Learning the CBGM by Design

Greek Paul Project Webinar
28 April 2022

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

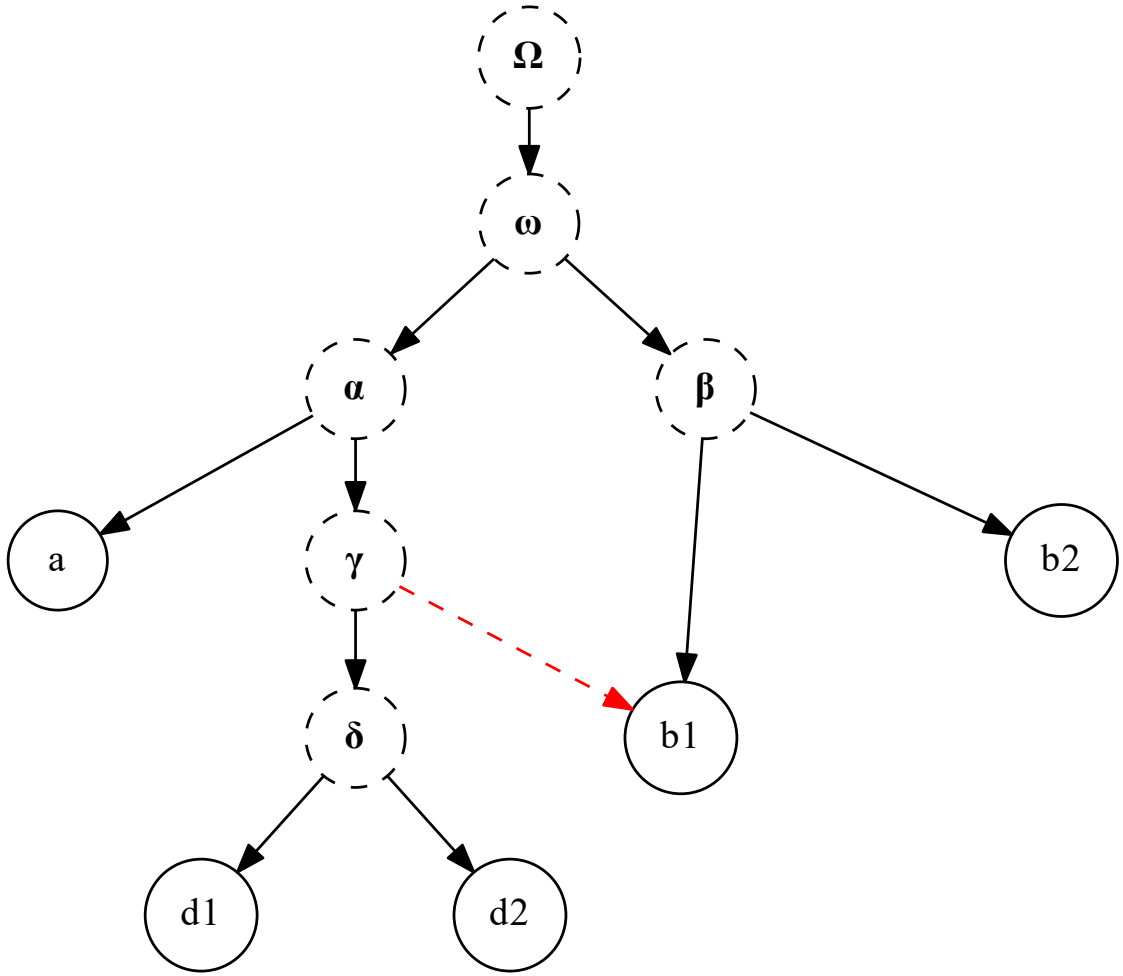
ACU INSTITUTE FOR RELIGION & CRITICAL INQUIRY

ACU INSTITUTE FOR RELIGION & CRITICAL INQUIRY

- 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

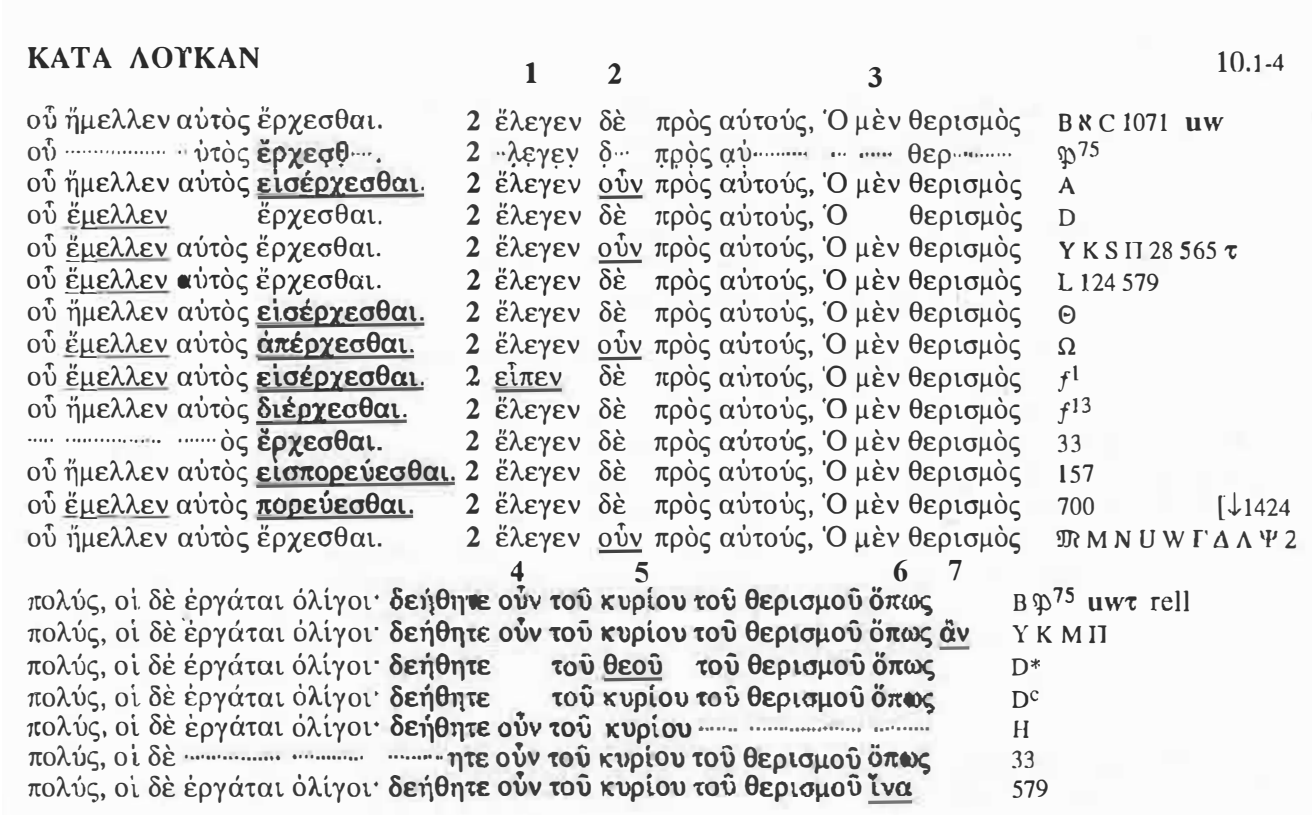
- Intended to solve *contamination*, or mixture across branches of the textual tradition



- Methodological assumptions:
  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.

ACU INSTITUTE FOR RELIGION & CRITICAL INQUIRY

- To compare manuscripts' texts, we must first 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 ἔρχεσθ···.	2	·λεγεν δ···	πρὸς αὐ····· ··· ····· θερ·····	𝔓75
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

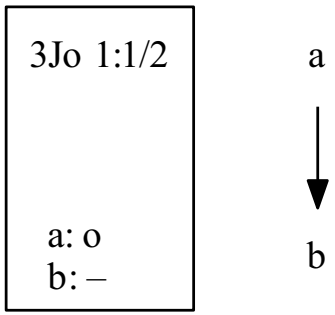
	4	5	6	7	
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε οὖν	τοῦ κυρίου	τοῦ θερισμοῦ	ὅπως	B 𝔓75 **uwτ** rell
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε οὖν	τοῦ κυρίου	τοῦ θερισμοῦ	ὅπως ἂν	Y K M Π
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε	τοῦ θεοῦ	τοῦ θερισμοῦ	ὅπως	D*
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε	τοῦ κυρίου	τοῦ θερισμοῦ	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

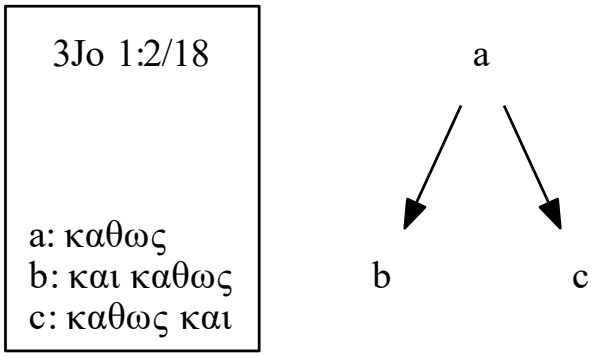
(Source: Swanson, *New Testament Greek Manuscripts, Luke*, 183)

- Variation units serve as our points of comparison between any two texts in the CBGM
- Think of them as the columns of a table and the witnesses as rows

	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

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

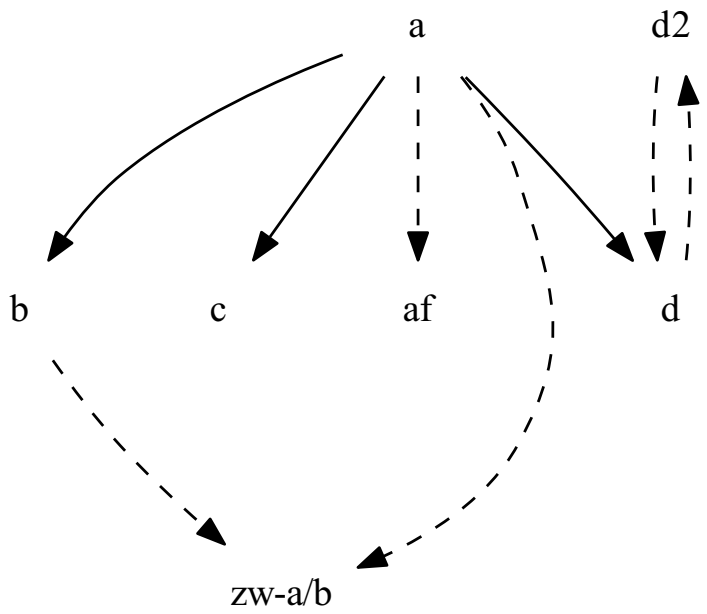




- Some are more complicated
  - *defective* readings (e.g., obvious misspellings)
  - *orthographic* readings (e.g., regional differences)
  - *split* attestations of the same reading (coincidental agreement)
  - *ambiguous* readings
- Some of these may be collapsed with other substantive readings

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



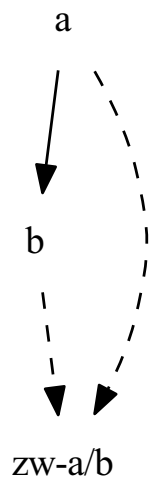
- For the CBGM's purposes, a *witness* is a sequence of readings
- Typically, the *text* of a known manuscript, minus the material baggage (date, provenance, etc.)
  - "How texts relate" ≠ "How manuscripts relate"

	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

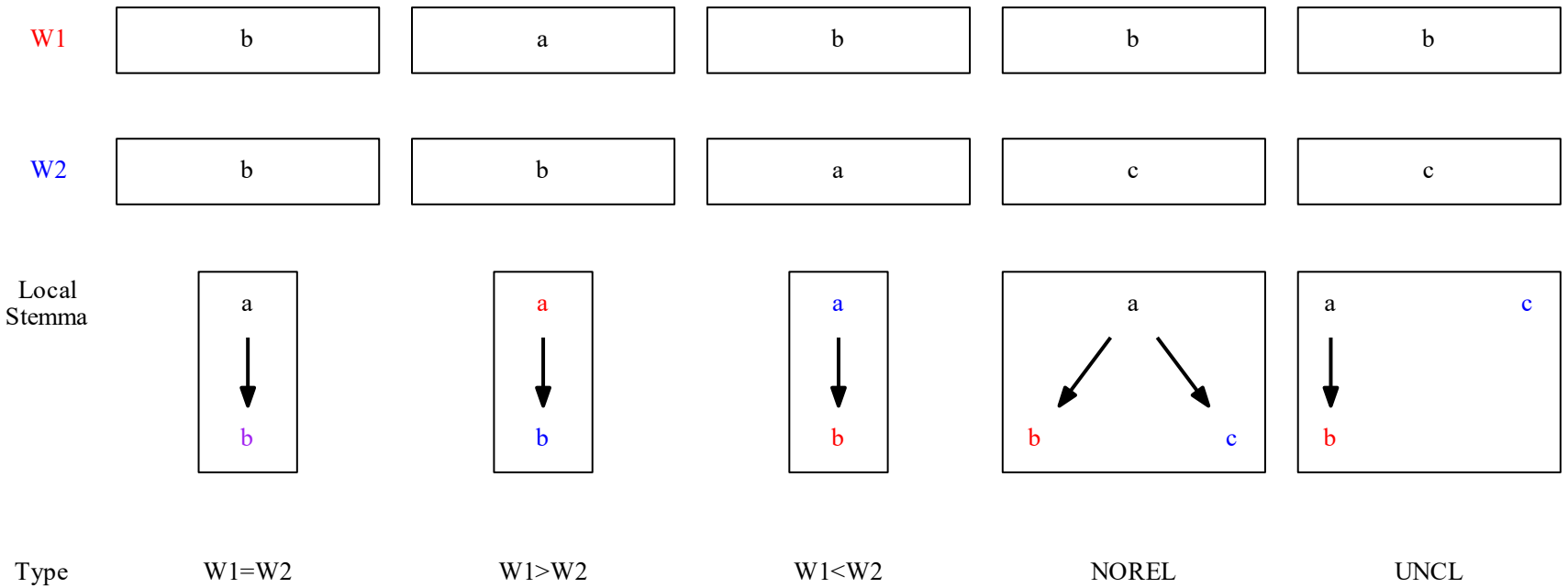
ACU INSTITUTE FOR RELIGION & CRITICAL INQUIRY

- Versions and fathers can also be treated as witnesses
- But back-translation may be ambiguous, and patristic citations may be "lacunose"

3Jo 1:1/2		
a:	o	... CosmIn. PsOec. S:H
b:	–	467. 2243. 2828
zw-a/b:		L:VT. K:SB. S:Ph

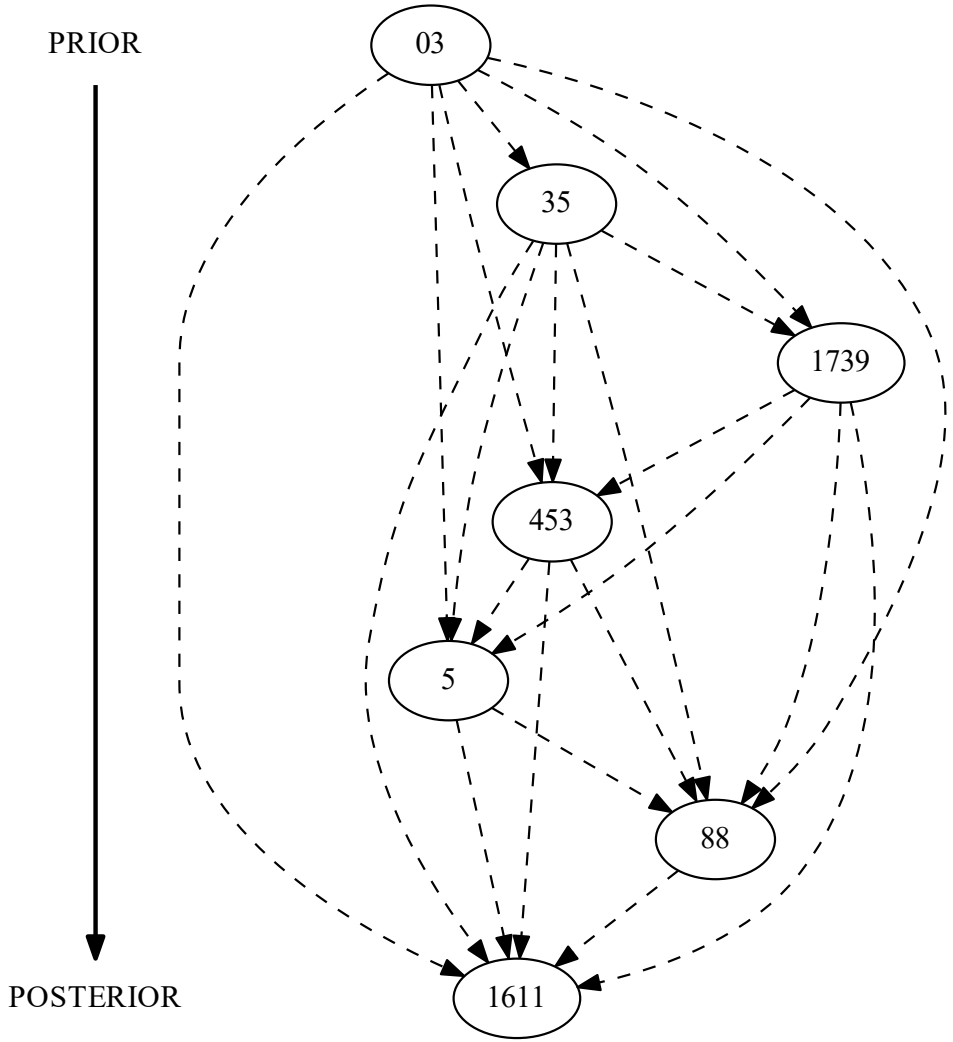


- The relationship of two witnesses is the overall pattern *of the relationships of their readings* at all variation units where both are extant



- The first three are the most important

- Potential ancestor = “more prior than posterior readings”

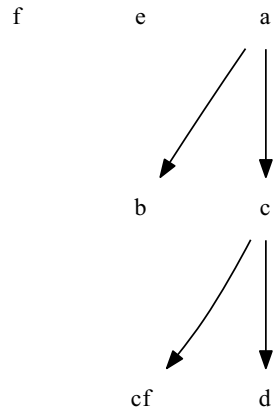


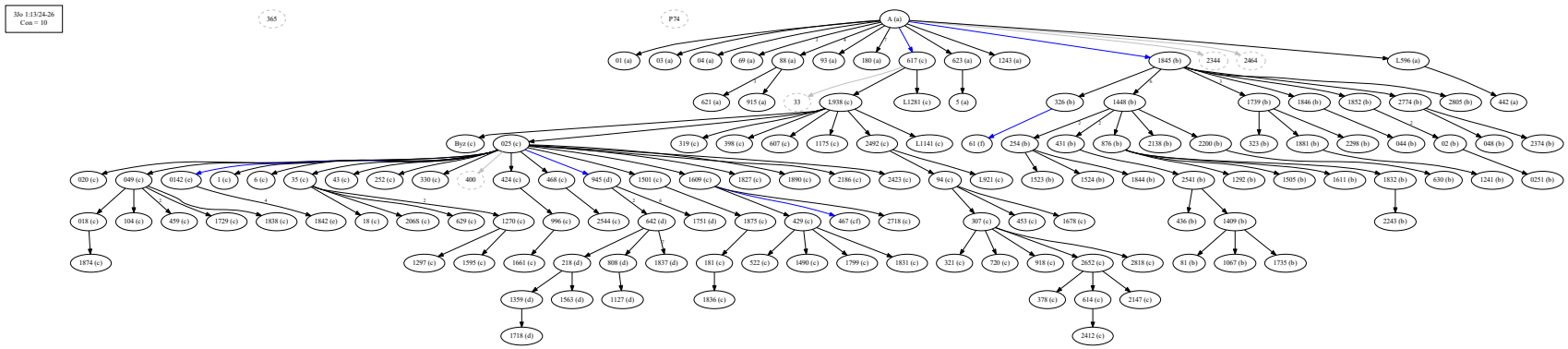
ACU INSTITUTE FOR RELIGION & CRITICAL INQUIRY

- *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

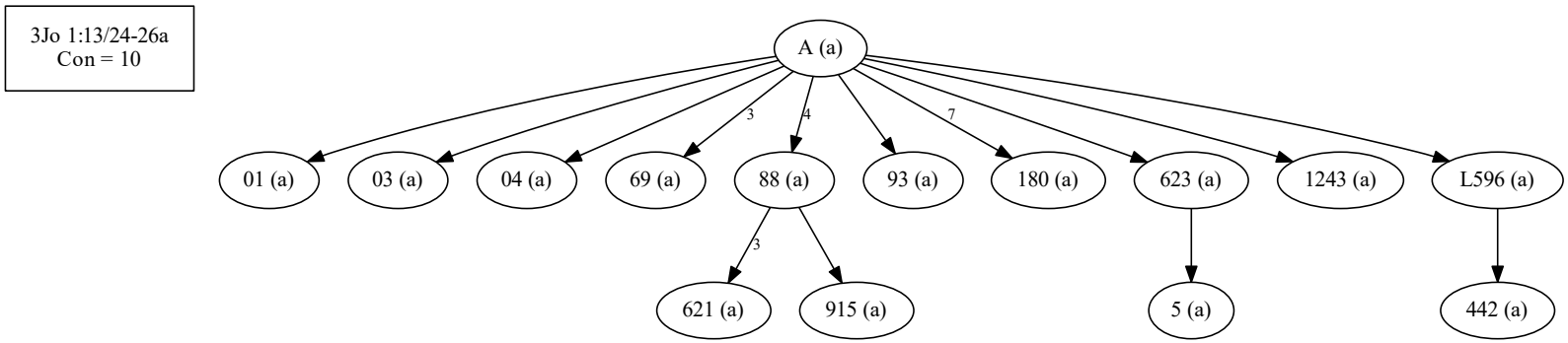
a: σοι γραφειν
b: γραφειν σοι
c: σοι γραψαι
cf: σοι σοι γραψαι
d: γραψαι σοι
e: γραψαι
f: –





- 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)

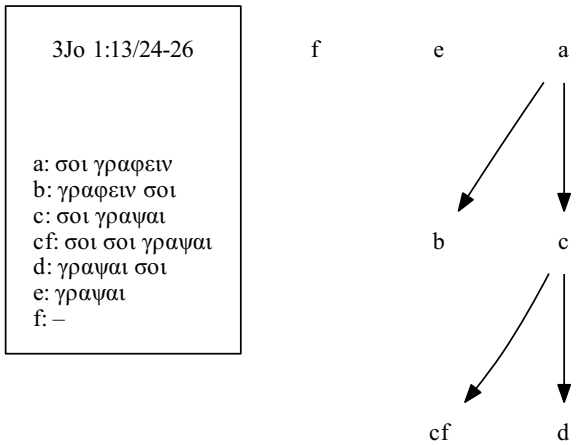
- Often, we just want to know the textual flow for witnesses with a specific reading

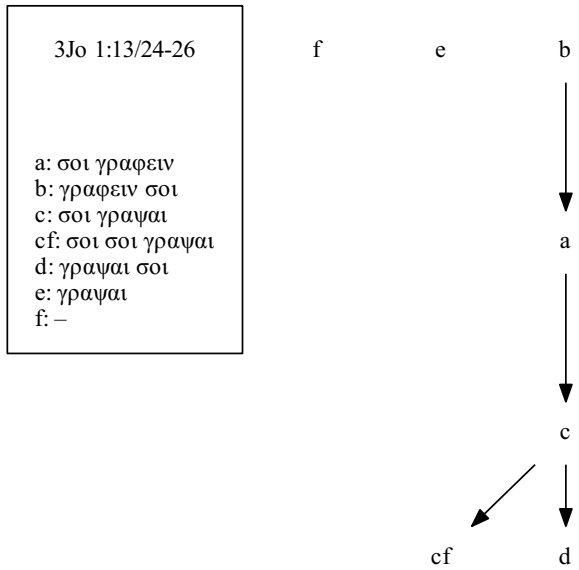


- (Numbers on edges represent the rank of the closest potential ancestor with the same reading, if it's not 1)

ACU INSTITUTE FOR RELIGION & CRITICAL INQUIRY

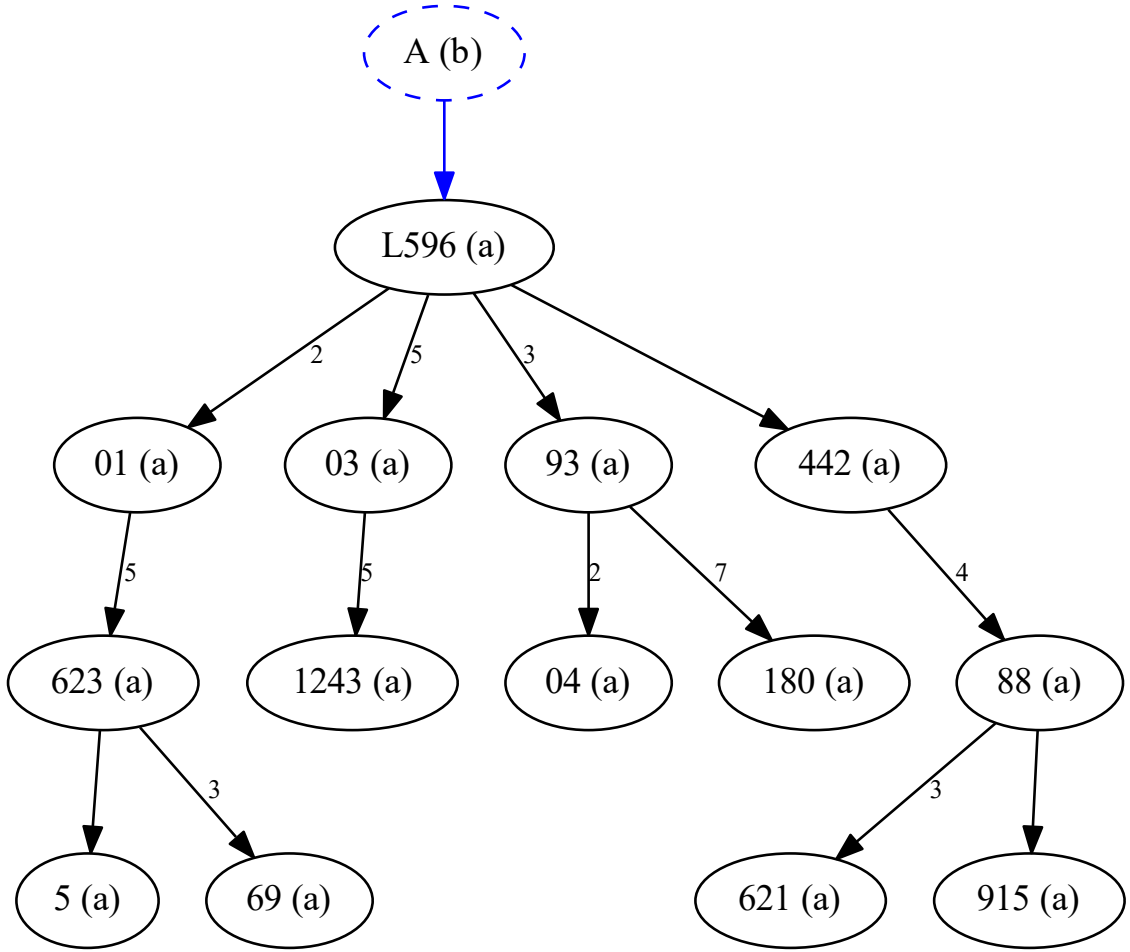
- We can use it to evaluate alternate hypotheses about the initial text (A)





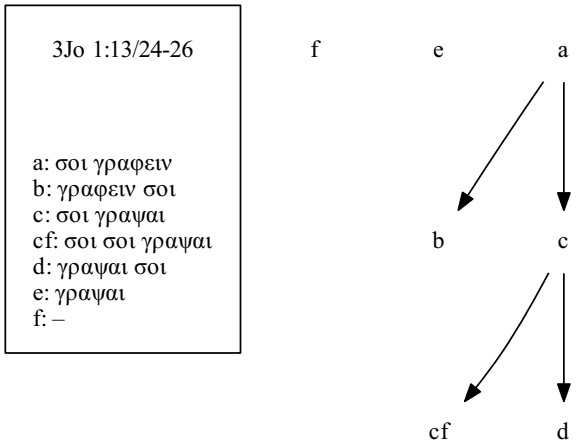
Textual Flow for a Variant Reading

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

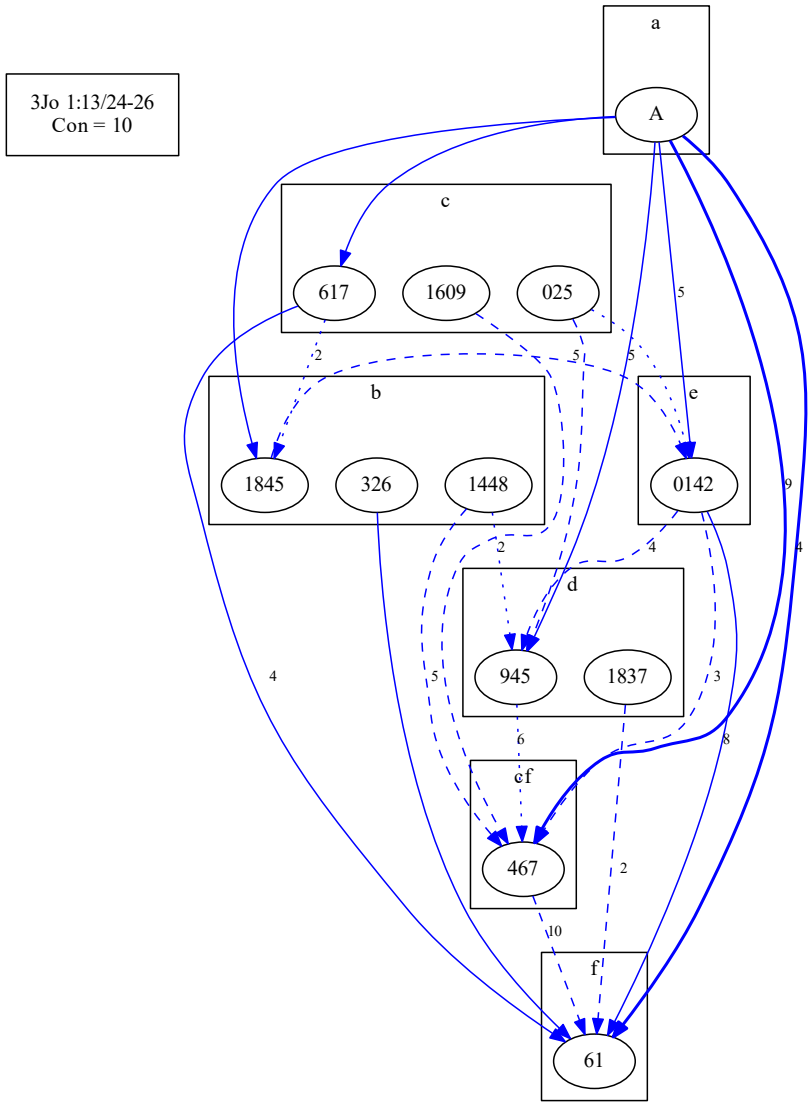


ACU INSTITUTE FOR RELIGION & CRITICAL INQUIRY

- 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*)

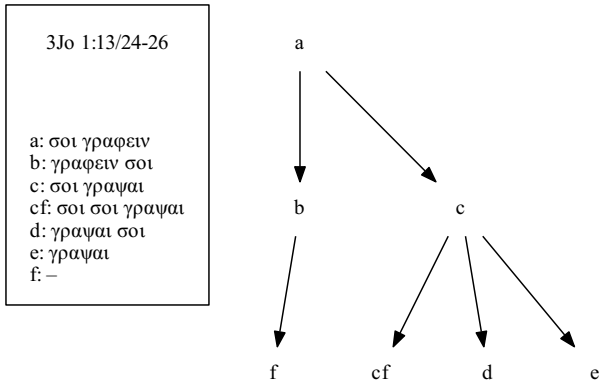


Textual Flow for a Variant Reading

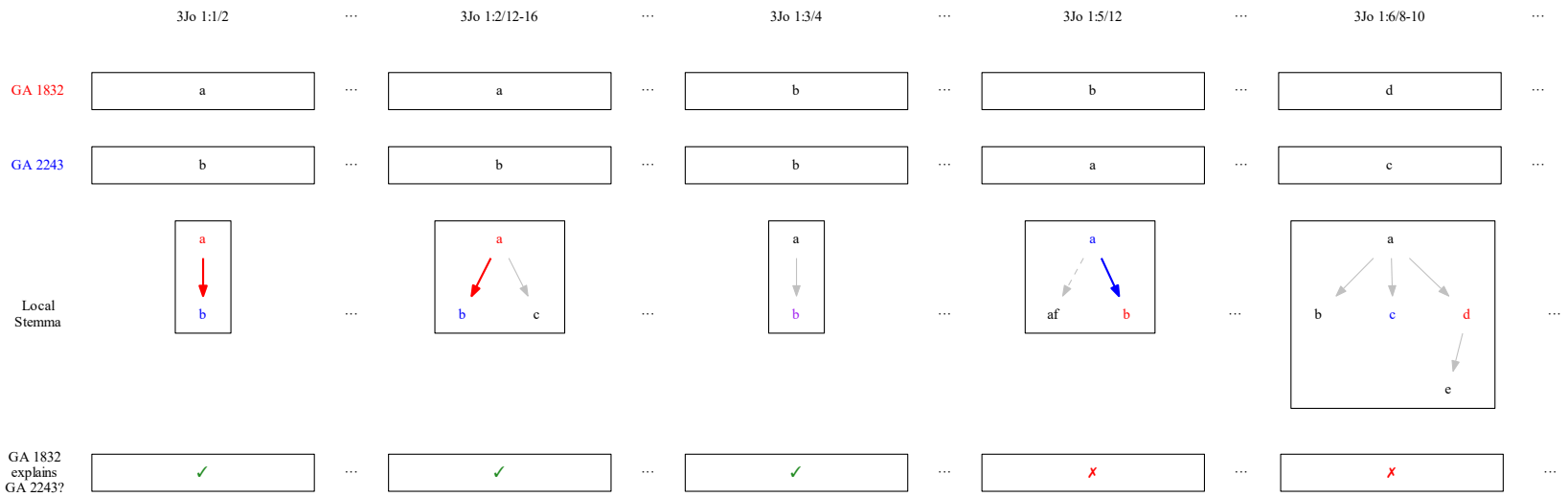


Textual Flow for a Variant Reading

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

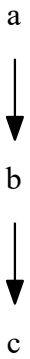


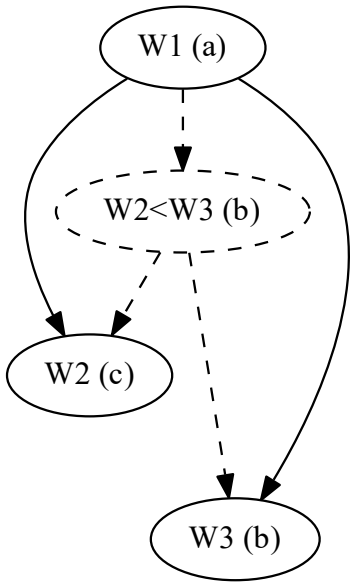
- We say that one reading *explains* another if
  - it is the same reading (explanation by agreement), or
  - there is an edge in the local stemma from it to the other reading



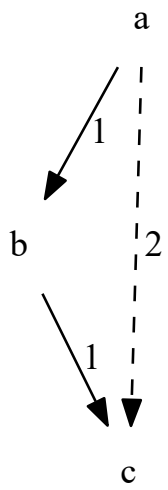
- Lacunae do not have to be explained, and they cannot explain readings

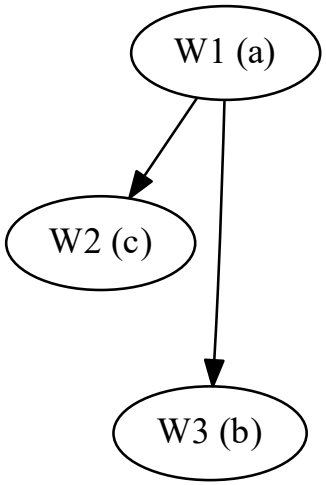
``Explained'' Readings





- 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



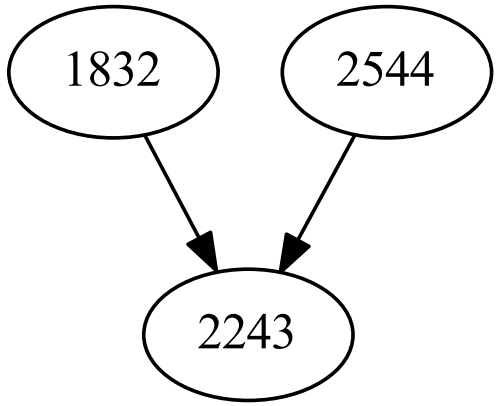


- If we instead allow *a* to explain *c*, but at a higher cost (more on this in the substemma slides), then we remove the need for intermediary nodes (although multiple changes in the same variation unit may be implied along an edge in the global stemma)

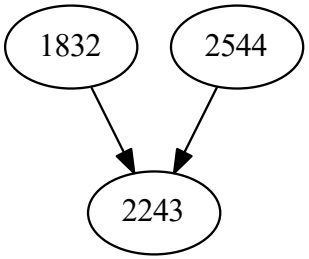
ACU INSTITUTE FOR RELIGION & CRITICAL INQUIRY

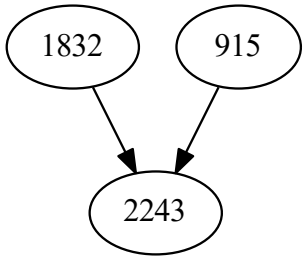
- The *substemma* of a witness is the portion of the global stemma consisting of the witness and its ancestors in the stemma
- Requirement: *every* extant reading in the witness must be explained by a reading in at least one of its ancestors

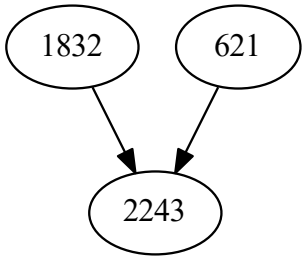
Explained by GA 1832	⋯	✗	✓	✓	✓	⋯
Explained by GA 2544	⋯	✓	✗	✗	✓	⋯
Explained by Either	⋯	✓	✓	✓	✓	⋯

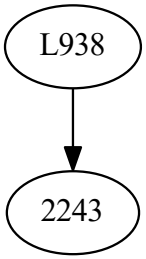


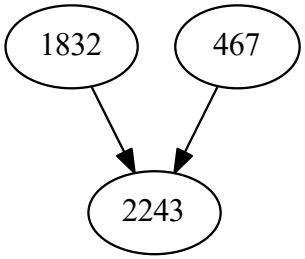
- A witness may have multiple valid substemma (i.e., ones that explain all of its readings), but some are better than others
- Two of the CBGM's methodological assumptions are important here:
  3. Scribes typically used fewer sources rather than many.
  4. Scribes typically used closely related sources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



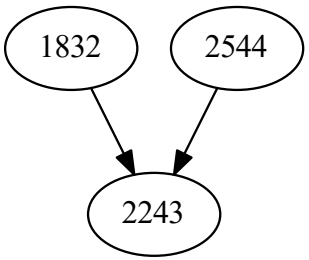


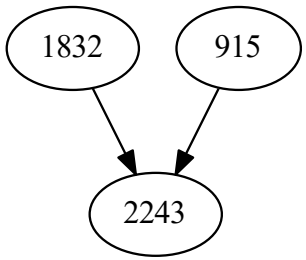


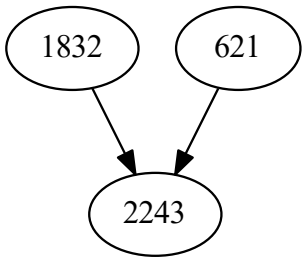


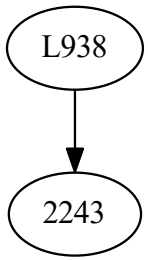


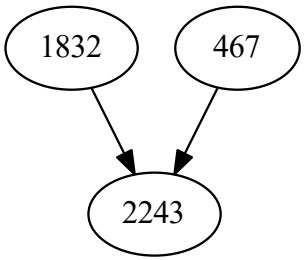
- Based on assumption 3, we should prefer substemmata with fewer ancestors (“parsimony”)
- Based on assumption 4, we should prefer substemmata with ancestors that agree as often as possible with the witness
- A balancing act: the substemma {L938} is more parsimonious, but may not explain as many readings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



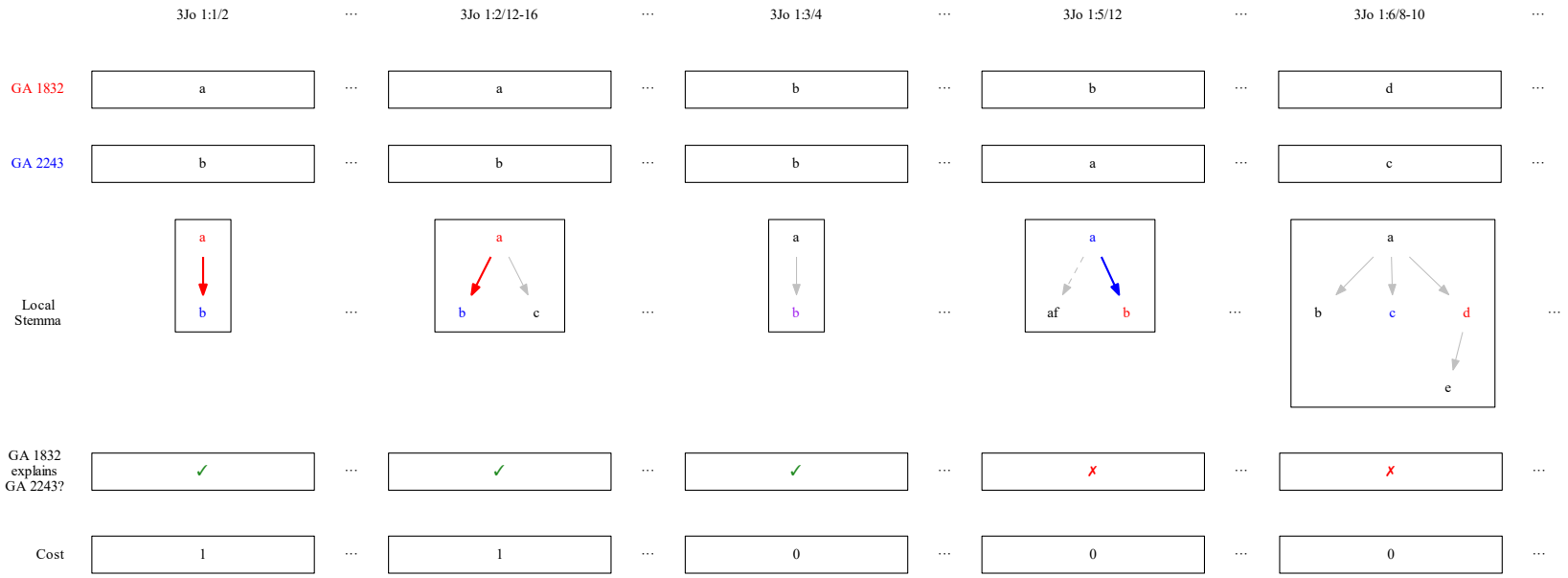




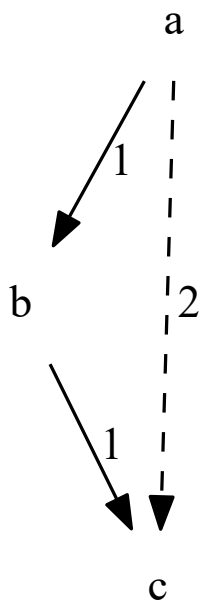




- A simple cost function for each ancestor is “the number of variation units where the ancestor explains the witness by descent and not agreement”



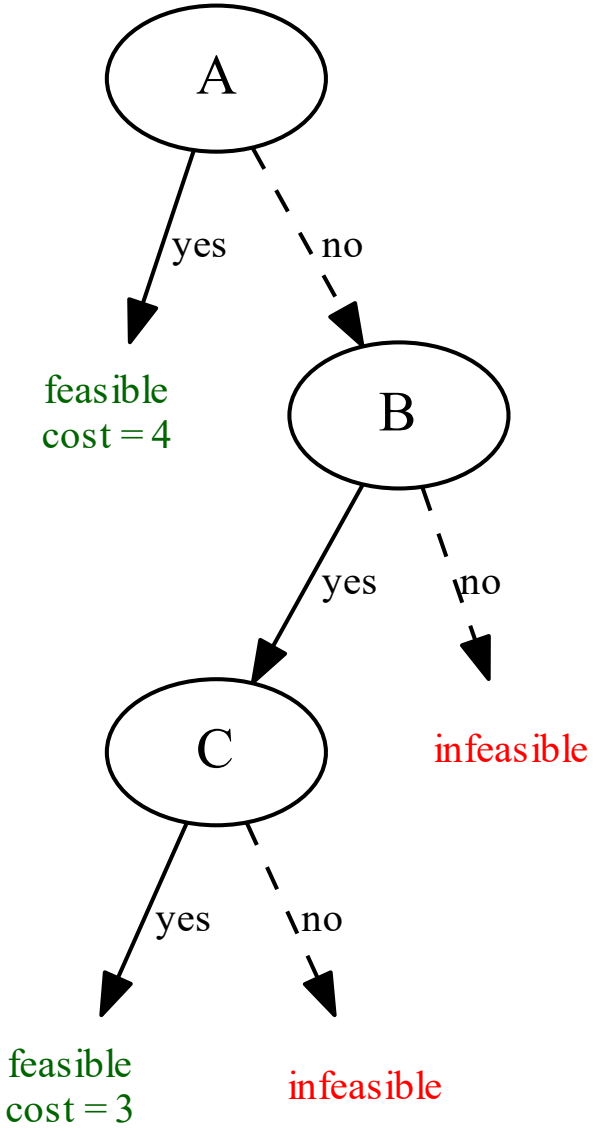
- If we allow a reading to explain any reading posterior to it, then a better cost per variation unit is the length of the path from the prior reading to the posterior one.



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

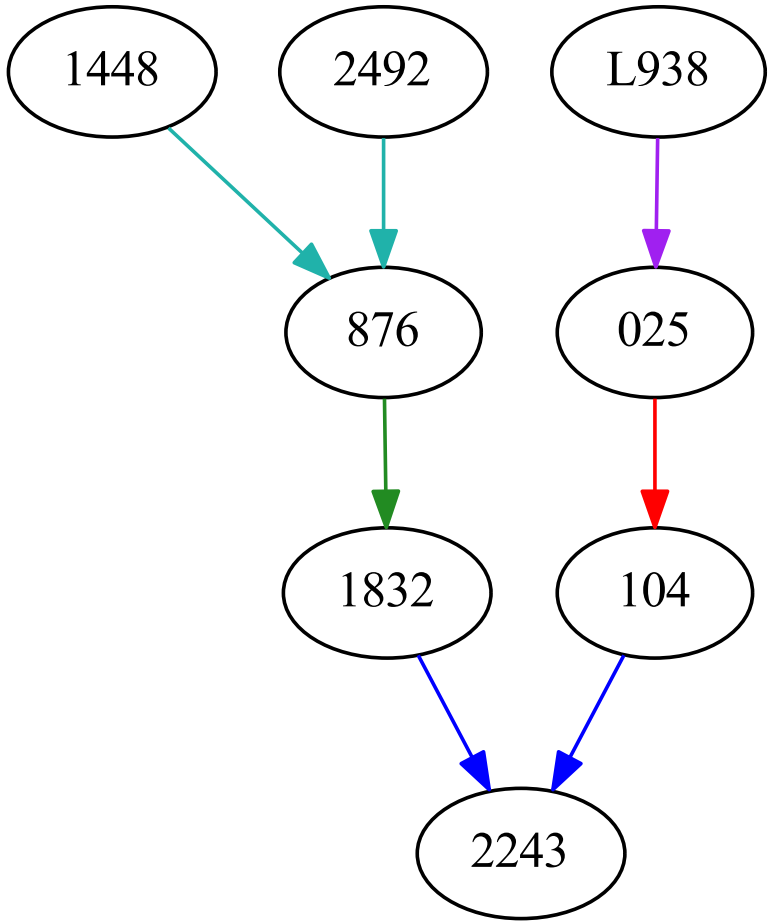
Substemma	Variation Units Explained				Cost
{A}	✓	✓	✓	✓	4
{B}	✓	✓	✗	✗	1
{C}	✗	✓	✓	✓	2
{A, B}	✓	✓	✓	✓	4+1=5
{A, C}	✓	✓	✓	✓	4+2=6
{B, C}	✓	✓	✓	✓	1+2=3
{A, B, C}	✓	✓	✓	✓	1+2+4=7

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

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

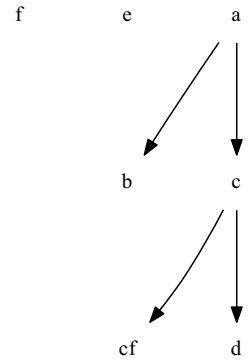


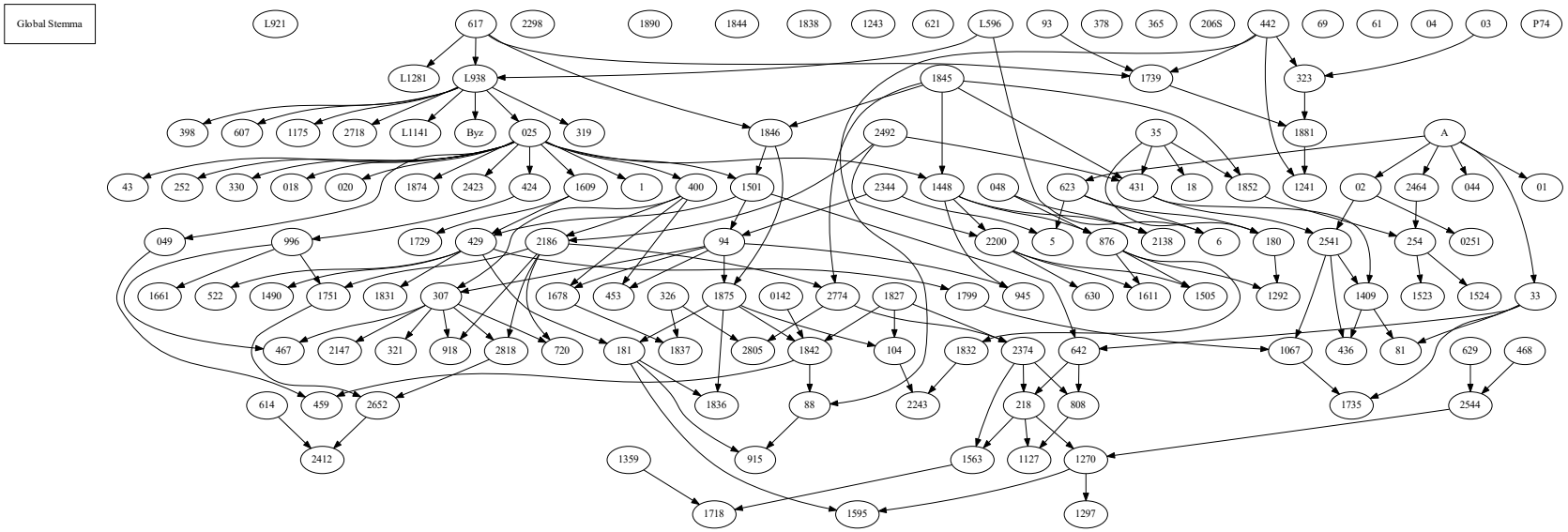
ACU INSTITUTE FOR RELIGION & CRITICAL INQUIRY

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

3Jo 1:13/24-26

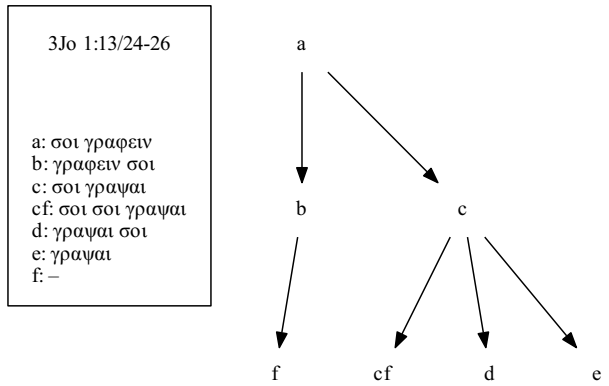
a: σοι γραφειν
b: γραφειν σοι
c: σοι γραψαι
cf: σοι σοι γραψαι
d: γραψαι σοι
e: γραψαι
f: –

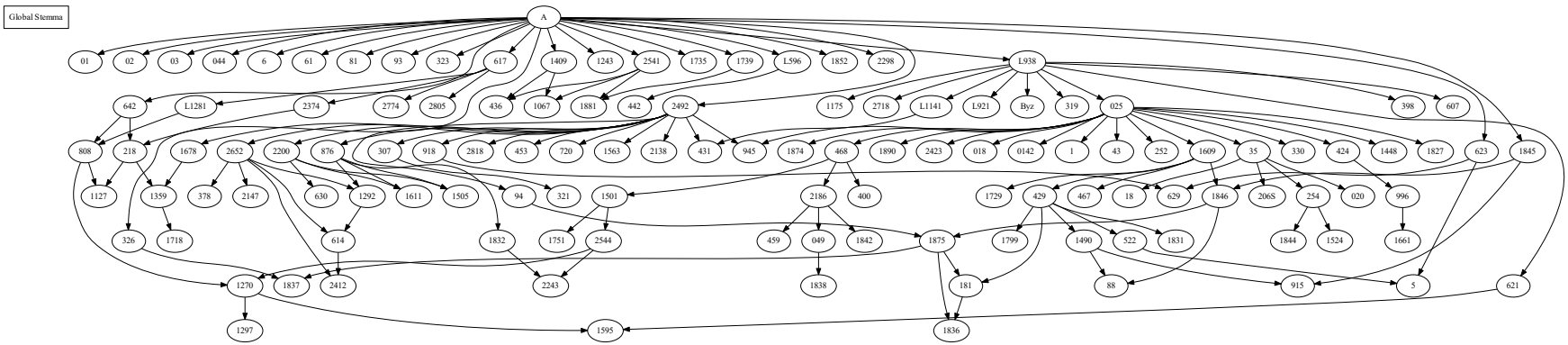




ACU INSTITUTE FOR RELIGION & CRITICAL INQUIRY

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





- 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

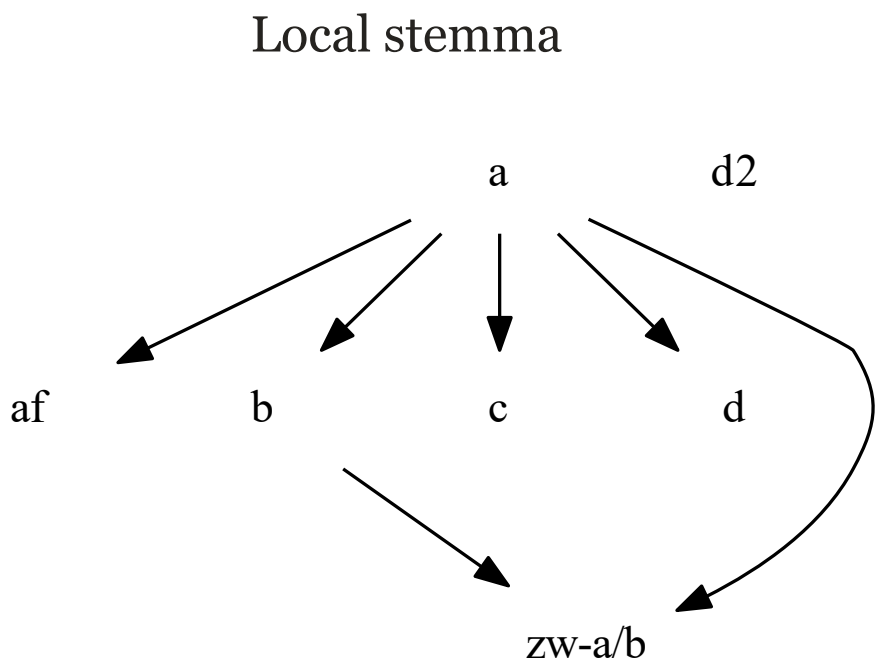
- Biggest idiosyncrasy: *no reconstruction of hypothetical ancestors* (because contamination is assumed to make this impossible)
  - (Personal opinion: this assumption is made for practical rather than theoretical reasons)
  - Texts of extant witnesses = bad representatives of ancestors of other extant 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?

- 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)

ACU INSTITUTE FOR RELIGION & CRITICAL INQUIRY

Collation

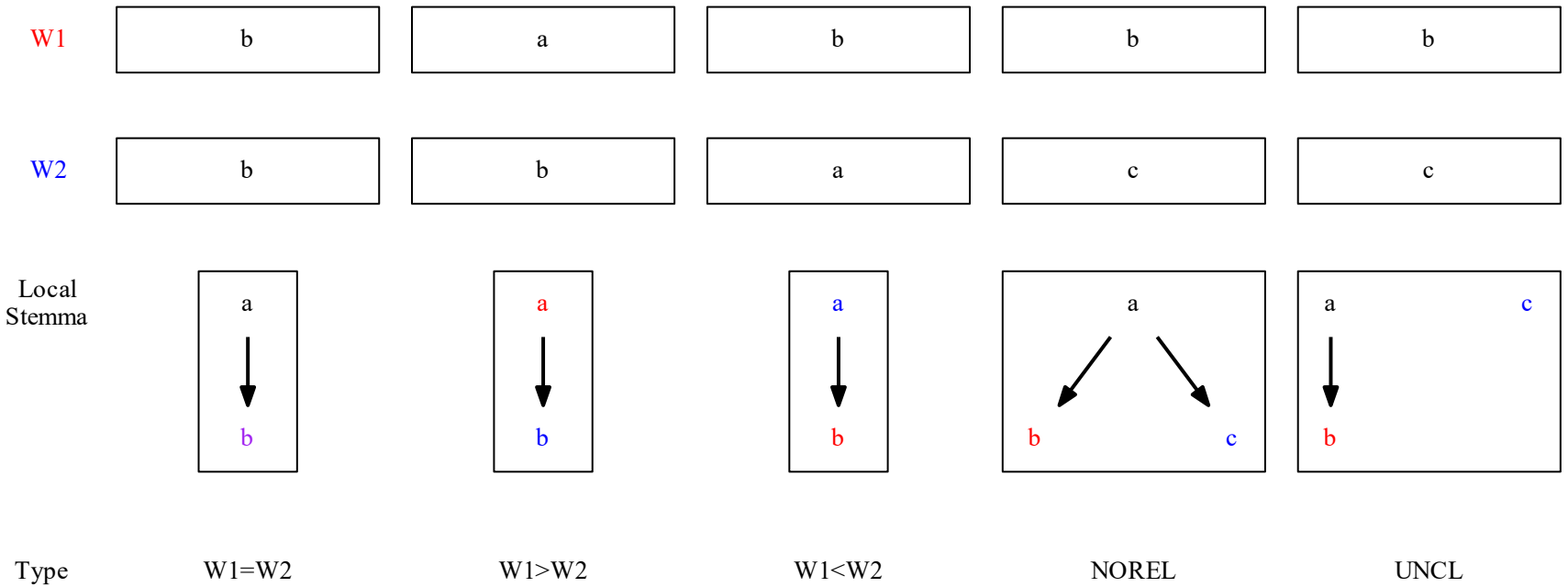
```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title>A minimal TEI XML collation file</title>
            </titleStmt>
            <publicationStmt>
                <p>Temporary publicationStmt for validation</p>
            </publicationStmt>
            <sourceDesc>
                <listWit>
                    <witness xml:id="A" />
                    <witness xml:id="B" />
                </listWit>
            </sourceDesc>
        </fileDesc>
    </teiHeader>
    <text xml:lang="GRC">
        <body>
            <div type="book" n="B00">
                <app n="B00K0V0U0">
                    <rdg n="a" wit="#A">ειπεν</rdg>
                    <rdg n="b" wit="#B">ελεγεν</rdg>
                    <note>
                        <label>Book Chapter:Verse/Unit</label>
                        <fs>
                            <f name="connectivity">
                                <numeric value="5" />
                            </f>
                        </fs>
                        <graph type="directed">
                            <node n="a" />
                            <node n="b" />
                            <arc from="a" to="b" />
                        </graph>
                    </note>
                </app>
            </div>
        </body>
    </text>
</TEI>
```




```
<graph type="directed">
    <node n="a" />
    <node n="af" />
    <node n="b" />
    <node n="c" />
    <node n="d" />
    <node n="d2" />
    <node n="zw-a/b" />
    <arc from="a" to="af" />
    <arc from="a" to="b" />
    <arc from="a" to="c" />
    <arc from="a" to="d" />
    <arc from="a" to="zw-a/b" />
    <arc from="b" to="zw-a/b" />
</graph>
```

Genealogical relationships

$$\texttt{agree} = [1, 0, 0, 0, 0] \qquad \texttt{norel} = [0, 0, 0, 1, 0]$$
$$\texttt{prior} = [0, 1, 0, 0, 0] \qquad \texttt{uncl} = [0, 0, 0, 0, 1]$$
$$\texttt{posterior} = [0, 0, 1, 0, 0] \qquad \texttt{expl} = [1, 0, 1, 0, 0]$$
$$\texttt{cost} = 1$$



- The `open-cbgm` library (my implementation of the CBGM, based on these principles) is freely available at https://github.com/jjmccollum/open-cbgm, and the standalone command-line utility is available at https://github.com/jjmccollum/open-cbgm-standalone
  - Supported on Windows, Mac, and Linux
- The INTF's official implementation (using a Docker container) is now also available (download and instructions at http://ntvmr.uni-muenster.de/intfblog/-/blogs/download-the-cbgm-docker-container)

Carlson, Stephen C. "A Bias at the Heart of the Coherence-Based Genealogical Method (CBGM)." *JBL* 139.2 (2020): 319–40.

Edmondson, Andrew Charles. "An Analysis of the Coherence-Based Genealogical Method Using Phylogenetics." PhD diss., University of Birmingham, 2019. https://etheses.bham.ac.uk/id/eprint/9150/.

Gurry, Peter. "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.

Gurry, Peter J. *A Critical Examination of the Coherence-Based Genealogical Method in New Testament Textual Criticism*. NTTSD 55. Leiden: Brill, 2017.

Jongkind, Dirk. "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.

McCollum, Joey. "Biclustering Readings and Manuscripts via Non-negative Matrix Factorization, with Application to the Text of Jude." *AUSS* 57.1 (2019): 61–89. https://digitalcommons.andrews.edu/auss/vol57/iss1/6/.

McCollum, Joey. “Luke 9,35 and the Power of Polygenesis.” *FilNeot* 33.53 (2020): 51–94.

———.“The open-cbgm Library: Design and Demonstration.” Paper presented at the Annual Meeting of the Society of Biblical Literature. Boston, MA, 9 December 2020.

———.“The Text and Margin of Gregory-Aland 274.” *TC* 26 (2021): 47–131.

Mink, Gerd. “Problems of a Highly Contaminated Tradition: The New Testament. Stemmata of Variants as a Source of a Genealogy for Witnesses.” Pages 13–85 in *Studies in Stemmatology II*. Edited by Pieter van Reenen, August den Hollander, and Margot van Mulken. Amsterdam: John Benjamins Publishing, 2004.

Swanson, Reuben J., ed. *New Testament Greek Manuscripts: Variant Readings Arranged in Horizontal Lines against Codex Vaticanus. Luke*. Sheffield: Sheffield Academic Press, 1995.

Wasserman, Tommy, 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.