Phylogenetics and the CBGM @ CSNTM

Center for the Study of New Testament Manuscripts
12 February 2024

Joey McCollum
Australian Catholic University
Institute for Religion and Critical Inquiry
james.mccollum@myacu.edu.au
@JoeyMcCollum
jjmccollum

ACU INSTITUTE FOR RELIGION & CRITICAL INQUIRY

- To compare textual witnesses, align them at independent *variation units*
- *Variant readings* occur at variation units

ΚΑΤΑ ΛΟΥΚΑΝ 10.1-4

		1	(2)	3	
οὗ ἤμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	B ℵ C 1071 uw
οὗ ····· ·· ὐτὸς ἔρχεφθ···.	2	··λεγεν δ···	πρὸς αὐ·····	··· ····· θερ·····	$\mathfrak{P}^{75}$
οὗ ἤμελλεν αὐτὸς εἰσέρχεσθαι.	2	ἔλεγεν οὖν	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	A
οὗ ἔμελλεν ἔρχεσθαι.	2	ἔλεγεν δὲ	πρὸς αὐτούς,	Ὁ θερισμὸς	D
οὗ ἔμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν οὖν	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	Y K S Π 28 565 τ
οὗ ἔμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	L 124 579
οὗ ἤμελλεν αὐτὸς εἰσέρχεσθαι.	2	ἔλεγεν δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	Θ
οὗ ἔμελλεν αὐτὸς ἀπέρχεσθαι.	2	ἔλεγεν οὖν	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	Ω
οὗ ἔμελλεν αὐτὸς εἰσέρχεσθαι.	2	εἶπεν δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	f^{1}
οὗ ἤμελλεν αὐτὸς διέρχεσθαι.	2	ἔλεγεν δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	f^{13}
····· ····· ·····ὸς ἔρχεσθαι.	2	ἔλεγεν δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	33
οὗ ἤμελλεν αὐτὸς εἰσπορεύεσθαι.	2	ἔλεγεν δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	157
οὗ ἔμελλεν αὐτὸς πορεύεσθαι.	2	ἔλεγεν δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	700 [↓1424
οὗ ἤμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν οὖν	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	𝔐 M N U W Γ Δ Λ Ψ 2

(4)	5	6	7	
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε οὖν	τοῦ κυρίου τοῦ θερισμοῦ	ὅπως	B $\mathfrak{P}^{75}$ uwτ rell
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε οὖν	τοῦ κυρίου τοῦ θερισμοῦ	ὅπως ἂν	Y K M Π
πολύς, οἱ δέ ἐργάται ὀλίγοι·	δεήθητε	τοῦ θεοῦ τοῦ θερισμοῦ	ὅπως	D*
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε	τοῦ κυρίου τοῦ θερισμοῦ	ὅπως	D^{c}
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε οὖν	τοῦ κυρίου ····· ·····	·····	H
πολύς, οἱ δὲ ····· ·····	·····ητε οὖν	τοῦ κυρίου τοῦ θερισμοῦ	ὅπως	33
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε οὖν	τοῦ κυρίου τοῦ θερισμοῦ	ἵνα	579

Collation of Luke 10:2 with variation units numbered above text (Source: Reuben J. Swanson, ed., *New Testament Greek Manuscripts: Variant Readings Arranged in Horizontal Lines against Codex Vaticanus. Luke* [Sheffield: Sheffield Academic Press, 1995], 183)

- Analogous to a DNA sequence alignment

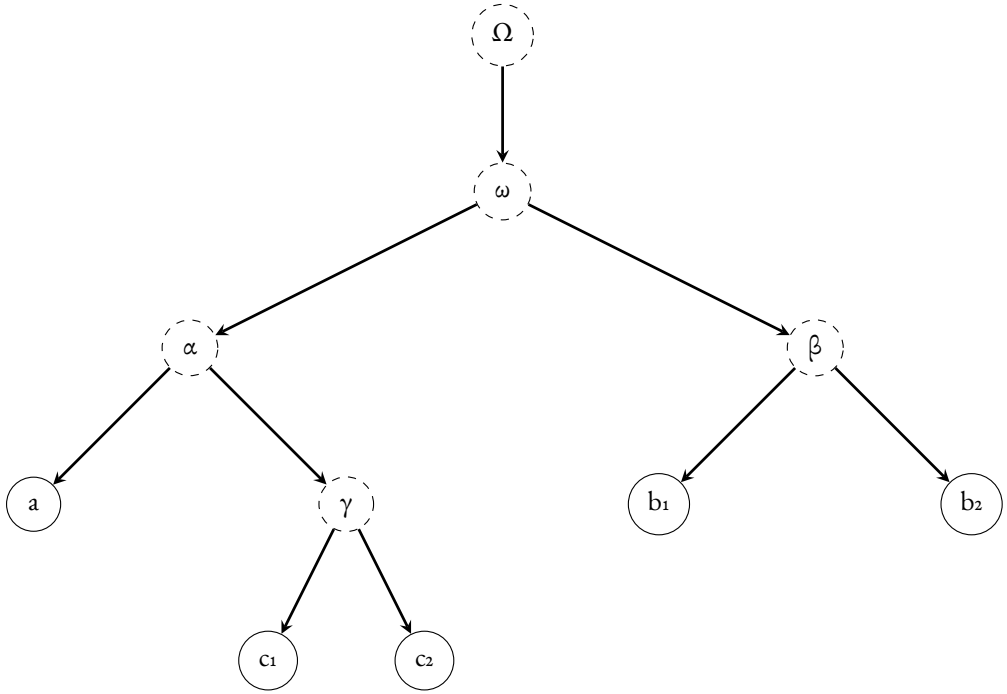
Scarites	C	T	T	A	G	A	T	C	G	T	A	C	C	A	A	-	-	-	A	A	T	A	T	T	A	C
Carenum	C	T	T	A	G	A	T	C	G	T	A	C	C	A	C	A	-	T	A	C	-	T	T	T	A	C
Pasimachus	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	T	A	T	A	A	G	T	T	T	A	C
Pheropsophus	C	T	T	A	G	A	T	C	G	T	T	C	C	A	C	-	-	-	A	C	A	T	A	T	A	C
Brachinus armiger	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	T	A	T	A	T	T	C
Brachinus hirsutus	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	T	A	T	A	T	A	C
Aptinus	C	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	C	A	A	T	T	A	C
Pseudomorpha	C	T	T	A	G	A	T	C	G	T	A	C	C	-	-	-	-	-	A	C	A	A	A	T	A	C

- Rows: *taxa* = witnesses
- Columns: *sites* = variation units
- Cells: *states* = variant readings (including omissions)
  - Lacunae and uncertain retroversions correspond to fully or partially *ambiguous states*

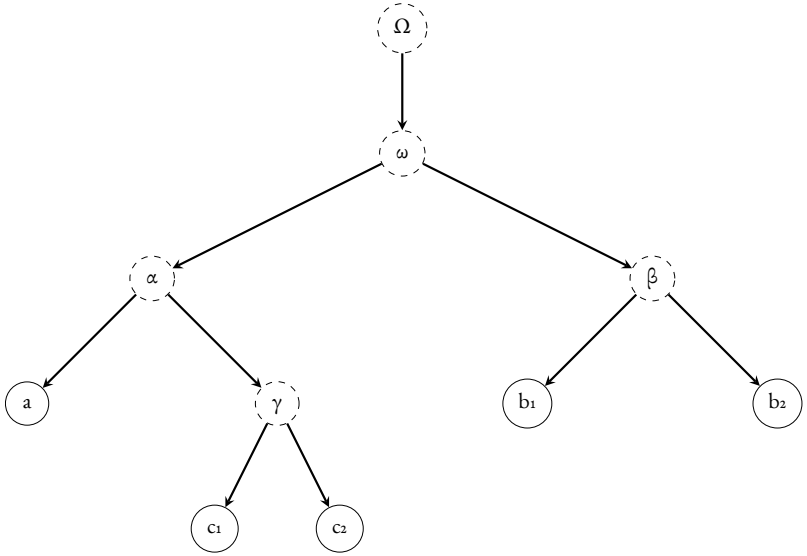
- At the most basic level, a *witness* is just a sequence of readings, a row in the collation

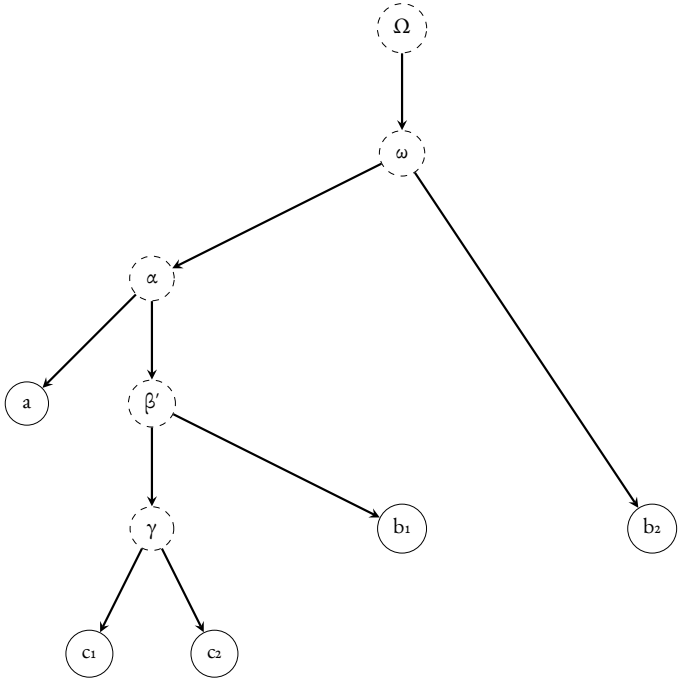
	3Jo 1:1/2	3Jo 1:1/6	3Jo 1:1/8	…	3Jo 1:15/23
GA 69	a	af1	a	…	a
GA 1739	a	a	b	…	a
GA 2243	b	a	a	…	a

- Paratextual features can be encoded in the same way
- In more complex phylogenetic approaches, age can also be incorporated



- A *stemma* is a "family tree" modeling the transmission of the text
- The *leaves* (solid circles) correspond to extant witnesses
- The *hyparchetypes* (dashed circles) are hypothetical (now-lost) ancestors, reconstructed from their descendents along the branches
- The *archetype* (ω) is the earliest reconstructible text
- The *root* (Ω) represents the authorial text



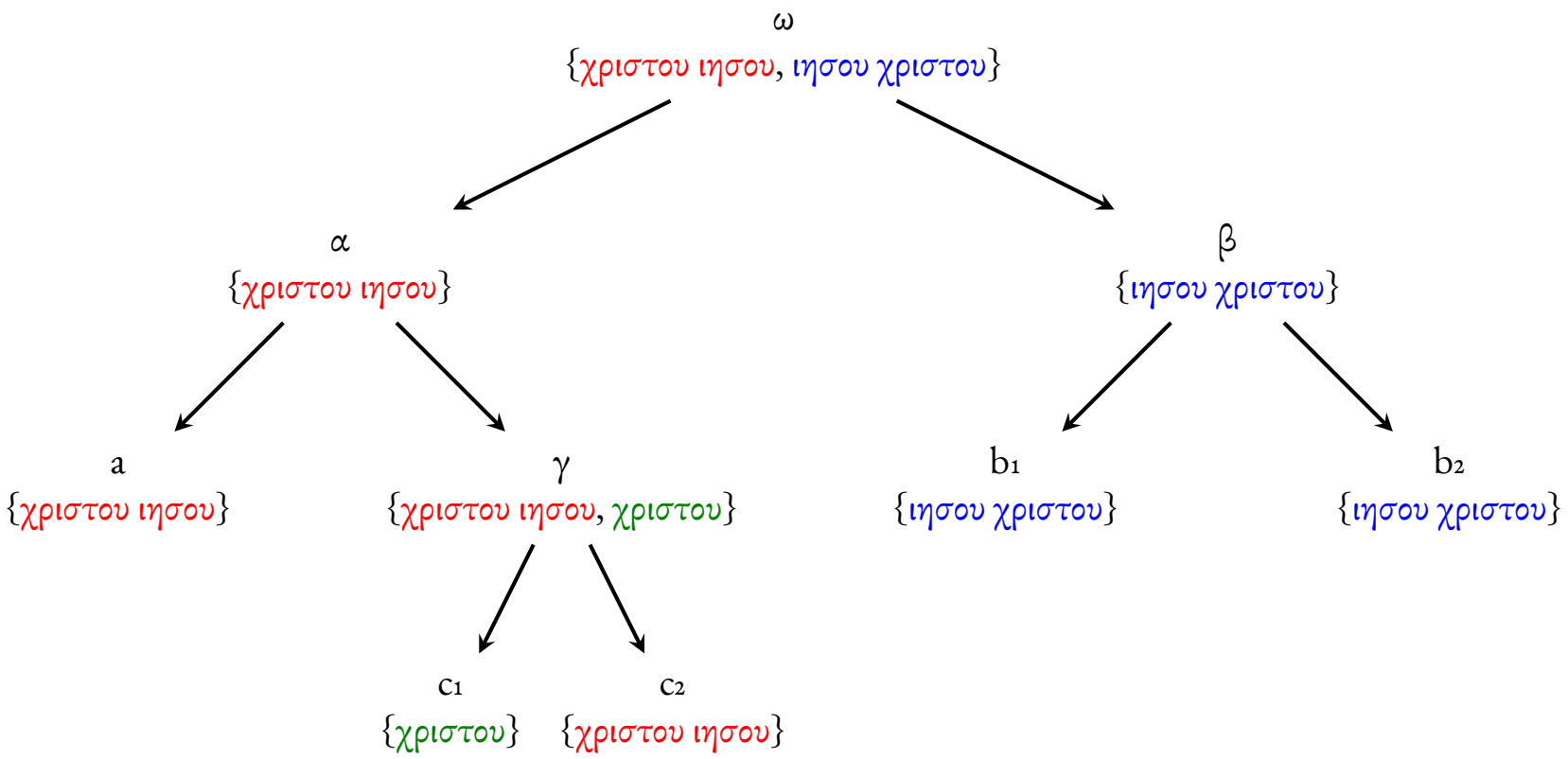


- A stemma represents a hypothesis about transmission history
- The goal is to determine which hypothesis (or hypotheses) best explain the extant data
- To do this, we need a numerical metric or “score” for the fitness of a given stemma

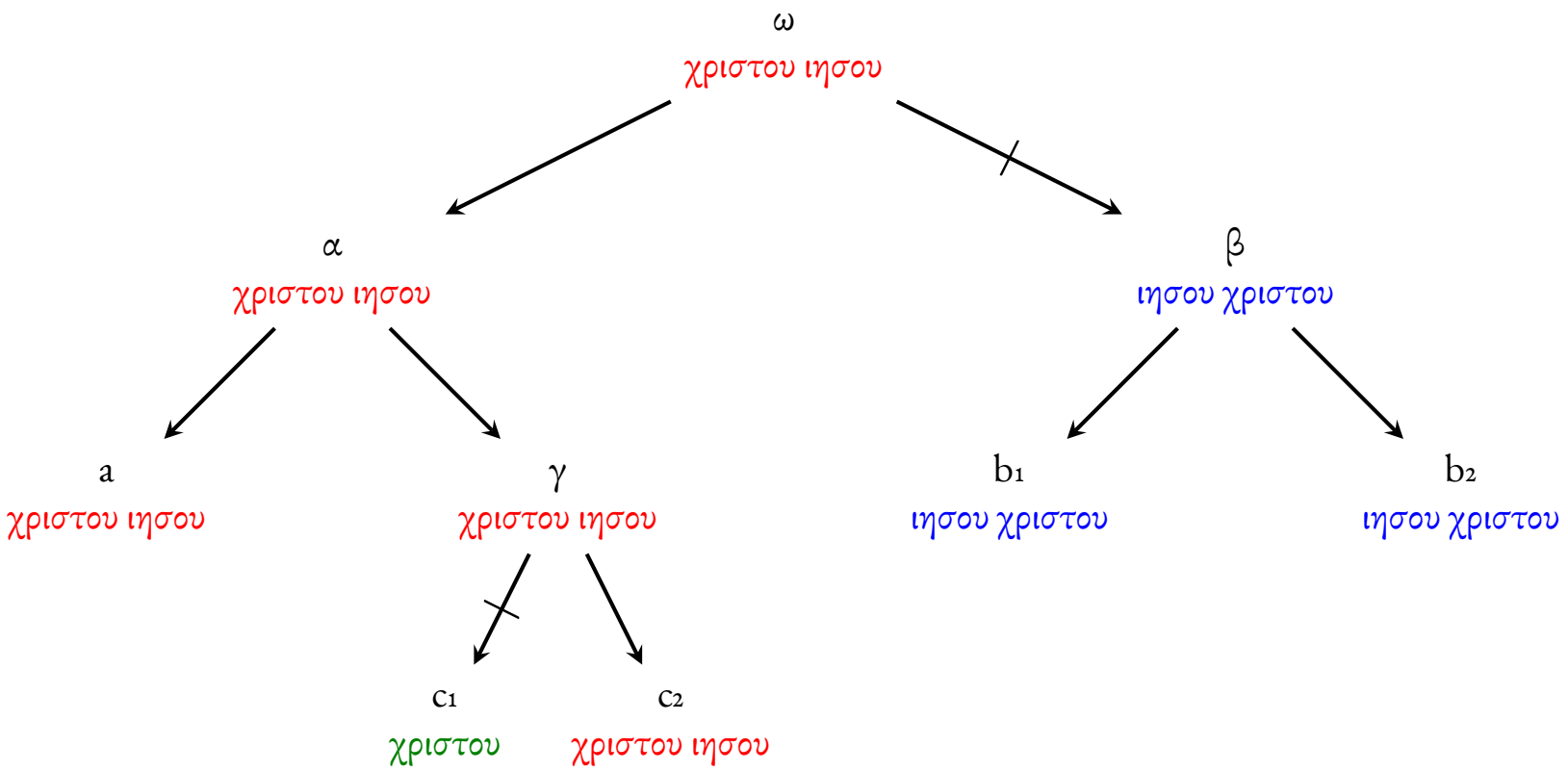
- One such metric is *parsimony*
- The smallest number of times one reading has to change to another along the branches of the stemma
- Motivated by Ockham's Razor
- Given a candidate stemma, we calculate its parsimony score for each variation unit independently, then add up the results
- Can be efficiently computed in a bottom-up fashion

Witness	Reading
a	χριστου ιησου
b_1	ιησου χριστου
b_2	ιησου χριστου
c_1	χριστου
c_2	χριστου ιησου

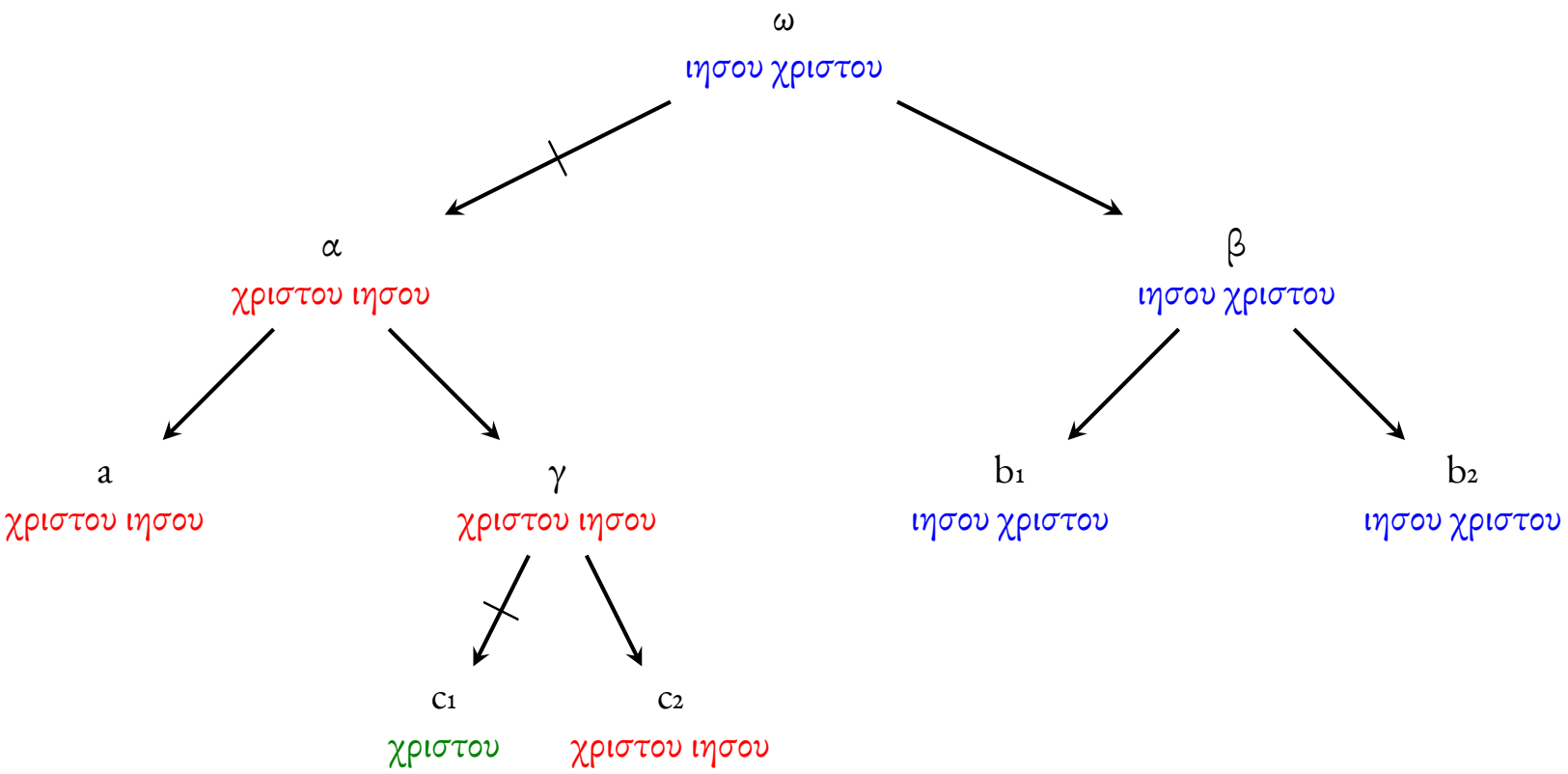
Cost for stemma 1:



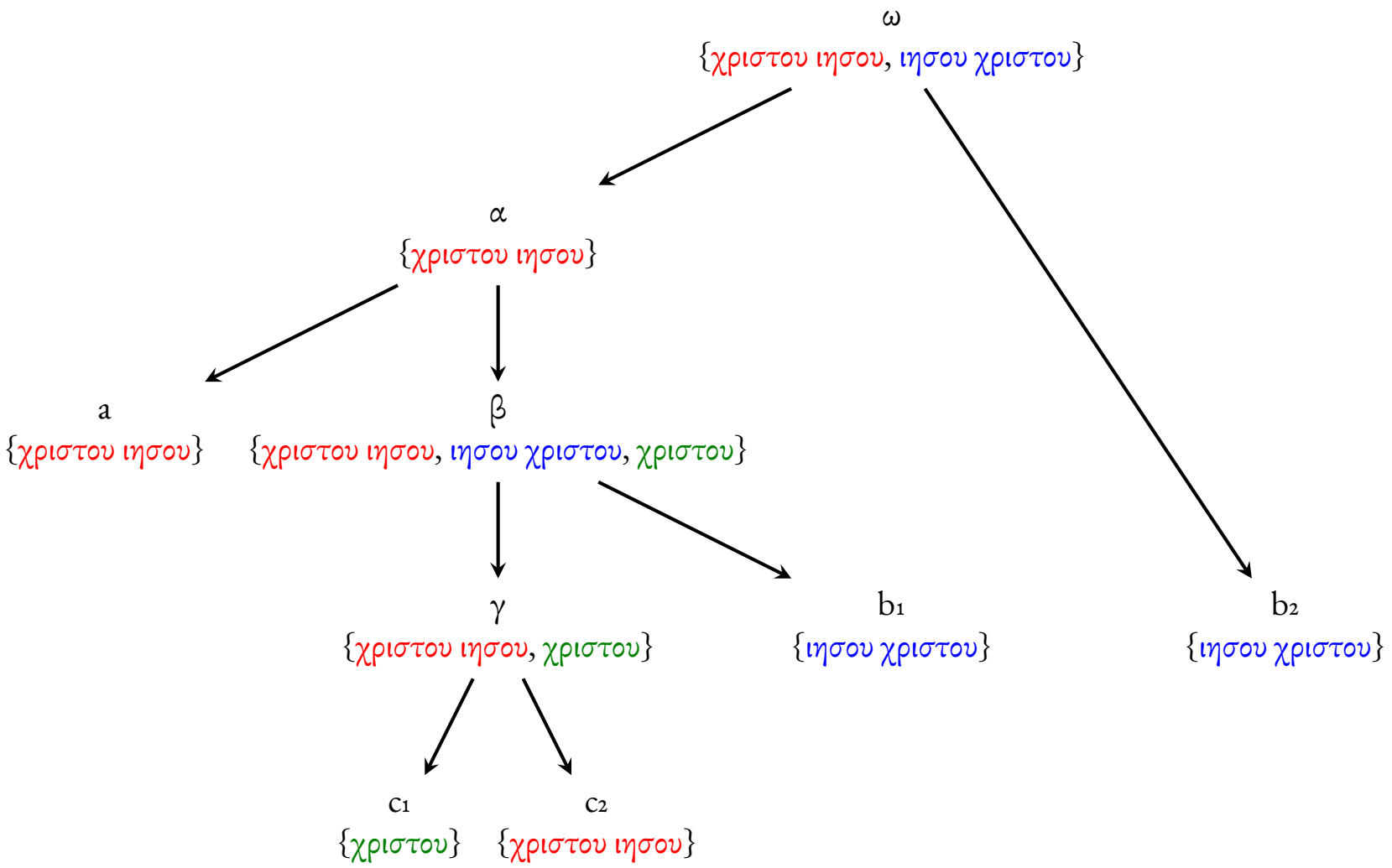
Cost for stemma 1: **2**



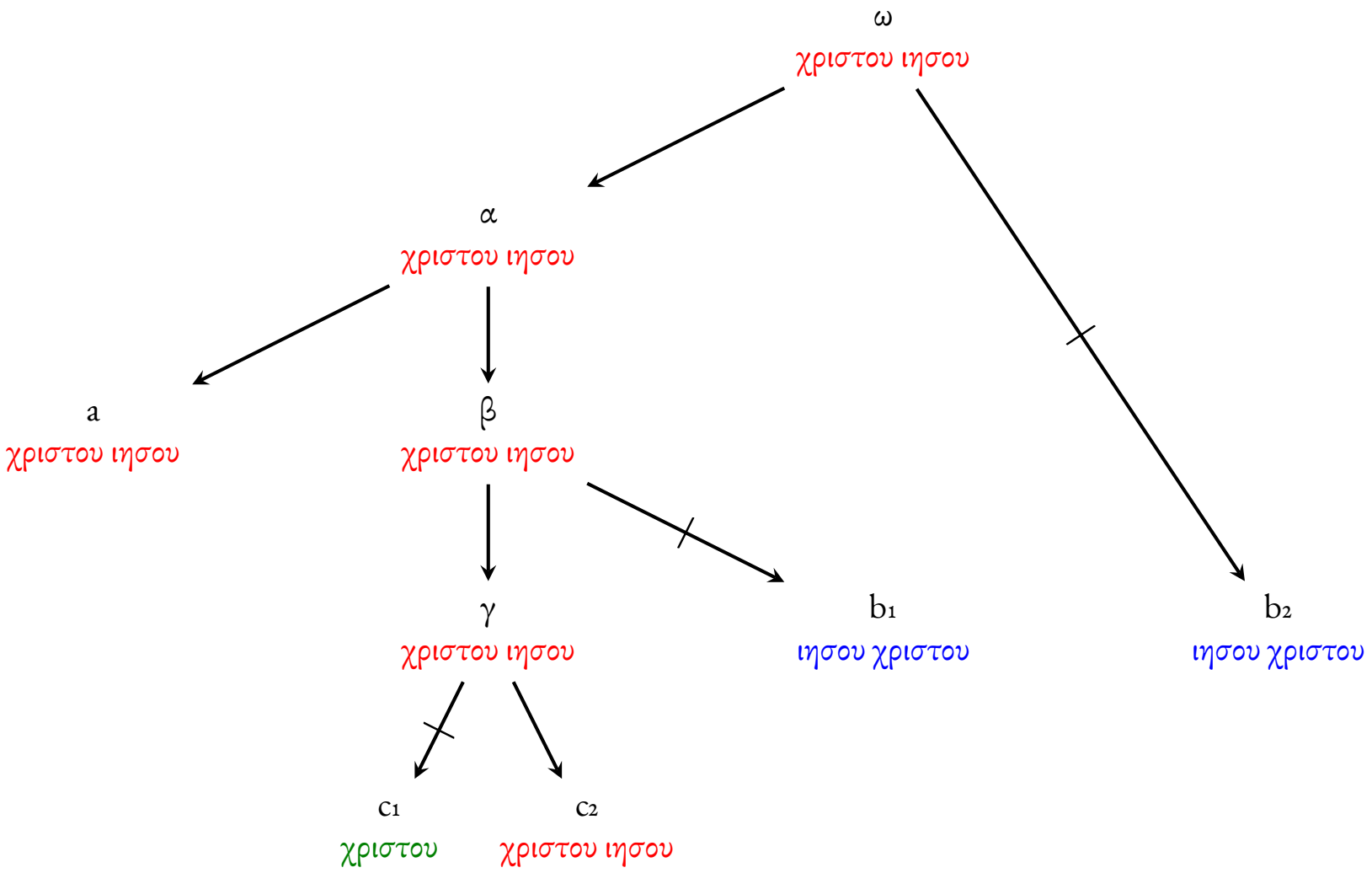
Cost for stemma 1: **2**



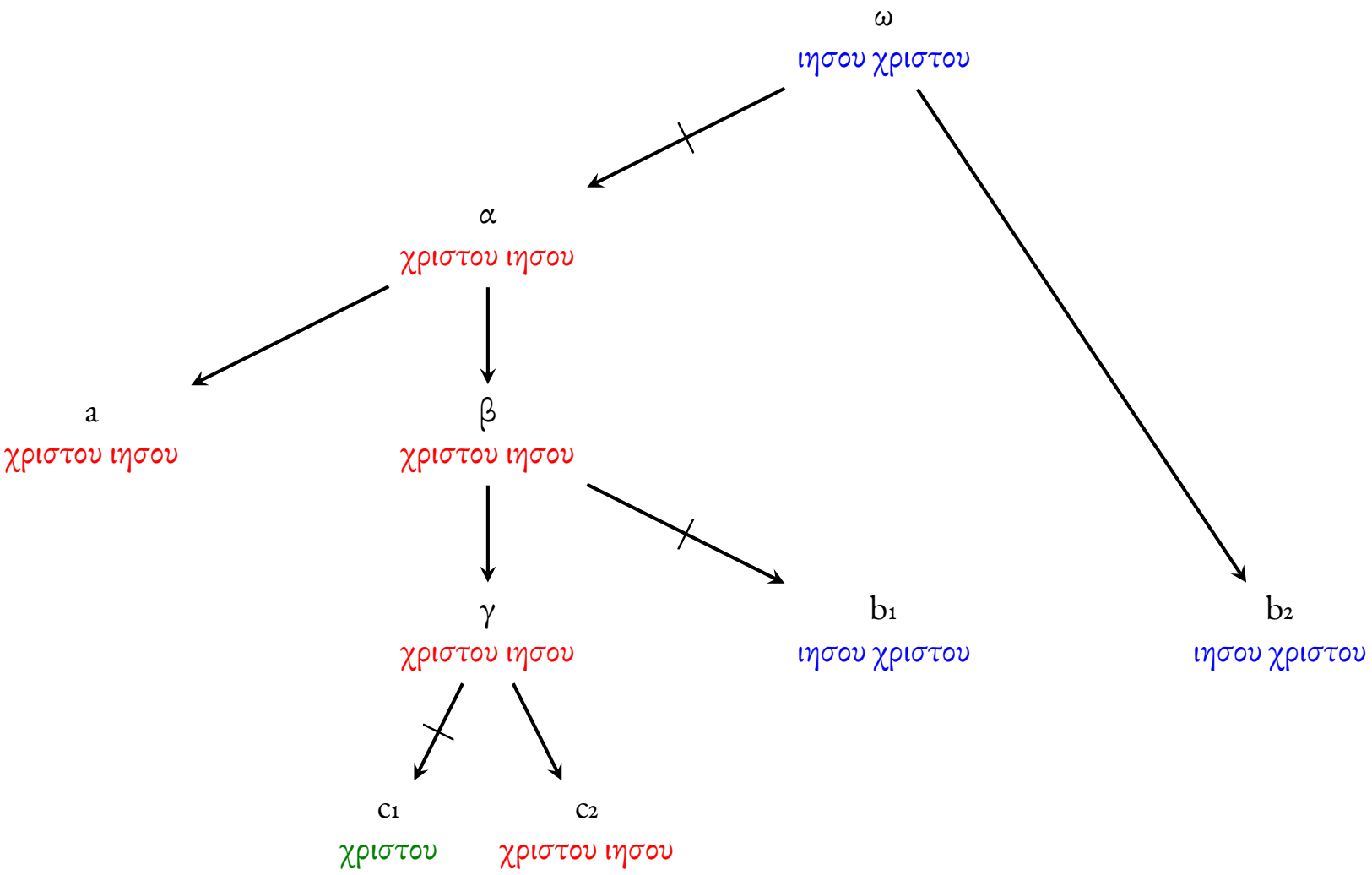
Cost for stemma 2:



Cost for stemma 2: **3**



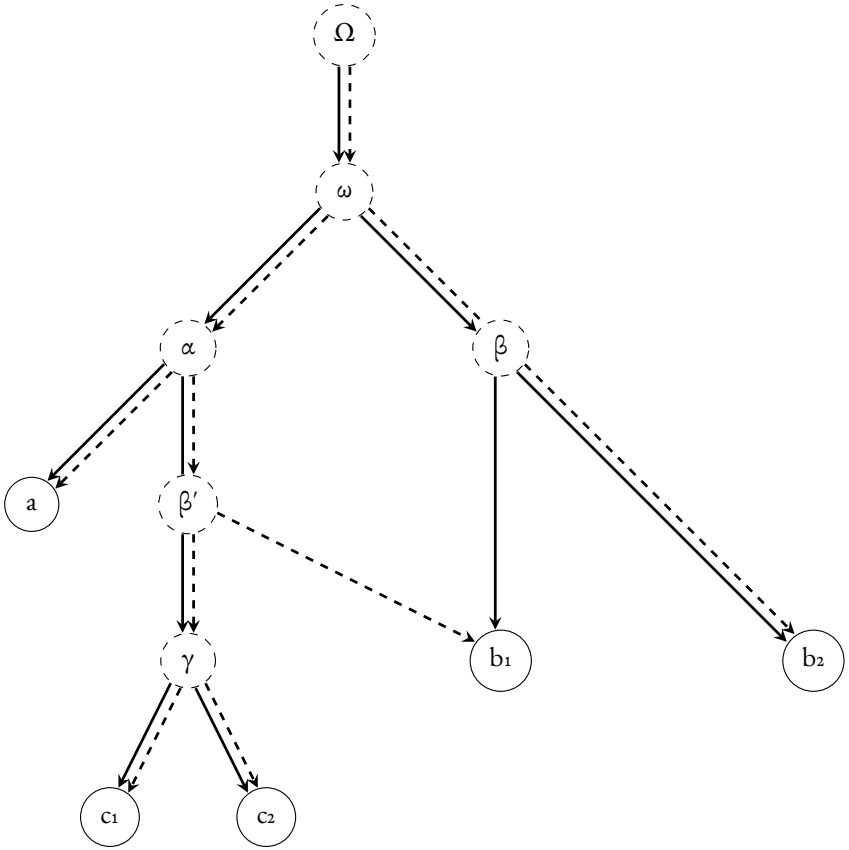
Cost for stemma 2: **3**



- Developed over thirty years by Gerd Mink, culminating in the latest updates to the *Editio Critica Maior* (*ECM*)
- Important reading:
  - Gerd Mink, "Problems of a Highly Contaminated Tradition: The New Testament. Stemmata of Variants as a Source of a Genealogy for Witnesses," in *Studies in Stemmatology II*, ed. Pieter van Reenen, August den Hollander, and Margot van Mulken (Amsterdam: John Benjamins Publishing, 2004), 13–85
  - Peter J. Gurry, *A Critical Examination of the Coherence-Based Genealogical Method in New Testament Textual Criticism*, NTTSD 55 (Leiden: Brill, 2017)
  - Tommy Wasserman and Peter J. Gurry, *A New Approach to Textual Criticism: An Introduction to the Coherence-Based Genealogical Method*, RBS 80 (Atlanta, GA: SBL Press, 2017)
  - Andrew Charles Edmondson, "An Analysis of the Coherence-Based Genealogical Method Using Phylogenetics" (PhD diss., University of Birmingham, 2019), https://etheses.bham.ac.uk/id/eprint/9150/

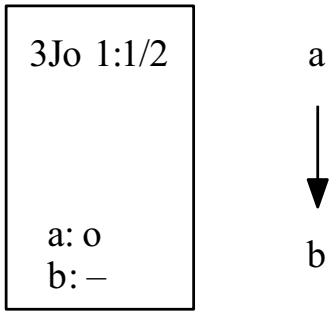
- *Not* a way to make computers do textual criticism, but a way for them to help us refine our judgments
- *Not* a new methodology for evaluating variant readings, but a "meta-approach" to be used on top of existing methods

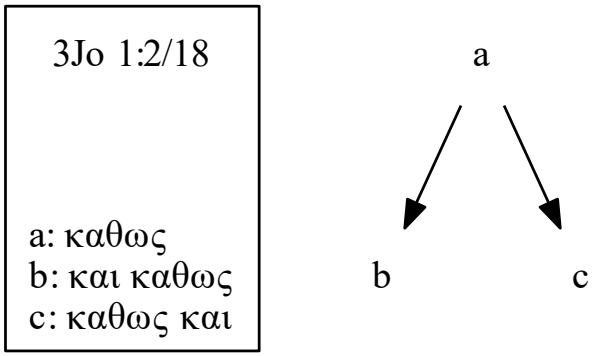
- Intended to solve *contamination*, or mixture across branches of the textual tradition



- Foundational principles:
  1. Scribes typically copied their exemplars with fidelity.
  2. If a scribe introduced a variant, then it came from some other reading.
  3. Scribes typically used fewer sources rather than many.
  4. Scribes typically used closely related sources rather than distant ones.
- Witnesses are *texts* (sequences of readings) minus the material baggage (date, provenance, etc.)
  - "How texts relate" $\neq$ "How manuscripts relate"

- The basic unit of comparison
- One for each variation unit
- A graphical representation of our judgments of readings

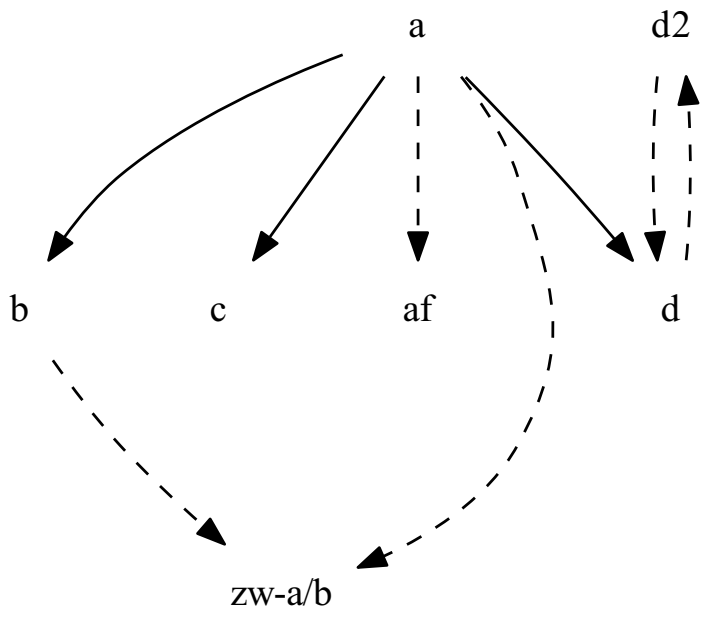




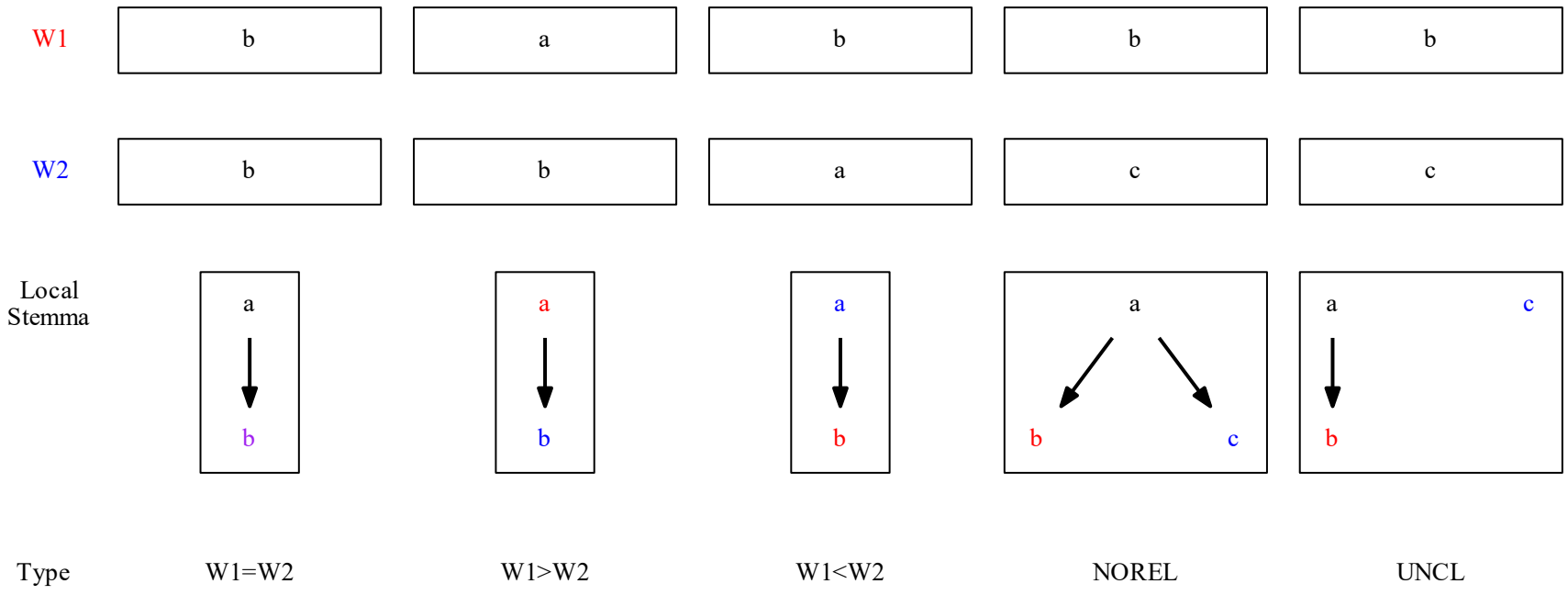
- Some are more complicated
  - *defective* readings (e.g., obvious misspellings)
  - *orthographic* readings (e.g., regional differences)
  - *split* attestations of the same reading (coincidental agreement)
  - *ambiguous* readings
- Some of these may be collapsed with other substantive readings

3Jo 1:4/22-26

a: εν αληθεια περιπατουντα
af: εν αληθεια περιπατουντο
b: εν τη αληθεια περιπατουντα
c: περιπατουντα εν αληθεια
d: τη αληθεια περιπατουντα
zw-a/b: εν [13-15]τουντα

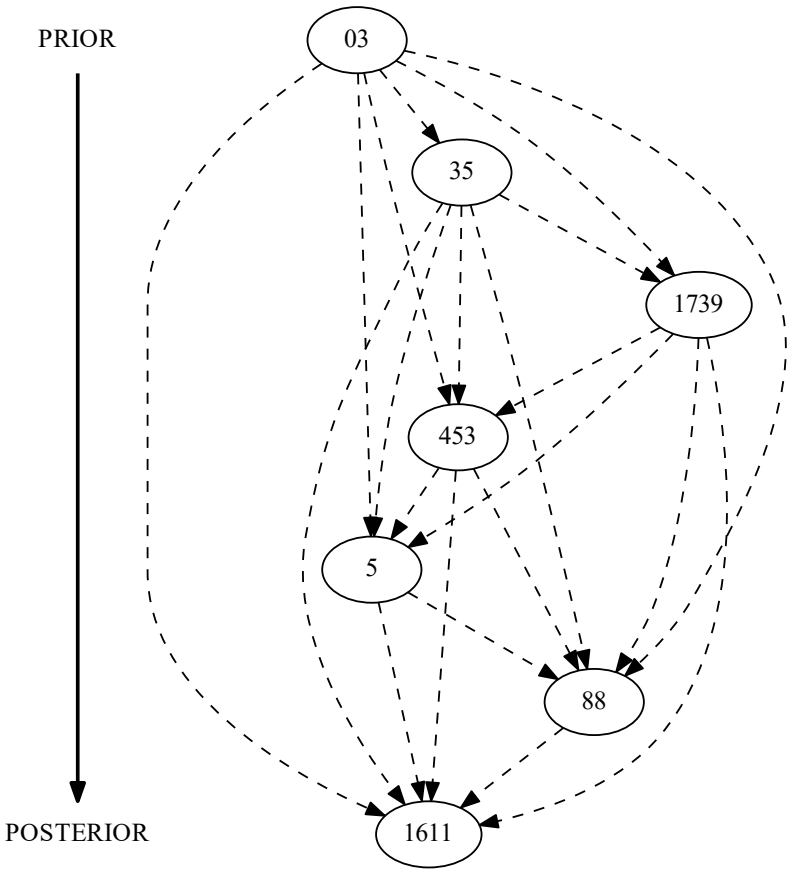


- The relationship of two witnesses is the overall pattern *of the relationships of their readings* at all variation units where both are extant



- The first three are the most important

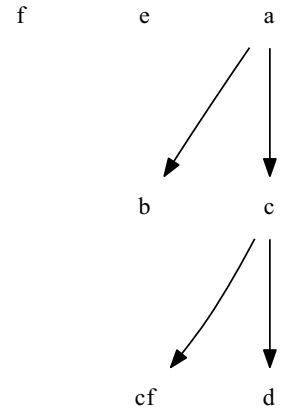
- Potential ancestor = “more prior than posterior readings”

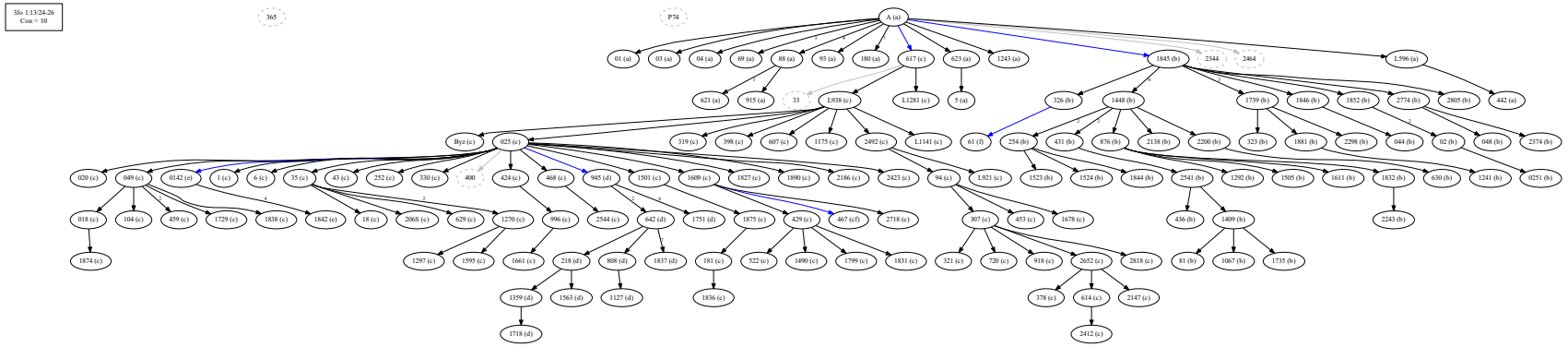


- *Textual flow* is a useful tool for helping us revise our judgments in a local stemma
- *Not* a global stemma (our ultimate goal), but still important

3Jo 1:13/24-26

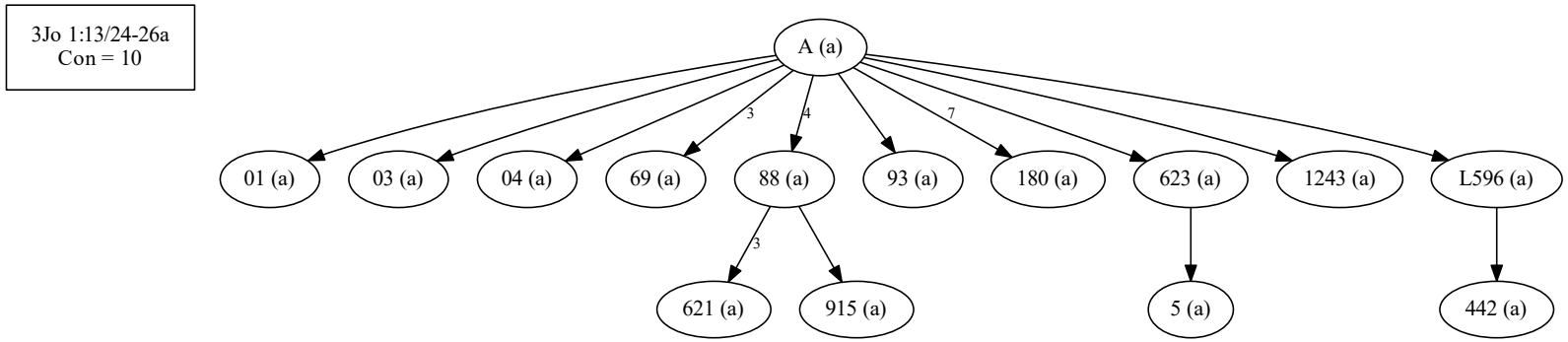
a: σοι γραφειν
b: γραφειν σοι
c: σοι γραψαι
cf: σοι σοι γραψαι
d: γραψαι σοι
e: γραψαι
f: –





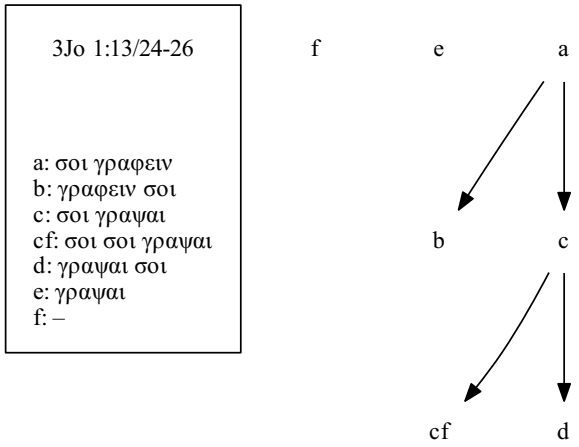
- How do we find a given witness's *textual flow ancestor*?
- We specify a *connectivity limit* κ (i.e., a radius of "close-enough" neighbors)
- Then, for each witness:
  1. List its potential ancestors, sorted from most agreement to least
  2. If one of the first κ has the same reading at this unit, then select it
  3. If not, then choose the first (non-lacunose) potential ancestor
- Core idea: use *general relationships* (between witnesses) to find *specific relationships* (between readings in a local stemma)

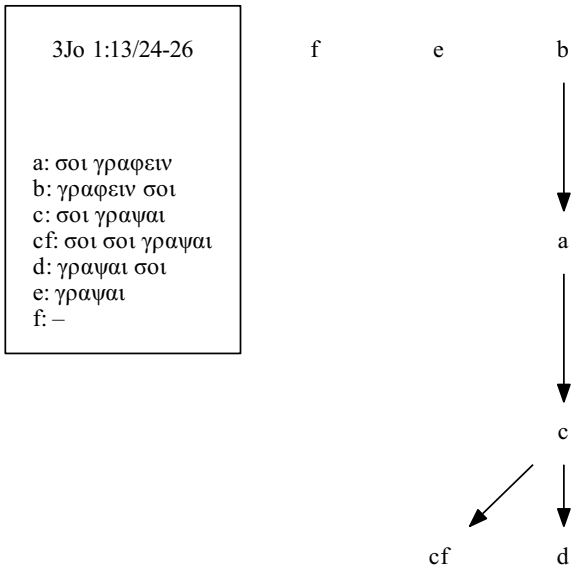
- Often, we just want to know the textual flow for witnesses with a specific reading



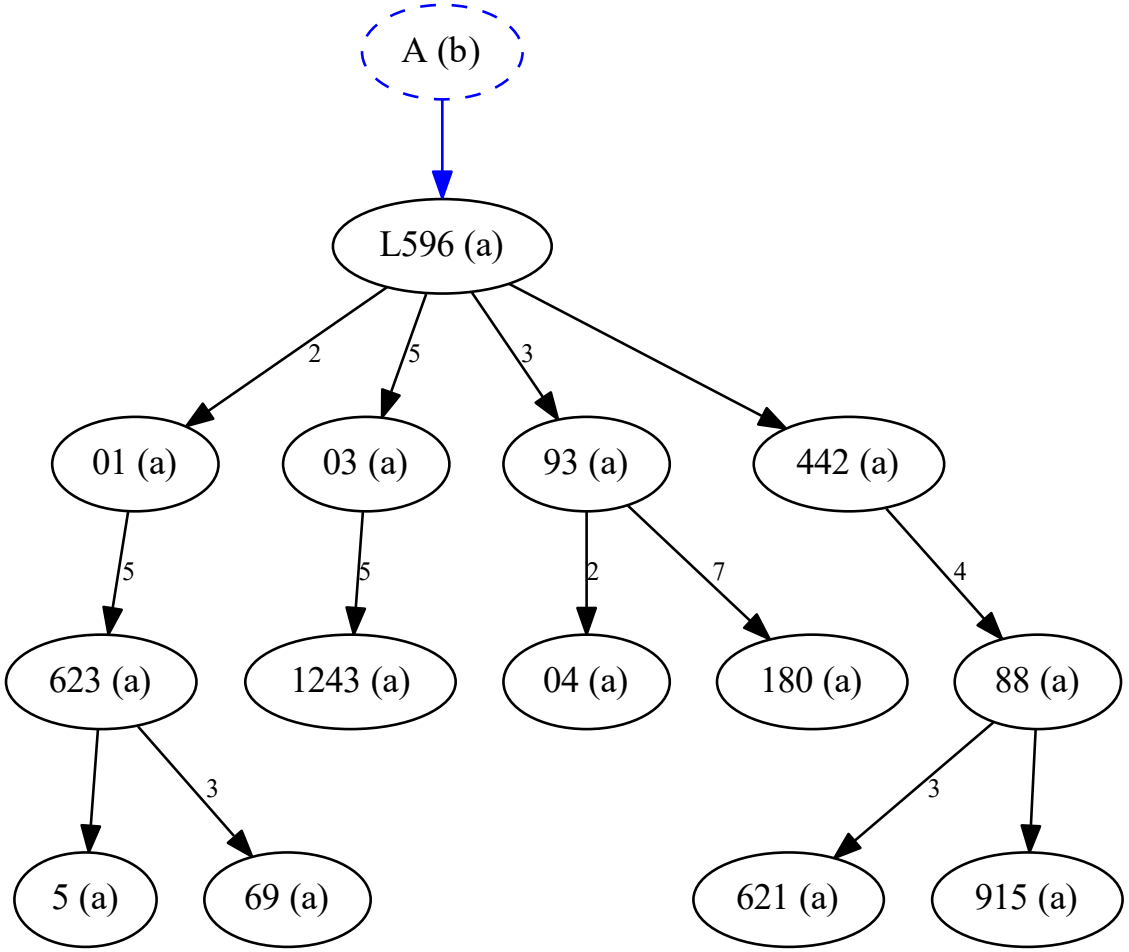
- (Numbers on edges represent the rank of the closest potential ancestor with the same reading, if it's not 1)

- We can use it to evaluate alternate hypotheses about the initial text (A)

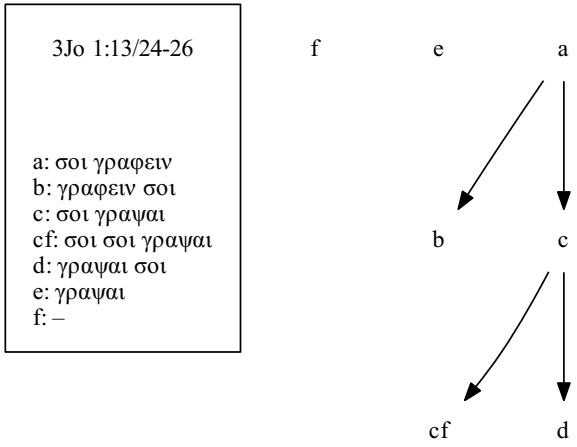


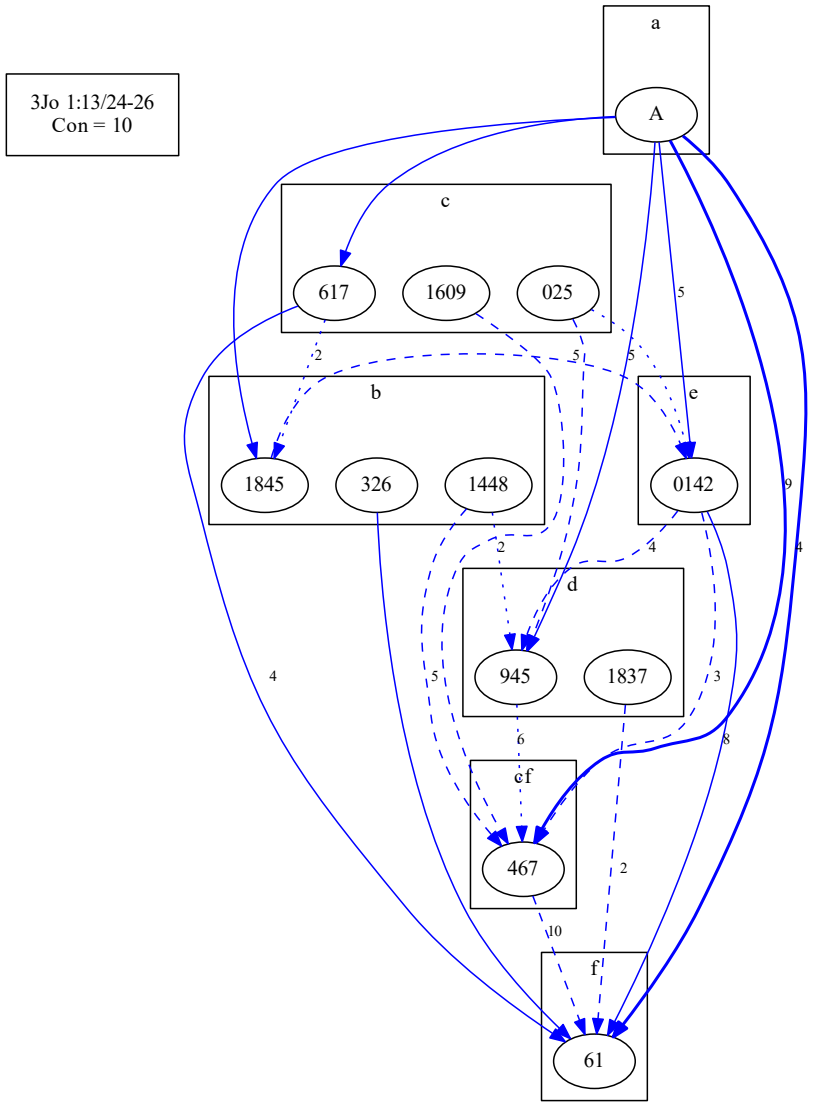


3Jo 1:13/24-26a
Con = 10



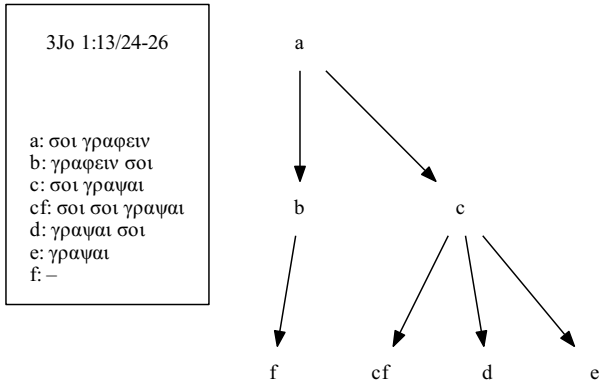
- Or, we can look only at the parts of textual flow where a reading gets changed to find the most likely sources of unexplained readings (*e* and *f*)



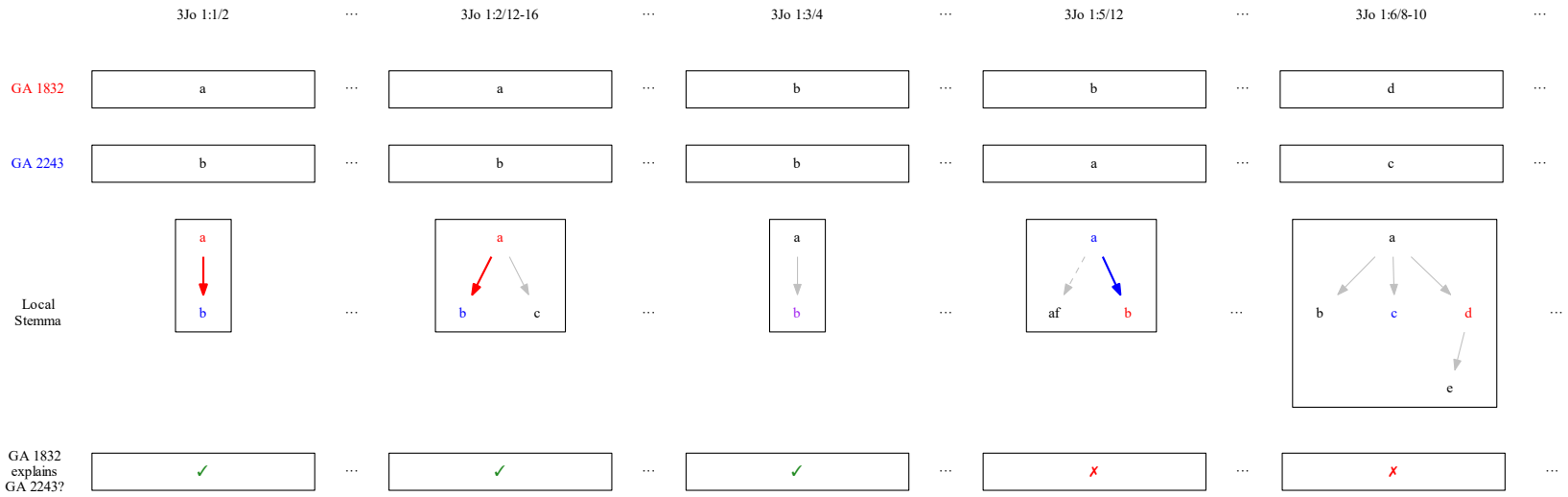

3Jo 1:13/24-26
Con = 10
a
A
c
617
1609
025
b
1845
326
1448
e
0142
d
945
1837
cf
467
f
61

Textual Flow for a Variant Reading

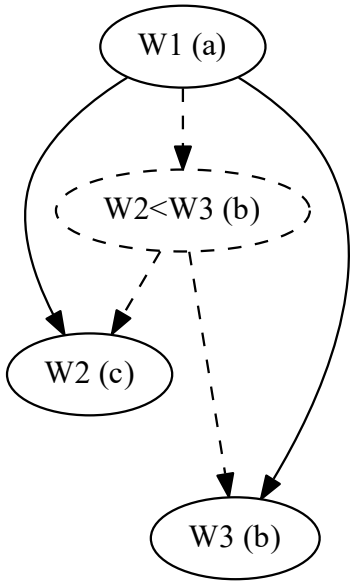
- Using this information, we can attempt to explain previous unexplained readings
- A necessary step for our ultimate goal of constructing a global stemma



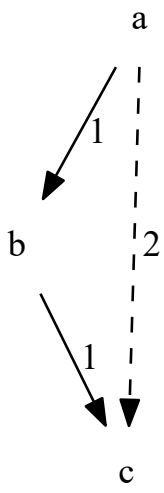
- We say that one reading *explains* another if
  - it is the same reading (explanation by agreement), or
  - there is an edge in the local stemma from it to the other reading

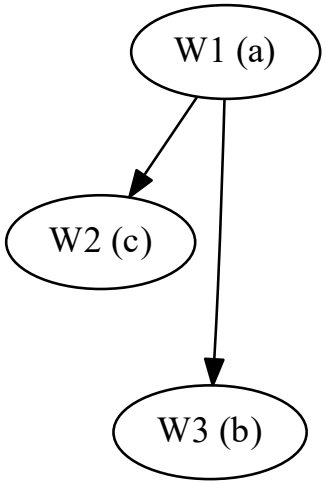


- Lacunae do not have to be explained, and they cannot explain readings



- Does a reading explain any of its posterior readings transitively (i.e., in the local stemma to the left, does *a* explain *c*)?
- As originally formulated, *no*: *a* explains *b* and *b* explains *c*, but *a* does not explain *c* (it's too many steps removed)
- Later, in the global stemma, *intermediary nodes* may be needed to ensure that all readings are explained



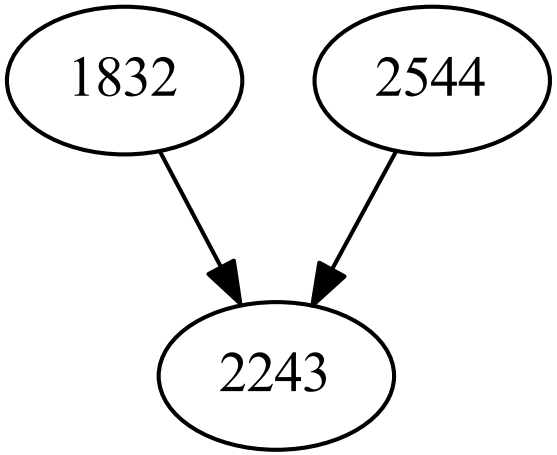


- If we instead allow *a* to explain *c*, but at a higher cost (more on this in the substemma slides), then we remove the need for intermediary nodes (although multiple changes in the same variation unit may be implied along an edge in the global stemma)

The Substemma(ta) of a Witness

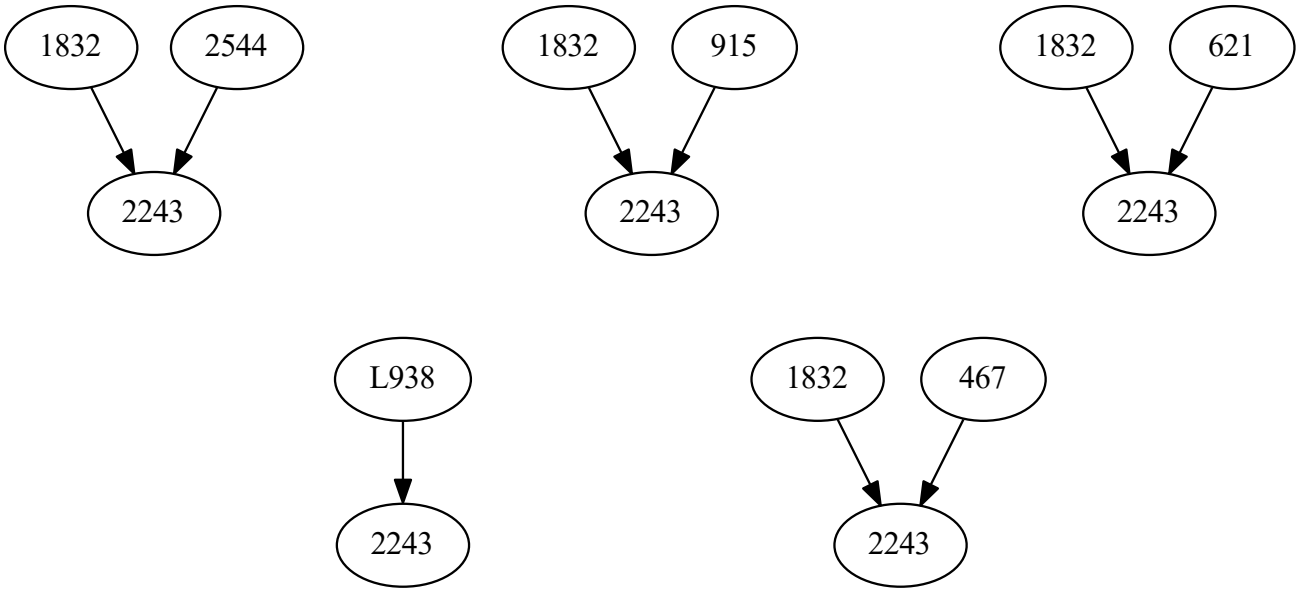
- The *substemma* of a witness is the portion of the global stemma consisting of the witness and its ancestors in the stemma
- Requirement: *every* extant reading in the witness must be explained by a reading in at least one of its ancestors

Explained by GA 1832	…	✗	✓	✓	✓	…
Explained by GA 2544	…	✓	✗	✗	✓	…
Explained by Either	…	✓	✓	✓	✓	…



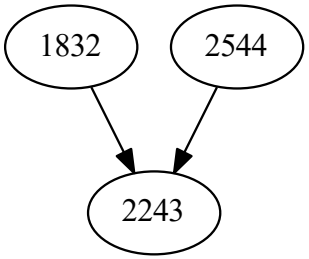
The Substemma(ta) of a Witness

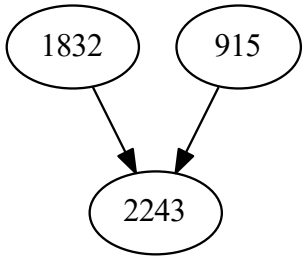
- A witness may have multiple valid substemma (i.e., ones that explain all of its readings), but some are better than others
- Two of the CBGM's methodological assumptions are important here:
  3. Scribes typically used fewer sources rather than many.
  4. Scribes typically used closely related sources rather than distant ones.

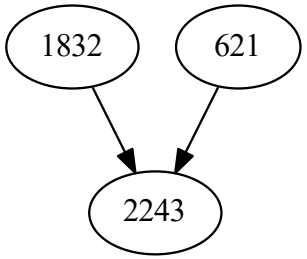


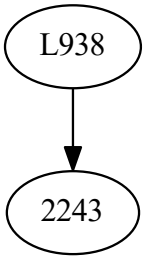
The Substemma(ta) of a Witness

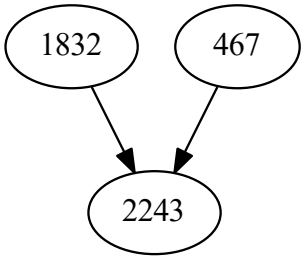
- Based on assumption 3, we should prefer substemmata with fewer ancestors ("parsimony")
- Based on assumption 4, we should prefer substemmata with ancestors that agree as often as possible with the witness
- A balancing act: the substemma {L938} is more parsimonious, but may not explain as many readings by agreement





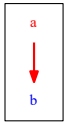
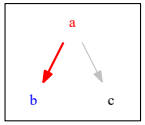
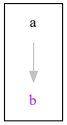
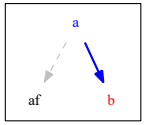
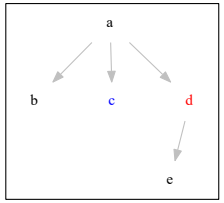






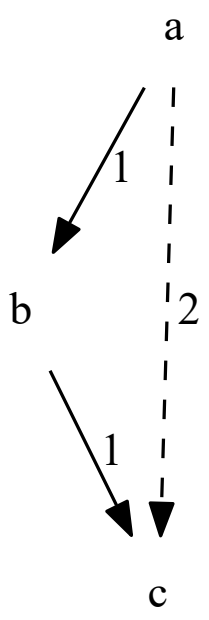
The Substemma(ta) of a Witness

- A simple cost function for each ancestor is “the number of variation units where the ancestor explains the witness by descent and not agreement”

	3Jo 1:1/2	…	3Jo 1:2/12-16	…	3Jo 1:3/4	…	3Jo 1:5/12	…	3Jo 1:6/8-10	…
GA 1832	a	…	a	…	b	…	b	…	d	…
GA 2243	b	…	b	…	b	…	a	…	c	…
Local Stemma		…		…		…		…		…
GA 1832 explains GA 2243?	✓	…	✓	…	✓	…	✗	…	✗	…
Cost	1	…	1	…	0	…	0	…	0	…

The Substemma(ta) of a Witness

- If we allow a reading to explain any reading posterior to it, then a better cost per variation unit is the length of the path from the prior reading to the posterior one.

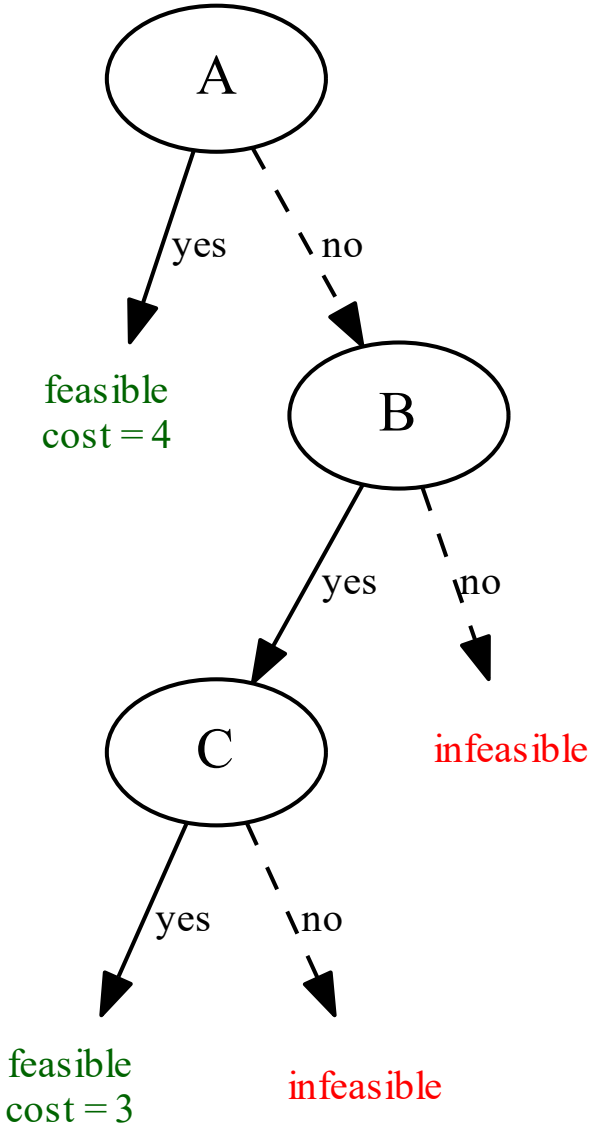


Finding a (Good) Substemma

- Also called *substemma optimization*
- For n potential ancestors, a *weighted set cover* problem with n sets (and $2^n - 1$ combinations to check!)

Substemma	Variation Units Explained				Cost
{A}	✓	✓	✓	✓	4
{B}	✓	✓	✗	✗	1
{C}	✗	✓	✓	✓	2
{A, B}	✓	✓	✓	✓	4+1=5
{A, C}	✓	✓	✓	✓	4+2=6
{B, C}	✓	✓	✓	✓	1+2=3
{A, B, C}	✓	✓	✓	✓	1+2+4=7

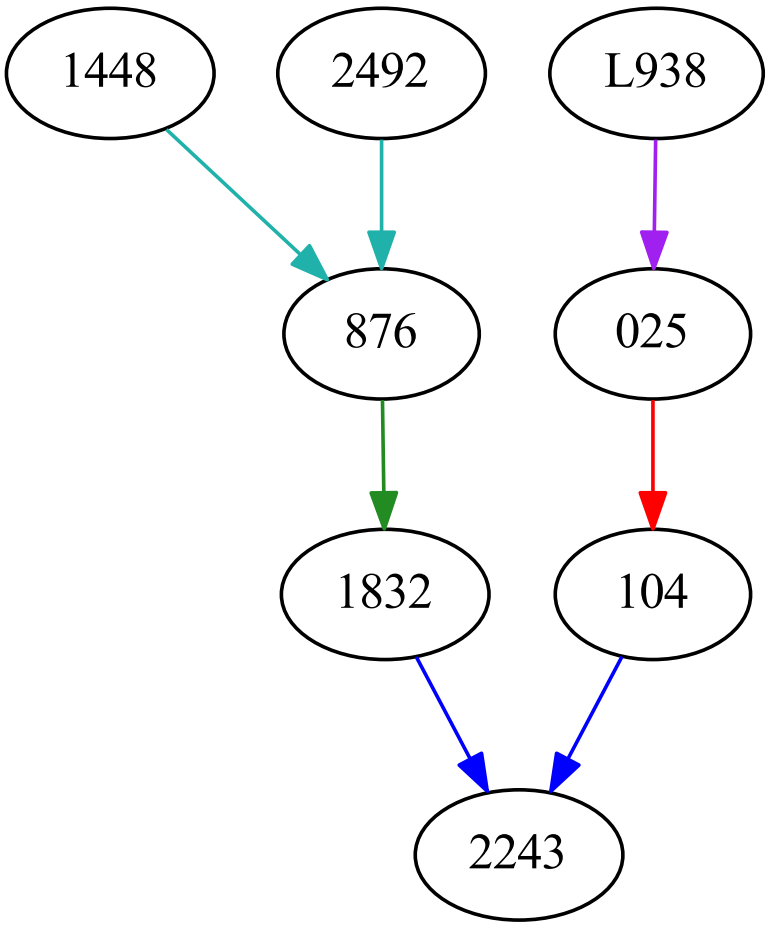
Finding a (Good) Substemma



- If a witness has many potential ancestors, then checking all $2^n - 1$ possible substemmata by brute force is prohibitive
- The *branch-and-bound* heuristic (pictured left) finds all minimum-cost substemmata quickly in practice
- Easily adapted to find all substemmata within a given cost

The Global Stemma

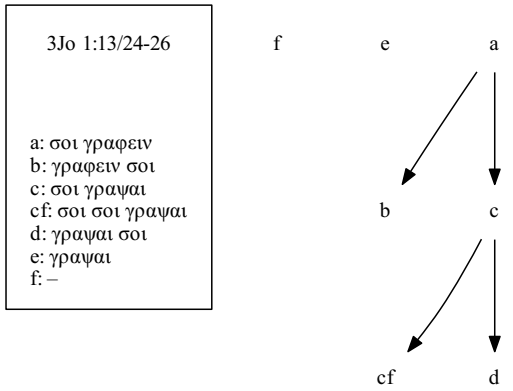
- Just as the local stemma relates readings, the *global stemma* relates witnesses
- Combination of all substemmata into a single graph

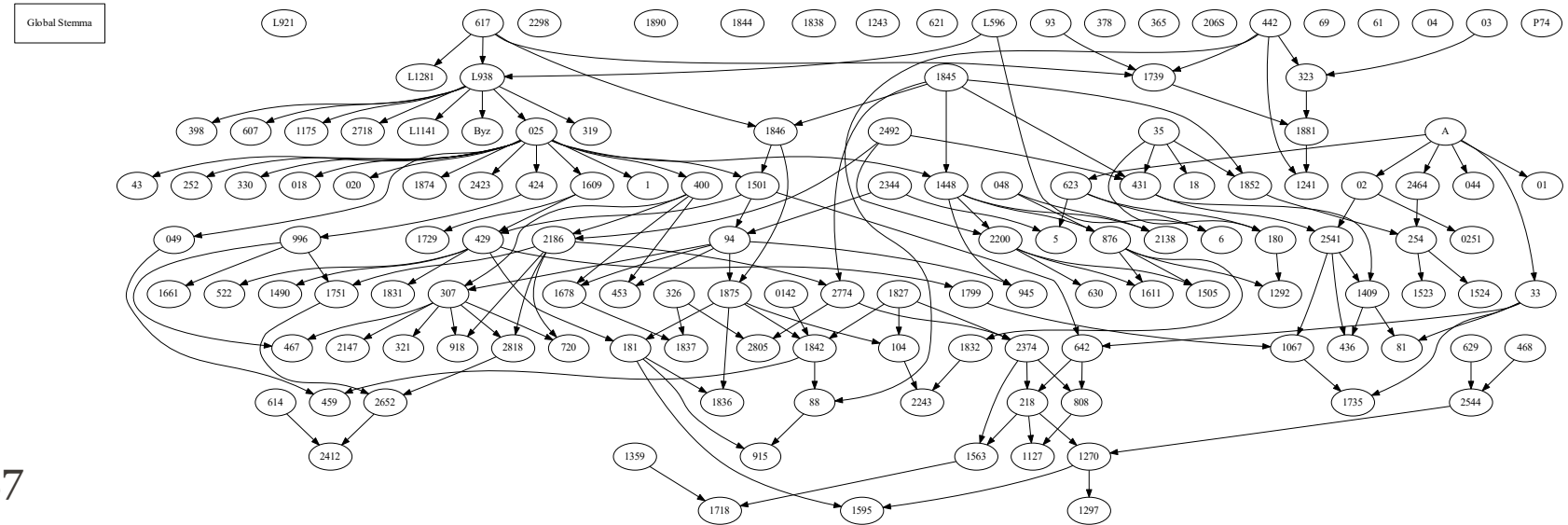


ACU INSTITUTE FOR RELIGION & CRITICAL INQUIRY

The Global Stemma

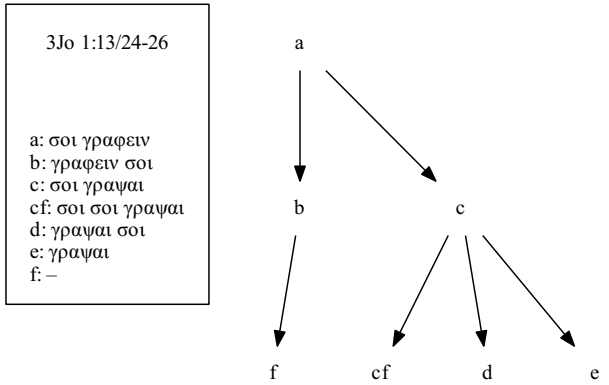
- But *every reading in every local stemma* except the initial one must be explained by another reading
- Otherwise…

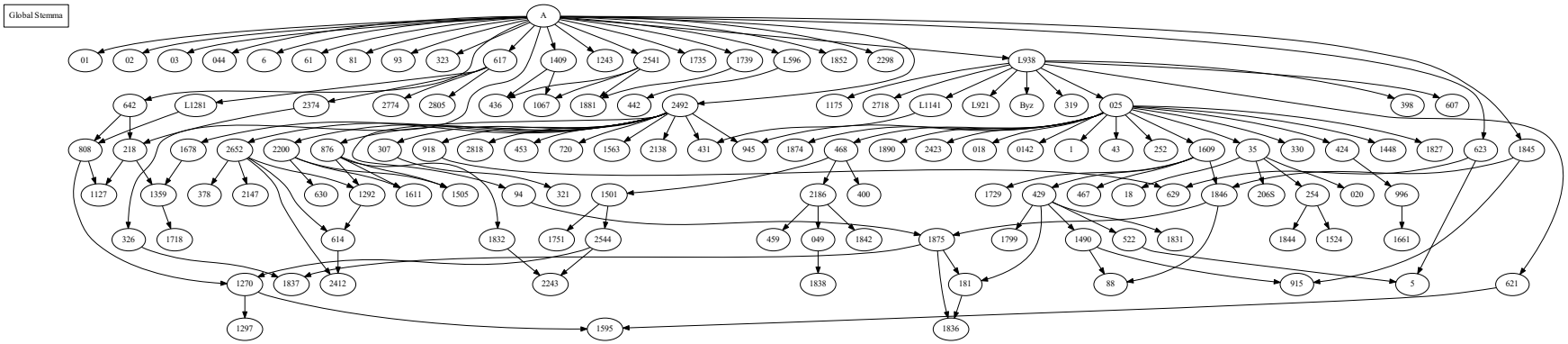




The Global Stemma

- If we “complete” every local stemma (and ignore or manually account for super fragmentary witnesses) …





The Global Stemma

- How is this different than a textual flow diagram?
  - A witness can have more than one ancestor
  - All readings in a witness must be explained by readings in its ancestor(s)
  - More computationally intensive, so takes a bit longer to produce

Field trip

Criticisms and Idiosyncrasies

- Biggest idiosyncrasy: *no reconstruction of hypothetical ancestors* (because contamination is assumed to make this impossible)
  - (Personal opinion: this assumption is made for practical rather than theoretical reasons)
  - Texts of extant witnesses = bad representatives of ancestors of other extant texts
  - CBGM may see "contamination" where there's just a gap in the textual tradition
  - Enough of the tradition is lost to make this a problem
- Can the global stemma be understood as a history of the text?

Criticisms and Idiosyncrasies

- Recommended reading:
  - Dirk Jongkind, "On the Nature and Limitations of the Coherence Based Genealogical Method" (paper presented at the Annual Meeting of the Society of Biblical Literature, San Diego, CA, 22 November 2014)
  - The special feature articles in *TC* 20 (2015)
  - Peter Gurry, "The Harklean Syriac and the Development of the Byzantine Text: A Historical Test for the Coherence-Based Genealogical Method (CBGM)," *NovT* 60.2 (2018): 358–75
  - Stephen C. Carlson, "A Bias at the Heart of the Coherence-Based Genealogical Method (CBGM)," *JBL* 139.2 (2020): 319–40 (but see Mink's response)