Authorship Verification

John Nerbonne
Rijksuniversiteit Groningen &
Albert-Ludwigs-Universität, Freiburg

Tübingen
WS 2023-24

Authorship verification

- Focus to-date on authorship attribution
  - Select which author wrote document d from n ($n \ll 10$) candidates
  - Who wrote the Federalist No. 10? Jay, Hamilton or Madison?
  - Fixed set of candidates

- In AUTHORSHIP VERIFICATION, the candidate set is less fixed. Did J.K. Rowling write *The Cuckoo's Calling*?
  - Published under the name 'Galbraith'

Verification is harder

- … than attribution
  - In attribution, we have a closed world of n candidates.
  - In attribution, it's enough to say that candidate x is more likely (than others)
  - In authorship verification, we might have to say, "No, the document was written by none of the candidates"
- Also attracts less attention, ◊ for this reason.
- Koppel, M., & Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, *65*(1), 178-187.
- K&W stick w. "similarity-based methods"
  - All the Δ-methods, all of what we've looked at

Data

- 1,000s of bloggers, 38 blogs/author, data over several yrs.
- Data in ordered pairs $< X, Y >$, where X first 500 wd. of a blog, Y last 500
  - X, Y may come from the same blogger
  - 500 wd. is a relatively short doc
- Corpus has 500 pairs $< X, Y >$
- Task: judge whether X and Y are by the same author

- Preprocessing: separate texts into 4-grams
  - The quick brown fox jumped … => ‘Theq’, ‘hequ’, ‘equi’, ‘quic’, ‘uick’ …
  - Spaces ignored! Why?
  - Collect frequencies of 100,000 most freq. 4-grams into vector

Processing

- Given texts separated into 4-grams
  - Frequencies of 100,000 most freq. 4-grams into vector

- Baseline1 (no training)
  - Compare vectors w. cosine (see earlier lectures) or MINMAX:

$$\text{Sim}(X,Y) = \text{minmax}(\vec{X},\vec{Y}) = \frac{\sum_{i=1}^{n}\min(x_i,y_i)}{\sum_{i=1}^{n}\max(x_i,y_i)}$$

  - Minmax tends to emphasize large differences
  - Accuracy in development: 70.6% cosine, 74.2% minmax

Processing

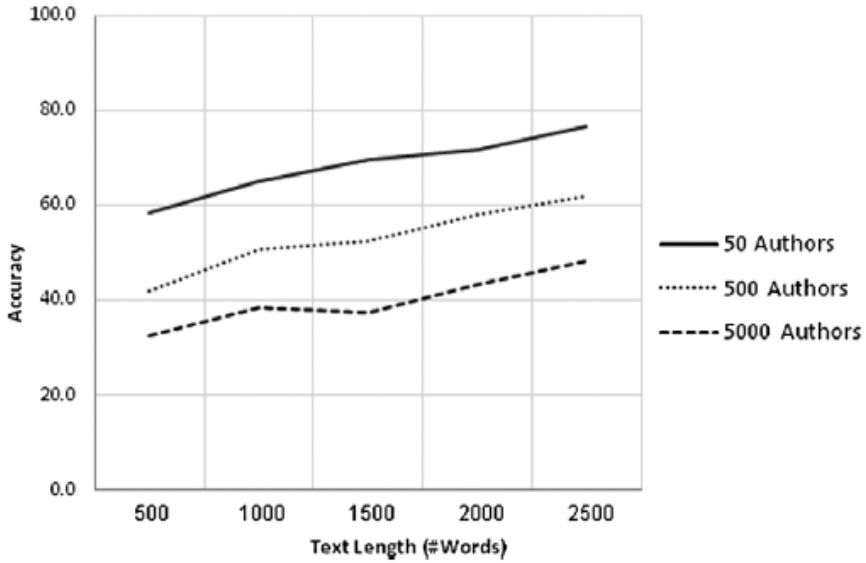
- Baseline 2 (with training)
  - Given $< X, Y >$, define

    $$\text{diff}(X, Y) =< |x_1 - y_1|, |x_2 - y_2|, \ldots |x_n - y_n| >$$

  - Assign each $< X, Y >$ to same-author or different-author
  - Use an ML classifier to learn same- vs. different
    - In fact SVM was used (Tübingen capital of SVM learning)
    - Experimented w. many parameters, best realized 79.8%

Many candidates problem

› Number of candidates crucial!



› Note baselines: 0.02% ≤ chance ≤ 2%

Many candidates

- 5K writers, judged same or different using most similar frequency vector, chance 0.02%
  - Using minmax for similarity
  - 32.5% accurate!
- Inadequate for application
- Instead, repeatedly choose best candidate based on randomly selected feature subset
  - A bit like the BOOTSTRAP method

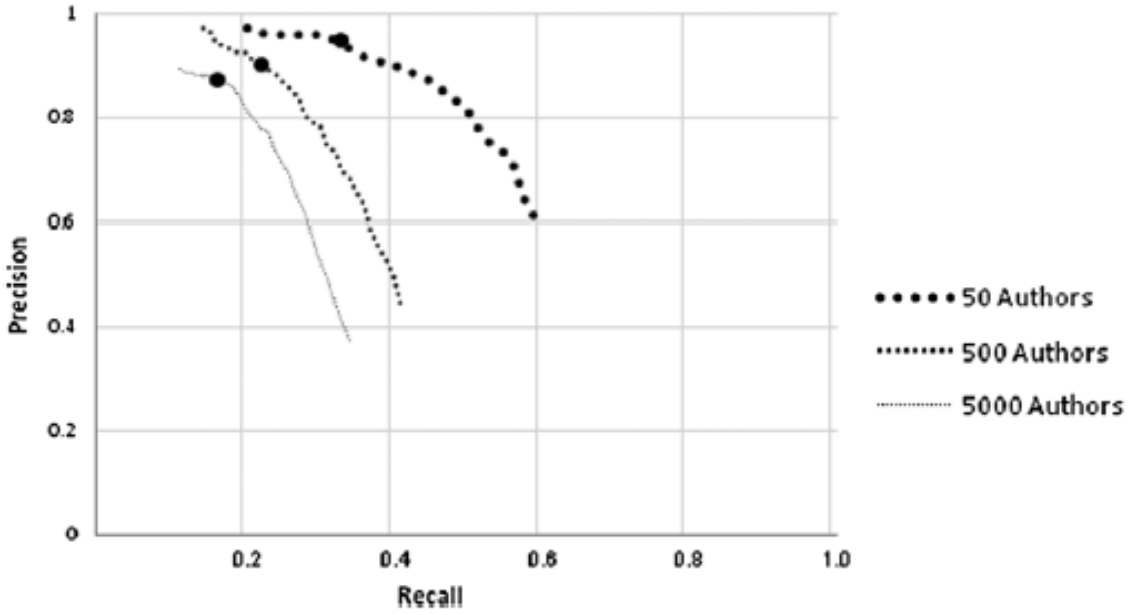
Algorithm

- Given snippet to be assigned, known texts (candidates C)
  - Repeat k times
    - Randomly choose ½ of features (4-grams)
    - Find best match using minmax, candidate c_i
    - Increment c_i's score of best matches
  - For candidate c_i
    - Score(c_i)= proportion of times c_i is best match
- If max Score(c_i) $\geq \sigma^*$, then Output: argmax$_{c_i}$ Score(c_i)
  Else Output 'Don't Know'

- Typically $100 \leq k \leq 1000$, σ^* depends on confidence needed

Idea of many candidates attack

- Reduce dependence on specific words
  - Precision %-age correct attributions,
  - Recall %-age texts attributed correctly
  - Dark dot: $\sigma^* = 0.8$
  - Results:
  - Recall low!



Even harder problem?

- Suppose that the real author is not among the candidates. Harder?
- NO! At $\sigma^* = 0.8$, 3.7% false positives of 5K candidates, 5.5% for 500 cand., 8.4% for 50
- Smaller candidate sets raise chance of consistently more similar text, leading to incorrect attribution!

- To be leveraged in the verification problem

Strategy of attack

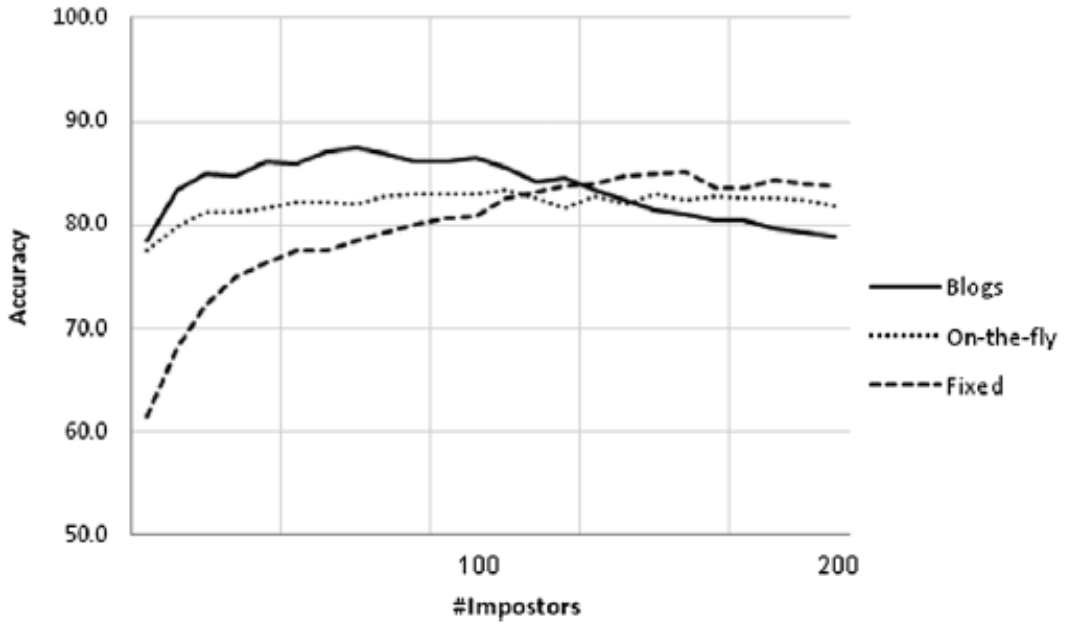
- Verification: given $< X, Y >$, are they by the same author?

- We reduce this to the many-candidates problem (discussed above), which asks which $c \in C$ wrote a given doc d

  - Introduce set "imposters" (*Hochstapler*) for Y: $\{Y_1, \ldots, Y_m\}$
  - Comp: $\text{Score}_X(Y)$= % feat. sets ∋ $Sim(X, Y) > Sim(X, Y_i)$
  - Similarly, gen. imposters for X: $\{X_1, \ldots, X_m\}$
  - …and comp. $\text{Score}_Y(X)$ analogously
  - If $(\text{Score}_X(Y) + \text{Score}_Y(X))/_2 \geq \sigma^*$, then $< X, Y >$ are co-authored

Parameters in approach

- How to choose imposters
  - Fixed: no special relation to doc pair
  - On-the-fly: based on docs returned by Google queries on med. freq. words from $< X, Y >$ [same content]
    - No knowledge needed
  - Blogs [same genre]
  - More imposters →
    - More false negative
    - Fewer false positives

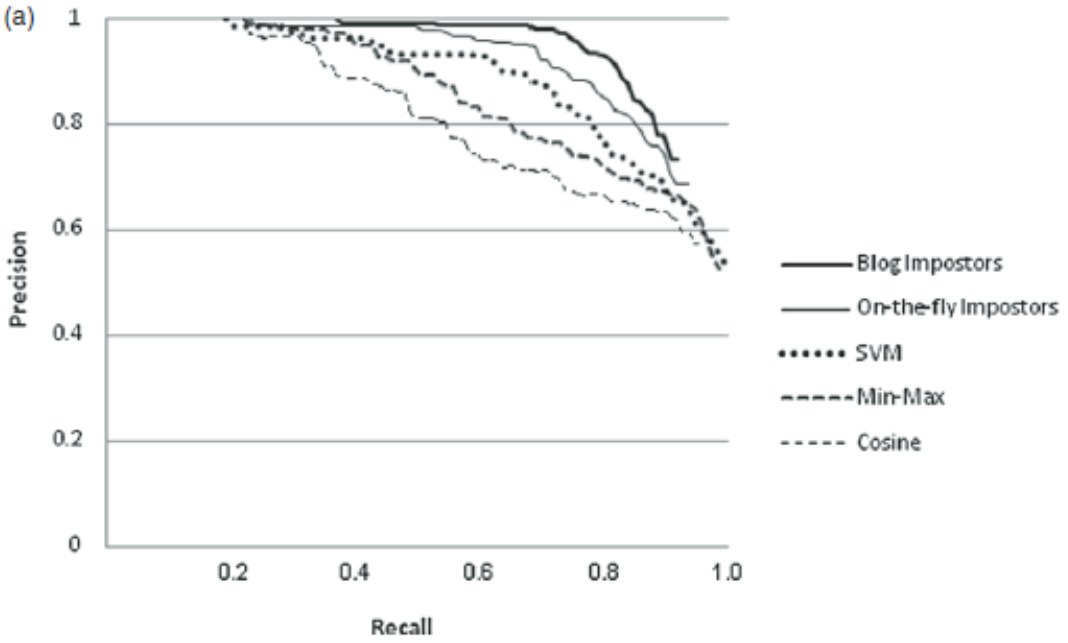


Experiments: Blogging

- Generate imposters: take the m most similar docs and randomly select n from these [potential vs. actual]
- Experiments with cosine & minmax thresholds, SVM (after training), imposters using both On-the-fly and Blog

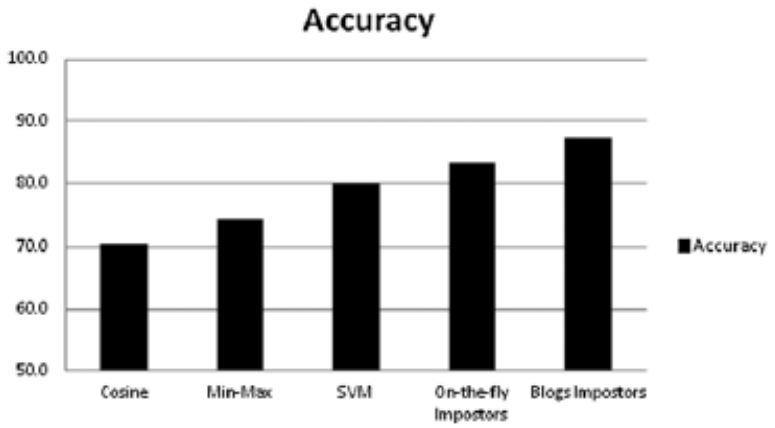
- Detecting co-authors
- In Blog method
  - Where Precision=0.9
  - Recall=0.83
  - σ^* = 0.13

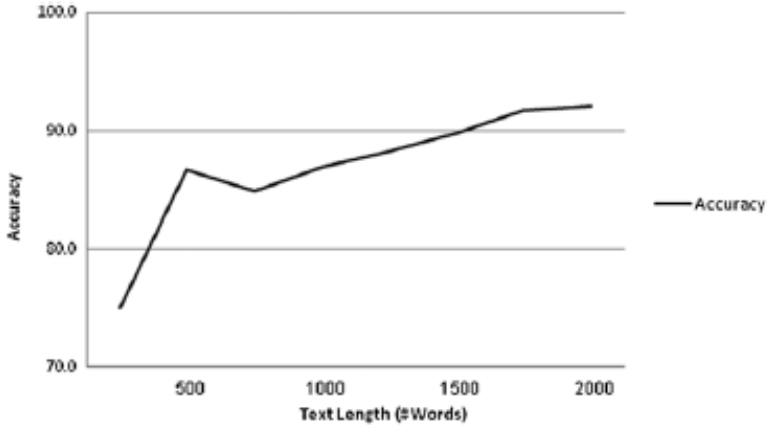
- Diff. auth. lower scores



Accuracy co-author detection

- Most pairs $< X, Y >$ not co-authored, so scores are high
- Average over σ^*
- Accuracy improves as longer texts are chosen
- Little sensitivity to number of potential/actual imposters



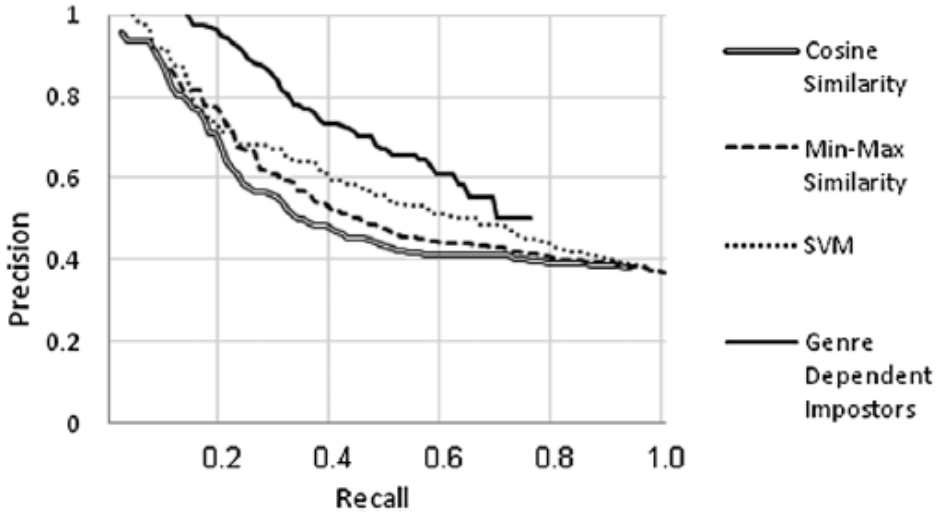


Blog imposters

Experiment 2: Plagiarism detection

- Similar to authorship verification, but not identical
  - Intentional distortions of authorship signal could make this different
- 4 essays each from 950 students, initial 500 wd. used
- 4 diff. topics, but all $< X, Y >$ pairs had diff. topics
- 2000 $< X, Y >$ pairs, imposters from same-topic essays

- Harder problem
- Imposters still best!



Impressive application

- *Wilhelmus* is Dutch national anthem
  - Anonymously written ca. 1570 (early in 80-yr war)
  - Popular since 18th cent., banned by Napoleon
  - Official national anthem since 1932
  - Usually attributed to Marnix of Antwerp
  - Very brief, only 15 couplets

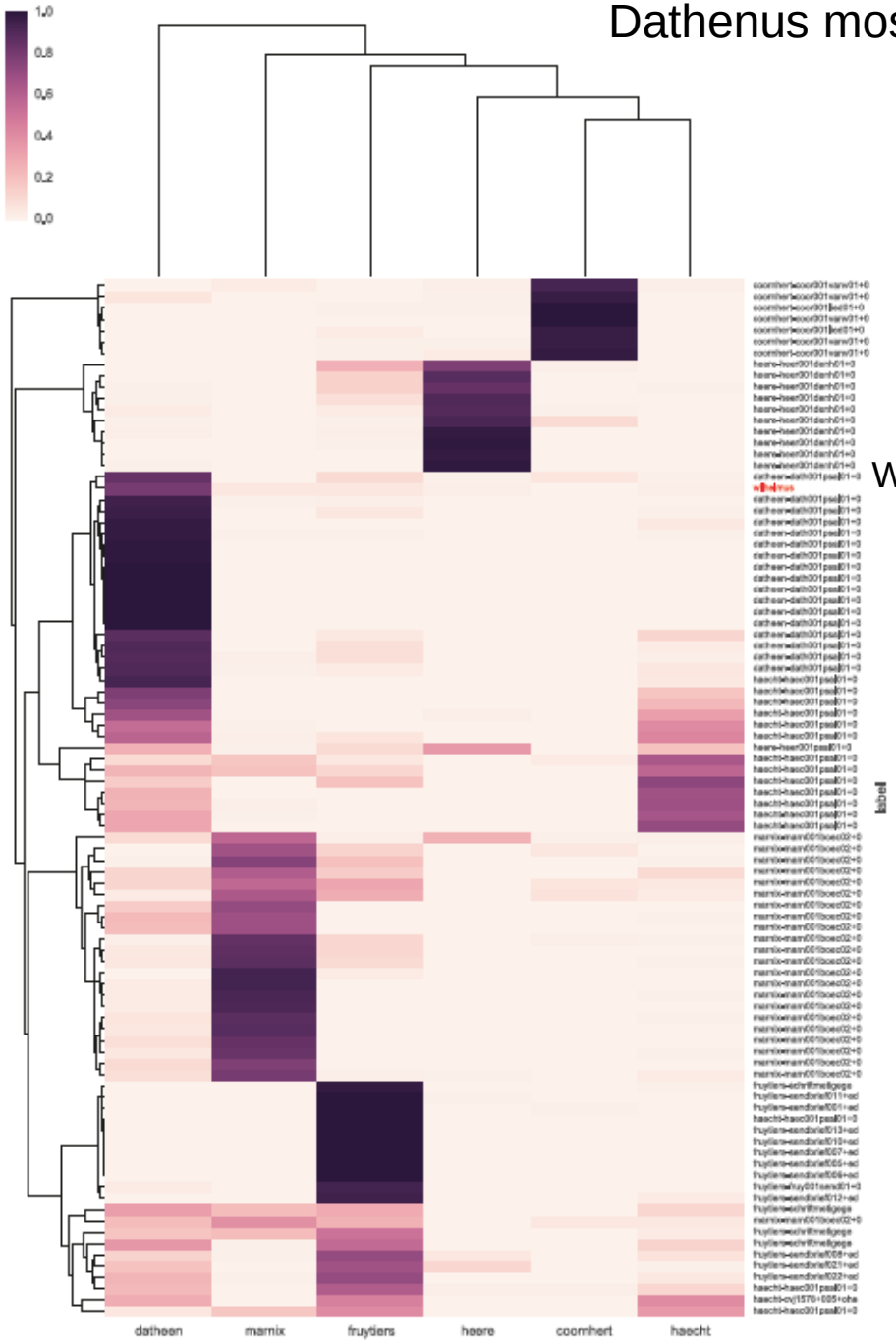
- Examined stylometrically
  - Kestemont, M., Stronks, E., De Bruin, M., & De Winkel, T. (2017). *Van wie is het Wilhelmus?: de auteur van het Nederlandse volkslied met de computer onderzocht*. Amsterdam University Press/ICAS Pubs.

Stylometry applied

- Added PoS tags to MFW, he-PRO, can-ModVb, …
- In addition to Marnix, Kestemont et al. examine 5 other contemporaries using usual stylometric methods
- "Imposters" method used
- Clear results point to Dathenus, never earlier a candidate
  - Best known for translating the *Psalms*
  - Referred to as "donkey-eared", for his poor poetry
- Verification solved a mystery, since Dathenus was present at the siege of Chartres, where the music originated

Dathenus most distinct!

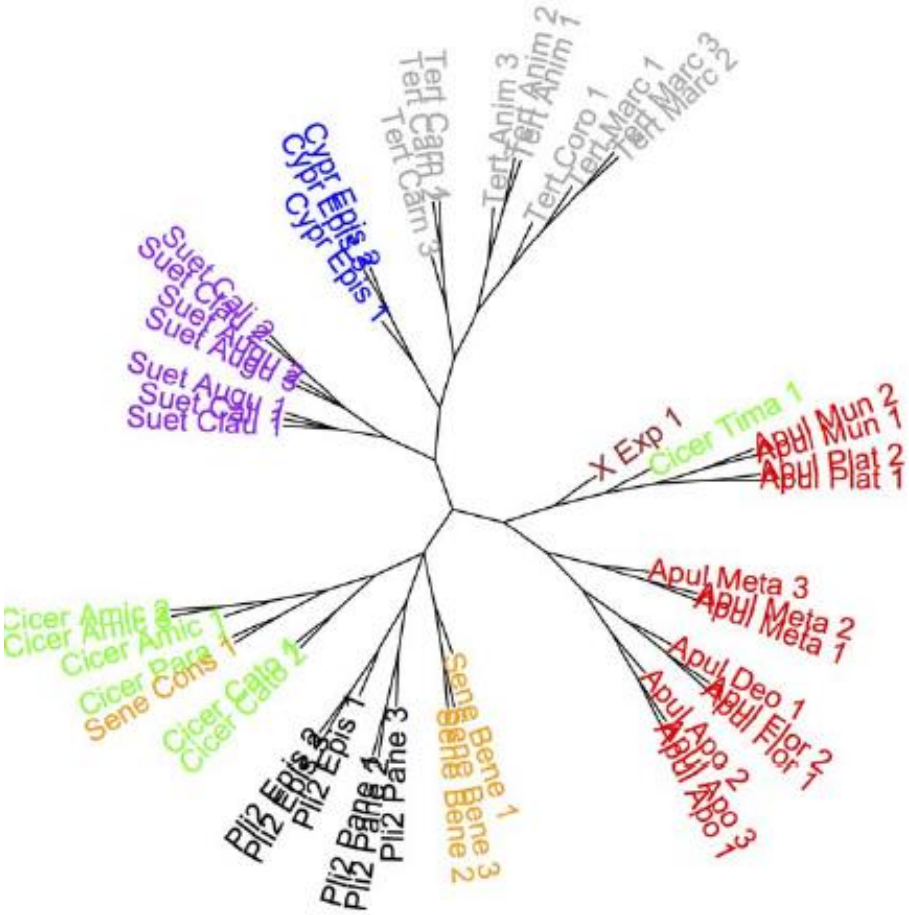
clcg



Wilhelmus, note no other plausible candidate

Application to new ms.

- Newly discovered anonymous ms. in Vatican Library, *Compendiosa expositio,* discussion of Plato's works
- Stylometry, "imposters"
- Apuleius of Madauros (today Algeria) singled out
- Stover, J. A., et al. (2016) Computational auth. verification […] attributes new work to 2nd century African author. *J. Assoc. Inf. Sci. & Tech. 67*(1), 239-242.



Summing up

- New efforts to overcome problems of limited scope are underway and promising
- They are shedding light on the applied problems of identifying blog authors and plagiarism, but also on authorship in classical (Apuleius) and early modern times (Wilhelmus)
- Next
  - Stylo Exercise
  - Bayesian foundations
  - Information theory (sometime)
  - Other views of identifying typical words