

# The open-cbgm Library: Design and Demonstration

---

Joey McCollum\*

9 December 2020

\*Virginia Polytechnic Institute and State University

✉ jjmccollum@vt.edu

🐦 @jamesjmccollum

🐙 jjmccollum

Note: A crash-course document on the CBGM is available at <https://vt.academia.edu/JoeyMcCollum>.

# Customizable, Standard-compliant Input

```
<listWit>
  <witness xml:id="A" />
  <witness xml:id="B" />
</listWit>
```

```
<app n="B00K0V0U0">
  <rdg n="a" wit="#A">ΕΙΠΕΝ</rdg>
  <rdg n="b" wit="#B">ΕΛΕΓΕΝ</rdg>
  <note>
    <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>
```

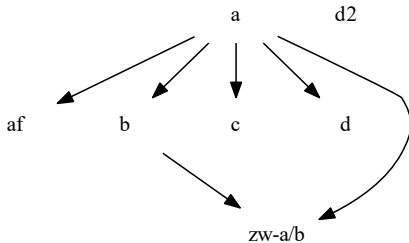
- Database is populated from a collation file in TEI XML format
- With a SQLite database, the entire process can be done locally
- This makes the distribution itself lightweight and portable

XML	Object	DB Table
witness	Witness	WITNESSES
app	Variant passage	VARIATION_ UNITS
rdg	Variant reading	READINGS
graph	Local stemma	READING_ RELATIONS

# Local Stemmata

```
<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>
```

- Local stemmata are fully customizable in the XML input



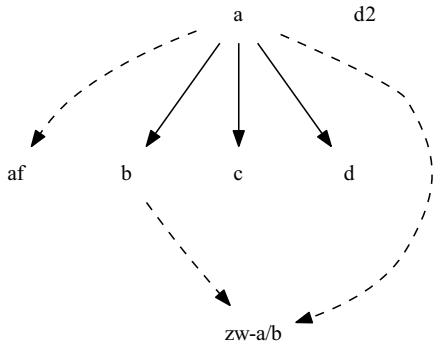
# Local Stemmata

- Reading types can be collapsed

-z defective -z ambiguous

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

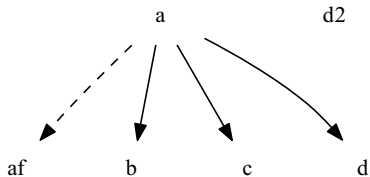


- Reading types can be dropped

-z defective -Z ambiguous

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



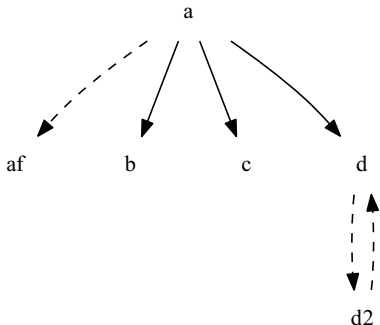
# Local Stemmata

- Split attestations can be merged

-z defective -Z ambiguous --merge-splits

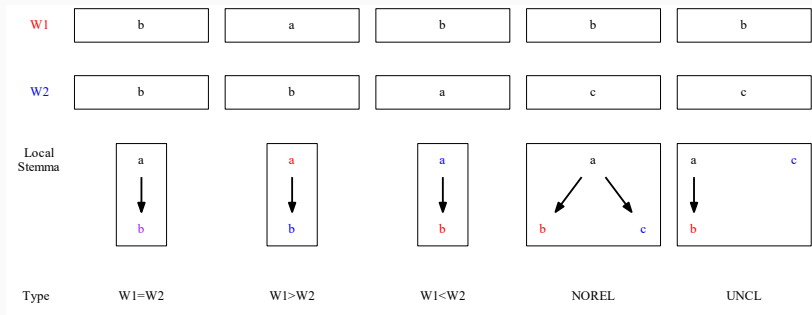
3Jo 1:4/22-26

a: εν αληθεια περιπατουντα  
af: εν αληθεια περιπατουντο  
b: εν τη αληθεια περιπατουντα  
c: περιπατουντα εν αληθεια  
d: τη αληθεια περιπατουντα



# Genealogical Relationships

- The **open-cbgm** library encodes genealogical relationships relative to a given witness as bitmaps (one bit per passage)



$\text{agree} = [1, 0, 0, 0, 0]^*$

$\text{prior} = [0, 1, 0, 0, 0]$

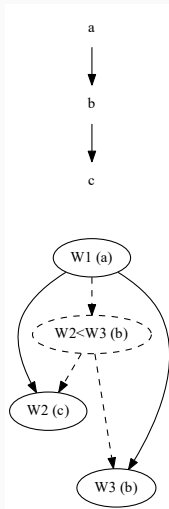
$\text{posterior} = [0, 0, 1, 0, 0]$

$\text{norel} = [0, 0, 0, 1, 0]$

$\text{uncl} = [0, 0, 0, 0, 1]$

$\text{expl} = [1, 0, 1, 0, 0]^*$

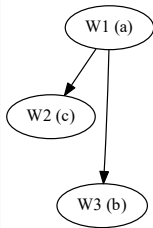
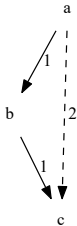
# Criteria for Explained Readings



- In the local stemma to the left, does *a* explain *c*?
- The “classic” rules of the CBGM as implemented by CCeH and Edmondson say *no*: the explaining reading must agree with or be directly prior to the explained reading
- Intermediary nodes may be needed to connect the global stemma



# Criteria for Explained Readings



- The **open-cbgm** implementation relaxes this criterion: any reading with a path in the local stemma to the reading in question explains it
- No intermediary nodes needed (but multiple changes in the same passage may be implied along the edges)

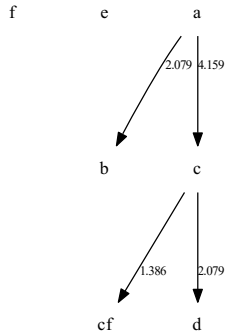
# Genealogical Relationship Costs

```
<graph type="directed">
  <node n="a" />
  <node n="b" />
  <node n="c" />
  <node n="cf" />
  <node n="d" />
  <node n="e" />
  <node n="f" />
  <arc from="a" to="b">
    <label>2.079</label>
  </arc>
  <arc from="a" to="c">
    <label>4.159</label>
  </arc>
  <arc from="c" to="cf">
    <label>1.386</label>
  </arc>
  <arc from="c" to="d">
    <label>2.079</label>
  </arc>
</graph>
```

- In **open-cbgm**, cost = shortest path length
- Edge weights in **label** element
- Example: scribal change  
log-likelihood  $w = -\log p$

3Jo 1:13/24-26

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

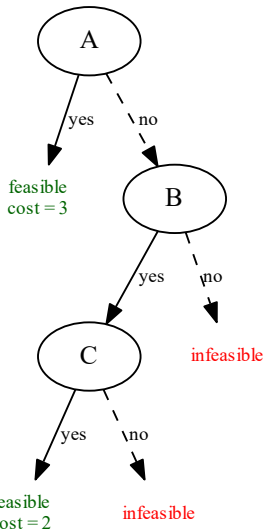


# Sublemma Optimization

Ancestor	Explained	Cost
A	[1, 1, 1, 1]	3
B	[1, 1, 0, 0]	1
C	[0, 1, 1, 1]	1

- Sublemma optimization can be cast as a *weighted set cover* problem
- Example: witness D, extant in four passages, has three potential ancestors A, B, and C
- The sublemma {A} is *feasible*, in that it covers all of D's passages (i.e., explains D's readings at all of them)
- But {B, C} is feasible and *optimal* in terms of total cost (2 rather than 3)

# Substemma Optimization



- For a witness with  $n$  potential ancestors, evaluating all of its  $2^n$  possible substemmata by brute force would be 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

DEMO

# Conclusion

- The **open-cbgm** core library 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
- A more user-friendly web interface is being co-developed with Tim and Jessie Stoel (Phoenix Seminary)
- Currently being used by David Flood (University of Edinburgh) for his PhD dissertation on GA 0150, 0151, 1506 and 2110

## References

- Edmondson, Andrew. *edmondac/CBGM*. Version 2.2. Zenodo, 2018.  
doi:10.5281/zenodo.1296288.
- Edmondson, Andrew Charles. “An Analysis of the Coherence-Based Genealogical Method Using Phylogenetics.” PhD diss., University of Birmingham, 2019.
- Flood, David. *d-flood/apparatus-explorer*.  
<https://github.com/d-flood/apparatus-explorer>.
- . *d-flood/Tendon*. <https://github.com/d-flood/Tendon>.
- Gurry, Peter J. *A Critical Examination of the Coherence-Based Genealogical Method in New Testament Textual Criticism*. NTTSD 55. Leiden: Brill, 2017.

McCollum, Joey. “Biclustering Readings and Manuscripts via Non-negative Matrix Factorization, with Application to the Text of Jude.” *AUSS* 57.1 (2019): 61–89.

———. *jimccollum/open-cbgm*. Version 1.4. Zenodo, 2020.  
doi:10.5281/zenodo.4282614.

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.

Perathoner, Marcello. *cceh/ntg*. <https://github.com/cceh/ntg>.

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.



## Appendix: Comparison of CBGM Implementations

	INTF/CCeH	Edmondson	open- cbgm
Input format	MySQL DB	Python script	TEI XML
Textual flow ancestry con- strained by local stemma edges?	No	Yes	No
Lacunose witnesses allowed in textual flow?	Yes (CCeH)	No	Yes

## Appendix: Comparison of CBGM Implementations

	INTF/CCeH	Edmondson	open-cbgm
Explanation only by same or parent reading?	Yes	Yes	No
Supports weighted input?	No	No	Yes
Genealogical cost function	Disagreement	Disagreement	Shortest path length
Intermediary nodes needed?	Yes	Yes	No