Persistent Topology of Syntax

Matilde Marcolli

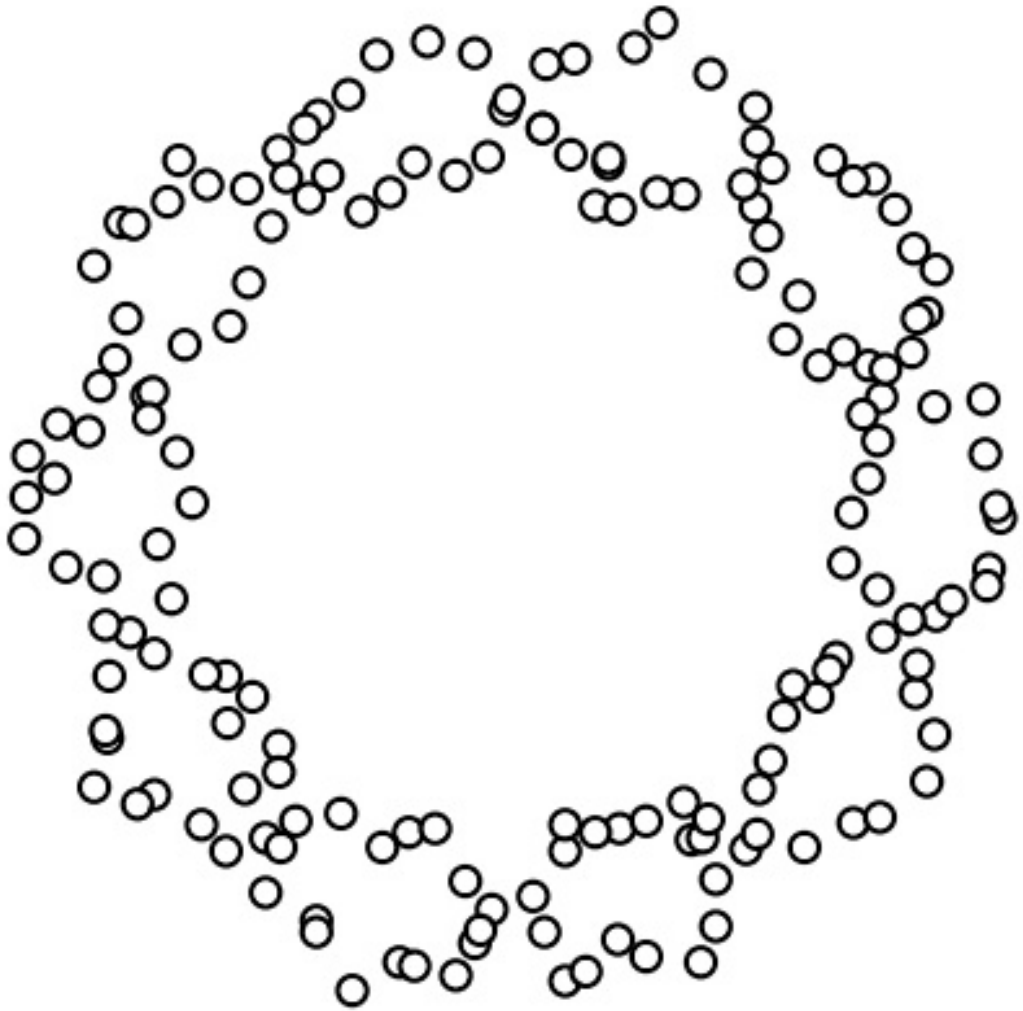
MAT1509HS: Mathematical and Computational Linguistics

University of Toronto, Winter 2019, T 4-6 and W 4, BA6180

This lecture based on:

- 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.

Persistent Topology of Data Sets

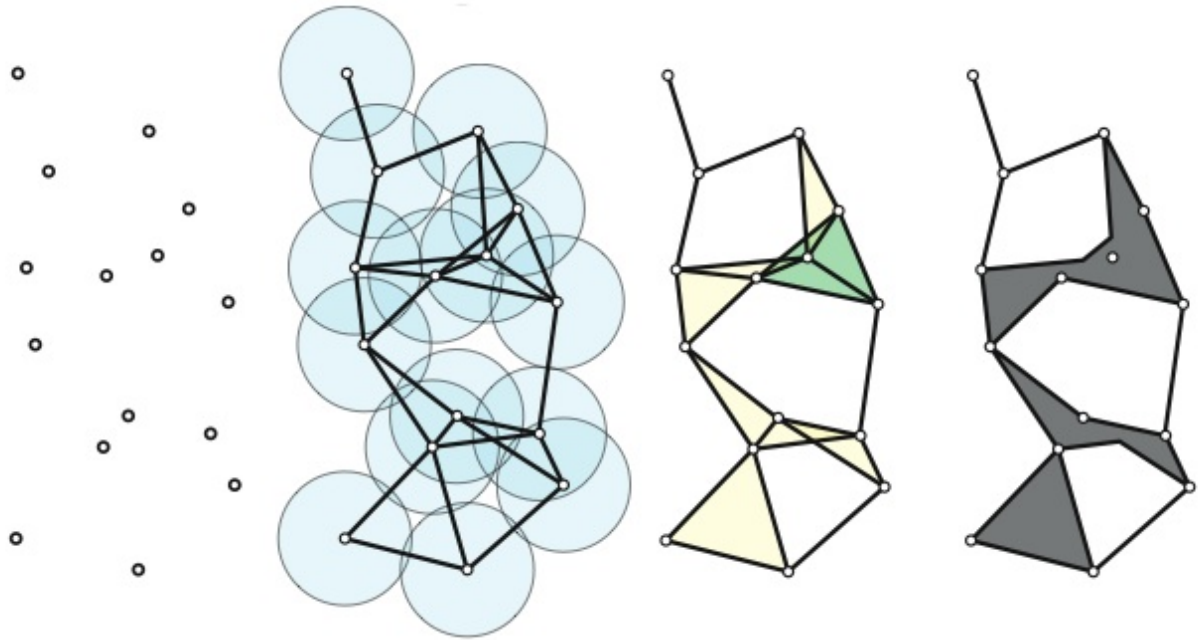


how data cluster around topological shapes at different scales

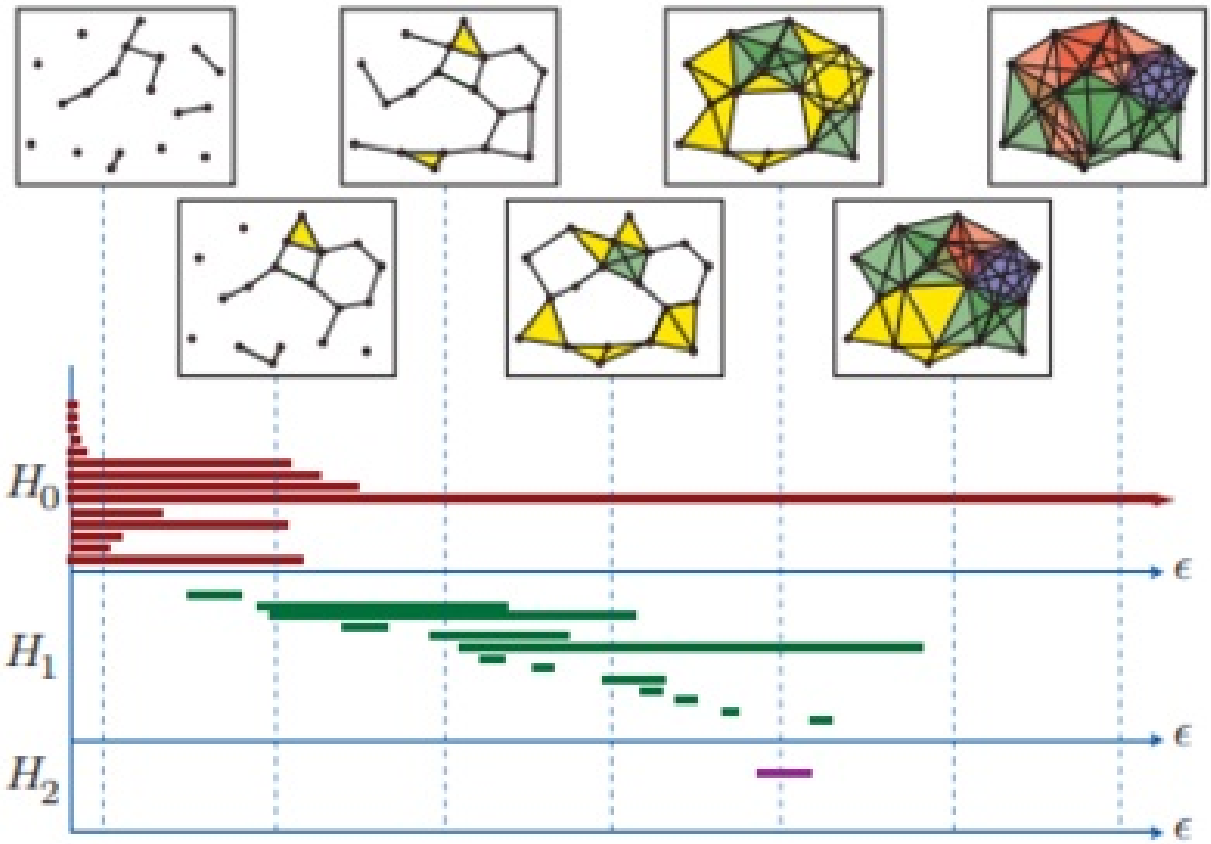
Vietoris–Rips complexes

- 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 $H_n(X, \epsilon_1) \to H_n(X, \epsilon_2)$

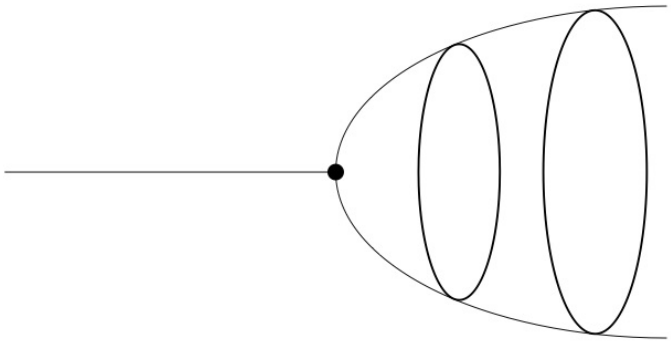


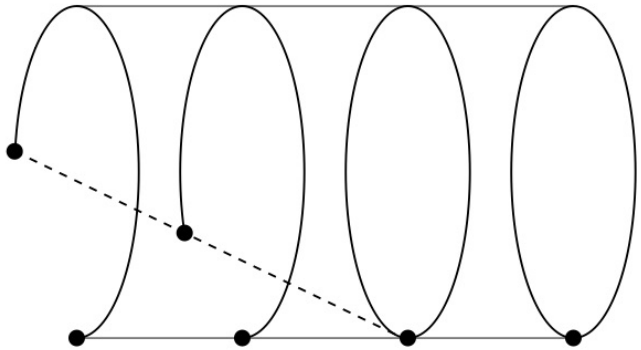
barcode diagrams: births and deaths of persistent generators

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

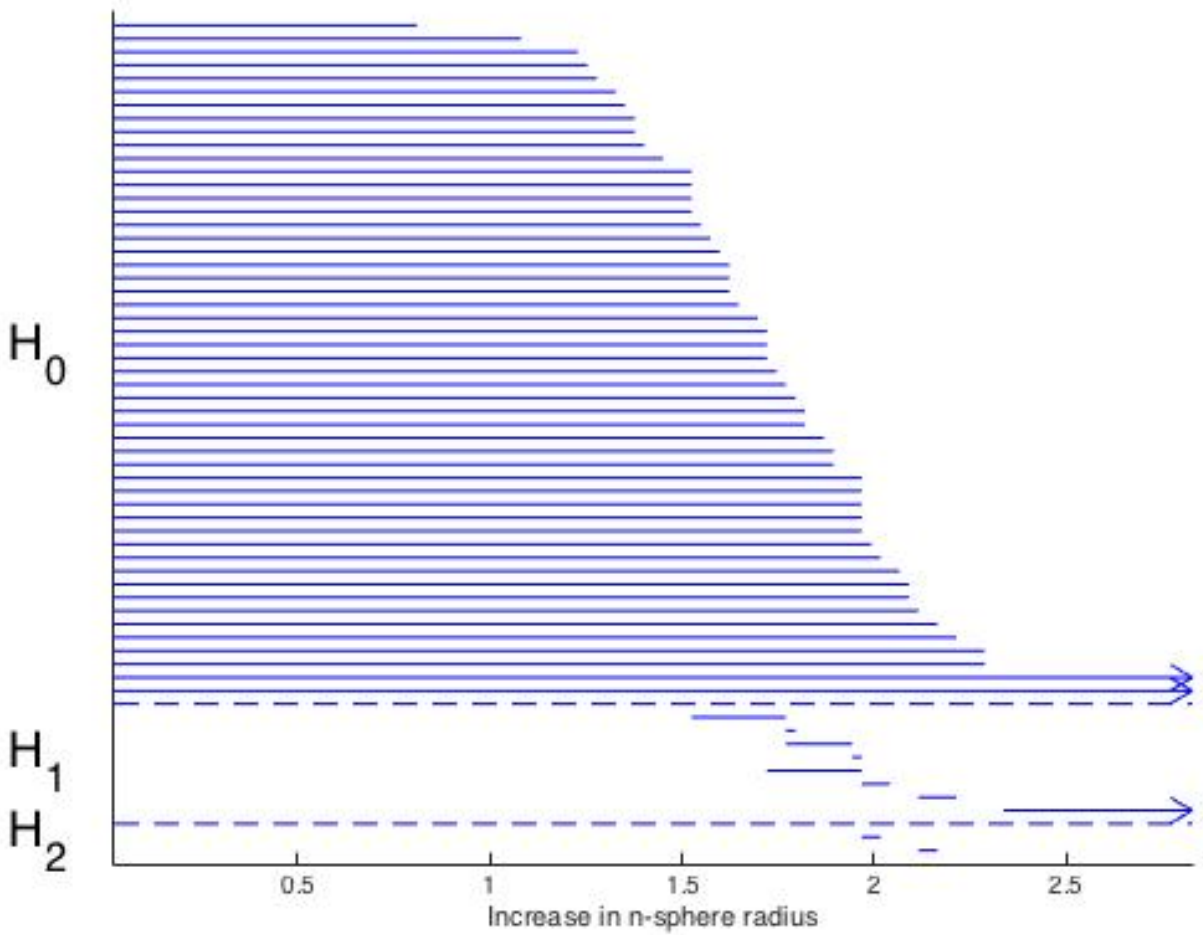




- "Hopf bifurcation" type phenomenon
- two different branches of a tree closing up in a loop

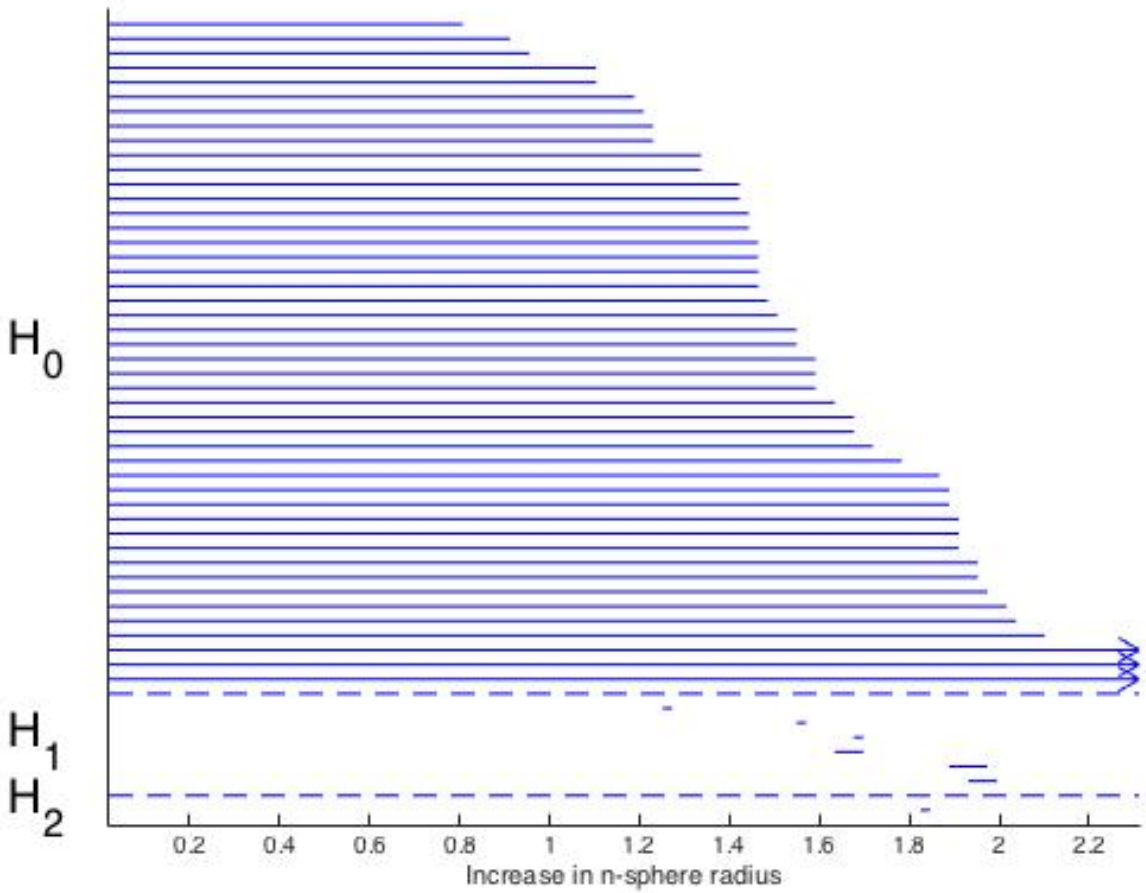
two different types of phenomena of historical linguistic development within a language family

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

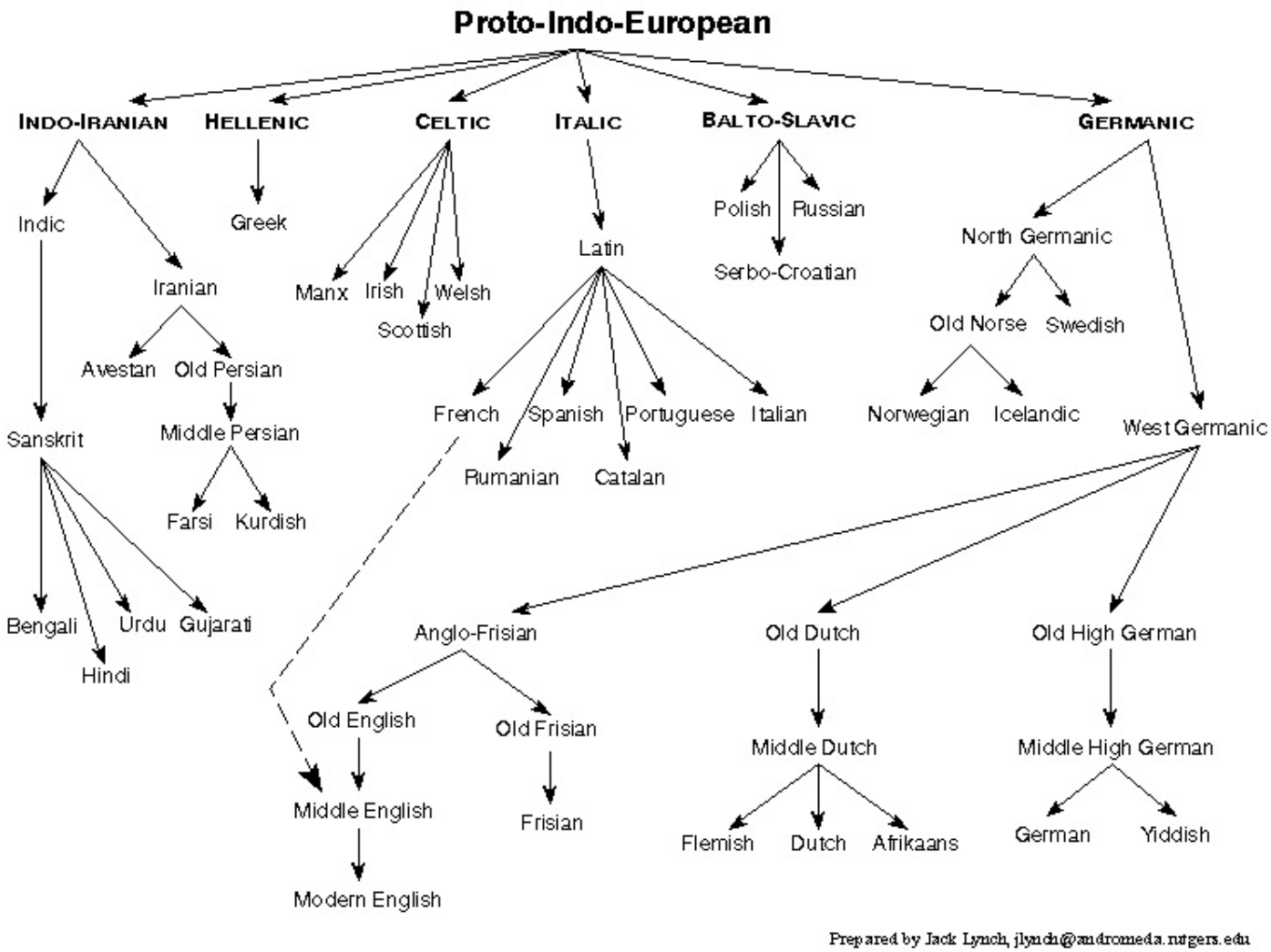


- Three persistent components of H_0

(Mande, Atlantic-Congo, Kordofanian)

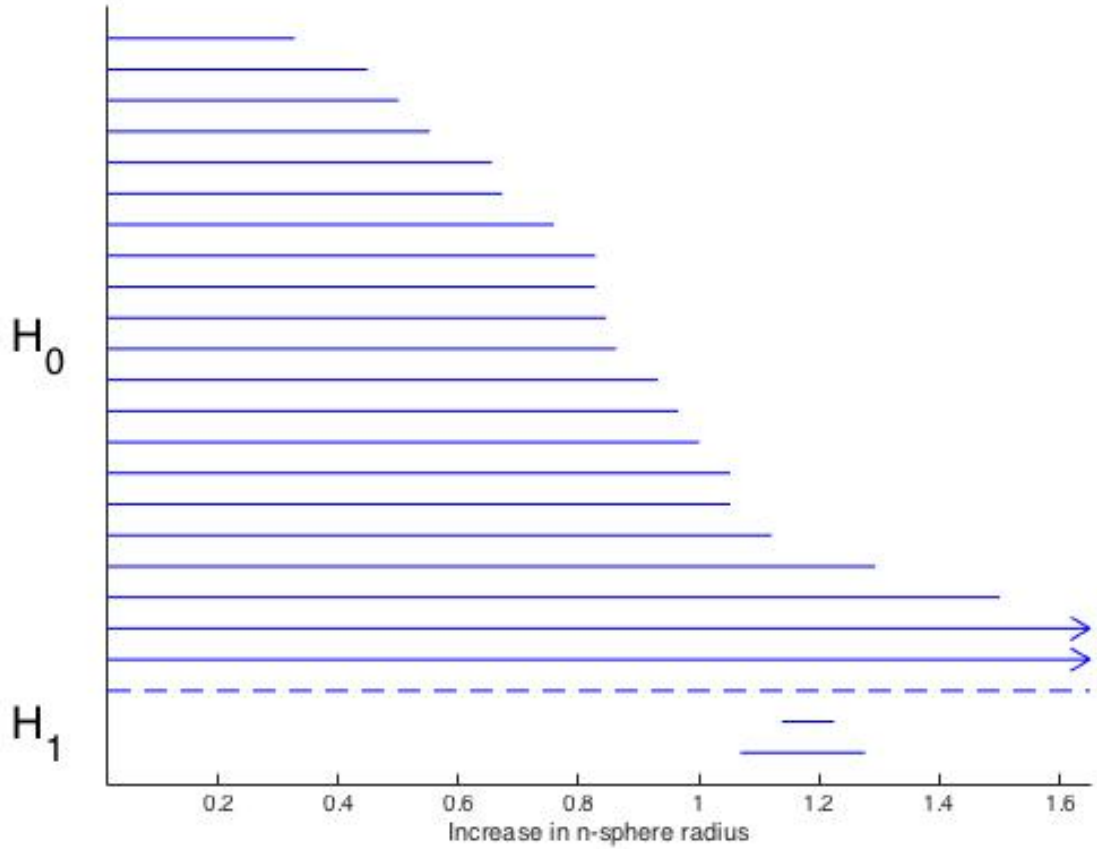
- No persistent H_1

The origin of persistent H_1 of Indo-European Languages?



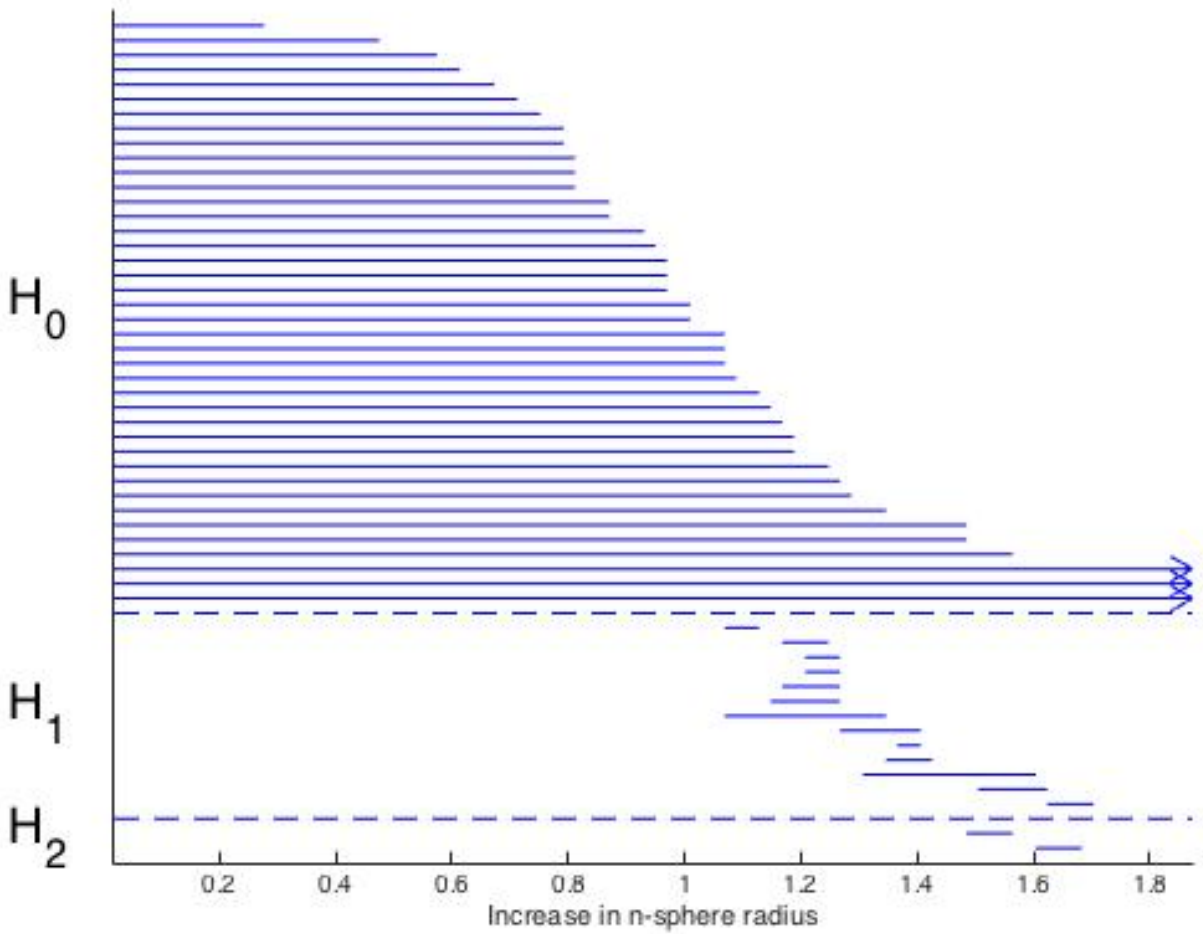
Naive guess: the Anglo-Norman bridge ... but lexical not syntactic!

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

Further work

- Alexander Port, Taelin Karidi, Matilde Marcolli, *Historical Linguistics via Topological Analysis of Syntactic Structures*, in preparation

- persistent H_0 (persistent connected components) as another method for the reconstruction of phylogenetic trees of language families
- persistent first homology group H_1 analyzed on all clusters and scales

Mathematical Digression on Persistent Homology

- Yuri I. Manin, Matilde Marcolli, *Nori Diagrams and Persistent Homology*, preprint 2019

- What kind of homology theory is Persistent Homology?
  - Does it have a good categorical formulation?
  - Based on simplicial sets (simplicial complexes) but with a scale parameter: what kind of homotopy theory?
  - Is there a good homotopy theory formalism? (Quillen model categories, etc.)

Category of data sets

- small category $\mathcal{P}_{\mathbb{E}}$: objects triples (A, f, P) finite set A, embedding $f : A \hookrightarrow \mathbb{E}$ in a fixed large Euclidean space, probability distribution P on A (probability on $\mathbb{E}$ supported on $f(A)$)
- morphisms in $\mathrm{Mor}_{\mathcal{P}_{\mathbb{E}}}((A, f, P), (A', f', P'))$ pairs $(\varphi, \tilde{\varphi})$ of Lipschitz map $\tilde{\varphi} : \mathbb{E} \to \mathbb{E}$ restricting to map $\varphi : A \to A'$ so that $P' = \varphi_* P$ pushforward measure

$$(\varphi_* P)_y = \sum_{x \in \varphi^{-1}(y)} P_x, \quad \forall y \in A'$$

- morphisms in $\mathcal{P}_{\mathbb{E}}$ are subsets of Lipschitz functions of $\mathbb{E}$
- probability distribution P finite set A assigns a degree of reliability to the points in the dataset: outlier points with low probability $P_x$regarded as errors in the data and discarded in the construction simplicial complex

Thin categories, posets, and categories of functors

- *thin category*: for any $x, y \in \mathrm{Obj}(\mathcal{C})$ the set of morphisms $\mathrm{Mor}_{\mathcal{C}}(x, y)$ consists of at most one element
- *poset*: $(S, \leq)$ partial order $x \leq x$, if $x \leq y$ and $y \leq z$ then $x \leq z$; if $x \leq y$ and $y \leq x$ then $x = y$
- *equivalence of categories*: a pair of functors $F : \mathcal{C} \leftrightarrows \mathcal{C}' : F'$ and natural isomorphisms $F \circ F' \simeq \mathrm{Id}_{\mathcal{C}'}$ and $F' \circ F \simeq \mathrm{Id}_{\mathcal{C}}$
- a thin category is equivalent to a poset

• Category of covariant functors (with natural transformations as morphisms) $\mathcal{F}(\mathcal{C}, \mathcal{C}') = \{F : \mathcal{C} \to \mathcal{C}' \text{ functors}\}$ and $\mathrm{Mor}_{\mathcal{F}(\mathcal{C},\mathcal{C}')}(F, F') = \{\eta : F \to F' \text{ natural transformations}\}$

• Category of "persistent modules"

$$\mathcal{A}^{(S,\leq)} := \mathcal{F}((S, \leq), \mathcal{A})$$

poset $(S, \leq)$ as a thin category; in particular will assume $\mathcal{A}$ an abelian category like $\mathrm{Vect}_{\mathbb{K}}$ finite dimensional vector spaces over a field $\mathbb{K}$; R-Mod modules over a commutative ring R

Vietoris–Rips functor

- choice of an error threshold $\Lambda \in [0, 1]$
- an object (A, f, P) in $\mathcal{P}_{\mathbb{E}}$: filter out error

$$X_\Lambda := \{x \in f(A) \subset \mathbb{E} \,|\, P_x \geq \Lambda\}$$

- choice of a scale $t \in \mathbb{R}_+^*$
- Vietoris–Rips complex $VR_\bullet(X_\Lambda, t)$ with $VR_n(X_\Lambda, t)$ span of $(n+1)$-tuples of points X_Λ with all pairwise distances $\mathrm{dist}(x_i, x_j) \leq t$

- choice of a morphism $(\varphi, \tilde{\varphi})$ in $\mathcal{P}_{\mathbb{E}}$, Lipschitz function $\tilde{\varphi} : \mathbb{E} \to \mathbb{E}$ with Lipschitz constant $K > 0$ and $\varphi : A \to A'$ with $m = \min_{y \in A'} \#\varphi^{-1}(y)$
- map sends X_Λ to $X'_{m\Lambda} = \{y \in f'(A') \,|\, P_y \geq m\Lambda\}$ and sends $x_i, x_j \in X_\Lambda$ with distance $\mathrm{dist}(x_i, x_j) \leq t$ to $\tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \in X'_{m\Lambda}$ with $\mathrm{dist}(\tilde{\varphi}(x_i), \tilde{\varphi}(x_j)) \leq Kt$
- it defines morphism of Vietoris–Rips complexes

$$VR(\varphi, \tilde{\varphi}) : VR_\bullet(X_\Lambda, t) \to VR_\bullet(X_{m\Lambda}, Kt)$$

- $S = \mathbb{R} \times [0,1]$ partial order structure $\leq$ given by the product order $(t, \Lambda) \leq (t', \Lambda')$ iff $t \leq t'$ and $\Lambda \geq \Lambda'$ (reverse ordering on $[0,1]$)
- category $\mathcal{M}^{(S,\leq)} = \mathcal{F}((S,\leq), \mathcal{M})$ with $\mathcal{M}$ either Ch_R (unbounded) chain complexes over a commutative ring R or $\Delta\mathcal{S}$ simplicial sets
- assignments $(A, f, P) \mapsto VR(A, f, P)$ and $(\varphi, \tilde{\varphi}) \mapsto VR(\varphi, \tilde{\varphi})$ as above determine a functor $VR : \mathcal{P}_{\mathbb{E}} \to \mathcal{M}^{(S,\leq)}$
- usual Vietoris–Rips construction when $\Lambda = 0$ (no filtering by probability)

Persistent homology: image of morphism
$H_\bullet(VR_\bullet^{t,\Lambda}(A, f, P)) \to H_\bullet(VR_\bullet^{t',\Lambda}(A, f, P))$ for $t \leq t'$ and fixed Λ

Quillen model categories: category $\mathcal{M}$ with three special classes of morphisms: weak equivalences, fibrations and cofibrations

Axioms

1. $\mathcal{M}$ has all small limits and colimits
2. if two among the three maps $f, g, g \circ f$ are weak equivalences the third also is
3. if f is a retract of g and g is a weak equivalence, fibration, or cofibration, then f also is
4. given a commutative diagram

$$\begin{array}{ccc} A & \longrightarrow & X \\ \downarrow{\scriptstyle \iota} & & \downarrow{\scriptstyle p} \\ B & \longrightarrow & Y \end{array}$$

   a lift $B \to X$ exists if either ι cofibration and p acyclic fibration (both fibration and weak equivalence) or ι acyclic cofibration (both cofibration and weak equivalence) and p fibration
5. morphisms g in $\mathcal{M}$ factor: $g = qi$ with q acyclic fibration and i cofibration or $g = pj$ with p fibration and j acyclic cofibration

Model structures on chain complexes and simplicial sets

• *projective model structure on chain complexes* Ch_R: weak equivalences are quasi-isomorphisms of chain complexes, fibrations chain maps $\varphi_\bullet : C_\bullet \to C'_\bullet$ where each $\varphi_n : C_n \to C'_n$ epimorphism of R-modules, cofibrations chain maps level-wise monomorphisms of R-modules with projective cokernel

• *Kan–Quillen model structure on simplicial sets* $\Delta\mathcal{S}$: weak equivalences are morphisms that induce a weak homotopy equivalence of topological spaces on geometric realizations, fibrations are Kan fibrations, and the cofibrations are monomorphisms of simplicial sets

• these model structures have good properties (cofibrantly generated)

Projective model category on category of functors

• if $\mathcal{D}$ cofibrantly generated model category have projective model structure on $\mathcal{F}(\mathcal{C}, \mathcal{D})$: weak equivalences and fibrations are natural transformations $\eta : F \to F'$ of functors $F, F' : \mathcal{C} \to \mathcal{D}$ such that, for all objects $X \in \mathrm{Obj}(\mathcal{C})$, the morphisms $\eta_X : F(X) \to F'(X)$ in $\mathcal{D}$ are weak equivalences and fibrations, respectively (also cofibrantly generated)

- cofibrantly generated model category $\mathcal{M}$
- category of functors $\mathcal{M} = \mathcal{F}((S, \leq), \mathcal{M})$
- $\mathcal{M}^{(S,\leq)}$ admits projective model structure (cofibrantly generated)

- Quillen pair $L : \mathcal{M} \leftrightarrow \mathcal{N} : R$ between model categories: adjoint pair of functors (L, R) where L preserves cofibrations and R preserves fibrations
- Dugger's universal model structure
  - small category $\mathcal{C}$
  - functor $F : \mathcal{C} \to \mathcal{M}$ to model category $\mathcal{M}$
  - exists universal model category $U(\mathcal{C})$ that factorizes $F : \mathcal{C} \to \mathcal{M}$
  - factorization: functor $J : \mathcal{C} \to U(\mathcal{C})$, Quillen pair $L : U(\mathcal{C}) \rightleftarrows \mathcal{M} : R$ and a natural weak equivalence $\eta : L \circ J \to F$
- applied to Vietoris–Rips functor $VR : \mathcal{P}_{\mathbb{E}} \to \mathcal{M}^{(S,\leq)}$ gives model category $U(\mathcal{P}_{\mathbb{E}}) \Rightarrow$ model structure for persistent topology