

# CBGM Q&A @ CSNTM

---

Joey McCollum\*

31 January 2022

\*Virginia Polytechnic Institute and State University

✉ [jjmccollum@vt.edu](mailto:jjmccollum@vt.edu)

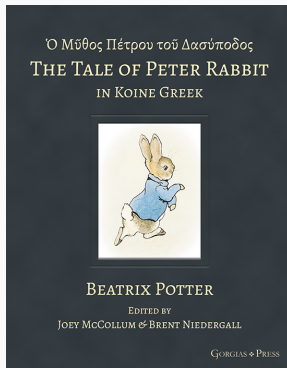
🐦 [@jamesjmccollum](https://twitter.com/jamesjmccollum)

🔗 [jjmccollum](https://github.com/jjmccollum)

Note: A more formal introduction to the CBGM (also written by me) is available for free at <https://vt.academia.edu/JoeyMcCollum>.

# About Me

- Research Associate at VT (for a few more days)
- Soon: PhD candidate at ACU (supervisor: Stephen Carlson)
- Publications/talks in *AUSS*, *FilNeot*, *TC*, *SBL* (see bibliography)
- PS guest talk: <https://www.youtube.com/watch?v=vbUnZZxww-o>



# Choose Your Own Adventure

About the CBGM

Collation

The Local Stemma

Witnesses

Genealogical Relationships between Witnesses

Potential Ancestors

Textual Flow at a Variation Unit

Textual Flow for a Variant Reading

“Explained” Readings

The Substemma(ta) of a Witness

Finding a (Good) Substemma

The Global Stemma

Criticisms and Idiosyncrasies

Making It Digital

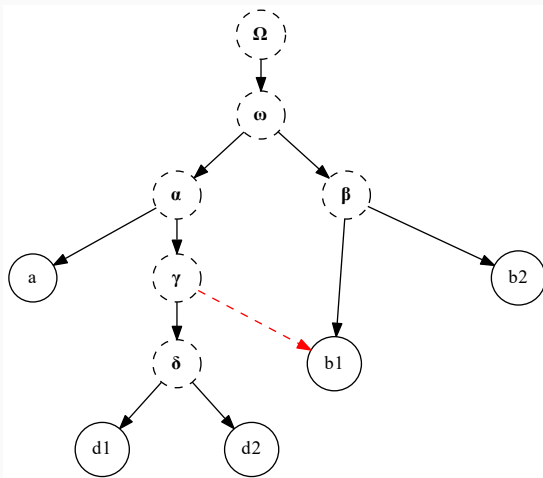
# About the CBGM

- 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

# About the CBGM

- 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.

(Return to start)

# Collation

- To compare manuscripts' texts, we must first align them at independent *variation units*
- *Variant readings* occur at variation units

ΚΑΤΑ ΛΟΥΚΑΝ	1	2	3	10.1-4
οὐ ἤμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	B K C 1071 uw
οὐ ἤμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	75
οὐ ἤμελλεν αὐτὸς <u>εἰσεργεσθαι</u> .	2	ἔλεγεν οὖν	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	A
οὐ ἔμελλεν ἔρχεσθαι.	2	ἔλεγεν δὲ	πρὸς αὐτούς, Ὁ θερισμὸς	D
οὐ ἔμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν οὖν	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	Y K S Π 28 565 τ
οὐ ἔμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	L 124 579
οὐ ἤμελλεν αὐτὸς <u>εἰσεργεσθαι</u> .	2	ἔλεγεν δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	Θ
οὐ ἔμελλεν αὐτὸς <u>ἀπεργεσθαι</u> .	2	ἔλεγεν οὖν	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	Ω
οὐ ἔμελλεν αὐτὸς <u>εἰσεργεσθαι</u> .	2	εἶπεν δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	f <sup>1</sup>
οὐ ἤμελλεν αὐτὸς <u>διεργεσθαι</u> .	2	ἔλεγεν δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	f <sup>13</sup>
οὐ ἤμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	33
οὐ ἤμελλεν αὐτὸς <u>εἰσπορεύεσθαι</u> .	2	ἔλεγεν δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	157
οὐ ἔμελλεν αὐτὸς <u>πορεύεσθαι</u> .	2	ἔλεγεν δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	700
οὐ ἤμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν οὖν	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	[↓1424 98 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 75 uw tell
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε οὖν τοῦ κυρίου τοῦ θερισμοῦ ὅπως <u>ἂν</u>				Y K M Π
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε τοῦ <u>θεοῦ</u> τοῦ θερισμοῦ ὅπως				D*
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε τοῦ κυρίου τοῦ θερισμοῦ ὅπως				D <sup>c</sup>
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε οὖν τοῦ κυρίου				H
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε οὖν τοῦ κυρίου τοῦ θερισμοῦ ὅπως				33
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε οὖν τοῦ κυρίου τοῦ θερισμοῦ <u>ἵνα</u>				579

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



# Collation

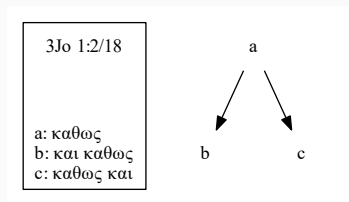
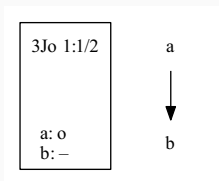
- Variation units serve as our points of comparison between any two texts in the CBGM
- Think of them as the columns of a table and the witnesses as rows

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

(Return to start)

# The Local Stemma

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

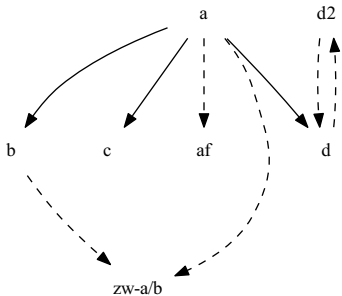


# The Local Stemma

- 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



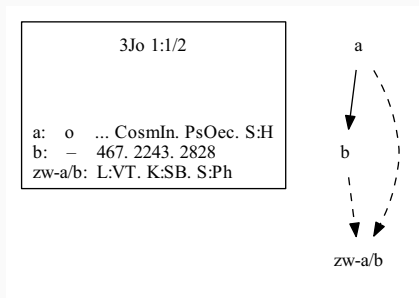
# Witnesses

- 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”  $\neq$  “How manuscripts relate”

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

# Witnesses




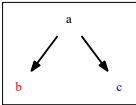
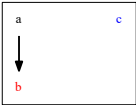
- Versions and fathers can also be treated as witnesses
- But back-translation may be ambiguous, and patristic citations may be “lacunose”



(Return to start)

# Genealogical Relationships between Witnesses

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

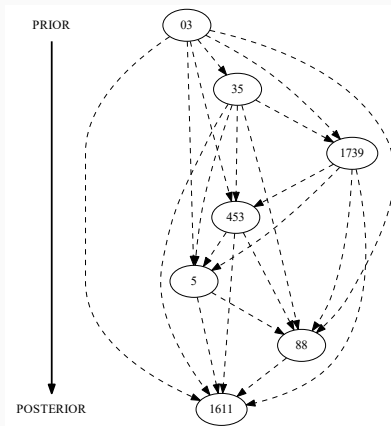
W1	b	a	b	b	b
W2	b	b	a	c	c
Local Stemma					
Type	W1=W2	W1>W2	W1<W2	NOREL	UNCL

- The first three are the most important

(Return to start)

# Potential Ancestors

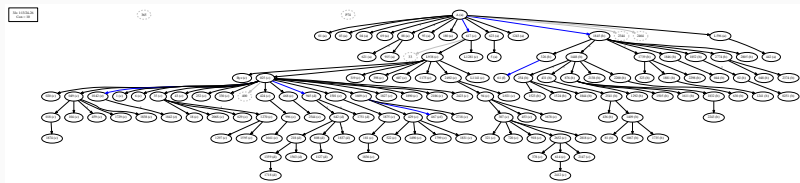
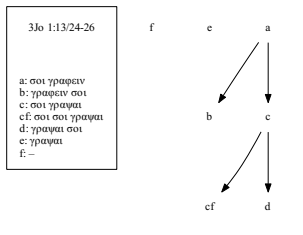
- Potential ancestor = “more prior than posterior readings”



(Return to start)

# Textual Flow at a Variation Unit

- *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





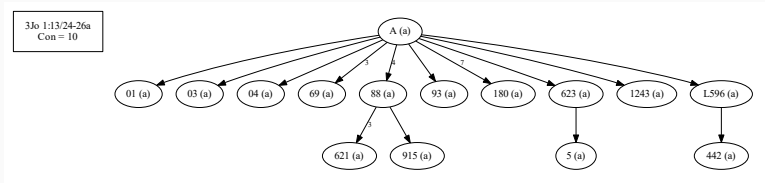
# Textual Flow at a Variation Unit

- How do we find a given witness's *textual flow ancestor*?
- We specify a *connectivity limit*  $\kappa$  (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  $\kappa$  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)

(Return to start)

# Textual Flow for a Variant Reading

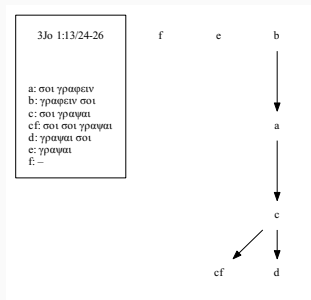
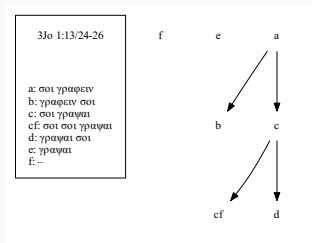
- Often, we just want to know the textual flow for witnesses with a specific reading



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

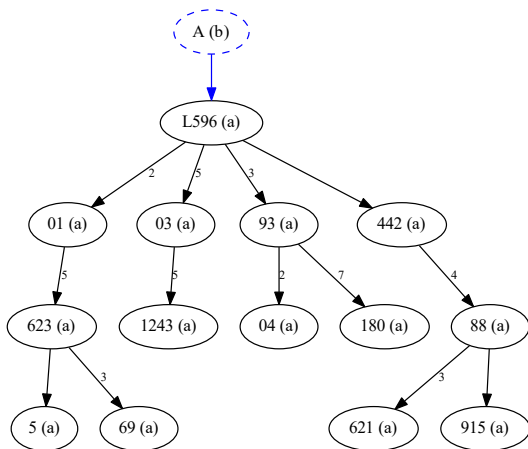
# Textual Flow for a Variant Reading

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



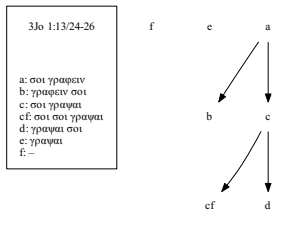
# Textual Flow for a Variant Reading

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

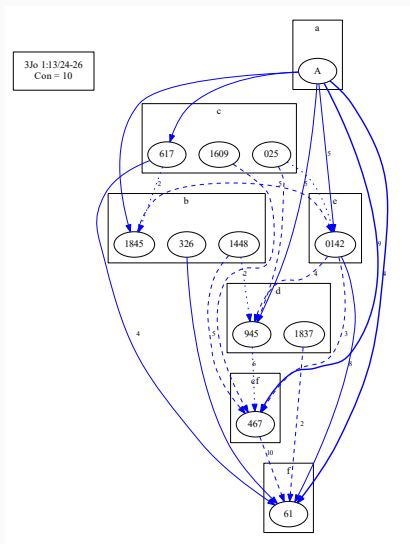


# Textual Flow for a Variant Reading

- 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*)

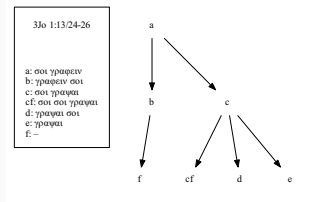


# Textual Flow for a Variant Reading



# Textual Flow for a Variant Reading

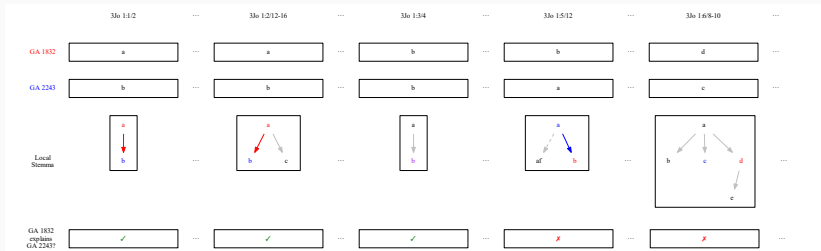
- Using this information, we can attempt to explain previous unexplained readings
- A necessary step for our ultimate goal of constructing a global stemma



(Return to start)

# “Explained” Readings

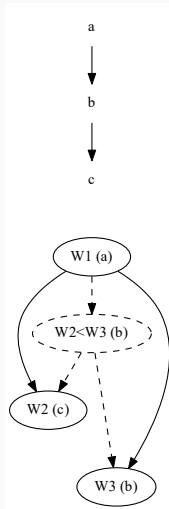
- We say that one reading *explains* another if
  - it is the same reading (explanation by agreement), or
  - there is an edge in the local stemma from it to the other reading



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

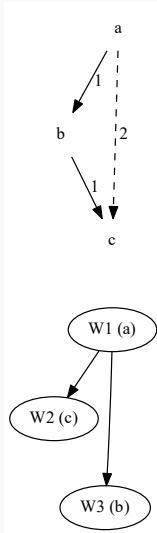


# “Explained” Readings



- 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

# “Explained” Readings



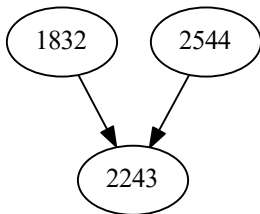
- If we instead allow *a* to explain *c*, but at a higher cost (more on this in the substemma slides), then we remove the need for intermediary nodes (although multiple changes in the same variation unit may be implied along an edge in the global stemma)

(Return to start)

## The Substemma(ta) of a Witness

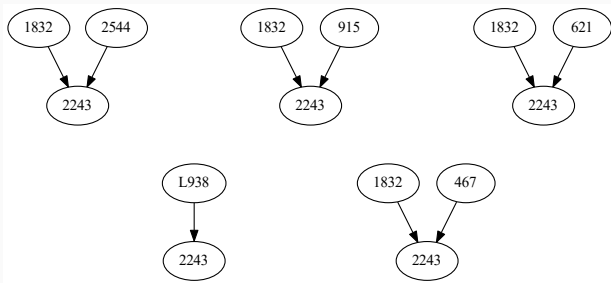
- The *substemma* of a witness is the portion of the global stemma consisting of the witness and its ancestors in the stemma
- Requirement: every extant reading in the witness must be explained by a reading in at least one of its ancestors

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



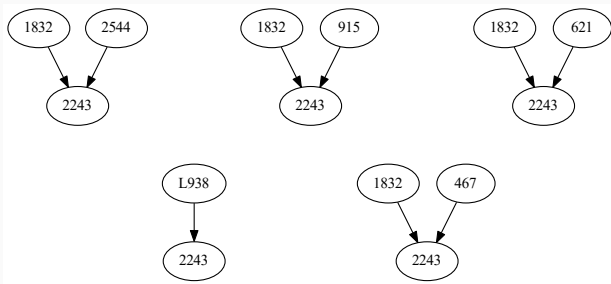
# The Substemma(ta) of a Witness

- 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.



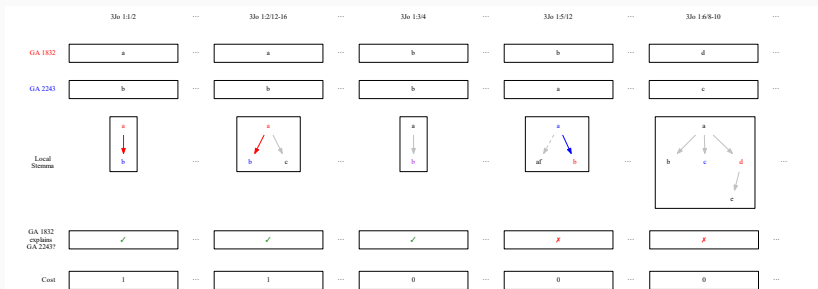
# The Substemma(ta) of a Witness

- 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



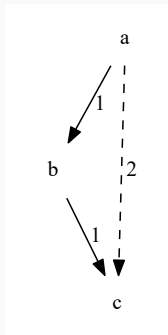
# The Substemma(ta) of a Witness

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



# 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.



(Return to start)

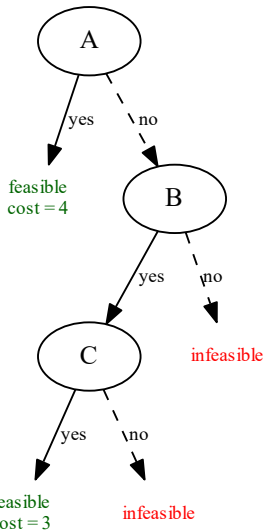
## Finding a (Good) Sublemma

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

Sublemma	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

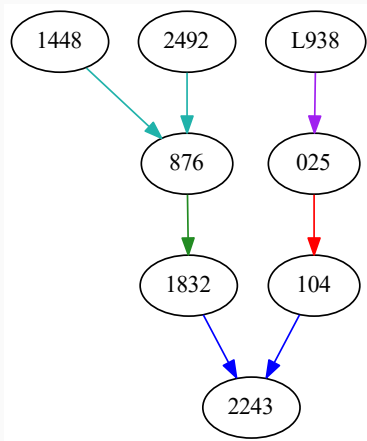


- If a witness has many potential ancestors, then checking all  $2^n - 1$  possible substemmata by brute force is prohibitive
- The *branch-and-bound* heuristic (pictured left) finds all minimum-cost substemmata quickly in practice
- Easily adapted to find all substemmata within a given cost

(Return to start)

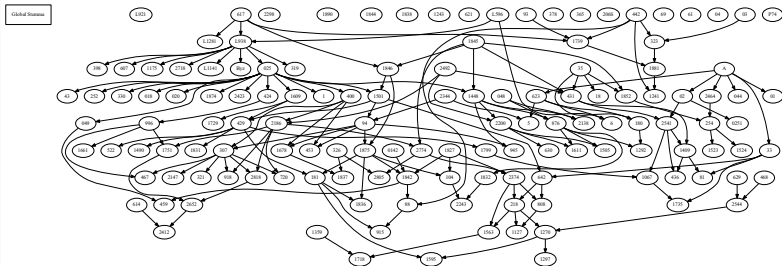
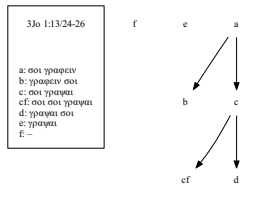
# The Global Stemma

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



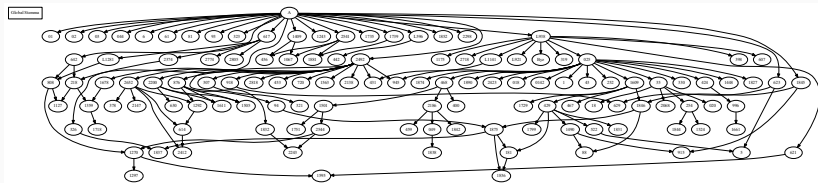
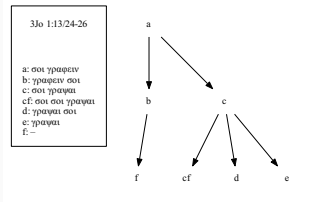
# The Global Stemma

- 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) ...



- 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\*

(Return to start)

# 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): 358–375
  - Stephen C. Carlson, “A Bias at the Heart of the Coherence-Based Genealogical Method (CBGM),” *JBL* 139.2 (2020): 319–340 (but see Mink’s response)

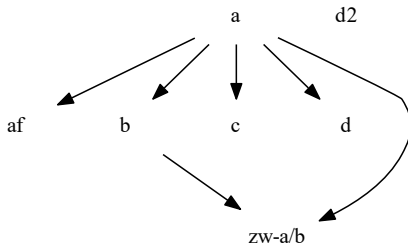
(Return to start)

## Collation

```
<?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">ενεε</rdg>
          <rdg n="b" wit="#B">ελεε</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>
```



Local stemma



```
<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

`agree` = [1, 0, 0, 0, 0]

`prior` = [0, 1, 0, 0, 0]

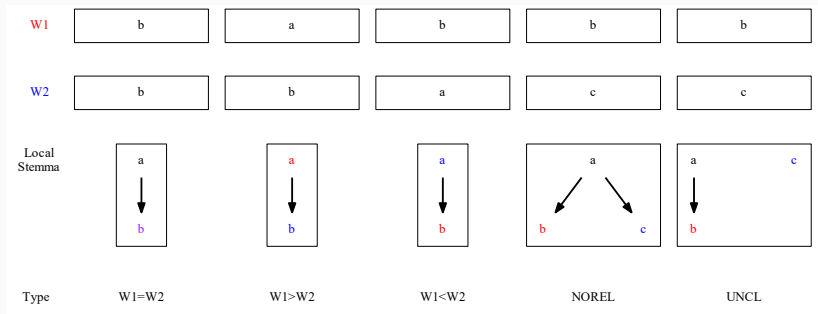
`posterior` = [0, 0, 1, 0, 0]

`norel` = [0, 0, 0, 1, 0]

`uncl` = [0, 0, 0, 0, 1]

`expl` = [1, 0, 1, 0, 0]

`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>)

(Return to start)

## References

- Carlson, Stephen C. “A Bias at the Heart of the Coherence-Based Genealogical Method (CBGM).” *JBL* 139.2 (2020): 319–340.
- 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): 358–375.
- 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/>.

———. "Luke 9,35 and the Power of Polygenesis." *FilNeot* 33.53 (2020): 51–94.

———. "The open-cbqm 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): 19–105.

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.