Who Wrote This Document?

Authorship Attribution by Computer

Charles Nicholas

Department of Computer Science and Electrical Engineering

Revised March 24, 2014

UMBC

AN HONORS UNIVERSITY IN MARYLAND

Summary

- Authorship questions are fascinating, but often complicated
- Linguistic or stylistic clues have been used for a long time
- Statistical and computer-based methods are now available
- Many questions remain!

Who cares?

- After all, documents usually list their authors
- But sometimes they don't
- And sometimes they don't tell the whole truth!

Example:

- The novel "Primary Colors" was in fact written by *Newsweek* columnist Joe Klein

- Professor Don Foster of Vassar College figured this out, and wrote his own book!
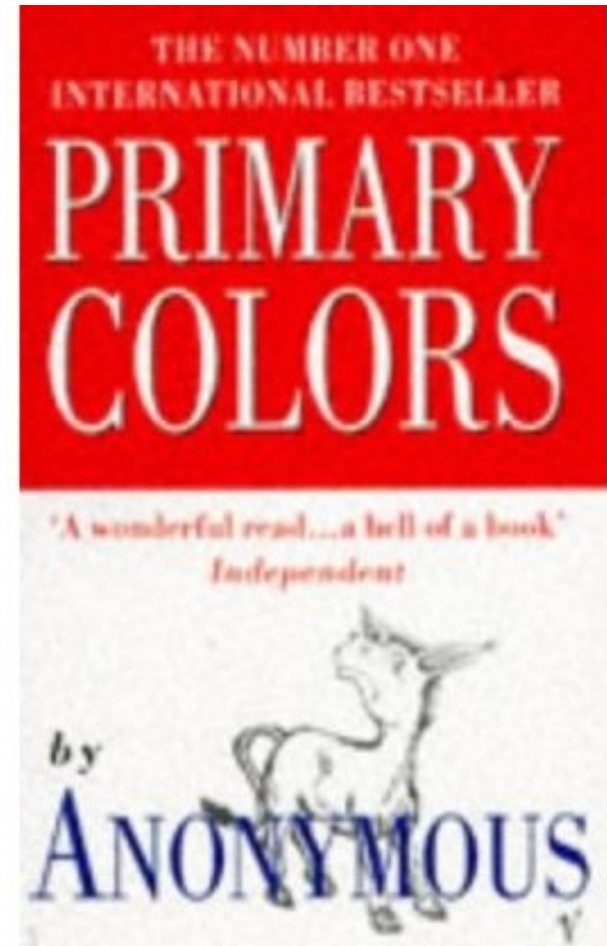
Foster Looks for Clues:

- Words and phrases repeatedly used
- Quirky expressions
- Patterns of punctuation
- Use of quotations
- Foster used on-line databases, but his methods were otherwise *not* automated

Author Unknown
On the Trail of Anonymous

DON FOSTER
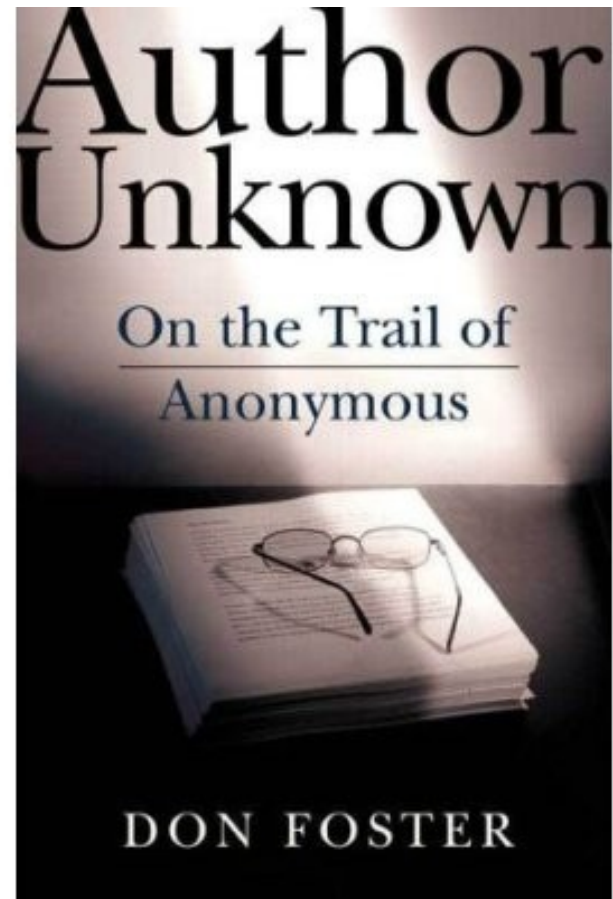
Lincoln's Letter to Mrs. Bixby

- Mrs. Bixby was thought to have lost f ve sons in the Civil War

- But maybe Lincoln didn't write this letter!

Not So Recent Examples

- The works of Shakespeare
  - Some plays seem to have more than one author!
- From the Christian New Testament
  - Who wrote the Letter to the Hebrews? The letter itself doesn't say!

How can we tell?

- Given a document, what forms of evidence can we use?
  - Knowledge of people, events or demonstrably earlier documents help us date documents
  - Linguistic evidence, such as vocabulary
  - Statistical evidence, such as consistency with other documents known to be by that author

Vocabulary

- In the Gospel of Mark, the Greek word *euthos* ("immediately") is used much more than in the rest of the NT

- More often than random chance would expect! $\chi^2=172$, signif cant at p<0.001

	Mark	rest of NT
	40	42
other words	11591	128640

One term or many?

- The frequency of a single term may be suff cient to suggest that document X was written by person Y, as in Mark's use of *euthos*

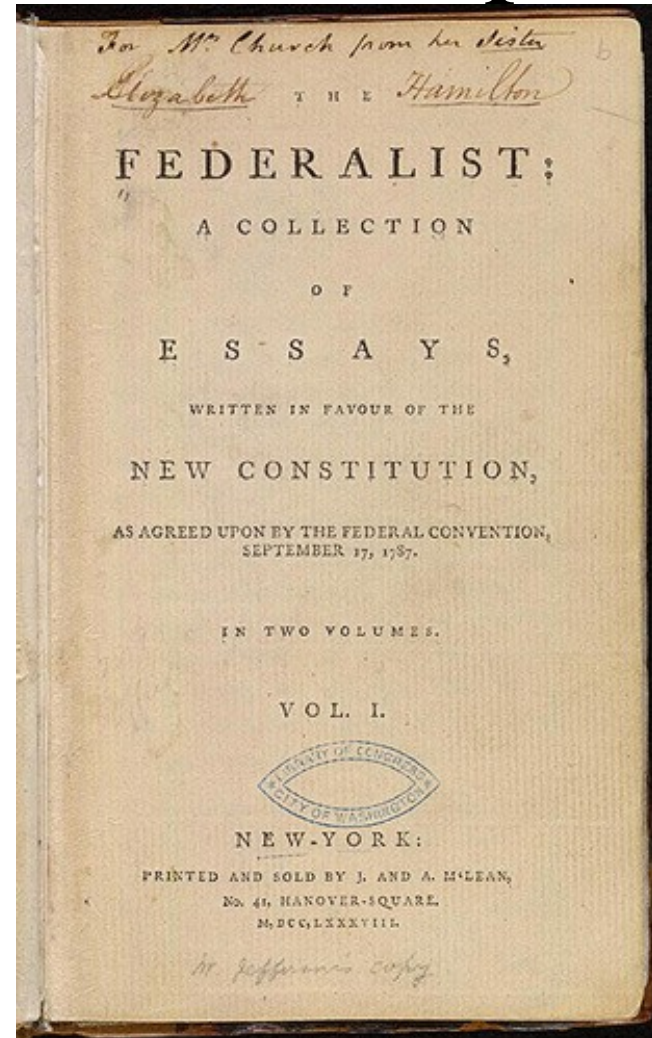- But the use of many terms is likely to be more convincing

Function Words

- Function words appear in most if not all documents written in a given language, regardless of topic

- Also known as "stop words" in Information Retrieval (IR)

- Since usage is independent of topic, patterns are likely to indicate authorship as opposed to other characteristics

Function Words Tell Us…

- **Inference and Disputed Authorship,** Mosteller and Wallace, 1964
- Using the Federalist papers as example, demonstrated how frequencies of function words can shed light on authorship questions.

Example: The Federalist Papers

- 85 essays written by James Madison, Alexander Hamilton, and John Jay under the pseudonym "Publius"
- Authorship of 11 has been disputed

FEDERALIST;

A COLLECTION

OF

E S S A Y S,

WRITTEN IN FAVOUR OF THE

NEW CONSTITUTION,

AS AGREED UPON BY THE FEDERAL CONVENTION,
SEPTEMBER 17, 1787.

IN TWO VOLUMES.

VOL. I.

NEW-YORK:

PRINTED AND SOLD BY J. AND A. M'LEAN,
No. 41, HANOVER-SQUARE.
M,DCC,LXXXVIII.

Hamilton appears on the $10 bill

Hamilton appears on the $10 bill

Madison appears on the $5000 bill

Function Words in the Federalist Papers

- Hamilton uses the word "upon" much more often than Madison

- Hamilton uses "while" (in the sense of "at the same time as") but Madison uses the (chief y British) "whilst"

- The disputed papers never use "while", and use "upon" and "whilst" in the same proportion as Madison

Matrix Methods Emerge

- Frequencies of these function words that distinguish one author from another can be analyzed using statistical tests, chi-square for example

- Methods such as singular value decompostion (SVD) and principal components analysis (PCA) can f nd combinations of terms with such distinguishing power

- Basic data structure is the Term-Document Matrix

Term-Document Matrix

- Create a matrix A, such that entry $a_{i,j}$ is the number of times term i occurs in document j
  - Terms can be words or n-grams
  - N-grams are best for noisy and/or multi-lingual
- The TDM is usually sparse; term weighting makes it more so
- Using function words reduces the rank of the TDM

Kjell and Frieder on the *FPs*

- Kjell and Frieder chose a set of 10 n-grams that most distinguished the sets of documents with known authorship in a training set
- Two clusters emerged in that term-document matrix, indicating Madisonian authorship of the eleven disputed Federalist Papers
- They used the KL-transform to reduce 10 dimensions to 2
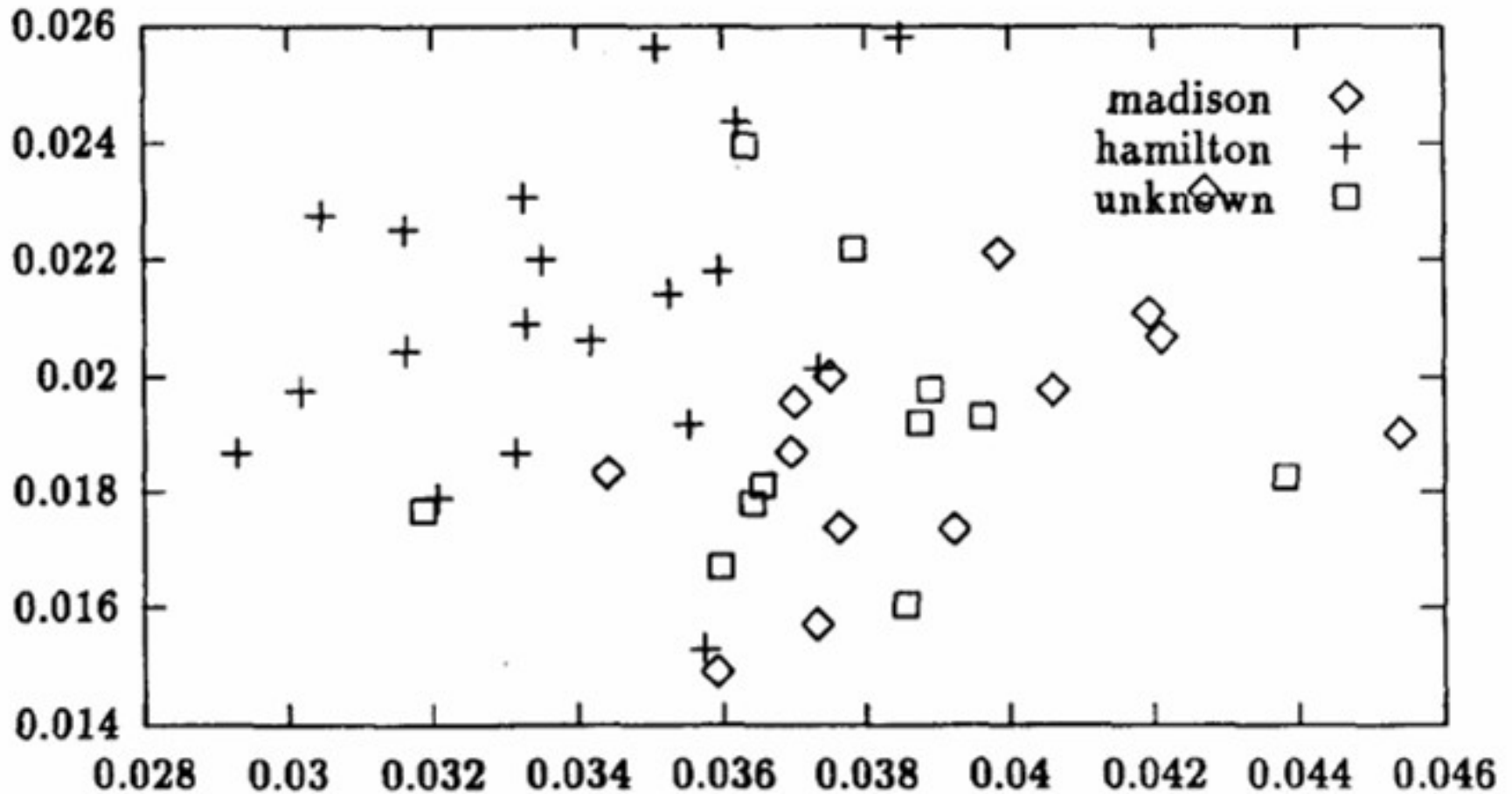
Kjell and Frieder's Findings

Figure 8: Plot of transformed feature vectors.

Observations on Kjell and Frieder

- The disputed documents are *mostly* in the Madison region, agreeing with other recent scholarship including Mosteller and Wallace
- Kjell and Frieder used a modest amount of data, i.e. the top ten most distinctive 2-grams
- Their analysis was computationally expensive at the time, but nowadays we have other options

15th book of Oz

- L. Frank Baum created the Wizard of Oz books, and wrote the f rst 14

- Ruth Plumly Thompson wrote installments 16-31

- The authorship of the 15th book was unclear
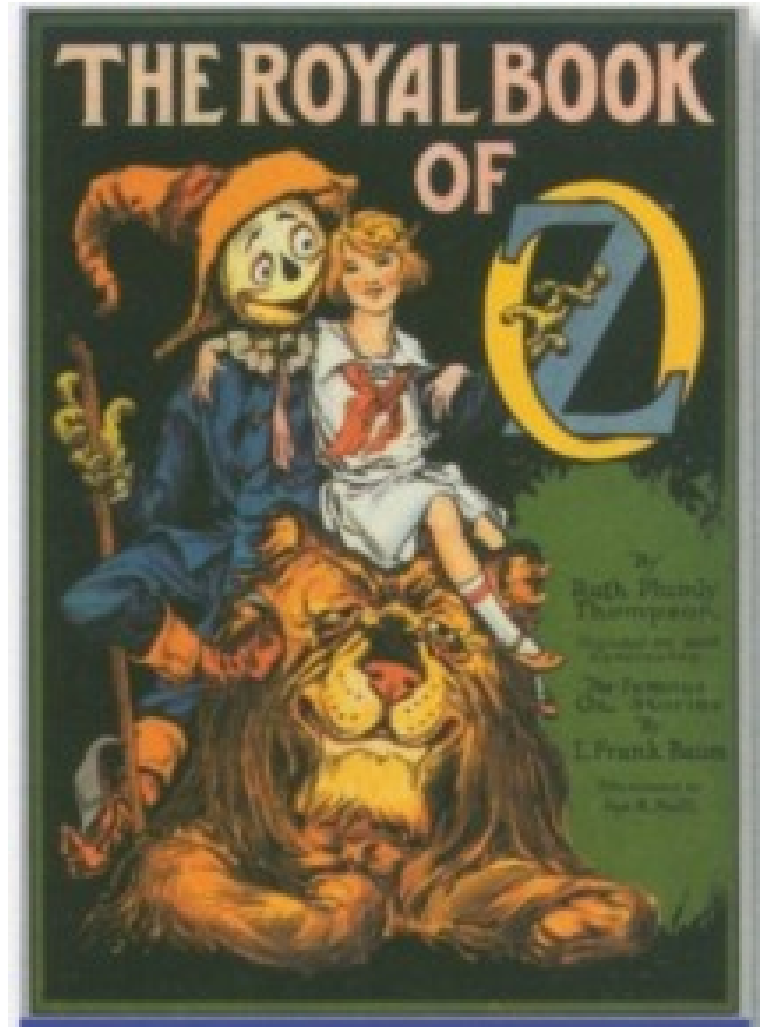
Binongo's use of PCA

- José Binongo took the whole Oz corpus, and built a term-document matrix using 223 text segments (documents) and 50 function words as terms

- The resulting matrix was subjected to PCA

- Plotting the data on the space spanned by the f rst two principal components

Thompson wrote the 15th volume

2nd PC

1st PC

- ● Baum
- ○ Thompson
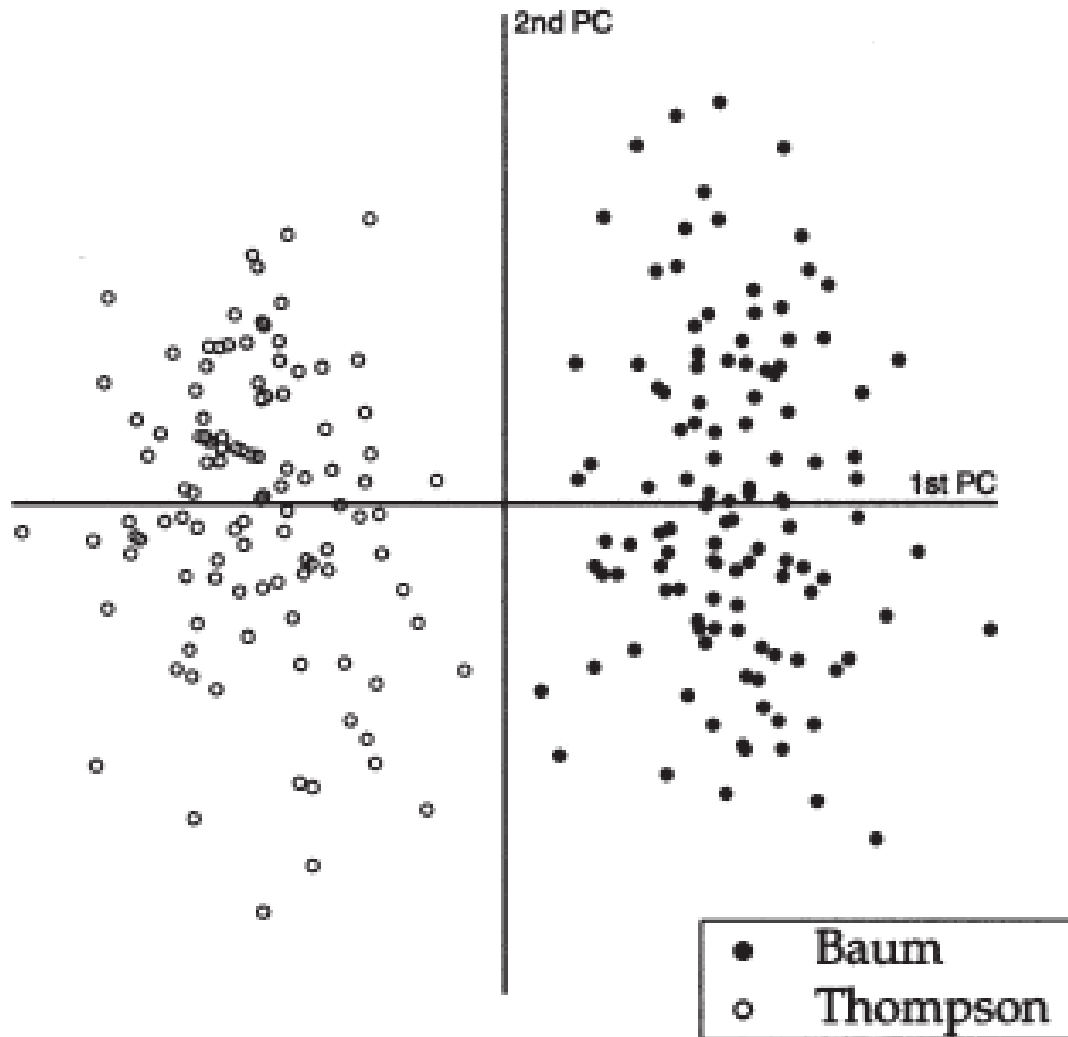
Can we spot other characteristics (besides authorship)?

- Soboroff and Nicholas looked at language, genre, and authorship as well as topic
- The SVD identifes patterns in the term document matrix, but the patterns still need interpretation
- Differences in language or dialect really stand out
- Examples from the Hebrew Bible

Singular Value Decomposition

- The SVD is an alternative to Principal Components Analysis
  - Easier to calculate
  - Finds patterns of terms
- Basis for latent semantic analysis used in IR
- Patterns of terms become dimensions in a vector space

Properties of the SVD

- SVD calculates matrices U, Σ, and V^T such that the term document matrix $A = U \Sigma V^T$

- The matrices U and V are *orthonormal*, i.e. the columns form a basis, and each column is length 1

- Complexity of full SVD is $O(n^3)$ for n non-zero entries in the matrix, so sparse is good

Interpreting U, Σ, and Vᵀ

- The columns of U are sets (or patterns) of terms that occur (or not) together.

- The *singular values* are the main diagonal entries in Σ, and they give the relative importance of these patterns

- Entries in the rows of Vᵀ are the coordinates of the documents in the space spanned by the columns of U

Ezra, Nehemiah, I and II Chronicles

- Attributed, by tradition, to Ezra
- We built a term-document matrix in which each chapter was a document, and Hebrew 3-grams were tabulated
- The SVD was calculated, and the f rst dimension (i.e. the X axis) was dominated by Hebrew function words
- So we projected the documents (chapters) onto the Y-Z plane

Ezra-Nehemia-Chronicles, 3-grams, Dimension 2+3
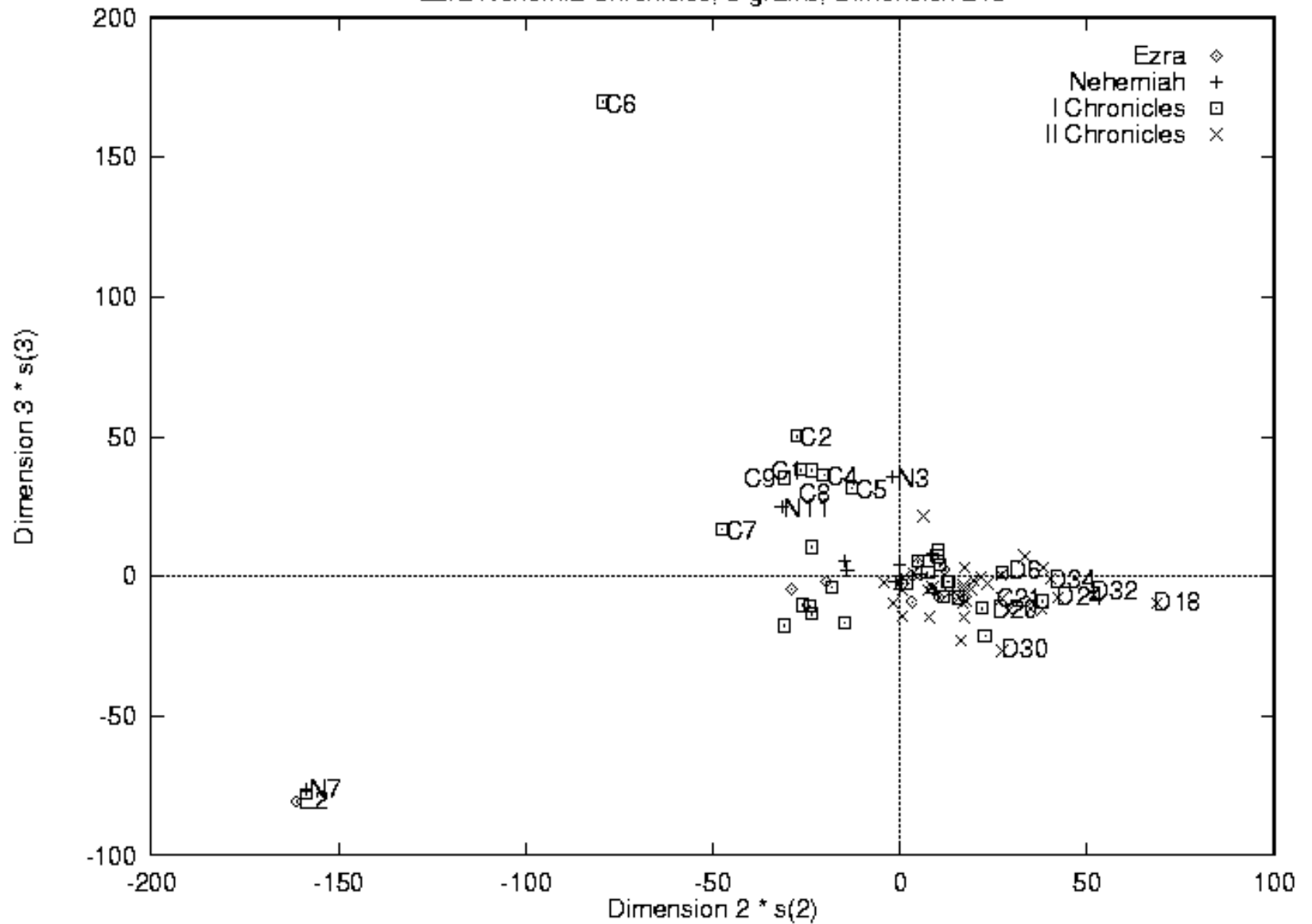
What does this graph say?

- Some chapters, such as Nehemiah 7 and Ezra 2, are different from the rest
  - Most of the text is narrative
  - Ezra 2 is a census, as is Nehemiah 7
- This plot is consistent with the (traditional) hypothesis that these books were written by the same person

Ecclesiastes, Song of Songs, and Daniel

- Ecclesiastes and Song of Songs are traditionally attributed to Solomon, and are poetic in nature

- Daniel dates from much later, and is more narrative (and apocalyptic) in nature

- Modern visualization tools let us squeeze multiple dimensions into a single image

Solomon texts and Daniel, 3-grams

o = Ecclesiastes
* = Song of Songs
+ = Daniel

Dimension 2 ' ϕ_2
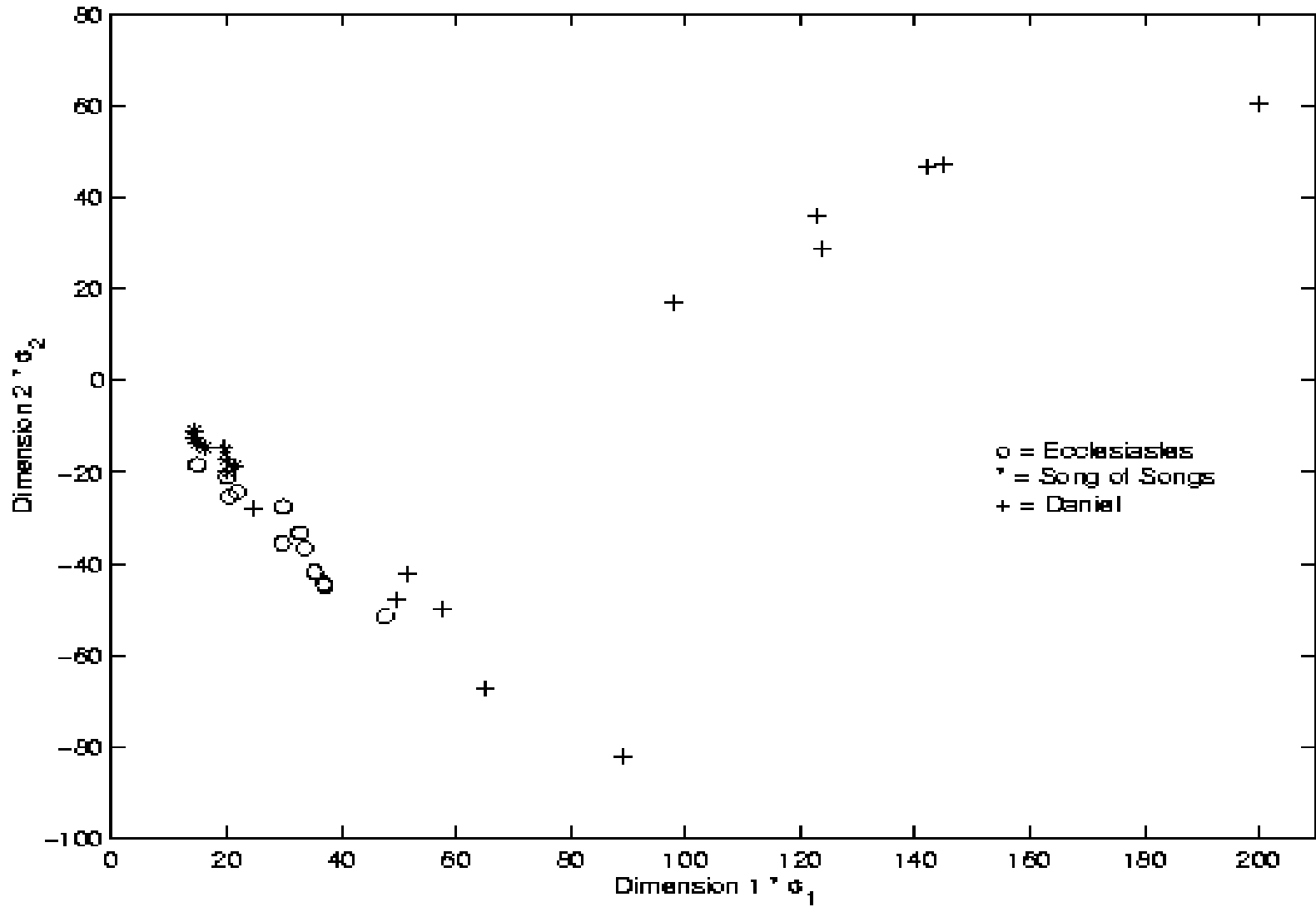
Dimension 1 ' ϕ_1

What does this graph say?

- Song of Songs and Ecclesiastes are clustered together, consistent with their poetic nature (and/or Solomonic authorship!)

- Chapters 2-7 of Daniel are in Aramaic!

- Choosing which dimension(s) to look at can be important!

Was there one Isaiah or more?


Isaiah, 3-grams, Dimension 3
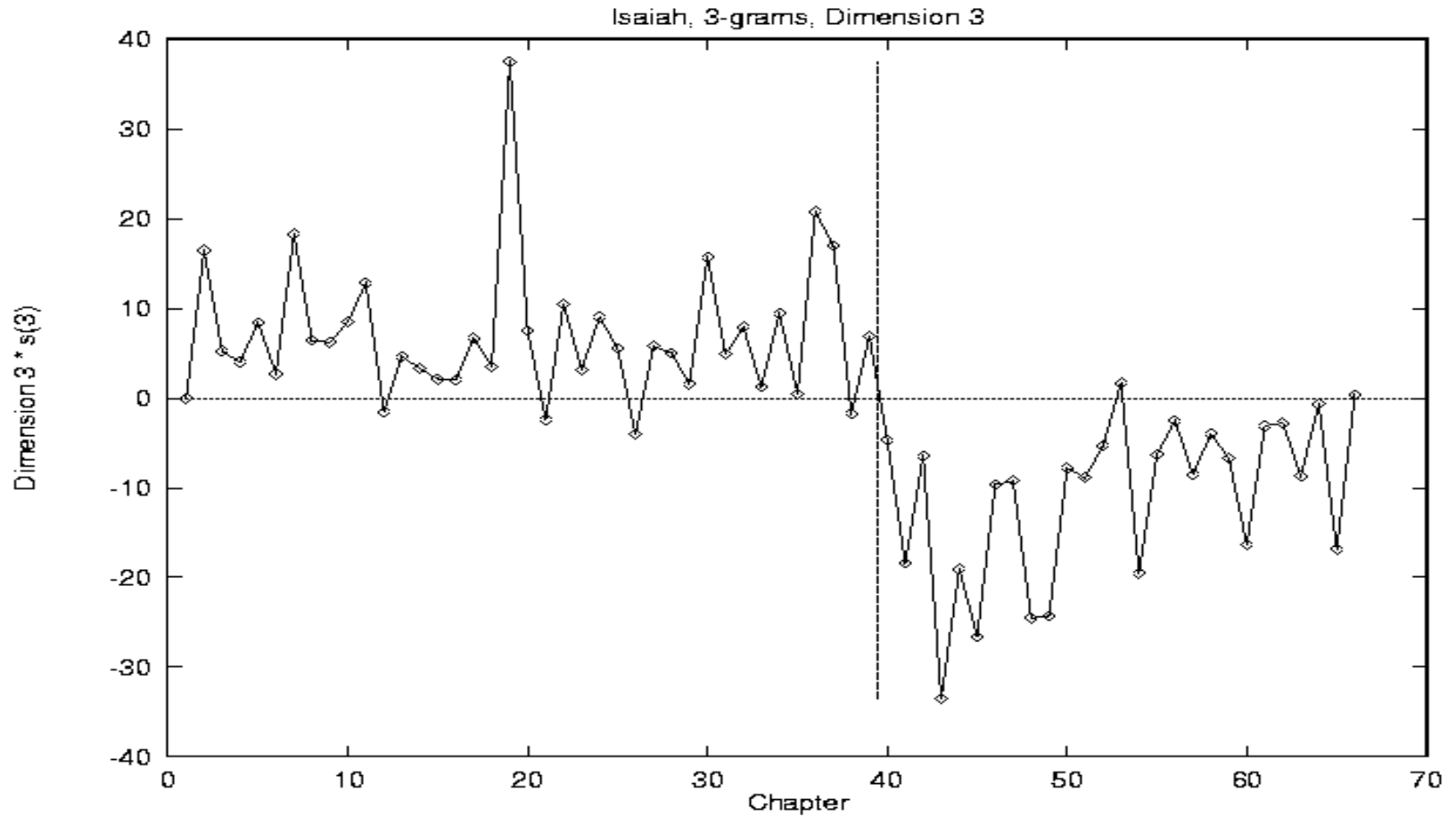
Dimensions of Isaiah

- In a monolingual corpus, the f rst dimension generated by the SVD will be dominated by function words

- The other dimensions can be inspected to see which terms are occurring together, or not, and in what proportion

- Some "new" pattern starts in Isaiah 40

Visualizing the New Testament

- The "synoptic problem" refers to the relationship between Matt, Mark, and Luke
- We can build a TDM of the most common words used in 1ˢᵗ Century CE Christian writing
- Kai ('and') is by far the most common term in the corpus, but its frequency of use varies signif cantly (anova F=23.3, p=0)
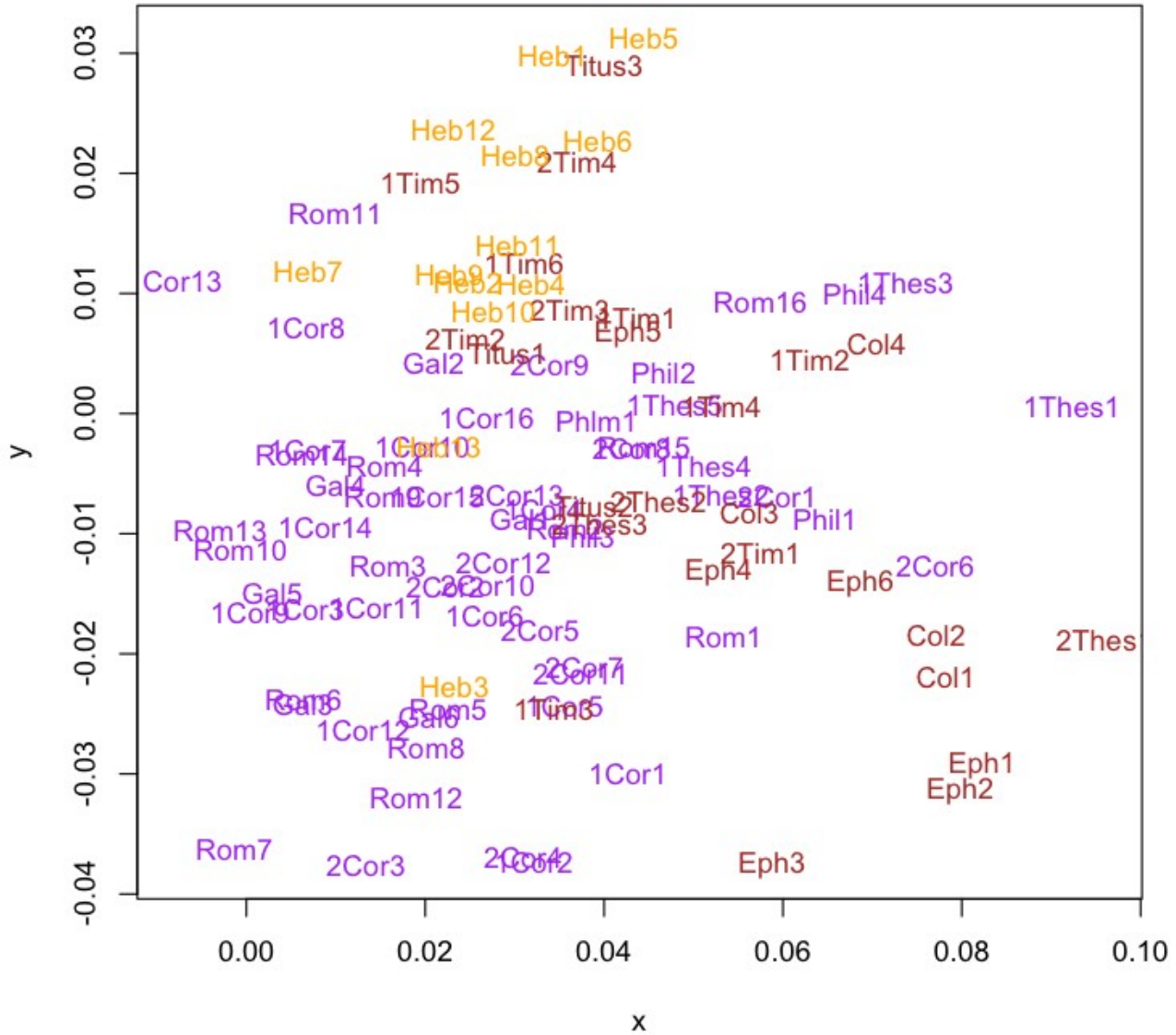
Usage of κα
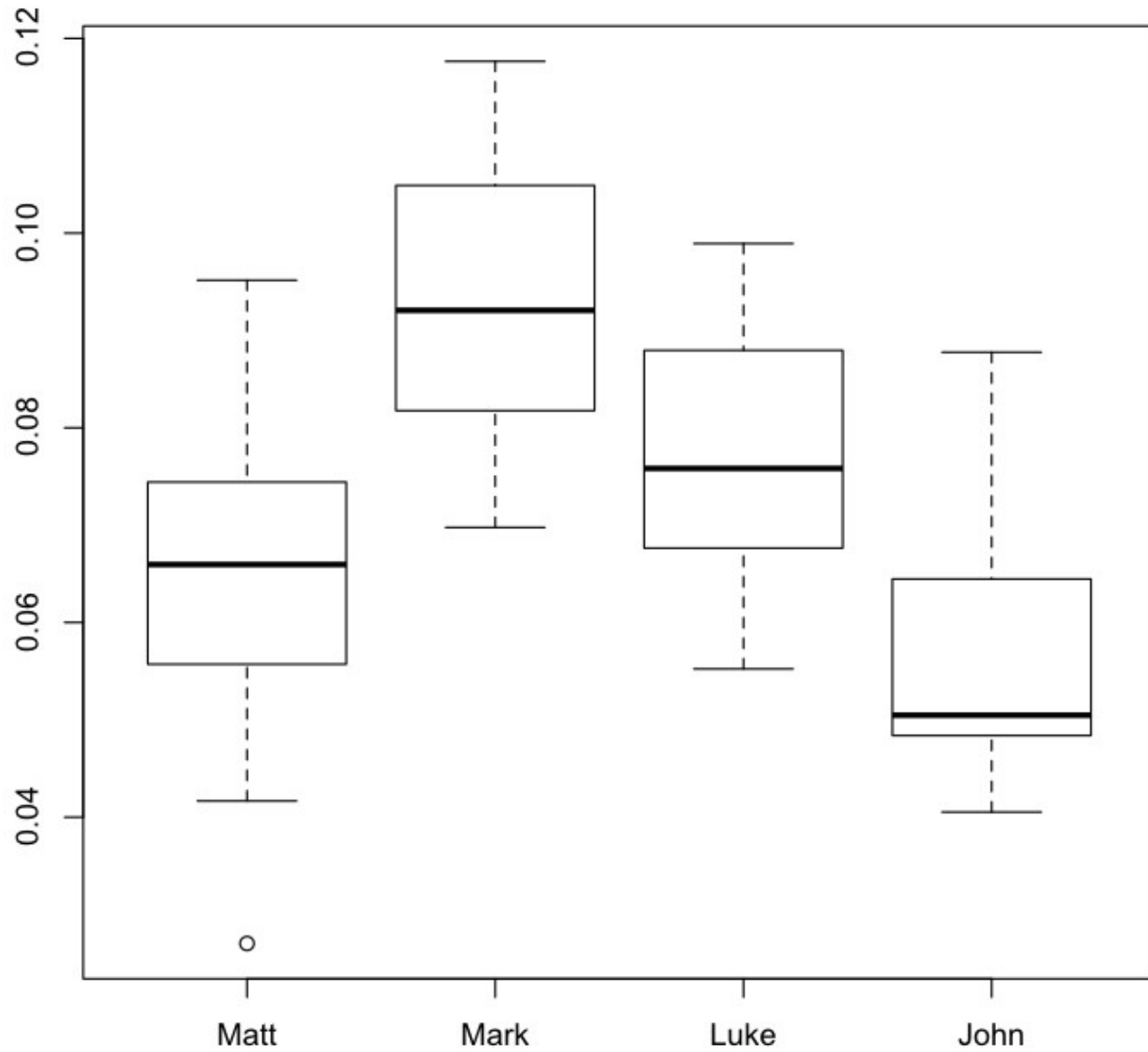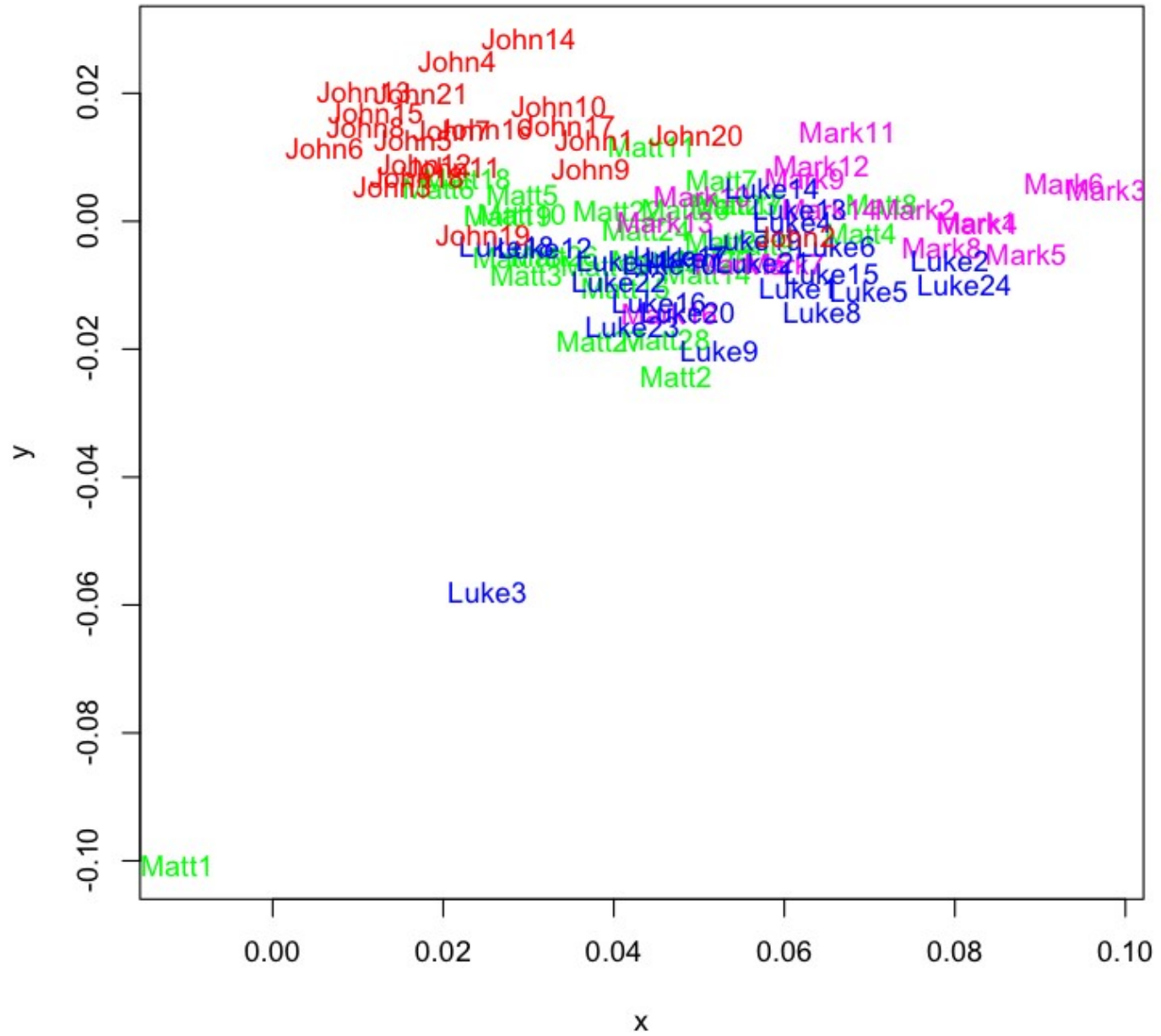
Paul, and Paul

- Several NT books are undoubtedly by Paul
  - Romans, 1&2 Cor, Gal, Phil, 1Thess, Phlm
- Some are attributed to Paul, but there's controversy
  - Eph, Col, 2 Thes, 1 Tim, 2Tim, Titus
- We don't know who wrote Hebrews, but Paul is one of several candidates

Limits of Existing Approaches

- Traditional methods of literary scholarship, based on history, language, or content, have limits
  - Patterns may defy easy description
  - Larger corpora are diff cult
- Statistical evidence needs to be interpreted in light of human understanding of language and history

Research Questions

- Some questions which apply to authorship study:
  - How can we represent features of an author's rhetorical style, as opposed to just vocabulary?
    - e.g. Markan "sandwich"
  - How can we represent what an author knows?
    - e.g. Judges' reference to the (then future) monarchy "In those days Israel had no king, and everybody did as they pleased."

More Research Issues

- How to deal with authorship in large corpora
  - Can we build a search engine that f nds documents with vocabulary or writing style similar to a given "query document"?

- How to represent more complicated features
  - Could a search engine f nd documents that mention f rst century CE people or events, but not second century?

Zoom back to the Present Day: Malware Analysis

- Can we use techniques like these to figure out who wrote a malware specimen, such as CryptoLocker?

- People are looking at such questions, but so far no easy answers

- We can compare malware specimens, though, using compression. (How?)

Work in Progress

- Can we use compression-based similarity to compare malware specimens? Yes

- But isn't compression kind of slow? Yes

- Can we cluster small malware collections anyway? Yes

- Will we have more to say later this year? Yes

"I think I've made one of the first steps toward unraveling the mysteries of the Old Testament. . . . I'm starting to read it!"

Selected References

- <u>Applied Bayesian and Classical Inference: The Case of *The Federalist* Papers,</u> Frederick Mosteller and David L. Wallace, Springer-Verlag 1984

- http://www.foundingfathers.info/federalistpapers/

- <u>Who Wrote the Bible?</u>, Richard Friedman, HarperSanFrancisco, 1997

- Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution, Jose Nilo G. Binongo, Chance 16(2) Spring 2003

More References

- <u>Statistics for Corpus Linguistics</u>, Michael Oakes, Edinburgh, esp. Chapter 5, Literary Detective Work
- Analyzing Worms and Network Traffic Using Compression, Stephanie Wehner, J. Comp. Security, 15(3), 2007, 303-320.

Still More References

- An article on the authenticity of Lincoln's letter to Mrs. Bixby appeared in the January 2006 issue of American Heritage
- Charles M. Schulz, The Complete Peanuts, 1950-1952, Fantagraphics Books, 2004, p. 329

Additonal Slides

The Matrix Approach

- Select subset of document terms to be considered (all words, n-grams, function words, or whatever)

- Build a term-document matrix

- Transform as needed to make any patterns visible

- Figure out what the patterns mean!

Dyadic Decomposition

- We can choose how much of the SVD to do

- For some k >= 1, we can calculate the rank k matrix $A_k \sim U_k \Sigma_k V_k^T$, where we compute only the first "k" of the singular values.

- The matrix A_k is the best (rank k) approximation to the original t-d matrix A.

- Choosing k=2 makes sense for a plot

Interpreting U

- Each column U_1, U_2, …, U_k of U represents a pattern of terms that tend to occur together

- Terms common to all documents collect into U_1

- A frequency plot can show these patterns of terms occurrence

- In an AP News corpus, of almost 100,000 terms, a relatively small number really stand out, thereby helping to characterize these term patterns
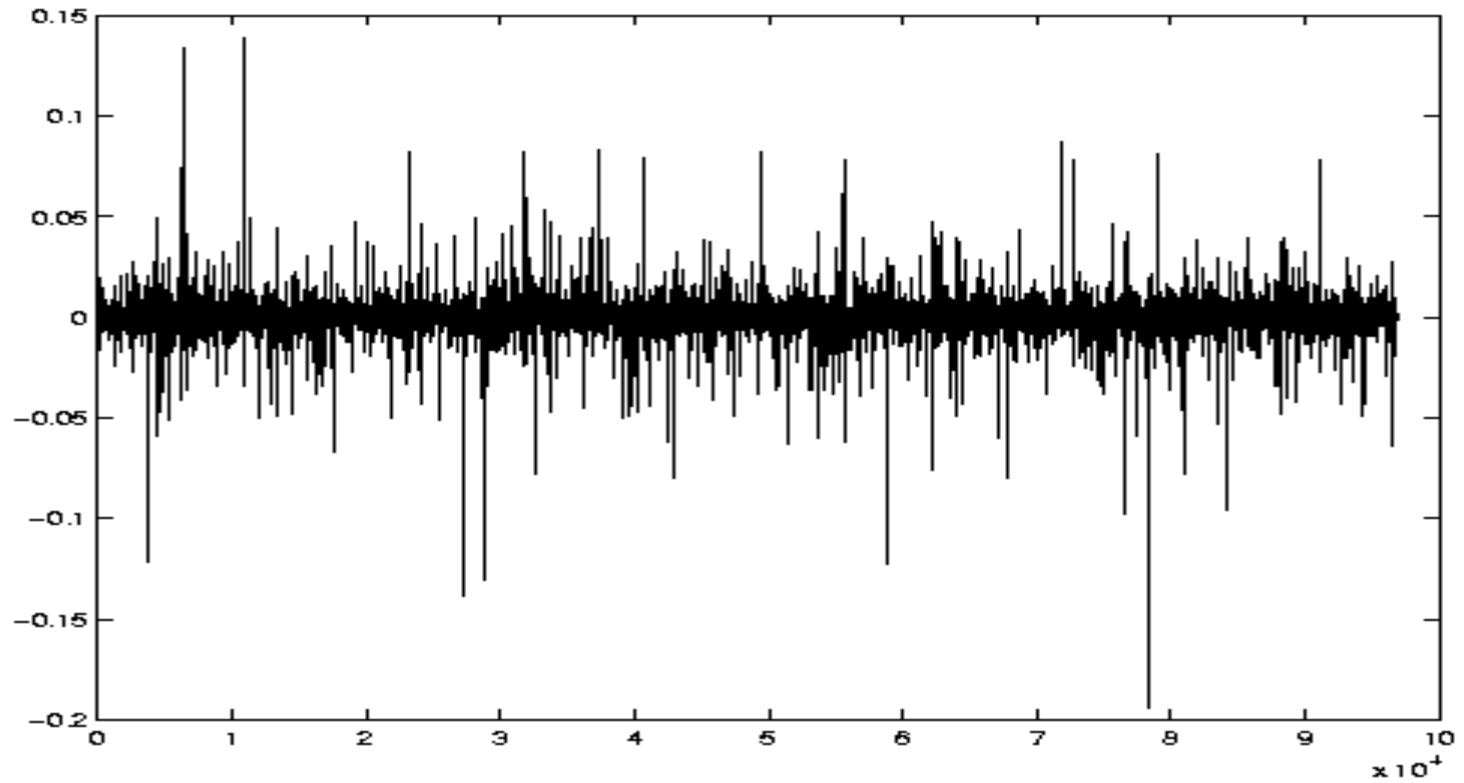
Interpreting V^T

- The columns of U form a basis, and the entries in row i of V^T are the coordinates of document i in the space spanned by the columns of U

- Documents that have large values in a certain dimension have many instances of the corresponding terms

# Example: Coordinates of documents in various dimensions
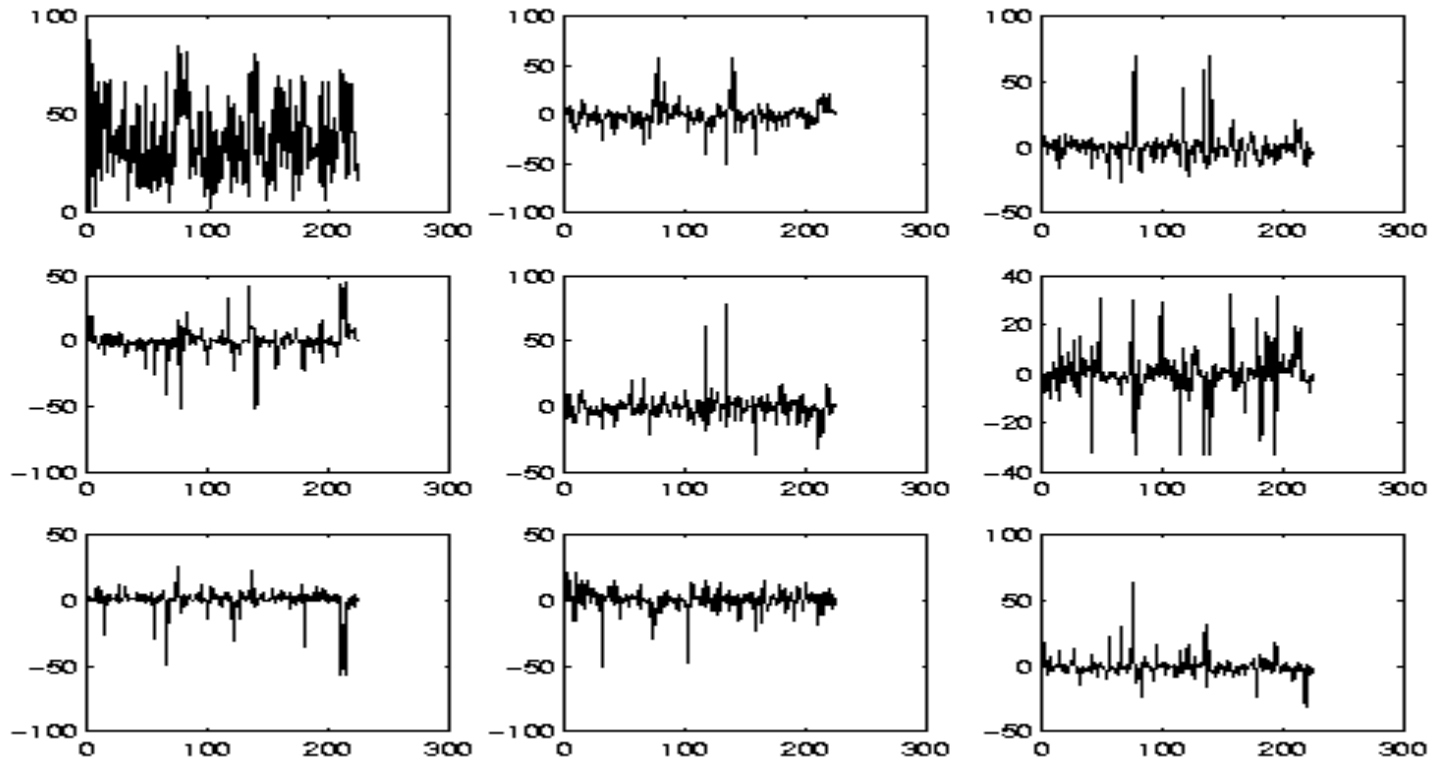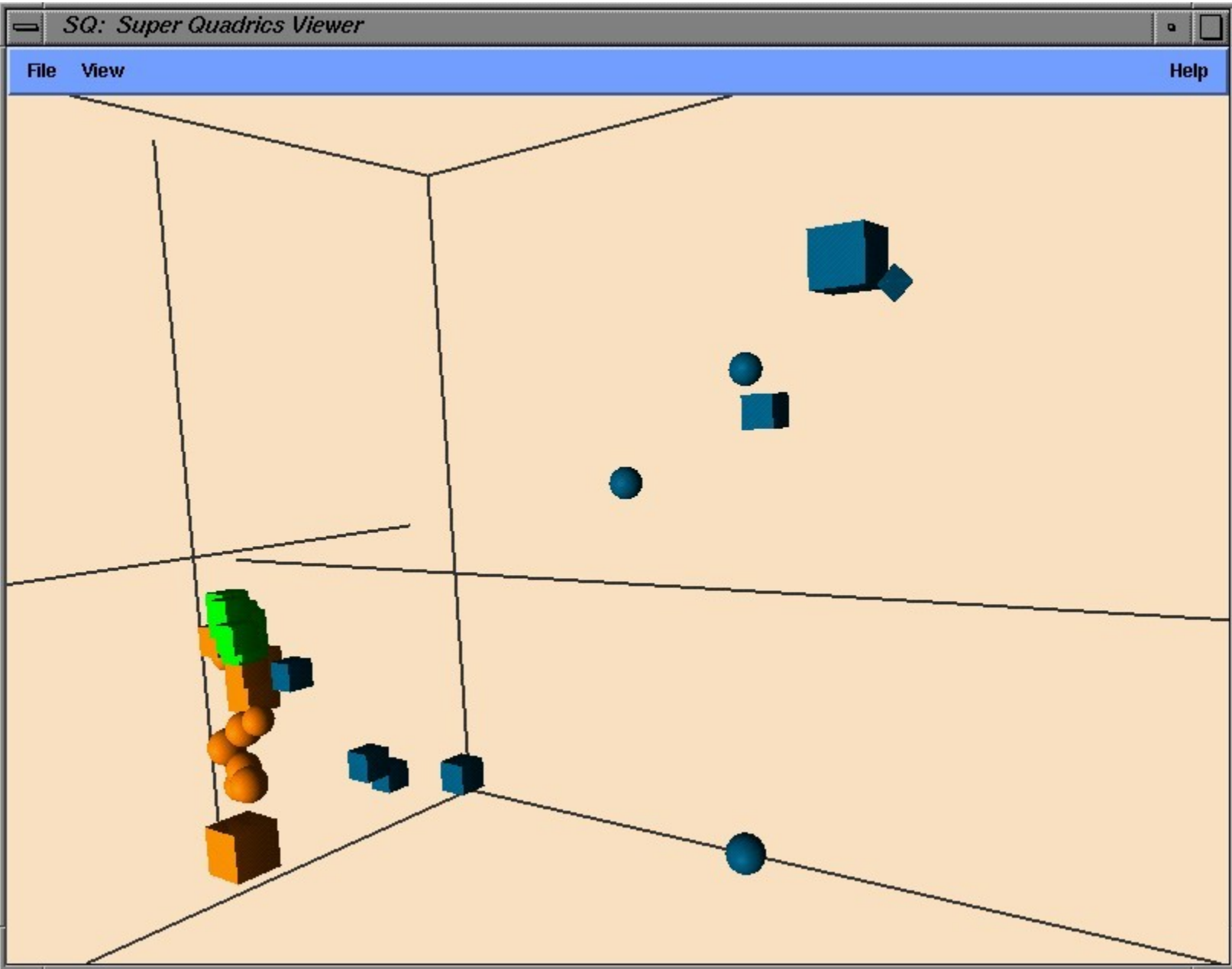
Example frequency distribution

The Entries in Σ

- The singular values are the squares of the eigenvalues of the matrix AAᵀ
- A plot of the singular values is revealing
  - a steep left/downward slope indicates a homogeneous corpus
  - a "jagged" left side indicates a heterogeneous (multi-lingual?) corpus

Example plot of singular values
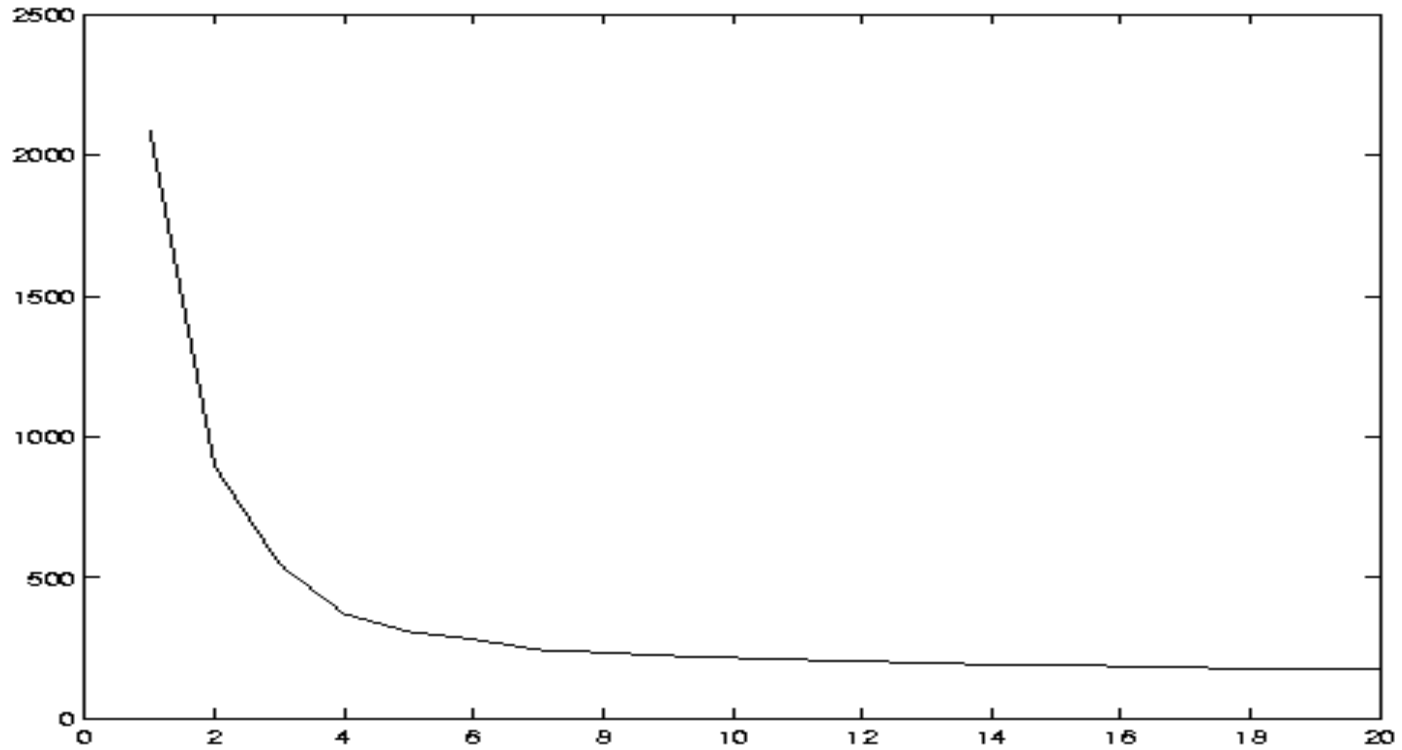
Authorship as Text Classif cation

- TC relies on features, such as where and how often a term appears
- Probabilistic (e.g. Naïve Bayes) or Information Theoretic (e.g. Maximum Entropy) models are used
- Usually assumes a reliable training corpus