The Geometry of Syntax

Matilde Marcolli
U. Toronto, Perimeter Institute, Caltech

Utrecht, April 2018
Mathematics Colloquium and Linguistics and AI Seminar

A Mathematical Physicist's adventures in Linguistics

- This talk is in two parts:
  - Today Math Colloquium: persistent topology of syntax, algebro-geometric historical linguistics, spin glass models of language evolution
  - Continuation Linguistics & AI Seminar: relations between syntactic parameters via Kanerva networks, coding theory, and via Belkin–Niyogi heat kernel dimensional reduction

References for this part:

- Alexander Port, Iulia Gheorghita, Daniel Guth, John M.Clark, Crystal Liang, Shival Dasu, Matilde Marcolli, *Persistent Topology of Syntax*, Mathematics in Computer Science, 12 (2018) no. 1, 33–50.
- Karthik Siva, Jim Tao, Matilde Marcolli, *Syntactic Parameters and Spin Glass Models of Language Change*, Linguistic Analysis, Vol. 41 (2017) N. 3-4, 559–608.
- Kevin Shu, Sharjeel Aziz, Vy-Luan Huynh, David Warrick, Matilde Marcolli, *Syntactic Phylogenetic Trees*, in "Foundations of Mathematics and Physics One Century After Hilbert" (J. Kouneiher, Ed.) Springer, 2018.
- 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.

What is Linguistics?

• *Linguistics* is the scientific study of language

- What is Language? (langage, lenguaje, ...)
- What is a Language? (lange, lengua,...)

Similar to 'What is Life?' or 'What is an organism?' in biology

• *natural* language
as opposed to artificial (formal, programming, ...) languages

• The point of view we will focus on:
Language is a kind of *Structure*

- It can be approached mathematically and computationally, like many other kinds of structures
- The main purpose of mathematics is the understanding of structures

• How are different languages related?
What does it mean that they come in *families*?

• How do languages evolve in time?
Phylogenetics, Historical Linguistics, Etymology

• How does the process of language acquisition work?
(Neuroscience)

• Semiotic viewpoint (mathematical theory of communication)

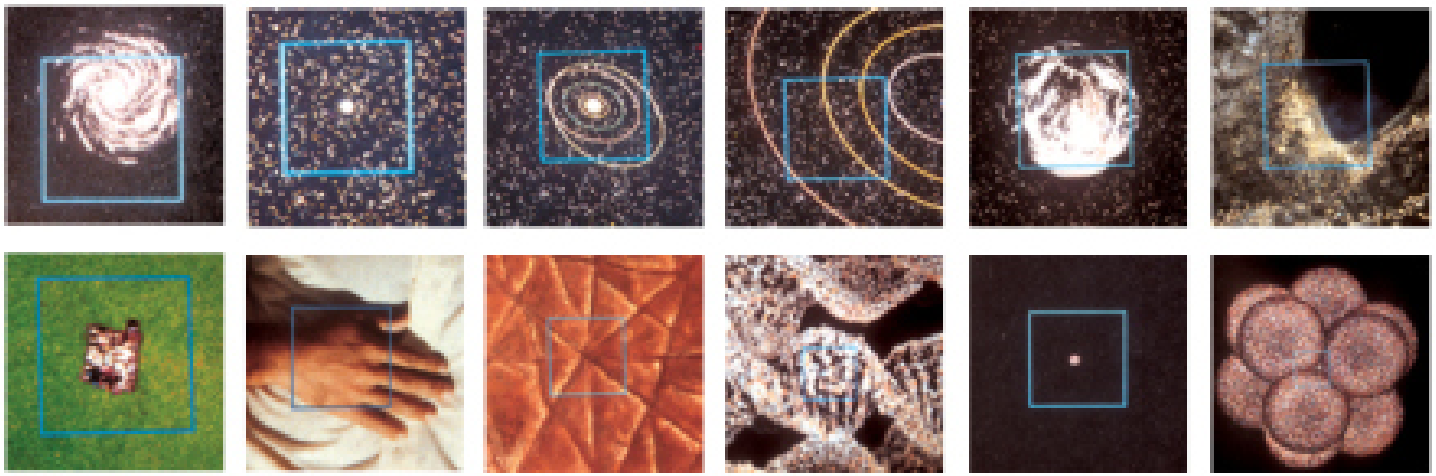
• Discrete versus Continuum
(probabilistic methods, versus discrete structures)

• Descriptive or Predictive?
to be predictive, a science needs good mathematical models

A language exists at many different levels of structure

An Analogy: Physics looks very different at different *scales*:

- General Relativity and Cosmology ($\geq 10^{10}$ m)
- Classical Physics (~ 1 m)
- Quantum Physics ($\leq 10^{-10}$ m)
- Quantum Gravity (10^{-35} m)



Despite dreams of a Unified Theory, we deal with different mathematical models for different levels of structure

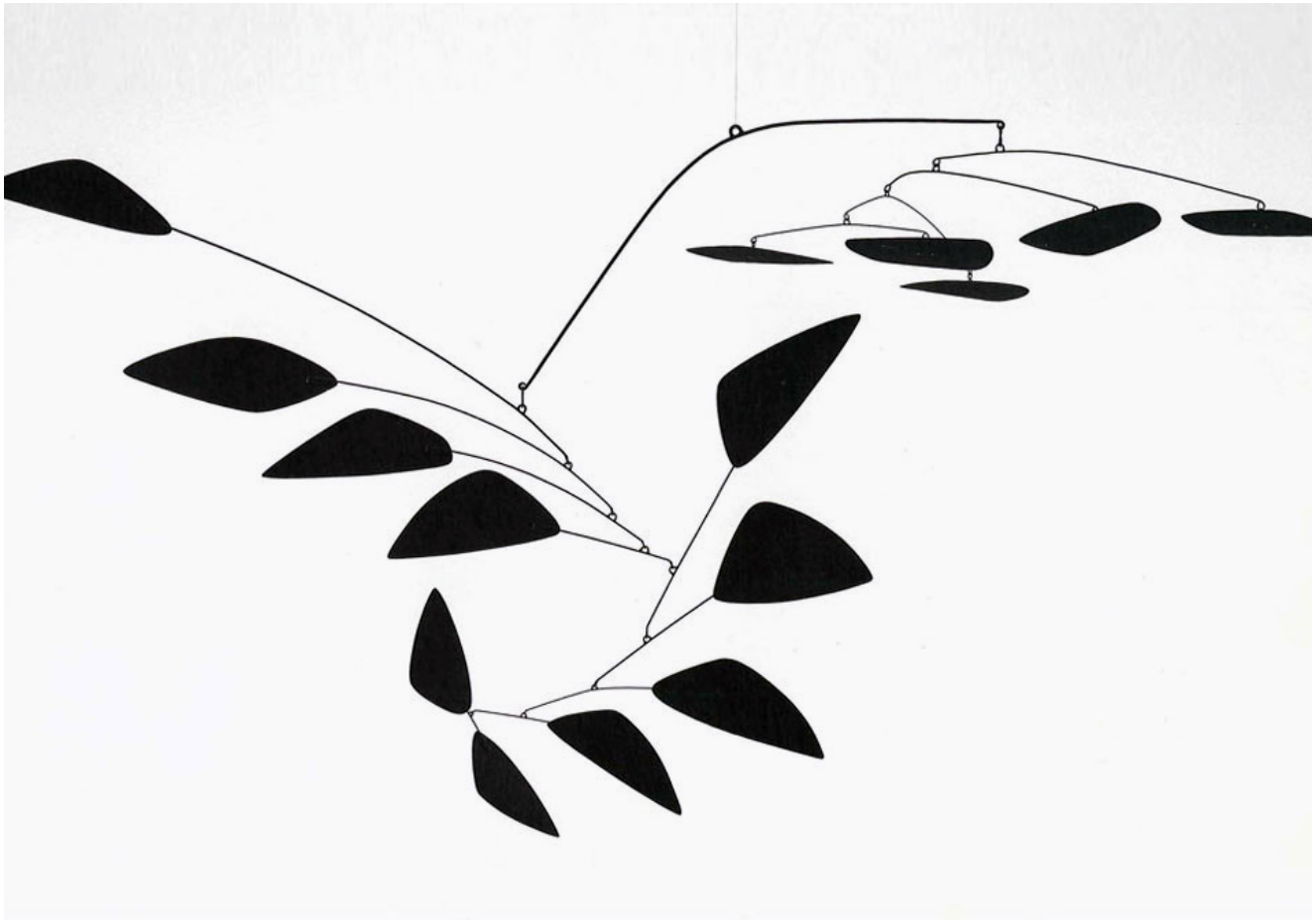
Similarly, we view language at different "scales":

- units of sound (phonology)
- words (morphology)
- sentences (syntax)
- global meaning (semantics)

We expect to be dealing with different mathematical structures and different models at these various different levels

Main level I will focus on: Syntax

Linguistics view of syntax kind of looks like this...



Alexander Calder, *Mobile*, 1960

Modern Syntactic Theory:

• grammaticality: judgement on whether a sentence is well formed (grammatical) in a given language, i-language gives people the capacity to decide on grammaticality

• generative grammar: produce a set of rules that correctly predict grammaticality of sentences

• universal grammar: ability to learn grammar is built in the human brain, e.g. properties like distinction between nouns and verbs are universal ... is universal grammar a falsifiable theory?

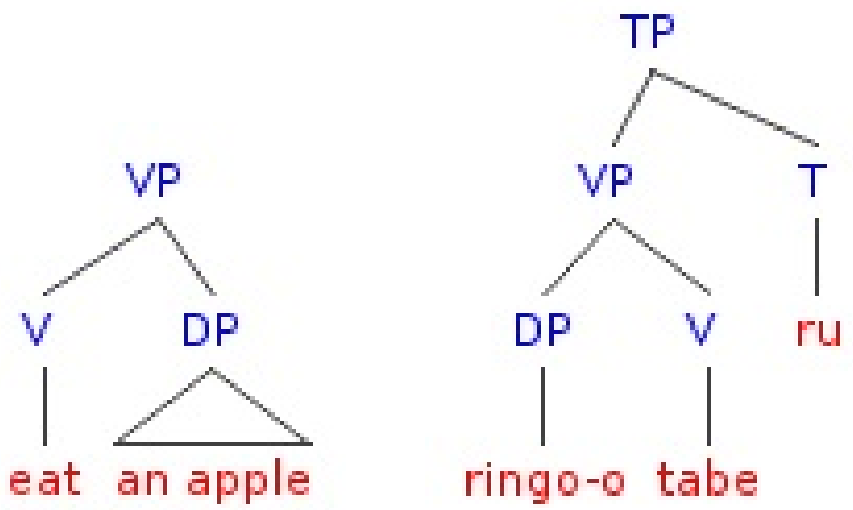
Principles and Parameters (Government and Binding)

(Chomsky, 1981)

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

(head-initial versus head-final)

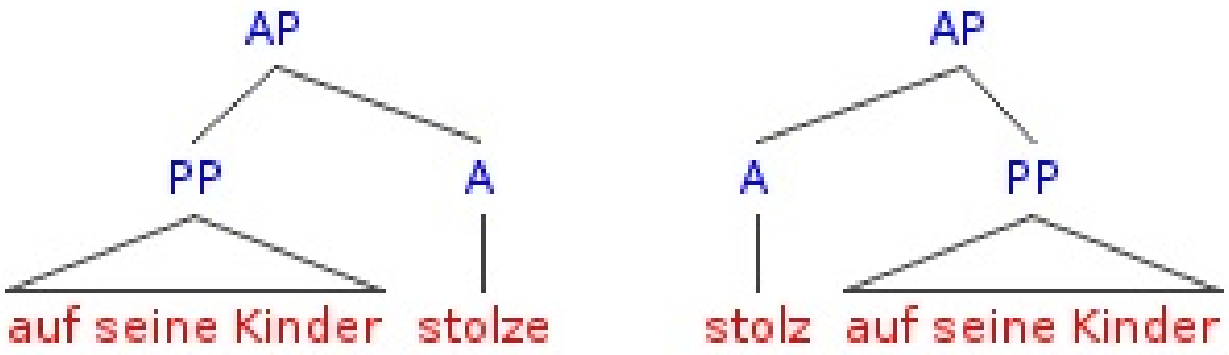
English is head-initial, Japanese is head-final



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

...but not always so clear-cut: German can use both structures

auf seine Kinder stolze Vater (head-final) or
er ist stolz auf seine Kinder (head-initial)



AP= adjective phrase, PP= prepositional phrase

• Corpora based statistical analysis of head-directionality (Haitao Liu, 2010): a continuum between head-initial and head-final

Examples of Parameters

- *Head-directionality*
- *Subject-side*
- *Pro-drop*
- *Null-subject*

Problems

- Interdependencies between parameters
- Diachronic changes of parameters in language evolution

Dependent parameters

- null-subject parameter: can drop subject

Example: among Latin languages, Italian and Spanish have null-subject (+), French does not (-)

it rains, piove, llueve, il pleut

- pro-drop parameter: can drop pronouns in sentences

不知道。喜欢吗?

Bù zhīdào. Xǐhuan ma? "*I* don't know. Do *you* like it?"

- Pro-drop controls Null-subject

How many independent parameters?

Geometry of the space of syntactic parameters?

Persistent Topology of Syntax

• Alexander Port, Iulia Gheorghita, Daniel Guth, John M.Clark, Crystal Liang, Shival Dasu, Matilde Marcolli, *Persistent Topology of Syntax*, Mathematics in Computer Science, 12 (2018) no. 1, 33–50

Databases of Syntactic Parameters of World Languages:

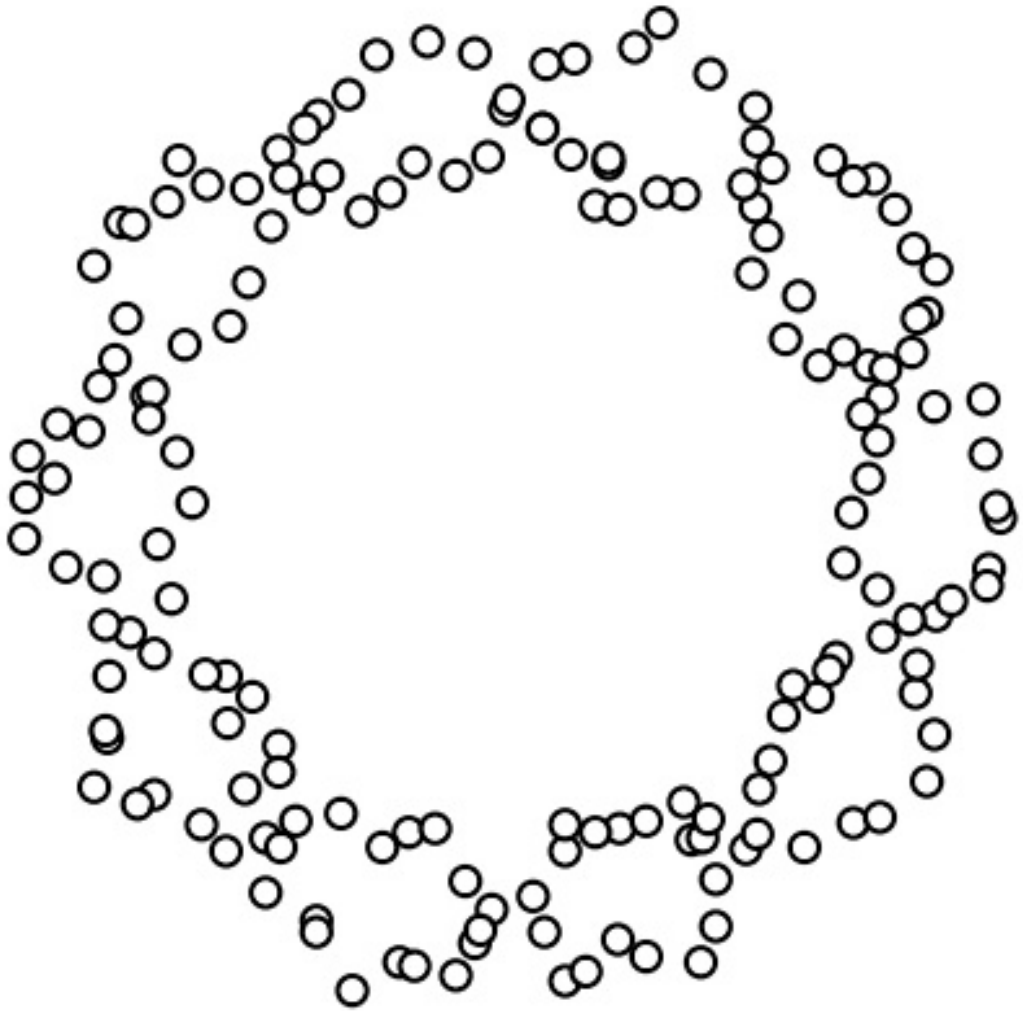
1. Syntactic Structures of World Languages (SSWL) `http://sswl.railsplayground.net/`
2. TerraLing `http://www.terraling.com/`

• currently 116 binary syntactic parameters and 253 world languages across several different language families

• *problem*: non-uniformly mapped across languages, need to deal with "missing data"

- replace missing binary data by 0.5 value
- establish a variable threshold: percentage of parameters completely mapped; select only languages above threshold and check range of threshold value for which results are stable

Persistent Topology of Data Sets

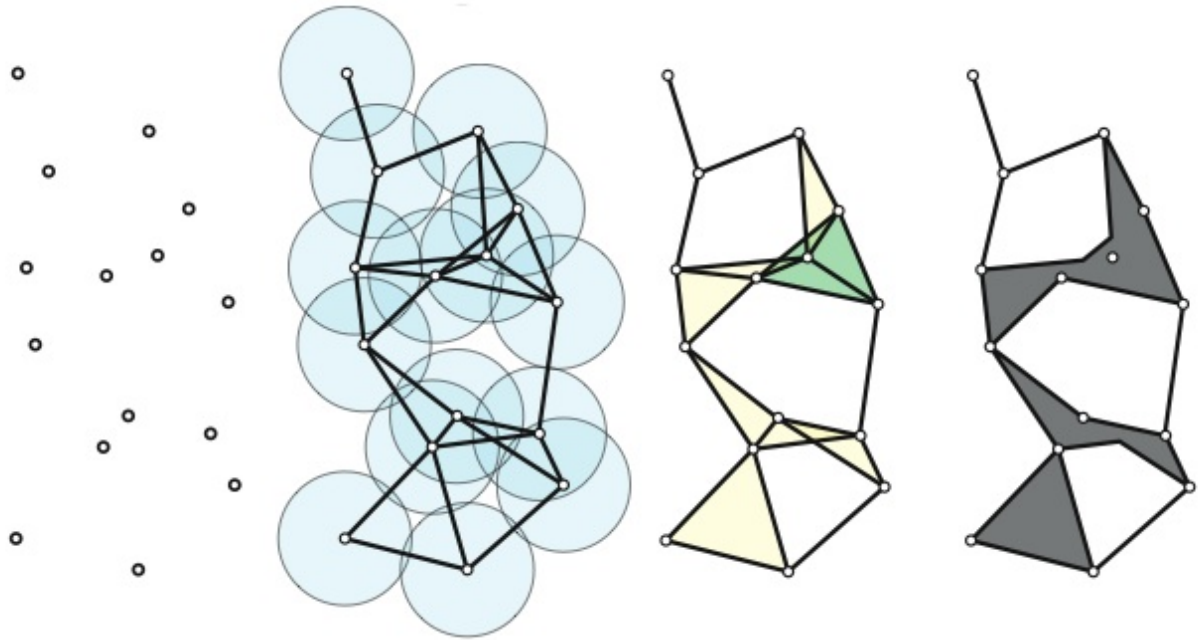


how data cluster around topological shapes at different scales

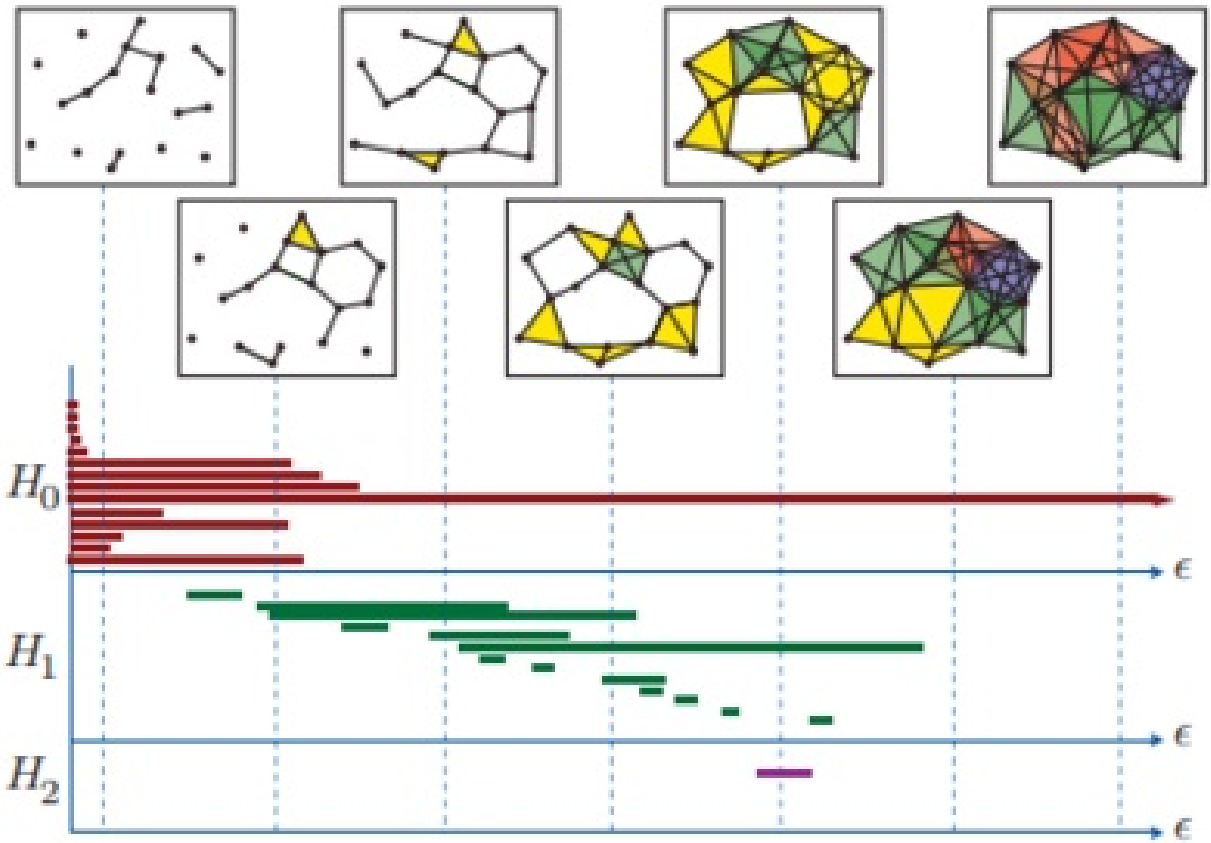
Vietoris–Rips complexes

- 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 $\{x_{\alpha_0}, x_{\alpha_1}, \ldots, x_{\alpha_n}\}$ in X where all pairs have distances $d(x_{\alpha_i}, x_{\alpha_j}) \leq \epsilon$



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

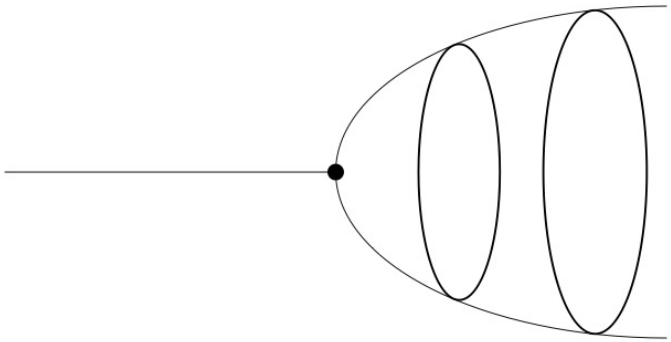


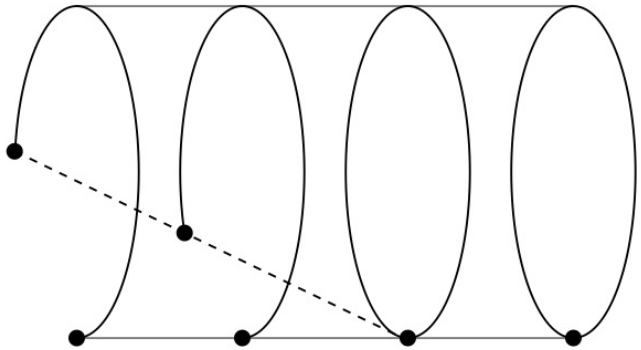
barcode diagrams: births and deaths of persistent generators

Persistent Topology of Syntactic Parameters

- Data: 253 languages from SSWL with 116 parameters
- if consider all world languages together too much noise in the persistent topology: subdivide by language families
- Principal Component Analysis: reduce dimensionality of data
- *Related Question*: what is the linguistic meaning of the principal components? (some admixture of different syntactic parameters)
- compute Vietoris–Rips complex and barcode diagrams
  - Persistent H_0: clustering of data in components – language subfamilies
  - Persistent H_1: clustering of data along closed curves (circles) – linguistic meaning?

Sources of Persistent H_1

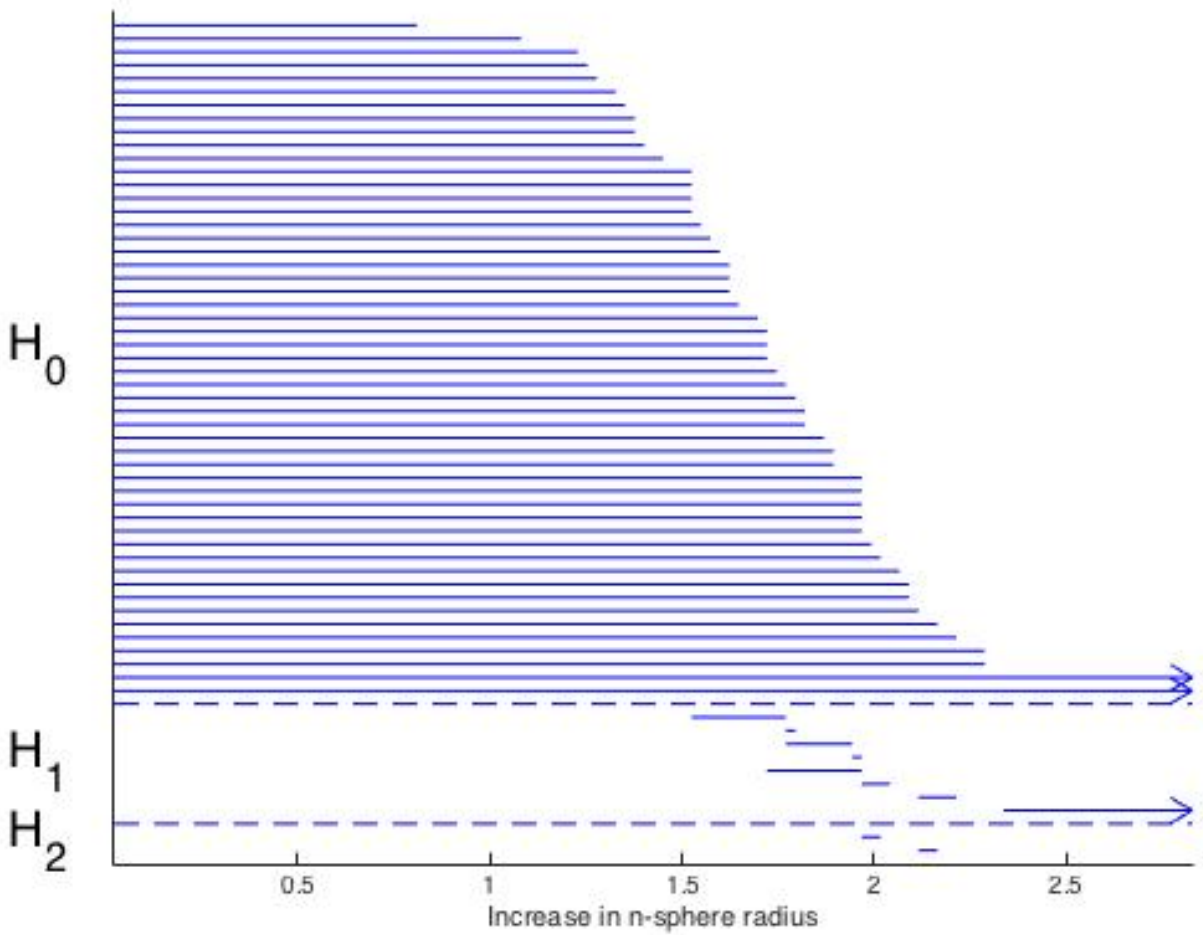




- "Hopf bifurcation" type phenomenon
- two different branches of a tree closing up in a loop

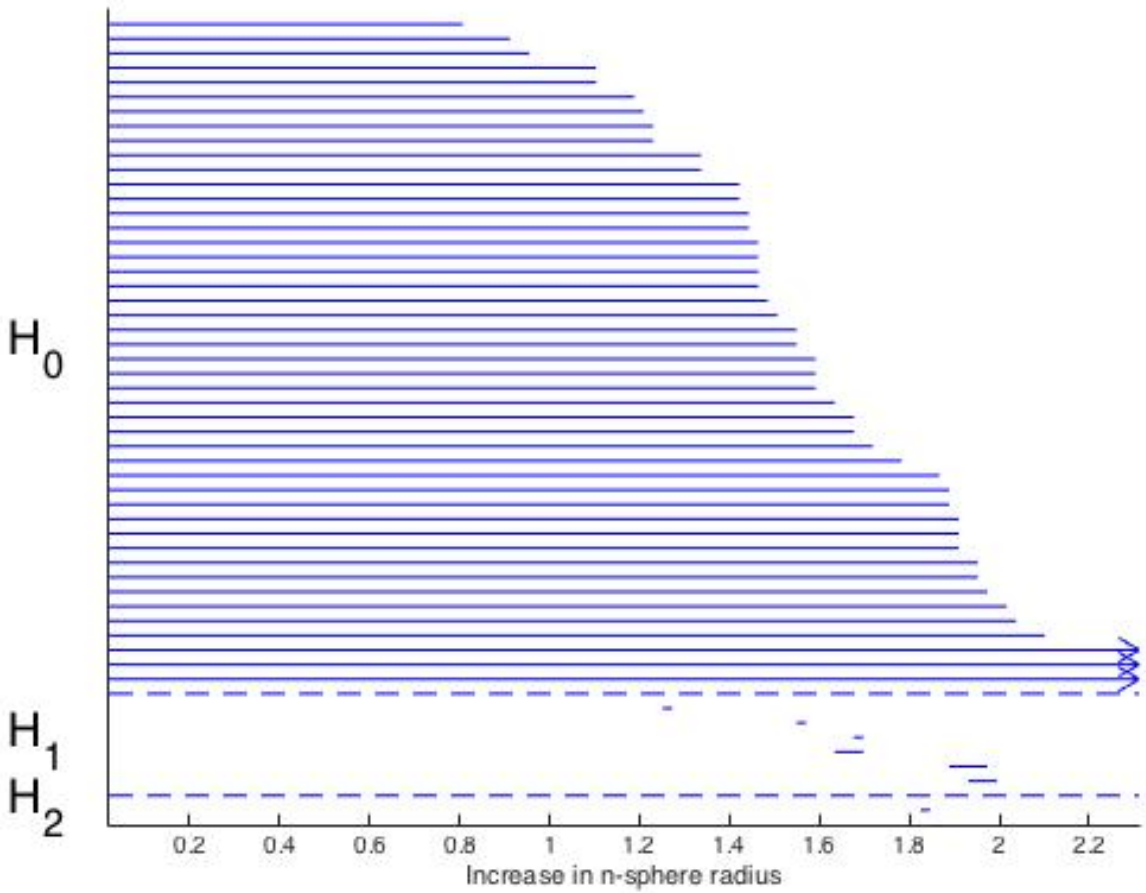
two different types of phenomena of historical linguistic development within a language family

Persistent Topology of Indo-European Languages



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

Persistent Topology of Niger–Congo Languages

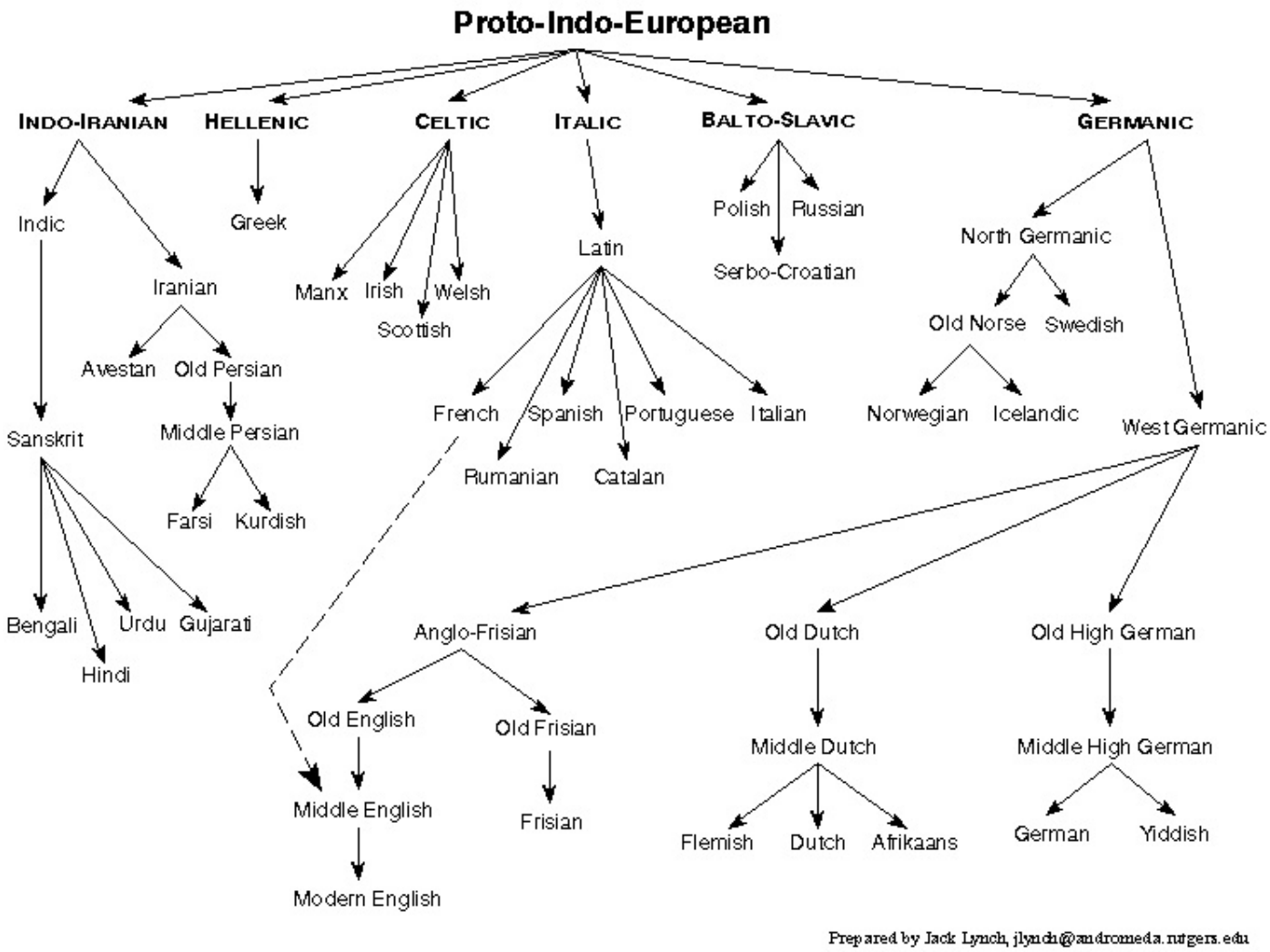


- Three persistent components of H_0

(Mande, Atlantic-Congo, Kordofanian)

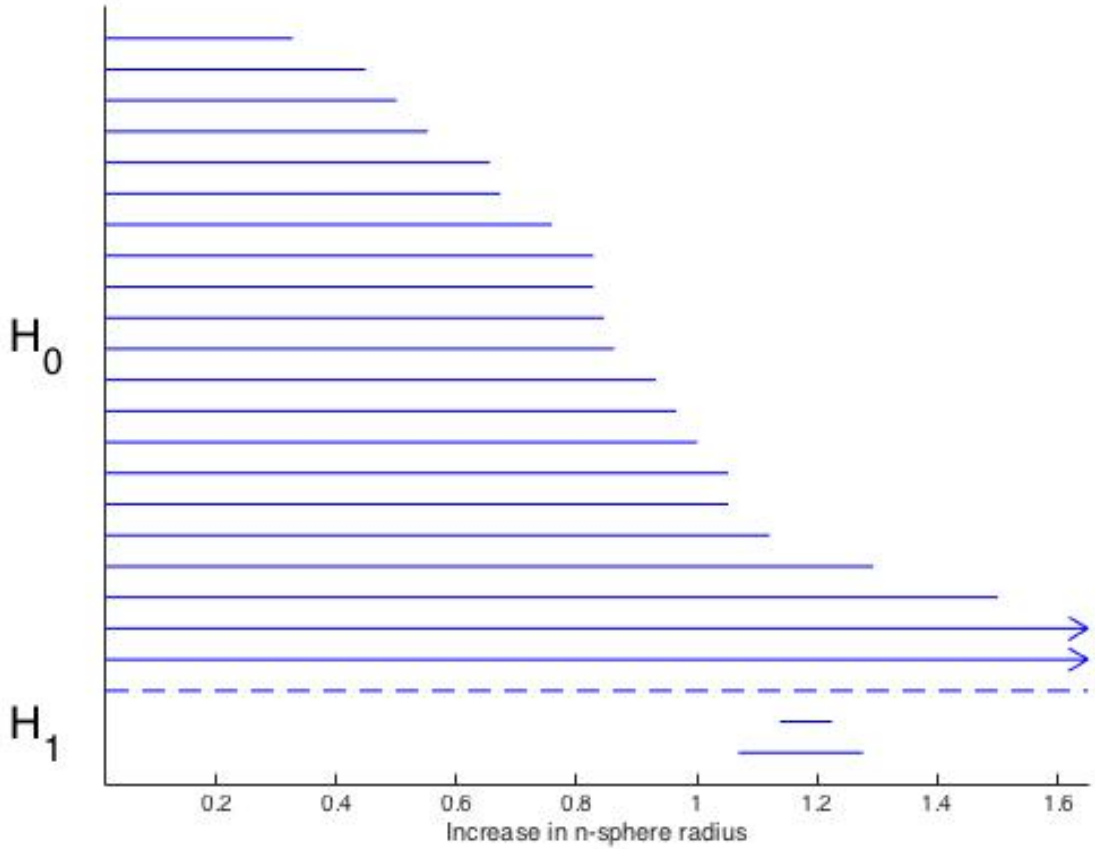
- No persistent H_1

The origin of persistent H_1 of Indo-European Languages?



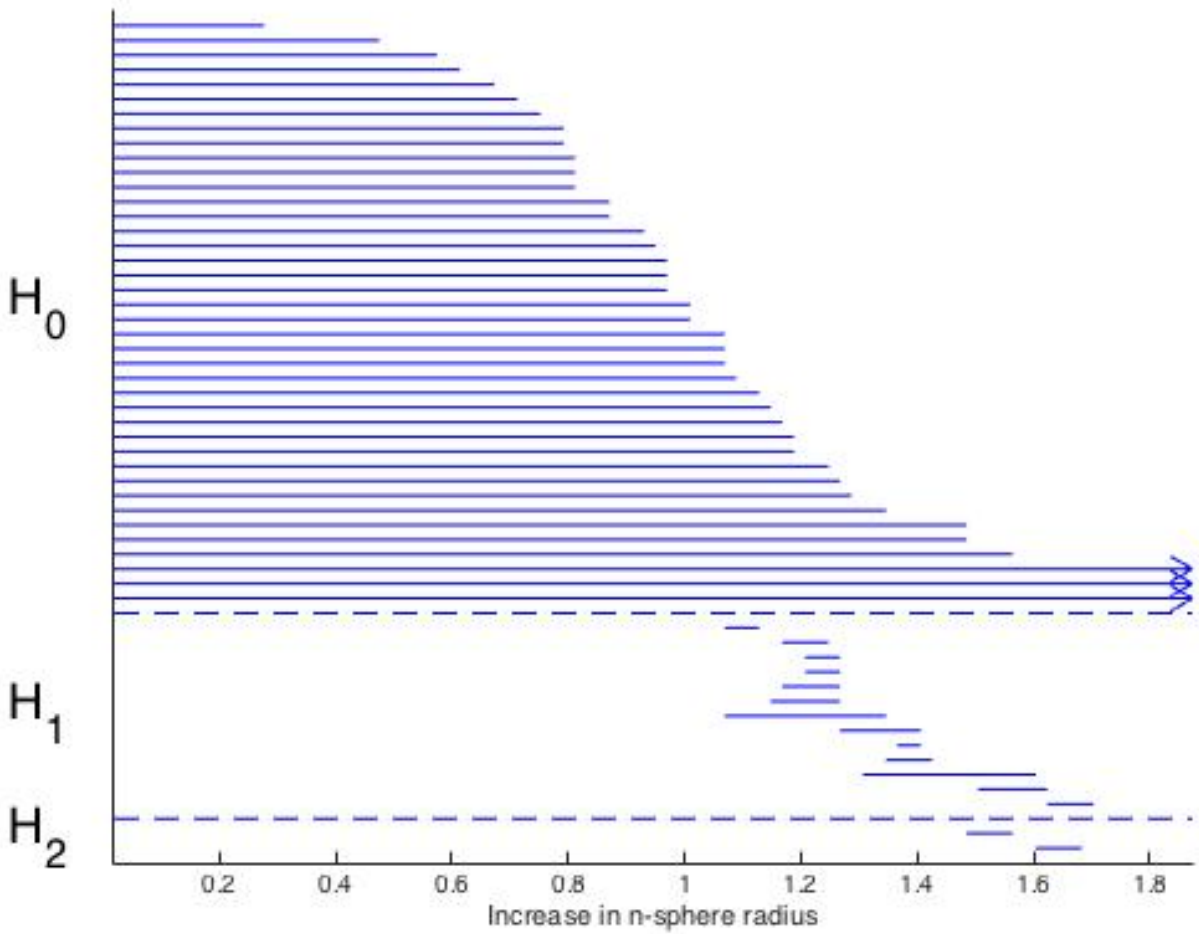
Naive guess: the Anglo-Norman bridge ... but lexical not syntactic!

Answer: No, it is not the Anglo-Norman bridge!



Persistent topology of the Germanic+Latin languages

Answer: It's all because of Ancient Greek!



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

So, what does topology tell us?

• H_1 of Indo-European languages related to influences (at the syntactic level) of the Hellenic branch on some Slavic languages (consistent with independent observations in new data by Longobardi, not analyzed yet topologically)

• Topology captures known historical-linguistics phenomena (clustering of syntactic structures by language families and sub-families)

• the barcode diagram for H_0 (persistent connected components) gives a splitting of a language family into finer and finer subfamilies: comparison with *phylogenetic trees* of historical linguistics!

• it is sensitive to more subtle phenomena, which are not seen in "phylogenetic trees" of languages: influences across different language sub-families (H_1 persistent generators)

• it can provide additional useful information on understanding how language (at the syntactic level) evolves

Syntactic Parameters as Dynamical Variables

- Example: Word Order: SOV, SVO, VSO, VOS, OVS, OSV

Word Orders	Percentage		
SOV	41.03%	Subject-initial	Specifier-Head
SVO	35.44%		
VSO	6.90%	Subject-medial	Head-Specifier
VOS	1.82%	Subject-final	
OVS	0.79%		
OSV	0.29%	Subject-medial	Specifier-Head

Very uneven distribution across world languages

• Word order distribution: a neuroscience explanation?

- D. Kemmerer, *The cross-linguistic prevalence of SOV and SVO word orders reflects the sequential and hierarchical representation of action in Broca's area*, Language and Linguistics Compass, 6 (2012) N.1, 50–66.

• Internal reasons for diachronic switch?

- F.Antinucci, A.Duranti, L.Gebert, *Relative clause structure, relative clause perception, and the change from SOV to SVO*, Cognition, Vol.7 (1979) N.2 145–176.

Changes over time in Word Order

• Ancient Greek: switched from Homeric to Classical
- A. Taylor, *The change from SOV to SVO in Ancient Greek*, Language Variation and Change, 6 (1994) 1–37

• Sanskrit: different word orders allowed, but prevalent one in Vedic Sanskrit is SOV (switched at least twice by influence of Dravidian languages)
- F.J. Staal, *Word Order in Sanskrit and Universal Grammar*, Springer, 1967

• English: switched from Old English (transitional between SOV and SVO) to Middle English (SVO)
- J. McLaughlin, *Old English Syntax: a handbook*, Walter de Gruyter, 1983.

Syntactic Parameters are Dynamical in Language Evolution

Two main types of questions on dynamical behavior of syntactic parameters

1. Reconstruct the past: phylogenetic trees of language families, historical linguistics
2. Predict the future: dynamical models of language change due to language interaction (bilingualism, code switching), dynamical models of language acquisition

- both questions considered with specific focus on syntax

Phylogenetic Algebraic Geometry of Languages

• 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.

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

• Usually via lexical and morphological analysis

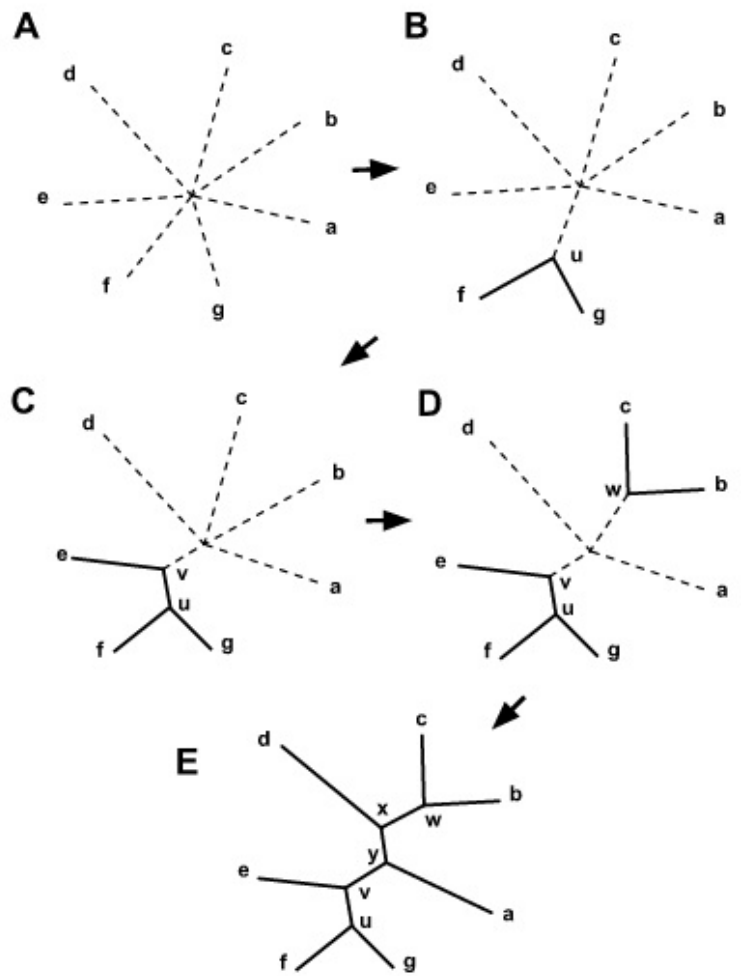
• Goal: understand the historical relatedness of different languages, subdivisions into families and sub-families, phylogenetic trees of language families

• 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?
- Can one reconstruct phylogenetic trees using syntactic parameters data? (Syntax is more stable than lexicon, slower changes, rare borrowing...)
- 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

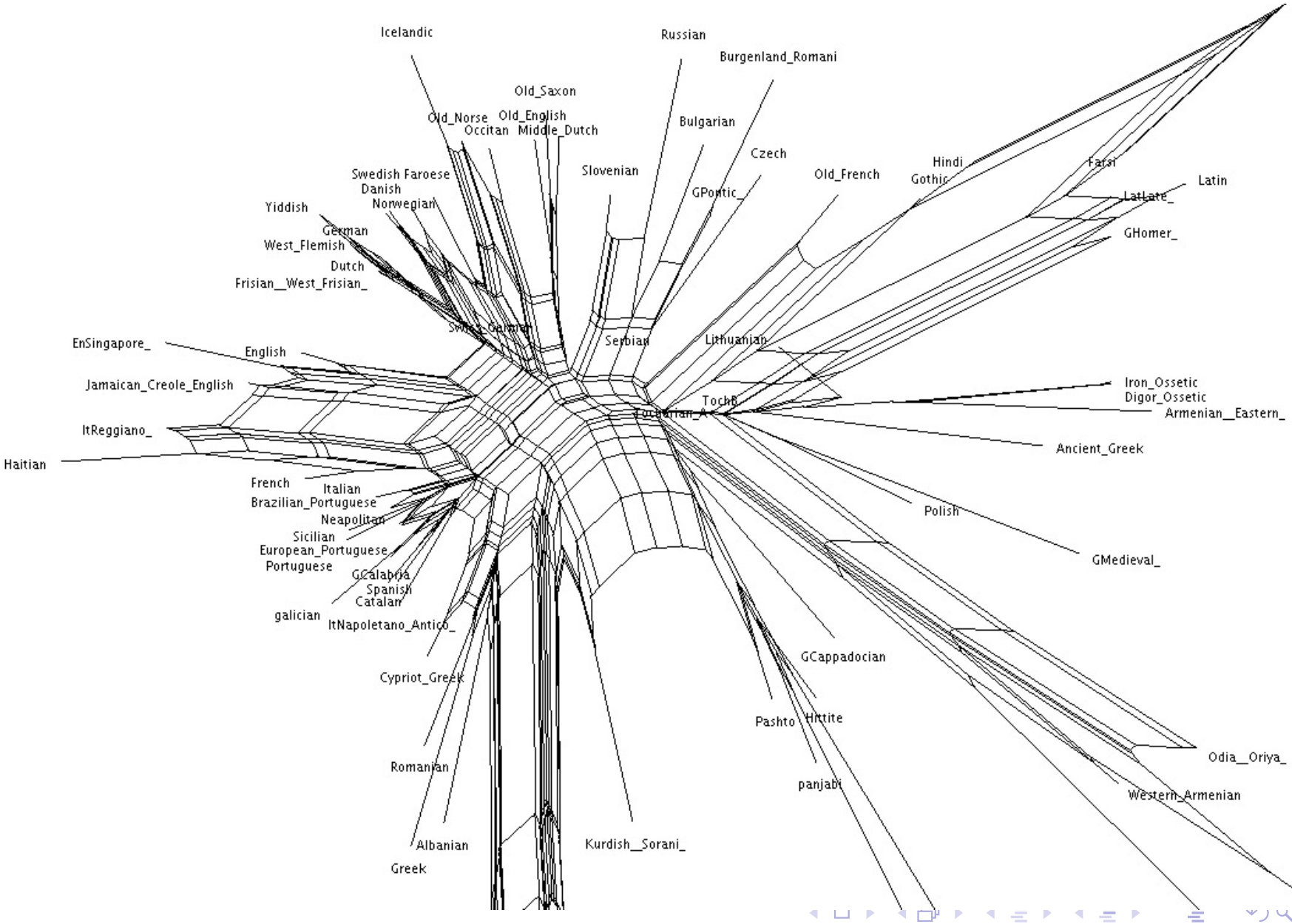


Expect problems: SSWL data and phylogenetic reconstructions

- 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, or use of H_0-method) and then use Phylogenetic Algebraic Geometry (Pachter, Sturmfels 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

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

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*

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

Main Toolbox

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

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

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 $P = (p_{i_1,\dots,i_n})$ given by data

• Euclidean distance of the point P from the variety V_T (in ambient affine space)

• Eckrat–Young formula: for a determinantal variety

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

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

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

Procedure

- set of languages $\mathcal{L} = \{\ell_1, \ldots, \ell_n\}$ (selected subfamily)
- set of SSWL 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 $\{\pi_i\}_{i=1}^N$

- 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 (how good a match P is to a Juke-Cantor model on T)
- evaluate Euclidean distance of P to V_T (or part of V_T that distinguishes candidate trees) to select max likelihood tree

Simple 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)

- 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`:

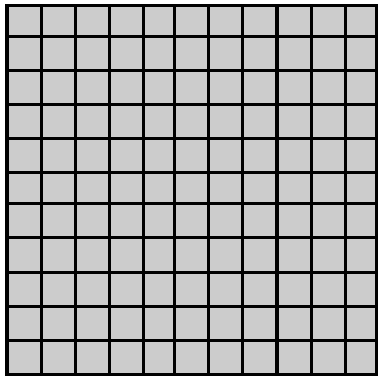
$$\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P) \quad 4\times 16\text{matrix}$$

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

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

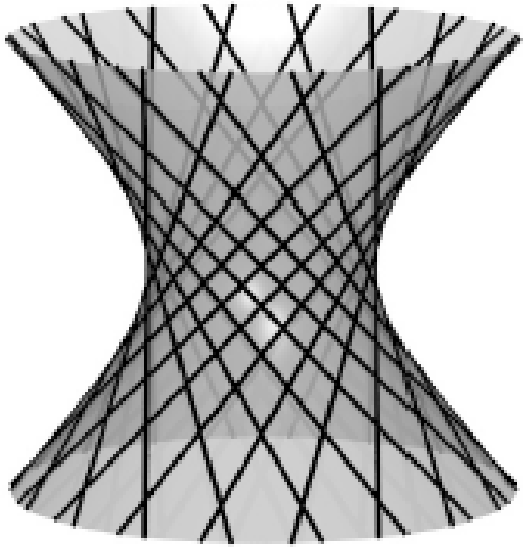
- $\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P)$ and $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3,\ell_6\}\cup\{\ell_4,\ell_5\}}(P)$ contribute to all candidate trees, do not discriminate between them
- left with simpler setting:
  - $F_1 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P)$ 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 is very simple:
  - `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 $u_{i_1,\ldots,i_6} = x_{i_1,i_2,i_4,i_5} y_{i_3,i_6}$.

Segre embeddings



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

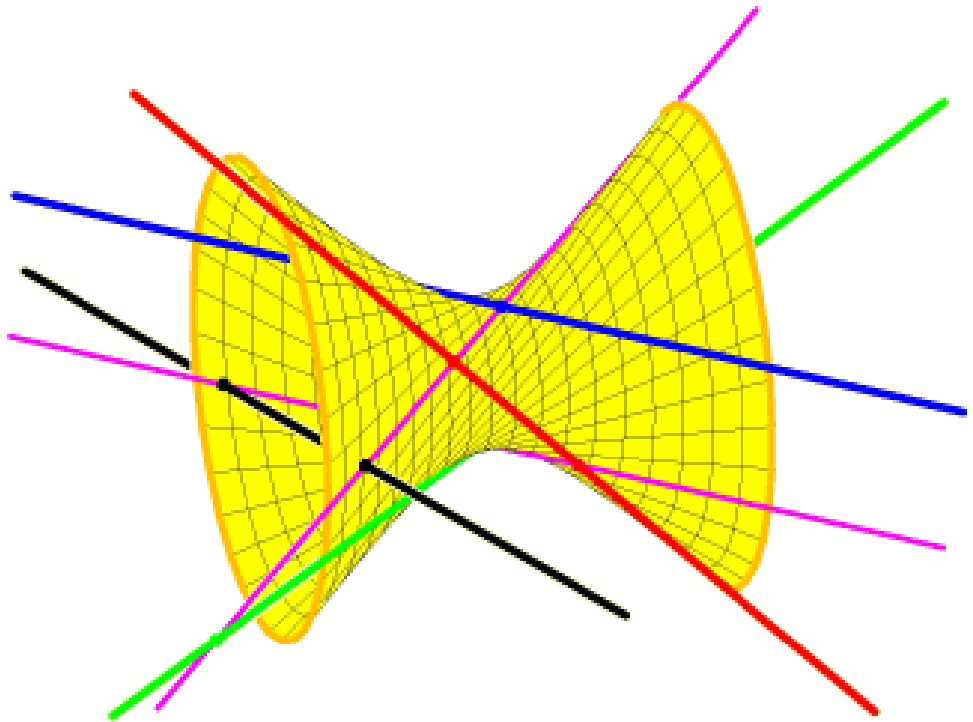
f



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

$\mathbb{P}^1 \times \mathbb{P}^1 \hookrightarrow \mathbb{P}^3$ 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

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

Phylogenetic invariants and Euclidean distance

- evalutation of the 3×3 minors of flattening matrices gives

$$\max_{\phi\, 3\times 3 \text{ minors}} |\phi(P)| = \frac{22}{18225} \sim 0.1207133059 \quad \text{for } \texttt{pars1} \text{ and } \texttt{pars3}$$

$$\max_{\phi} |\phi(P)| = \frac{419}{364500} \sim 0.1149519890 \quad \text{for } \texttt{pars2}$$

- Euclidean distance: favors `pars2` as most likely
  - `pars1`

$$\begin{aligned} \mathrm{dist}(F_1, \mathrm{Sec}(\mathcal{S}(8,8)))^2 &= \sigma_3^2 + \cdots + \sigma_8^2 \\ &= 0.46768 \times 10^{-3} \end{aligned}$$

  - `pars2`

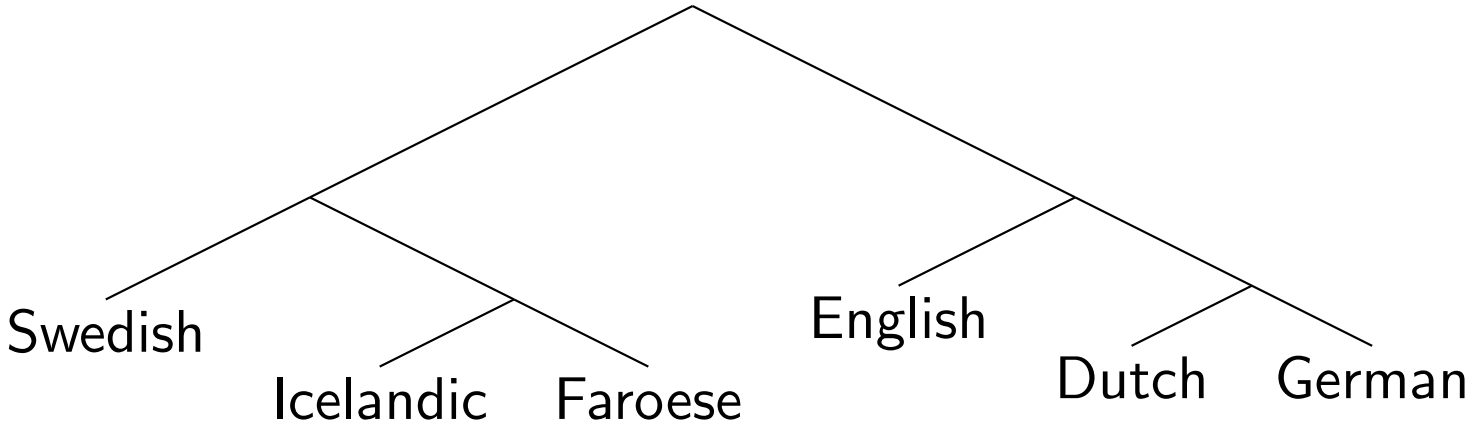
$$\begin{aligned} \mathrm{dist}(F_2, \mathrm{Sec}(\mathcal{S}(8,8)))^2 &= \sigma_3^2 + \cdots + \sigma_8^2 \\ &= 0.24424 \times 10^{-3} \end{aligned}$$

  - `pars3`

$$\begin{aligned} \mathrm{dist}(F_3, \mathrm{Sec}(\mathcal{S}(16,4)))^2 &= \sigma_3^2 + \sigma_4^2 \\ &= 0.51457 \times 10^{-3} \end{aligned}$$

Result

- correctly identifies the West Germanic/North Germanic split



- other PHYLIP candidate trees misplaced it

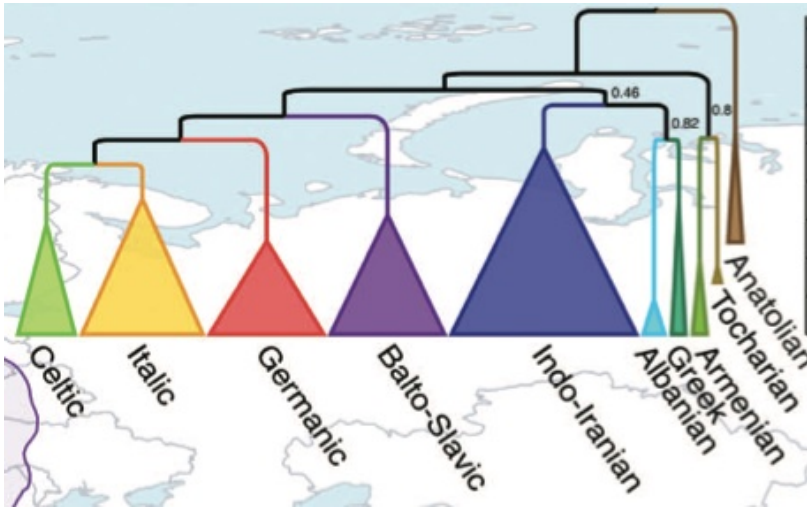
Main Question: can one use this method to obtain new results on the "Indo-European controversy"?

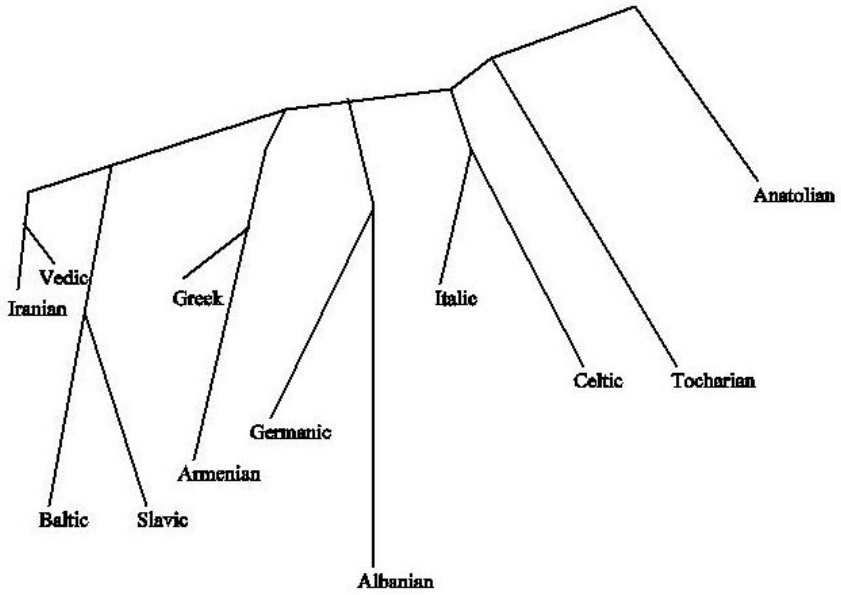
• What is the controversy? Early branches of the tree of Indo-European languages

- 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)

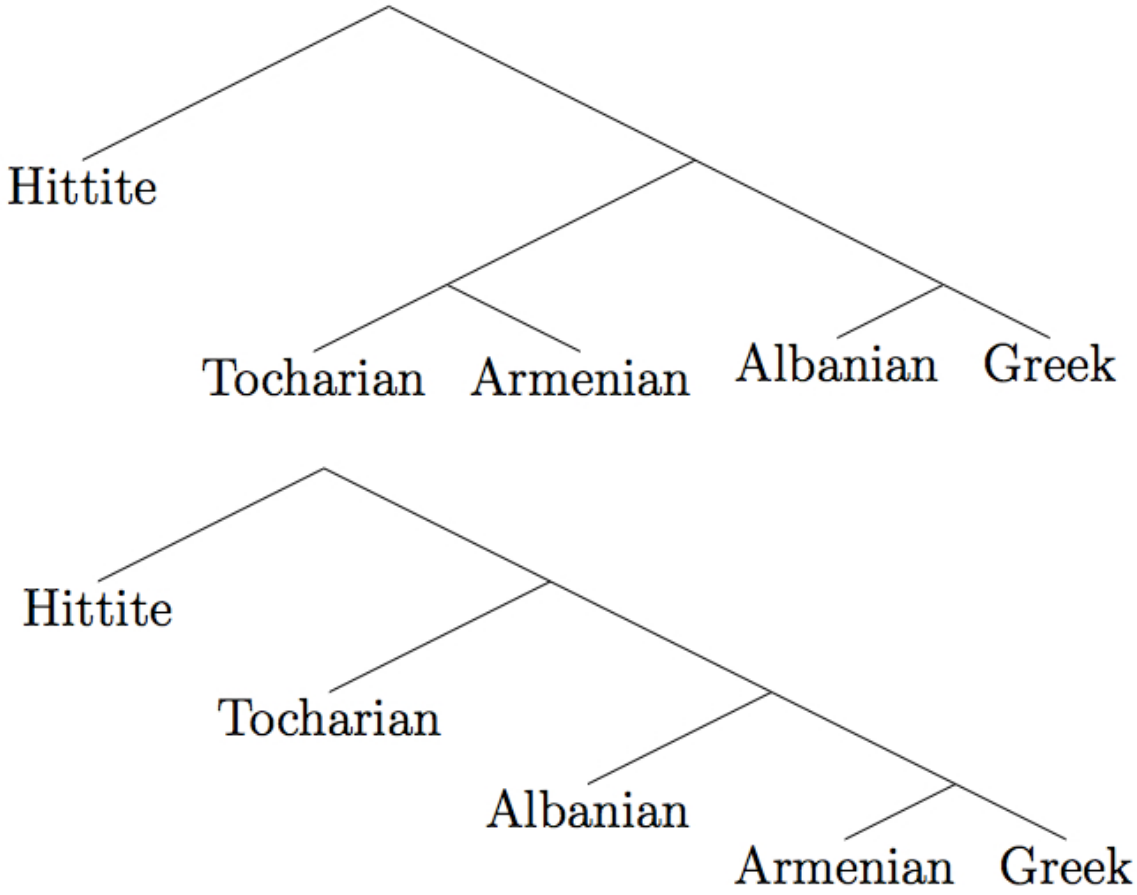
• A. Perelysvaig, M.W. Lewis, *The Indo-European controversy: facts and fallacies in Historical Linguistics*, Cambridge University Press, 2015.





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

Focus on this part of the tree:

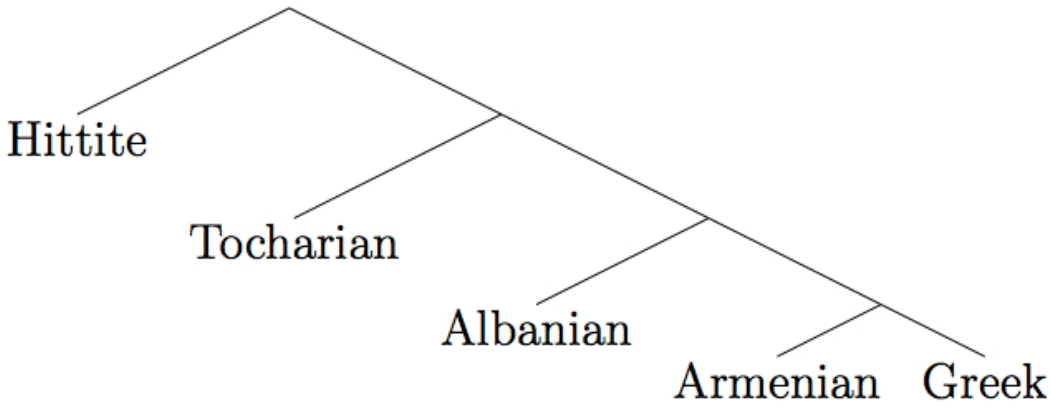


Can detect the difference from syntactic parameters?

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 slightly favor the Ringe–Warnow–Taylor tree over 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)



How to improve the syntactic phylogenetic models?

• the hypothesis that individual syntactic parameters behave like identically distributed independent random variables for a Markov process on a tree needs to be revised: relations between parameters need to be included in the model

• part of the relations can only be detected statistically (see discussion in the other half of this talk)

• need to correct the boundary distribution at the leaves of the tree by a different weight for different parameters that corresponds to different amount of "recoverability" from other parameters (amount of independence)

Spin Glass Models of Syntax

• Karthik Siva, Jim Tao, Matilde Marcolli, *Syntactic Parameters and Spin Glass Models of Language Change*, Linguistic Analysis, Vol. 41 (2017) N. 3-4, 559–608.

• historical examples: Sanskrit flipped some syntactic parameters by influence of Dravidian languages...

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

– focus on linguistic change caused by language interactions

– think of syntactic parameters as spin variables

– spin interaction tends to align (ferromagnet)

– strength of interaction proportional to bilingualism (MediaLab)

– role of temperature parameter: probabilistic interpretation of parameters & amount of code-switching in bilingual populations

– not all parameters are independent: entailment relations

– Metropolis–Hastings algorithm: simulate evolution

The Ising Model of spin systems on a graph G

• graph: vertices = languages, edges = language interaction (strength proportional to bilingual population); over each vertex a set of spin variables (syntactic parameters)

• configurations of spins $s : V(G) \to \{\pm 1\}$

• magnetic field B and correlation strength J: Hamiltonian

$$H(s) = -J \sum_{e \in E(G): \partial(e) = \{v, v'\}} s_v s_{v'} - B \sum_{v \in V(G)} s_v$$

• first term measures degree of alignment of nearby spins

• second term measures alignment of spins with direction of magnetic field

Equilibrium Probability Distribution

- Partition Function $Z_G(\beta)$

$$Z_G(\beta) = \sum_{s: V(G) \to \{\pm 1\}} \exp(-\beta H(s))$$

- Probability distribution on the configuration space: Gibbs measure

$$\mathbb{P}_{G,\beta}(s) = \frac{e^{-\beta H(s)}}{Z_G(\beta)}$$

- low energy states weight most

- at low temperature (large β): ground state dominates; at higher temperature (β small) higher energy states also contribute

Average Spin Magnetization

$$M_G(\beta) = \frac{1}{\#V(G)} \sum_{s:V(G)\to\{\pm 1\}} \sum_{v\in V(G)} s_v \, \mathbb{P}(s)$$

- Free energy $F_G(\beta, B) = \log Z_G(\beta, B)$

$$M_G(\beta) = \frac{1}{\#V(G)} \frac{1}{\beta} \left(\frac{\partial F_G(\beta, B)}{\partial B} \right) |_{B=0}$$

- if all syntactic parameters were independent: just have several uncoupled Ising models (low temperature: converge to more prevalent up/down state in initial configuration; high temperature fluctuations around zero magnetization state)

Syntactic Parameters and Ising/Potts Models

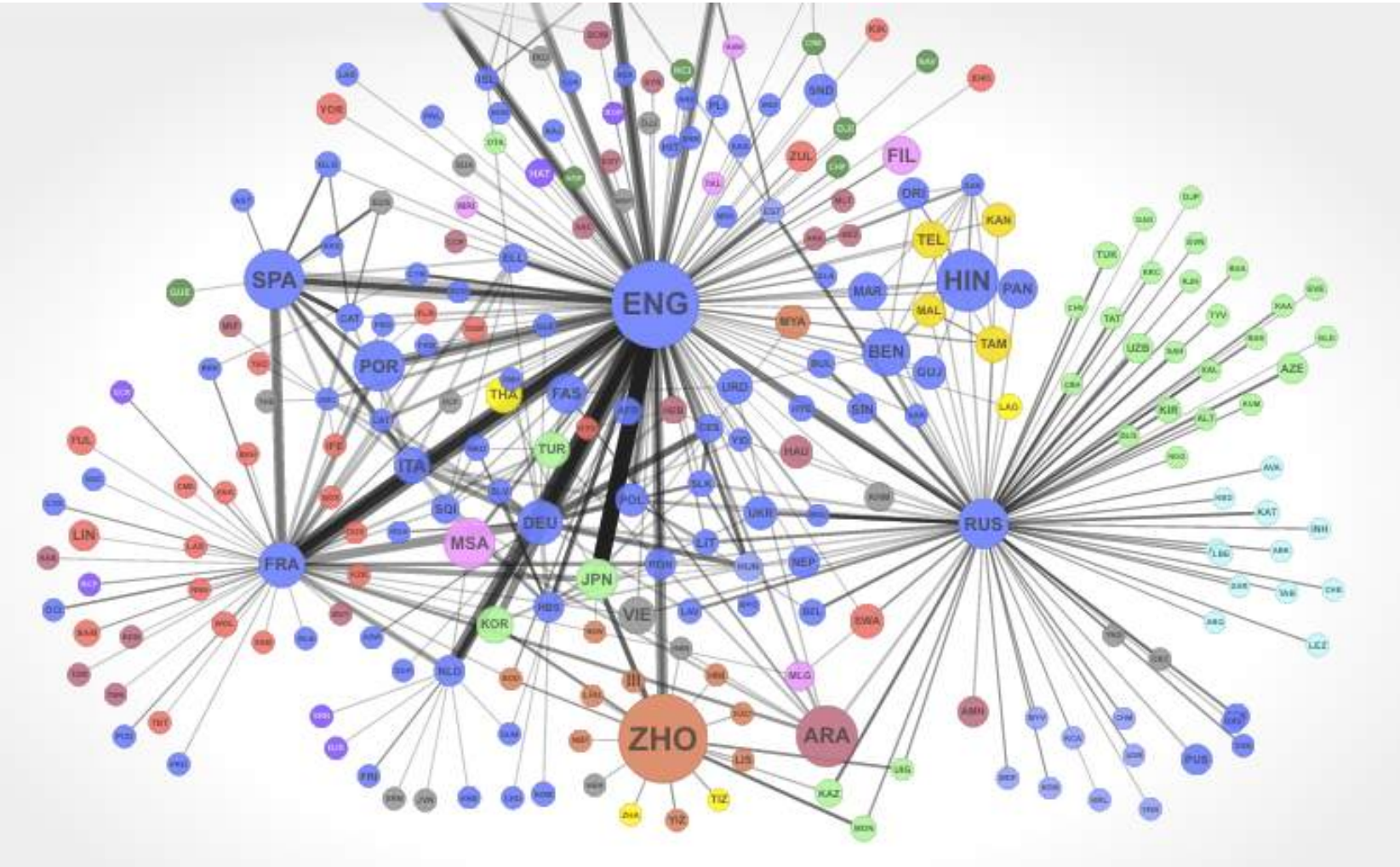
• characterize set of $n = 2^N$ languages $\mathcal{L}_i$ by binary strings of N syntactic parameters (Ising model)

• or by ternary strings (Potts model) if take values ± 1 for parameters that are set and 0 for parameters that are not defined in a certain language

• a system of n interacting languages = graph G with $n = \#V(G)$

• languages $\mathcal{L}_i$ = vertices of the graph (e.g. language that occupies a certain geographic area)

• languages that have interaction with each other = edges $E(G)$ (geographical proximity, or high volume of exchange for other reasons)



graph of language interaction (detail) from Global Language Network of MIT MediaLab, with interaction strengths J_e on edges based on number of book translations (or Wikipedia edits)

• if only one syntactic parameter, would have an Ising model on the graph G: configurations $s : V(G) \to \{\pm 1\}$ set the parameter at all the locations on the graph

• variable interaction energies along edges (some pairs of languages interact more than others) • magnetic field B and correlation strength J: Hamiltonian

$$H(s) = \; - \sum_{e \in E(G) : \partial(e) = \{v, v'\}} \sum_{i=1}^{N} J_e \, s_{v,i} \, s_{v',i}$$

• if N parameters, configurations

$$\underline{s} = (s_1, \ldots, s_N) : V(G) \to \{\pm 1\}^N$$

• if all N parameters are independent, then it would be like having N non-interacting copies of a Ising model on the same graph G (or N independent choices of an initial state in an Ising model on G)

Metropolis–Hastings

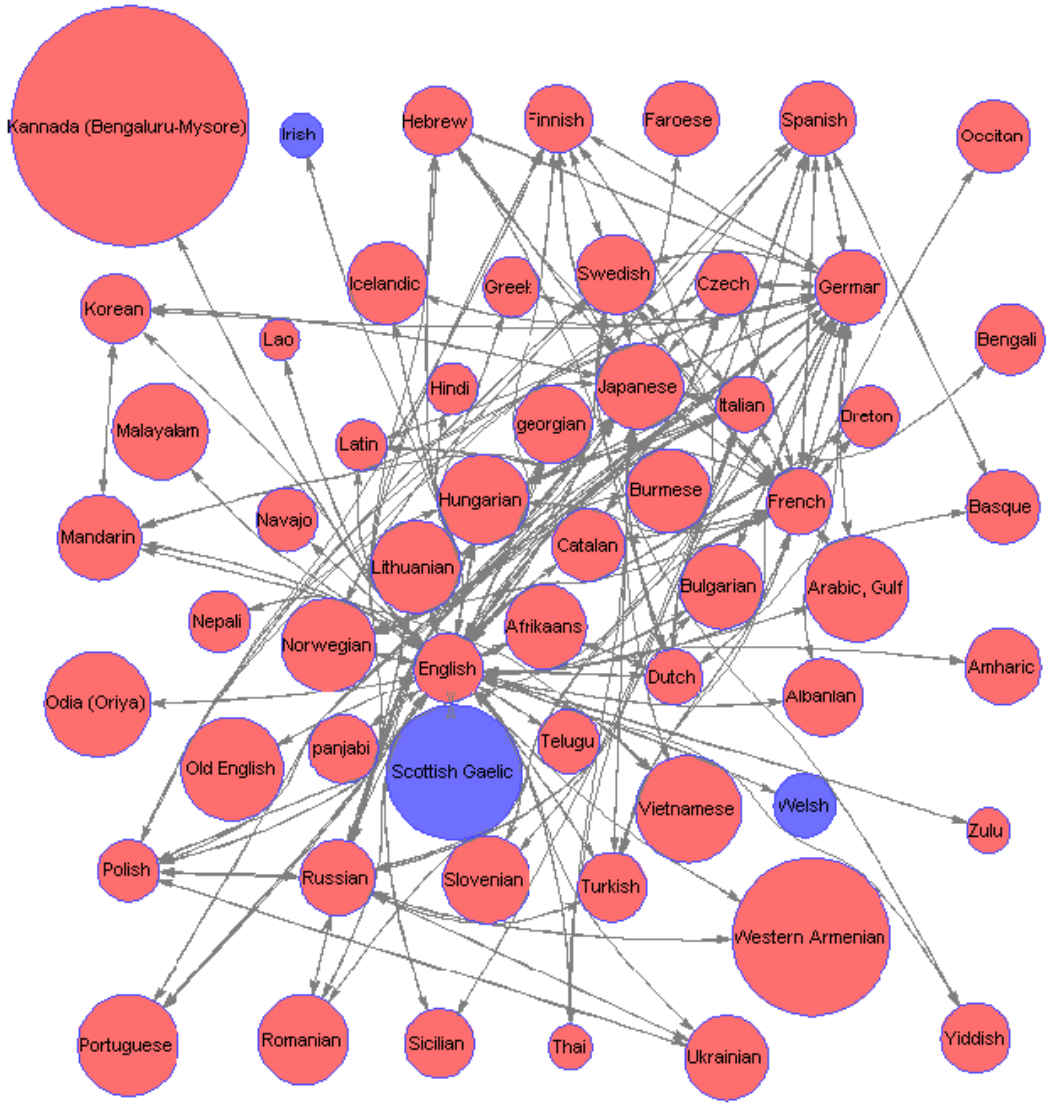
- detailed balance condition $\mathbb{P}(s)\mathbb{P}(s \to s') = \mathbb{P}(s')\mathbb{P}(s' \to s)$ for probabilities of transitioning between states (Markov process)
- transition probabilities $\mathbb{P}(s \to s') = \pi_A(s \to s') \cdot \pi(s \to s')$ with $\pi(s \to s')$ conditional probability of proposing state s' given state s and $\pi_A(s \to s')$ conditional probability of accepting it
- Metropolis–Hastings choice of acceptance distribution (Gibbs)

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

satisfying detailed balance

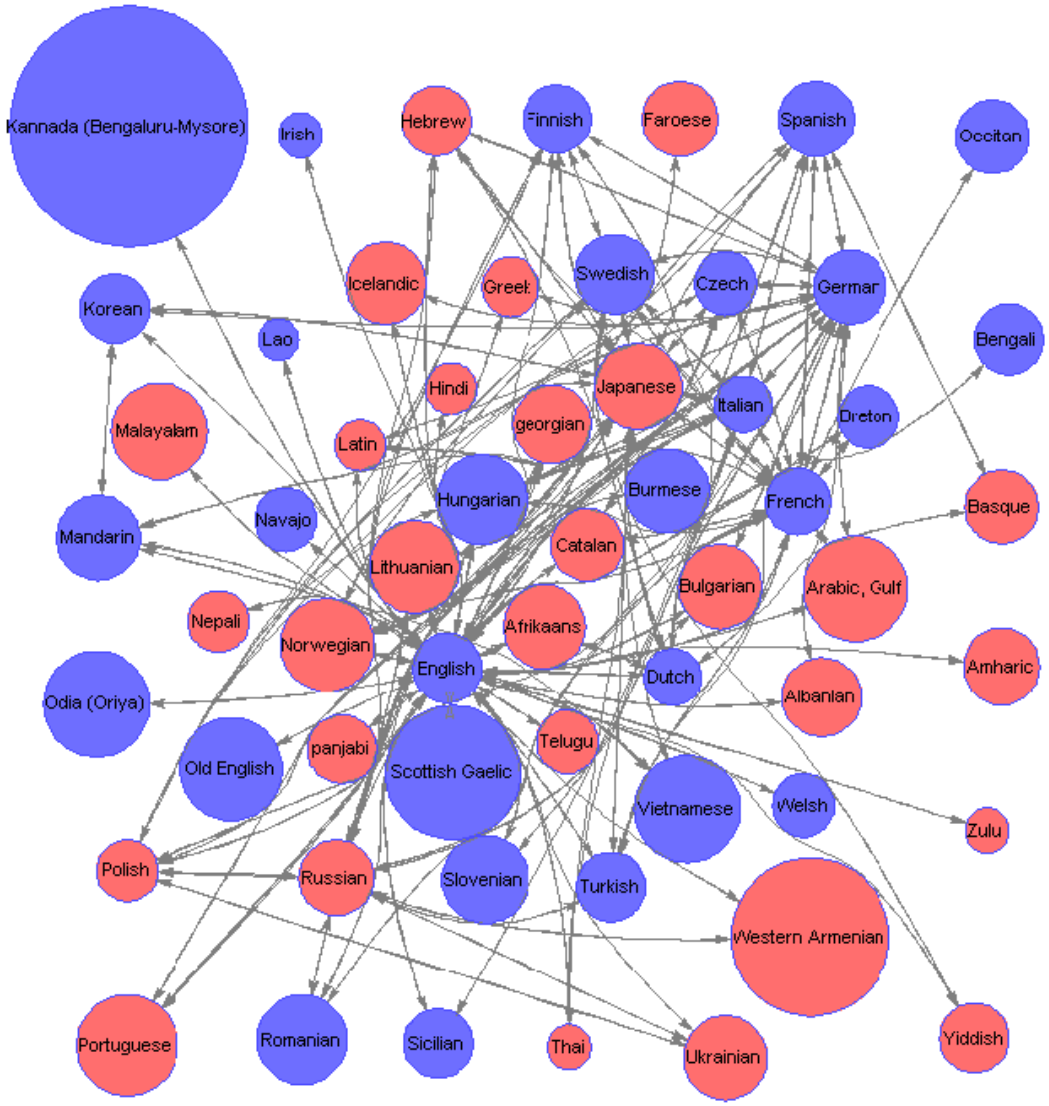
- selection probabilities $\pi(s \to s')$ single-spin-flip dynamics
- ergodicity of Markov process $\Rightarrow$ unique stationary distribution

Example: Single parameter dynamics *Subject-Verb* parameter



Initial configuration: most languages in SSWL have $+1$ for *Subject-Verb*; use interaction energies from MediaLab data

Equilibrium: low temperature all aligned to $+1$; high temperature:



Temperature: fluctuations in bilingual users between different structures ("code-switching" in Linguistics)

Entailment relations among parameters

- Example: $\{p_1, p_2\} = \{\text{Strong Deixis}, \text{Strong Anaphoricity}\}$

	p_1	p_2
ℓ_1	$+1$	$+1$
ℓ_2	-1	0
ℓ_3	$+1$	$+1$
ℓ_4	$+1$	-1

$\{\ell_1, \ell_2, \ell_3, \ell_4\} = \{\text{English}, \text{Welsh}, \text{Russian}, \text{Bulgarian}\}$

Strong Deixis $+1$: governs possible positions of demonstratives in the nominal domain

Strong Anaphoricity $+1$: obligatory dependence on an antecedent in a local and asymmetric relation to anaphor

Modeling Entailment

- 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

Acceptance probabilities

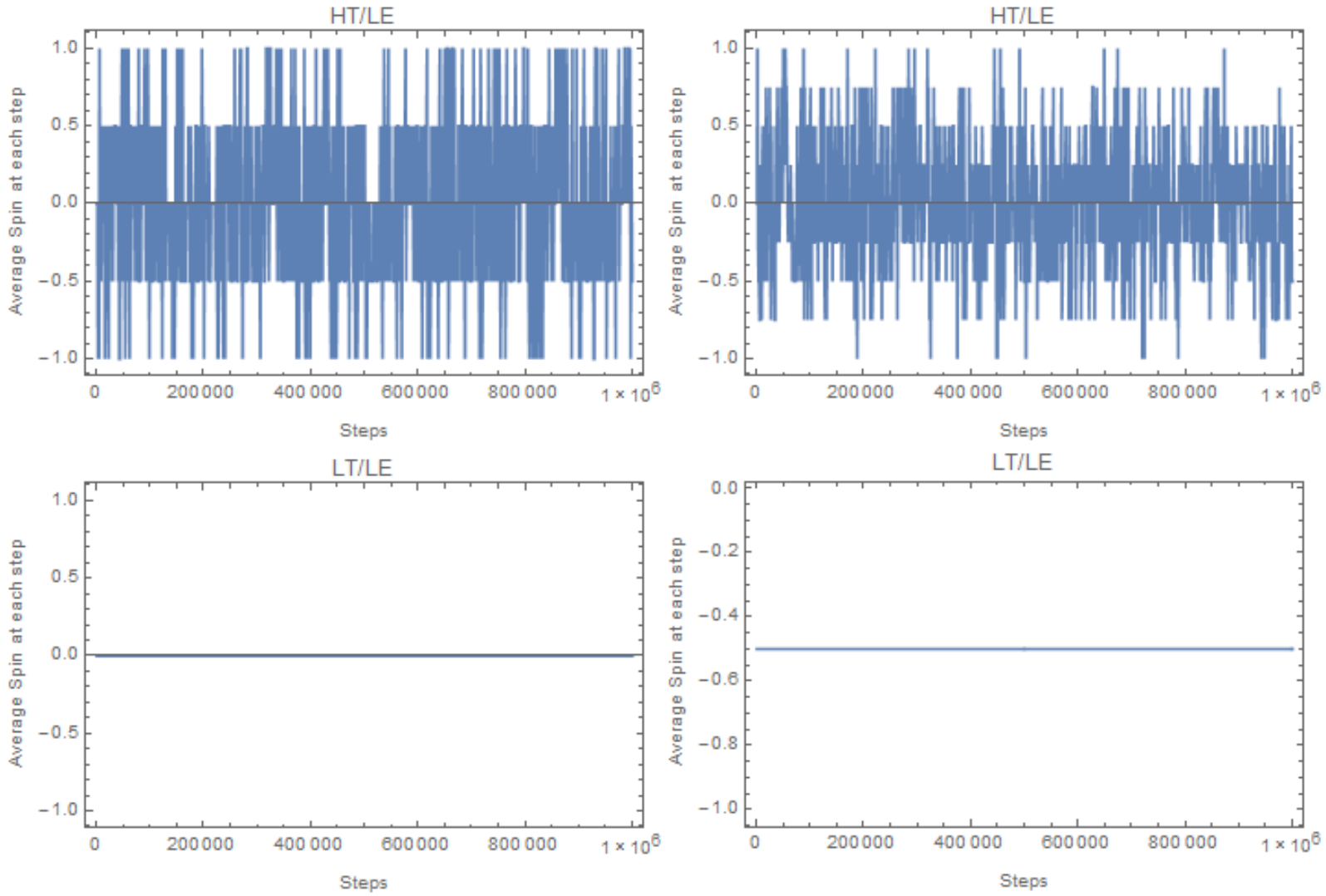
$$\pi_A(s \to s \pm 1 \,(\mathrm{mod}\, 3)) = \begin{cases} 1 & \text{if } \Delta_H \leq 0 \\ \exp(-\beta \Delta_H) & \text{if } \Delta_H > 0. \end{cases}$$

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

Equilibrium configuration

(p_1, p_2)	HT/HE	HT/LE	LT/HE	LT/LE
ℓ_1	$(+1, 0)$	$(+1, -1)$	$(+1, +1)$	$(+1, -1)$
ℓ_2	$(+1, -1)$	$(-1, -1)$	$(+1, +1)$	$(+1, -1)$
ℓ_3	$(-1, 0)$	$(-1, +1)$	$(+1, +1)$	$(-1, 0)$
ℓ_4	$(+1, +1)$	$(-1, -1)$	$(+1, +1)$	$(-1, 0)$

Average value of spin



p_1 left and p_2 right in low entailment energy case

• when consider more realistic models (at least the 28 languages and 63 parameters of Longobardi–Guardiano with all their 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

• in the other part of this talk: discuss a coding theory perspective on code parameters; induced dynamics on the space of codes shows more easily long term behavior of the system

How to improve this dynamical model?

- language change is related to mechanisms of language acquisition
- dynamical systems models of language acquisition were proposed by Berwick and Niyogi based on a Markov model on a space of possible grammats (in the formal languages sense)
- would like to couple the spin glass dynamics capturing language interaction through code-switching and bilingualism to a dynamical model of language acquisition

Next Episode: how to detect relations between syntactic parameters? what is the manifold of syntax?