The Colour of Magic by Numbers

Matthew Roughan

ARC CoE for Mathematical & Statistical Frontiers,
University of Adelaide, Australia.

matthew.roughan@adelaide.edu.au

THE UNIVERSITY of ADELAIDE

SUB CRUCE LUMEN

ACEMS

AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Distant Reading

- Some corpora of text are very large:
  - Discworld *novels*
    ~ 41 books, 4 million words
  - Australian WW1 diaries
    ~ 1000 documents, 15 million words
- We can’t just read them
  - You could read them, but by the time you finish, you’d need to start again
  - We can’t hold the Whole in our mind at once
- In distant reading we apply computational, statistical and mathematical techniques to gain insights

The Good, the Bad and the ...

- The Good
  - Distant reading is (more) objective
  - Distant reading allows us to access larger corpora
- The Bad
  - Algorithmic analysis of language is in its infancy
  - Close reading takes advantage of the human brain
- SO
  - The two approaches are complementary
    - One can generate questions for the other to answer
  - Distant reading should be seen as enabling

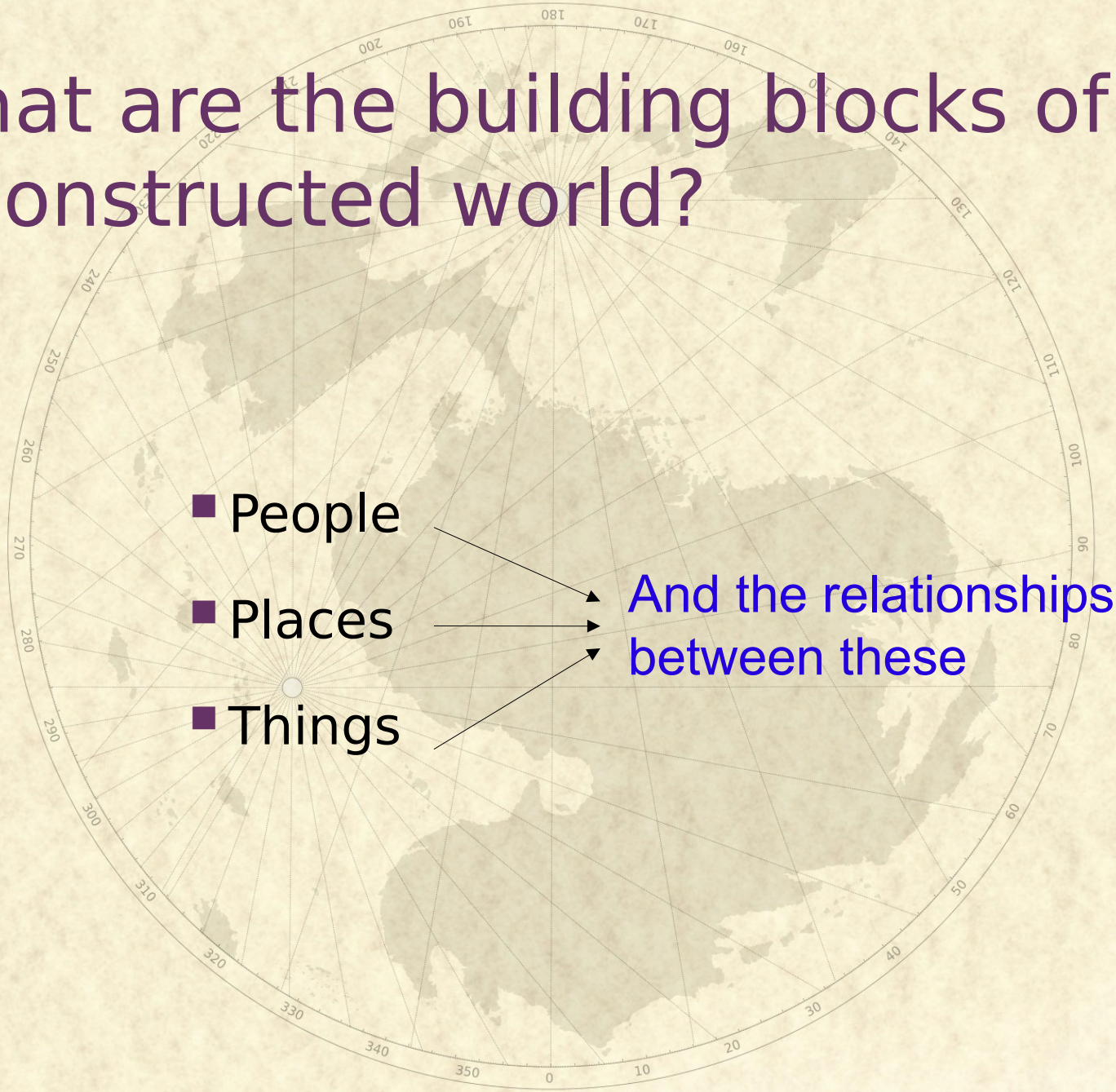
Natural Language Processing (NLP)

- This is a HUGE Meta-area
  - Quantitative Linguistics
  - Natural Language Understanding
  - ...
- Techniques
  - Sentiment analysis
    - Identifying archetypal story arcs
  - Topic modelling
    - Linking stories by theme
  - ...

Fantasy in the context of NLP

- Almost untouched
  - Some work on mythology
    - Icelandic Sagas
    - Irish Mythology
  - Fantasy and SciFi often neglected
- Interesting questions
  - Why are we drawn to “made up” contexts from the time we are children?
  - **Is it important that constructed worlds are realistic in some sense?**

What are the building blocks of a constructed world?



What are the building blocks of a constructed world?

- People
  - “Ordinary” people
  - Gods, demons, …
  - Important animals
- Places
  - Geopolitical entities: cities, nations, …
  - Geographic features
  - Man-made locales: buildings, roads, …
- Things
  - Languages
  - Organisations
  - Products
  - Works of art
  - Events

We call these NAMED ENTITIES

Named Entity Recognition (NER)

- NER has several components
  - Recognition – automatically extracting entities
  - Linking – combining “aliases”
  - Classification – into Person, Location, Artefact, …
- NER is a standard technique in NLP
  - Existing tools
    - spaCy: https://spacy.io/
    - NLTK: https://www.nltk.org/
  - They failed pretty badly :-(

Who Knew English was Complicated?

- Modern NER techniques are based on AI (Machine learning) and trained on large LABELLED corpora, e.g., from newsprint
- Fantasy corpora (and Pratchett's writing in particular) are **different**
  - Much more dialogue
  - Made up words:
    - Proper, e.g., Rincewind, or
    - Common, e.g., octarine, reannual
  - Common nouns used as names, e.g., I-Don't-Know Jack AND C.M.O.T. Dibbler
  - Words used in non-standard ways, e.g., puns such as "Equal Rites"
  - Names including weird punctuation, e.g., Lio!rt
  - Common nouns raised to proper status as concepts, e.g., Time
  - Non-standard text to represent foreign language: "?H 0 ryu latruin mr ii?"
  - Capitals used for effect: e.g., "AIR, Air, air" or for Golems speech
  - Frequent use of aliases, nicknames, sobriquets, epithets, abbreviations, misspellings (deliberate or otherwise)

Back to the Drawing Board

- I wrote my own code
  - Implemented in Julia (which is the cool new kid on the block)
  - Using an “old-fashioned” structural approach,

    e.g., look for capitals ...
  - Semi-supervised learning
  - Goal >95% classifications correct
- Status

Back to the Drawing Board

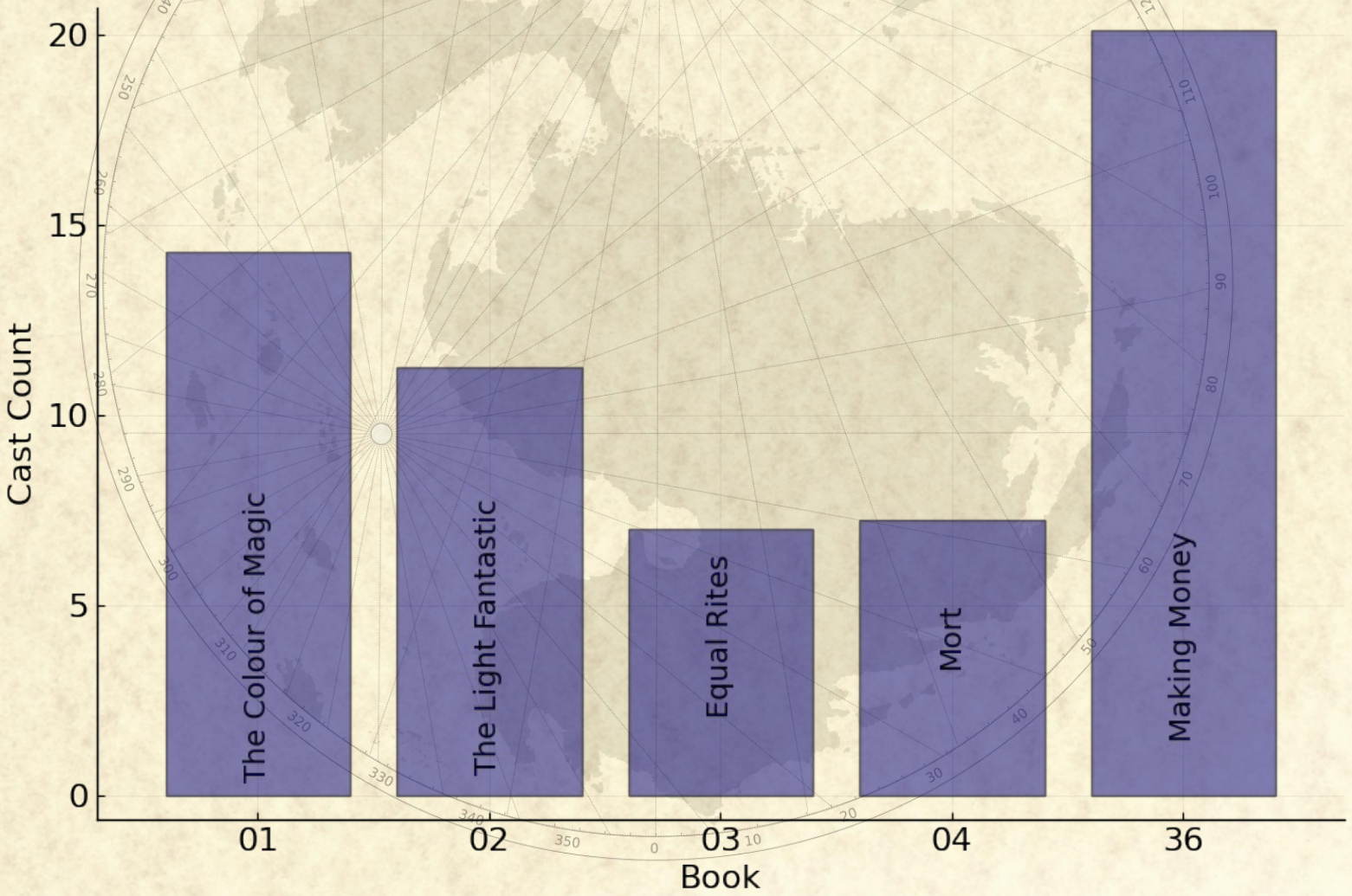
- I wrote my own code
  - Implemented in Julia (which is the cool new kid on the block)
  - Using an “old-fashioned” structural approach,

    e.g., look for capitals ...
  - Semi-supervised learning
  - Goal >95% classifications correct
- Status

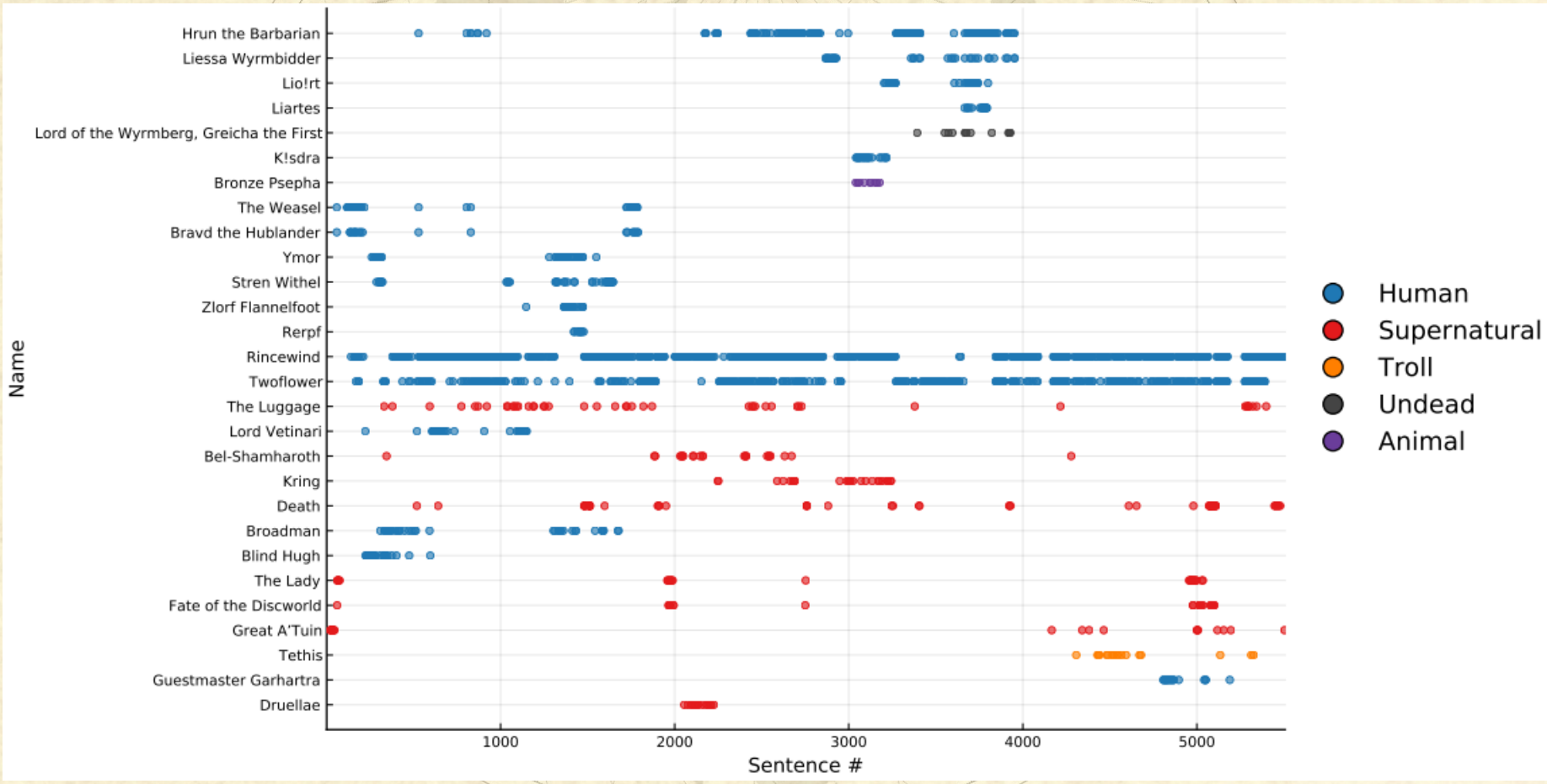
Book	Mentions	Persons	Locations	Artefacts
01 - The Colour of Magic	2474	70	62	37
02 - The Light Fantastic	1721	50	20	26
03 - Equal Rites	2044	37	20	27
04 - Mort	2272	42	45	24
36 - Making Money	4608	124	37	72

How the Avengers assemble: Ecological modelling of effective cast sizes for movies, Roughan et al., PLOS ONE https://doi.org/10.1371/journal.pone.0223833

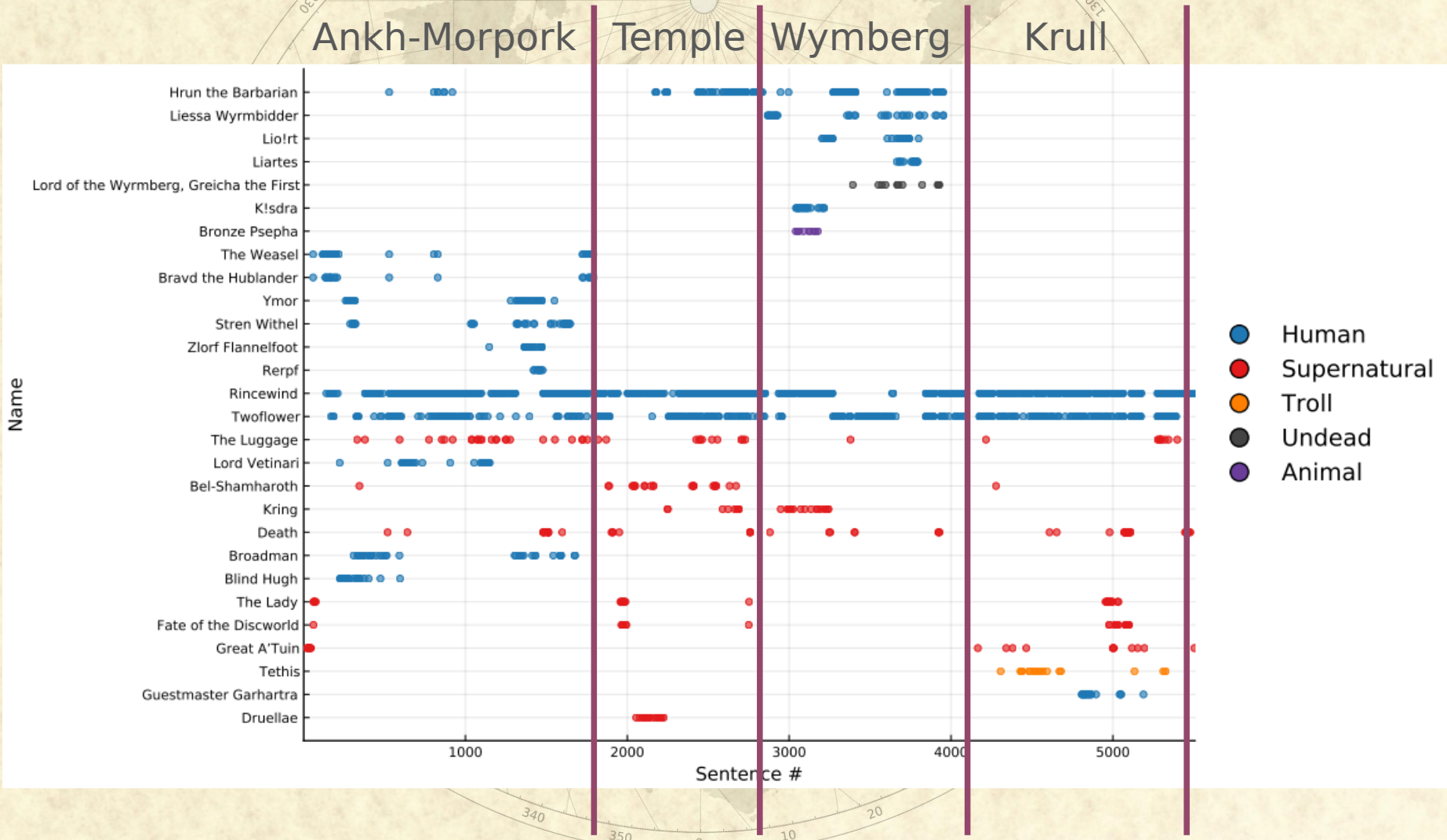
Robust measure of the number of characters in each book



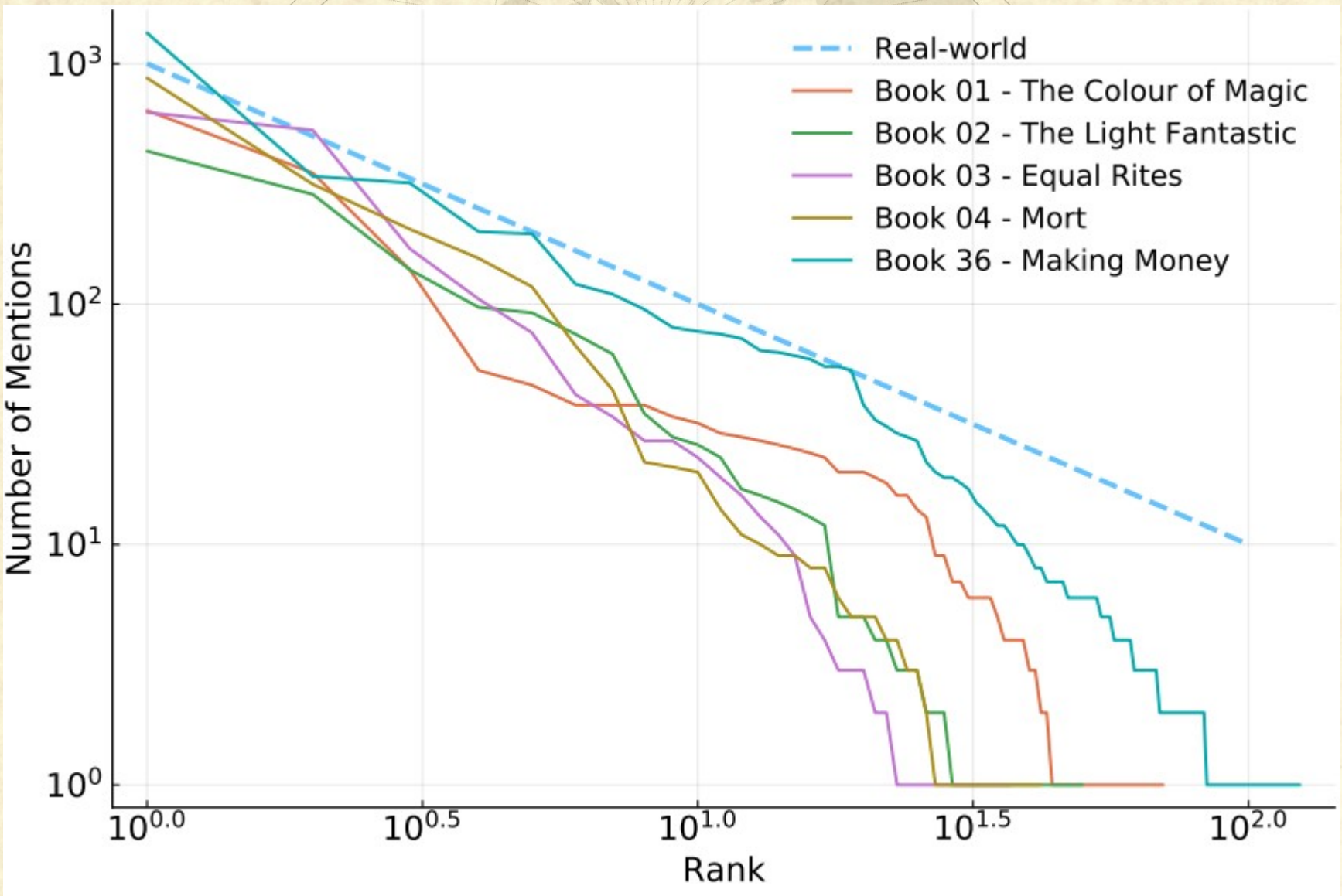
Timeline: The Colour of Magic



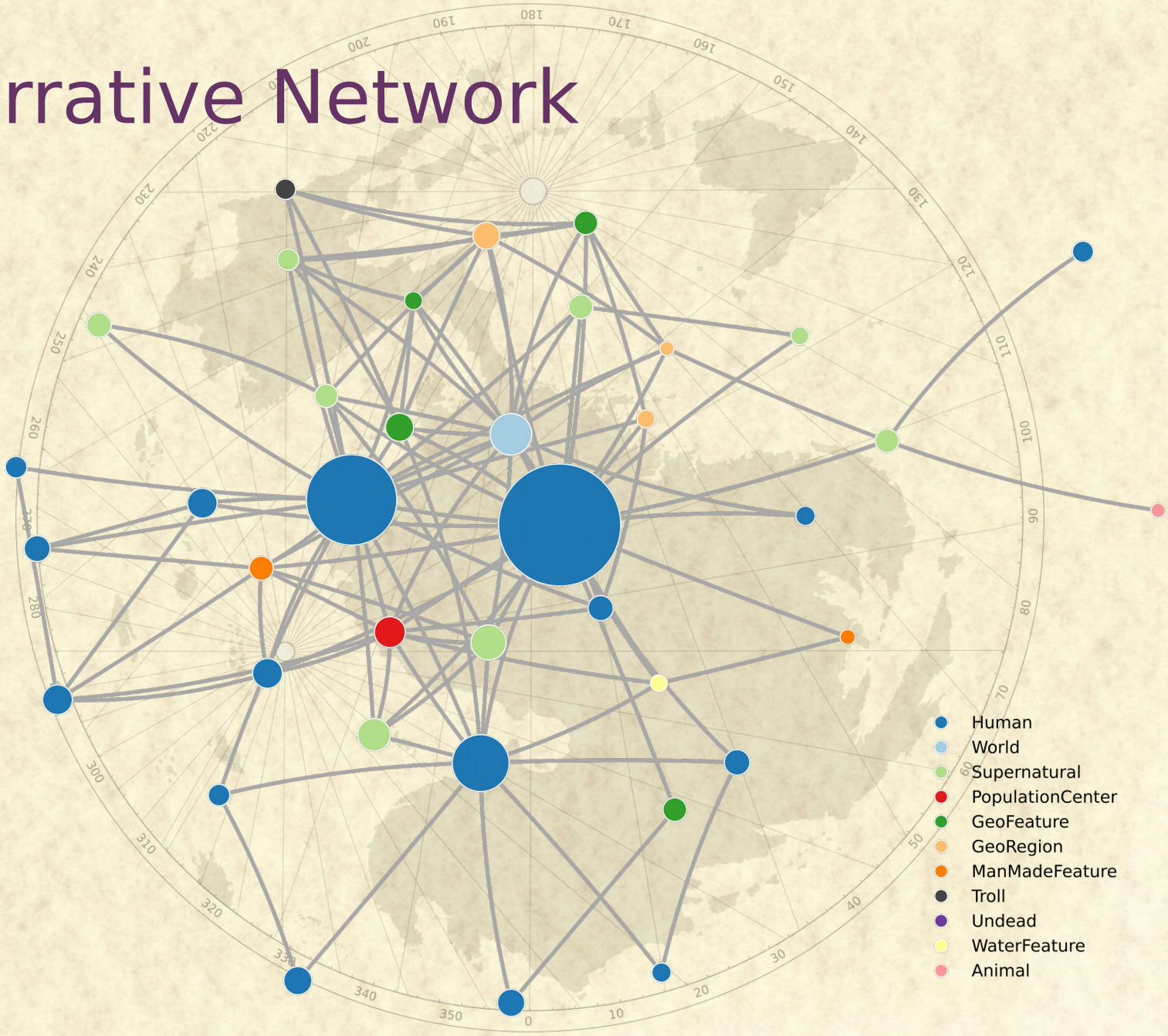
Timeline: The Colour of Magic



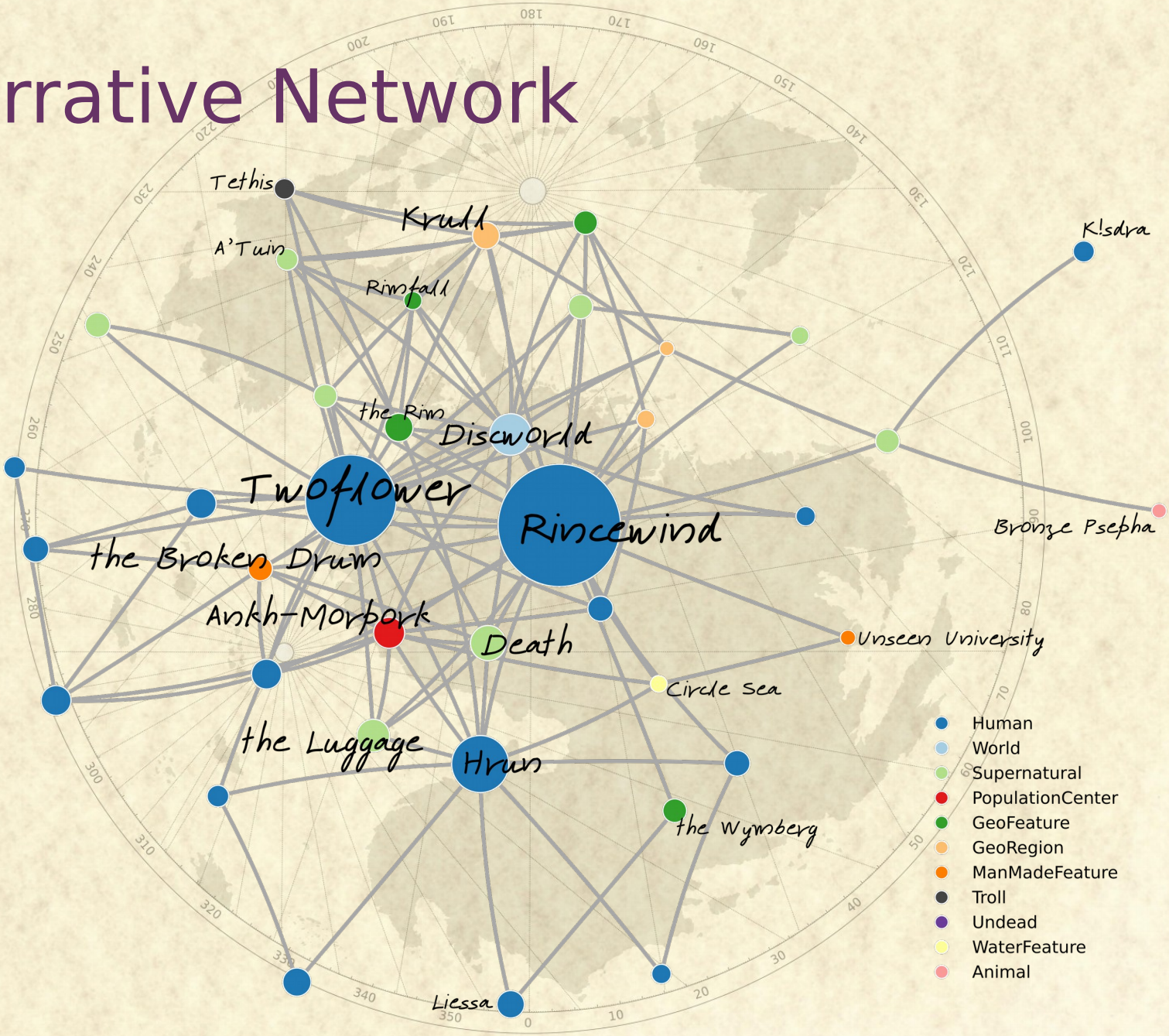
The Rich Get Richer



Narrative Network



Narrative Network



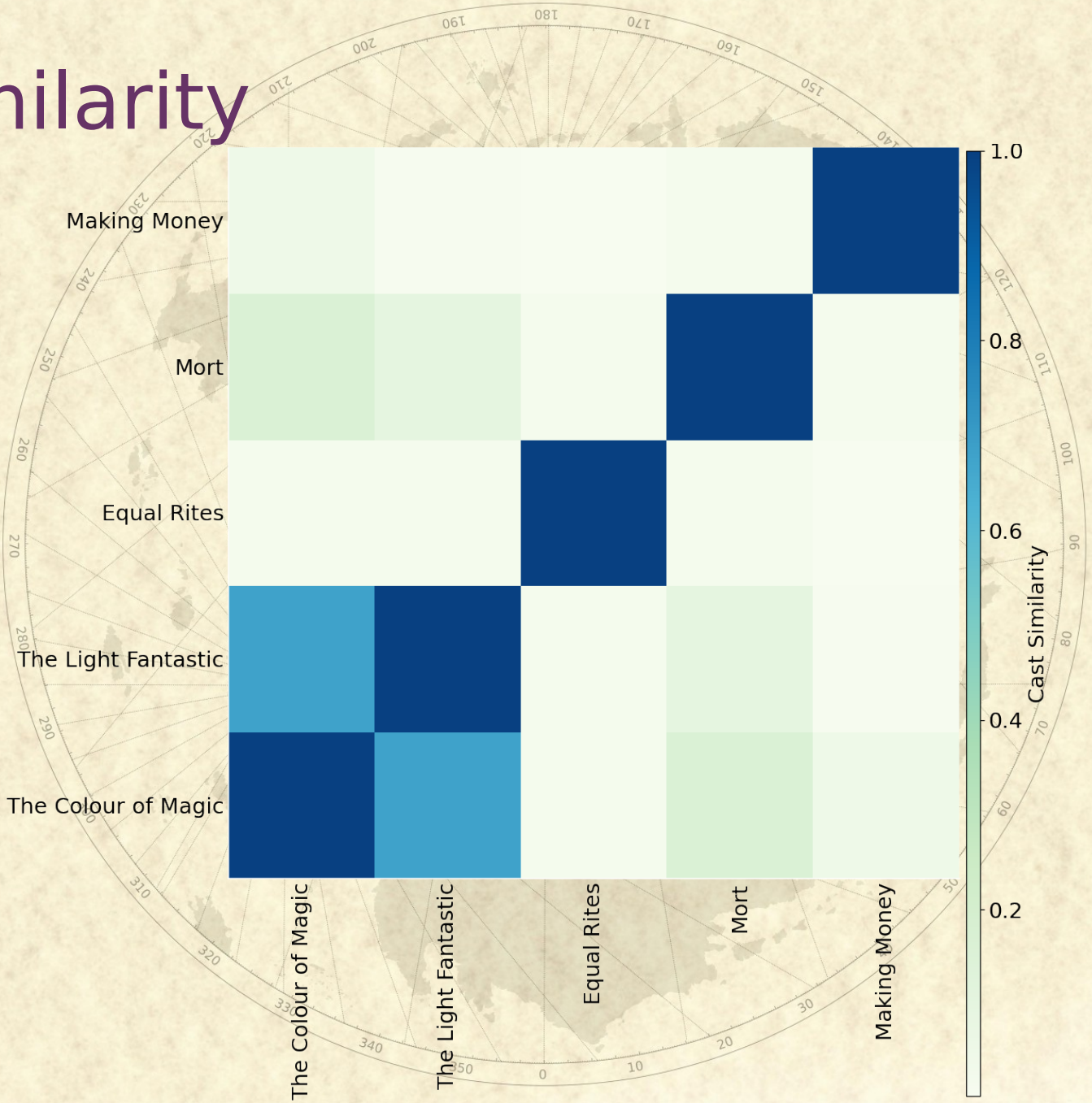
Conclusion

- One drawcard of successful fantasy series has been the worlds in which they are set
  - E.g.
    - Discworld
    - LoTR
    - GoT
  - We can learn a lot about ourselves from understanding the constructed worlds to which we are drawn
  - LOTS more to come
  - THE DATA WILL BE OPEN

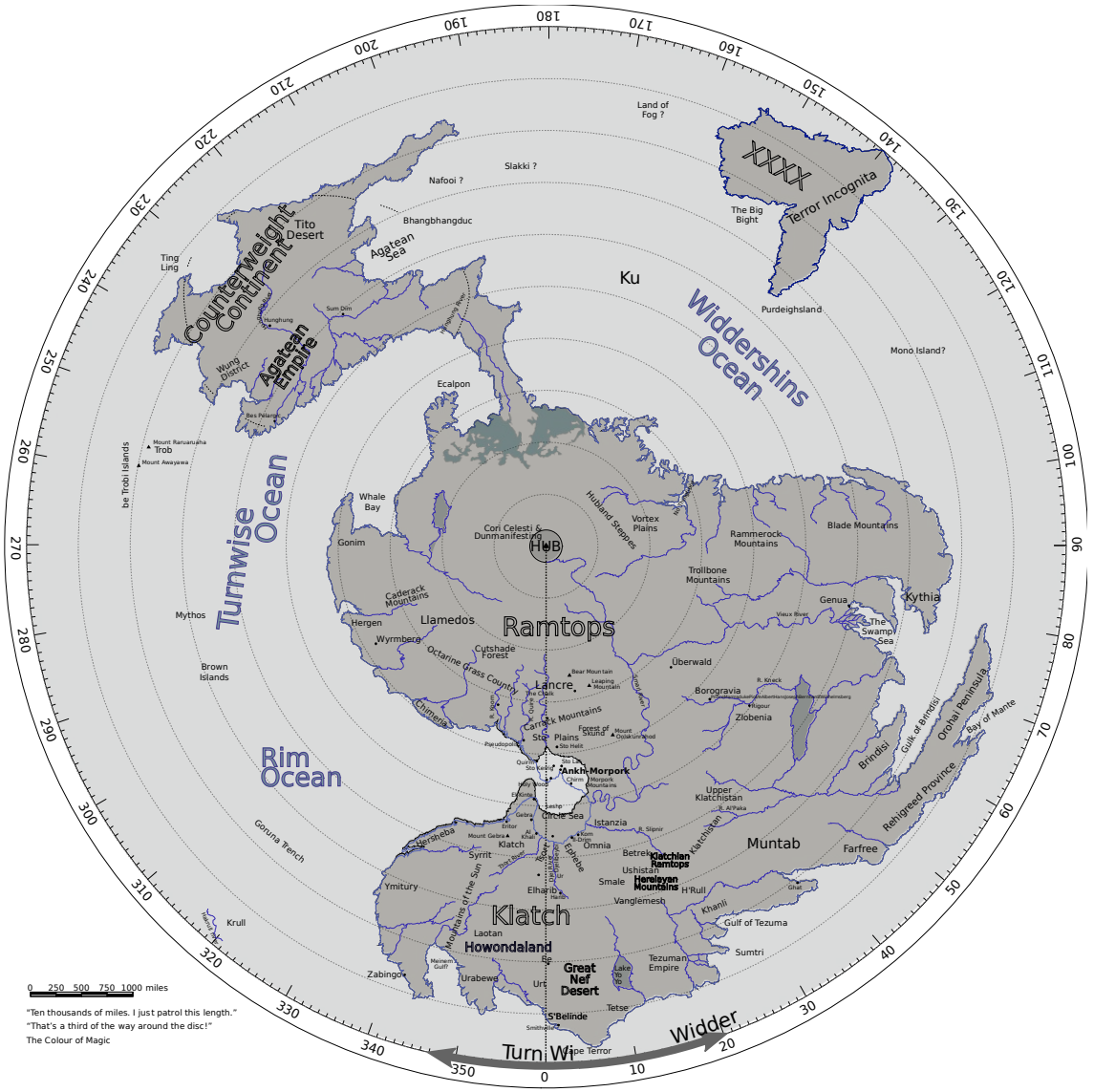
Extra Slides

- Some bonus material

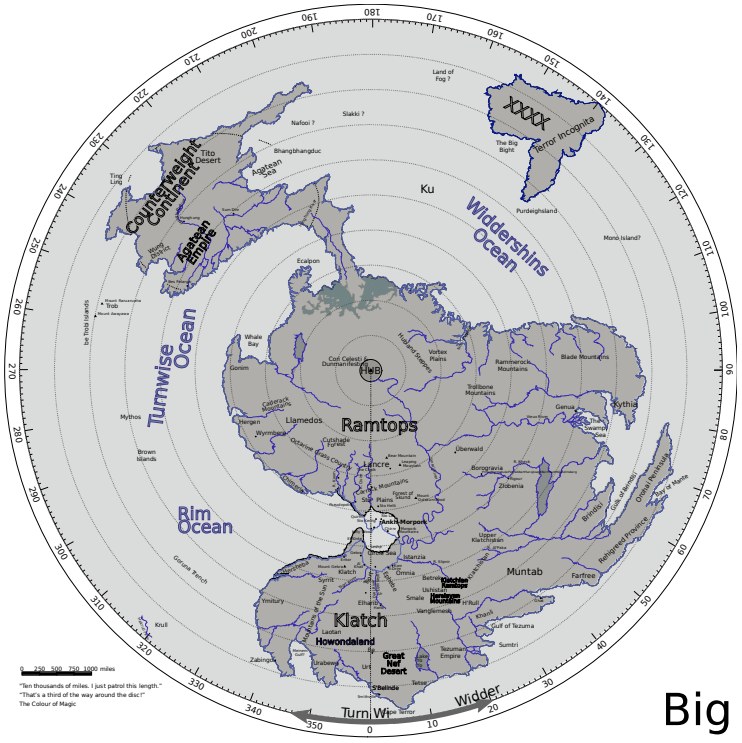
Similarity

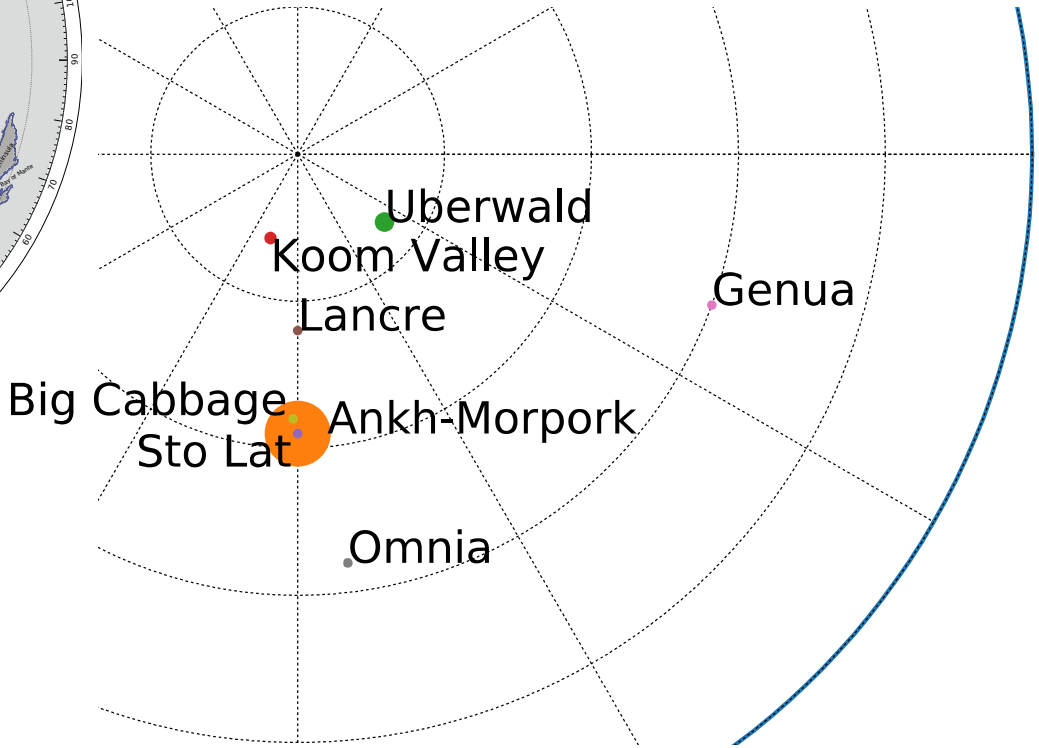

Making Money
Mort
Equal Rites
The Light Fantastic
The Colour of Magic
The Colour of Magic
The Light Fantastic
Equal Rites
Mort
Making Money
1.0
0.8
0.6
0.4
0.2
Cast Similarity

The Discworld

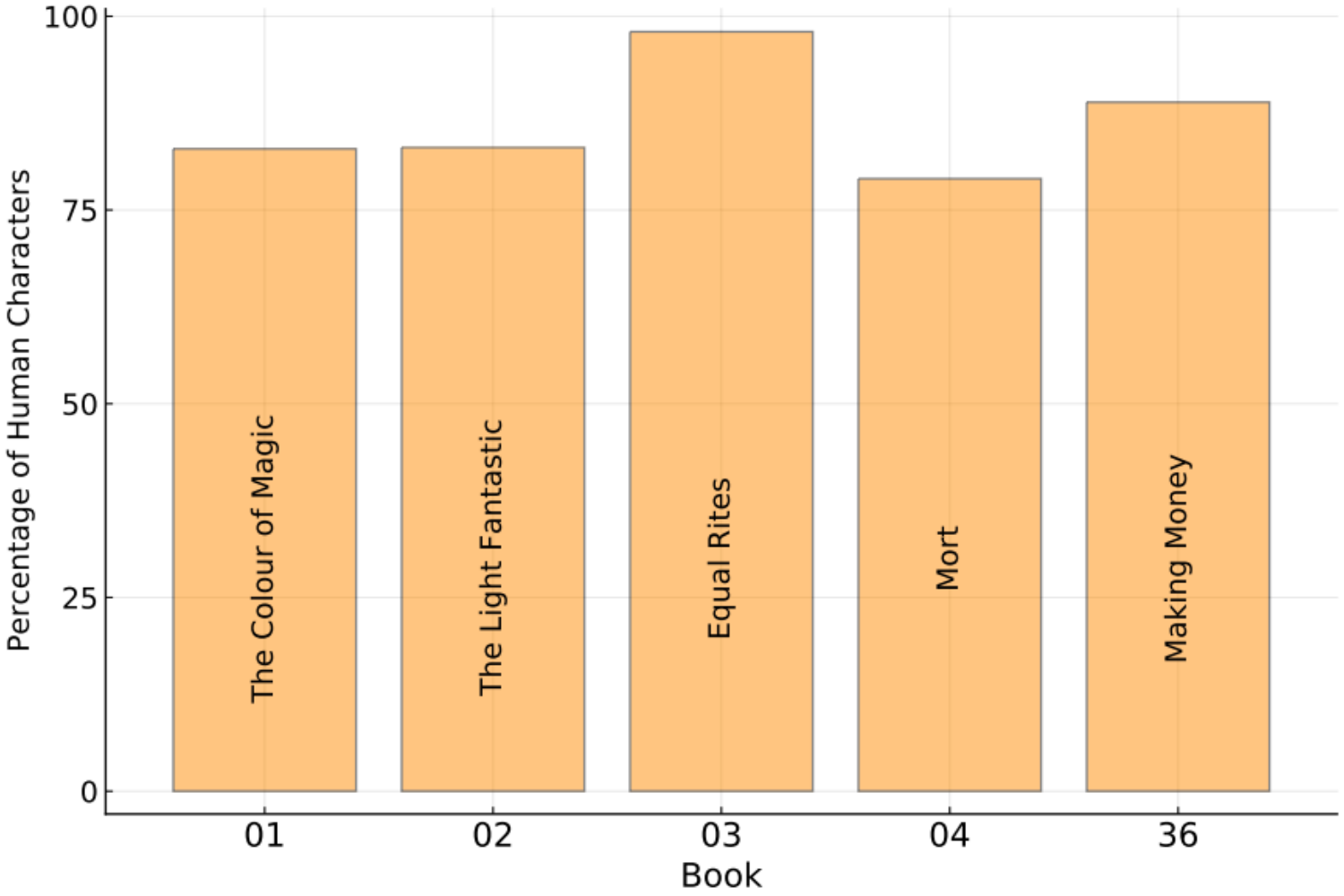


Making Money

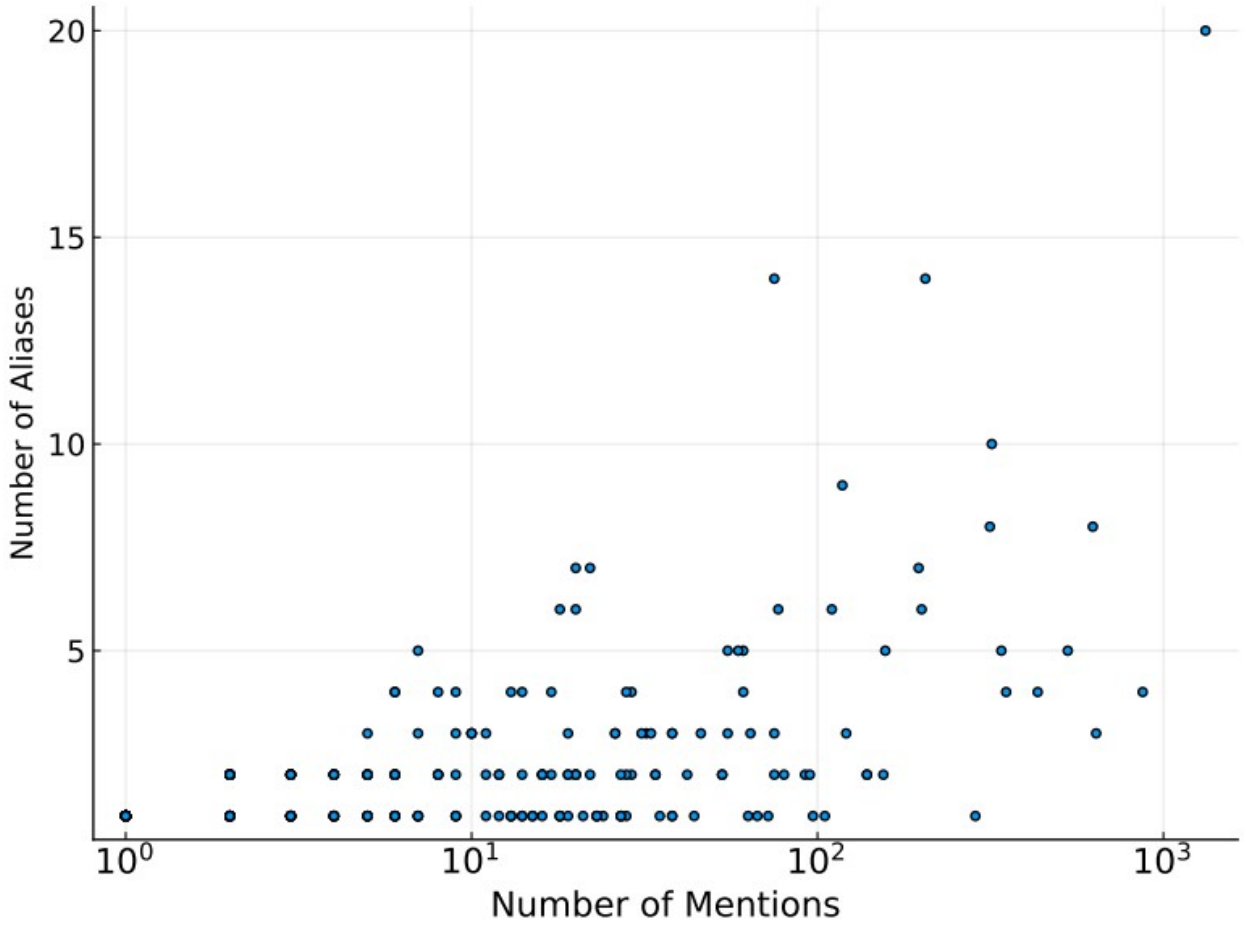




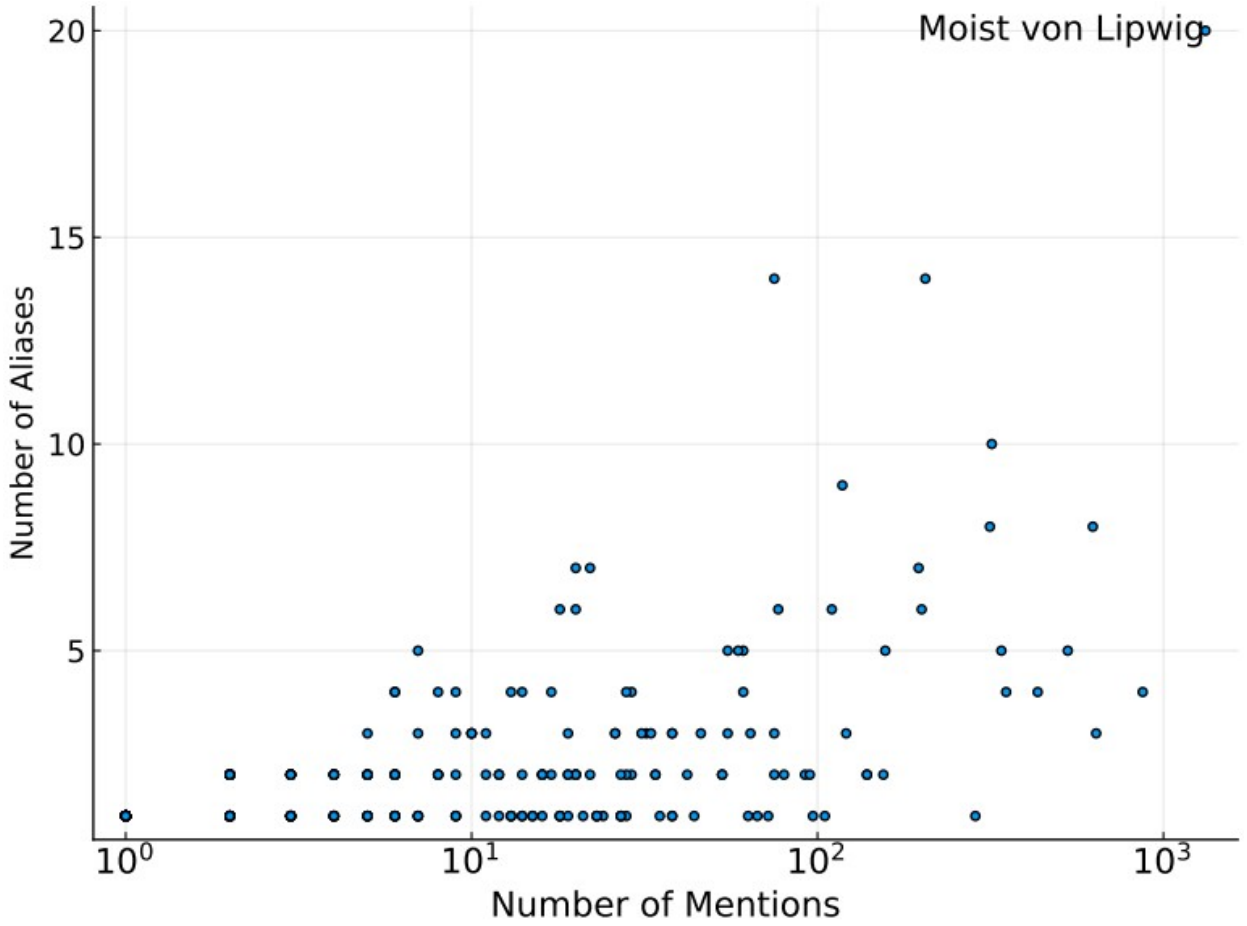
How does Pratchett's use of non-human characters change?



How many aliases?

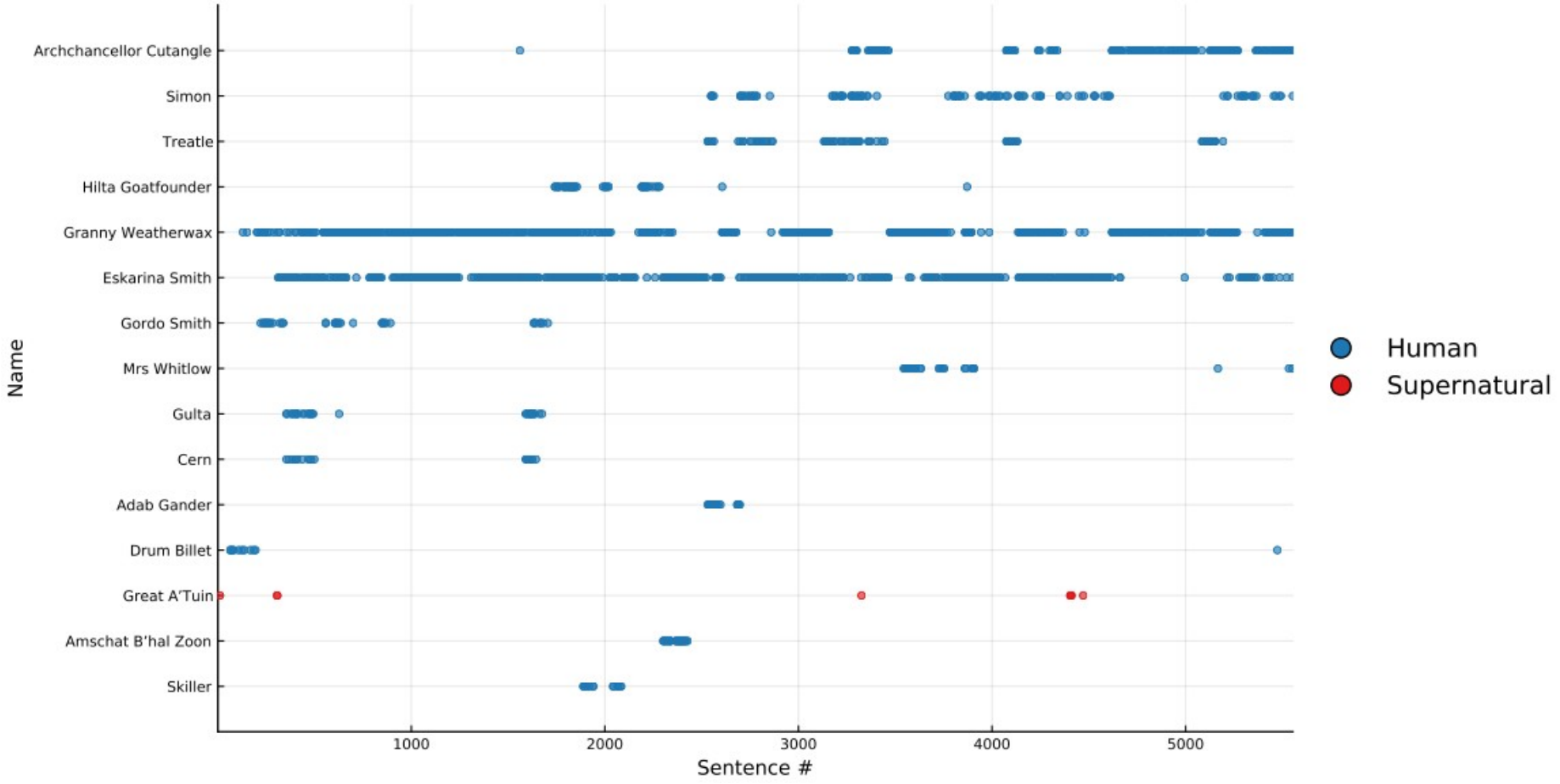


How many aliases?

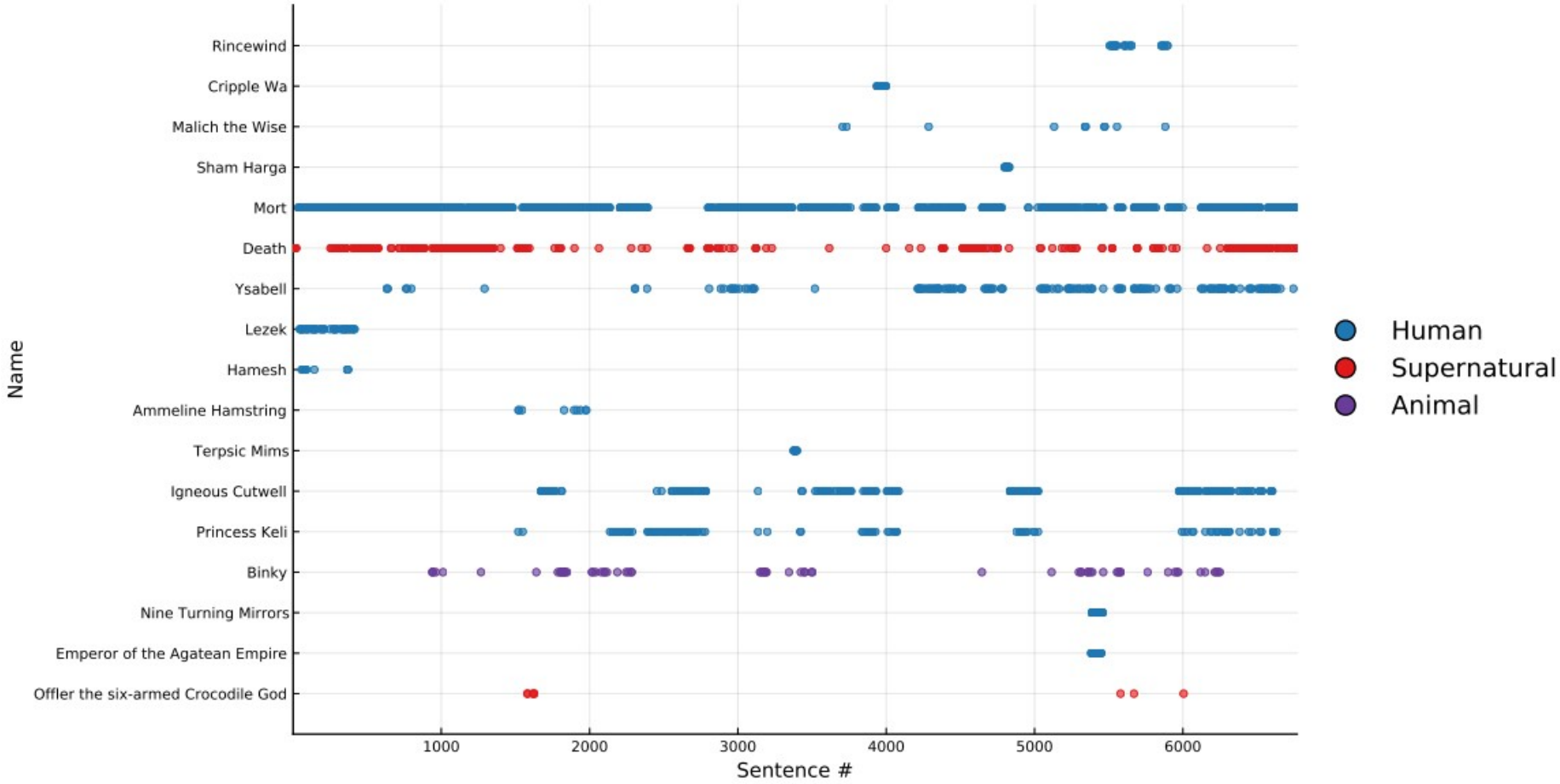


Timeline: The Light Fantastic

Timeline: Equal Rites



Timeline: Mort



Timeline: Making Money

Random Graphs and SERNs