

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](https://twitter.com/jamesjmccollum)

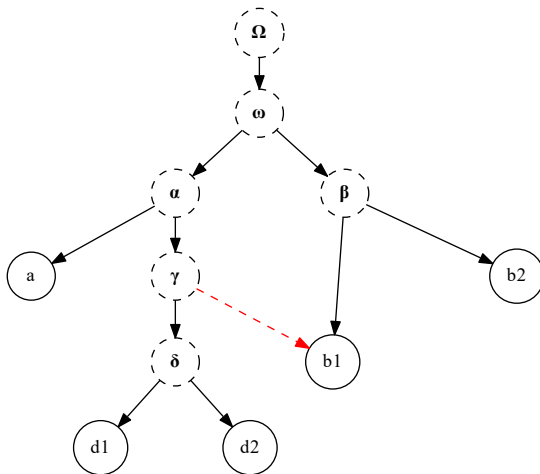
 [jjmccollum](https://github.com/jjmccollum)



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 Stemmata 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/>

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



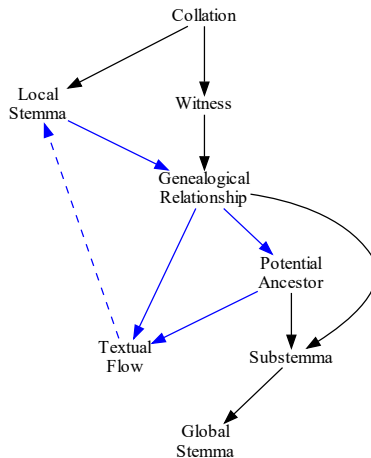


- Methodological assumptions:
 1. Scribes typically copied their exemplars with fidelity.
 2. If a scribe introduced a variant, then it came from some other reading.
 3. Scribes typically used fewer sources rather than many.
 4. Scribes typically used closely related sources rather than distant ones.



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

- “Iterative workflow” highlighted in blue



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

	1	2	3	10.1-4
οὐ ἤμελλον αὐτοῦς ἔρχεσθαι.	2 ἔλεγεν	δὲ πρὸς αὐτούς,	Ὁ μὲν θερσιμός	B K C 1071 uw
οὐ ἤμελλον αὐτοῦς ἔρχεσθαι.	2 ἔλεγεν	δὲ πρὸς αὐτούς,	Ὁ μὲν θερ-	φ ⁷⁵
οὐ ἤμελλον αὐτοῦς εἰσεργεσθαι.	2 ἔλεγεν	οὖν πρὸς αὐτούς,	Ὁ μὲν θερσιμός	A
οὐ ἔμελλον ἔρχεσθαι.	2 ἔλεγεν	δὲ πρὸς αὐτούς,	Ὁ θερσιμός	D
οὐ ἔμελλον αὐτοῦς ἔρχεσθαι.	2 ἔλεγεν	οὖν πρὸς αὐτούς,	Ὁ μὲν θερσιμός	Y K S II 28 565 τ
οὐ ἔμελλον αὐτοῦς ἔρχεσθαι.	2 ἔλεγεν	δὲ πρὸς αὐτούς,	Ὁ μὲν θερσιμός	L 124 579
οὐ ἤμελλον αὐτοῦς εἰσεργεσθαι.	2 ἔλεγεν	δὲ πρὸς αὐτούς,	Ὁ μὲν θερσιμός	Θ
οὐ ἔμελλον αὐτοῦς ἀπεργεσθαι.	2 ἔλεγεν	οὖν πρὸς αὐτούς,	Ὁ μὲν θερσιμός	Ω
οὐ ἔμελλον αὐτοῦς εἰσεργεσθαι.	2 εἶπεν	δὲ πρὸς αὐτούς,	Ὁ μὲν θερσιμός	f ¹
οὐ ἤμελλον αὐτοῦς διεργεσθαι.	2 ἔλεγεν	δὲ πρὸς αὐτούς,	Ὁ μὲν θερσιμός	f ¹³
..... ὅς ἐρχεσθαι.	2 ἔλεγεν	δὲ πρὸς αὐτούς,	Ὁ μὲν θερσιμός	33
οὐ ἤμελλον αὐτοῦς εἰσπορεύεσθαι.	2 ἔλεγεν	δὲ πρὸς αὐτούς,	Ὁ μὲν θερσιμός	157
οὐ ἔμελλον αὐτοῦς πορευέσθαι.	2 ἔλεγεν	δὲ πρὸς αὐτούς,	Ὁ μὲν θερσιμός	700 [↓124]
οὐ ἤμελλον αὐτοῦς ἔρχεσθαι.	2 ἔλεγεν	οὖν πρὸς αὐτούς,	Ὁ μὲν θερσιμός	¶ M N U W Γ Δ Λ Ψ 2
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεθήθητε	4 οὖν τοῦ κυρίου	τοῦ θηρισμοῦ ὅπως	αν	B φ ⁷⁵ uwτ tell
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεθήθητε	5 οὖν τοῦ κυρίου	τοῦ θηρισμοῦ ὅπως	αν	Y K M JJ
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεθήθητε	6 τοῦ θεοῦ	τοῦ θηρισμοῦ ὅπως		D*
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεθήθητε	7 τοῦ κυρίου	τοῦ θηρισμοῦ ὅπως		D ^c
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεθήθητε	8 οὖν τοῦ κυρίου	τοῦ θηρισμοῦ ὅπως		H
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεθήθητε	9 οὖν τοῦ κυρίου	τοῦ θηρισμοῦ ὅπως		33
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεθήθητε	10 οὖν τοῦ κυρίου	τοῦ θηρισμοῦ ὅπως		579

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

- Variation units serve as our points of comparison between witnesses 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

- This is readily encoded in TEI XML format

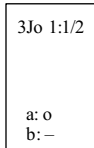
```

1 <?xml version='1.0' encoding='UTF-8'?>
2 <TEI xmlns="http://www.tei-c.org/ns/1.0">
3   <teiHeader>
4     <fileDesc>
5       <titleStmt>
6         <title>A collation of Luke 10:2 in Swanson</title>
7       </titleStmt>
8       <publicationStmt>
9         <p>Swanson, Reuben J., ed. <emph>New Testament Greek Manuscripts: Variant
10       </publicationStmt>
11       <sourceDesc>
12         <listWit>
13           <witness n="P75"/>
14           <witness n="f1"/>
15           <witness n="f13"/>
16         </listWit>
17       </sourceDesc>
18     </fileDesc>
19   </teiHeader>
20   <text xml:lang="GRC">
21     <body>
22       <div type="book" n="B03">
23         <div type="chapter" n="B03K10">
24           <ab n="B03K10V2">
25             <app n="B03K10V2U2">
26               <rdg n="1" wit="f13"><w>ελεγεεν</w></rdg>
27               <rdg n="1-f1" type="defective" wit="P75">
28                 <w><gap/><unclear>λε</unclear>υ<unclear>εν</unclear></w>
29               </rdg>
30               <rdg n="2" wit="f1"><w>ελεεν</w></rdg>
31             </app>
32           </ab>
33         </div>
34       </div>
35     </body>
36   </text>
37 </TEI>

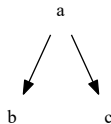
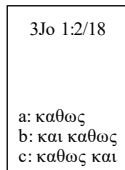
```

```
<app n="B03K10V2U2">
  <rdg n="1" wit="f13"><w>ελεγεν</w></rdg>
  <rdg n="1-f1" type="defective" wit="P75">
    <w><gap></unclear>λε</unclear>γ<unclear>εν</unclear></w>
  </rdg>
  <rdg n="2" wit="f1"><w>επιτεν</w></rdg>
</app>
```

- The basic unit of comparison
- One for each variation unit
- A graphical representation of our judgments of readings
- Kurt Aland's "local genealogical" principle



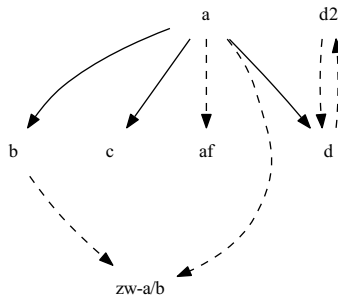
a
↓
b



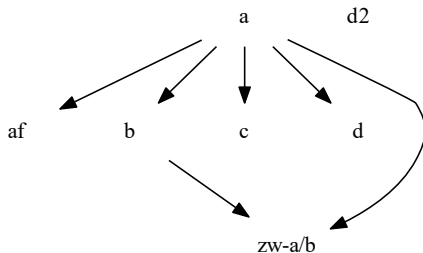
- Some are more complicated
 - defective* readings (e.g., misspellings, reconstructions)
 - orthographic* readings (e.g., regional differences)
 - split* attestations of the same reading (coincidental agreement)
 - ambiguous* readings (can be reconstructed as more than one reading)
- Some of these may 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



- Computationally, just a directed graph.



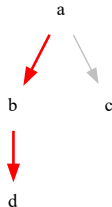
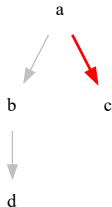
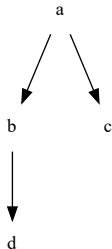
```

<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>
    
```

- Relationships between readings are determined by checking for a path between them
- $a = b$: path of length 0
- $a > b$: path of length > 0 from a to b
- NOREL: no path from a to b , but both have a *common ancestor*
- UNCL: same as NOREL, but no common ancestor (reserved for when we don't know where a reading fits in the stemma)

Acts 2:5/26-30

a: ὑπο τὸν οὐρανὸν
b: ὑπο τῶν οὐρανῶν
c: ὑπο τοῦ οὐρανοῦ
d: ὑπ οὐρανὸν



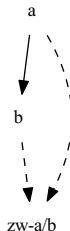
- For the CBGM's purposes, a *witness* is a sequence of readings
- Typically, the *text* of a known manuscript, minus the material baggage (date, provenance, etc.)
 - “How texts relate” \neq “How manuscripts relate”

	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

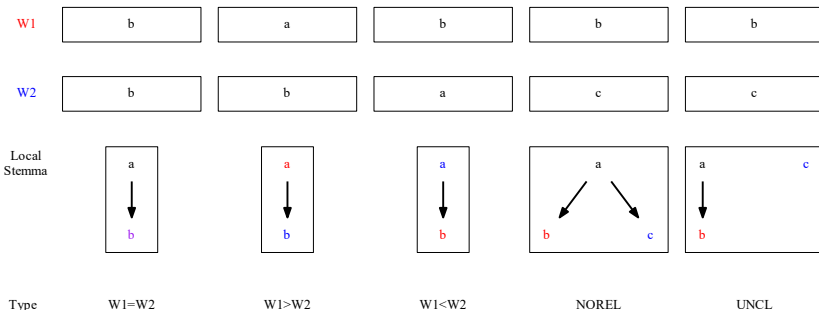
- Versions and fathers can also be treated as witnesses
- But back-translation may be ambiguous, and patristic citations may be “lacunose”

3Jo 1:1/2

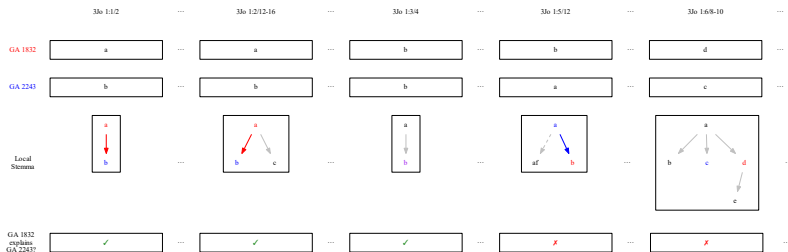
a: o ... CosmIn. PsOec. S:H
 b: — 467. 2243. 2828
 zw-a/b: L:VT. K:SB. S:Ph



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



- 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

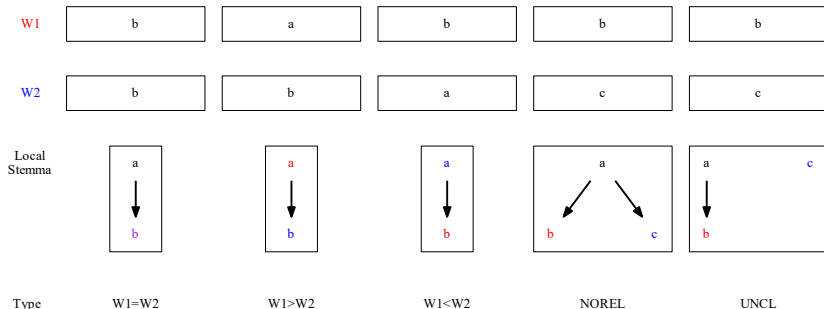


- Lacunae do not have to be explained, and they cannot explain readings
- The *cost* of a genealogical relationship is the number of explained readings that are not agreements (so the cost in the example above is 2)

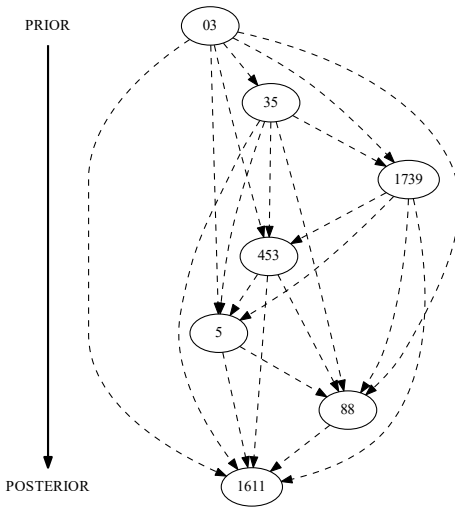
It is convenient to encode genealogical relationships with *bit-arrays*

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



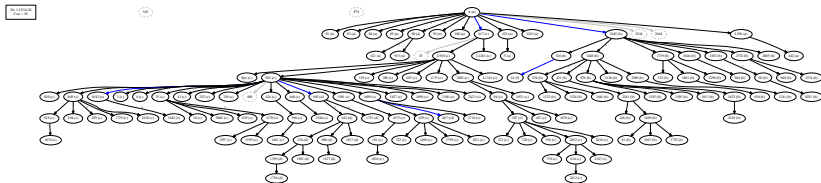
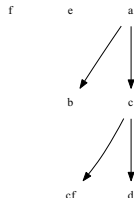
- Potential ancestor = “more prior than posterior readings”



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

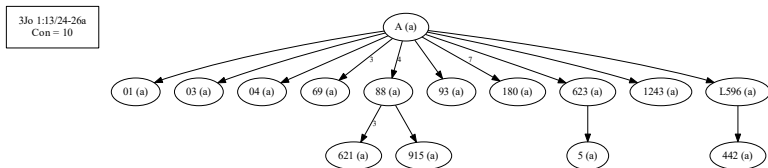
3Jo 1:13/24-26

a: σοι γραφειν
b: γραφειν σοι
c: σοι γραφαι
cf: σοι σοι γραφαι
d: γραφαι σοι
e: γραφαι
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

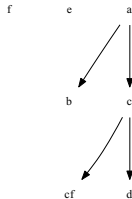


- (Numbers on edges represent the rank of the closest potential ancestor with the same reading, if it's not 1)

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

3Jo 1:13/24-26

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

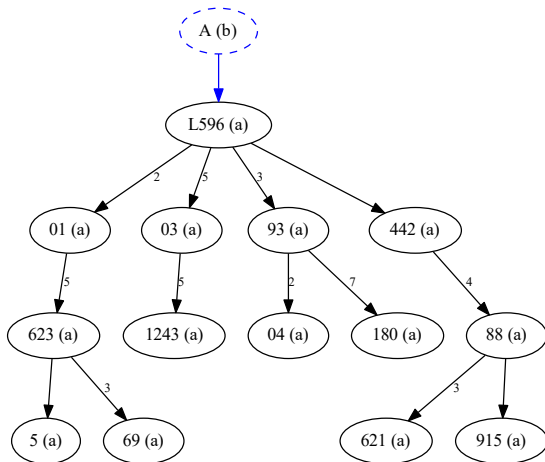


3Jo 1:13/24-26

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



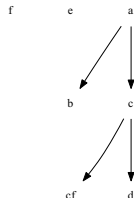
3Jo 1:13/24-26a
Con = 10

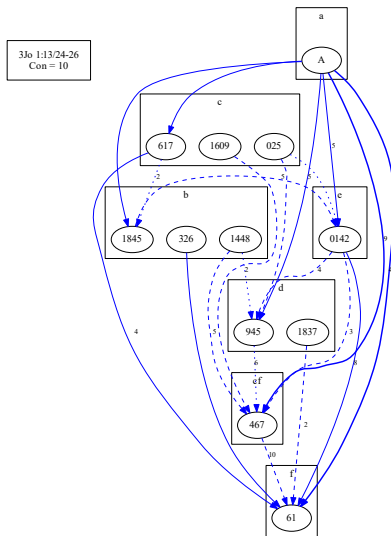


- Or, we can look only at the parts of textual flow where a reading gets changed to find the most likely sources of unexplained readings (*e* and *f*)

3Jo 1:13/24-26

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

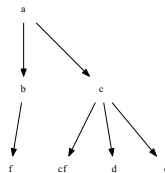




- Between coherence (a form of external evidence) and internal evidence, we can attempt to explain previous unexplained readings
- A necessary step for our ultimate goal of constructing a global stemma

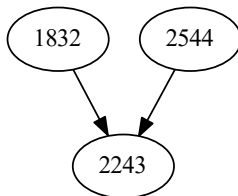
3Jo 1:13/24-26

a: σοι γραφειν
b: γραφειν σοι
c: σοι γραφαι
cf: σοι σοι γραφαι
d: γραφαι σοι
e: γραφαι
f: —

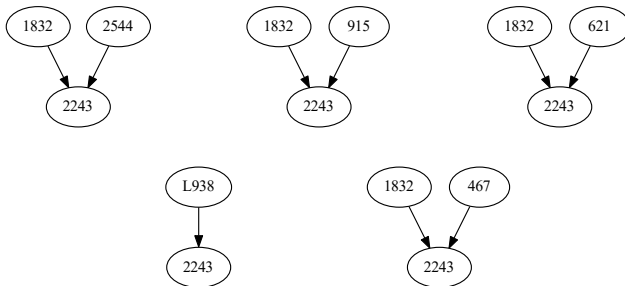


- The *substemma* of a witness is the portion of the global stemma consisting of the witness and its ancestors in the stemma
- Requirement: *every* extant reading in the witness must be explained by a reading in at least one of its ancestors

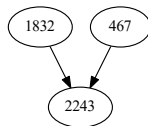
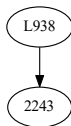
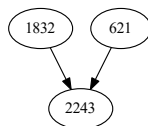
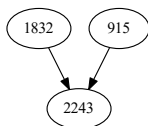
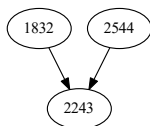
Explained by GA 1832	...	✗	✓	✓	✓	...
Explained by GA 2544	...	✓	✗	✗	✓	...
Explained by Either	...	✓	✓	✓	✓	...



- 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

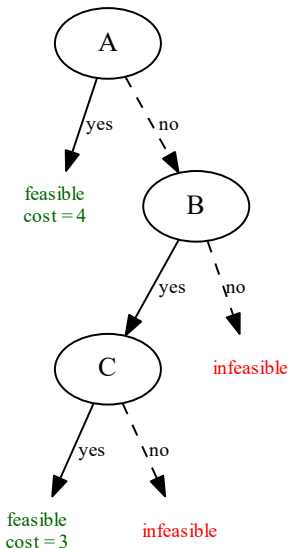


- A simple cost function for each ancestor is “the number of variation units where the ancestor explains the witness by descent and not agreement”

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

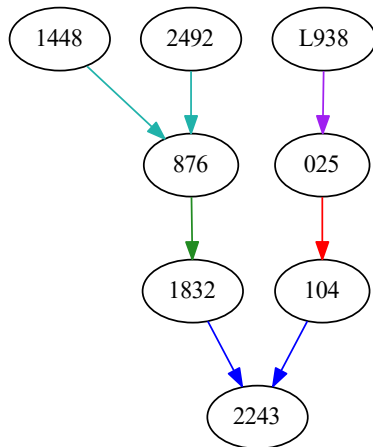
- Also called *sublemma optimization*
- For n potential ancestors, a *weighted set cover* problem with n sets (and $2^n - 1$ combinations to check!)

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



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

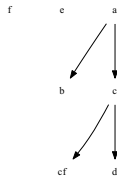
- Just as the local stemma relates readings, the *global stemma* relates witnesses
- Combination of all substemmata into a single graph



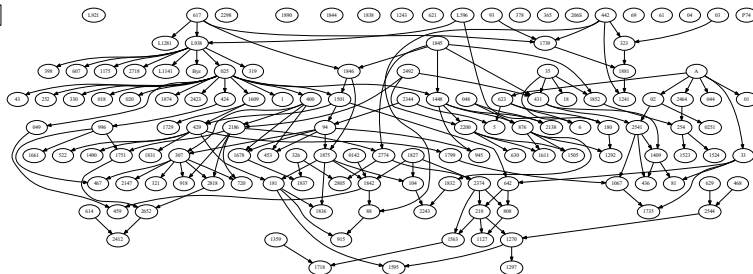
- But *every reading in every local stemma* except the initial one must be explained by another reading
- Otherwise...

3Jo 1:13/24-26

a: σοι γραφειν
b: γραφειν σοι
c: σοι γραφαι
cf: σοι σοι γραφαι
d: γραφαι σοι
e: γραφαι
f: -



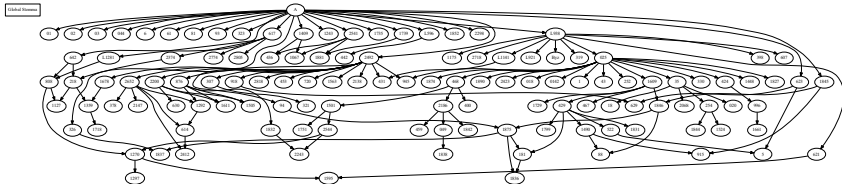
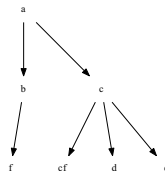
Global Stemma



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

3Jo 1:13/24-26

a: σοι γραφειν
b: γραφειν σοι
c: σοι γραφαι
cf: σοι σοι γραφαι
d: γραφαι σοι
e: γραφαι
f: -





- 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



- 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-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): 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.