

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)

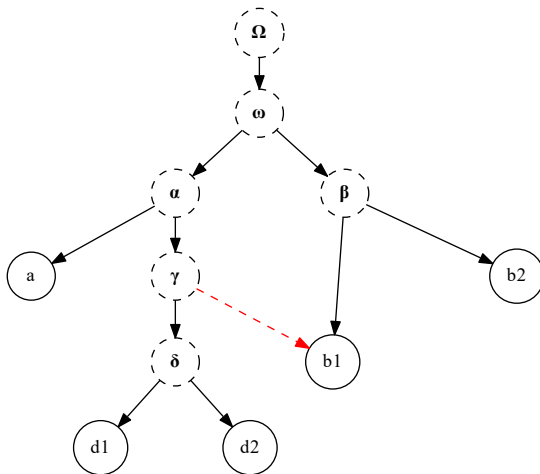
 [jjmccollum](https://github.com/jjmccollum)



ACU INSTITUTE FOR
RELIGION &
CRITICAL INQUIRY

- Developed over thirty years by Gerd Mink, culminating in the latest updates to the *Editio Critica Maior* (ECM)
- Recommended 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





- Key assumption: *no hypothetical ancestors* (except the *Ausgangstext A*)
- Other important 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.

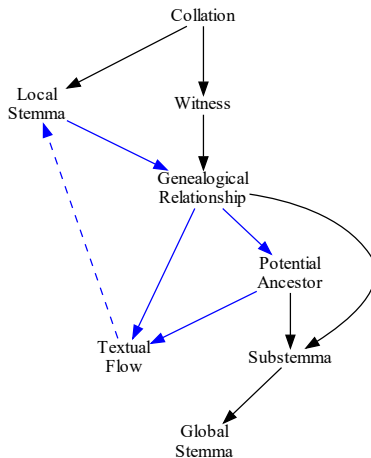


- Key assumption: *no hypothetical ancestors* (except the *Ausgangstext A*)
- Other important assumptions:
 1. Scribes typically copied their exemplars with fidelity.
 2. If a scribe introduced a variant, then it came from some other reading.
 3. Scribes typically used fewer sources rather than many.
 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

ΚΑΤΑ ΛΟΥΚΑΝ		10.1-4	
	1	2	3
οὐ ἤμελλεν αὐτοὺς ἔρχεσθαι.	2 ἔλεγεν	δὲ πρὸς αὐτούς, Ὁ μὲν	θηρισμὸς B K C 1071 uw
οὐ ἤμελλεν αὐτοὺς ἔρχεσθαι.	2 ἔλεγεν	δὲ πρὸς αὐτούς, Ὁ μὲν	θηρ- 75
οὐ ἤμελλεν αὐτοὺς <u>εἰσέρχεσθαι</u> .	2 ἔλεγεν	οὖν πρὸς αὐτούς, Ὁ μὲν	θηρισμὸς A
οὐ ἔμελλεν ἔρχεσθαι.	2 ἔλεγεν	δὲ πρὸς αὐτούς, Ὁ	θηρισμὸς D
οὐ ἔμελλεν αὐτοὺς ἔρχεσθαι.	2 ἔλεγεν	οὖν πρὸς αὐτούς, Ὁ μὲν	θηρισμὸς Y K S II 28 565 τ
οὐ ἔμελλεν αὐτοὺς ἔρχεσθαι.	2 ἔλεγεν	δὲ πρὸς αὐτούς, Ὁ μὲν	θηρισμὸς L 124 579
οὐ ἤμελλεν αὐτοὺς <u>εἰσέρχεσθαι</u> .	2 ἔλεγεν	δὲ πρὸς αὐτούς, Ὁ μὲν	θηρισμὸς Θ
οὐ ἔμελλεν αὐτοὺς <u>ἀπερχεσθαι</u> .	2 ἔλεγεν	οὖν πρὸς αὐτούς, Ὁ μὲν	θηρισμὸς Ω
οὐ ἔμελλεν αὐτοὺς <u>εἰσέρχεσθαι</u> .	2 εἶπεν	δὲ πρὸς αὐτούς, Ὁ μὲν	θηρισμὸς f ¹
οὐ ἤμελλεν αὐτοὺς <u>διέρχεσθαι</u> .	2 ἔλεγεν	δὲ πρὸς αὐτούς, Ὁ μὲν	θηρισμὸς f ¹³
οὐ ἤμελλεν αὐτοὺς ἔρχεσθαι.	2 ἔλεγεν	δὲ πρὸς αὐτούς, Ὁ μὲν	θηρισμὸς 33
οὐ ἤμελλεν αὐτοὺς <u>εἰσπορεύεσθαι</u> .	2 ἔλεγεν	δὲ πρὸς αὐτούς, Ὁ μὲν	θηρισμὸς 157
οὐ ἔμελλεν αὐτοὺς <u>πορεύεσθαι</u> .	2 ἔλεγεν	δὲ πρὸς αὐτούς, Ὁ μὲν	θηρισμὸς 700 [J 1242]
οὐ ἤμελλεν αὐτοὺς ἔρχεσθαι.	2 ἔλεγεν	οὖν πρὸς αὐτούς, Ὁ μὲν	θηρισμὸς 98 M N U W Γ Δ Λ Ψ 2
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε	4 οὖν	τοῦ κυρίου	τοῦ θρισμοῦ ὅπως B φ ⁷⁵ uwτ tell
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε	4 οὖν	τοῦ κυρίου	τοῦ θρισμοῦ ὅπως αν Y K M JJ
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε	4 οὖν	τοῦ θεοῦ	τοῦ θρισμοῦ ὅπως D* 75
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε	4 οὖν	τοῦ κυρίου	τοῦ θρισμοῦ ὅπως D ^c
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε	4 οὖν	τοῦ κυρίου	τοῦ θρισμοῦ ὅπως H
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε	4 οὖν	τοῦ κυρίου	τοῦ θρισμοῦ ὅπως 33
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε	4 οὖν	τοῦ κυρίου	τοῦ θρισμοῦ 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>

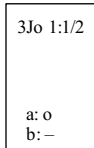
```

ἐλεγεν	B K C 1071 uw
ἐλεγεν	ϕ ⁷⁵
ἐλεγεν	A
ἐλεγεν	D
ἐλεγεν	Y K S II 28 565 τ
ἐλεγεν	L 124 579
ἐλεγεν	Θ
ἐλεγεν	Ω
ἐλεγεν	f ¹
ἐλεγεν	f ¹³
ἐλεγεν	33
ἐλεγεν	157
ἐλεγεν	700 [↓1424
ἐλεγεν	Ϡ R M N U W Γ Δ Λ Ψ 2

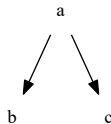
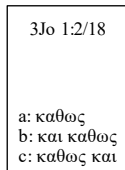
```
<app n="B03K10V2U2">
  <rdg n="1" wit="f13"><w>ἐλεγεν</w></rdg>
  <rdg n="1-f1" type="defective" wit="P75">
    <w><gap/><unclear>λε</unclear>γ<unclear>εν</unclear></w>
  </rdg>
  <rdg n="2" wit="f1"><w>ελεγεν</w></rdg>
</app>
```

```
reading_support = {
    "f13": "1",
    "P75": "1-f1",
    "f1": "2"
}
```

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



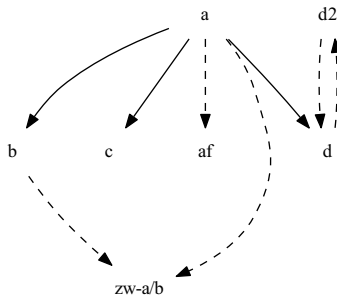
a
↓
b



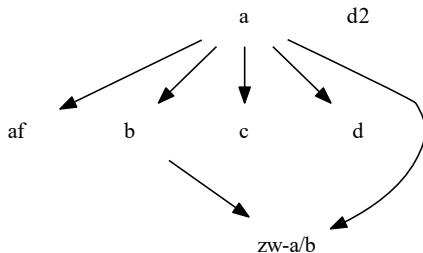
- Some are more complicated
 - *defective* readings (e.g., misspellings, reconstructions)
 - *orthographic* readings (e.g., regional differences)
 - *split* attestations of the same reading (coincidental emergence)
 - *ambiguous* readings (can be reconstructed as more than one reading)
- Some of these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



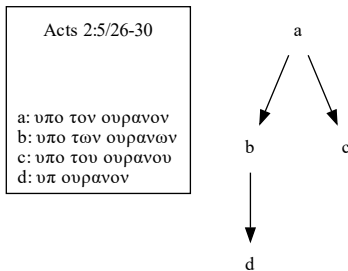
- Computationally, just a directed graph.



```

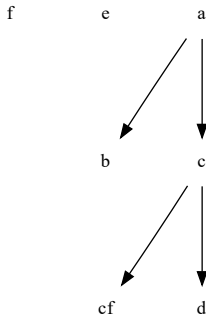
<graph type="directed">
  <node n="a" />
  <node n="af" />
  <node n="b" />
  <node n="c" />
  <node n="d" />
  <node n="d2" />
  <node n="zw-a/b" />
  <arc from="a" to="af" />
  <arc from="a" to="b" />
  <arc from="a" to="c" />
  <arc from="a" to="d" />
  <arc from="a" to="zw-a/b" />
  <arc from="b"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

- Relationships between readings are determined by checking for a path between them
- $a = b$ (agreement): path of length 0
- $a > b$ (prior): path of length > 0 from a to b
- $a < b$ (posterior): path of length > 0 from b to a
- NOREL (no relationship): no path from a to b , but both have a *common ancestor*



- UNCL (unclear): same as NOREL, but no common ancestor (reserved for when we don't know where a reading fits in the stemma)
- We say that one reading *explains* another if
 - it is the same reading (explanation by agreement), or
 - there is a path of length 1 from it to the other reading
- Lacunae do not have to be explained, and they cannot explain readings

3Jo 1:13/24-26
a: σοι γραφειν
b: γραφειν σοι
c: σοι γραψαι
cf: σοι σοι γραψαι
d: γραψαι σοι
e: γραψαι
f: —



- 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

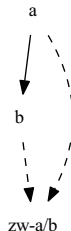
- Versions and fathers can also be treated as witnesses
- But back-translation may be ambiguous, and patristic citations may be “lacunose”

3Jo 1:1/2

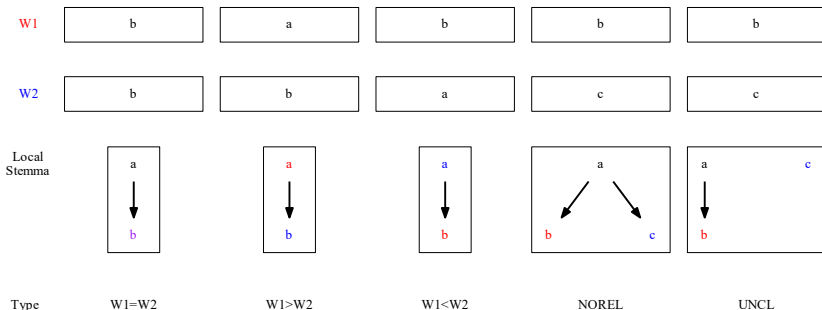
a: o ... CosmIn. PsOec. S:H

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* where both are extant
- The *cost* of a genealogical relationship is the number of explained readings that are not agreements (so the cost in the example below is 1)



- It is convenient to encode genealogical relationships with *bitmaps*

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

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

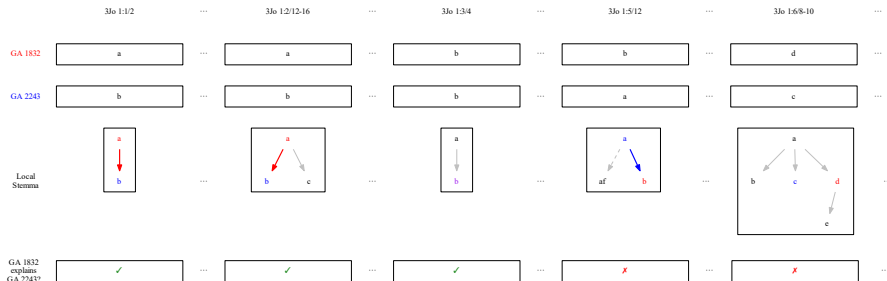
posterior = [0, 0, 1, 0, 0]

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

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

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

cost = 1

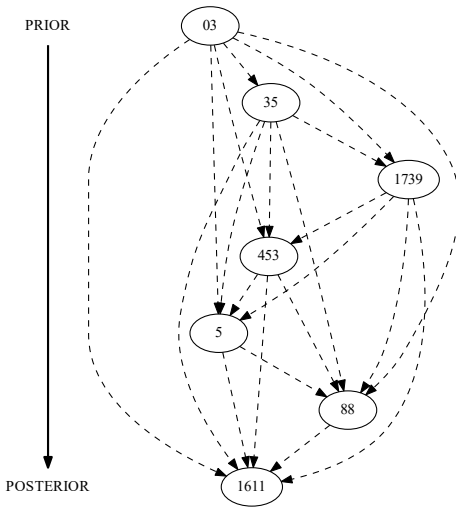


- For d units and n witnesses, $\sim n^2 d$ comparisons as one-time work
- The `compare_witnesses` module (below) presents this computed data

Genealogical comparisons for W1 = 5 (116 extant passages):

W2	DIR	NR	PASS	EQ	W1>W2	W1<W2	NOREL	UNCL	EXPL	COST
623	>	1	116	112 (96.552%)	0	3	1	0	115	3.000
A	>	2	116	104 (89.655%)	0	12	0	0	116	12.000
025	>	3	116	103 (88.793%)	4	8	0	1	111	8.000
319	>	3	116	103 (88.793%)	4	8	0	1	111	8.000
398	>	3	116	103 (88.793%)	4	8	0	1	111	8.000
607	>	3	116	103 (88.793%)	4	8	0	1	111	8.000
617	>	3	116	103 (88.793%)	3	9	0	1	112	9.000
1175	>	3	116	103 (88.793%)	4	8	0	1	111	8.000
1890	>	3	116	103 (88.793%)	3	7	1	2	110	7.000
Byz	>	4	114	102 (89.474%)	3	8	0	1	110	8.000
049	>	4	116	102 (87.931%)	6	7	0	1	109	7.000
0142	>	4	116	102 (87.931%)	4	8	0	2	110	8.000
1	>	4	116	102 (87.931%)	5	8	0	1	110	8.000
35	>	4	116	102 (87.931%)	4	8	0	2	110	8.000
326	>	4	116	102 (87.931%)	6	7	0	1	109	7.000
424	>	4	116	102 (87.931%)	5	8	0	1	110	8.000
468	>	4	116	102 (87.931%)	4	8	0	2	110	8.000
1448	>	4	116	102 (87.931%)	6	8	0	0	110	8.000
1609	>	4	116	102 (87.931%)	5	8	0	1	110	8.000
2186	>	4	116	102 (87.931%)	6	7	0	1	109	7.000
2423	>	4	116	102 (87.931%)	5	8	0	1	110	8.000
2805	=		115	102 (88.696%)	6	6	0	1	108	
L938	>	4	116	102 (87.931%)	4	9	0	1	111	9.000
018	=		116	101 (87.069%)	7	7	0	1	108	
18	>	5	116	101 (87.069%)	5	8	0	2	109	8.000
43	=		116	101 (87.069%)	7	7	0	1	108	

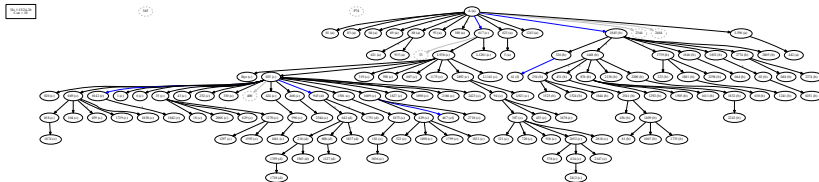
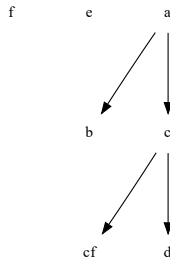
- Potential ancestor = “more prior than posterior readings”



- *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

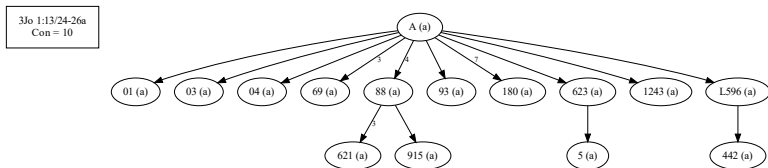
3Jo 1:13/24-26

a: σοι γραφειν
b: γραφειν σοι
c: σοι 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

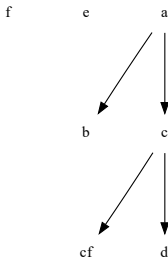


- (Numbers on edges represent the rank of the closest potential ancestor with the same reading, if it's not 1)

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

3Jo 1:13/24-26

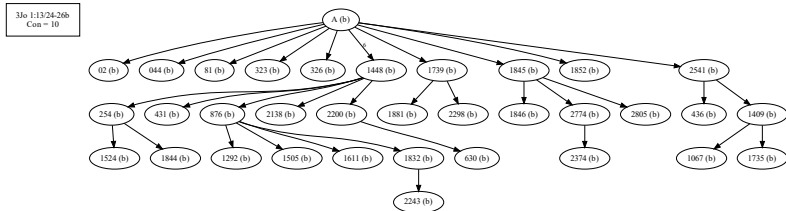
a: σοι γραφειν
b: γραφειν σοι
c: σοι γραψαι
cf: σοι σοι γραψαι
d: γραψαι σοι
e: γραψαι
f: —



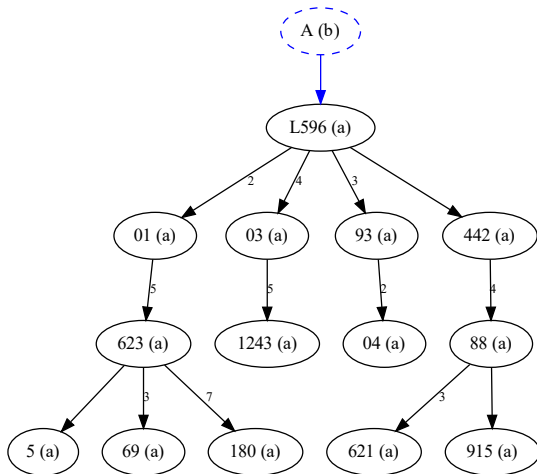
3Jo 1:13/24-26

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





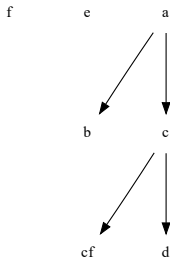
3Jo 1:13/24-26a
Con = 10



- 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

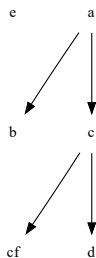
a: σοι γραφειν
b: γραφειν σοι
c: σοι γραψαι
cf: σοι σοι γραψαι
d: γραψαι σοι
e: γραψαι
f: —



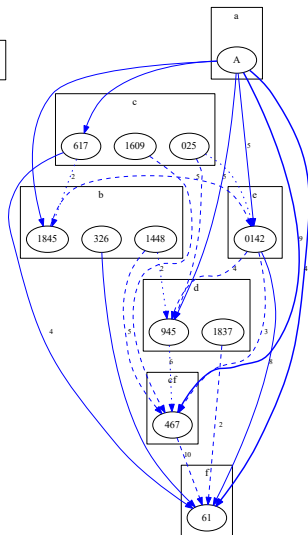
3Jo 1:13/24-26

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

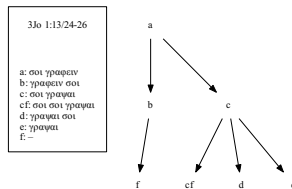
f



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

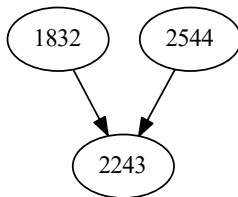


- Between coherence (a form of external evidence) and internal evidence, we can attempt to explain previous unexplained readings
- A necessary step for our ultimate goal of constructing a global stemma

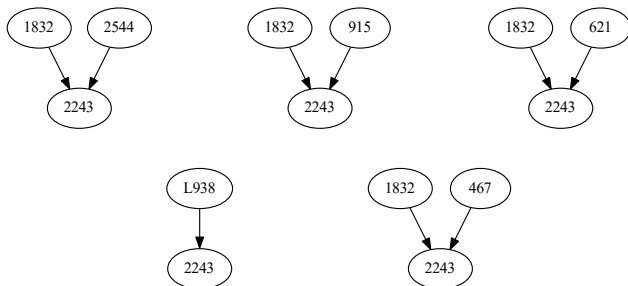


- The *substemma* of a witness is the portion of the global stemma consisting of the witness and its ancestors in the stemma
- Requirement: *every* extant reading in the witness must be explained by a reading in at least one of its ancestors

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

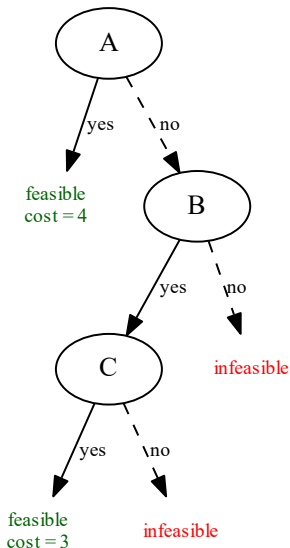


- A witness may have multiple valid substemma (i.e., ones that explain all of its readings), but some are better than others
- Two of the CBGM's methodological assumptions are important here:
 3. Scribes typically used fewer sources rather than many.
 4. Scribes typically used closely related sources rather than distant ones.
- A balancing act: the substemma {L938} is more parsimonious, but may not explain as many readings by agreement



- 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



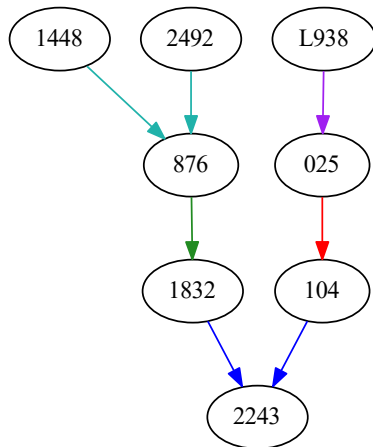
- If a witness has many potential ancestors, then checking all $2^n - 1$ possible substemmata by brute force is prohibitive
- The *branch-and-bound* heuristic (pictured left) finds all minimum-cost substemmata quickly in practice
- Easily adapted to find all substemmata within a given cost

- In practice, the `optimize_substemmata` module finds all desired substemmata in seconds.

```
Substemmata for witness W1 = 5 (116 extant passages):
```

ANCESTORS	COST	AGREE
623, 2344	9	113
623, 69	9	112
623, 33	9	112
623, 326	10	113
623, 2541	10	113
623, 1890	10	112
623, 049	10	112
623, 2186	10	112
623, 307	10	112
623, 400	10	112
623, 429	10	112
623, 453	10	112
623, 918	10	112
623, 180	10	112

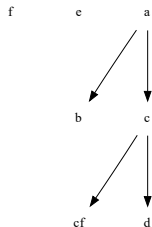
- Just as the local stemma relates readings, the *global stemma* relates witnesses
- Combination of all substemmata into a single graph



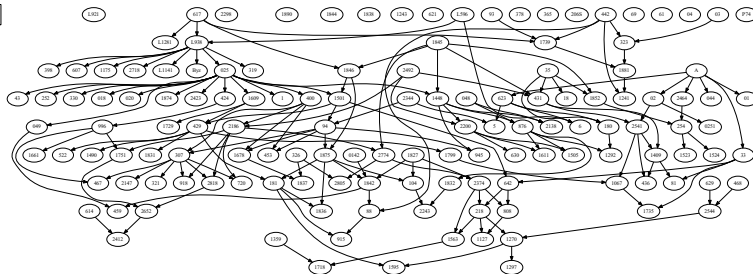
- But *every reading in every local stemma* except the initial one must be explained by another reading
- Otherwise...

3Jo 1:13/24-26

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



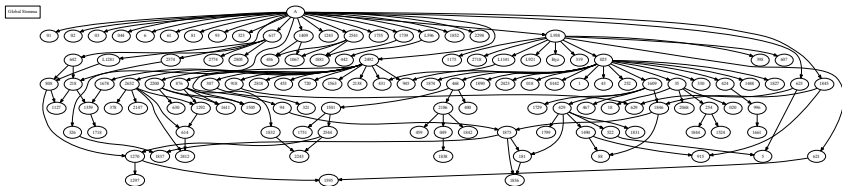
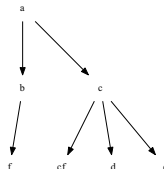
Global Stemma



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

3Jo 1:13/24-26

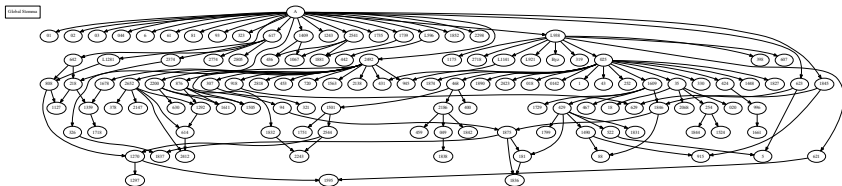
a: σοι γραφειν
b: γραφειν σοι
c: σοι γραφαι
cf: σοι σοι γραφαι
d: γραφαι σοι
e: γραφαι
f: -



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



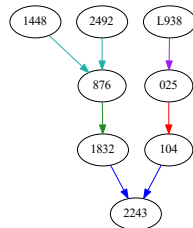
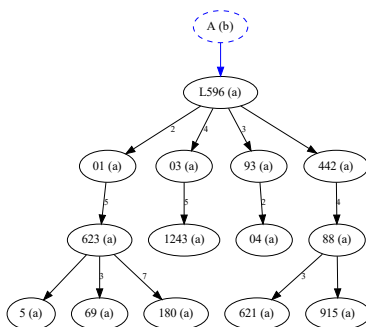
(Source gallica.bnf.fr / Bibliothèque nationale de France)



Field trip

- 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

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





- 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>)

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