Automatic Detection of Borrowings in Lexicostatistic Datasets

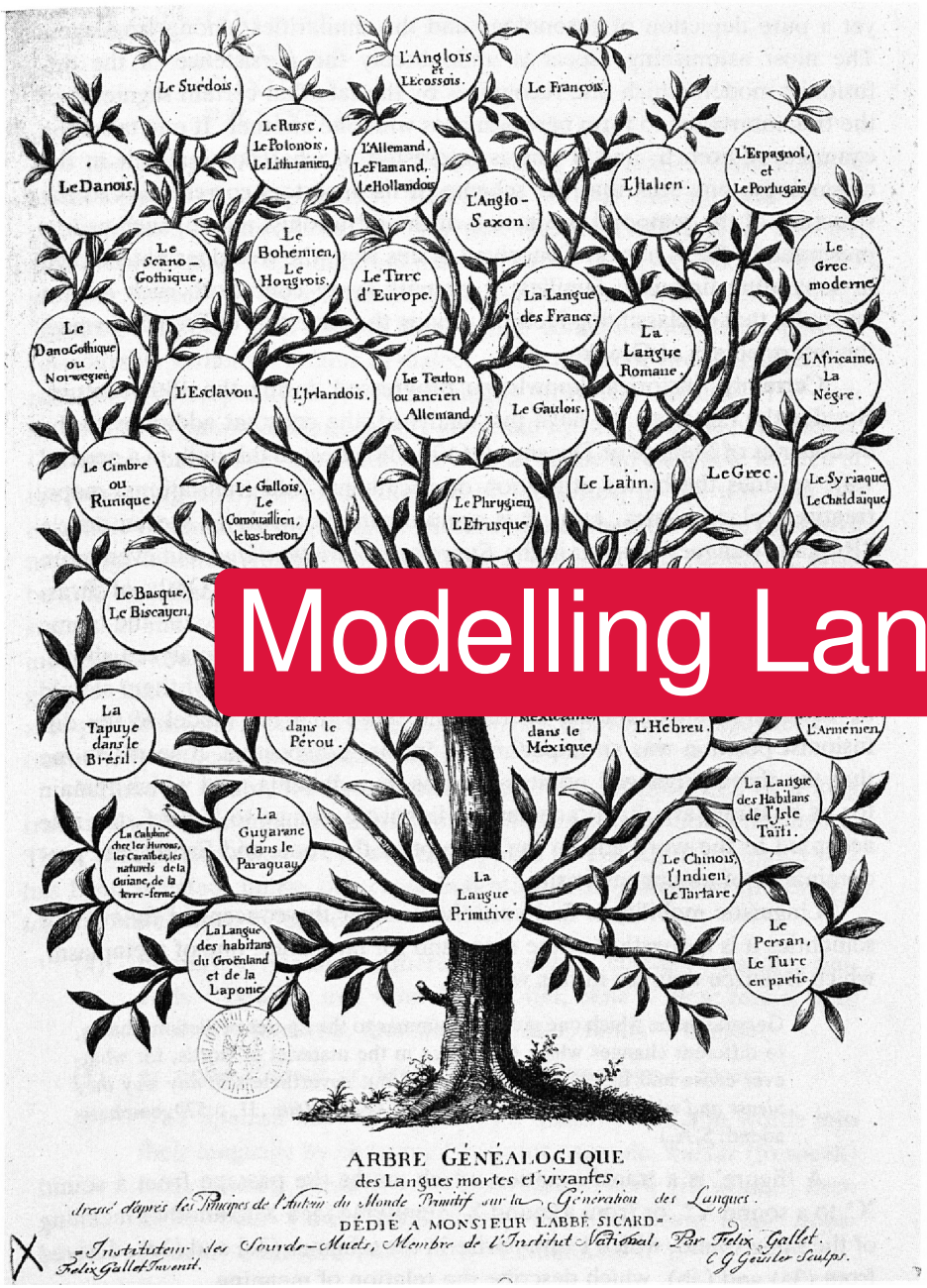
Johann-Mattis List[1], Steven Moran[1,2] & Jelena Prokić[1]

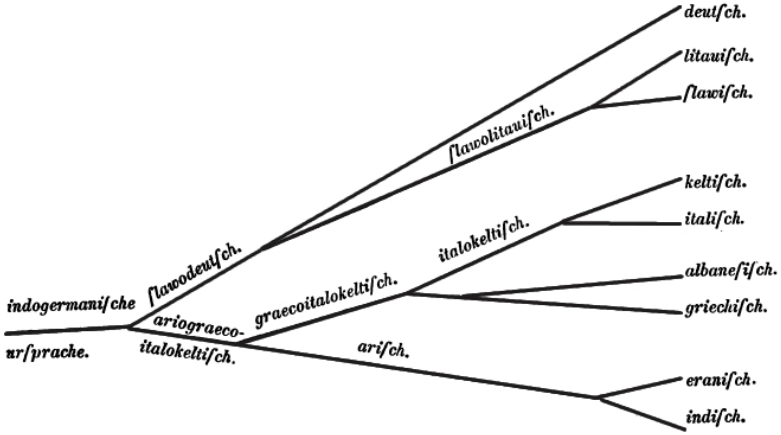
[1]Research Unit *Quantitative Language Comparison*
Philipps-University Marburg
[2]Linguistics Department, University of Zürich

December 13, 2012

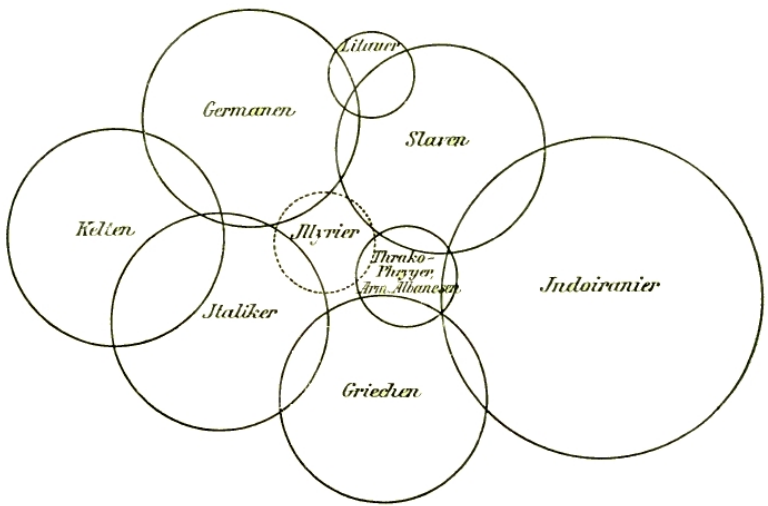
Structure of the Talk

1. Modelling Language History
   - Trees
   - Waves
   - Networks
2. Borrowing
   - Complexity of Borrowing Processes
   - Phyletic Patterns
   - Borrowing Detection
3. Application
   - Material
   - Methods
   - Results
4. Discussion
   - Natural Findings or Artifacts?
   - Limits
   - Examples





Modelling Language History



Dendrophilia

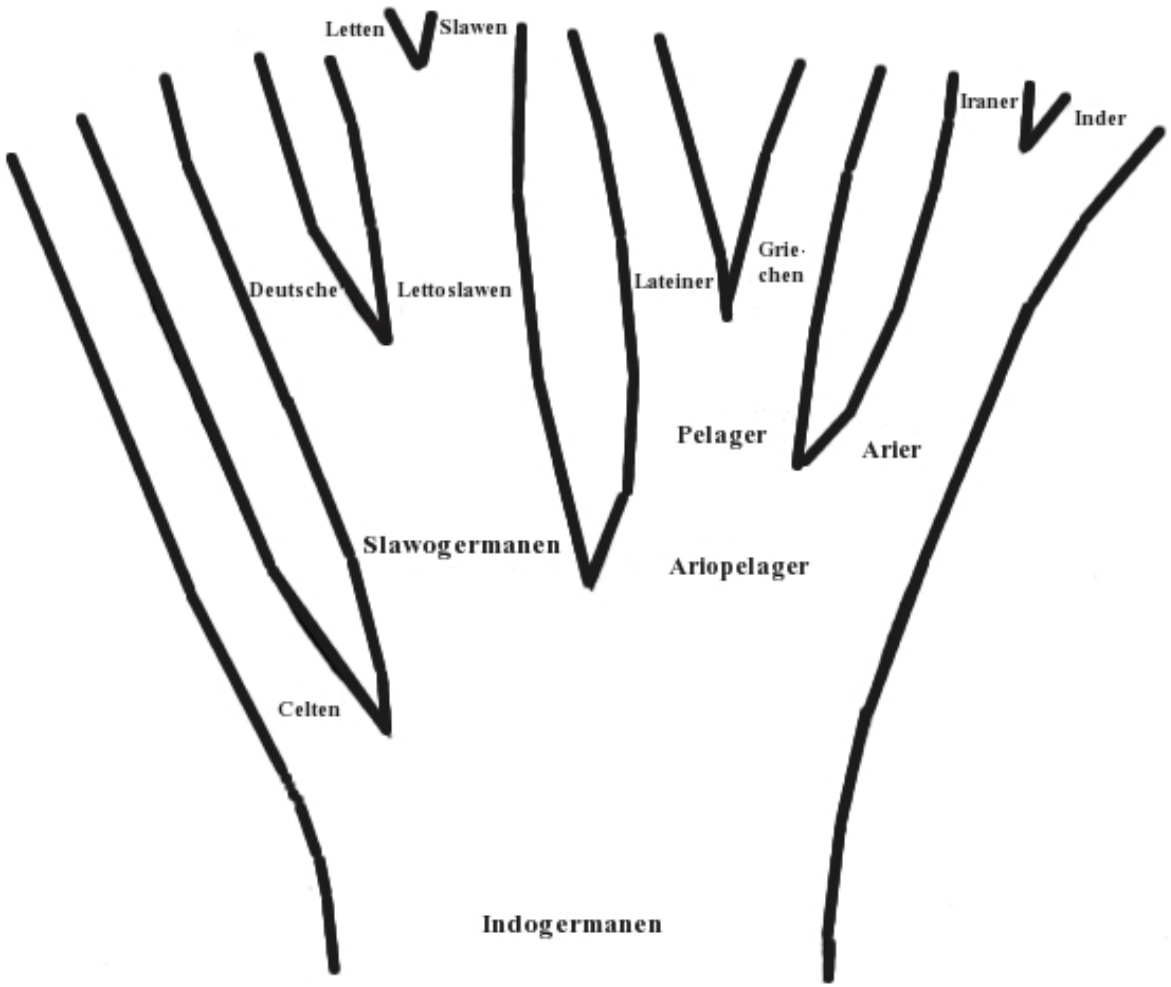
August Schleicher
(1821-1868)

Dendrophilia

These assumptions that logically follow from the results of our research can be best illustrated with help of a branching tree. (Schleicher 1853: 787, translation JML)

August Schleicher
(1821-1868)

Dendrophilia



Schleicher (1853)

Dendrophobia

Johannes Schmidt
(1843-1901)

Dendrophobia

No matter how we look at it, as long as we stick to the assumption that today's languages originated from their common proto-language via multiple furcation, we will never be able to explain all facts in a scientifically adequate way. (Schmidt 1872: 17, translation JML)

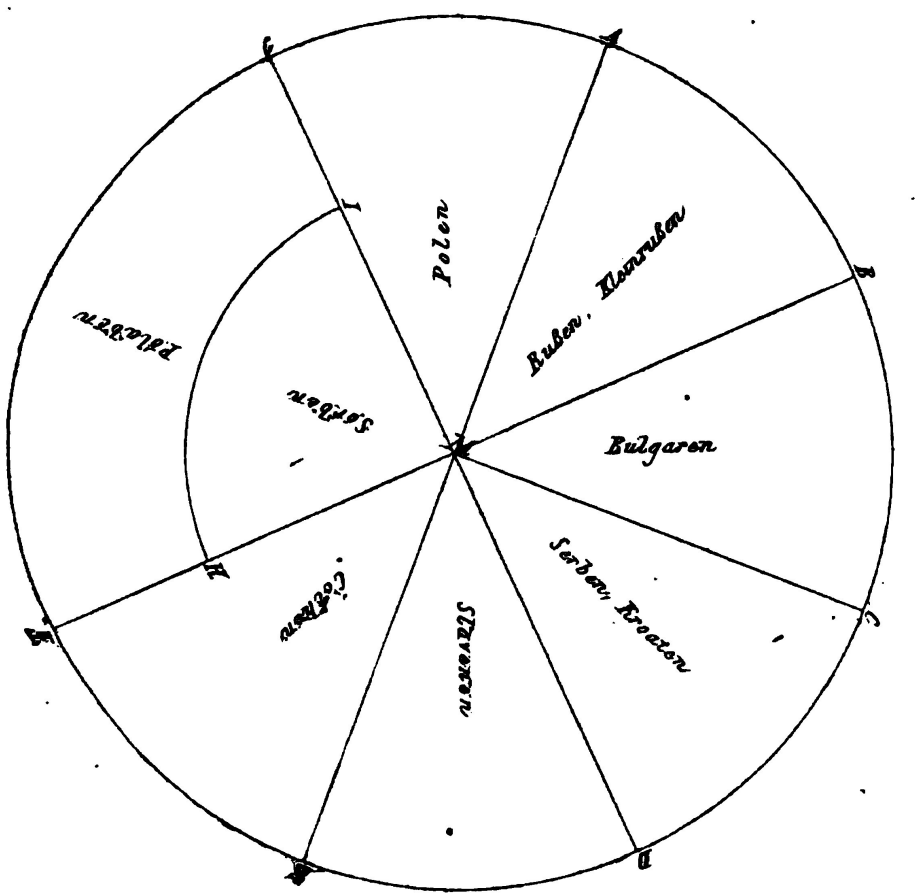
Johannes Schmidt
(1843-1901)

Dendrophobia

I want to replace [the tree] *by the image of a wave that spreads out from the center in concentric circles becoming weaker and weaker the farther they get away from the center.*
(Schmidt 1872: 27, translation JML)

Johannes Schmidt
(1843-1901)

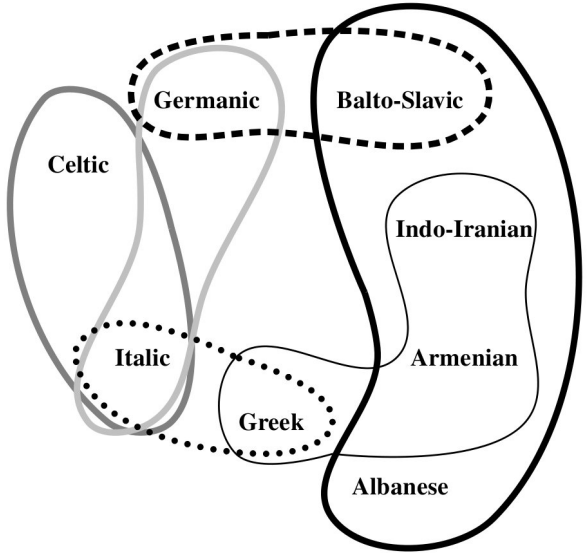
Dendrophobia



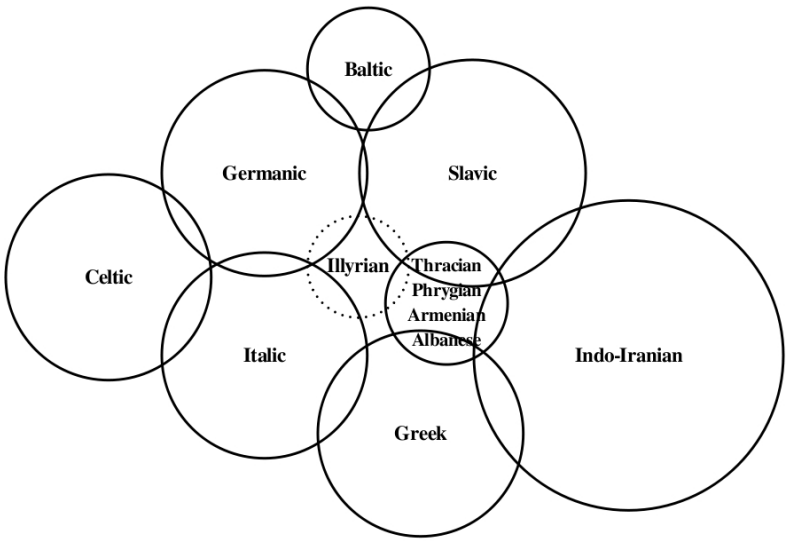
Schmidt (1875)

Dendrophobia

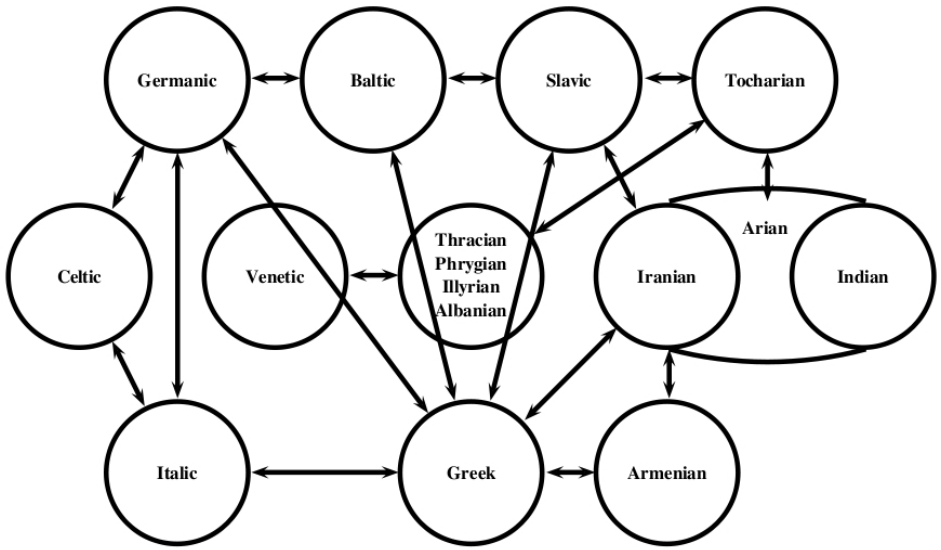
Meillet (1908)



Bloomfield (1933)



Hirt (1905)



Bonfante (1931)

Phylogenetic Networks

Trees are bad because

Phylogenetic Networks

Trees are bad because

- they are difficult to reconstruct

Phylogenetic Networks

Trees are bad because

- they are difficult to reconstruct
- languages do not separate in split processes

Phylogenetic Networks

Trees are bad because

- they are difficult to reconstruct
- languages do not separate in split processes
- they are boring, since they only capture certain aspects of language history, namely the vertical relations

Phylogenetic Networks

Trees are bad because

- they are difficult to reconstruct
- languages do not separate in split processes
- they are boring, since they only capture certain aspects of language history, namely the vertical relations

Waves are bad because

- nobody knows how to reconstruct them

Phylogenetic Networks

Trees are bad because

- they are difficult to reconstruct
- languages do not separate in split processes
- they are boring, since they only capture certain aspects of language history, namely the vertical relations

Waves are bad because

- nobody knows how to reconstruct them
- languages still separate, even if not in split processes

Phylogenetic Networks

Trees are bad because

- they are difficult to reconstruct
- languages do not separate in split processes
- they are boring, since they only capture certain aspects of language history, namely the vertical relations

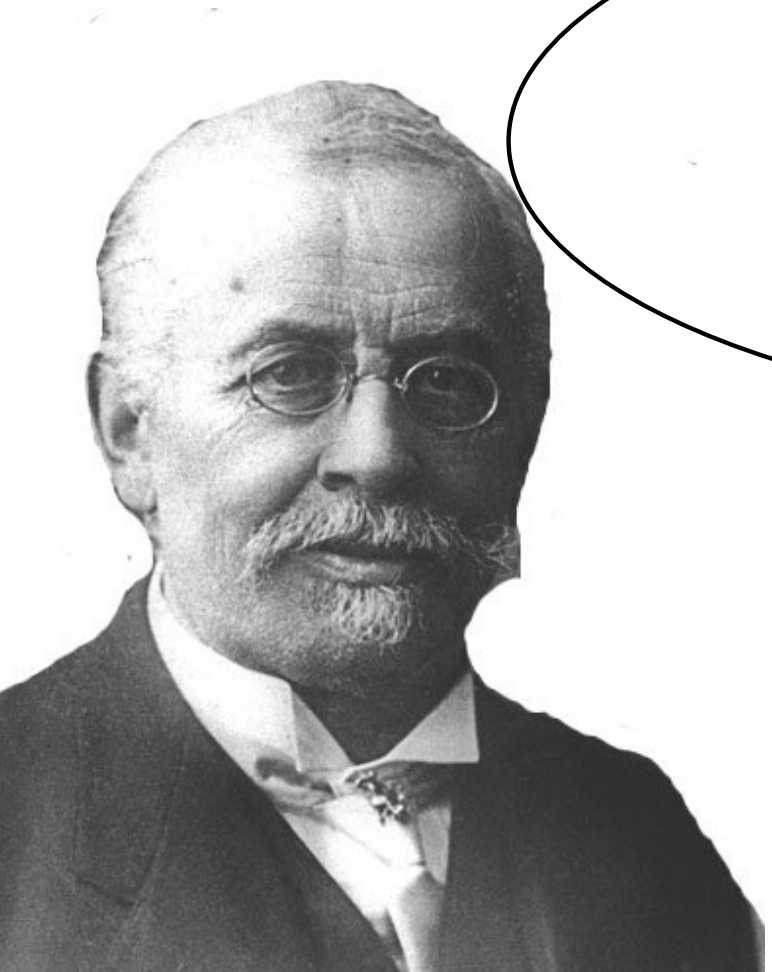
Waves are bad because

- nobody knows how to reconstruct them
- languages still separate, even if not in split processes
- they are boring, since they only capture certain aspects of language history, namely, the horizontal relations

Phylogenetic Networks

Hugo Schuchardt
(1842-1927)

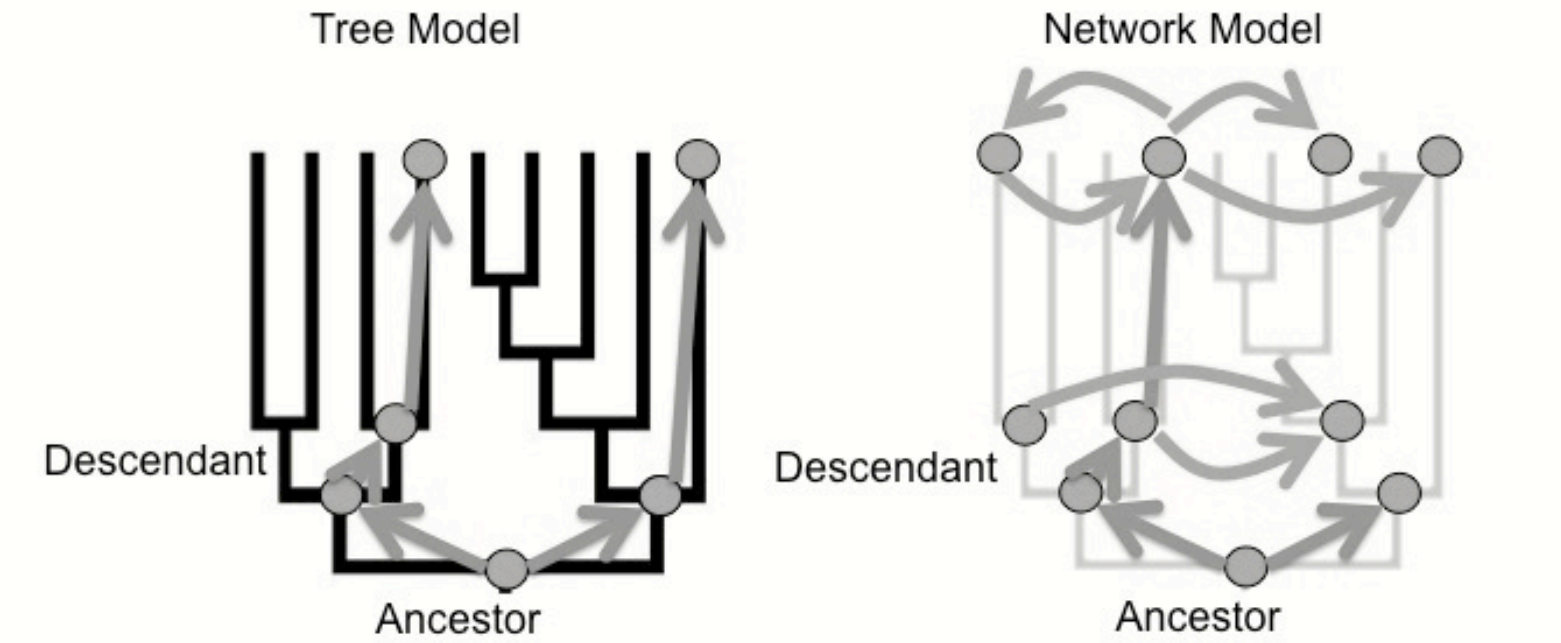
Phylogenetic Networks



We connect the branches and twigs of the tree with countless horizontal lines and it ceases to be a tree (Schuchardt 1870 [1900]: 11, translation JML)

Hugo Schuchardt (1842-1927)

Phylogenetic Networks



Borrowing

Complexity of Borrowing Processes

Complexity of Borrowing Processes

expected Mandarin $[\mathrm{ma}_{55}\mathrm{po}_{21}\mathrm{lou}]$

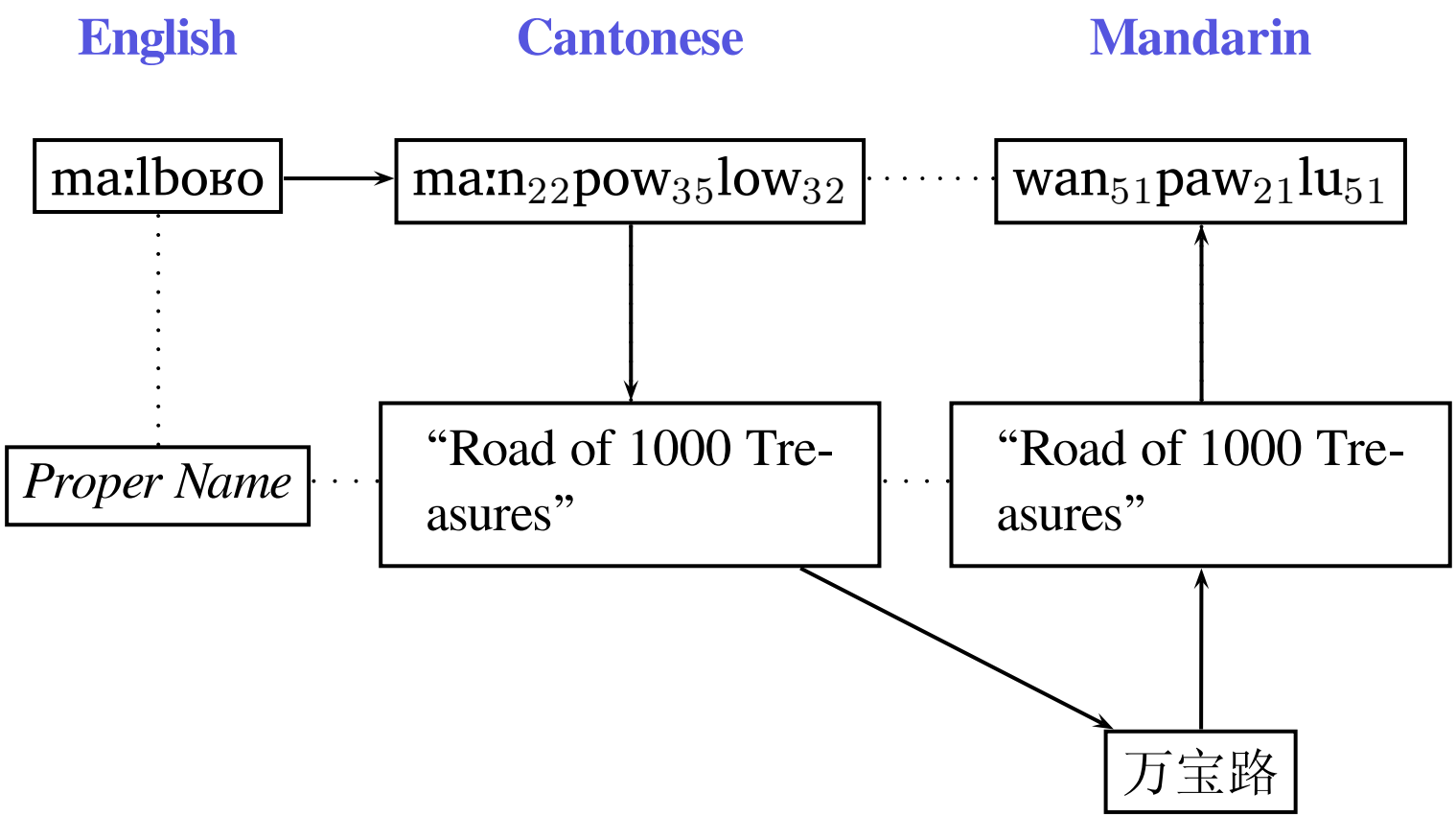
Complexity of Borrowing Processes

expected	Mandarin	$[\mathrm{ma}_{55}\mathrm{po}_{21}\mathrm{lou}]$
attested	Mandarin	$[\mathrm{wan}_{51}\mathrm{paw}_{21}\mathrm{lu}_{51}]$

Complexity of Borrowing Processes

expected	Mandarin	[$ma_{55}po_{21}lou$]
attested	Mandarin	[$wan_{51}paw_{21}lu_{51}$]
explanation	Cantonese	[$ma{:}n_{22}pow_{35}low_{32}$]

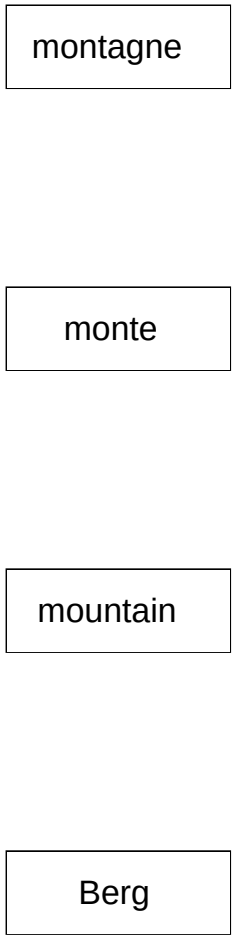
Complexity of Borrowing Processes



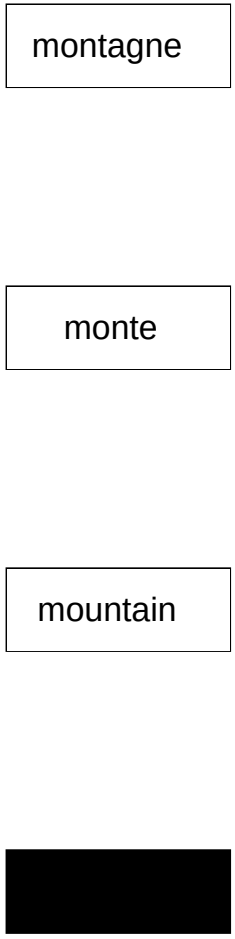
Patchy Distributions in Phyletic Patterns

Borrowing processes can be incredibly complex. Nevertheless, they always leave direct traces, in so far as the borrowed word is usually phonetically quite similar to the donor word. Furthermore, since the borrowing process is not tree-like, borrowings may – if they are mistaken for cognates – show up as “patchy distributions” in phyletic patterns of genetically related languages.

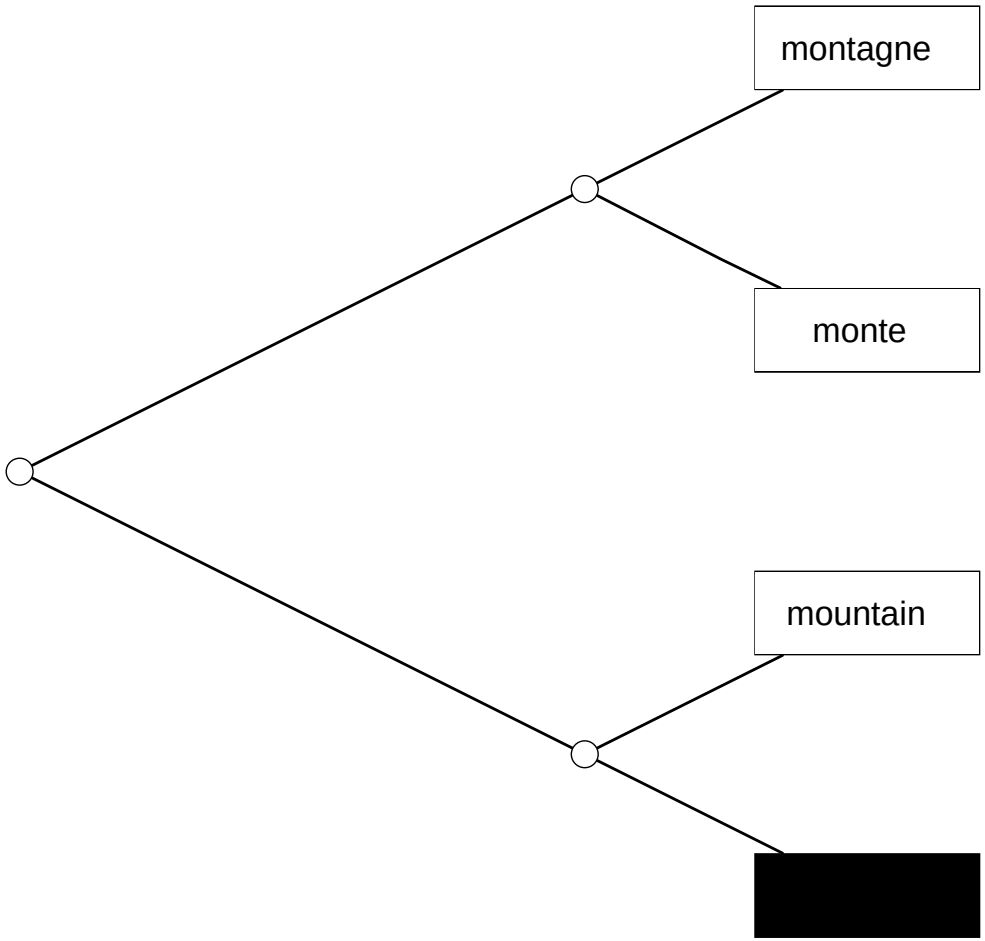
Patchy Distributions in Phyletic Patterns



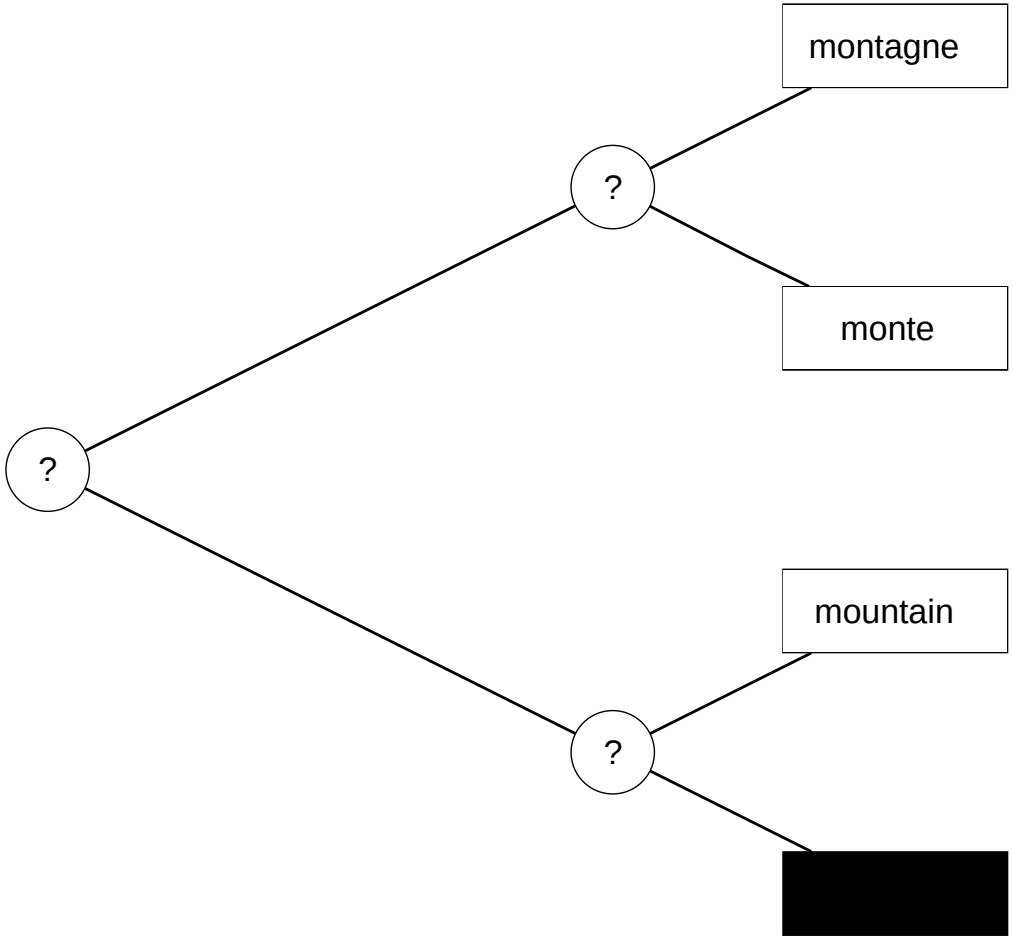
Patchy Distributions in Phyletic Patterns



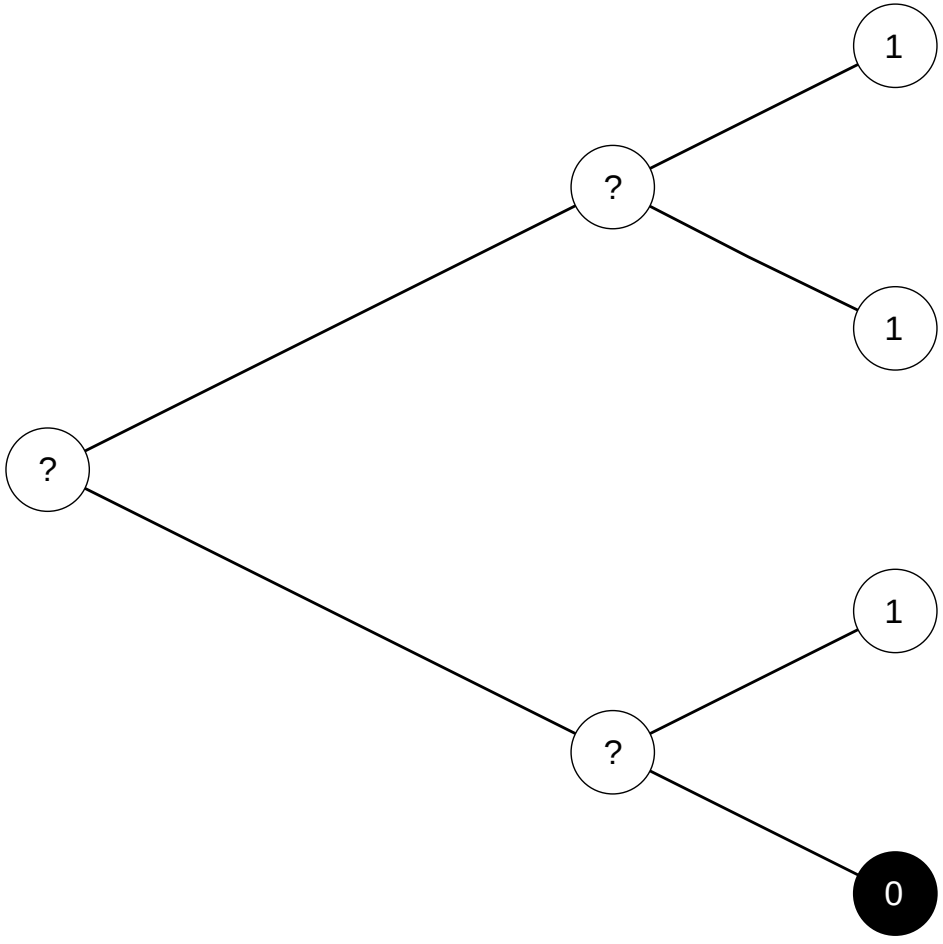
Patchy Distributions in Phyletic Patterns



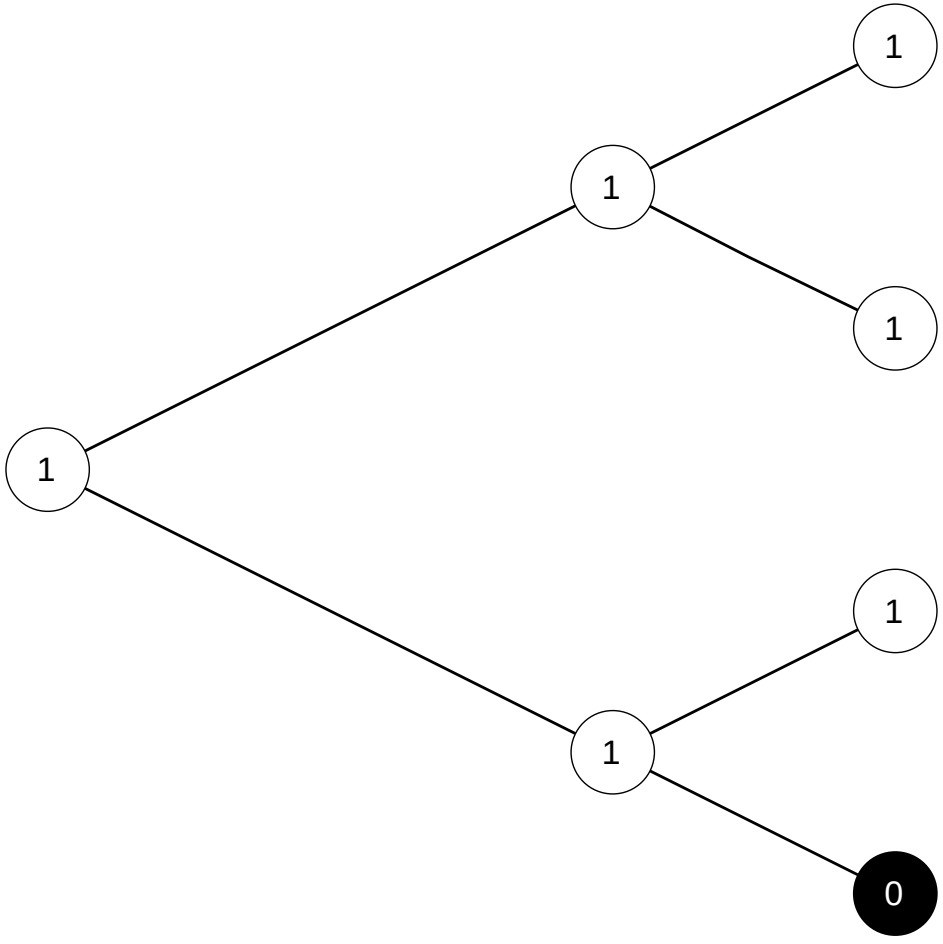
Patchy Distributions in Phyletic Patterns



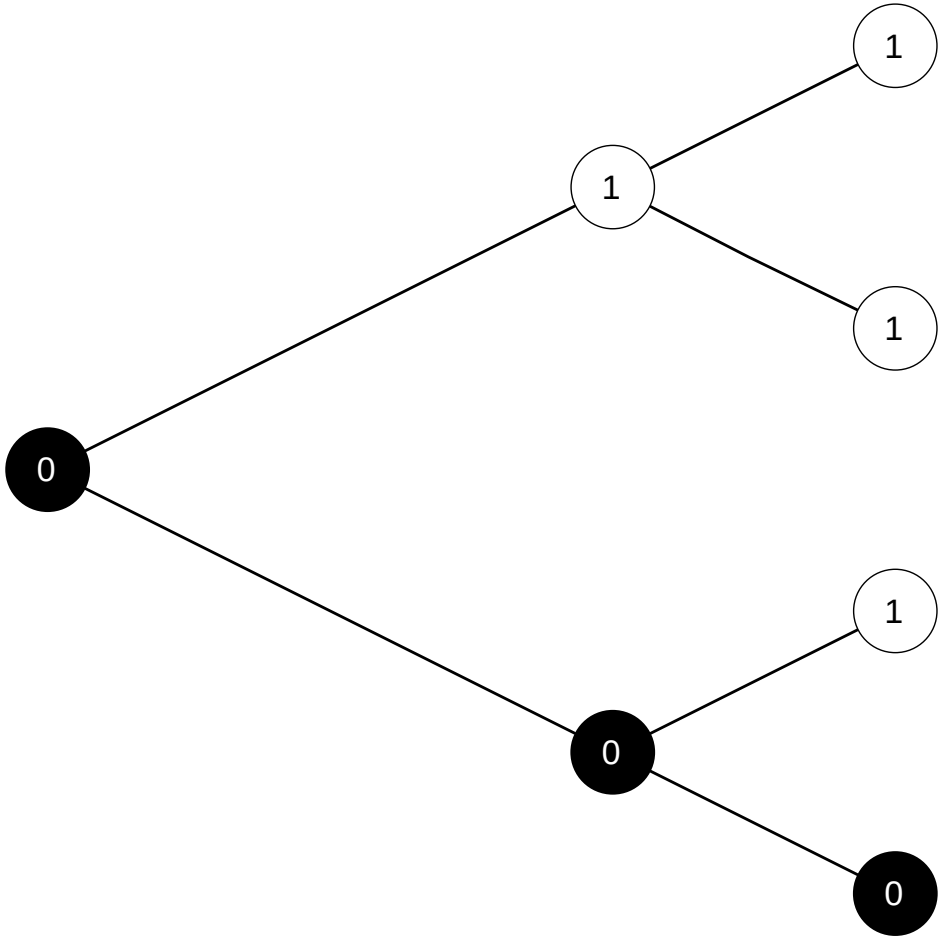
Patchy Distributions in Phyletic Patterns



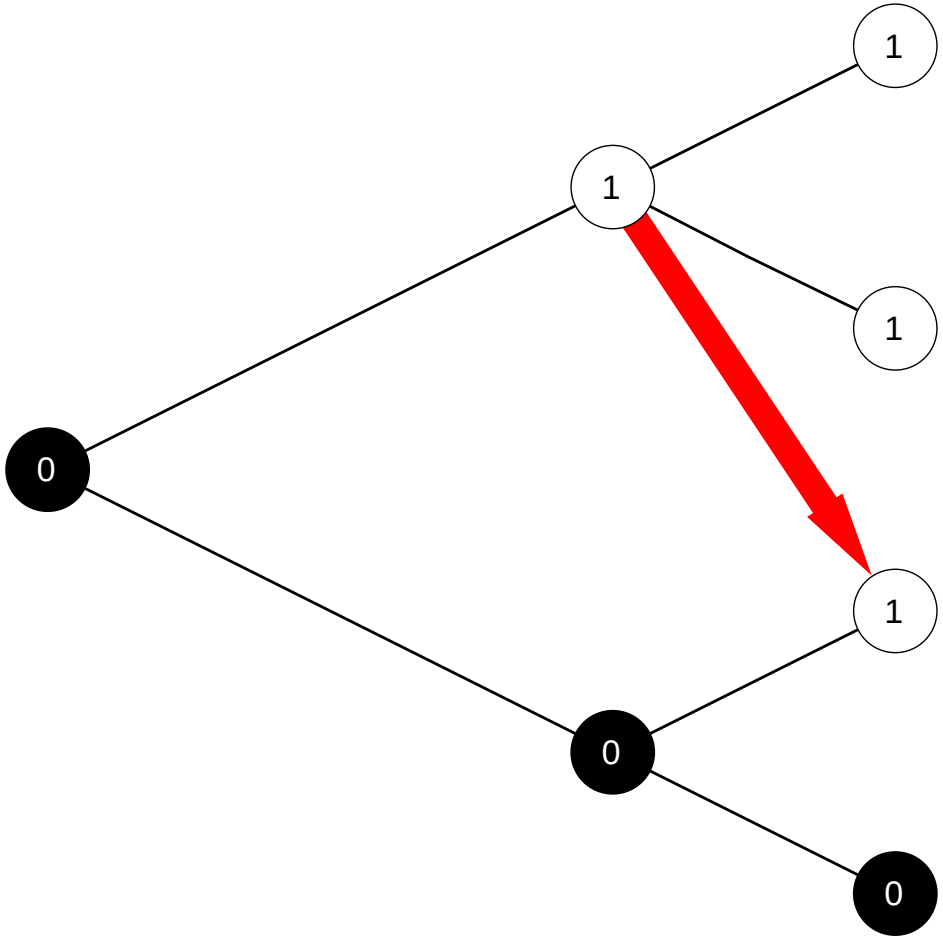
Patchy Distributions in Phyletic Patterns



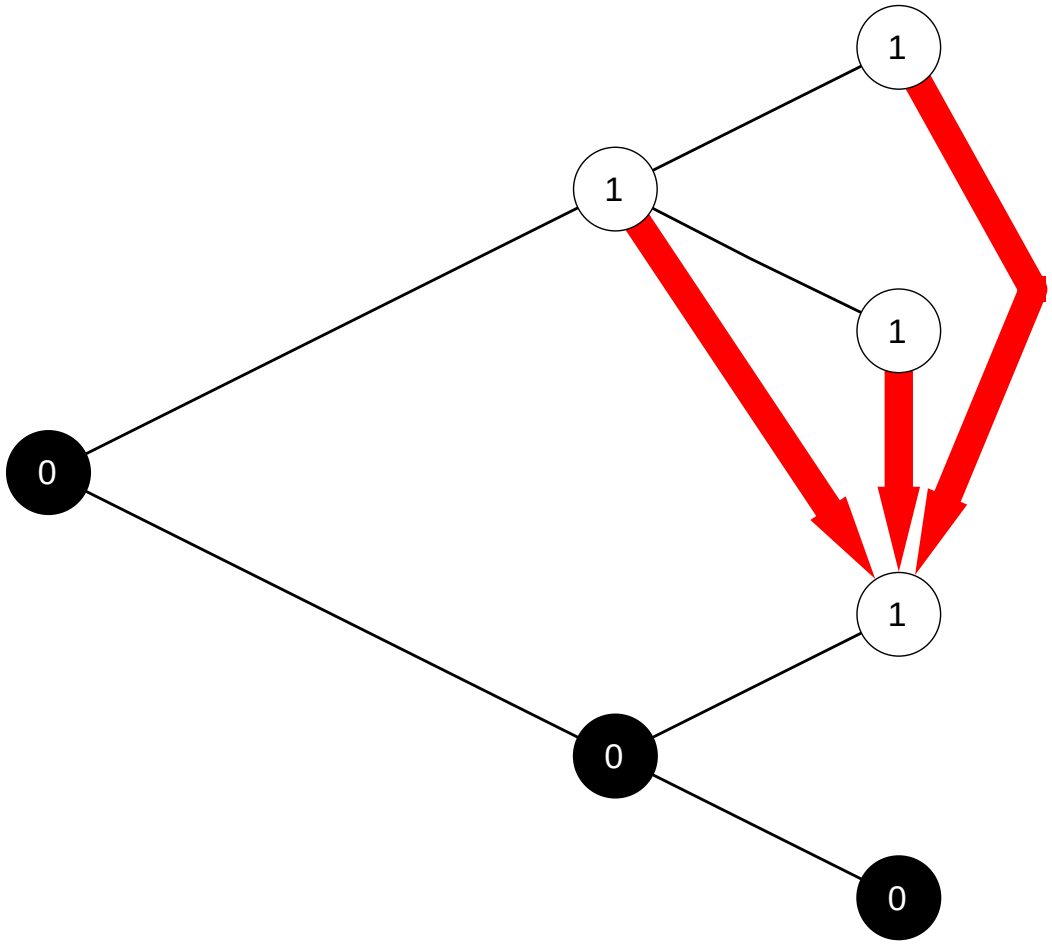
Patchy Distributions in Phyletic Patterns



Patchy Distributions in Phyletic Patterns



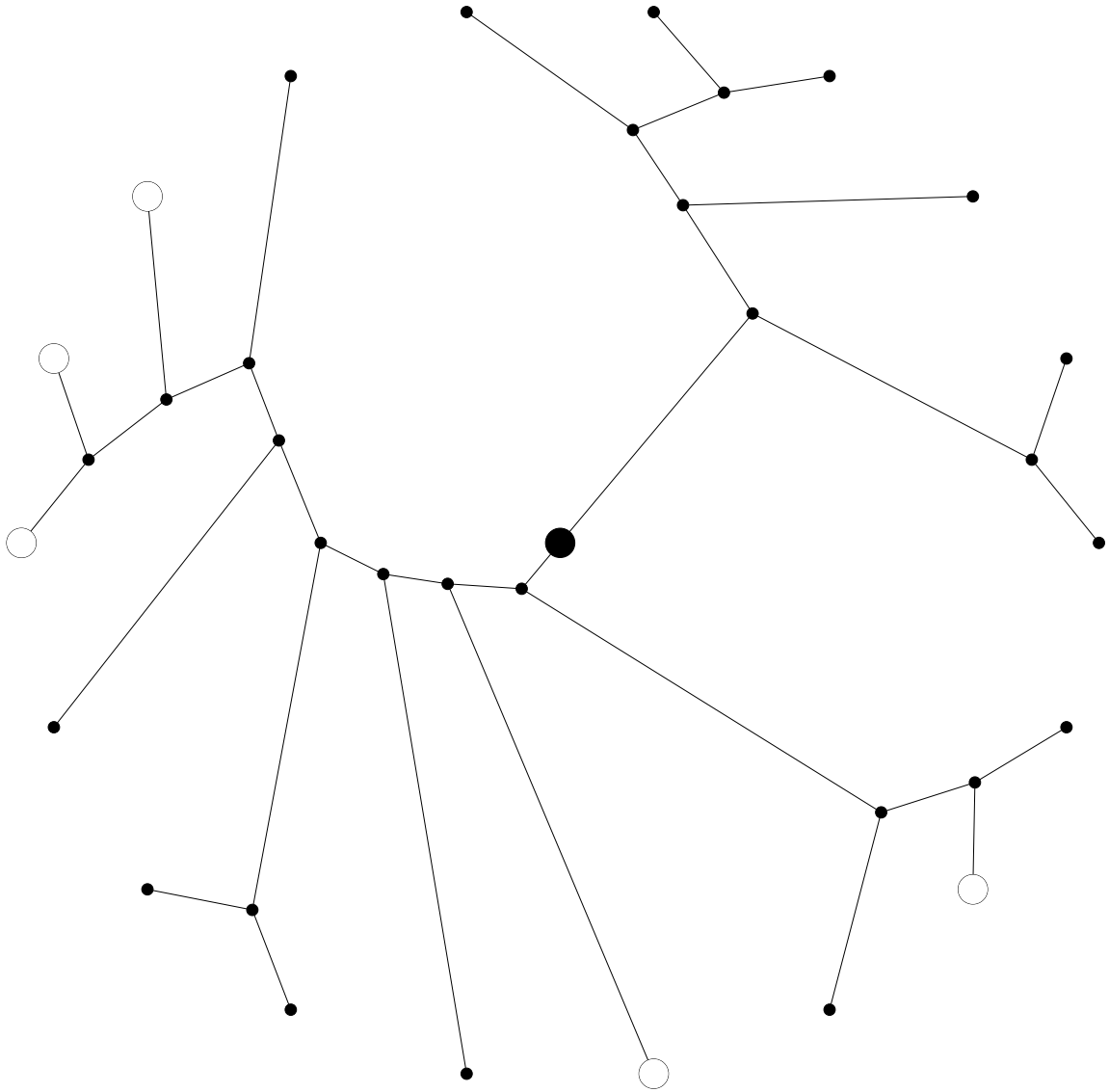
Patchy Distributions in Phyletic Patterns



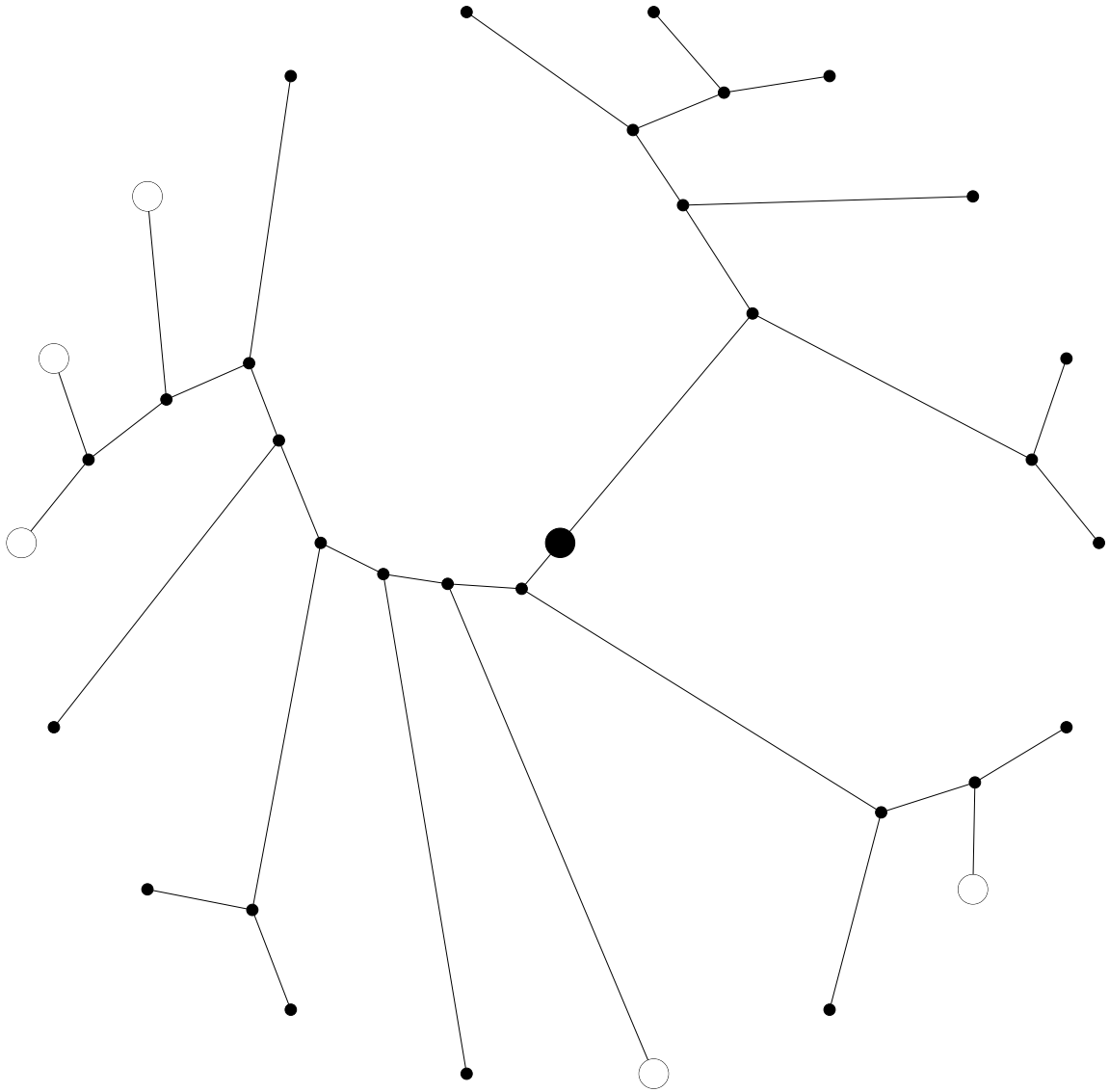
Gain Loss Mapping

Patchy distributions in phyletic patterns can serve as a heuristic for borrowing detection. Patchily distributed cognates can be identified with help of *gain loss mapping approaches* (Mirkin et al. 2003, Dagan & Martin 2007, Cohen et al. 2008) by which *phyletic patterns* are plotted to a *reference tree*.

Gain Loss Mapping



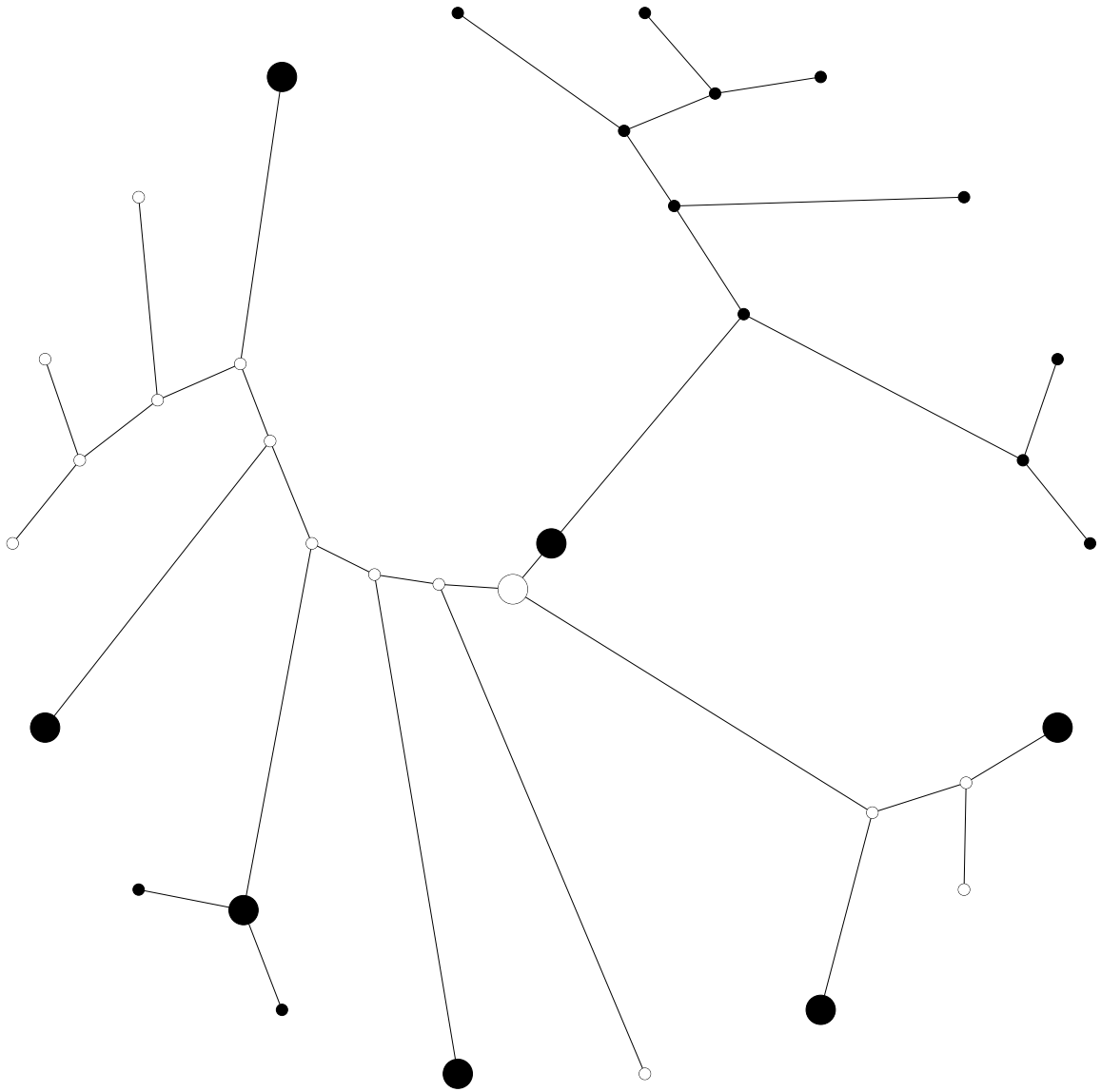
Gain Loss Mapping



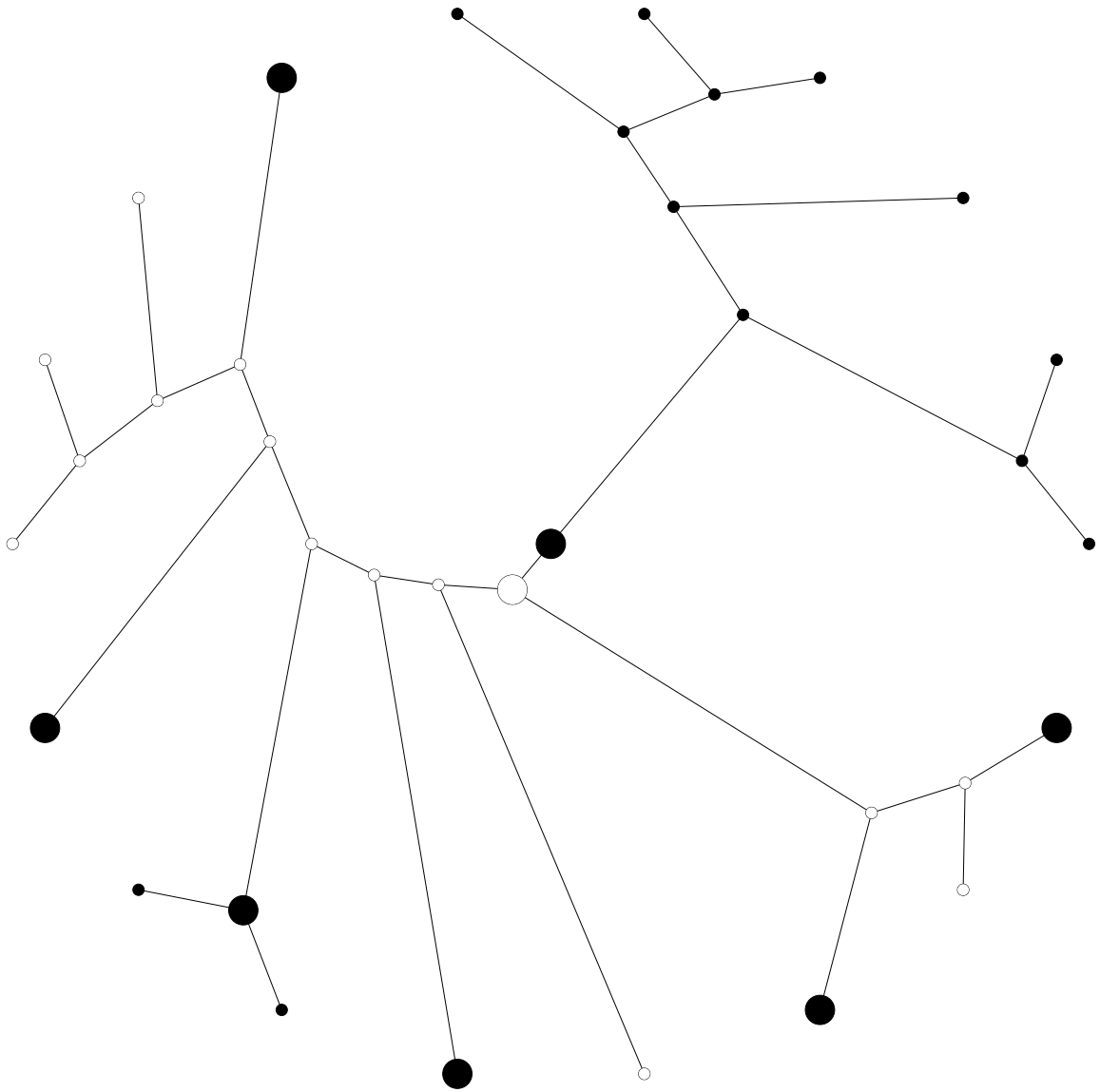
5 gains

0 losses

Gain Loss Mapping



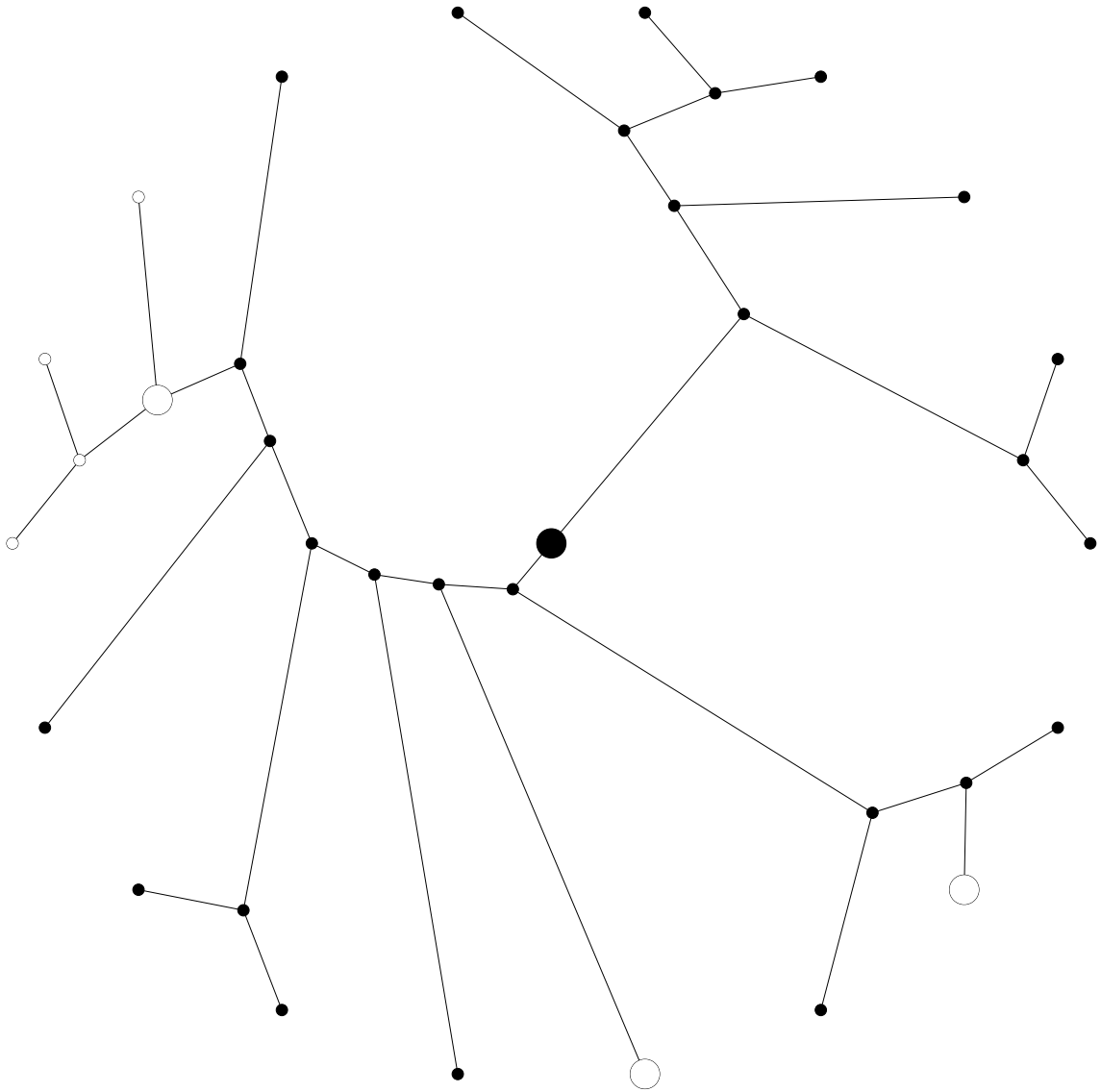
Gain Loss Mapping



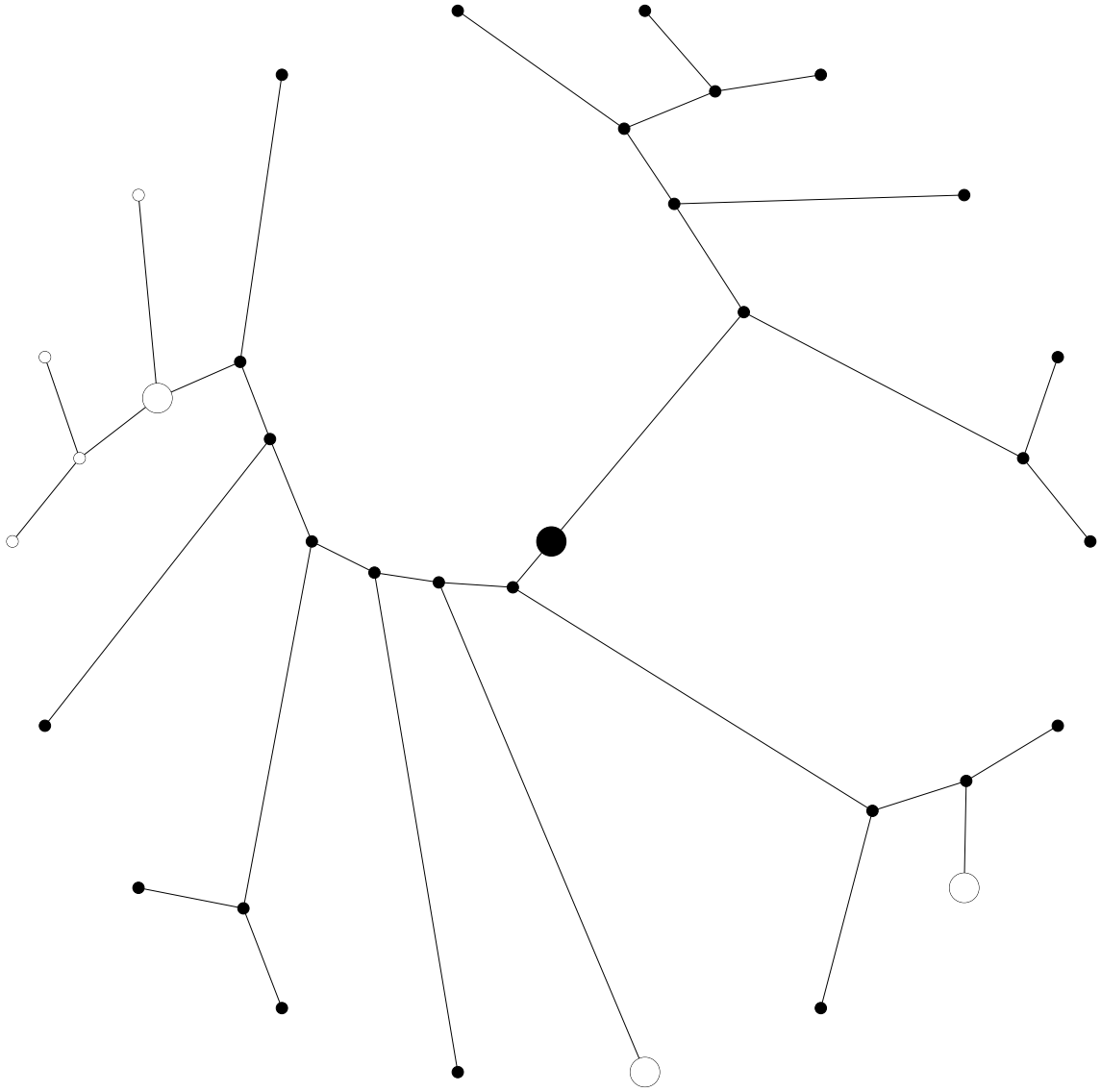
1 gain

6 losses

Gain Loss Mapping



Gain Loss Mapping



3 gains

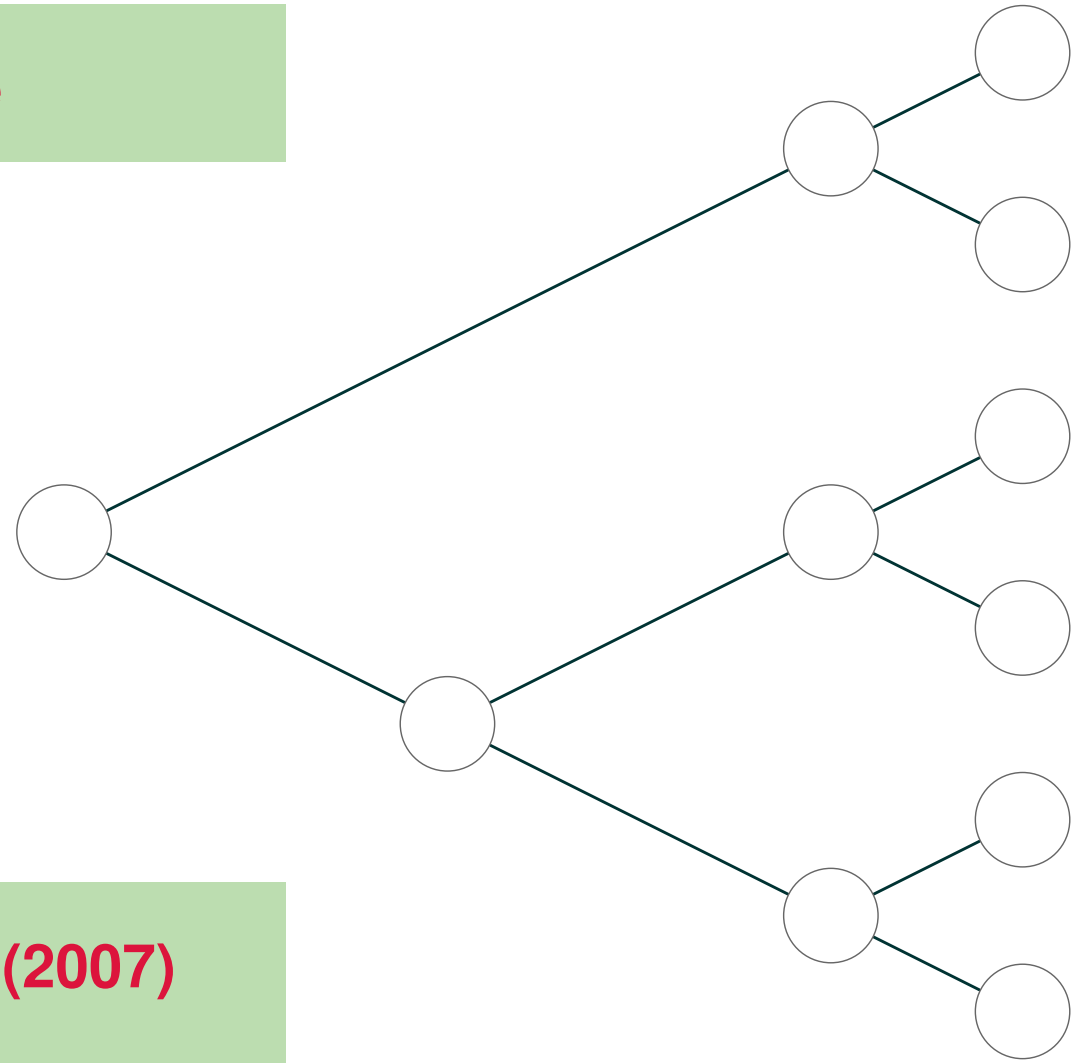
0 losses

Ancestral Vocabulary Distributions

Gain loss mapping is useful to test possible scenarios of character evolution. However, as long as there is no direct criterion that helps to choose the "best" of many different solutions, the method hardly gives us any new insights. Nelson-Sathi et al. (2011) use *ancestral vocabulary sizes* as a criterion to determine the right model. Here, we introduce *ancestral vocabulary distributions*, i.e. the form-meaning ratio of ancestral taxa, as a new criterion.

Ancestral Vocabulary Distributions

Vocabulary Size

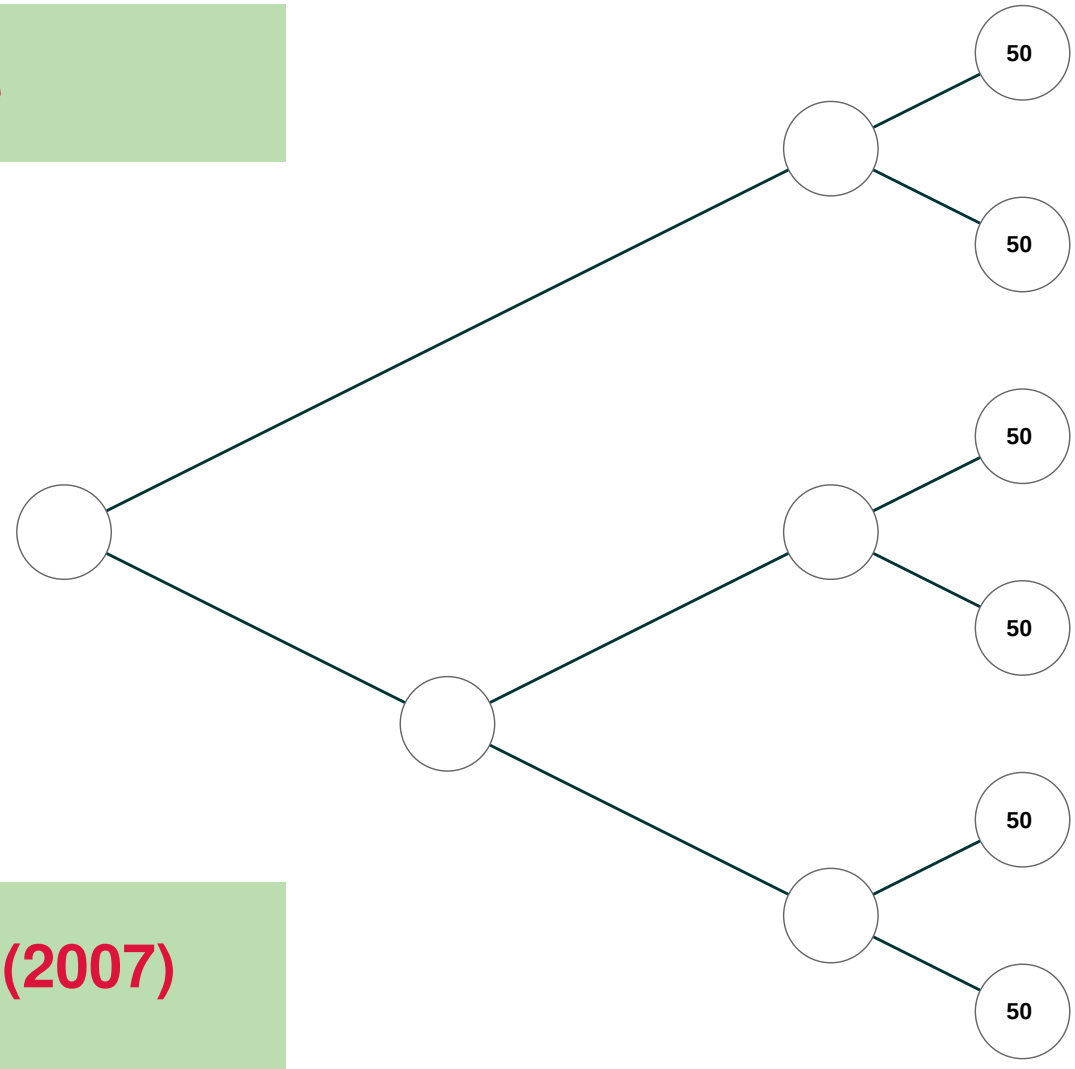


Dagan & Martin (2007)

Nelson-Sathi et al. (2011)

Ancestral Vocabulary Distributions

Vocabulary Size

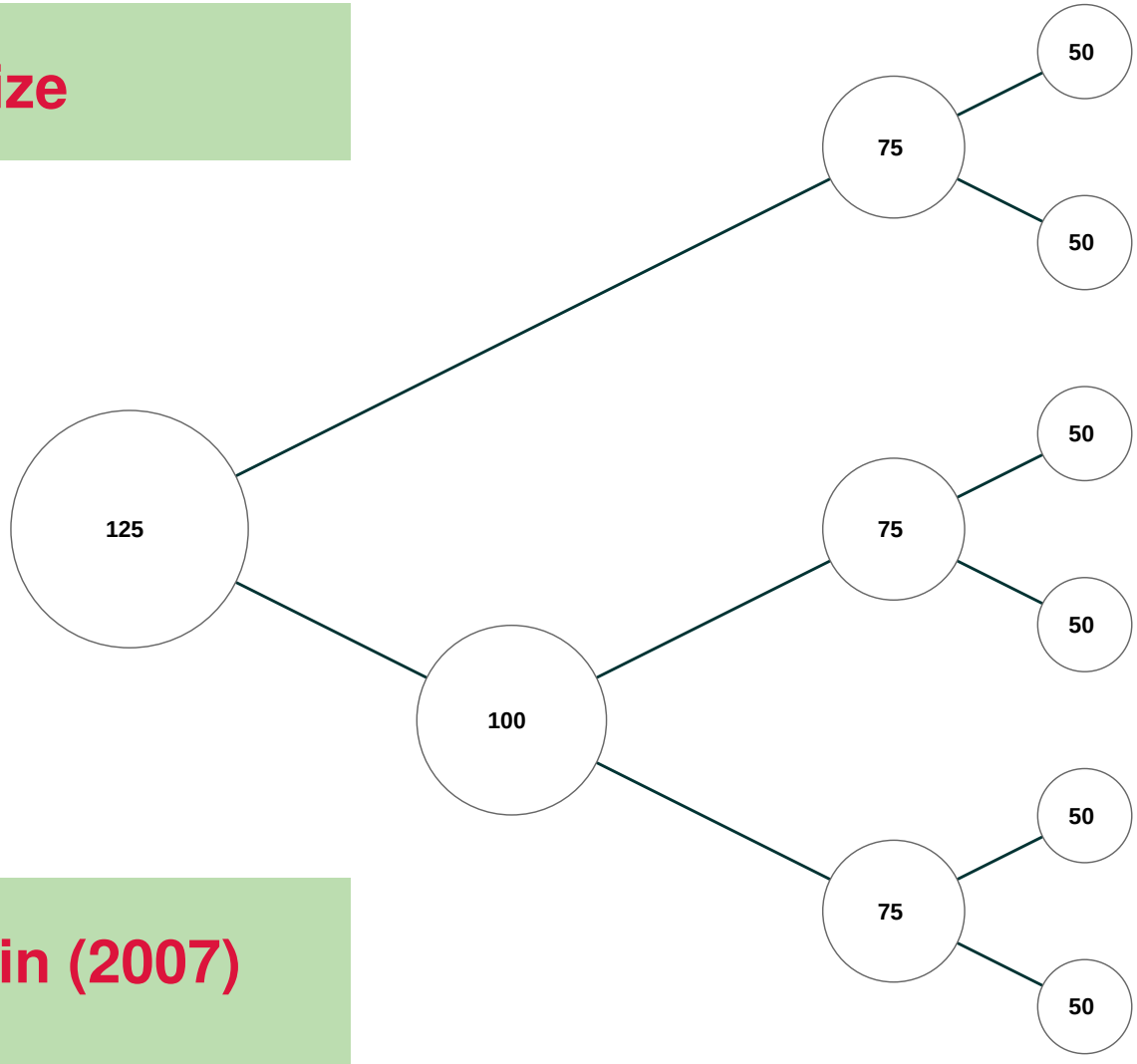


Dagan & Martin (2007)

Nelson-Sathi et al. (2011)

Ancestral Vocabulary Distributions

Vocabulary Size

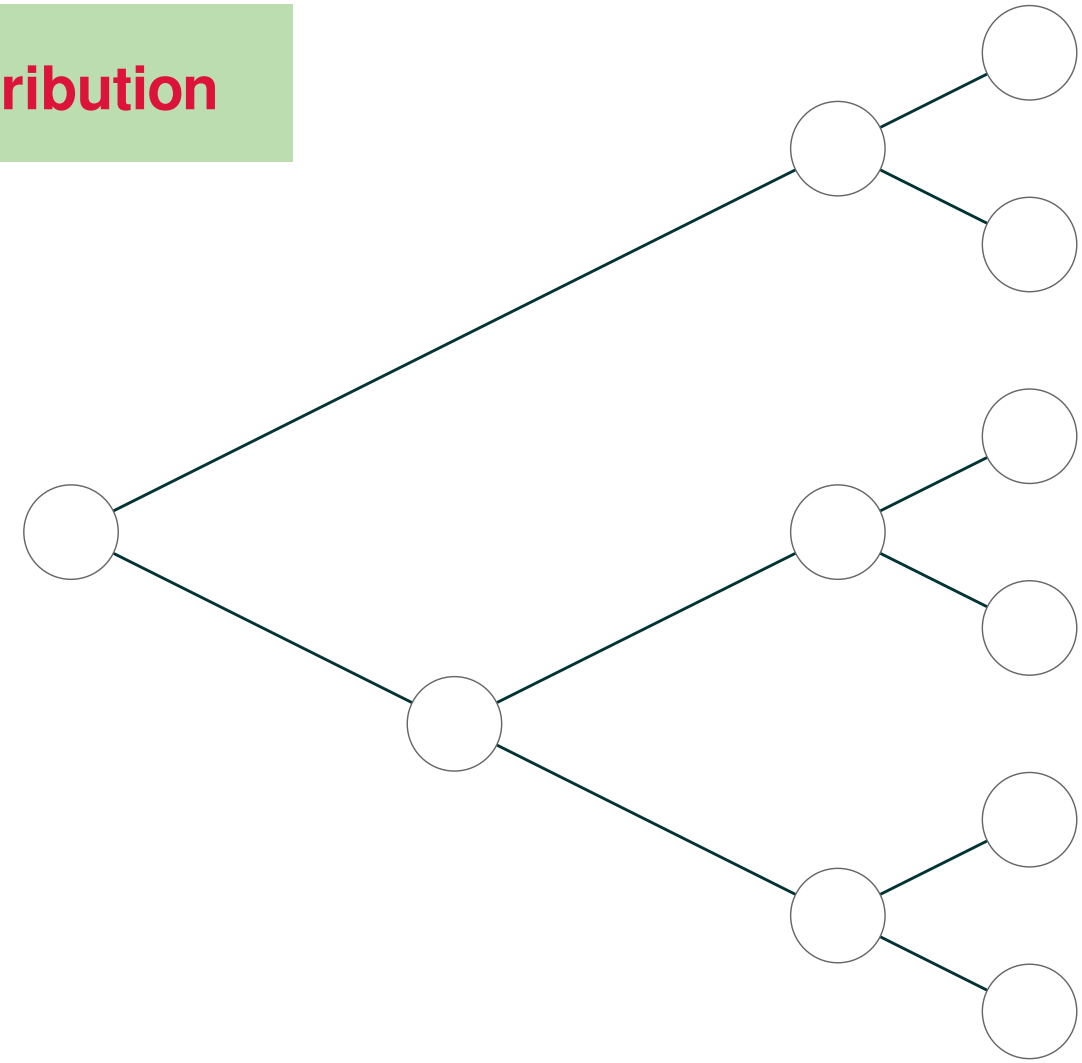


Dagan & Martin (2007)

Nelson-Sathi et al. (2011)

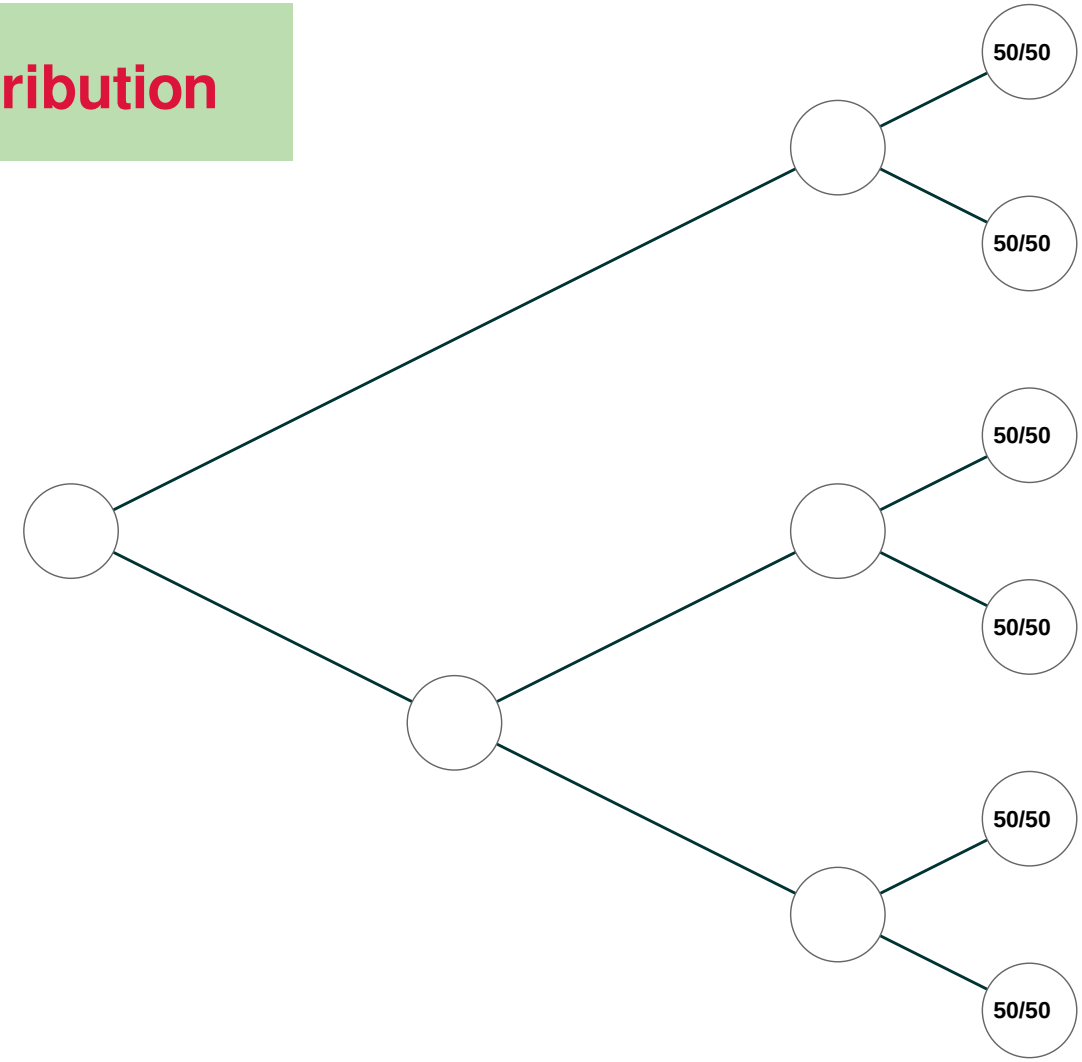
Ancestral Vocabulary Distributions

Vocabulary Distribution



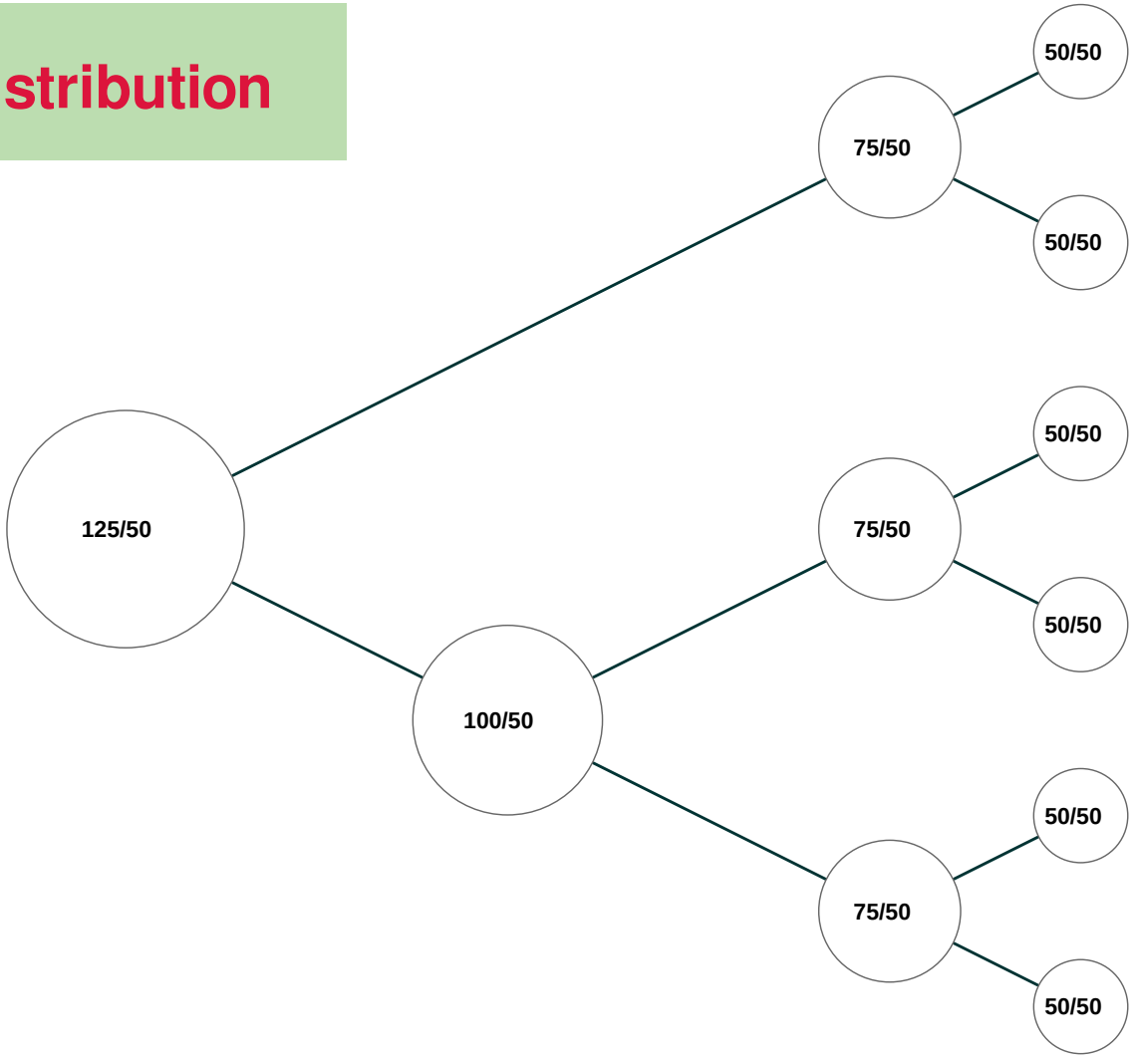
Ancestral Vocabulary Distributions

Vocabulary Distribution



Ancestral Vocabulary Distributions

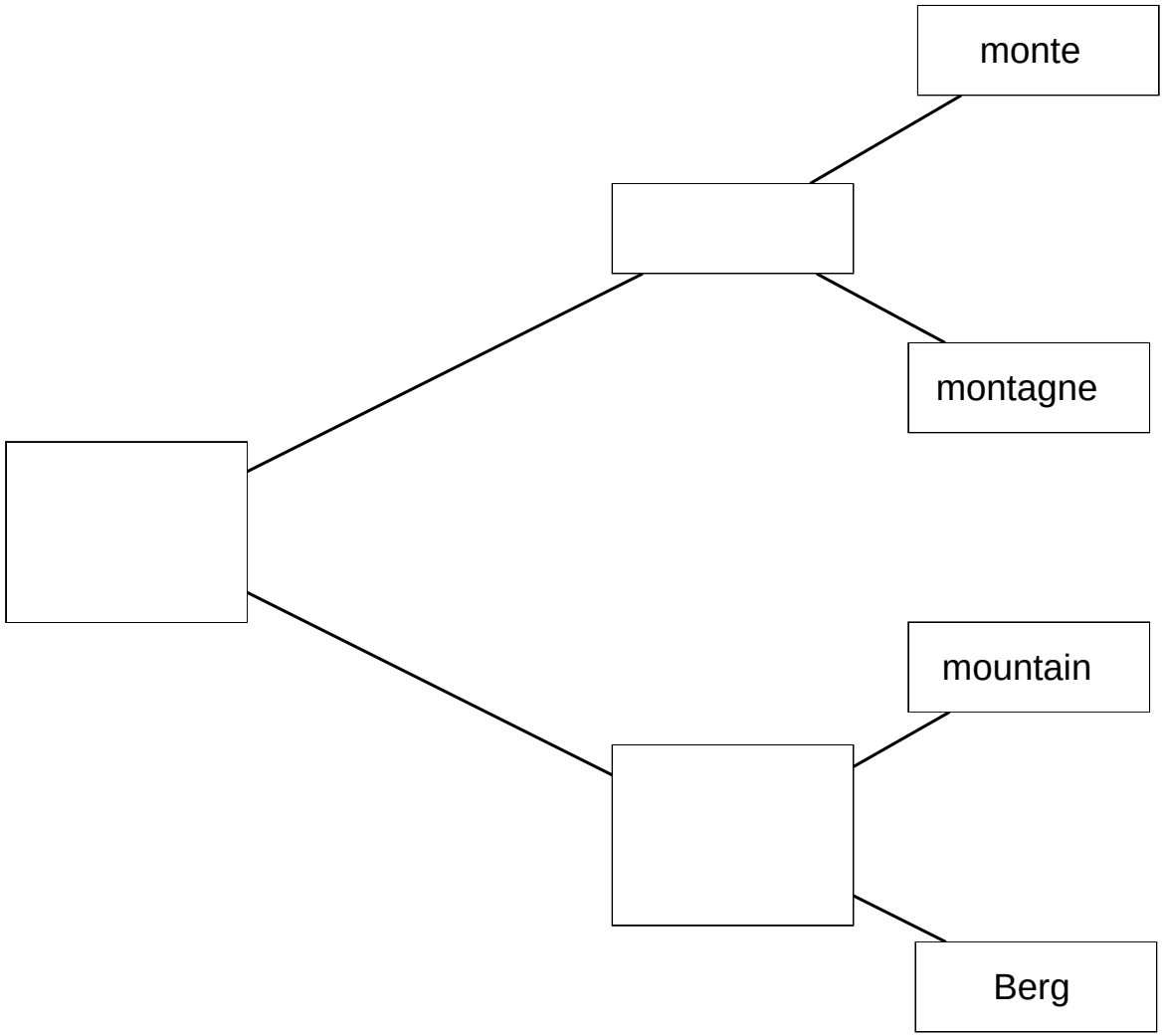
Vocabulary Distribution



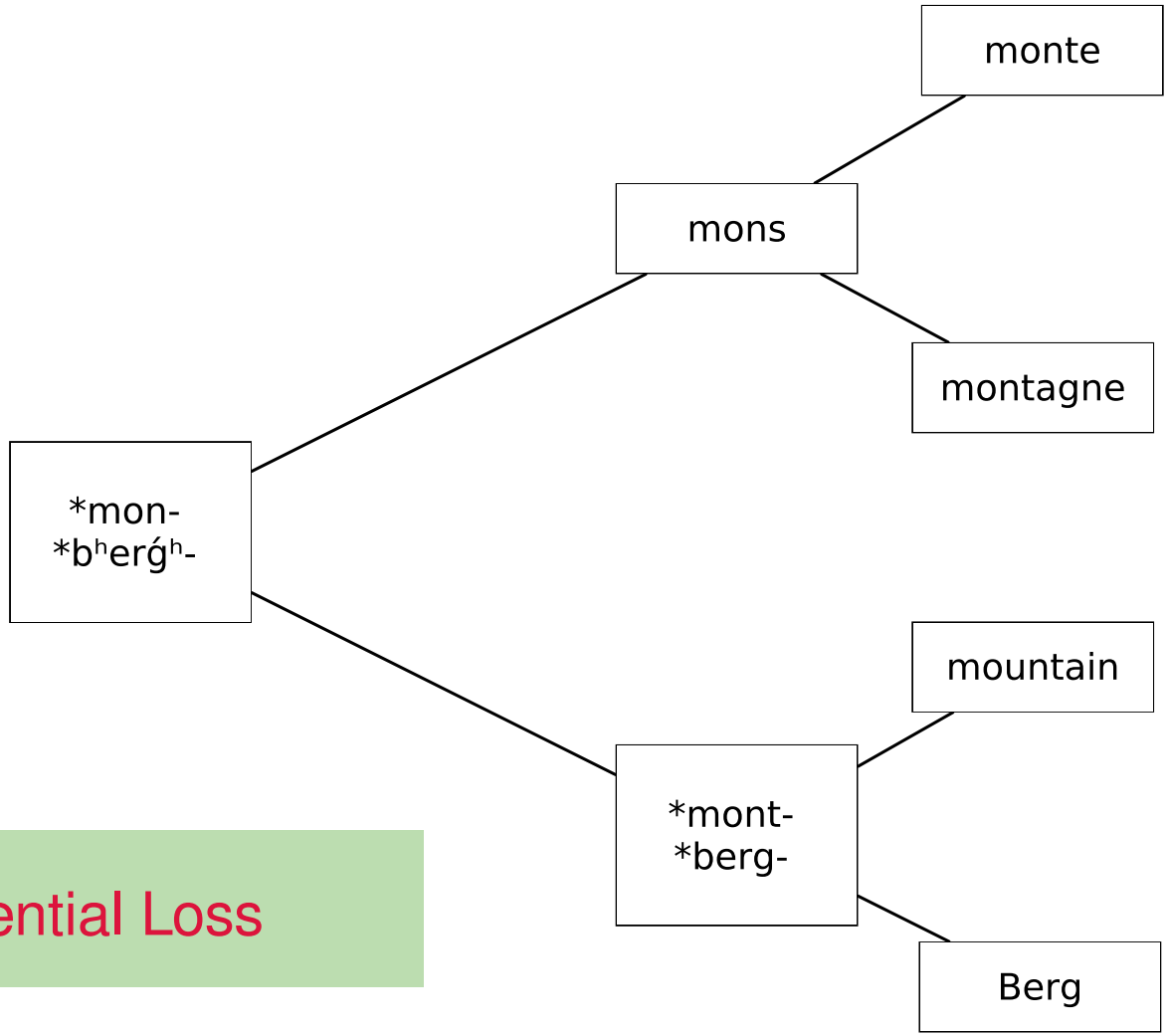
Ancestral Vocabulary Distributions

Favoring ancestral vocabulary distributions over ancestral vocabulary sizes comes quite closer to linguistic needs, since we know that languages cannot be measured in terms of their “size”, while it is reasonable to assume that languages do not allow for an unlimited amount of synonyms. Furthermore, ancestral vocabulary distributions help to avoid problems resulting from semantic shift.

Differential Loss and Semantic Shift

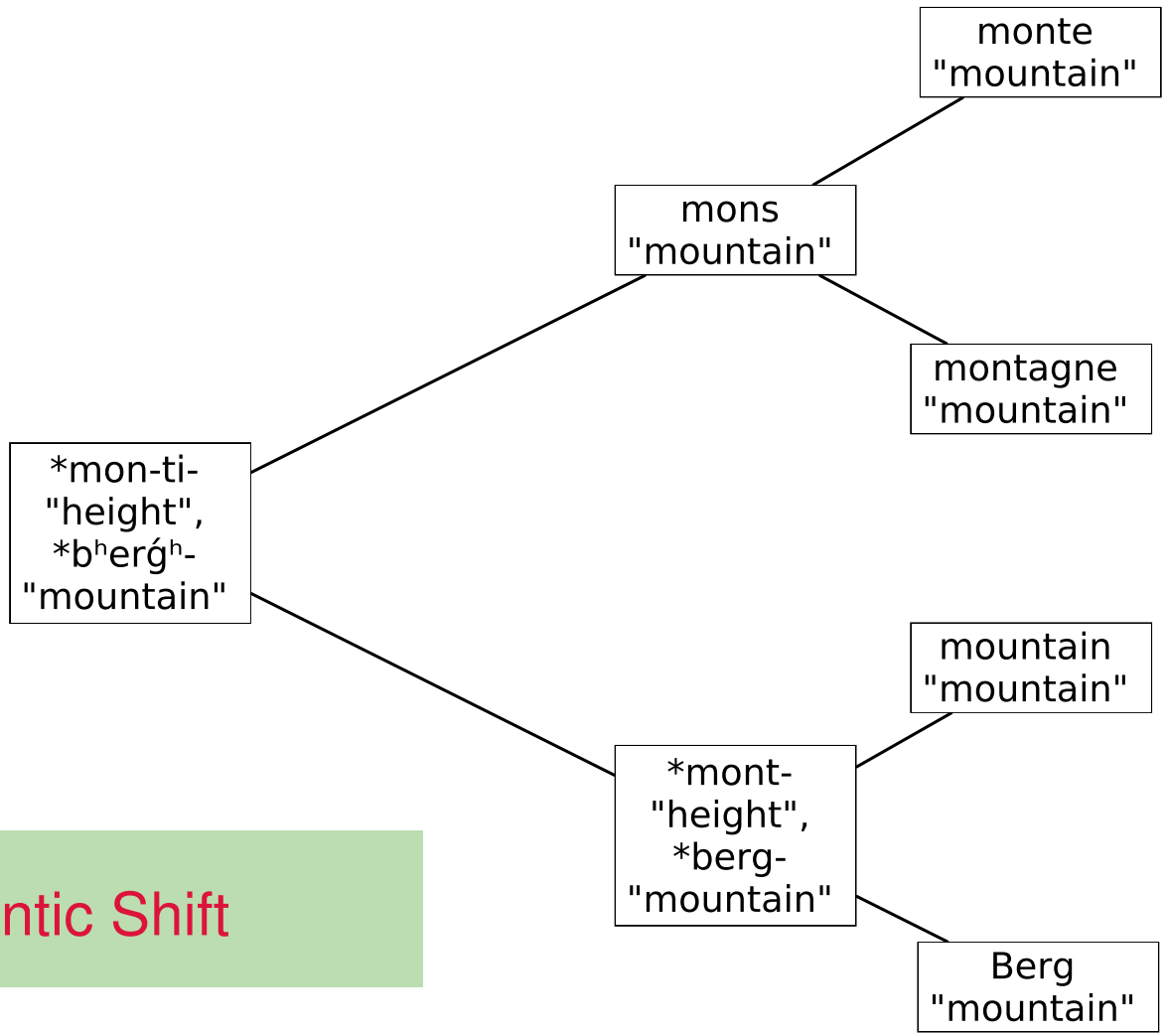


Differential Loss and Semantic Shift



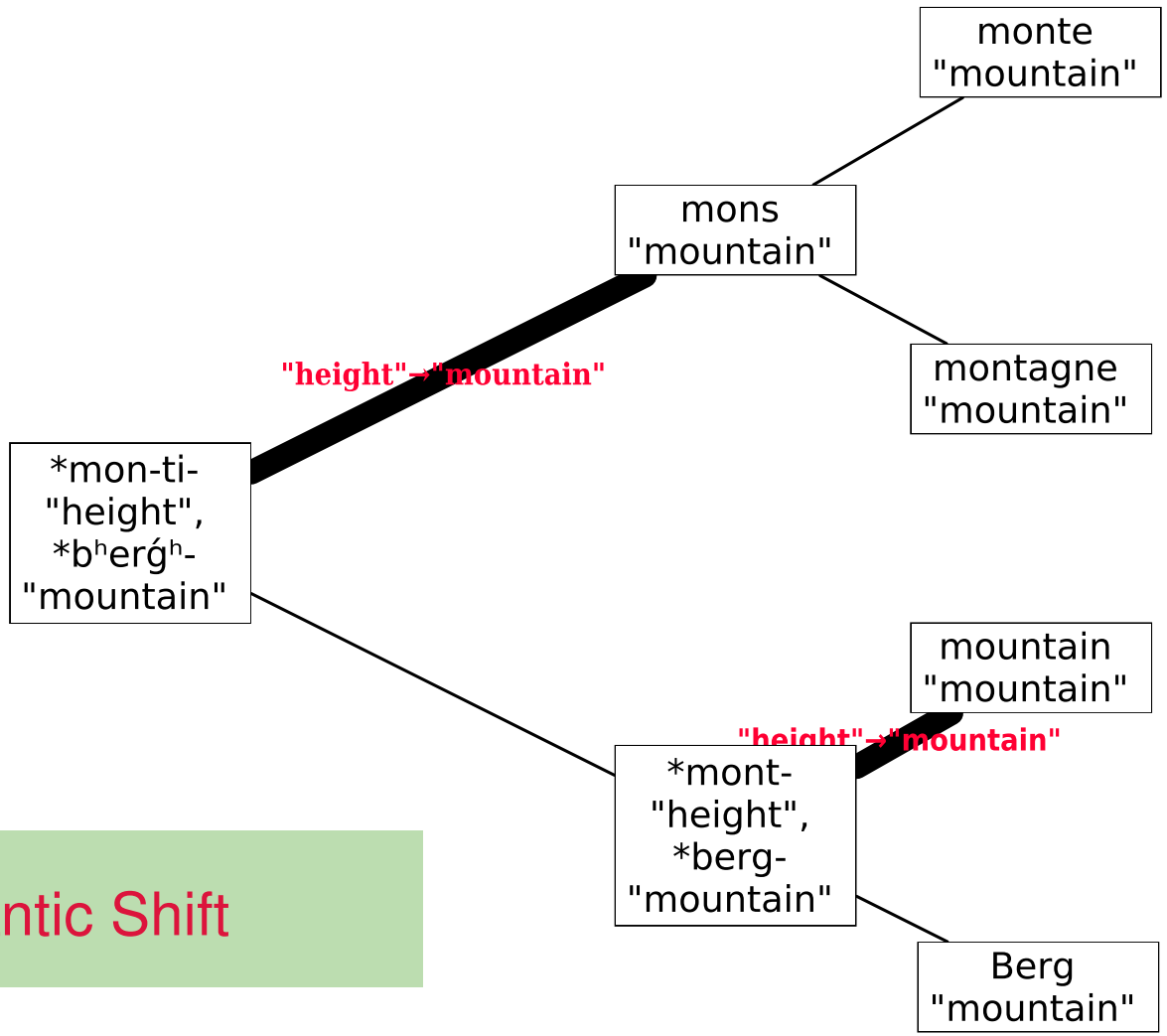
Differential Loss

Differential Loss and Semantic Shift



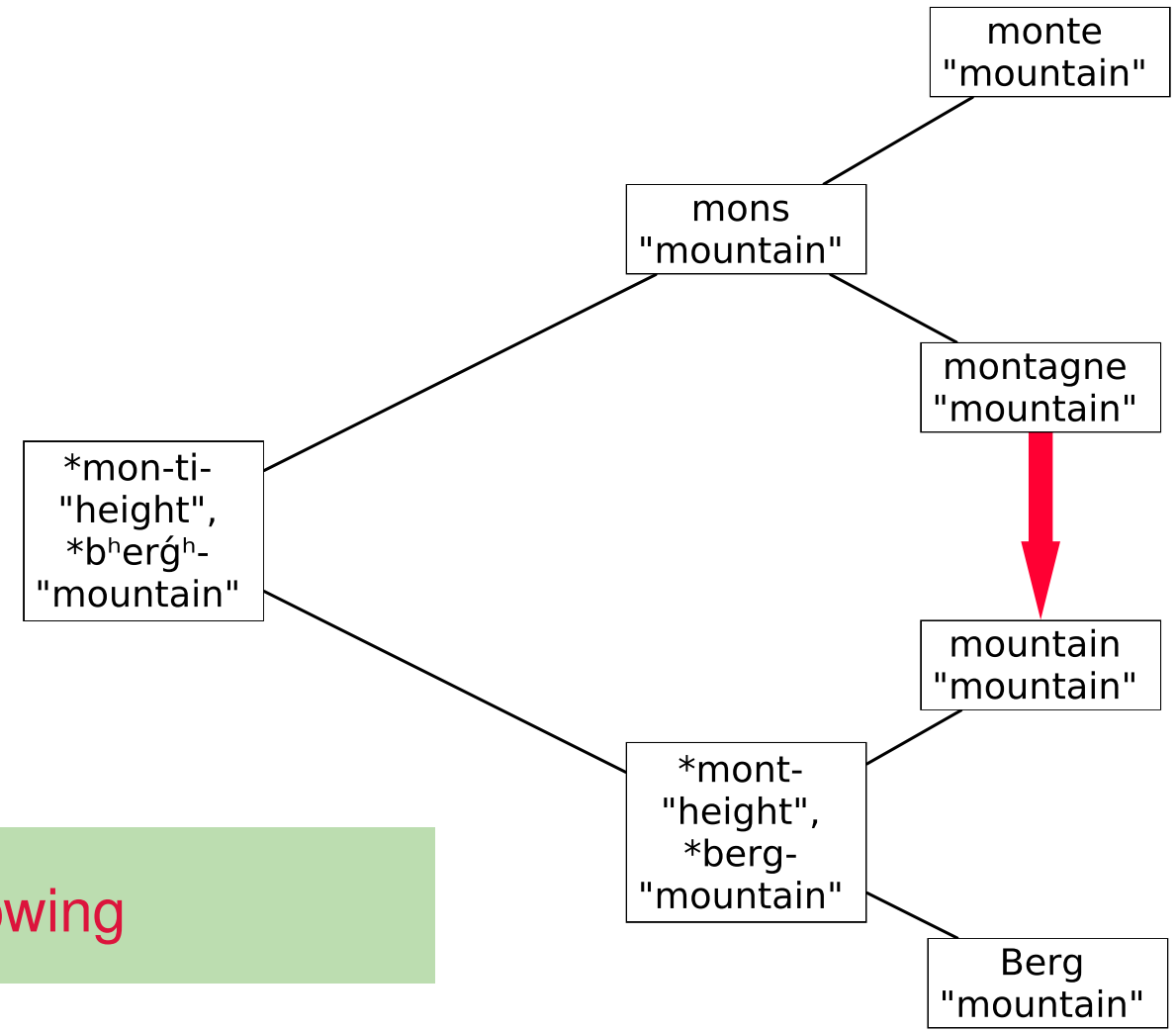
Semantic Shift

Differential Loss and Semantic Shift



Semantic Shift

Differential Loss and Semantic Shift



Borrowing

Differential Loss and Semantic Shift

Parallel semantic shift is not improbable *per se*. However, parallel semantic shift involving the same *source forms* in independent branches of a language family is rather unlikely.

Gain Loss Mapping Approach to Borrowing Detection

Gain Loss Mapping Approach to Borrowing Detection

Input

(a) lexicostatistic dataset (cognate sets)

(b) presence-absence matrix (phyletic patterns)

(c) reference tree

Gain Loss Mapping Approach to Borrowing Detection

Input

(a) lexicostatistic dataset (cognate sets)

(b) presence-absence matrix (phyletic patterns)

(c) reference tree

---

1 **Gain Loss Mapping**

Apply parsimony-based gain loss mapping analysis using different models with varying ratio of weights for gains and losses.

Gain Loss Mapping Approach to Borrowing Detection

Input

(a) lexicostatistic dataset (cognate sets)

(b) presence-absence matrix (phyletic patterns)

(c) reference tree

---

1 **Gain Loss Mapping**

Apply parsimony-based gain loss mapping analysis using different models with varying ratio of weights for gains and losses.

2 **Model Selection**

Choose the most probable model by comparing the ancestral vocabulary distributions with the contemporary ones using the Mann-Whitney-U test.

Gain Loss Mapping Approach to Borrowing Detection

Input

(a) lexicostatistic dataset (cognate sets)

(b) presence-absence matrix (phyletic patterns)

(c) reference tree

---

1 **Gain Loss Mapping**

Apply parsimony-based gain loss mapping analysis using different models with varying ratio of weights for gains and losses.

2 **Model Selection**

Choose the most probable model by comparing the ancestral vocabulary distributions with the contemporary ones using the Mann-Whitney-U test.

3 **Patchy Cognate Detection**

Split all cognate sets for which more than one origin was inferred by the best model into subsets of common origin.

Gain Loss Mapping Approach to Borrowing Detection

Input	(a) lexicostatistic dataset (cognate sets) (b) presence-absence matrix (phyletic patterns) (c) reference tree
1 Gain Loss Mapping	Apply parsimony-based gain loss mapping analysis using different models with varying ratio of weights for gains and losses.
2 Model Selection	Choose the most probable model by comparing the ancestral vocabulary distributions with the contemporary ones using the Mann-Whitney-U test.
3 Patchy Cognate Detection	Split all cognate sets for which more than one origin was inferred by the best model into subsets of common origin.
4 Network Reconstruction	Connect the separate origins of all patchy cognate sets by calculating a weighted minimum spanning tree and add all links as edges to the reference tree, whereby the edge weight reflects the number of inferred links.

Application

Dogon Languages

Dogon Languages

- The Dogon language family consists of about 20 distinct (mutually unintelligible) languages.
- The internal structure of the language family is largely unknown. Some scholars propose a split in an Eastern and a Western branch.
- The Dogon Languages Project (DLP, `http://dogonlanguages.org`) provides a lexical spreadsheet consisting of 23 language varieties submitted by 5 authors.
- The spreadsheet consists of 9000 semantic items translated into the respective varieties, but only a small amount of the items (less than 200) is translated into all languages.

Dogon Data

From the Dogon spreadsheet, we extracted:

- 325 semantic items (“concepts”), translated into
- 18 varieties (“doculects”), yielding a total amount of
- 4883 words (“counterparts”)

The main criterion for the data selection was to maximize the number of semantically aligned words in the given varieties in order to avoid large amounts of gaps in the data.

QLC-LingPy

- All analyses were conducted using the development version of QLC-LingPy.
- QLC-LingPy is a Python library currently being developed in Michael Cysouw's reasearch unit "Quantitative Language Comparison" (Philipps-University Marburg).
- QLC-LingPy supersedes the independently developed QLC and LingPy libraries by merging their specific features into a common framework, while extending its functionality.
- Our goal is to provide a Python toolkit that is easy to use for non-experts in programming, while at the same time offering up-to-date proposals for common tasks in quantitative historical linguistics.

Workflow

Workflow

Input

(a) Dogon spreadsheet

(b) Reference trees (DLP, MrBayes, Neighbor-Joining)

Workflow

Input

(a) Dogon spreadsheet

(b) Reference trees (DLP, MrBayes, Neighbor-Joining)

---

1 **Preprocessing**

Orthographic parsing (IPA conversion) and tokenization using the Orthography Profile Approach (Moran & Cysouw in prep.).

Workflow

Input

(a) Dogon spreadsheet

(b) Reference trees (DLP, MrBayes, Neighbor-Joining)

---

1 **Preprocessing**

Orthographic parsing (IPA conversion) and tokenization using the Orthography Profile Approach (Moran & Cysouw in prep.).

2 **Cognate Detection**

Identification of etymologically related words (cognates and borrowings, i.e. “homologs”) using the LexStat method (List 2012) with a low threshold (0.4) in order to minimize the number of false positives.

Workflow

Input

(a) Dogon spreadsheet

(b) Reference trees (DLP, MrBayes, Neighbor-Joining)

---

1 **Preprocessing**

Orthographic parsing (IPA conversion) and tokenization using the Orthography Profile Approach (Moran & Cysouw in prep.).

2 **Cognate Detection**

Identification of etymologically related words (cognates and borrowings, i.e. “homologs”) using the LexStat method (List 2012) with a low threshold (0.4) in order to minimize the number of false positives.

3 **Borrowing Detection**

Identification of patchy phyletic patterns using the improved gain loss mapping approach (ten different gain-loss models, favoring varying amounts of origins).

Workflow

Input

(a) Dogon spreadsheet

(b) Reference trees (DLP, MrBayes, Neighbor-Joining)

---

1 **Preprocessing**

Orthographic parsing (IPA conversion) and tokenization using the Orthography Profile Approach (Moran & Cysouw in prep.).

2 **Cognate Detection**

Identification of etymologically related words (cognates and borrowings, i.e. “homologs”) using the LexStat method (List 2012) with a low threshold (0.4) in order to minimize the number of false positives.

3 **Borrowing Detection**

Identification of patchy phyletic patterns using the improved gain loss mapping approach (ten different gain-loss models, favoring varying amounts of origins).

Output

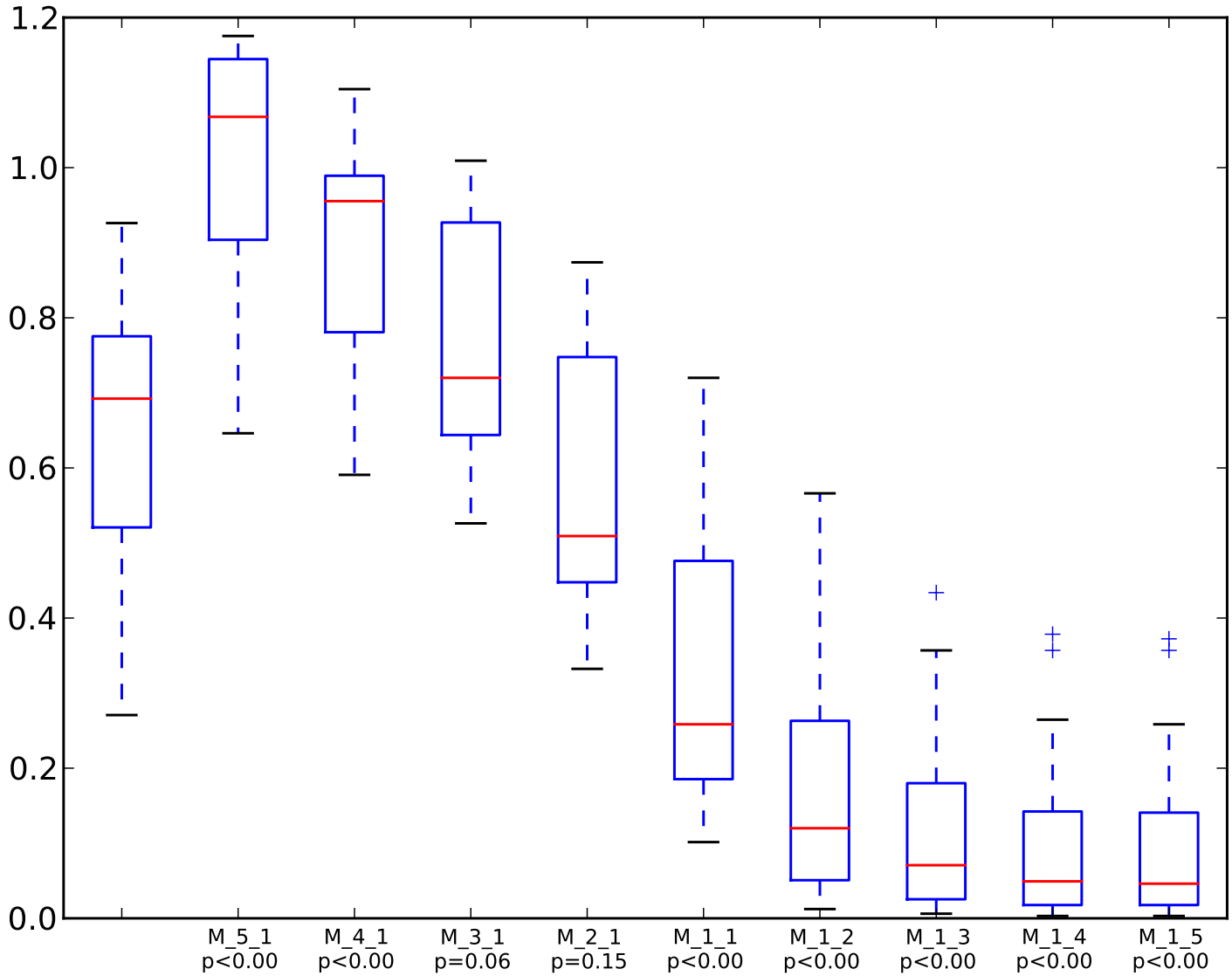
(a) Cognate sets

(b) Patchy cognate sets

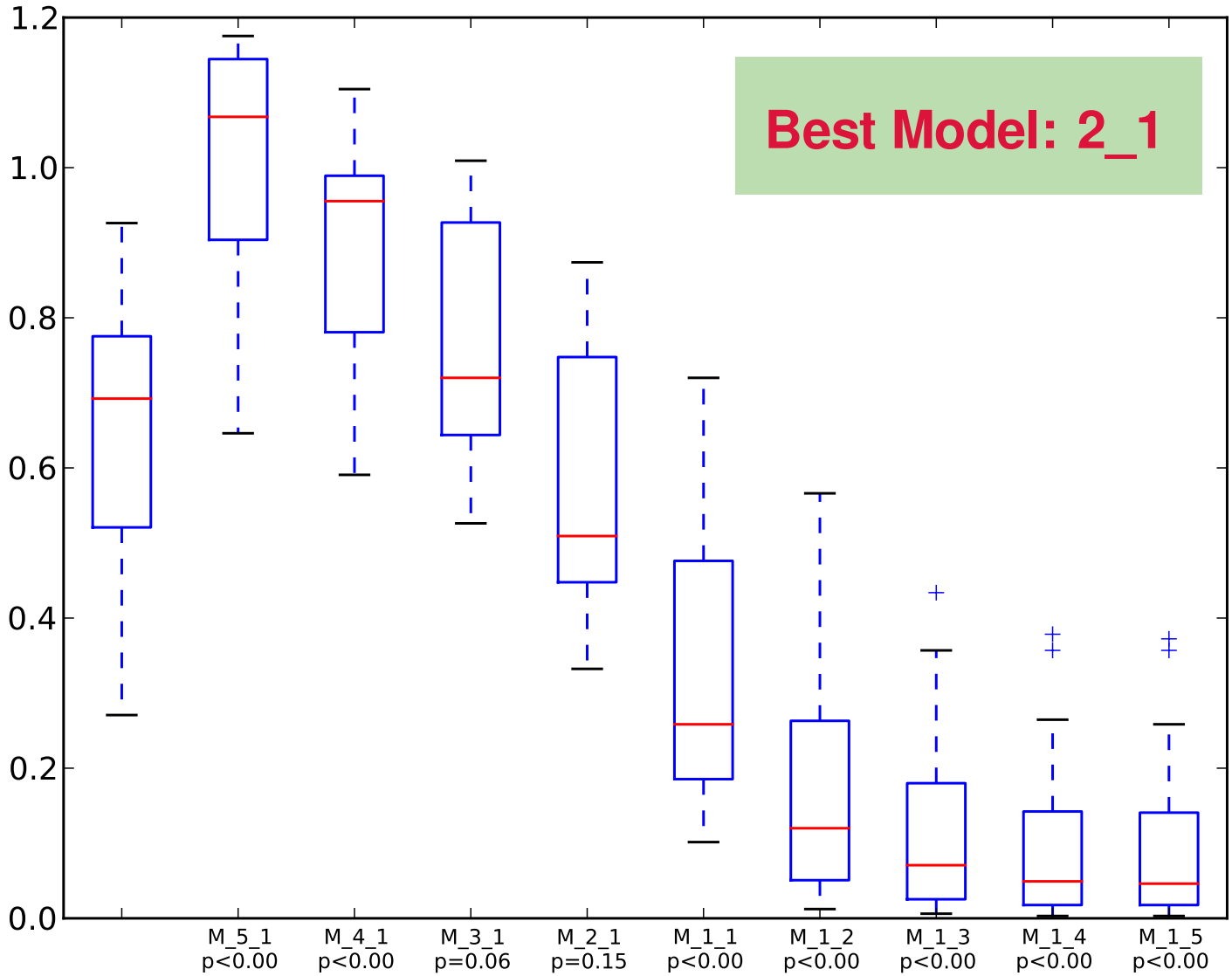
(c) Phylogenetic network

Models

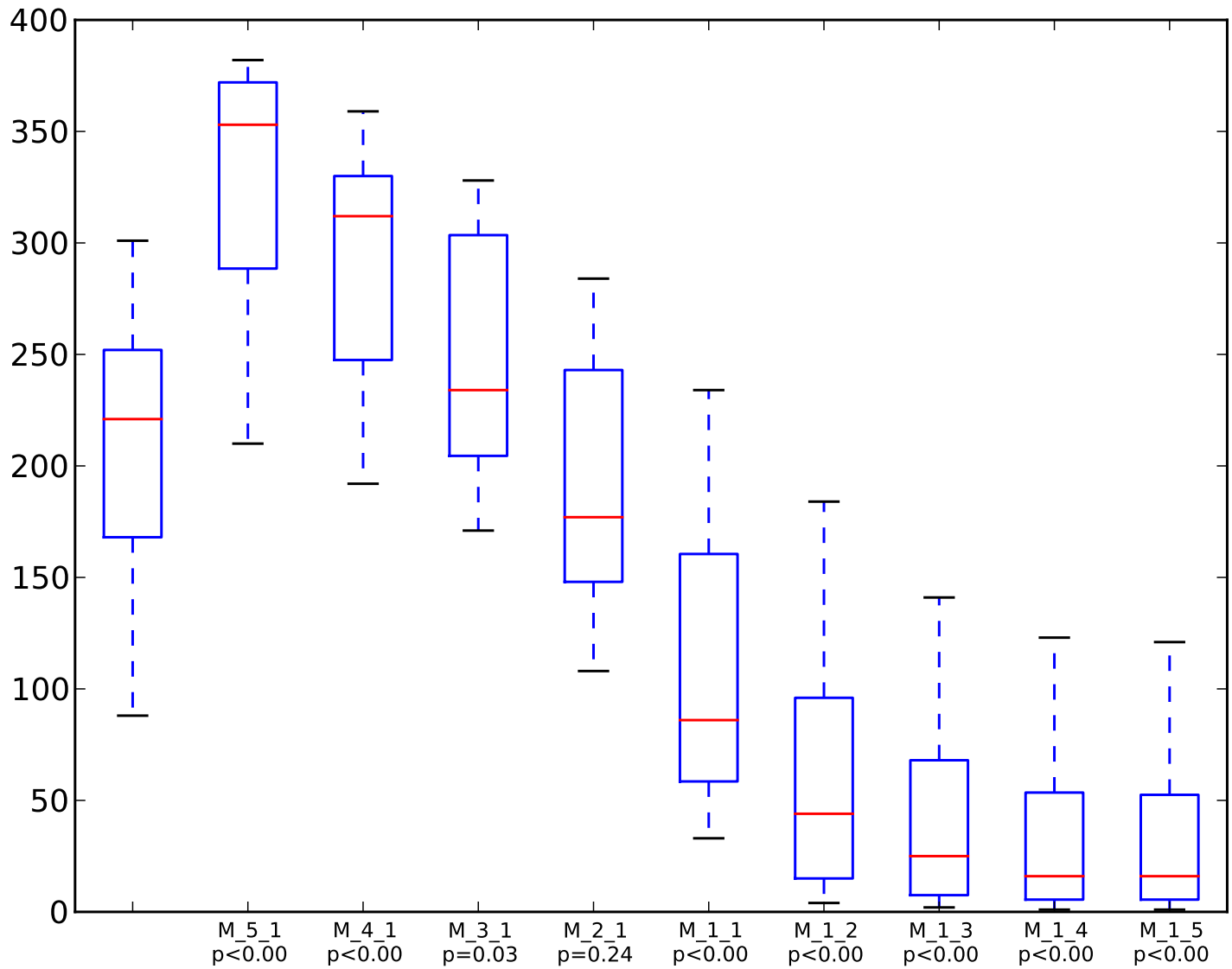
Models


1.2
1.0
0.8
0.6
0.4
0.2
0.0
M_5_1
p<0.00
M_4_1
p<0.00
M_3_1
p=0.06
M_2_1
p=0.15
M_1_1
p<0.00
M_1_2
p<0.00
M_1_3
p<0.00
M_1_4
p<0.00
M_1_5
p<0.00

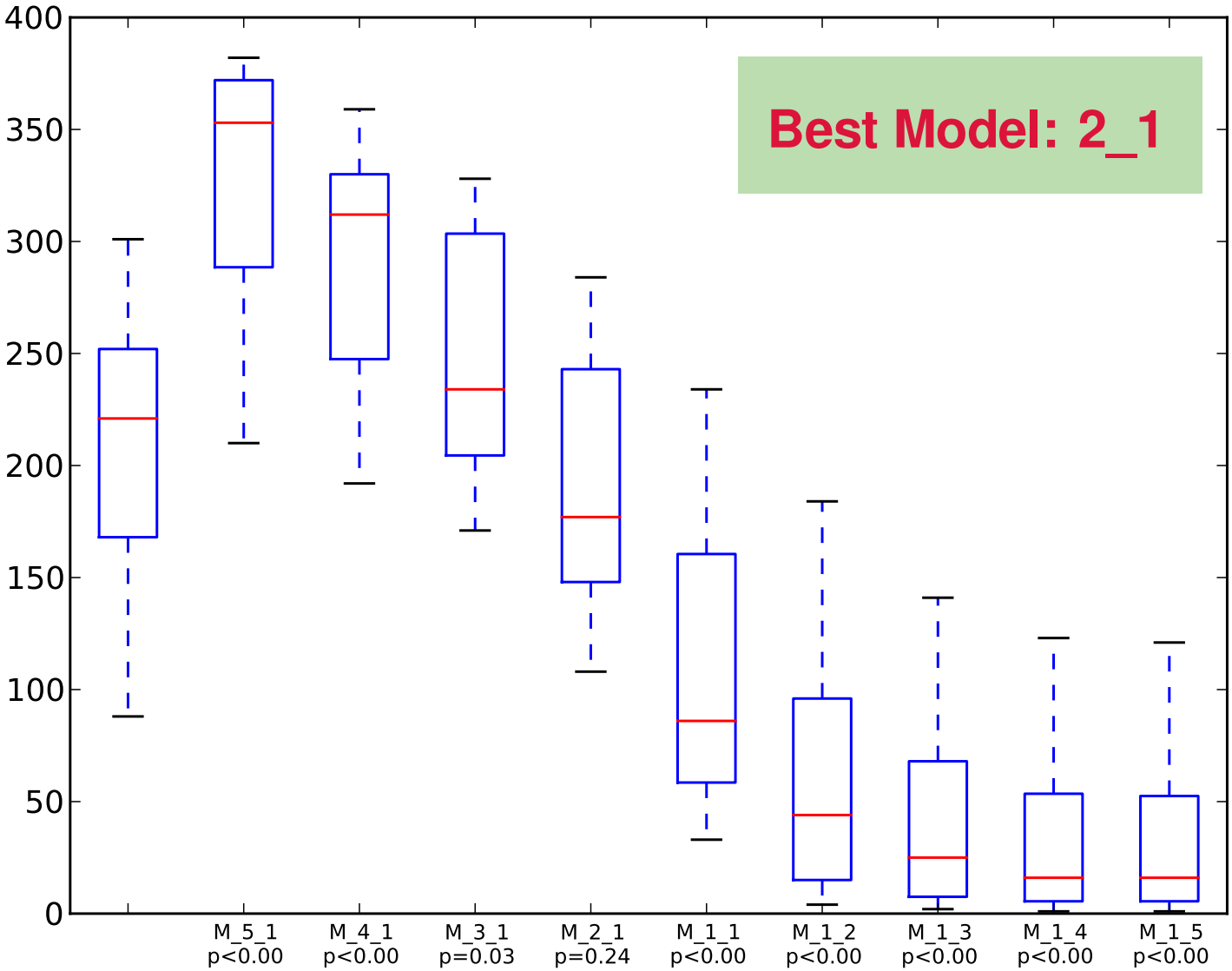
Models


Best Model: 2_1
1.2
1.0
0.8
0.6
0.4
0.2
0.0
M_5_1 p<0.00
M_4_1 p<0.00
M_3_1 p=0.06
M_2_1 p=0.15
M_1_1 p<0.00
M_1_2 p<0.00
M_1_3 p<0.00
M_1_4 p<0.00
M_1_5 p<0.00

Models


400
350
300
250
200
150
100
50
0
M_5_1
p<0.00
M_4_1
p<0.00
M_3_1
p=0.03
M_2_1
p=0.24
M_1_1
p<0.00
M_1_2
p<0.00
M_1_3
p<0.00
M_1_4
p<0.00
M_1_5
p<0.00

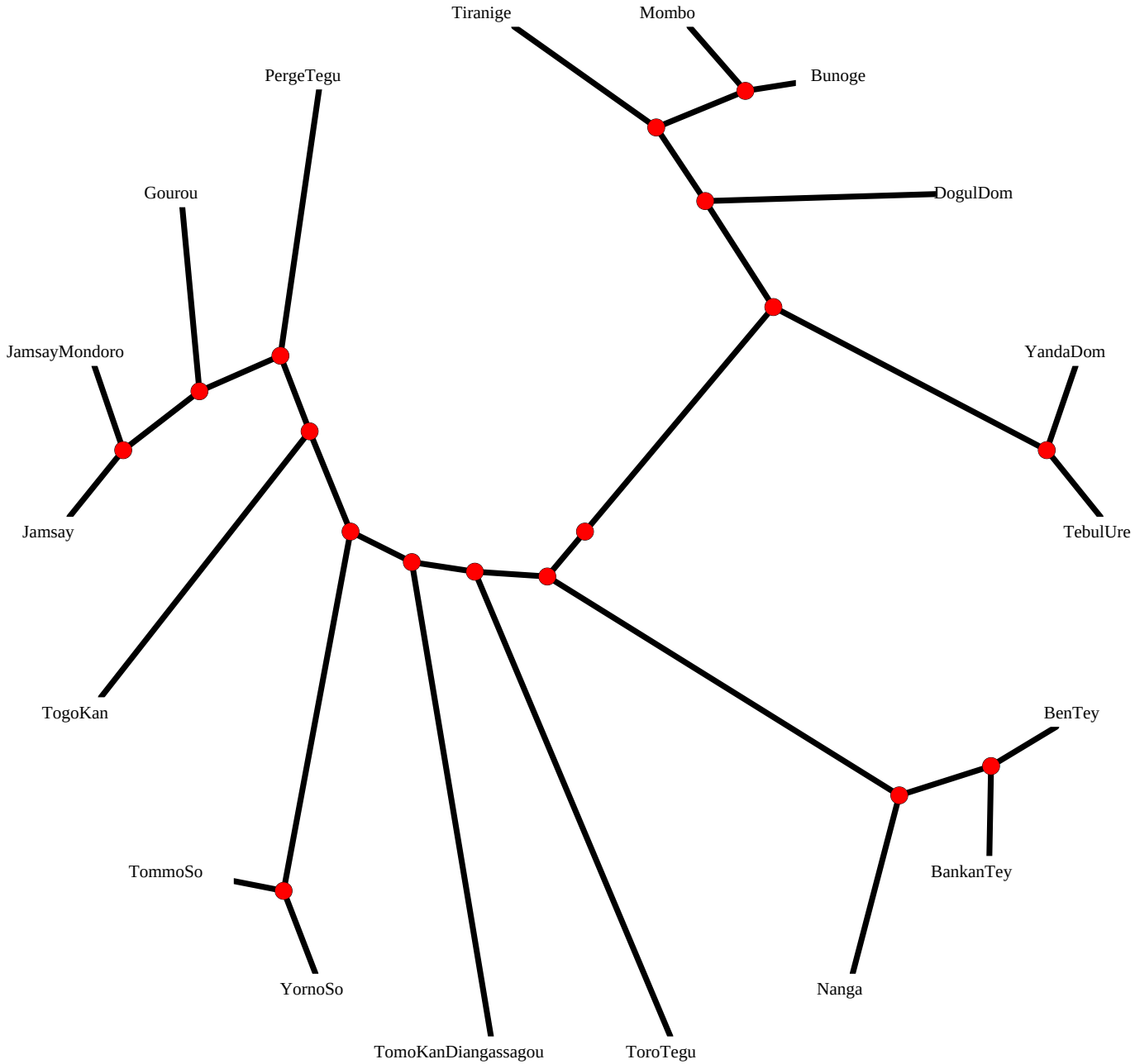
Models


Best Model: 2_1
400
350
300
250
200
150
100
50
0
M_5_1
p<0.00
M_4_1
p<0.00
M_3_1
p=0.03
M_2_1
p=0.24
M_1_1
p<0.00
M_1_2
p<0.00
M_1_3
p<0.00
M_1_4
p<0.00
M_1_5
p<0.00

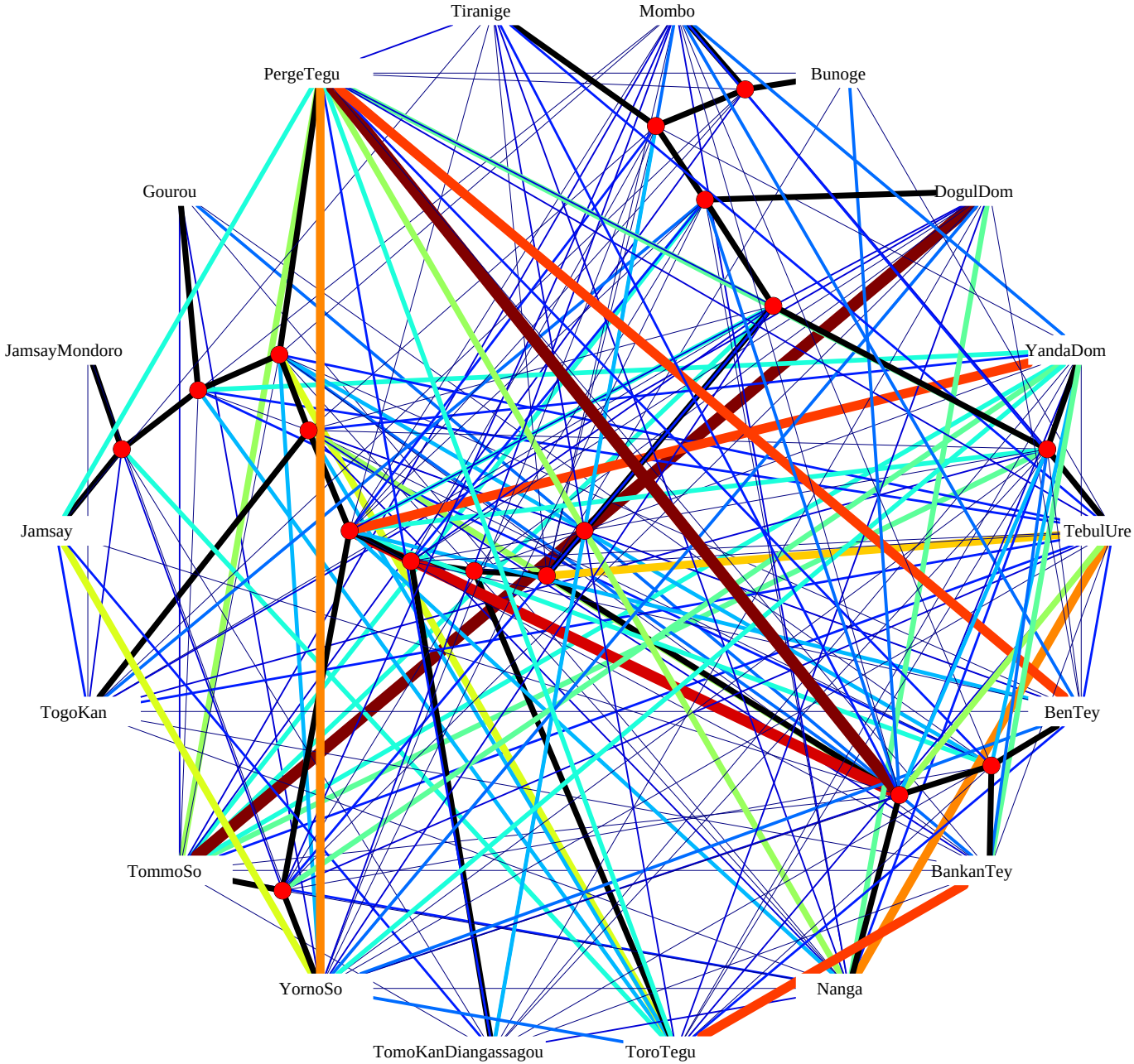
Numbers

Tree	Model	Origins (Ø)	MaxO	p-value
DLP	2_1	1.68	5	0.15
MrBayes	2_1	1.67	5	0.50
NeighborJoining	2_1	1.69	5	0.16

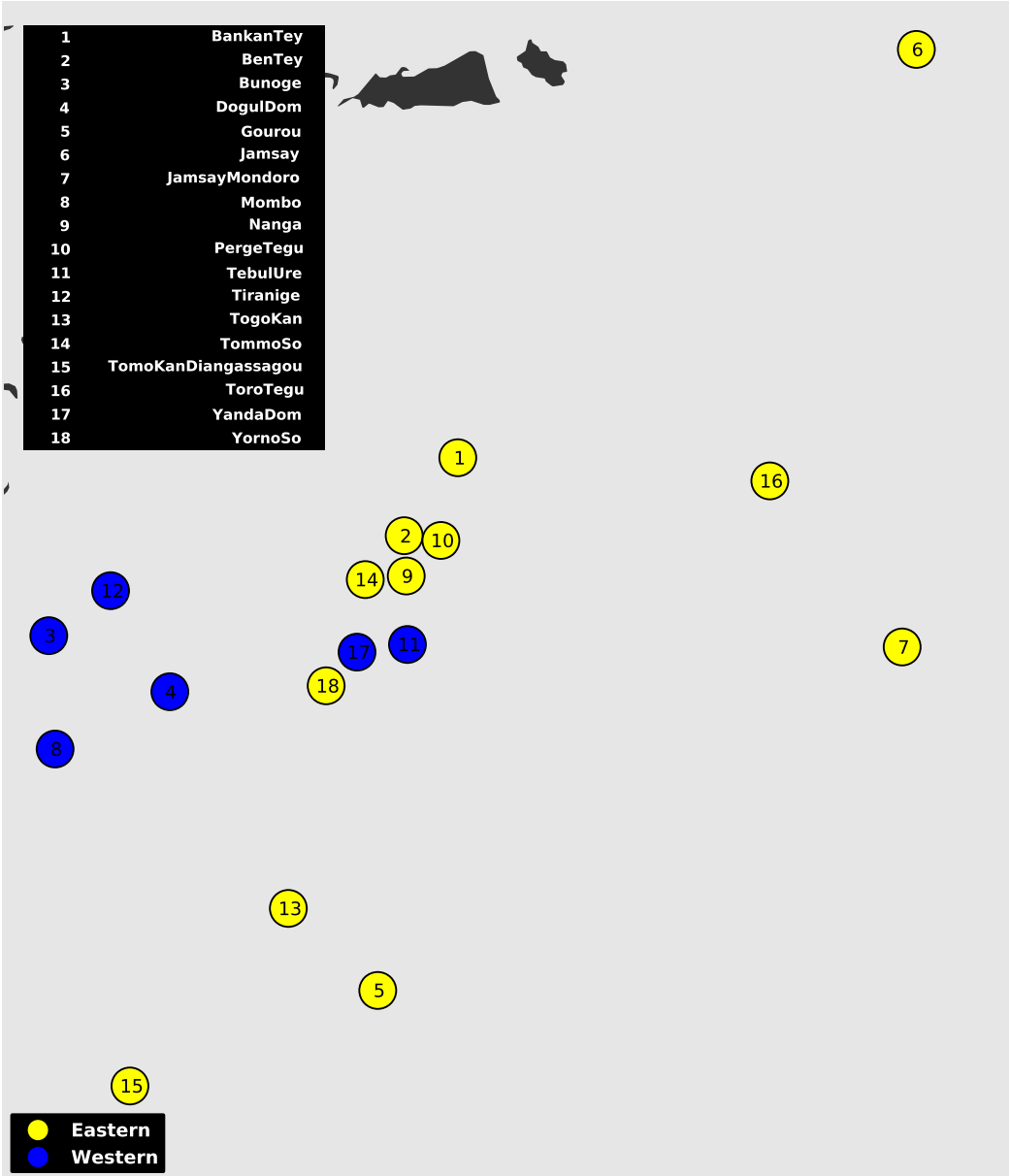
Phylogenetic Network



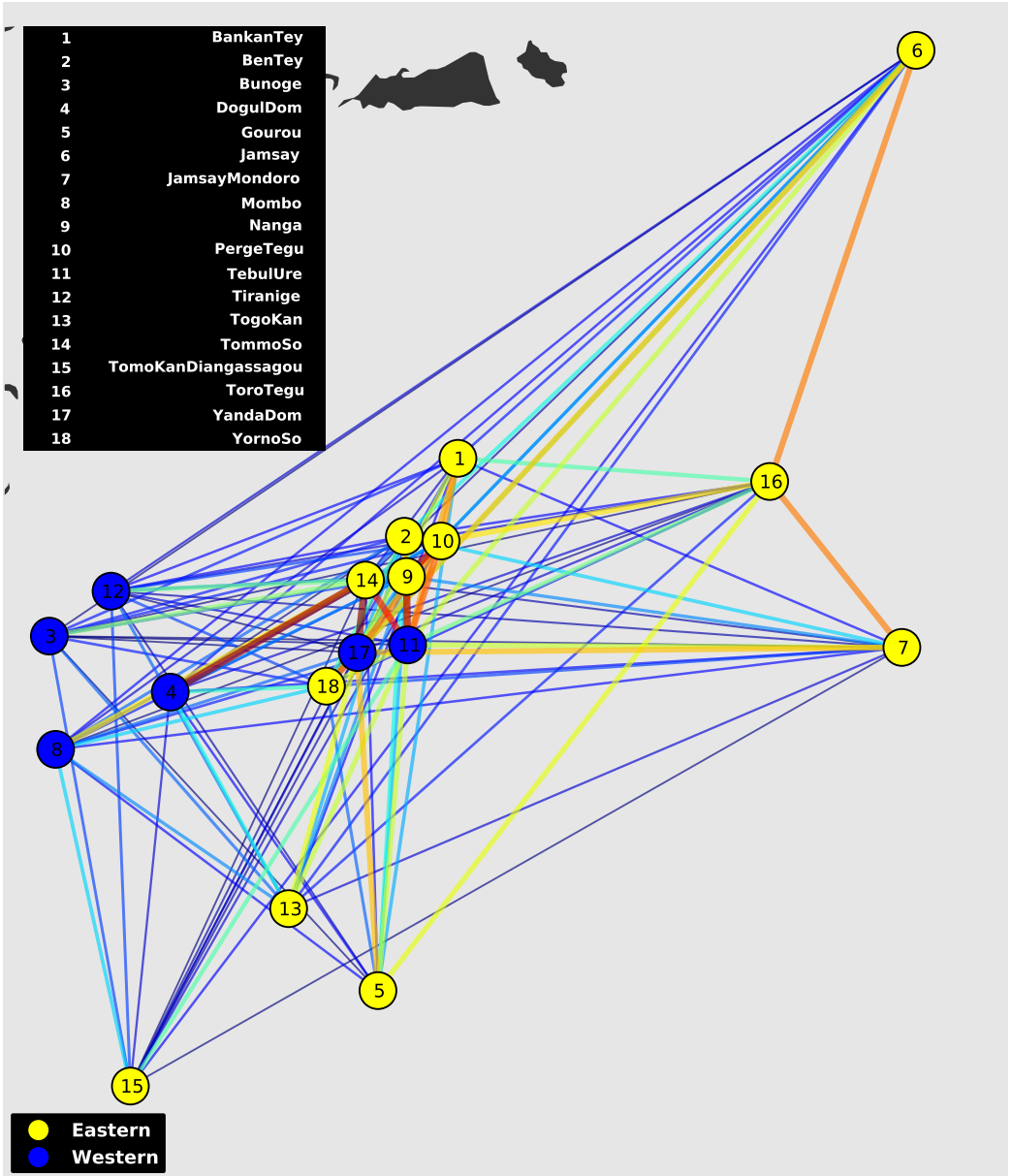
Phylogenetic Network

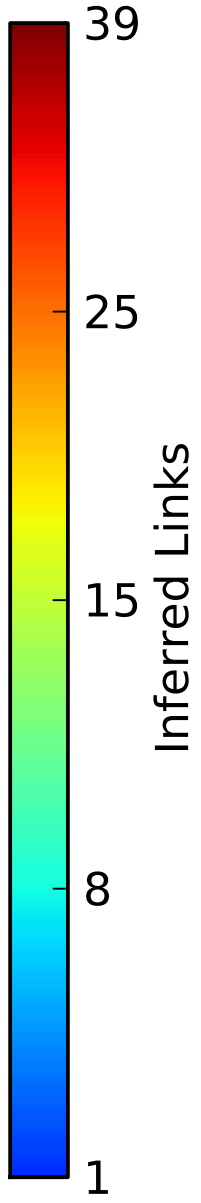


Areal Perspective

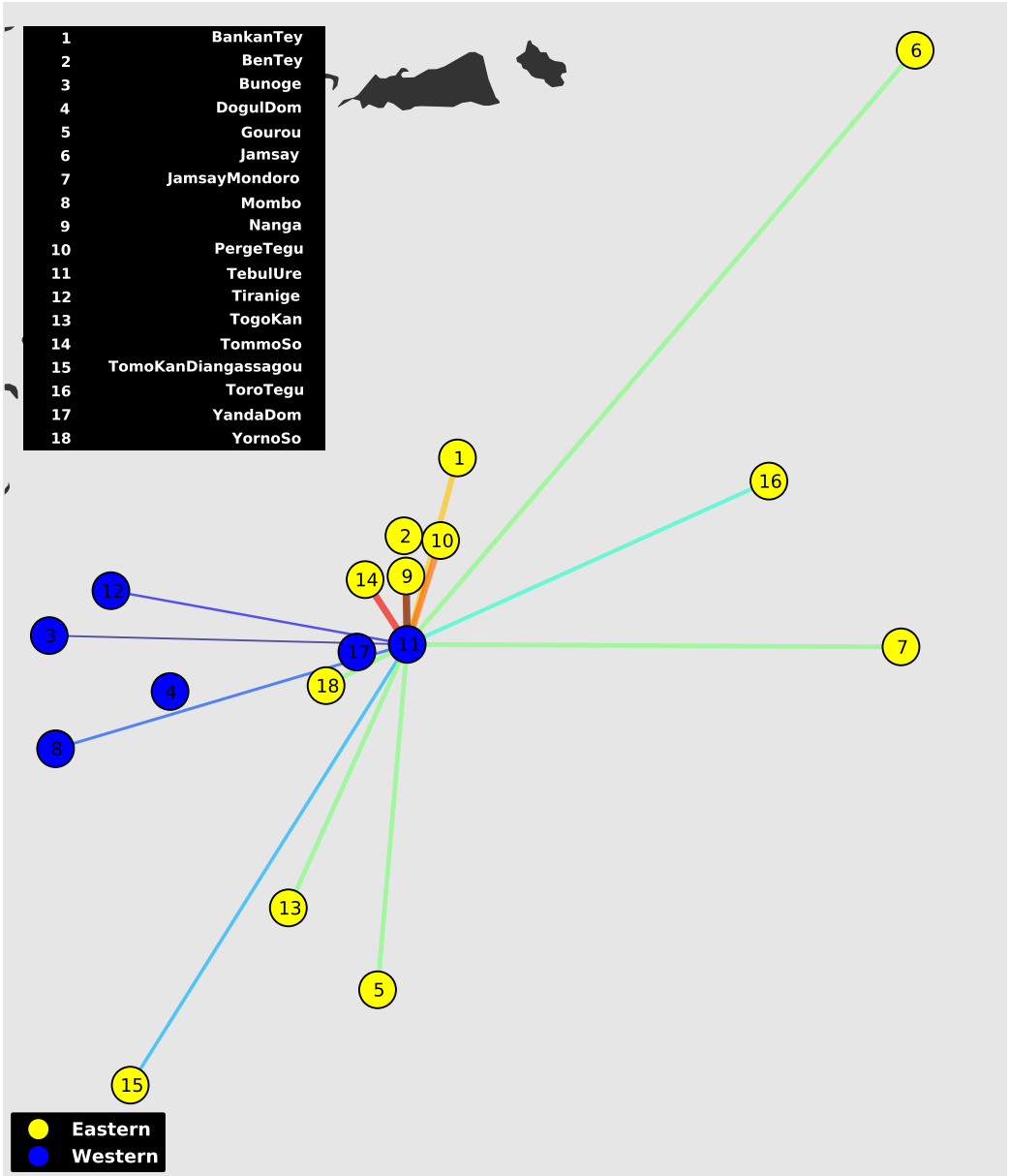


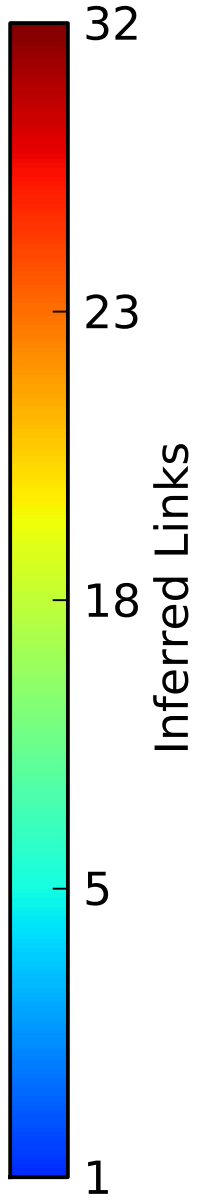
Areal Perspective



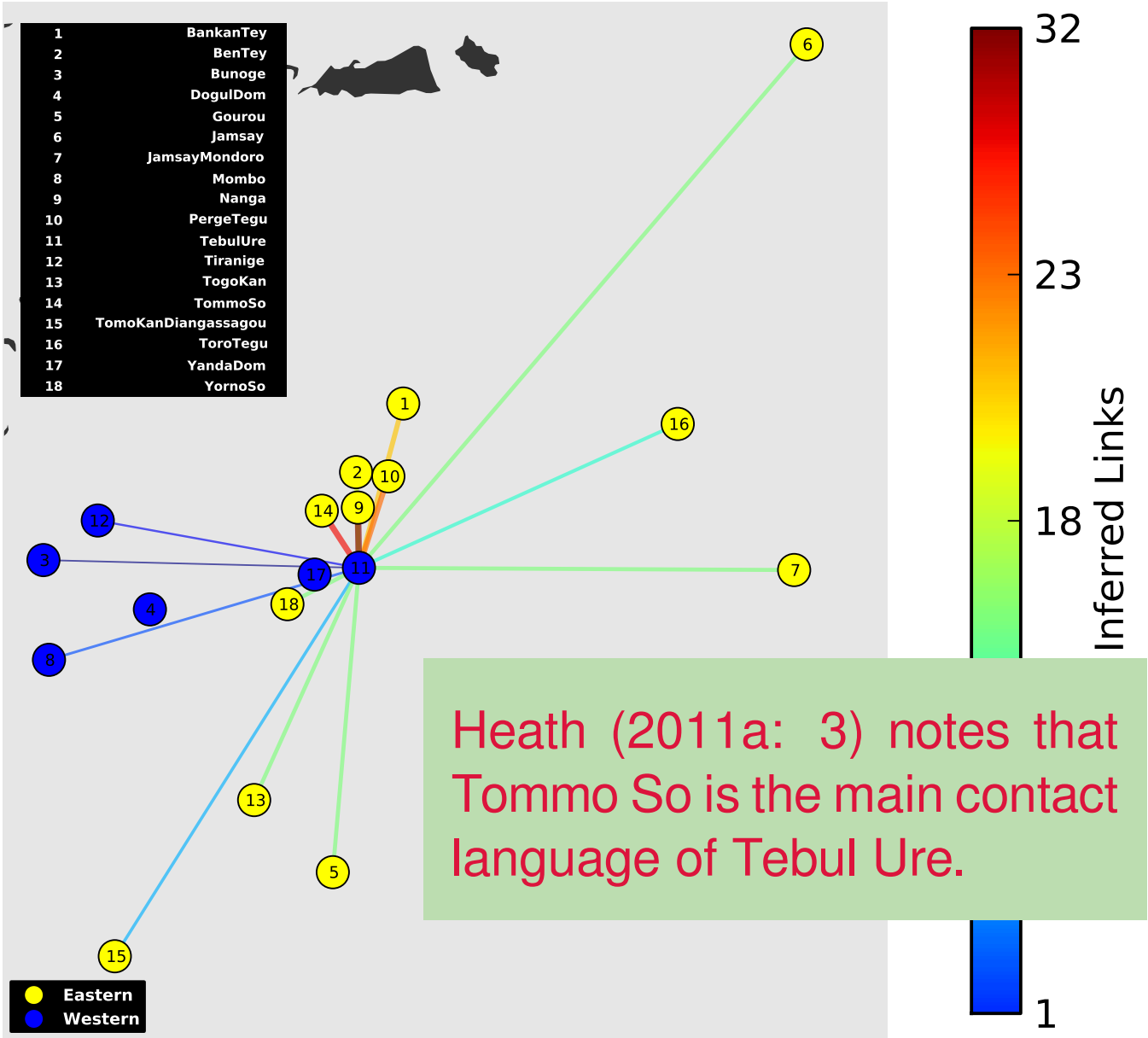


Areal Perspective: Tebul Ure



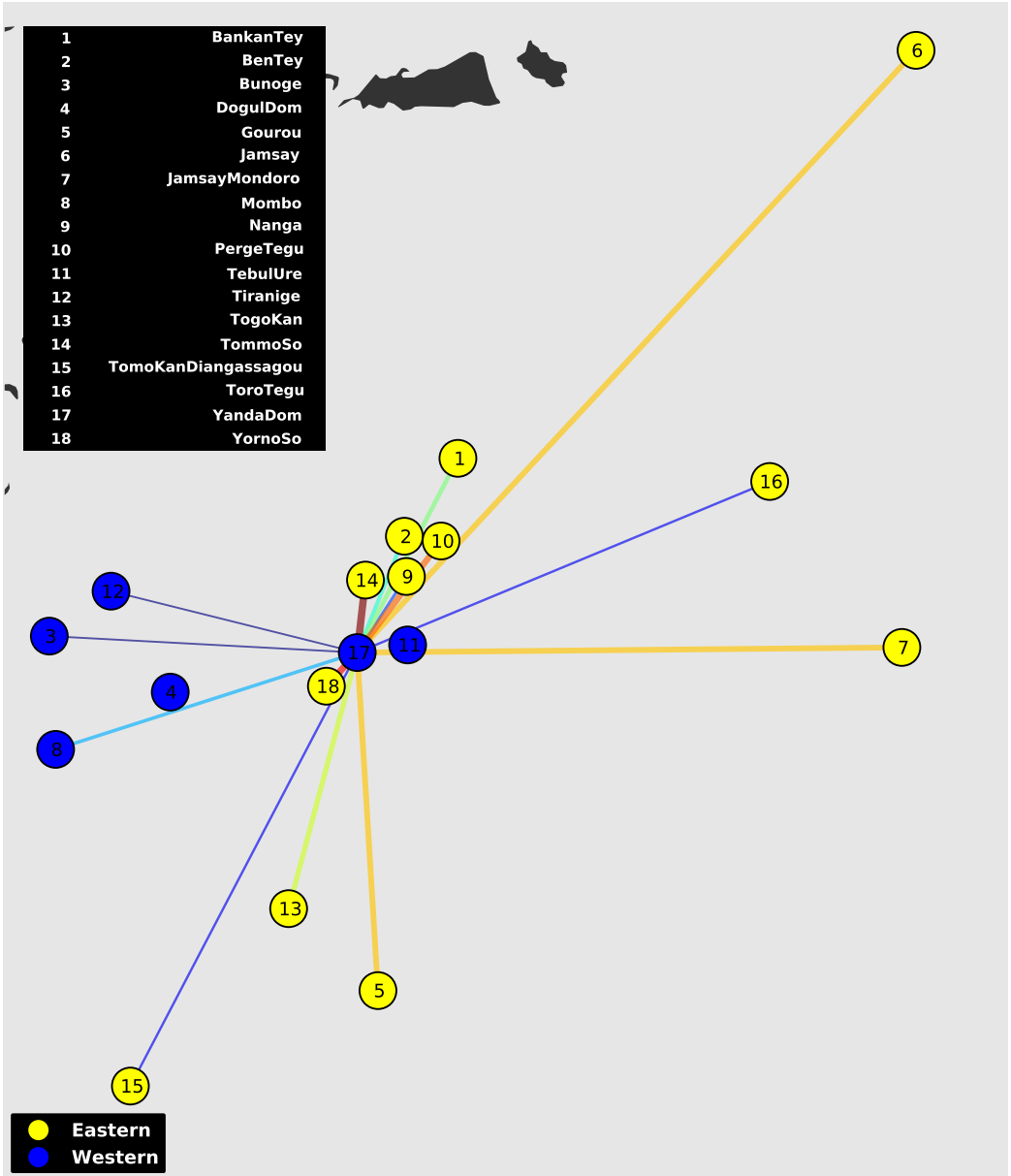


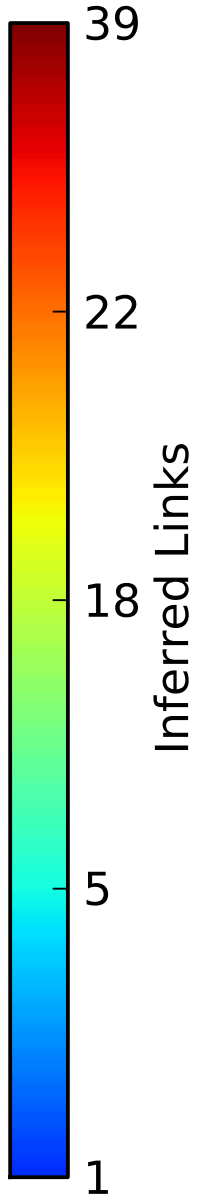
Areal Perspective: Tebul Ure



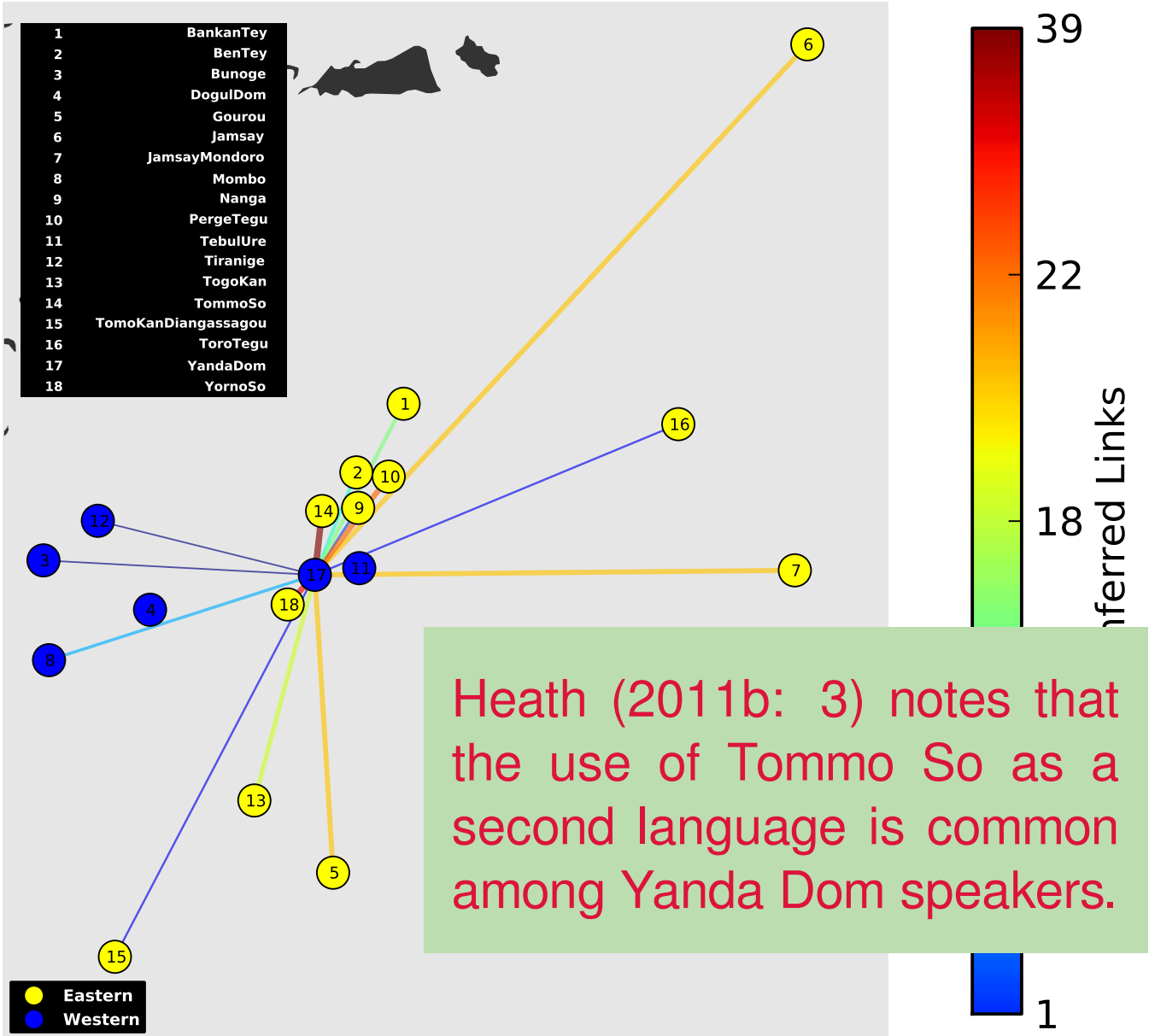
Heath (2011a: 3) notes that Tommo So is the main contact language of Tebul Ure.

Areal Perspective: Yanda Dom



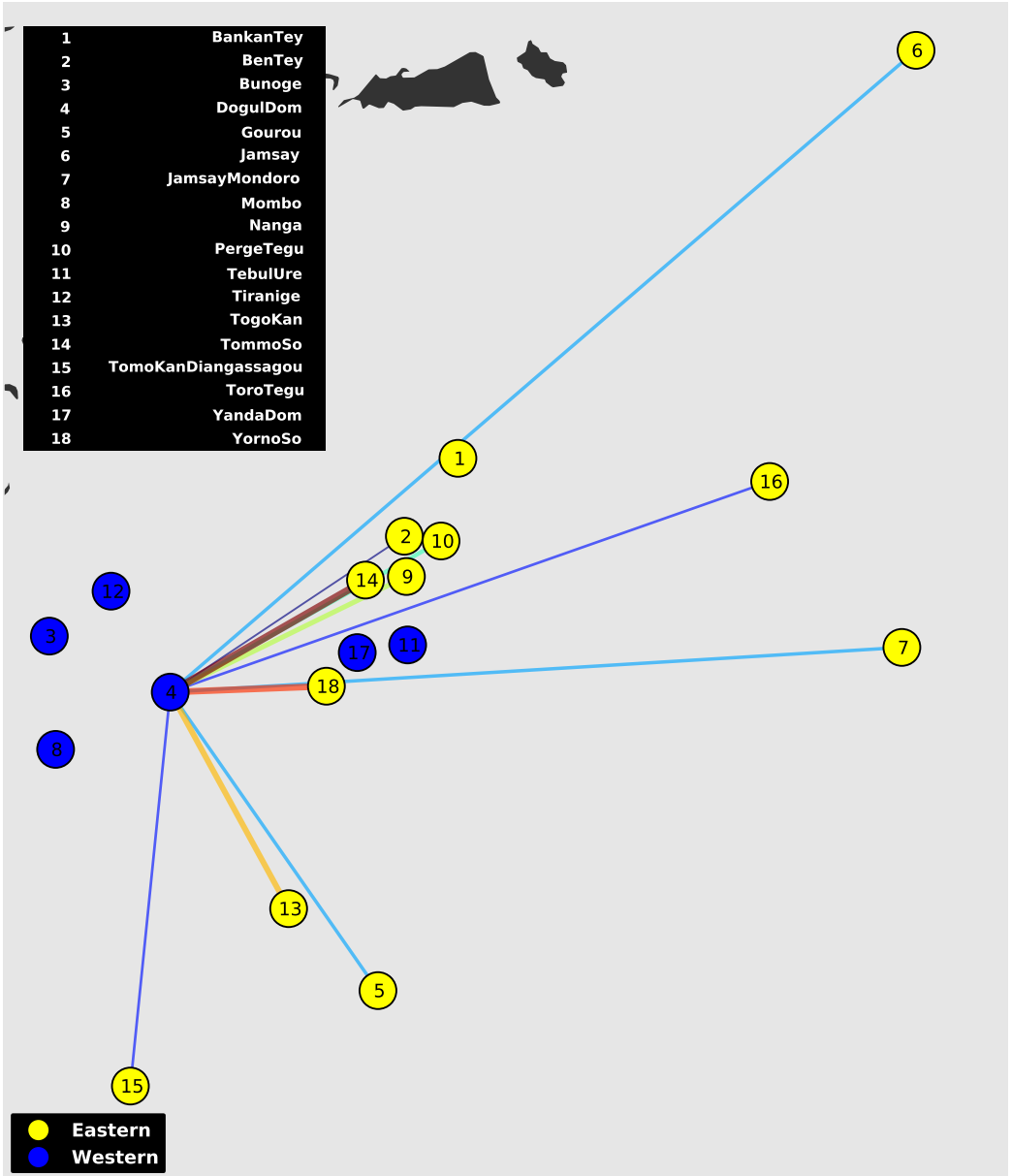


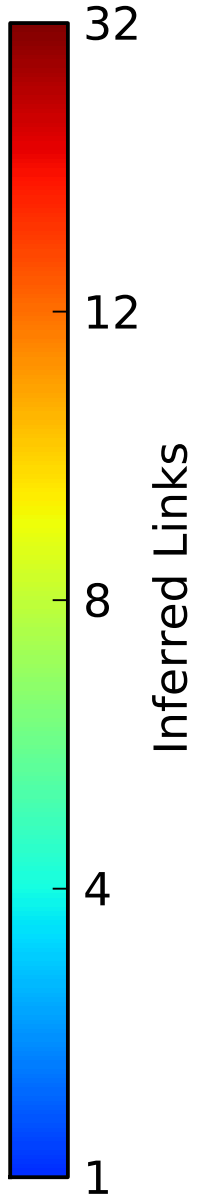
Areal Perspective: Yanda Dom



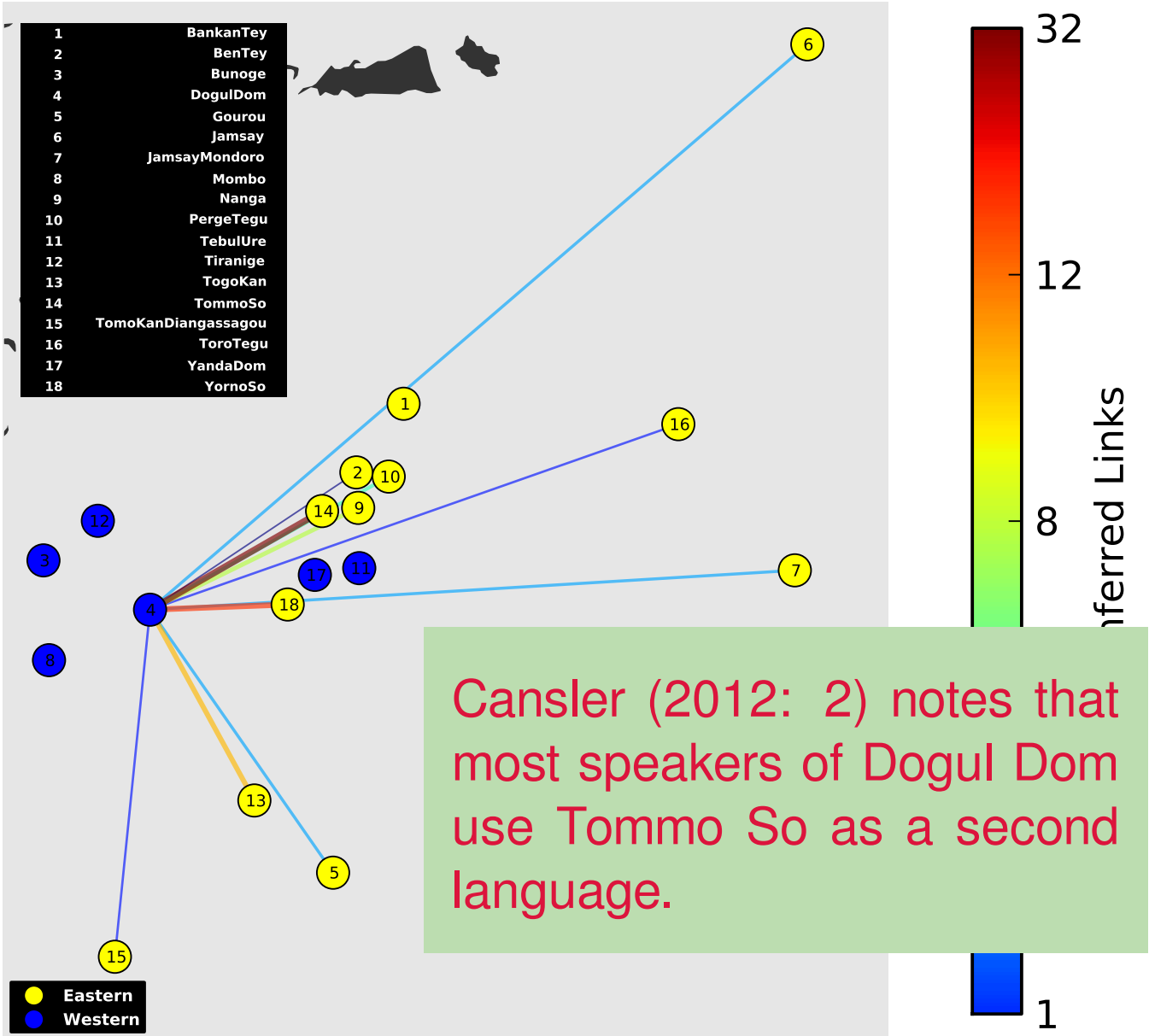
Heath (2011b: 3) notes that the use of Tommo So as a second language is common among Yanda Dom speakers.

Areal Perspective: Dogul Dom





Areal Perspective: Dogul Dom



Cansler (2012: 2) notes that most speakers of Dogul Dom use Tommo So as a second language.

?!?

!!!

Discussion

!?!

???

Natural Findings or Artifacts?

On the large scale, the results seem to confirm the method. However, given the multitude of possible errors that may have influenced our results, how can we be sure that these findings are “natural” and not artifacts of our methods?

Natural Findings or Artifacts?

Well, we can't! At least not for sure. But we can say, that our results are consistent throughout a couple of varying parameters, which makes us rather confident that it is worth pursuing our work with the methods...

Natural Findings or Artifacts?

TreeA	TreeB	B-Cubed F-Score
DLP	MrBayes	0.9539
DLP	Neighbor-Joining	0.9401
MrBayes	Neighbor-Joining	0.9464

Comparing the Impact of Varying Reference Trees

Natural Findings or Artifacts?

Varying the reference trees does only marginally change the concrete predictions of the method. Although the trees created from the data (MrBayes & Neighbor) do *not* reflect the East-West distinction of the DLP tree, the dominating role of Tommo So can still be inferred.

Natural Findings or Artifacts?

Threshold	Best Model	Origins (Ø)	MaxO	*p*-Value
0.2	3_1	1.43	4	0.35
0.3	2_1	1.64	5	0.31
0.4	2_1	1.68	5	0.15
0.5	2_1	1.65	5	0.42
0.6	1_1	2.35	7	0.45

Varying the Threshold for Cognate Detection

Natural Findings or Artifacts?

Varying the thresholds for cognate (homolog) detection clearly changes the results. The higher the threshold, the higher the amount of false positives proposed by the LexStat method. False positives, however, also often show up as patchy distributions.

Limits of the Method

Limits of the Method

- Patchily distributed cognate sets do not necessarily result from borrowings but may likewise result from
  (a) missing data,
  (b) false positives, or
  (c) coincidence.

Limits of the Method

- Patchily distributed cognate sets do not necessarily result from borrowings but may likewise result from
  (a) missing data,
  (b) false positives, or
  (c) coincidence.
- Borrowing processes do not necessarily result in patchily distributed cognate sets, especially if they occur
  (a) outside the group of languages being compared,
  (b) so frequently that they are “masked” as non-patchy distributions, or
  (c) between languages that are genetically close on the reference tree.

Examples

Basic Concept: *file (tool)* (ID: 5137)

CogID	Language	Entry	Aligned Entry
68	Bankan_Tey	kírâl	k í r â l
68	Toro_Tegu	kí:rà	k í: r à -
69	Ben_Tey	dì:sî:	d ì: s î:
69	Gourou	dì:zú	d ì: z ú
69	Jamsay	dì:jú	d ì: j ú
69	Jamsay_Mondoro	dì:jú	d ì: j ú
69	Nanga	dì:sî	d ì: s î
69	Perge_Tegu	dì:sí	d ì: s í
69	Togo_Kan	dì:sí	d ì: s í
69	Yanda_Dom	dì:zù	d ì: z ù
69	Yorno_So	dì:jú	d ì: j ú
70	Dogul_Dom	bìmbú	b ì m b ú -
70	Mombo	bí:mbyé	b í: m b y é
70	Tommo_So	bìmbú	b ì m b ú -

Examples

Basic Concept: *file (tool)* (ID: 5137)

CogID	Language	Entry	Aligned Entry
68.1	Bankan_Tey	kírâl	k í r â l
68.2	Toro_Tegu	kí:rà	k í: r à -
69.1	Gourou	dì:zú	d ì: z ú
69.1	Jamsay	dì:jú	d ì: j ú
69.1	Jamsay_Mondoro	dì:jú	d ì: j ú
69.1	Perge_Tegu	dì:sí	d ì: s í
69.1	Togo_Kan	dì:sí	d ì: s í
69.1	Yorno_So	dì:jú	d ì: j ú
69.2	Ben_Tey	dì:sî:	d ì: s î:
69.2	Nanga	dì:sî	d ì: s î
69.3	Yanda_Dom	dì:zù	d ì: z ù
70.1	Tommo_So	bìmbú	b ì m b ú -
70.2	Dogul_Dom	bìmbú	b ì m b ú -
70.2	Mombo	bí:mbyé	b í: m b y é


Thank You for Listening!