The Geometry of Syntax

Matilde Marcolli

TGSI: Topological and Geometrical Structures of Information
CIRM Luminy
August 28–September 1, 2017

Talk based on some recent and ongoing work

1. Matilde Marcolli, Robert Berwick, Kevin Shu, *Phylogenetics of Indo-European Language families via an Algebro-Geometric Analysis of their Syntactic Structures*, in preparation;
2. Andrew Ortegaray, Matilde Marcolli, *A Heat Kernel analysis of the space of Syntactic Parameters*, in preparation;
3. Alexander Port, Matilde Marcolli, *Persistent topology and syntactic parameters of Indo-European languages*, in preparation.
4. Matilde Marcolli, *Syntactic Parameters and a Coding Theory Perspective on Entropy and Complexity of Language Families*, Entropy 2016, 18(4), 110
5. Kevin Shu, Matilde Marcolli, *Syntactic Structures and Code Parameters*, Math. Comput. Sci. 11 (2017), no. 1, 79–90.

and some revious work

1. Alexander Port, Iulia Gheorghita, Daniel Guth, John M.Clark, Crystal Liang, Shival Dasu, Matilde Marcolli, *Persistent Topology of Syntax*, arXiv:1507.05134
2. Karthik Siva, Jim Tao, Matilde Marcolli, *Spin Glass Models of Syntax and Language Evolution*, arXiv:1508.00504
3. 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*, arXiv:1510.06342
4. Kevin Shu, Sharjeel Aziz, Vy-Luan Huynh, David Warrick, Matilde Marcolli, *Syntactic Phylogenetic Trees*, arXiv:1607.02791

General Question: Language and Machines

- Natural Language Processing has made enormous progress in problems like automated translation

- but can we use computational (mathematical) techniques to better understand how the human brain processes language?

- some of the main questions:
  - Language acquisition (poverty of the stimulus): how does the learning brain converge to *one* grammar?
  - How is language (in particular syntax) stored in the brain?
  - How do languages change and evolve in time? quantitative, predictive modeling?

- Plan: approach these questions from a mathematical perspective, using tools from geometry and theoretical physics

- focus on the "large scale structure" of language: syntax

Syntax and Syntactic Parameters

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

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

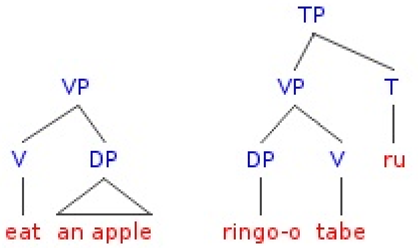
• this idea is very appealing for a mathematician: at the level of syntax a language can be described by a set of coordinates given by binary variables

• however, surprisingly no mathematical model of Principles and Parameters formulation of Linguistics has been developed so far

What are the 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

- there are hundreds of such binary syntactic variables, but not all of them are "true" syntactic parameters (conflations of deep/surface structure)

- 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. another set of data from Longobardi–Guardiano, Lingua 119 (2009) 1679-1706
5. more complete set of data by Giuseppe Longobardi, 2016

- Data Analysis of syntax of world languages with various mathematical tools (persistent topology, etc.)

Problems of SSWL data

- Very non-uniformly mapped across the languages of the database: some are 100% mapped, while for some only very few of the 116 parameters are mapped
- Linguists criticize the choice of binary variable (not all of them should count as "true" parameters)

• the data of Longobardi–Guardiano and the more recent data of Longobardi are more reliable: 83 parameters and 62 languages (mostly Indo-European), more completely mapped, "true" parameters

• linguistic question: can languages that are far away in terms of historical linguistics end up being close in terms of syntactic parameters?

Phylogenetic Algebraic Geometry of Languages (ongoing work with R.Berwick, K.Shu)

- Linguistics has studied in depth how languages change over time (Philology, Historical Linguistics)

- Usually via lexical and morphological analysis

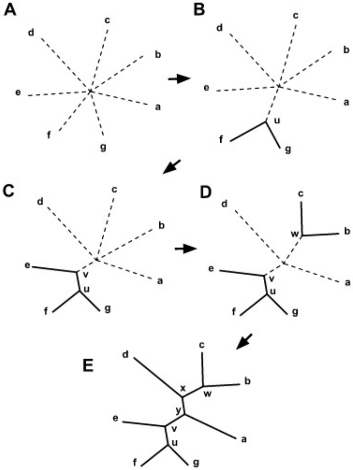
- Goal: understand the historical relatedness of different languages, subdivisions into families and sub-families, phylogenetic trees of language families

- Historical Linguistics techniques work best for language families where enough ancient languages are known (Indo-European and very few other families)

- Can one reconstruct phylogenetic trees computationally using only information on the modern languages?

- controversial results about the Indo-European tree based on *lexical data*: Swadesh lists of lexical items compared on the existence of cognate words (many problems: synonyms, loan words, false positives)

- Some phylogenetic tree reconstructions using syntactic parameters by Longobardi–Guardiano using their parameter data
- Hamming distance between binary string of parameter values + neighborhood joining method

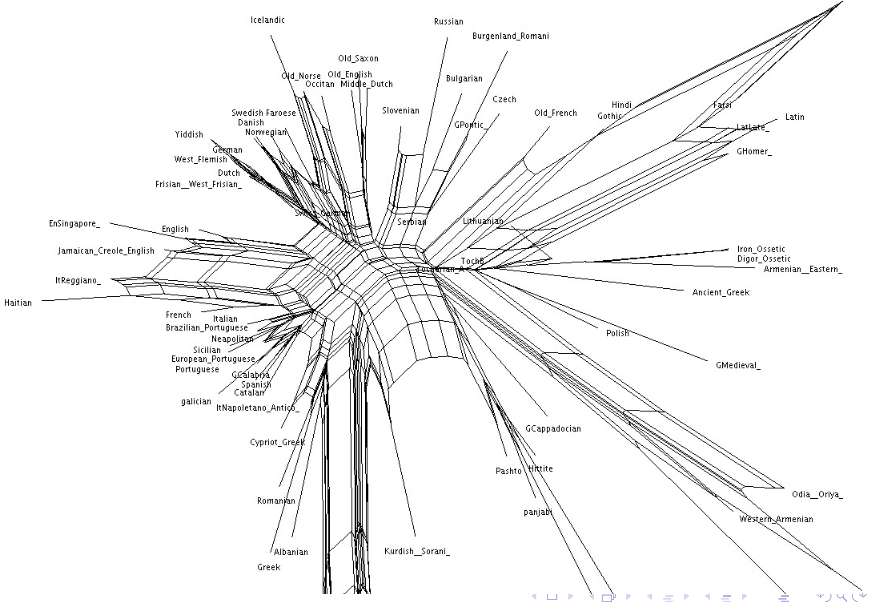


Expect problems: SSWL data and phylogenetic reconstructions

- known problems related to the use of Hamming metric for phylogenetic reconstruction
- SSWL problems mentioned above (especially non-uniform mapping)
- dependence among parameters (not independent random variables)
- syntactic proximity of some unrelated languages

- Phylogeny Programs for trees and networks
  - PHYLIP
  - Splittree 4
  - Network 5

Checking on the Indo-European tree where good Historical-Linguistics



Indeed Problems

- misplacement of languages within the correct family subtree
- placement of languages in the wrong subfamily tree
- proximity of languages from unrelated families (all SSWL)
- incorrect position of the ancient languages

• different approach: subdivide into subfamilies (some a priori knowledge from morpholexical linguistic data) and use Phylogenetic Algebraic Geometry (Sturmfels, Pachter, et al.) for statistical inference of phylogenetic reconstruction

General Idea of Phylogenetic Algebraic Geometry

- Markov process on a binary rooted tree (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

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

with sum over "histories" consistent with data at leaves

- polynomial map that assigns

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

defines an *algebraic variety*

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

- (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

Procedure

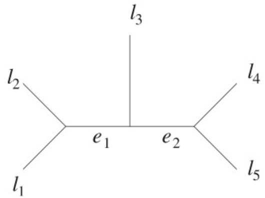
- set of languages $\mathcal{L} = \{\ell_1, \dots, \ell_n\}$ (selected subfamily)
- set of SSWL (or Longobardi) syntactic parameters mapped for all languages in the set: π_i, $i = 1, \dots, N$
- gives vectors $\pi_i = (\pi_i(\ell_j)) \in \mathbb{F}_2^n$
- compute frequencies

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

  with $N_{i_1,\dots,i_n}$ = number of occurrences of binary string $(i_1, \dots, i_n) \in \mathbb{F}_2^n$ among the $\{\pi_i\}_{i=1}^N$
- Given a *candidate tree* T, compute all 3×3 minors of each flattening matrix $Flat_{e,T}(P)$, for each edge
- evaluate $\phi_T(P)$ minimum absolute value of these minors
- smallest $\phi_T(P)$ selects best among candidate trees

Simple examples

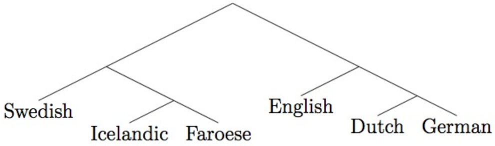
- PHYLIP and Splittree 4 misplace the position of Portuguese among the Latin languages, but phylogenetic invariants identify the correct tree ($\ell_1 =$ French, $\ell_2 =$ Italian, $\ell_3 =$ Latin, $\ell_4 =$ Spanish, $\ell_5 =$ Portuguese)



$$\mathrm{Flat}_{e_1}(P) = \begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} & p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} & p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} & p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} & p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix}$$

$$\mathrm{Flat}_{e_2}(P) = \begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} \\ p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} \\ p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} \\ p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} \\ p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix}$$

- PHYLIP and Splittree 4 misplace the relative position of sub-branches of the Germanic languages, but phylogenetic invariants identify the correct tree (similar computation)



with correct subdivision into North Germanic and West Germanic sub-branches

Conclusion: work with smaller subfamilies, then paste together subtrees; use PHYLIP to generate candidate subtrees and phylogenetic algebraic geometry to select among them

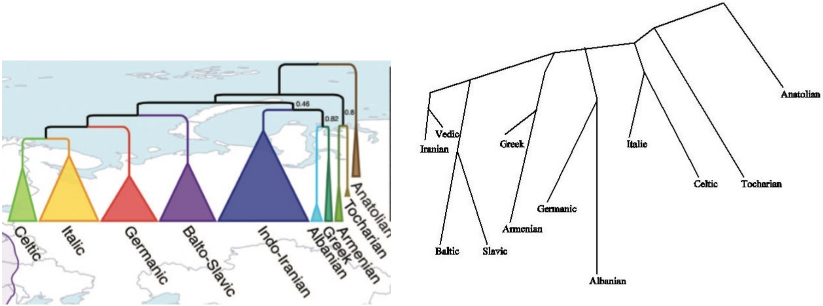
Main Question: can one use this method to obtain new results on the "Indo-European controversy"?

• What is the controversy? 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.

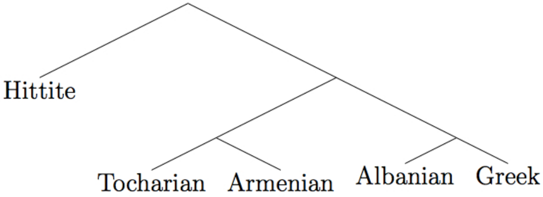
• Controversial claims by Gray and Atkinson (Nature, 2003); disputed via morphological analysis (Ringe, Warnow, Taylor, 2002)

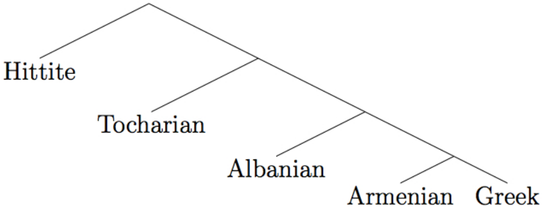
• A. Perelysvaig, M.W. Lewis, *The Indo-European controversy: facts and fallacies in Historical Linguistics*, Cambridge University Press, 2015.



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

Focus on this part of the tree:



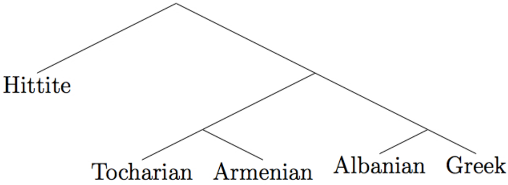


Can detect the difference from syntactic parameters?

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 data.

- the SSWL data appear to favor the Atkinson–Gray tree, *but the data is too problematic to be trusted!* ...need better syntactic data on these languages (especially Hittite and Tocharian that are poorly mapped in all available databases)



Coding Theory to study how syntactic structures differ across the landscape of human languages

• Kevin Shu, Matilde Marcolli, *Syntactic Structures and Code Parameters*, Math. Comput. Sci. 11 (2017), no. 1, 79–90.

• Matilde Marcolli, *Syntactic Parameters and a Coding Theory Perspective on Entropy and Complexity of Language Families*, Entropy 2016, 18(4), 110

- select a group of languages $\mathcal{L} = \{\ell_1, \ldots, \ell_N\}$
- with the binary strings of n syntactic parameters form a code $\mathcal{C}(\mathcal{L}) \subset \mathbb{F}_2^n$
- compute code parameters $(R(\mathcal{C}), \delta(\mathcal{C}))$ code rate and relative minimum distance
- analyze position of (R, δ) in space of code parameters
- get information about “syntactic complexity” of $\mathcal{L}$

code parameters $\mathcal{C} \subset \mathbb{F}_2^n$

- transmission rate (encoding)

$$R(\mathcal{C}) = \frac{k}{n}, \quad k = \log_2(\#\mathcal{C}) = \log_2(N)$$

for q-ary codes in $\mathbb{F}_q^n$ take $k = \log_q(N)$

- relative minimum distance (decoding)

$$\delta(\mathcal{C}) = \frac{d}{n}, \quad d = \min_{\ell_1 \neq \ell_2} d_H(\ell_1, \ell_2)$$

Hamming distance of binary strings of ℓ_1 and ℓ_2

- error correcting codes: optimize for maximal R and δ but constraints that make them inversely correlated

- bounds in the space of code parameters (R, δ)

Bounds on code parameters

- Gilbert-Varshamov curve (q-ary codes)

$$R = 1 - H_q(\delta), \quad H_q(\delta) = \delta \log_q(q-1) - \delta \log_q \delta - (1-\delta) \log_q(1-\delta)$$

q-ary Shannon entropy: asymptotic behavior of volumes of Hamming balls for large n

- The Gilbert-Varshamov curve represents the typical behavior of large random codes (Shannon Random Code Ensemble)

- Plotkin curve $R = 1 - \delta/q$: asymptotically codes below Plotkin curve $R \leq 1 - \delta/q$

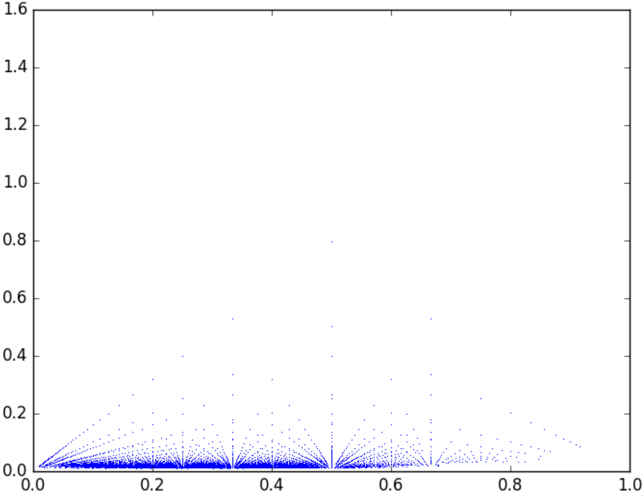
- more significant asymptotic bound (Manin '82) between Gilbert-Varshamov and Plotkin curve

$$1 - H_q(\delta) \leq \alpha_q(\delta) \leq 1 - \delta/q$$

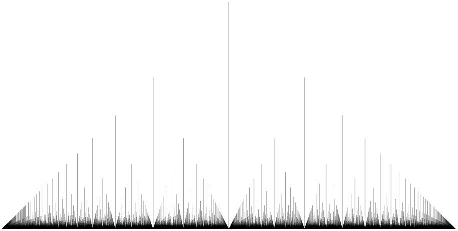
separates a region with dense code points with infinite multipliciites (below) and one with isolated code points with finite multiplicity (good codes above): difficult to find examples

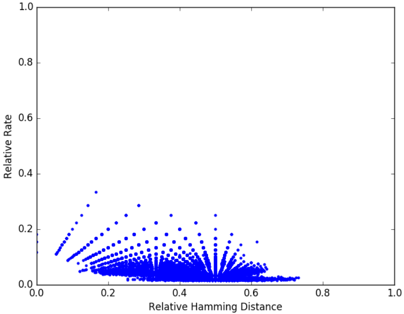
- asymptotic bound not explicitly computable (related to Kolmogorov complexity of codes, Manin–Marcolli)
- difficult to construct codes above the asymptotic bound: examples from algebro-geometric codes from curves (but only for $q \geq 49$ otherwise entirely below the GV curve)

- look at the distribution of code parameters for small sets of languages (pairs or triples) and SSWL data



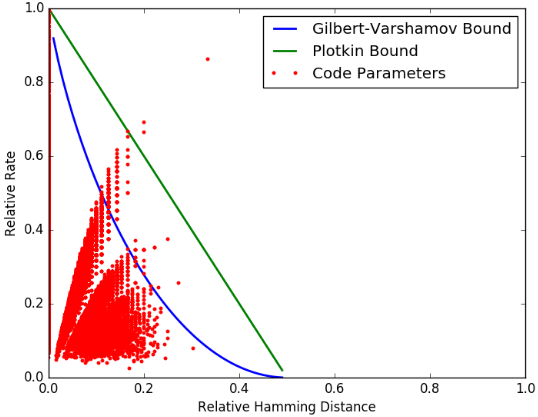
- in lower region of code parameter space a superposition of two Thomae functions ($f(x) = 1/q$ for $x = p/q$ coprime, zero on irrationals)





and behaves like the case of random codes with fixed $k = \log_2(N)$

- more interesting what happens in the upper regions of the code parameter space
- take larger sets of randomly selected languages and syntactic parameters in the SSWL database



codes better than algebro-geometric above GV, asymptotic, and Plotkin

Spin Glass model of Language Evolution

• Karthik Siva, Jim Tao, Matilde Marcolli, *Spin Glass Models of Syntax and Language Evolution*, arXiv:1508.00504, to appear in Linguistic Analysis

• syntactic parameters are dynamical: change over time, effects of *interaction between languages* (Ancient Greek switched SOV to SVO from Homeric to Classical; Sanskrit also switched by influence of Dravidian languages; also Old English to Middle English)

• physicist viewpoint: binary variables (up/down spins) that flip by effect of interactions: Spin Glass Model

Building a Spin Glass Model

- graph: vertices = languages, edges = language interaction (strength proportional to bilingual population)
- over each vertex a set of spin variables (syntactic parameters)
- if all syntactic parameters independent: uncoupled Ising models (low temperature: converge to more prevalent up/down state in initial configuration; high temperature fluctuations around zero magnetization state)
- role of temperature: fluctuations in bilingual users between different structures ("code-switching" in Linguistics)
- Interactions between parameters! .... *coupled* Ising models
- Hamiltonian modeling *entailment relations* in Longobardi–Guardiano data (case where one state of a parameter can make another parameter undefined)

- variables: $S_{\ell,p_1} = \exp(\pi i X_{\ell,p_1}) \in \{\pm 1\}$, $S_{\ell,p_2} \in \{\pm 1, 0\}$ and $Y_{\ell,p_2} = |S_{\ell,p_2}| \in \{0,1\}$

- Hamiltonian $H = H_E + H_V$

$$H_E = H_{p_1} + H_{p_2} = - \sum_{\ell,\ell' \in \text{languages}} J_{\ell\ell'} \left(\delta_{S_{\ell,p_1},S_{\ell',p_1}} + \delta_{S_{\ell,p_2},S_{\ell',p_2}} \right)$$

$$H_V = \sum_{\ell} H_{V,\ell} = \sum_{\ell} J_{\ell}\, \delta_{X_{\ell,p_1},Y_{\ell,p_2}}$$

$J_\ell > 0$ anti-ferromagnetic

- two parameters: *temperature* as before and coupling *energy of entailment*

- if freeze p_1 and evolution for p_2: Potts model with external magnetic field

- Metropolis–Hastings dynamics (some binary some ternary variables)

$$\pi_A(s \to s \pm 1 \pmod 3) = \begin{cases} 1 & \text{if } \Delta_H \leq 0 \\ \exp(-\beta \Delta_H) & \text{if } \Delta_H > 0. \end{cases}$$

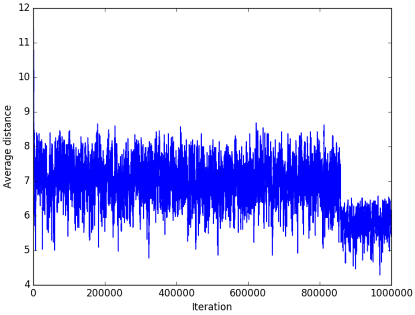
$$\Delta_H := \min\{H(s+1 \pmod 3), H(s-1 \pmod 3)\} - H(s)$$

- obtain interesting dynamics in the case of a small number of languages and parameters with strong entailment relations

- Problem: when consider more realistic models (28 languages and 63 parameters of Longobardi–Guardiano with all the entailment relations) *very slow convergence* of the Metropolis–Hastings dynamics, even for low temperature

- how to get better information on the dynamics? consider set of languages as codes and an *induced dynamics in the space of code parameters*

- Spin Glass Model dynamics for a set of languages $\mathcal{L}$ induces dynamics on codes $\mathcal{C}(\mathcal{L})$ and on code parameters (R, δ)
- without entailment (independent parameters) fixed R and δ flows towards zero (spoiling code)
- with entailment parameters dynamics can improve code making δ larger (R fixed)
- in some cases can see better the dynamics on code parameter than with average magnetization of spin glass model



The Manifold of Syntax? looking for relations between parameters (ongoing work with Andrew Ortegaray)

- Geometric methods of dimensional reduction: *Belkin–Niyogi heat kernel method*

- 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

- *Want* more efficient methods than Principal Component Analysis

- *Main Idea*: build a graph with neighborhood information, use Laplacian of graph, obtain low dimensional representation that maintains the local neighborhood information using eigenfunctions of the Laplacian

- setting: data points $x_1, \dots, x_k \in \mathcal{M} \subset \mathbb{R}^\ell$ on a manifold; find points $y_1, \dots, y_k$ in a low dimensional $\mathbb{R}^m$ ($m << \ell$) that *represent* the data points x_i

- Step 1 (a): adjacency graph (ϵ-neighborhood): an edge e_{ij} between x_i and x_j if $\|x_i - x_j\|_{\mathbb{R}^\ell} < \epsilon$

- Step 1 (b): adjacency graph (n nearest neighborhood): egde e_{ij} between x_i and x_j if x_i is among the n nearest neighbors of x_j or viceversa

- Step 2: weights on edges: heat kernel

$$W_{ij} = \exp\left(- \frac{\|x_i - x_j\|^2}{t} \right)$$

if edge e_{ij} and $W_{ij} = 0$ otherwise; heat kernel parameter $t > 0$

- Step 3: Eigenfunctions for connected graph (or on each component)

$$L\psi = \lambda D\psi$$

diagonal matrix of weights $D_{ii} = \sum_j W_{ji}$; Laplacian $L = D - W$ with $W = (W_{ij})$; eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{k-1}$ and ψ_j eigenfuctions

$$\psi_i : \{1, \dots, k\} \to \mathbb{R}$$

defined on set of vertices of graph

- Step 4: Mapping by Laplace eigenfunctions

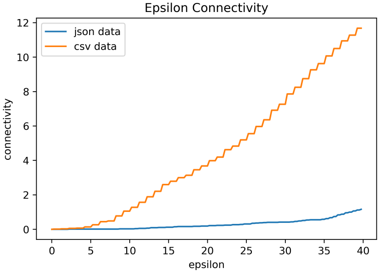
$$\mathbb{R}^\ell \supset \mathcal{M} \ni x_i \mapsto (\psi_1(i), \dots, \psi_m(i)) \in \mathbb{R}^m$$

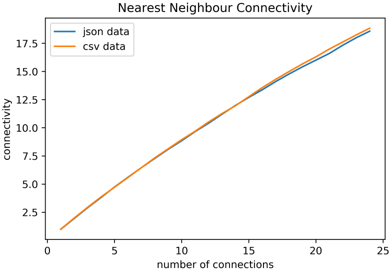
map by first m eigenfunctions

- Belkin–Niyogi: *optimality* of embedding by Laplace eigenfunctions

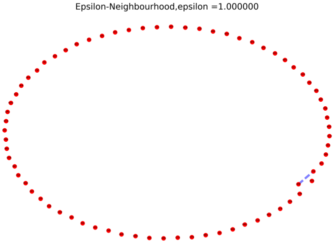
Heat Kernel analysis of Syntactic Parameters

- Connectivity in ϵ-neighborhood and nearest-neighbor (difference between SSWL data (json) and Longobardi data (csv)

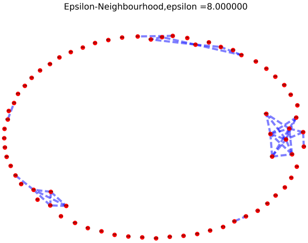




Graphs with ϵ-neighborhood Longobardi data



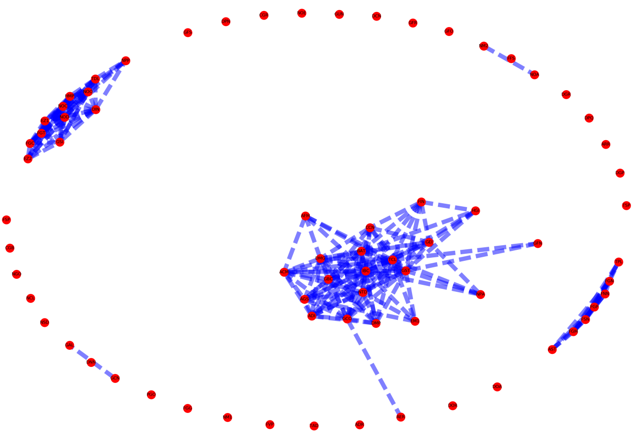
Epsilon-Neighbourhood,epsilon =1.000000



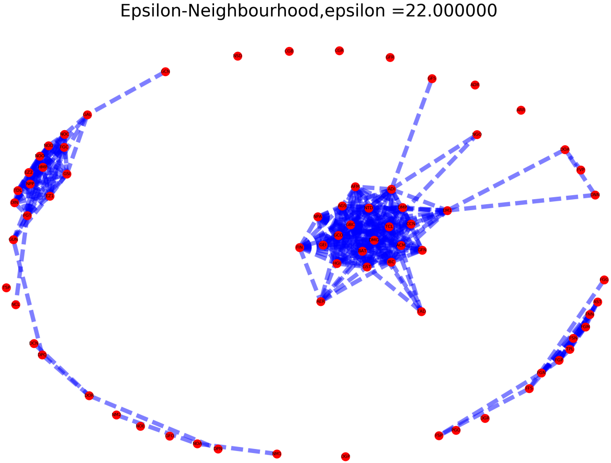
Epsilon-Neighbourhood,epsilon =8.000000

Graphs with ϵ-neighborhood Longobardi data

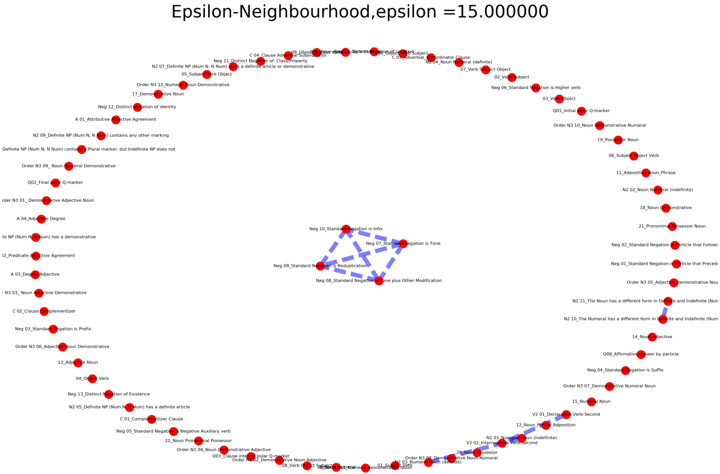
Epsilon-Neighbourhood,epsilon =15.000000



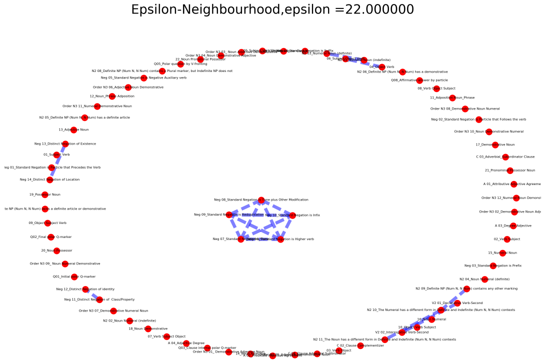
Graphs with ϵ-neighborhood Longobardi data



Graphs with ϵ-neighborhood SSWL data



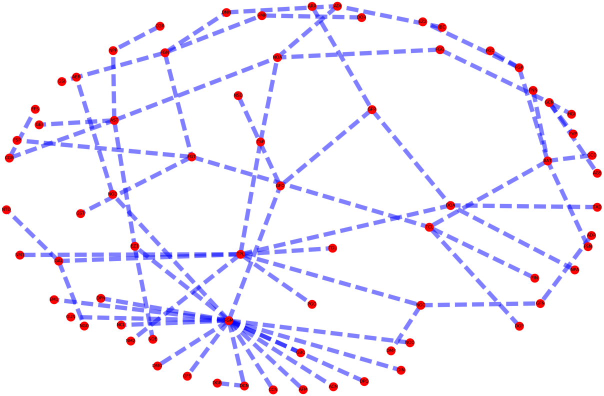
Graphs with ϵ-neighborhood SSWL data


Epsilon-Neighbourhood,epsilon =22.000000
Order N3 03_ Noun Adjective Demonstrative
Order N3 04_Noun Demonstrative Adjective
22_Noun Pronominal Possessor
Q05_Polar question by V-fronting
N2 08_Definite NP (Num N, N Num) contains a Plural marker, but Indefinite NP does not
Neg 05_Standard Negation is Negative Auxiliary verb
Order N3 06_Adjective Noun Demonstrative
12_Noun_Phrase Adposition
Order N3 11_Numeral Demonstrative Noun
N2 05_Definite NP (Num N, N Num) has a definite article
13_Adjective Noun
Neg 13_Distinct Negation of Existence
01_Subject Verb
leg 01_Standard Negation is Particle that Precedes the Verb
Neg 14_Distinct Negation of Location
19_Possessor Noun
te NP (Num N, N Num) lacks a definite article or demonstrative
09_Object Subject Verb
Q02_Final polar Q-marker
20_Noun Possessor
Order N3 09_ Noun Numeral Demonstrative
Q01_Initial polar Q-marker
Neg 12_Distinct Negation of identity
Neg 11_Distinct Negation of Class/Property
Order N3 07_Demonstrative Numeral Noun
N2 02_Noun Numeral (indefinite)
18_Noun Demonstrative
07_Verb Subject Object
A 04_Adjective Degree
Q03_Clause internal polar Q-marker
Order N3 01_ Demonstrative Adjective Noun
Neg 04_Standard Negation is Suffix
N2 03_Numeral Noun (definite)
06_Subject Object Verb
N2 01_Numeral Noun (indefinite)
04_Object Verb
N2 06_Definite NP (Num N, N Num) has a demonstrative
Q08_Affirmative answer by particle
08_Verb Object Subject
11_Adposition Noun_Phrase
Order N3 08_Demonstrative Noun Numeral
Neg 02_Standard Negation is Particle that Follows the verb
Order N3 10_Noun Demonstrative Numeral
17_Demonstrative Noun
C 03_Adverbial_Subordinator Clause
21_Pronominal Possessor Noun
A 01_Attributive Adjective Agreeme
Order N3 12_Numeral Noun Demonst
Order N3 02_Demonstrative Noun Adj
A 03_Degree Adjective
02_Verb Subject
15_Numeral Noun
Neg 03_Standard Negation is Prefix
N2 04_Noun Numeral (definite)
N2 09_Definite NP (Num N, N Num) contains any other marking
V2 01_Declarative Verb-Second
N2 10_The Numeral has a different form in Definite and Indefinite (Num N, N Num) contexts
16_Noun Numeral
10_Object Verb Subject
V2 02_Interrogative Verb-Second
N2 11_The Noun has a different form in Definite and Indefinite (Num N, N Num) contexts
C 02_Clause Complementizer
03_Verb Object
C 01_Clause Adverbial Subordinator
Neg 08_Standard Negation is Tone plus Other Modification
Neg 09_Standard Negation is Reduplication
Neg 10_Standard Negation is Infix
Neg 07_Standard Negation is Tone
Neg 06_Standard Negation is Higher verb

The ϵ-neighborhood construction is better suited to gain connectivity information in the Longobardi data: the SSWL data remain highly disconnected (only small local structures)

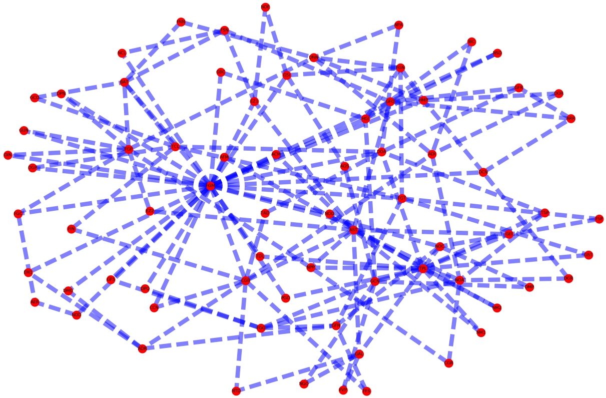
Graphs with *n*-neighborhood Longobardi data

Nearest 1 Connections



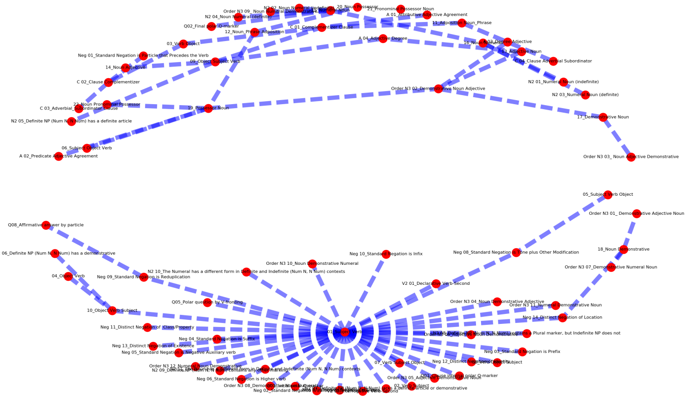
Graphs with *n*-neighborhood Longobardi data

Nearest 2 Connections



Graphs with *n*-neighborhood SSWL data

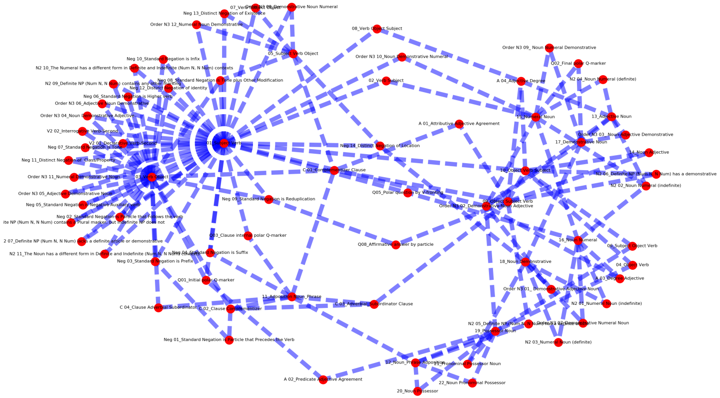
Nearest 1 Connections



N2 02_Noun Numeral (indefinite)
20_Noun Possessor
Order N3 09_ Noun Numeral Demonstrative
15_Numeral Noun
21_Pronominal Possessor Noun
N2 04_Noun Numeral (definite)
A 01_Attributive Adjective Agreement
Q02_Final polar Q-marker
C 01_Complementizer Clause
11_Adposition Noun_Phrase
12_Noun_Phrase Adposition
A 04_Adjective Degree
03_Verb Object
16_Noun Numeral
A 03_Degree Adjective
13_Adjective Noun
Neg 01_Standard Negation is Particle that Precedes the Verb
09_Object Subject Verb
C 04_Clause Adverbial Subordinator
14_Noun Adjective
C 02_Clause Complementizer
N2 01_Numeral Noun (indefinite)
Order N3 02_Demonstrative Noun Adjective
N2 03_Numeral Noun (definite)
22_Noun Pronominal Possessor
C 03_Adverbial_Subordinator Clause
19_Possessor Noun
17_Demonstrative Noun
N2 05_Definite NP (Num N, N Num) has a definite article
06_Subject Object Verb
A 02_Predicate Adjective Agreement
Order N3 03_ Noun Adjective Demonstrative
05_Subject Verb Object
Order N3 01_ Demonstrative Adjective Noun
Q08_Affirmative answer by particle
18_Noun Demonstrative
N2 06_Definite NP (Num N, N Num) has a demonstrative
Neg 10_Standard Negation is Infix
Neg 08_Standard Negation is Tone plus Other Modification
Order N3 10_Noun Demonstrative Numeral
Order N3 07_Demonstrative Numeral Noun
04_Object Verb
N2 10_The Numeral has a different form in Definite and Indefinite (Num N, N Num) contexts
Neg 09_Standard Negation is Reduplication
V2 01_Declarative Verb-Second
Q05_Polar question by V-fronting
Order N3 04_Noun Demonstrative Adjective
Order N3 11_Numeral Demonstrative Noun
10_Object Verb Subject
Neg 14_Distinct Negation of Location
Neg 11_Distinct Negation of Class/Property
01_Subject Verb
N2 08_Definite NP (Num N, N Num) contains a Plural marker, but Indefinite NP does not
Order N3 06_Adjective Noun Demonstrative
Neg 04_Standard Negation is Suffix
Neg 13_Distinct Negation of Existence
Neg 03_Standard Negation is Prefix
Neg 05_Standard Negation is Negative Auxiliary verb
07_Verb Subject Object
Neg 12_Distinct Negation of Identity
08_Verb Object Subject
Order N3 12_Numeral Noun Demonstrative
N2 11_The Noun has a different form in Definite and Indefinite (Num N, N Num) contexts
N2 09_Definite NP (Num N, N Num) contains any other marking
Q03_Clause internal polar Q-marker
Neg 06_Standard Negation is Higher verb
Order N3 05_Adjective Demonstrative Noun
02_Verb Subject
Order N3 08_Demonstrative Noun Numeral
N2 07_Definite NP (Num N, N Num) lacks a definite article or demonstrative
Neg 02_Standard Negation
V2 02_Interrogative Verb-Second

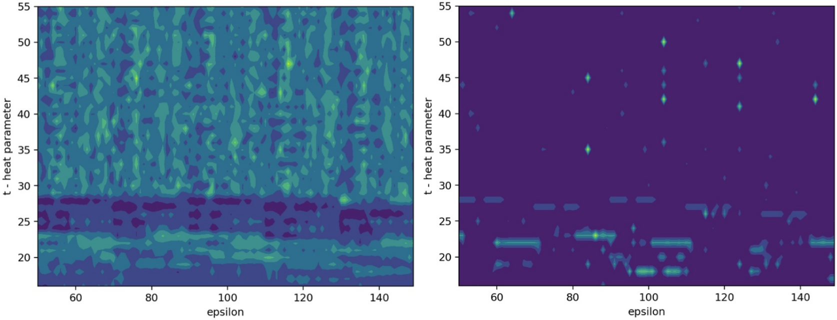
Graphs with *n*-neighborhood SSWL data

Nearest 2 Connections


07_Verb Subject Object
Order N3 08_Demonstrative Noun Numeral
Neg 13_Distinct Negation of Existence
Order N3 12_Numeral Noun Demonstrative
08_Verb Object Subject
05_Subject Verb Object
Order N3 09_ Noun Numeral Demonstrative
Neg 10_Standard Negation is Infix
Order N3 10_Noun Demonstrative Numeral
Q02_Final polar Q-marker
N2 10_The Numeral has a different form in Definite and Indefinite (Num N, N Num) contexts
Neg 08_Standard Negation is Tone plus Other Modification
02_Verb Subject
A 04_Adjective Degree
N2 04_Noun Numeral (definite)
N2 09_Definite NP (Num N, N Num) contains any other marking
Neg 12_Distinct Negation of identity
Neg 06_Standard Negation is Higher verb
Order N3 06_Adjective Noun Demonstrative
Order N3 04_Noun Demonstrative Adjective
15_Numeral Noun
13_Adjective Noun
A 01_Attributive Adjective Agreement
V2 02_Interrogative Verb-Second
Order N3 03_ Noun Adjective Demonstrative
17_Demonstrative Noun
V2 01_Declarative Verb-Second
01_Subject Verb
Neg 07_Standard Negation is Tone
Neg 14_Distinct Negation of Location
14_Noun Adjective
Neg 11_Distinct Negation of Class/Property
C 01_Complementizer Clause
10_Object Verb Subject
N2 06_Definite NP (Num N, N Num) has a demonstrative
Order N3 11_Numeral Demonstrative Noun
03_Verb Object
N2 02_Noun Numeral (indefinite)
Q05_Polar question by V-fronting
Order N3 05_Adjective Demonstrative Noun
09_Object Subject Verb
Order N3 02_Demonstrative Noun Adjective
Neg 09_Standard Negation is Reduplication
Neg 05_Standard Negation is Negative Auxiliary verb
Neg 02_Standard Negation is Particle that Follows the verb
ite NP (Num N, N Num) contains a Plural marker, but Indefinite NP does not
Q03_Clause internal polar Q-marker
16_Noun Numeral
2 07_Definite NP (Num N, N Num) lacks a definite article or demonstrative
Q08_Affirmative answer by particle
06_Subject Object Verb
N2 11_The Noun has a different form in Definite and Indefinite (Num N, N Num) contexts
Neg 04_Standard Negation is Suffix
Neg 03_Standard Negation is Prefix
18_Noun Demonstrative
04_Object Verb
A 03_Degree Adjective
Q01_Initial polar Q-marker
Order N3 01_ Demonstrative Adjective Noun
11_Adposition Noun_Phrase
C 03_Adverbial_Subordinator Clause
N2 01_Numeral Noun (indefinite)
C 04_Clause Adverbial Subordinator
C 02_Clause Complementizer
N2 05_Definite NP (Num N, N Num) has a definite article
Order N3 07_Demonstrative Numeral Noun
19_Possessor Noun
N2 03_Numeral Noun (definite)
Neg 01_Standard Negation is Particle that Precedes the Verb
12_Noun_Phrase Adposition
21_Pronominal Possessor Noun
A 02_Predicate Adjective Agreement
22_Noun Pronominal Possessor
20_Noun Possessor

Regions of ϵ-t space

- Graphs depend on ϵ-neighborhood and on t-heat kernel variable
- explore ϵ-t space: determine regions where high variance in distribution of each parameter under the heat kernel mapping
- high variance in a parameter suggests it is highly independent (similar to PCA method)
- contour plots of variance; plots of number of outliers produced in set of coordinates for a given parameter


55
50
45
40
35
30
25
20
t - heat parameter
60
80
100
120
140
epsilon
55
50
45
40
35
30
25
20
t - heat parameter
60
80
100
120
140
epsilon

Further Questions

- an in depth linguistic analysis of the meaning of these clustering structures is still needed (ongoing work)

- comparison of the heat kernel technique with other dimensional reduction techniques (PCA etc.)

- more detailed discussion of different regions of the ϵ-t space in the heat kernel model (specific parameters with high independence measure)

- manifold $\mathcal{M}$ reconstruction? Belkin-Niyogi results

Topological Structures of Syntactic Parameters

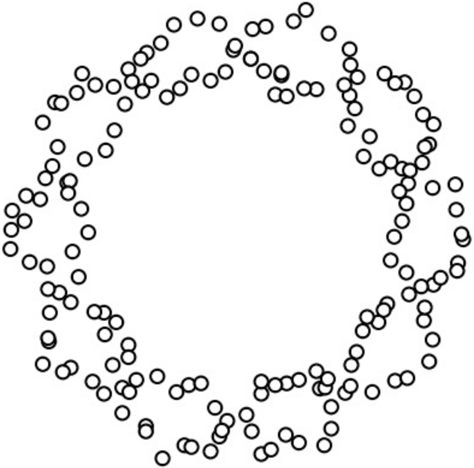
persistent homology (ongoing work with Alex Port)

- previous work computing persistent homology of SSWL data

Alexander Port, Iulia Gheorghita, Daniel Guth, John M.Clark, Crystal Liang, Shival Dasu, Matilde Marcolli, *Persistent Topology of Syntax*, arXiv:1507.05134

- ongoing work: persistent homology in the new Longobardi data
- main questions:
  - persistent generators of H_0 and phylogenetic trees?
  - meaning of persistent generators of H_1?

Persistent Topology of Data Sets

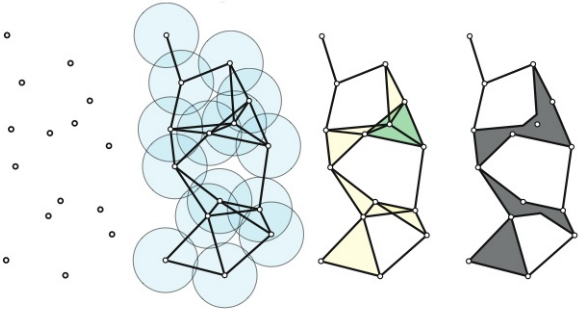


how data cluster around topological shapes at different scales

Vietoris–Rips complexes

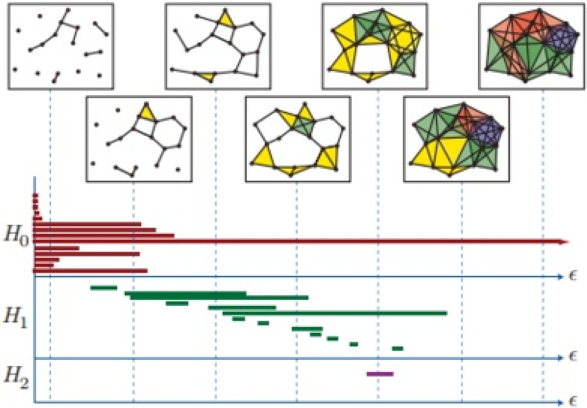
- set $X = \{x_\alpha\}$ of points in Euclidean space $\mathbb{E}^N$, distance $d(x,y) = \|x-y\| = (\sum_{j=1}^N (x_j - y_j)^2)^{1/2}$
- Vietoris-Rips complex $R(X,\epsilon)$ of scale ϵ over field $\mathbb{K}$:

$R_n(X,\epsilon)$ is $\mathbb{K}$-vector space spanned by all unordered $(n+1)$-tuples of points $\{x_{\alpha_0}, x_{\alpha_1}, \dots, x_{\alpha_n}\}$ in X where all pairs have distances $d(x_{\alpha_i}, x_{\alpha_j}) \leq \epsilon$



(image by Jeff Erickson)

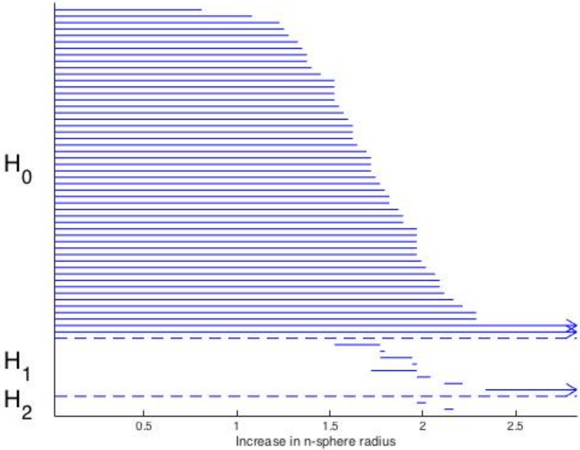
- inclusion maps $R(X,\epsilon_1) \hookrightarrow R(X,\epsilon_2)$ for $\epsilon_1 < \epsilon_2$ induce maps in homology by functoriality $H_n(X,\epsilon_1) \to H_n(X,\epsilon_2)$



(image by `forty.to`)

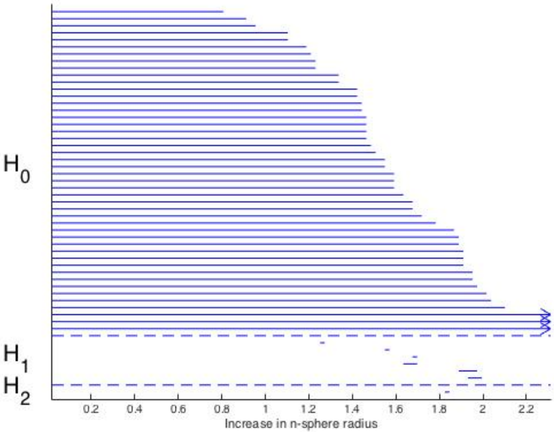
barcode diagrams: births and deaths of persistent generators

Persistent Topology of Indo-European Languages (SSWL data)



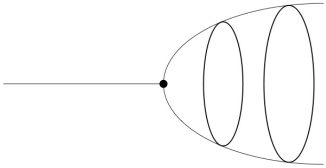
- Two persistent generators of H_0 (Indo-Iranian, European)
- One persistent generator of H_1

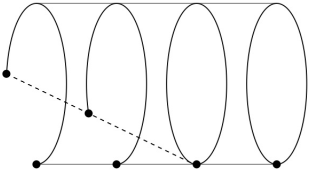
Persistent Topology of Niger–Congo Languages (SSWL data)



- Three persistent components of H_0 (Mande, Atlantic-Congo, Kordofanian)
- No persistent H_1

Sources of Persistent H_1

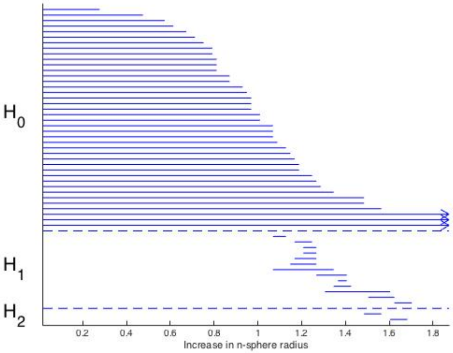




- "Hopf bifurcation" type phenomenon
- two different branches of a tree closing up in a loop

two different types of phenomena of historical linguistic development within a language family

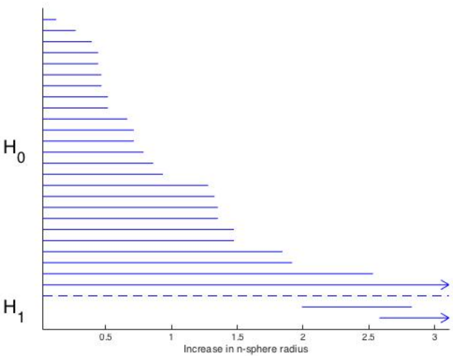
What is the Indo-European H_1?



Persistent topology with Hellenic (and Indo-Iranic) branch removed

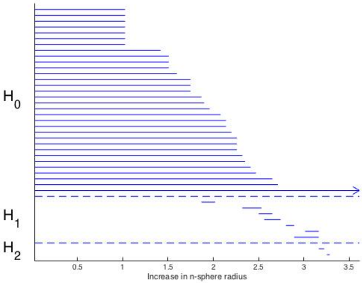
- related to influences (at the syntactic level) of the Hellenic branch on languages in other branches (like some Slavic languages)
- consistent with some previous independent linguistic observations by Longobardi
- what about the new Longobardi data analyzed topologically?

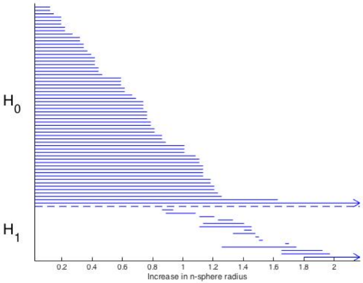
Topological Analysis of New Syntactic Data
(ongoing with Alex Port)



Longobardi data (2016): persistent generator of H_1 still present

Evidence of further structures at different scales





Linguistic interpretation: behavior of H_0 versus phylogenetic trees; interpretation of generators of H_1?

Longer Term Goals

- import a set of different mathematical techniques (phylogenetic algebraic geometry, persistent topology, coding theory, statistical mechanics, geometric models of associative memory) in order to *study natural languages as dynamical objects*

- create mathematical and computational models of

1. how languages are acquired?
2. how languages are stored in the brain?
3. how languages change and evolve dynamically in time?

for human languages viewed at the level of their syntactic structures