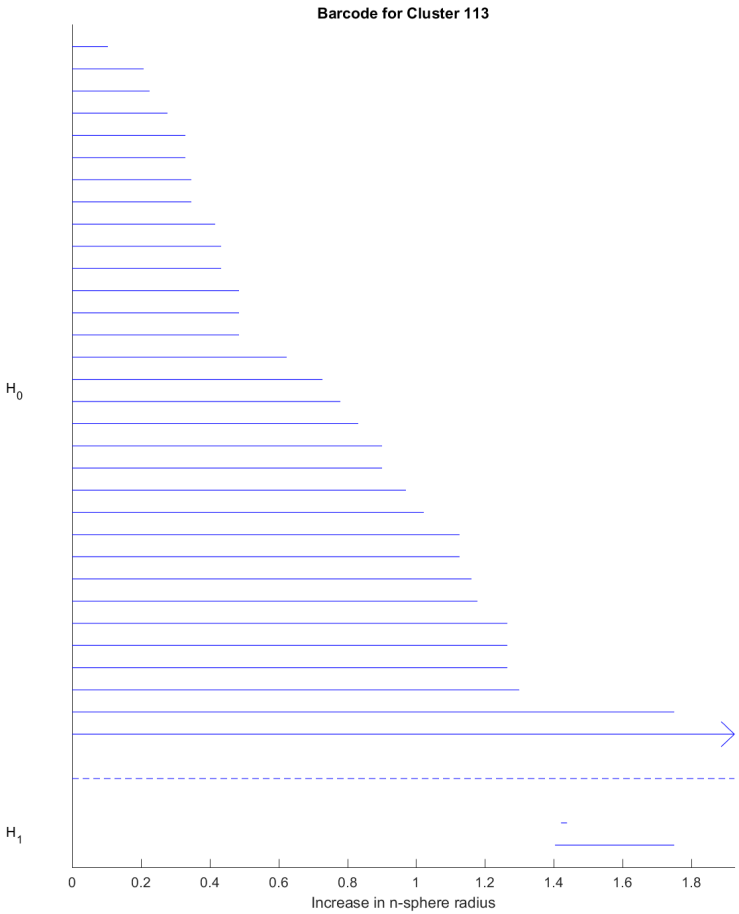
- LanGeLin data: Indo-European + Ural and Altaic languages

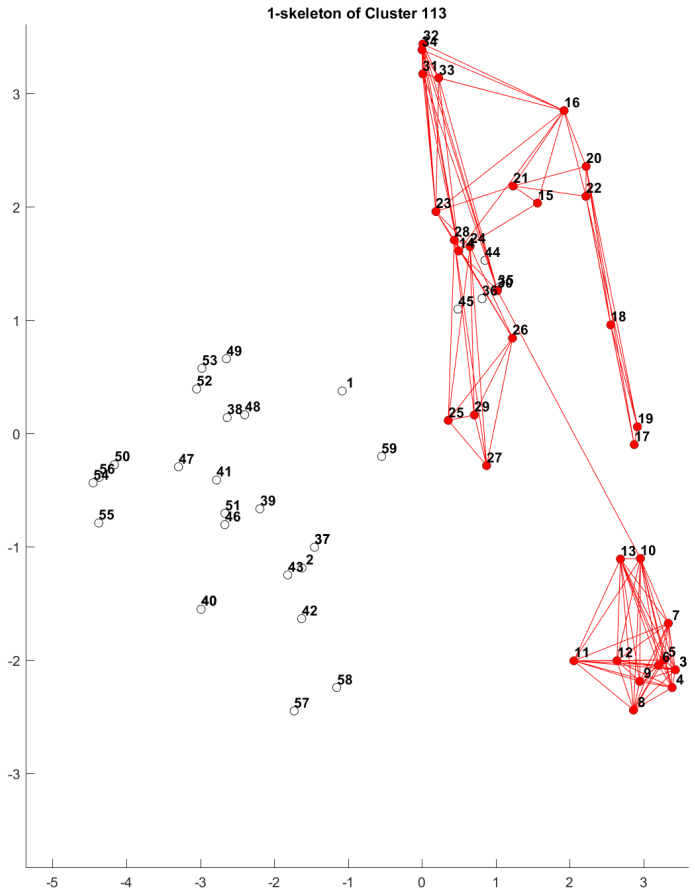




Gothic–Slavic–Greek loop (Example with historical linguistic explanation)

- in the LanGeLin data





Identify cycle representative for persistent H_1-generator

1. identify first cluster of persistent components where new H_1-generator appears
2. list languages (vertices) added and all new cycles added in Vietoris-Rips 1-skeleton
3. in turn remove the languages belonging to one of the new cycles and recompute
4. if new generator disappears have a cycle representative
5. homologous cycles (remove all at once)

Gothic–Slavic–Greek loop: forms in the Indo-European languages between New Testament Greek, Romeyka Pontic Greek, Gothic, and Slavic languages (need to remove all Slavic languages together: homologous cycles)

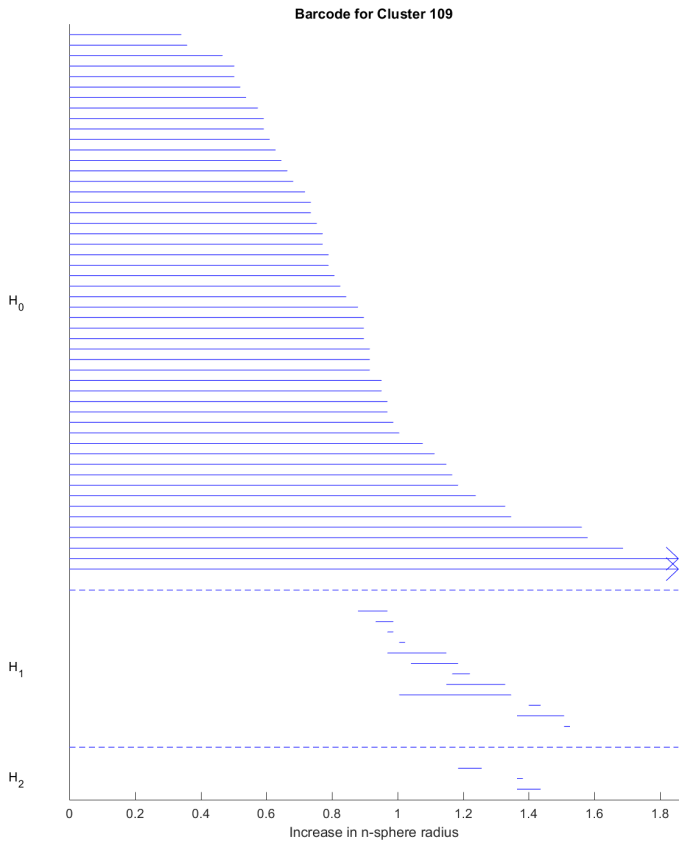
Possible historical linguistic explanation for this H_1-generator

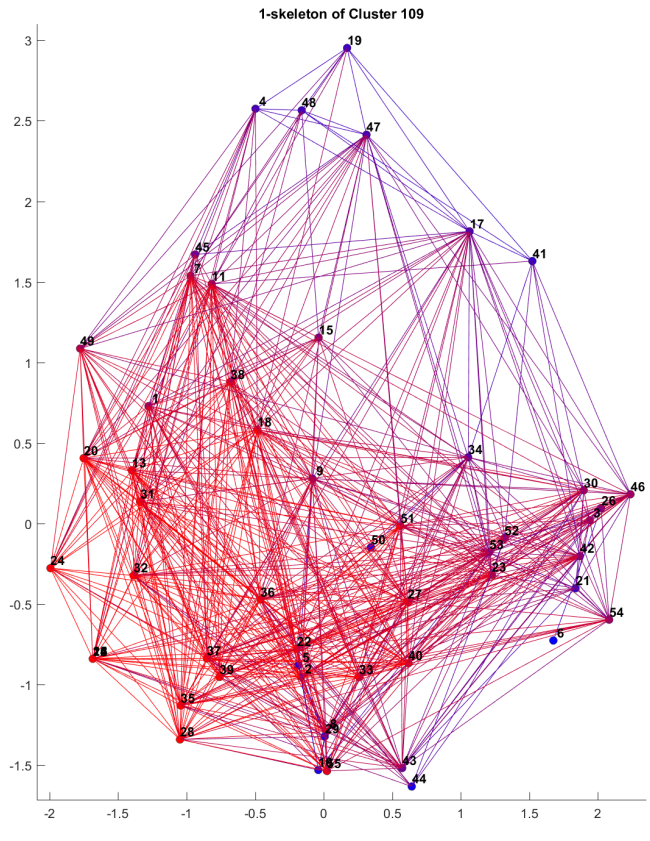
- influence (also at syntactic level) between Greek languages and South Slavic languages
- syntactic influence of New Testament Greek on Gothic (observed calques of Greek constructions in Gothic syntax)
- Proto-Slavic borrowing (influence of Gothic mostly lexical, but indications of morpho-syntactic borrowing as well)

- Some References:
  - O. Mišeska-Tomić, *Balkan Sprachbund. Morpho-syntactic Features*, Dordrecht, Springer 2006
  - J.D. Gliesche, *Gothic Syntax*, lecture notes `http://users.clas.ufl.edu/drjdg/oe/pubs/gothicsyntax.pdf`
  - R. Genis, *Comparing verbal aspect in Slavic and Gothic*, Amsterdam contributions to Scandinavian studies; No. 8, (2012) 59–80.

- other H_1-generators may reflect only homoplasy phenomena

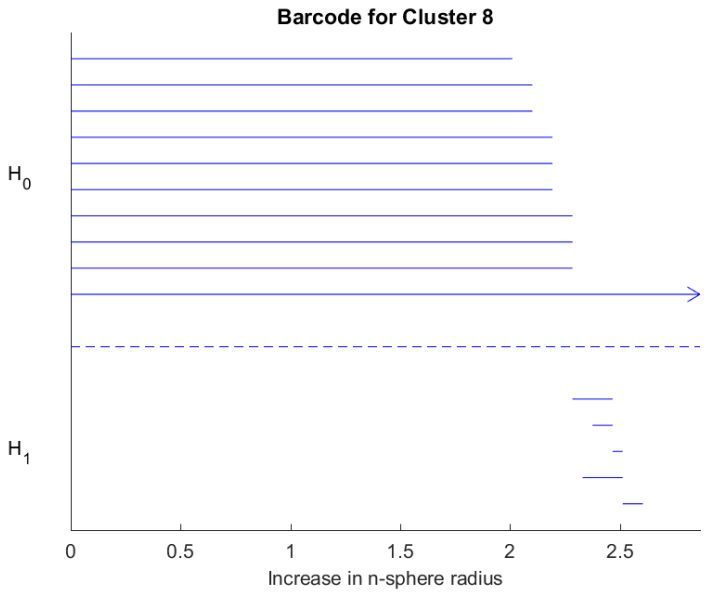
Persistent Homology of the Niger-Congo languages

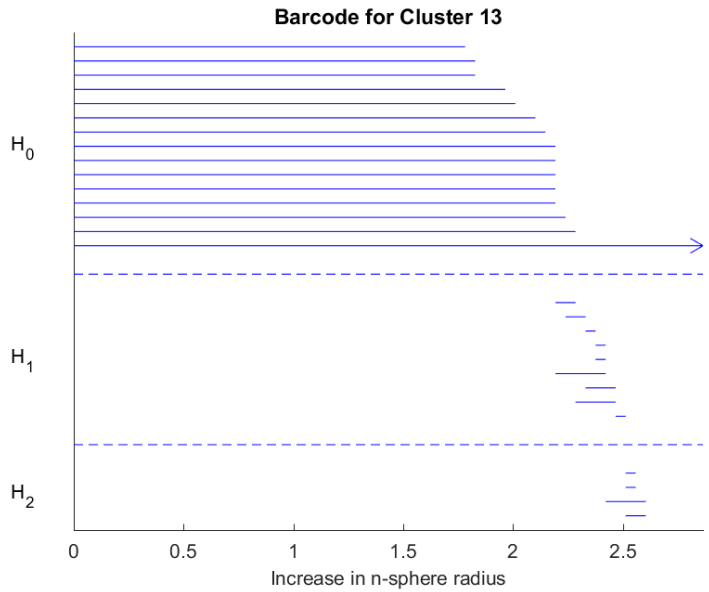




- More persistent homology in high clusters than other language families (not seen in previous work where only a few clusters analyzed)
- Only example found so far of non-trivial persistent H_2: what does it mean???

General Observation: Comparison with homology of random simplicial sets





main differences: in random case H_1 occurs already in small clusters, shorter persistence, more stacking of many generators

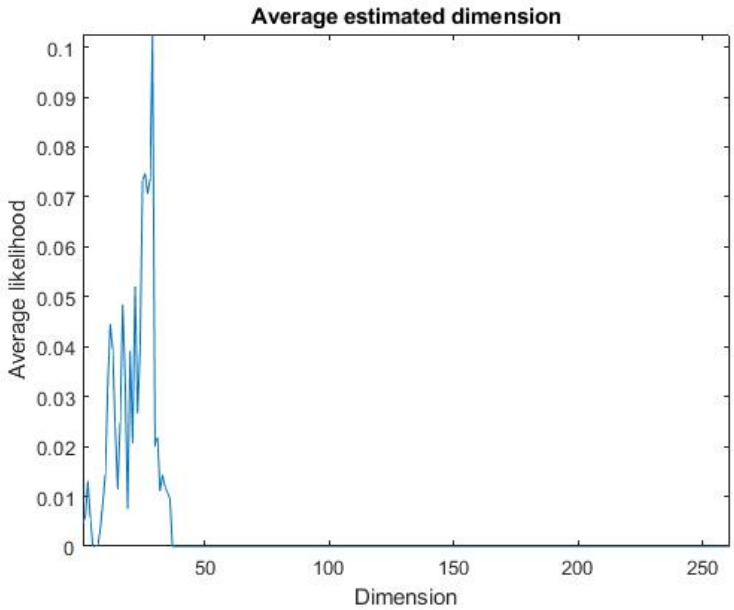
Conclusions: more work (computationally hard) to identify representative cycles for all these H_1-generators for different language families, separate those that appear caused by homoplasy from those with historical linguistic significance; linguistic interpretation of H_2?

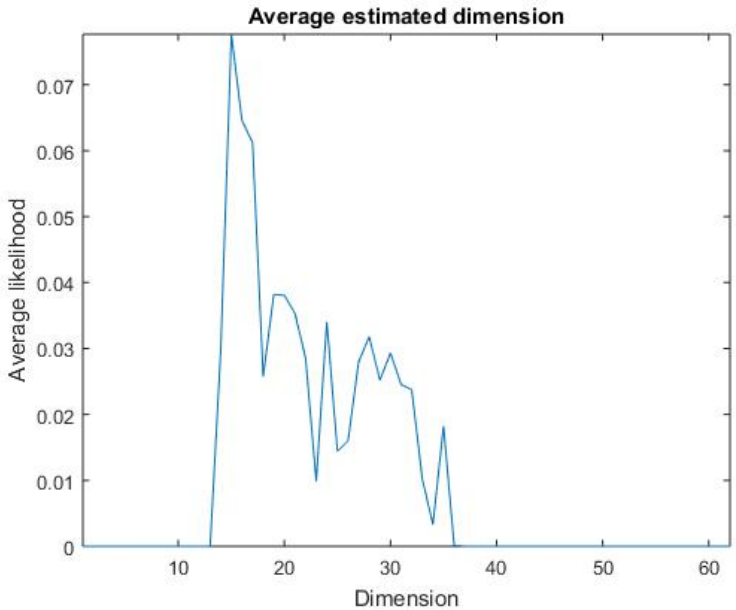
Shift of perspective: seek relations between syntactic parameters instead of relations between languages, using values of parameters on the given set of languages.

- What is the estimated dimension of the space of syntactic parameters?
- Are there additional relations that are specific to language families and do not hold universally? (drop in dimension)
- Which parameters cluster together?
- Do syntactic parameters span a manifold?
- What is the (persistent) topology of this manifold?
- Are some parameters more easily recoverable from others?

Estimated dimension of syntactic parameters

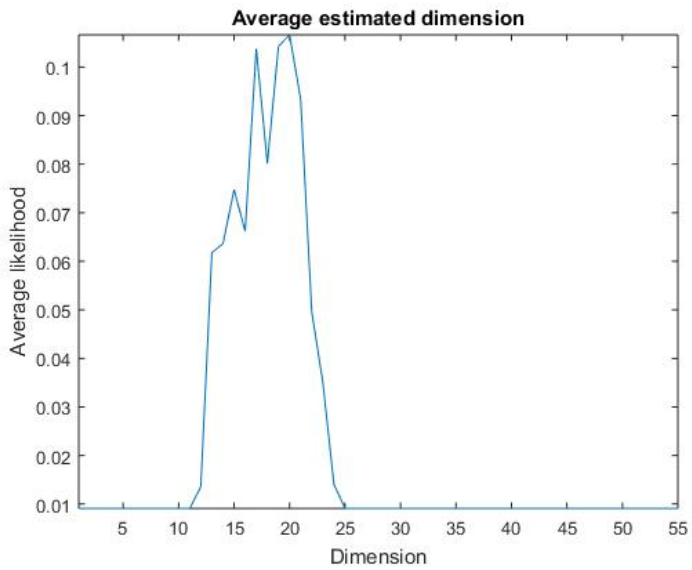
* Alexander Port, Taelin Karidi, Matilde Marcolli, *Topological Analysis of Syntactic Structures*, arXiv:1903.05181

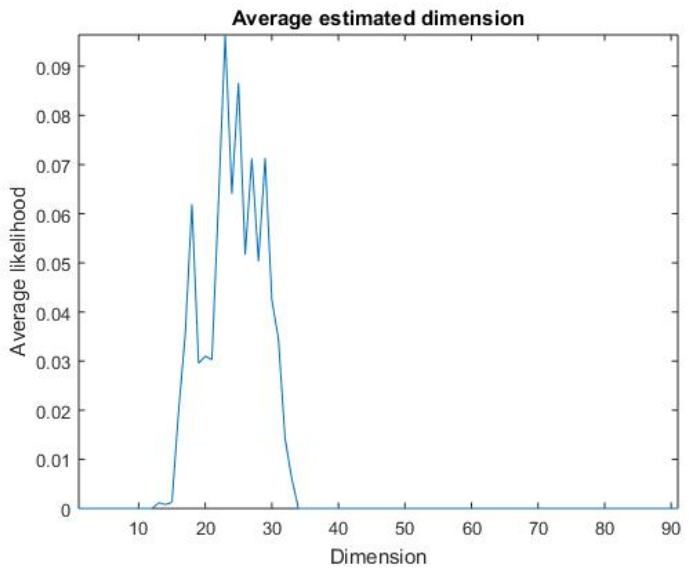




- Dimension of SSWL syntactic variables peak $d \sim 30$ (116 dim ambient space)
- Dimension of LanGeLin syntactic variables peak $d \sim 15$ (83 dim ambient space)

Family specific relations: dimension drop from $d \sim 30$ of SSWL





- Niger-Congo languages (SSWL data) $d \sim 20$
- Indo-European languages (SSWL data) $d \sim 23$

More work on this upcoming:

* Sitanshu Gakkhar, Matilde Marcolli, *Metric Distortion, Graph Expansion, and Syntactic Structures*, in preparation, 2019

Heat Kernel Method

* Andrew Ortegaray, Robert C. Berwick, Matilde Marcolli, *Heat Kernel analysis of Syntactic Structures*, arXiv:1803.09832

Geometric methods of dimensional reduction:

- M. Belkin, P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Comput. 15 (6) (2003) 1373–1396

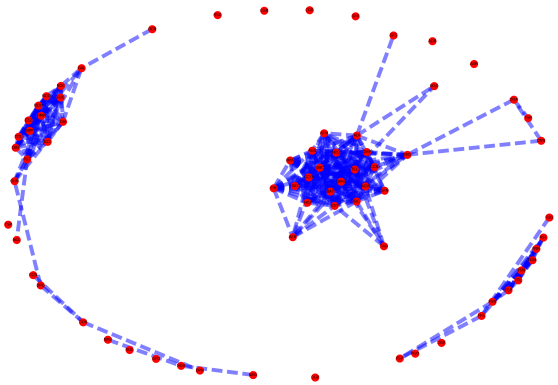
• *Problem*: low dimensional representations of data sampled from a probability distribution on a manifold

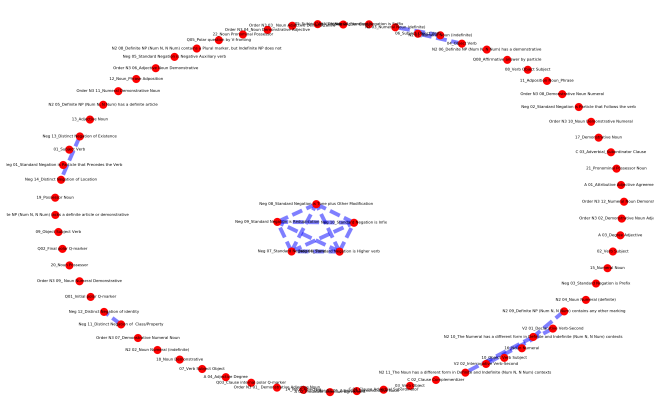
• *Main Idea*: build a graph with neighborhood information, use Laplacian of graph, want low dimensional representation that maintains local neighborhood information

• *Key Result*: graph Laplacian for a set of data point sampled from a uniform distribution on a manifold converges to Laplace–Beltrami operator on the manifold for large sets (using heat kernel and relation to Laplacian)

• Use to construct optimal (preserving information on manifold geometry) mapping of data sets to low dimensional spaces via eigenfunctions of Laplacian

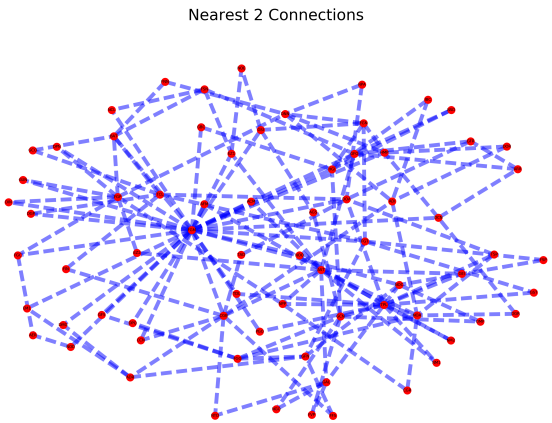
Parameter clustering via heat kernel method

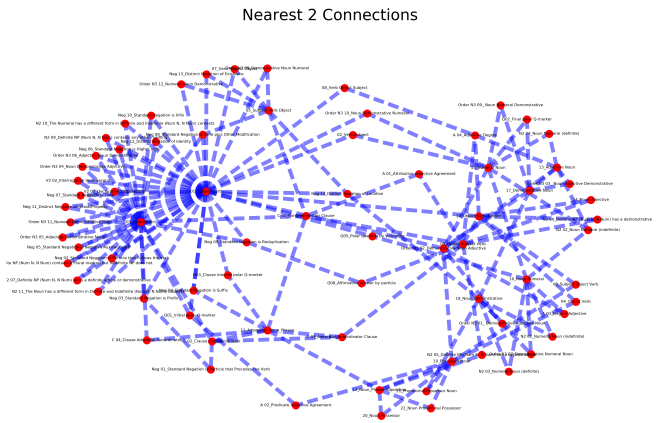




- ϵ-neighbors graphs for LanGeLin and SSWL data

Parameter clustering via heat kernel method



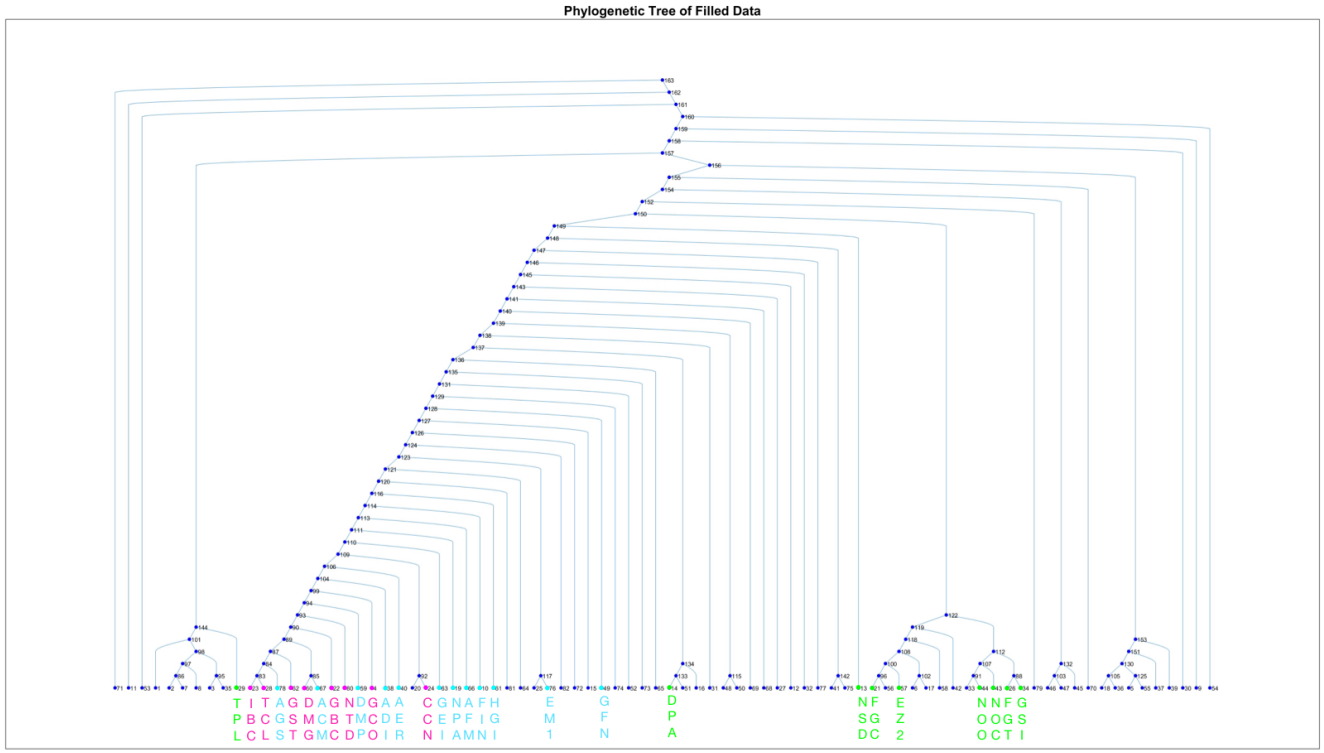


- n-nearest neighbors graphs for LanGeLin and SSWL data, $n = 2$

Comparison with clustering of parameters via persistent connected components in

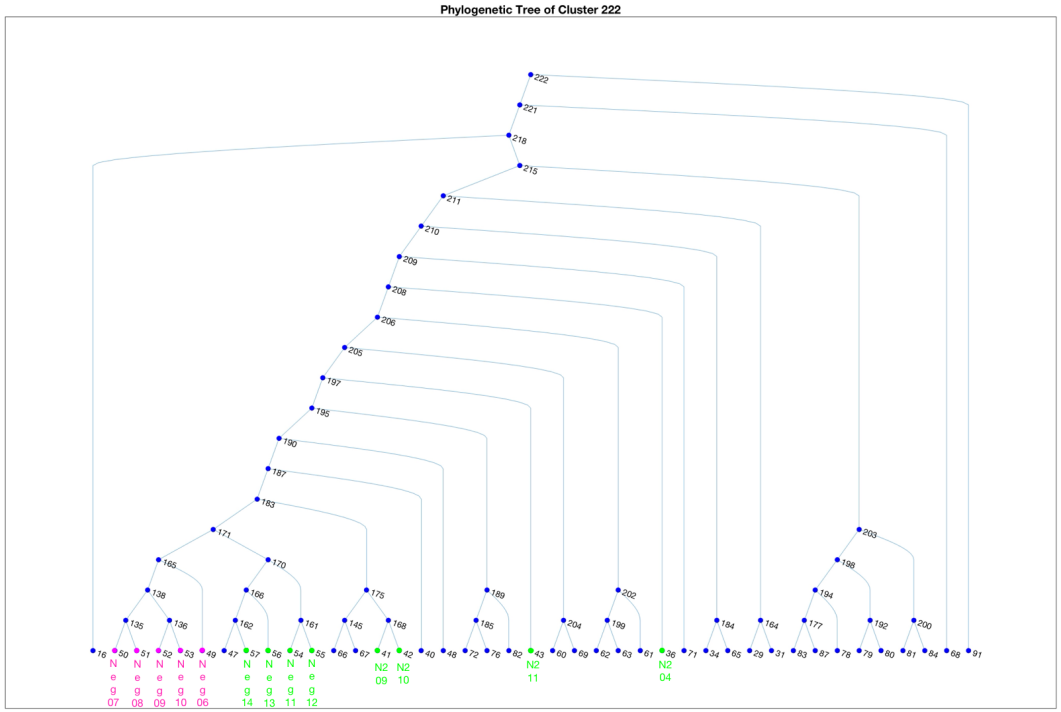
* Alexander Port, Taelin Karidi, Matilde Marcolli, *Topological Analysis of Syntactic Structures*, arXiv:1903.05181

Parameter clustering via persistent connected components
Persistent Components Tree of LanGeLin parameters compared to heat kernel clusters



pink-colored and blue-colored: same as in two main sub-clusters of first cluster with heat-kernel method; green-colored second smaller cluster in heat-kernel method

Persistent Components Tree of SSWL syntactic variables compared to heat ker



pink-colored: first heat kernel cluster; green-colored: second

More ongoing work on parameter clustering:

* Sitanshu Gakkhar, Matilde Marcolli, *Metric Distortion, Graph Expansion, and Syntactic Structures*, in preparation, 2019

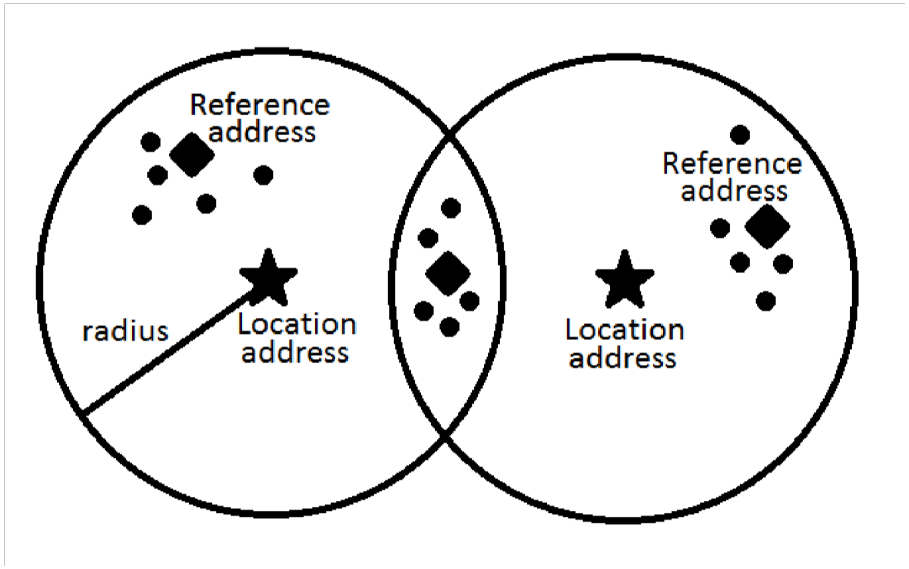
Recoverability in Kanerva Networks

* Jeong Joon Park, Ronnel Boettcher, Andrew Zhao, Alex Mun, Kevin Yuh, Vibhor Kumar, Matilde Marcolli, *Prevalence and recoverability of syntactic parameters in sparse distributed memories*, in "Geometric Structures of Information 2017", Lecture Notes in Computer Science, Vol. 10589 (2017) 1–8

Kanerva networks (sparse distributed memories)

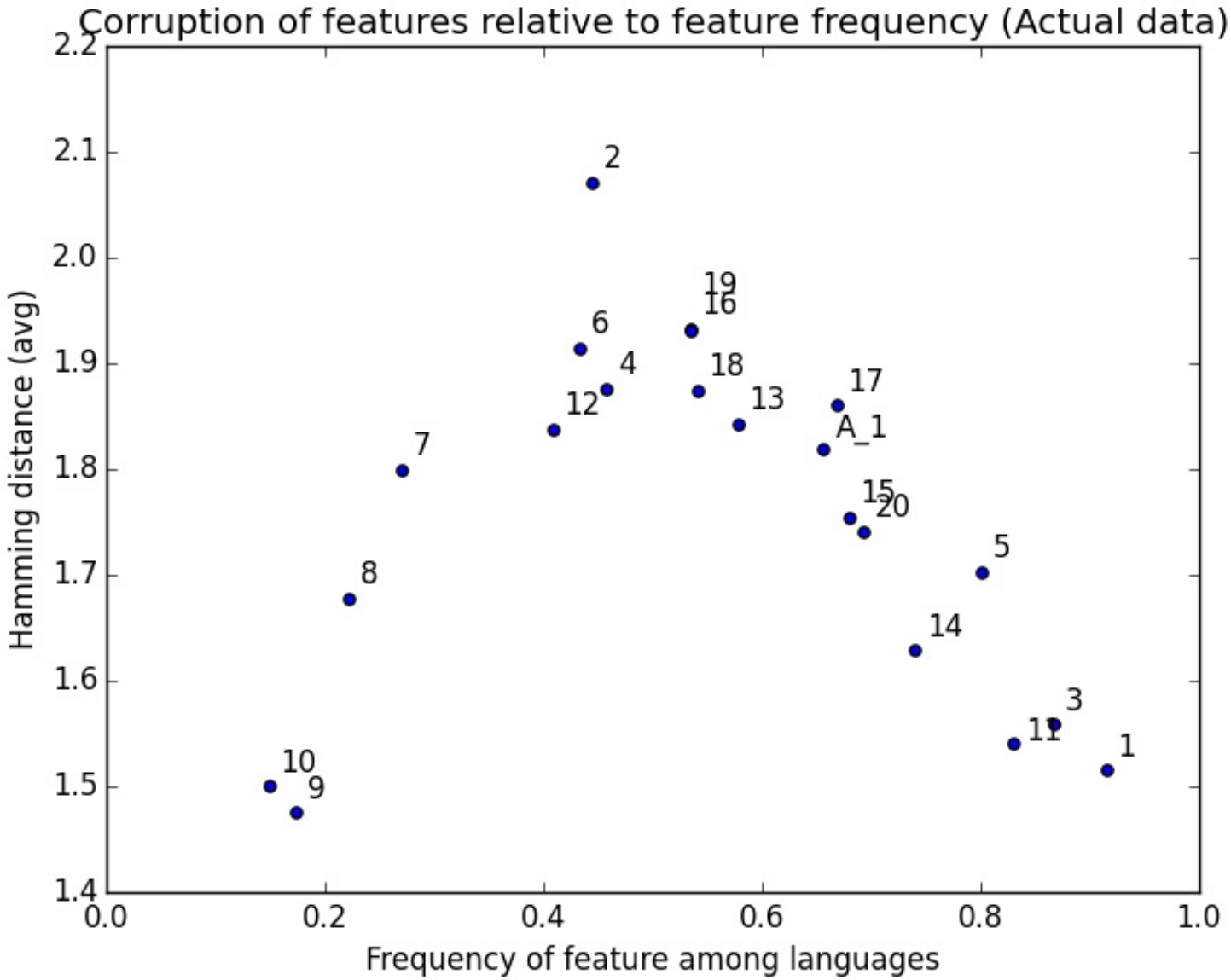
- P. Kanerva, *Sparse Distributed Memory*, MIT Press, 1988.
- field $\mathbb{F}_2 = \{0, 1\}$, vector space $\mathbb{F}_2^N$ large N
- uniform random sample of 2^k hard locations with $2^k << 2^N$
- access sphere: Hamming spheres around location, radius slightly larger than median
- *writing to network*: storing datum $X \in \mathbb{F}_2^N$, each hard location in access sphere of X gets i-th coordinate incremented depending on i-th entry of X
- *reading at a location*: i-th entry determined by majority rule of i-th entries of all stored data in hard locations within access sphere

Kanerva networks are good at reconstructing corrupted data

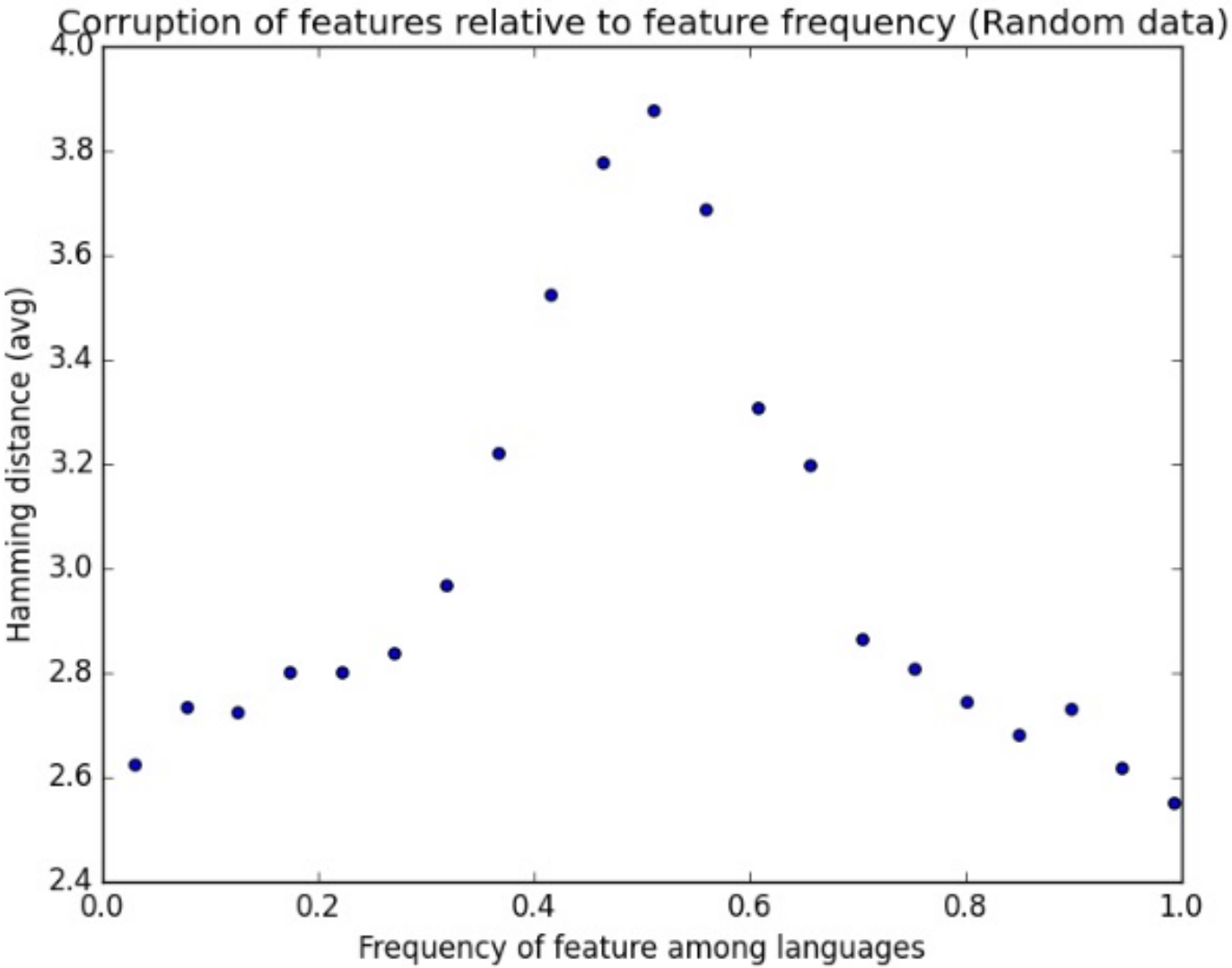


Procedure

- 165 data points (languages) stored in a Kanerva Network in $\mathbb{F}_2^{21}$ (choice of 21 SSWL parameters)
- corrupting one parameter at a time: analyze recoverability
- language bit-string with a single corrupted bit used as read location and resulting bit string compared to original bit-string (Hamming distance)
- resulting average Hamming distance used as score of recoverability (lowest = most easily recoverable parameter)

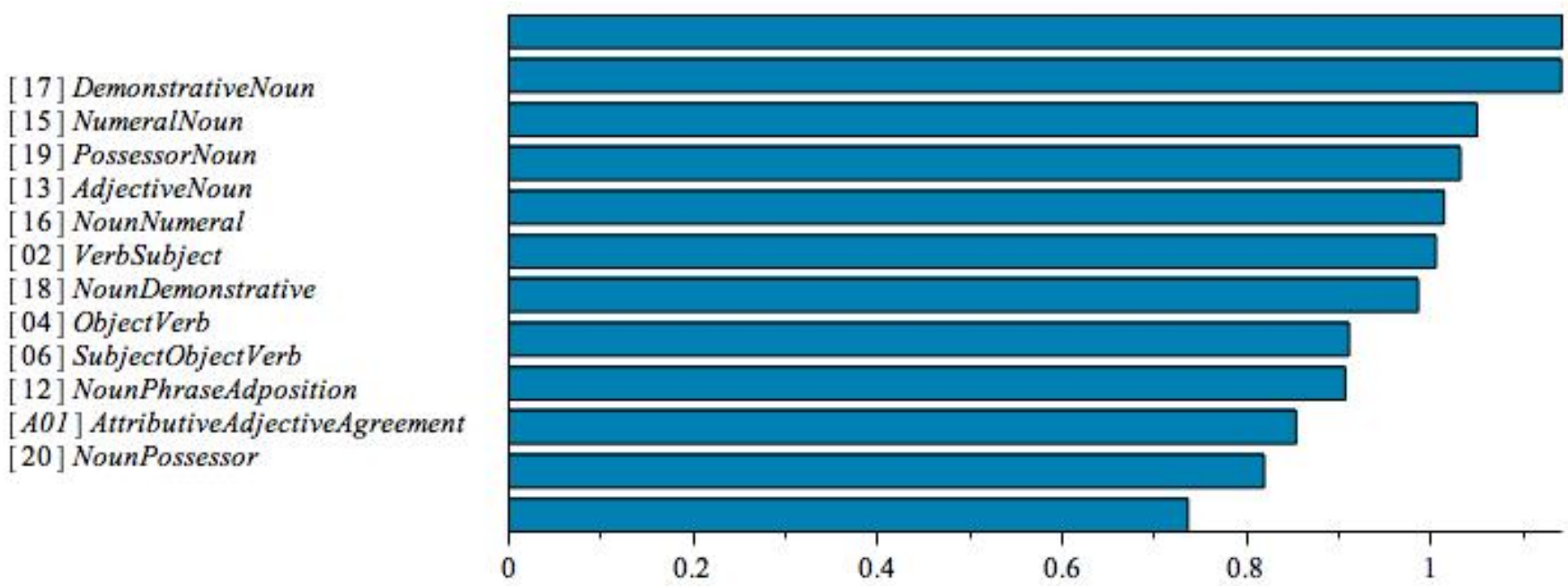


Specific effects due to individual parameters



Overall effect related to relative prevalence of a parameter (frequency of expression among languages)

More refined effect after normalizing for prelavence (syntactic dependencies)



What does this tell us? some SSWL syntactic variables have a much higher degree of recoverability than others: consider them dependent variables; can sparse distributed memories model how syntax is in fact stored in the human brain?

A Couple of Future Questions:

• Can one infer possible *algebraic* relations between syntactic parameters based on information such as persistent homology and dimension?

- Paul Breiding, Sara Kalisnik Verovsek, Bernd Sturmfels, Madeleine Weinstein, *Learning Algebraic Varieties from Samples*, arXiv:1802.09436
- Emilie Dufresne, Parker B. Edwards, Heather A. Harrington, Jonathan D. Hauenstein, *Sampling real algebraic varieties for topological data analysis*, arXiv:1802.07716

• Can form dynamical models of language change improving on the Markov models on trees of phylogenetic algebraic geometry by incorporating relations between parameters?

- Karthik Siva, Jim Tao, Matilde Marcolli, *Spin Glass Models of Syntax and Language Evolution*, Linguistic Analysis, Vol.41 (2017) N.3-4, 559-608.
- P. Niyogi, R.C. Berwick, *A Dynamical Systems Model for Language Change*, Complex Systems. 11 (1996)