EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät
FB Neuphilologie
Seminar für Sprachwissenschaft

EVOLAEMP
Language Evolution:
The Empirical Turn

Causal Inference of Evolutionary Networks

Phylogenetic Methods in Historical Linguistics

Tübingen, March 30, 2017

Johannes Dellert

This research has been supported by the ERC Advanced Grant 324246 EVOLAEMP.

Table of Contents

Introduction: General Idea

General ideas behind my talk:

- current evolutionary network inference methods do not scale well, or are not general enough
- we can treat **languages as information-theoretic variables**, and the cognate sets employed for each concept as samples
- cognacy overlaps define information geometry over languages
- vanishing conditional mutual information can be used to test for **conditional independence between languages**
- principles of causal inference sometimes allow us to infer that one language “causes” another
- directionality of causal signal between languages can be interpreted as the **dominant direction of lexical flow**

Table of Contents

Phylogenetic Lexical Flow Inference

A map of the linguistic history of a region should include

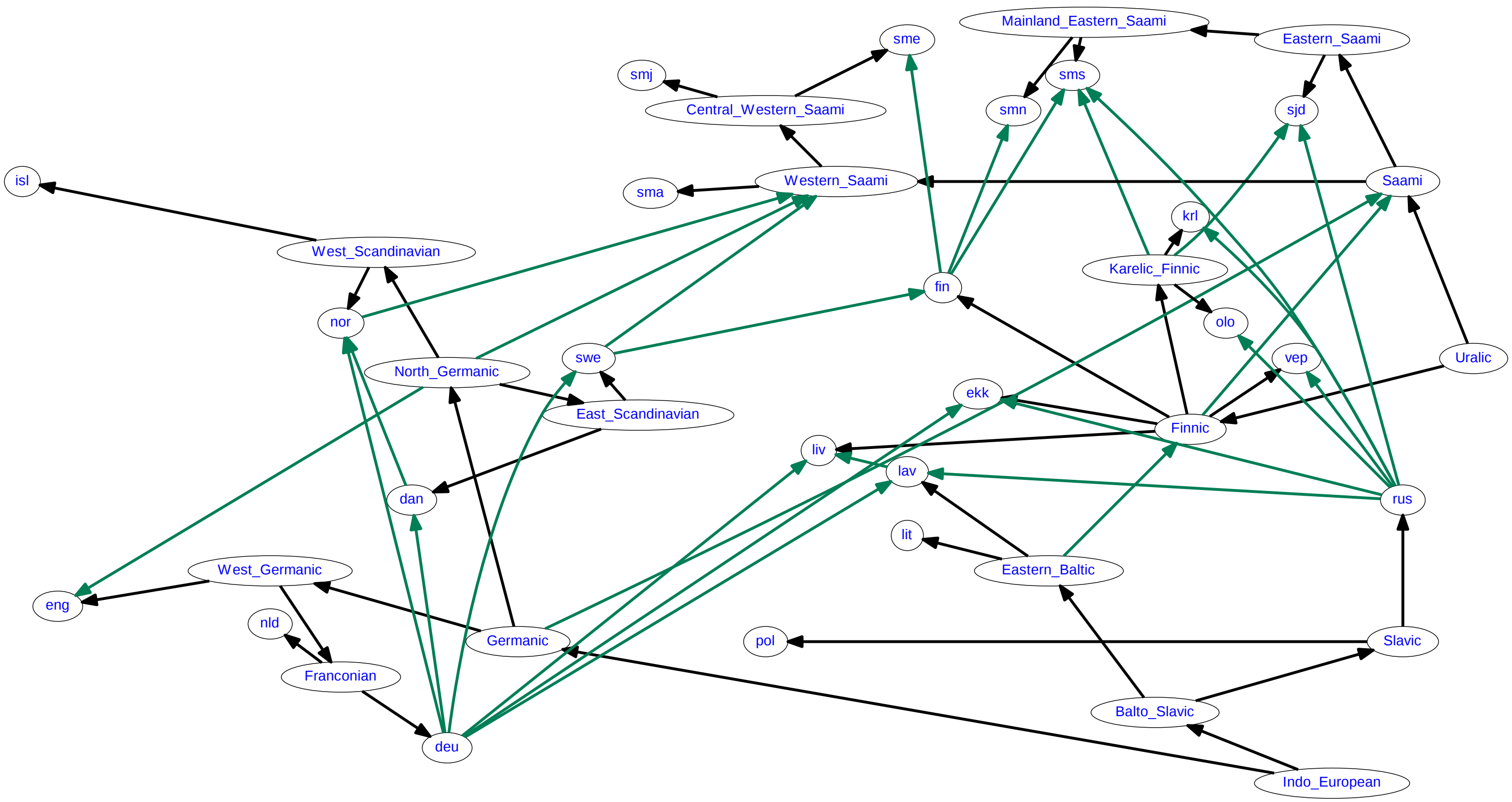
- the paths on which lexical material was inherited (i.e. a phylogenetic tree)
- the paths on which lexical material was borrowed (both among ancestral and living languages)
- taken together, all the paths on which lexical material has “flown” to produce the observable situation (**lexical flow**)

Simplifying assumptions taken in my approach:

- some phylogenetic tree is known (good inference methods exist)
- we have a usable reconstruction of the cognacy classes present at each proto-language (derived by historical linguists, or using some automated reconstruction method)

Phylogenetic Lexical Flow Inference: Example

Desired result for the region around the Baltic Sea:



Existing Phylogenetic Network Methods

Morrison (2011): two main types of phylogenetic network

- **data-display networks**
  - ▷ generalize unrooted trees
  - ▷ use additional virtual nodes to visualize conflicting signals
  - ▷ examples: median network, neighbor-net
- **evolutionary networks**
  - ▷ generalize rooted trees
  - ▷ all nodes represent some (ancestral) language
  - ▷ lateral connections are directed
  - ▷ examples: galled tree, galled network, hybridization network

Existing Phylogenetic Network Methods

Evolutionary network inference is still in its infancy:

- **probabilistic models** are very complex and need a lot of strong modeling assumptions; inference methods do not scale well to large networks, 7 species is the limit hit by Wen et al. (2016)
- models for more languages restrict the search space rather heavily, usually in terms of reticulation cycles
- **galled trees** do not allow node sharing between reticulation cycles ($\Rightarrow$ multiple donor languages not possible)
- **galled networks** allow reticulation cycles to share nodes, but only reticulation nodes, i.e. multi-way colliders are possible (BUT deu $\leftarrow$ eng $\rightarrow$ hin still not representable)
- **hybridization networks** are only slightly more general (they allow leaves as source languages)

Table of Contents

Causal Inference: Basic Idea

- algorithmic techniques to infer causal relationships between variables from observational data alone (Pearl, 2009)
- not possible for two variables: “correlation is not causation”
- but: interaction between more than two variables often provides hints about underlying causal scenario
- underlying theory (Reichenbach’s **Common Cause Principle**) states that whenever two variables are correlated, there must be either a directed causal path in exactly one direction, or a common cause (“no correlation without causation”)
- model causal scenarios as **causal DAGs** (directed acyclic graphs) over the variables, systematically exploit hints to infer properties of the underlying causal DAG

Conditional Independence and Causal Graphs

- core building block: a **conditional independence** relation
- $(X \perp\!\!\!\perp Y \mid Z)$ intuitively means:
  "any dependence between the variables X and Y can be explained by the influence of Z"
- PC algorithm: sequence of conditional independence tests reduces a complete graph to a **causal skeleton**, where no link can be explained away by conditioning on other variables
- removal of link $X - Y$ relies on finding a **separating set**, i.e. a set of variables $\{Z_1, \ldots, Z_n\}$ such that $(X \perp\!\!\!\perp Y \mid Z_1, \ldots, Z_n)$
- example: $(\mathit{sma} \perp\!\!\!\perp \mathit{fin} \mid \mathit{swe}, \mathit{Uralic})$

Unshielded Collider Criterion

- directionality inference on the causal skeleton
- for each pattern of the form $X - Z - Y$ (**unshielded triple**), ask whether the central variable was part of the separating set that was used for explaining away the link $X - Y$
- underlying idea: if Z was not necessary to explain away $X - Y$, this excludes all patterns except $X \rightarrow Z \leftarrow Y$ (a **v-structure**)
- reason: we would expect some information flow in all three scenarios $X \leftarrow Z \rightarrow Y$, $X \leftarrow Z \leftarrow Y$, and $X \rightarrow Z \rightarrow Y$
- this relies on a causal **faithfulness** assumption: we can measure $(X \perp\!\!\!\perp Y \mid Z)$ iff this is implied by the true causal graph
- example: *swe* – *fin* – *Fennic*, (*swe* $\perp\!\!\!\perp$ *Fennic*), i.e. Finnish not necessary to separate Swedish from Fennic, therefore *swe* $\rightarrow$ *fin* $\leftarrow$ *Fennic*

Propagating Directionality Information

- if all possible common causes are measured, the faithfulness assumption implies we can be sure to have detected exactly the true v-structures
- this provides an inference rule $X \rightarrow Z - Y \Rightarrow X \rightarrow Z \rightarrow Y$
- the PC algorithm uses this rule to **propagate directionality information** through the graph, in many case assigning a direction to each node in the causal skeleton
- example: Glottolog gives us $\mathit{Franconian} \rightarrow \mathit{deu}$, we found it impossible to separate $\mathit{deu} - \mathit{liv}$, but $(\mathit{Franconian} \not\perp\!\!\!\perp \mathit{liv})$ and $(\mathit{Franconian} \perp\!\!\!\perp \mathit{liv} \mid \mathit{deu})$, no v-structure, therefore $\mathit{deu} \rightarrow \mathit{liv}$

Conditional Independence between Languages

- joint information measure for sets of languages $L_1, \ldots, L_n$:

$$R(L_1, \ldots, L_n) := \left| \bigcup_{i=1}^{n} cog(L_i) \right|$$

- from this we get **conditional mutual information between languages** given a set of languages $\mathbf{S} := \{S_1, \ldots, S_n\}$:

$$I(L_i, L_j; \mathbf{S}) := R(L_i, S_1, \ldots, S_n) + R(L_j, S_1, \ldots, S_n) \\ - R(L_i, L_j, S_1, \ldots, S_n) - R(S_1, \ldots, S_n)$$

- R is **submodular**; Steudel et al. (2010) show that checking for non-zero I gives us a consistent conditional independence test
- intuitively: how many cognates between L_i and L_j cannot be explained away by also being cognate to a word in one of the languages in **S**?

Skeleton Inference: Standard PC variants

- testing exponentially many possible sepsets: intractable
- decisive ideas behind **PC algorithm** (Spirtes et al., 2000):
  - ▷ search for minimal separating sets by increasing cardinality
  - ▷ any information flow must involve the remaining neighbors of either node, we only need to consider separating set candidates composed of such neighbors
- **PC*** variant: only build candidate sepsets from neighbors on connecting paths between X and Y

Skeleton Inference: Flow separation criterion

Explicit discrete information units allow us to

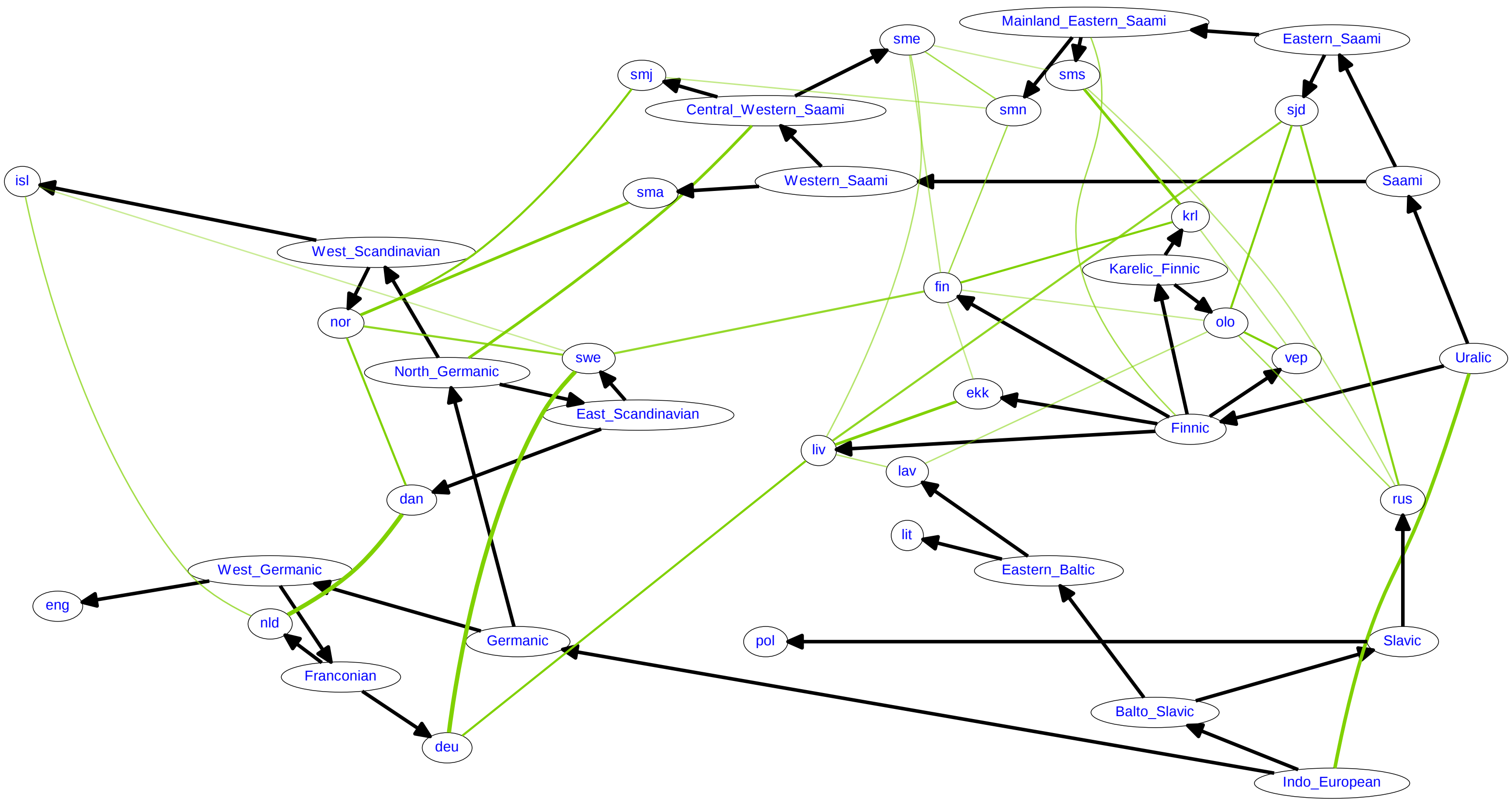
- compose all separating set candidates of connecting paths (not just neighbors, but all nodes on the paths)
- decide for every single shared cognate set whether the sepset includes a path by which the shared material could have traveled

This leads to a **Flow Separation (FS)** criterion:

- separation only occurs if there are is a concrete alternative path for every single cognate shared between X and Y
- some threshold is still necessary in practice to correct for dirty cognacy judgments, and semantic change withering away the traces; 2% in my tests (meaning that contacts which replaced less than 20 out of 1000 words will never appear in the network)

Phylogenetic Lexical Flow Inference: Example

Example result of FS in region around the Baltic Sea:



Directionality Inference: Standard variants

- **PC**: v-structure $X \rightarrow Z \leftarrow Y$ iff Z not needed to separate X, Y, i.e. there is one separating set S with $Z \notin S$
- **Stable PC**: compare how many minimal sepsets contain or do not contain Z, make decision by majority rule
- Despite the name, all PC variants have stability problems! Workaround in Dellert (2016):
  - ▷ aggregate evidence from different unshielded triples into a **Triangle Sum Score (TSS)** measuring the signal on each link
  - ▷ this causes some errors to cancel out, arrows with high aggregate scores are much more reliable
  - ▷ TSS can be used independently of the skeleton, the two inference steps do not depend on each other! (more stability)

Directionality Inference: Unique Flow Ratio (UFR)

New alternative:

- define a score for unshielded triples for making the collider decisions, based on the same intuitions plus a flow criterion
- propagate the decisions by the PC propagation rules

Details of the **Unique Flow Ratio (UFR)** score:

- idea: quantify the notion of "Z needed to remove X — Y"
- let cog_{XYZ} be the cognates shared between between X, Y, Z
- cog_{XYZ*}: the cognates which no path excluding Z could have transported between X and Y (**unique flow**)
- $ufr_1 := \frac{\frac{|cog_{XYZ*}|}{\min(|cog_X|,|cog_Y|,|cog_Z|)}}{\frac{|cog_{XZ}|}{\min(|cog_X|,|cog_Z|)} \cdot \frac{|cog_{YZ}|}{\min(|cog_Y|,|cog_Z|)}}$ ("as much UF as expected?")
- $ufr_2 := cog_{XYZ*} / cog_{XYZ}$ ("how relevant is flow through Z?")
- $ufr := ufr_1 \cdot ufr_2$, v-structures will typically have $ufr < 0.02$

Phylogenetic Lexical Flow Inference: Example

Example result of TSS in region around the Baltic Sea:

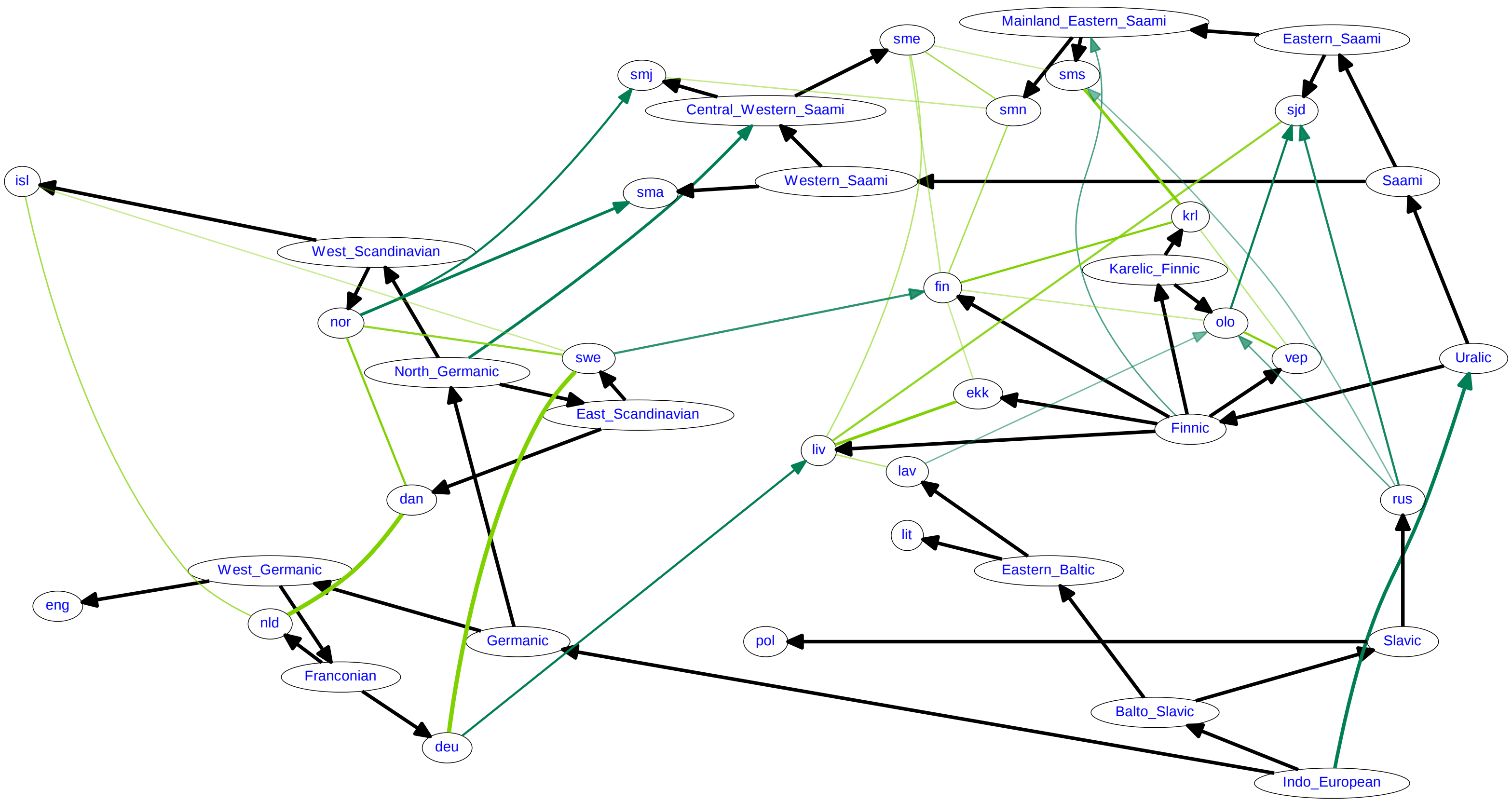


Table of Contents

Generating Testset Data by Simulation

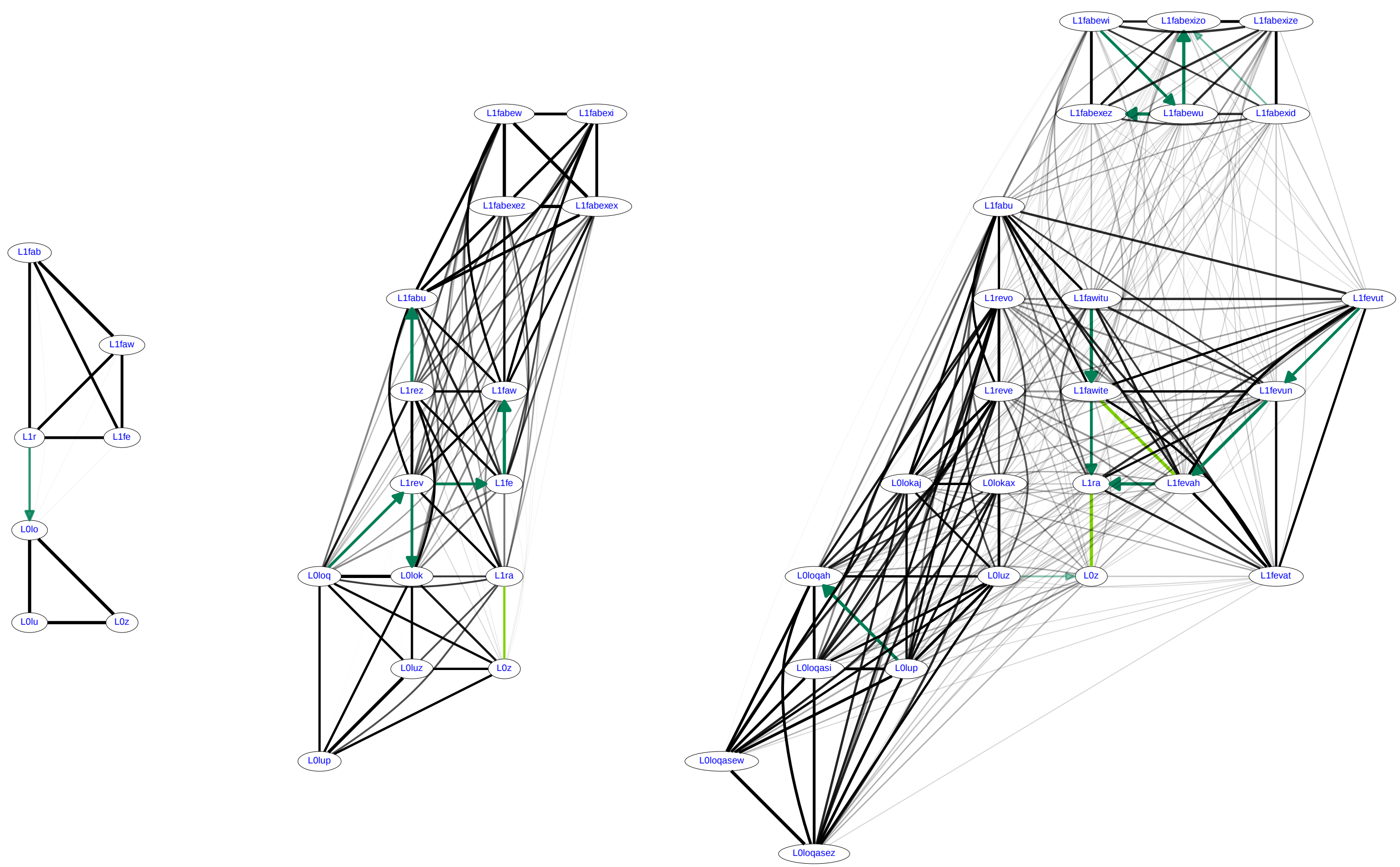
Advantages of using simulations:

- arbitrary amount of test data
- abstract away from problems caused by error-prone cognate detection, tree inference, and ancestral state reconstruction

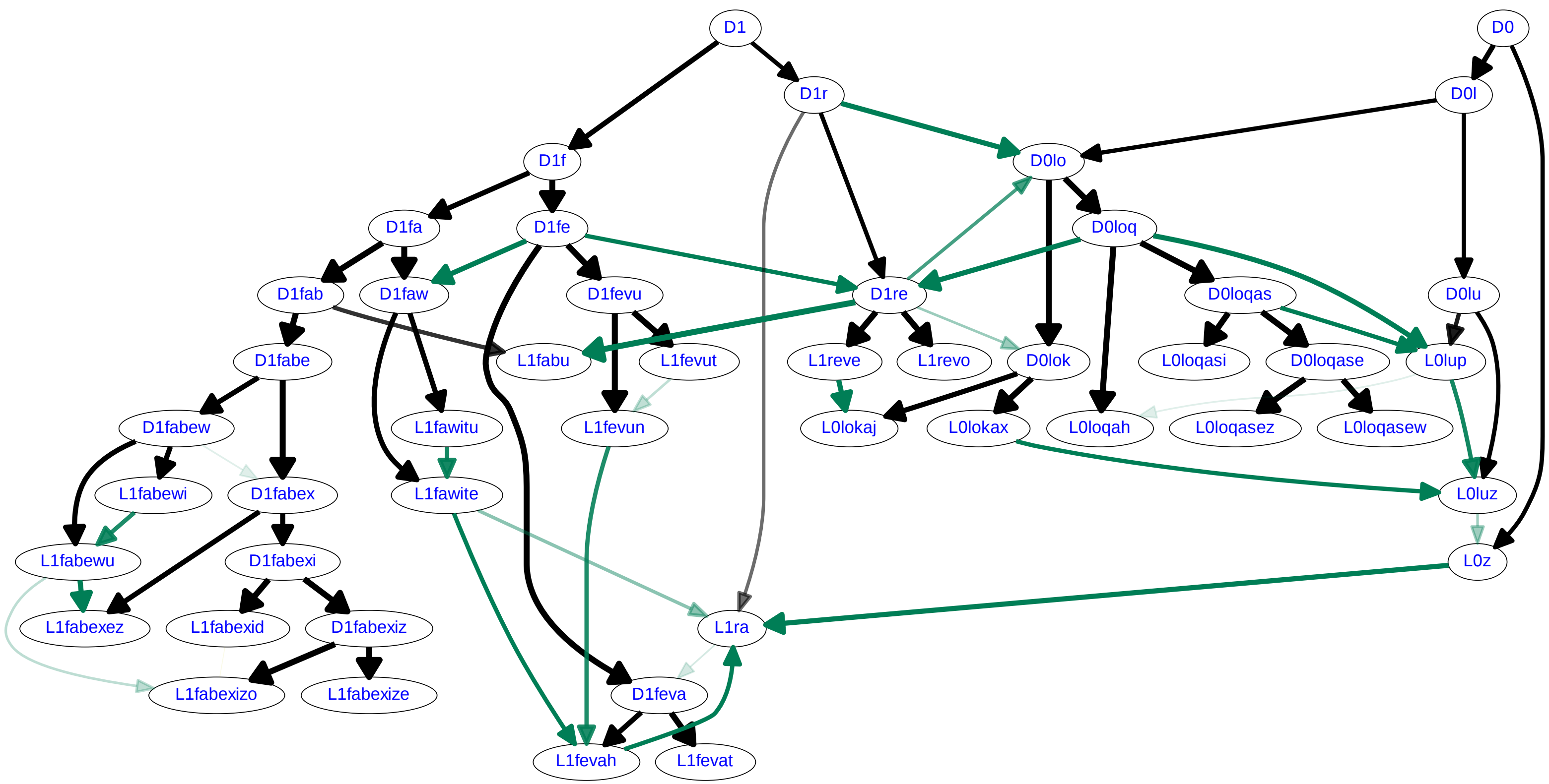
Core design decisions of my simulation model:

- languages split at random intervals, filling a continent
- a language does not become extinct without reason, it only gets replaced if a neighboring language splits into its territory
- we explicitly model lexical replacement in each language (longer splits will lead to less cognate set overlap)
- monodirectional contact channel can open at any time between neighbors, on which cognate IDs are randomly copied over
- every single event modifying the data is tracked, we retain access to complete knowledge

Example: The Simulation Process



Example: A Simulated Flow Network



Skeleton Inference: Evaluation Measures

Evaluation measures can be defined in a very straightforward way:

- **skeleton recall**: which percentage of the lateral connections in the gold standard are also in the inferred skeleton?
- **skeleton precision**: which percentage of the inferred lateral connections are justified by the gold standard?
- **skeleton f-score**: harmonic mean of skeleton precision and recall, i.e. $2 \cdot \frac{SkPr \cdot SkRc}{SkPr + SkRc}$

Skeleton Inference: Comparison on 5 scenarios

	PC	PC*	FS
skeleton recall	0.894	**0.972**	0.897
skeleton precision	0.648	0.687	**0.763**
skeleton f-score	0.752	0.805	**0.825**

- skeletons tend to include almost all relevant lateral connections, but about one fourth of lateral connections are spurious
- clear ranking: PC* better than PC, and FS more precise
- for all the experiments, the flow separation-based skeleton and separating sets will be used

Directionality Inference: Evaluation Measures

Evaluation measures for directionality more difficult to define:

- problem for defining precision and recall:
  we have three options in both the gold standard and the result!
- mapping these to the four basic categories is non-trivial
- my proposal for counting the instances:

	$\rightarrow$ **in result**	$\leftarrow$ **in result**	$\leftrightarrow$ **in result**
$\rightarrow$ **in standard**	*tp* + *tn*	*fp* + *fn*	*tp* + *fp*
$\circ\!\!\rightarrow$ **in standard**	*tp*	*fn*	*tp* + *tp*
$\leftrightarrow$ **in standard**	*tp* + *fn*	*tp* + *fn*	*tp* + *tp*

- **arrow recall**: $tp/(tp + fn)$, as usual
- **arrow precision**: $tp/(tp + fp)$, as usual
- **arrow f-score**: harmonic mean of arrow precision and recall, i.e. $2 \cdot \frac{ArPr \cdot ArRc}{ArPr + ArRc}$

Directionality Inference: Comparison on 5 scenarios

Comparison on the best skeleton (derived by FS):

	PC	Stable PC	UFR	TSS
arrow recall	0.758	**0.805**	0.798	0.637
arrow precision	0.878	0.854	0.866	**0.909**
arrow f-score	0.814	0.829	**0.831**	0.749

- directionality inference on the true arcs is quite satisfactory
- clearly the worst method: triangle score sum,
  though the fewer arrows it infers are quite reliable
- vanilla PC quite reasonable, not much worse than best variants
- stable PC and UFR best, very comparable in performance

Acknowledgments

- Gerhard Jäger (supervision)
- Igor Yanovich (detailed discussions)
- all other members of the EVOLAEMP team (feedback at many stages)
- the ERC (Advanced Grant 324246)

References

Dellert, J. (2016). Uralic and its Neighbors as a Test Case for a Lexical Flow Model of Language Contact. Second International Workshop on Computational Linguistics for Uralic Languages.

Morrison, D. A. (2011). *An introduction to phylogenetic networks*. RJR Productions.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press, 2nd edition.

Steudel, B., Janzing, D., and Schölkopf, B. (2010). Causal markov condition for submodular information measures. In Kalai, A. and Mohri, M., editors, *Proceedings of the 23rd Annual Conference on Learning Theory*, pages 464–476, Madison, WI, USA. OmniPress.

Wen, D., Yu, Y., and Nakhleh, L. (2016). Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genet*, 12(5):e1006006.

Directionality Inference: Triangle Score Sum (TSS)

Details of the **Triangle Score Sum (TSS)** score:

- consider each unshielded triple $l_1 \rightarrow l_2 \leftarrow l_3$
- define $w(l_1 \rightarrow l_2; l_3) := \frac{|cog(l_1) \cap cog(l_2)| \cdot |cog(l_2) \cap cog(l_3)|}{|cog(l_2)|}$, i.e. the cognate overlap between l_1 and l_3 we would have expected if the true pattern had been $l_1 \leftarrow l_2 \rightarrow l_3$ or $l_1 \leftarrow l_2 \leftarrow l_3$
- aggregate from all triples into $sc(l_1 \rightarrow l_2) := \sum_{l_3} w(l_1 \rightarrow l_2; l_3)$, use threshold on $sc(l_1 \rightarrow l_2)/sc(l_2 \rightarrow l_1)$ to make decision