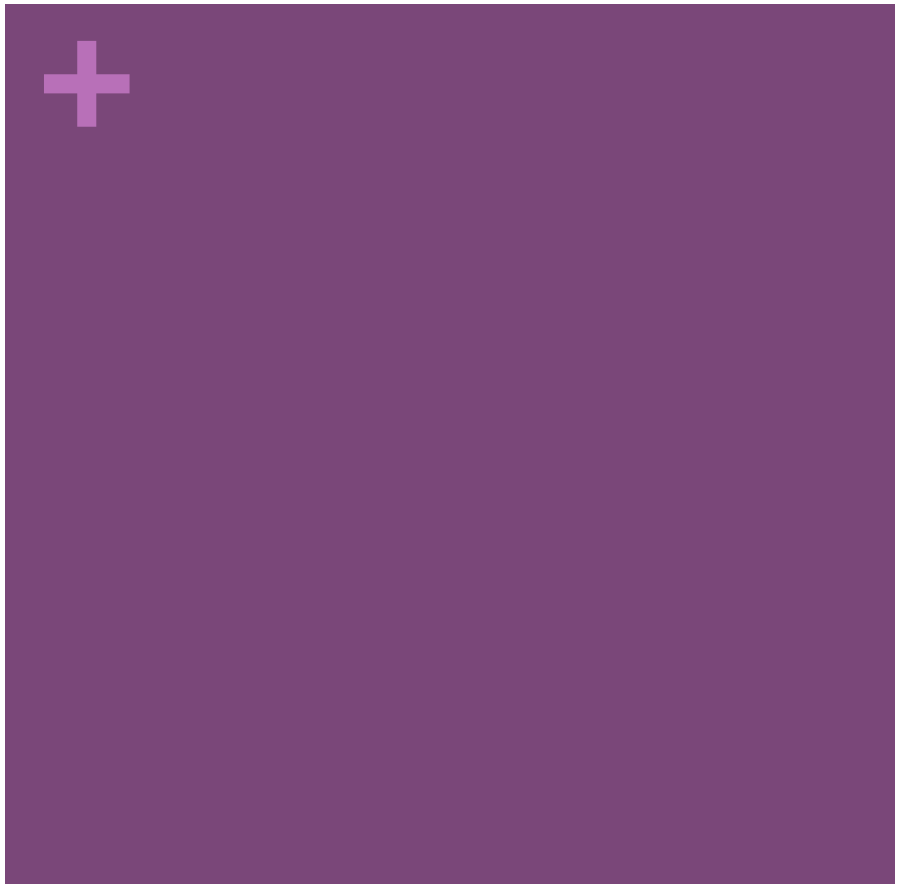
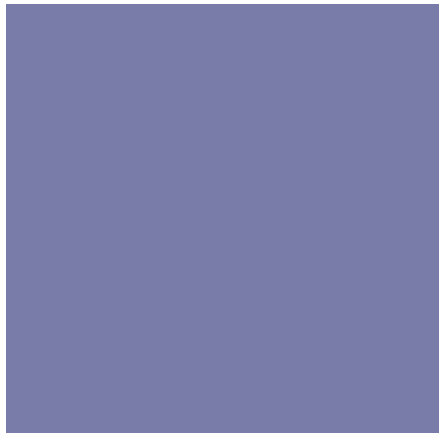
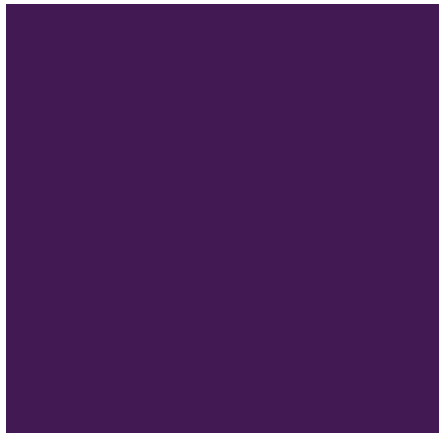




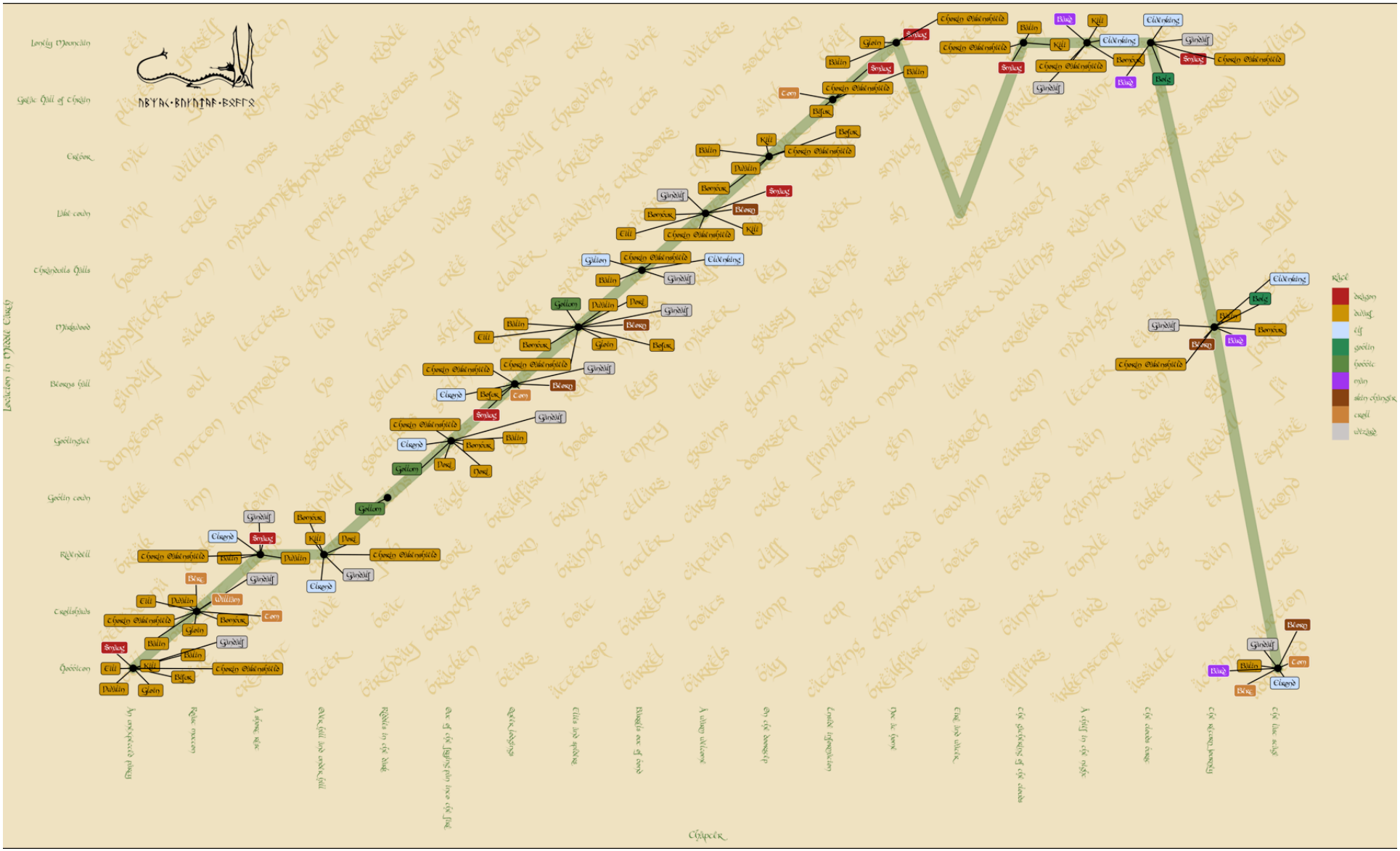


Pratchett, the Discworld, and Narrative Networks

Matthew Roughan

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

matthew.roughan@adelaide.edu.au



Courtesy of Jono Tuke

Spatially Embedded Random Graphs (SERNs)

- There is a cost for longer (physical) links
  - **$** for physical cables
  - Power in wireless networks
  - Time/effort to maintain longer links
- Model this by
  - Generate random point/nodes in a metric space
  - Link nodes with probability dependent on distance

$$p_{ij} = h(d_{ij})$$

  - Many examples (with many different names)
    - Geometric random graphs
    - Waxman
    - …

The Waxman Random Graph

- Waxman used an exponential *distance deterrence function*

$$p_{ij} = q \exp(-s d_{ij})$$

- Why this particular case
  - One of the earlier models (1988) and very often used
  - Tractable
  - Lots of generalisability
- Notes
  - My parameterization is a bit different
  - Waxman, despite the exponential function, is not an ERGM
  - Distances are not latent

Maximum Likelihood Est (MLE)

- Assume independence
  - Crank the MLE handle*
  - Sufficient statistics
    - Number of nodes
    - Average edge length
  - Get an equations to solve

$$\frac{G'(s)}{G(s)} = -\bar{d}$$

- Pros
  - MLEs come with many useful results (asymptotic normality)*
  - Input needed is very small
    - Can even work with sampled links
  - O(E) calculation
- Cons
  - Model dependent
  - Numerically heavy (unless you are careful)

* - there are some subtle technical issues

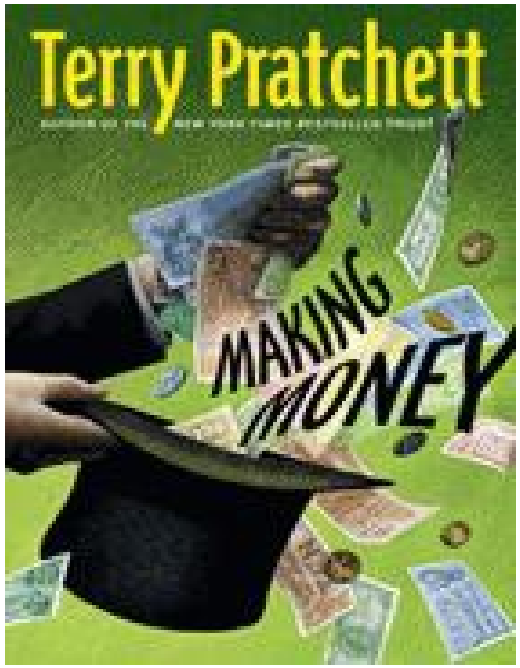
Internet Data

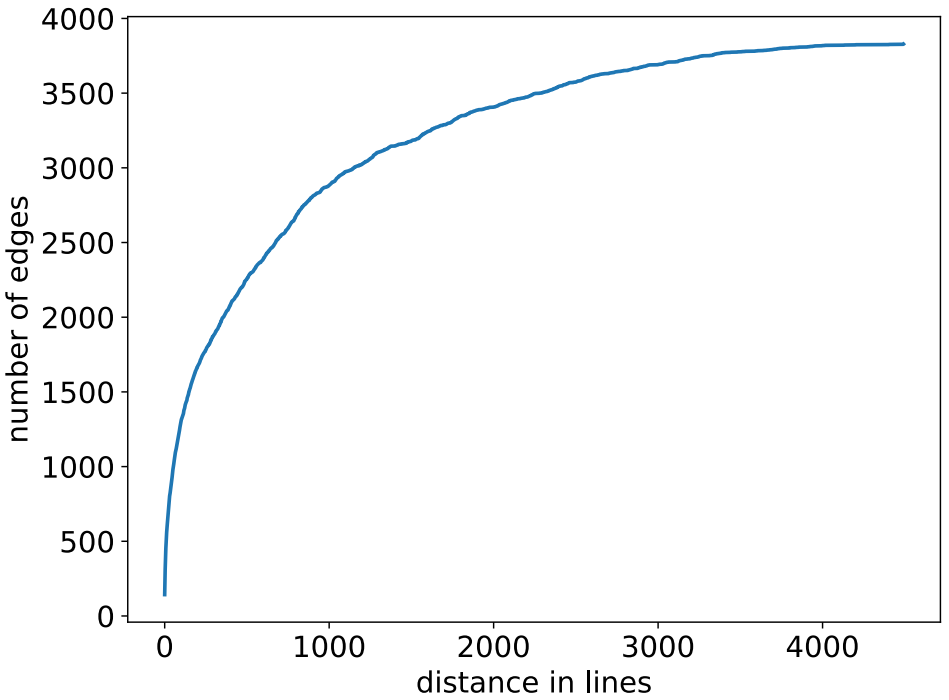
- Mercator Internet data: https://ieeexplore.ieee.org/document/832534
  - Traceroutes (IP level 3 topology)
  - Data problems
    - Missing data
    - Aliasing
    - Geolocation errors => short distances are wrong
    - Timing estimates for “non” links are bad, so can’t use

Region	N	E	Average length (miles)	S
USA	123426	152602	384.7	6.63
EUROPE	32928	30049	319.5	10.09
JAPAN	14318	16665	317.6	7.30

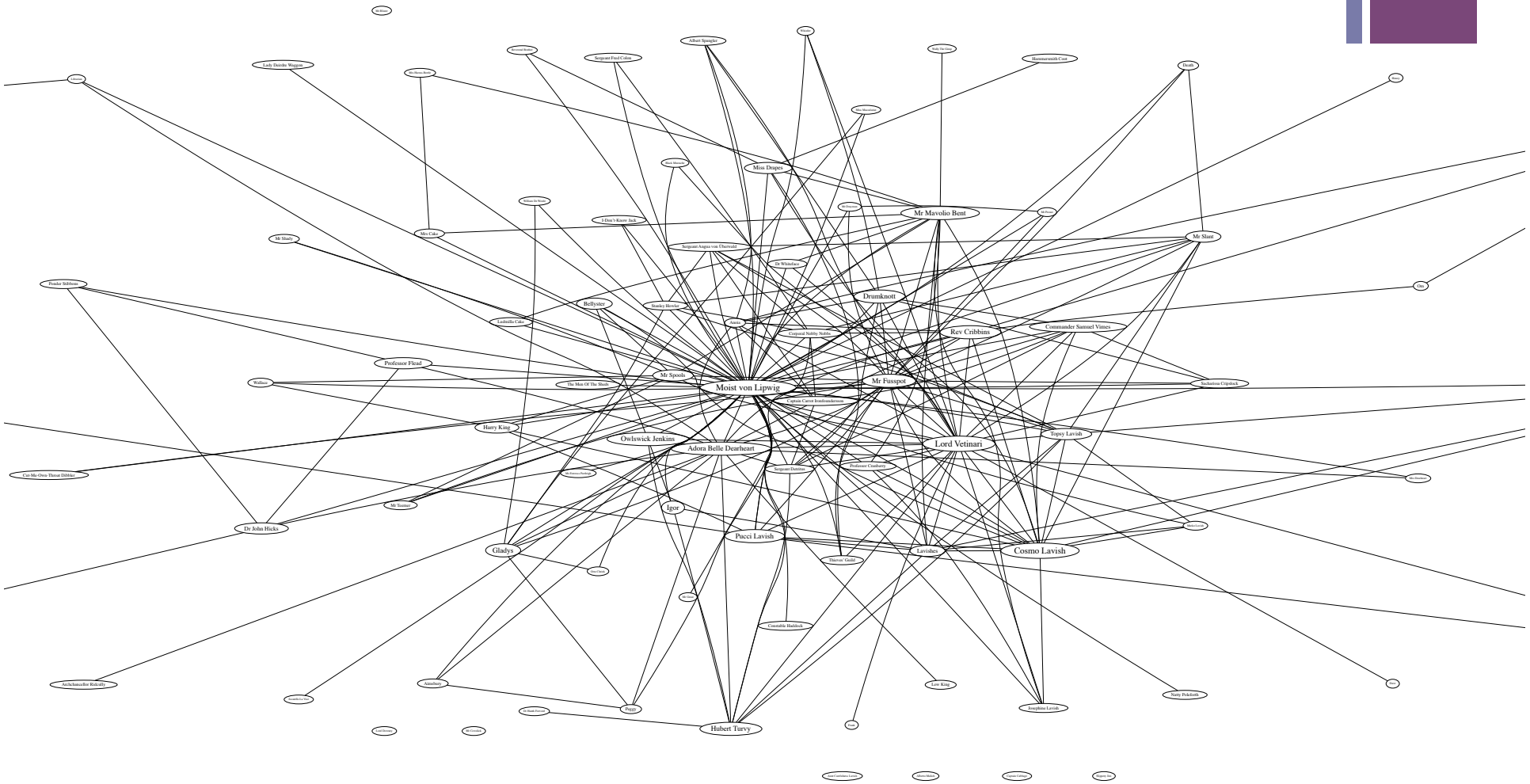
Narrative Networks

- Typically
  - Form a network of characters
    - Nodes = characters (88 identified)
    - Edges = by association (often proximity in text)

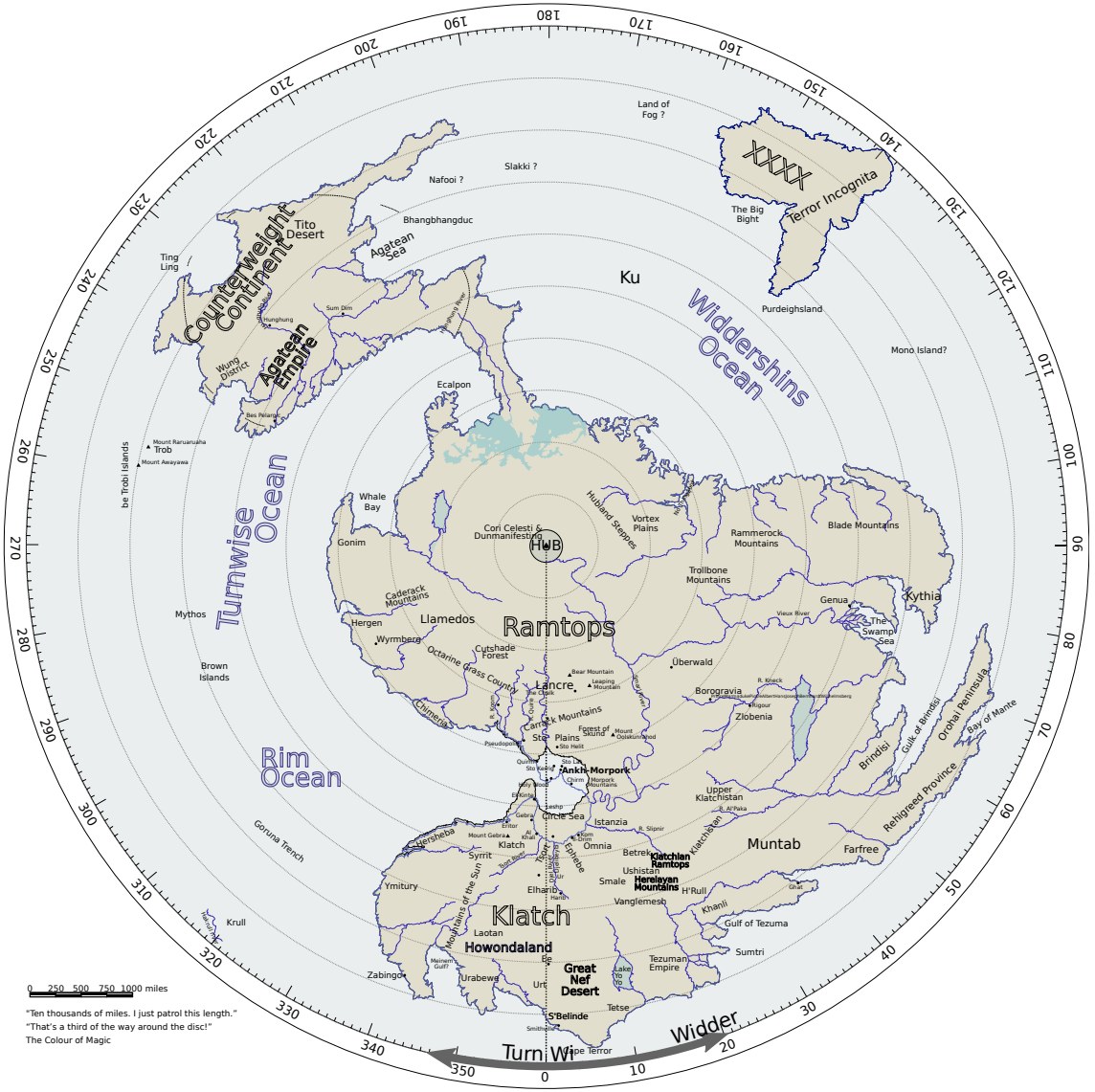




Making Money Network



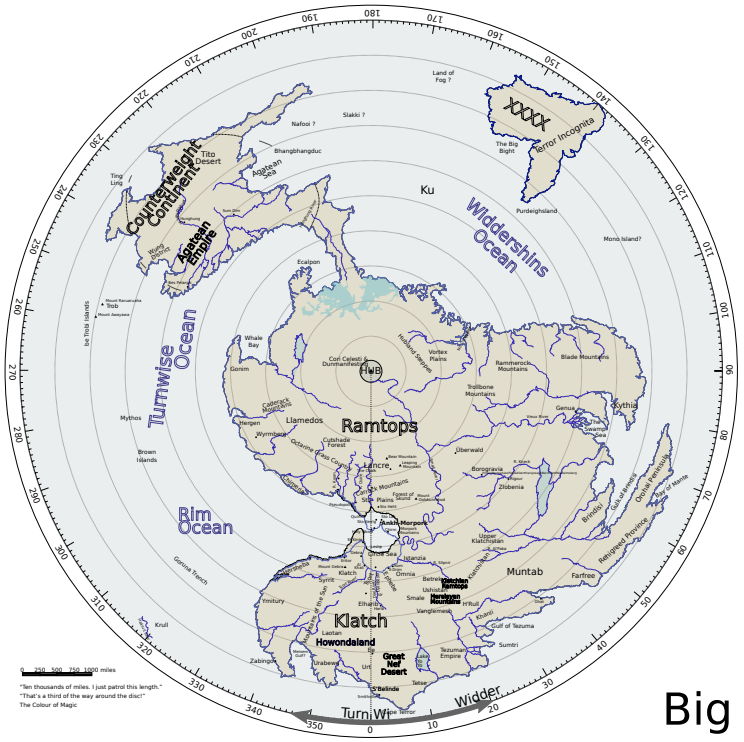
The Discworld

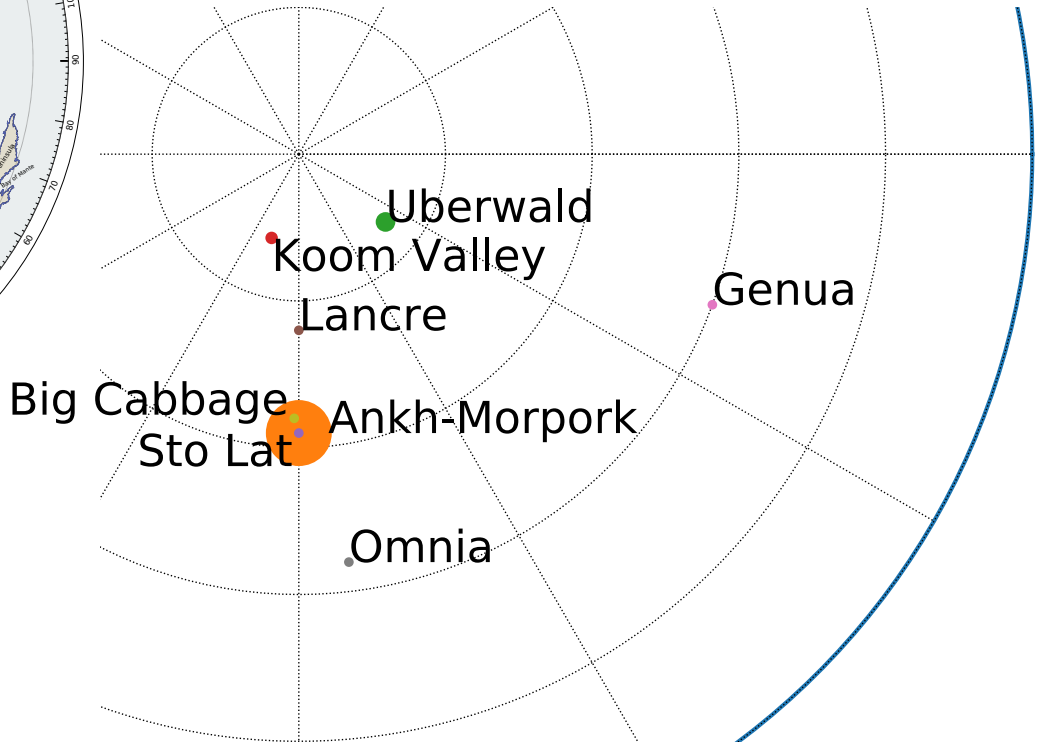


Extracting locations

- Proper nouns
  - People
  - Things (Dates, Titles of Books, Organisations,…)
  - Locations
- Proper nouns are in Title Case
  - Should be easy?
  - But
    - Sentences begin with capitals (but could still be a proper noun)
    - Clauses, e.g., “We start dialogue with a capital”
    - Pratchett (and others) use capitals For Emphasis
    - Numbers can be part of an address
    - Multiword Proper Nouns can involve Joining Words
    - Weird punctuation (interactions of quotations, brackets, and other)
    - OCR errors

Making Money





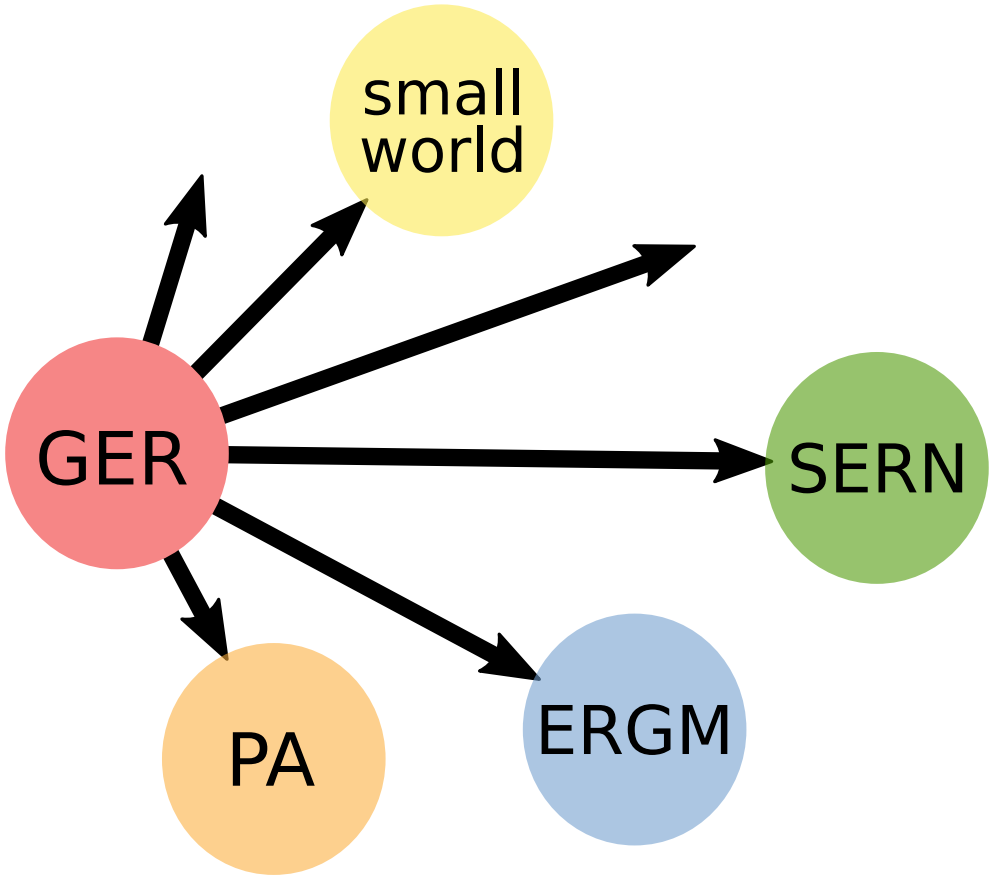
Conclusion

- We can easily work with SERNs
  - MLE estimate
- Many created worlds are rich and complex
  - E.g.
    - Discworld
    - LoTR
    - GoT
  - Part of that is creating realism in relationships
  - Maybe we will see this in spatial relationships
    - “Making Money” wasn’t the best place to start
- Watch this space

Random Graphs

- Why random-graph models?
  - Insights into formation of real networks
  - Simulation models (in higher-level network sims)
  - Understanding growth
  - Predicting missing links
  - Measurement and sampling methodologies
  - …
- Tension between
  - Realism
  - Simplicity/tractability

Random Graphs and SERNs



Possible Inputs

- All possible data
  - Locations of all nodes, and which are linked
  - Hardest to measure
  - $\Omega(n^2)$
- All distances
  - Distance between all pairs of nodes, and which are linked
  - Hard to measure, not necessarily distance metric
  - $\Omega(n^2)$
- All distance of existing links
  - No information about “non” link distances, but assumed metric space
  - $\Omega(e)$
- Sampled link distances

GLM

- Assume independence and treat link/no link as a binary random variable dependent on the covariate of distance
  - Link function is exponential
  - Standard GLM fitting
- Pros
  - Accurate*
- Cons
  - Input: all distances
  - Costly: $\Omega(n^2)$ computation and memory

* - there are some subtle technical issues

Log-linear regression

- Create scaled histogram of frequencies of distances as estimate of distance deterrence function

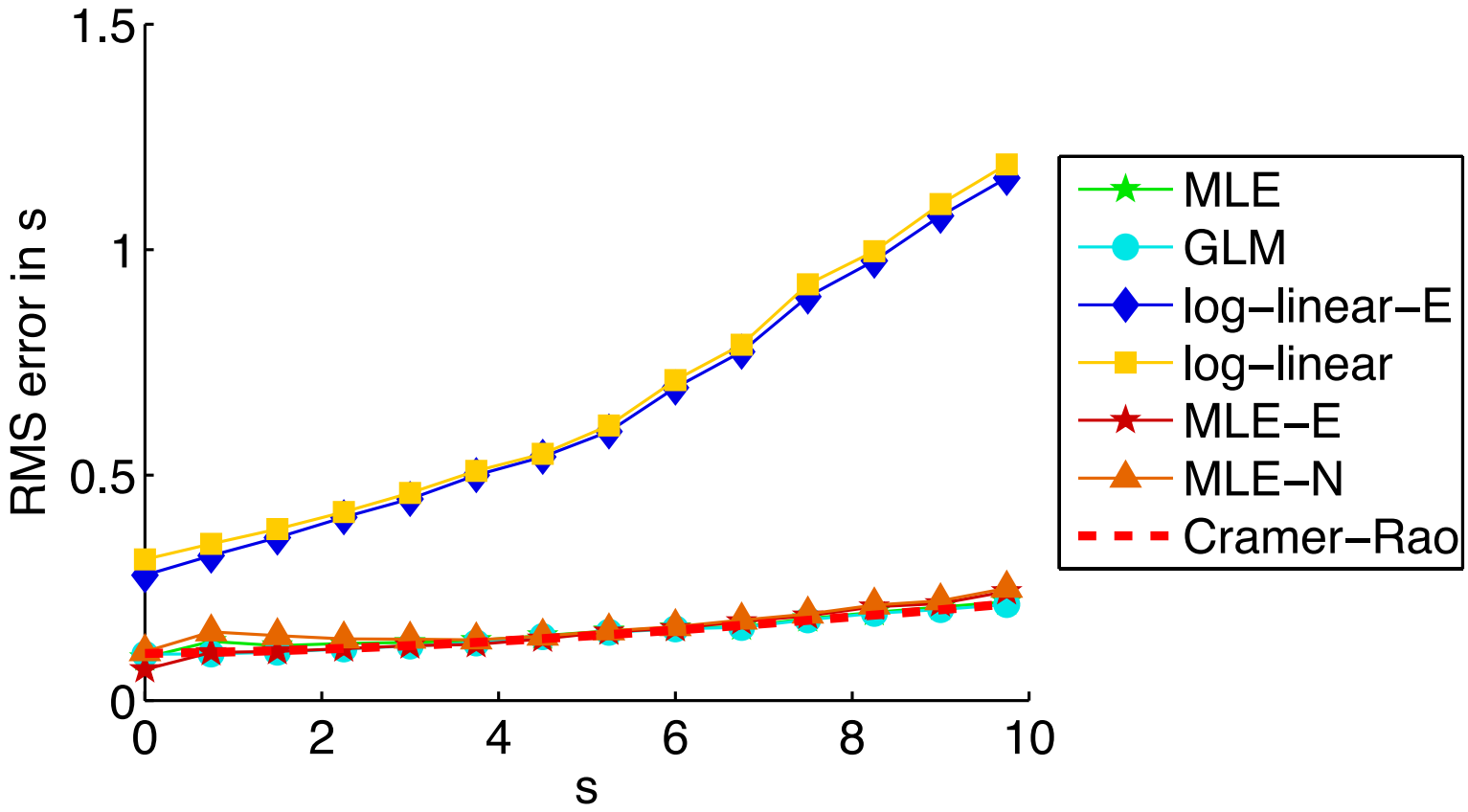
$$q\exp(-sd) \propto freq(d \,||\, s,q) / g(d)$$

  - Scaled histogram should be proportional to *distance function*
  - Simple log-linear regression

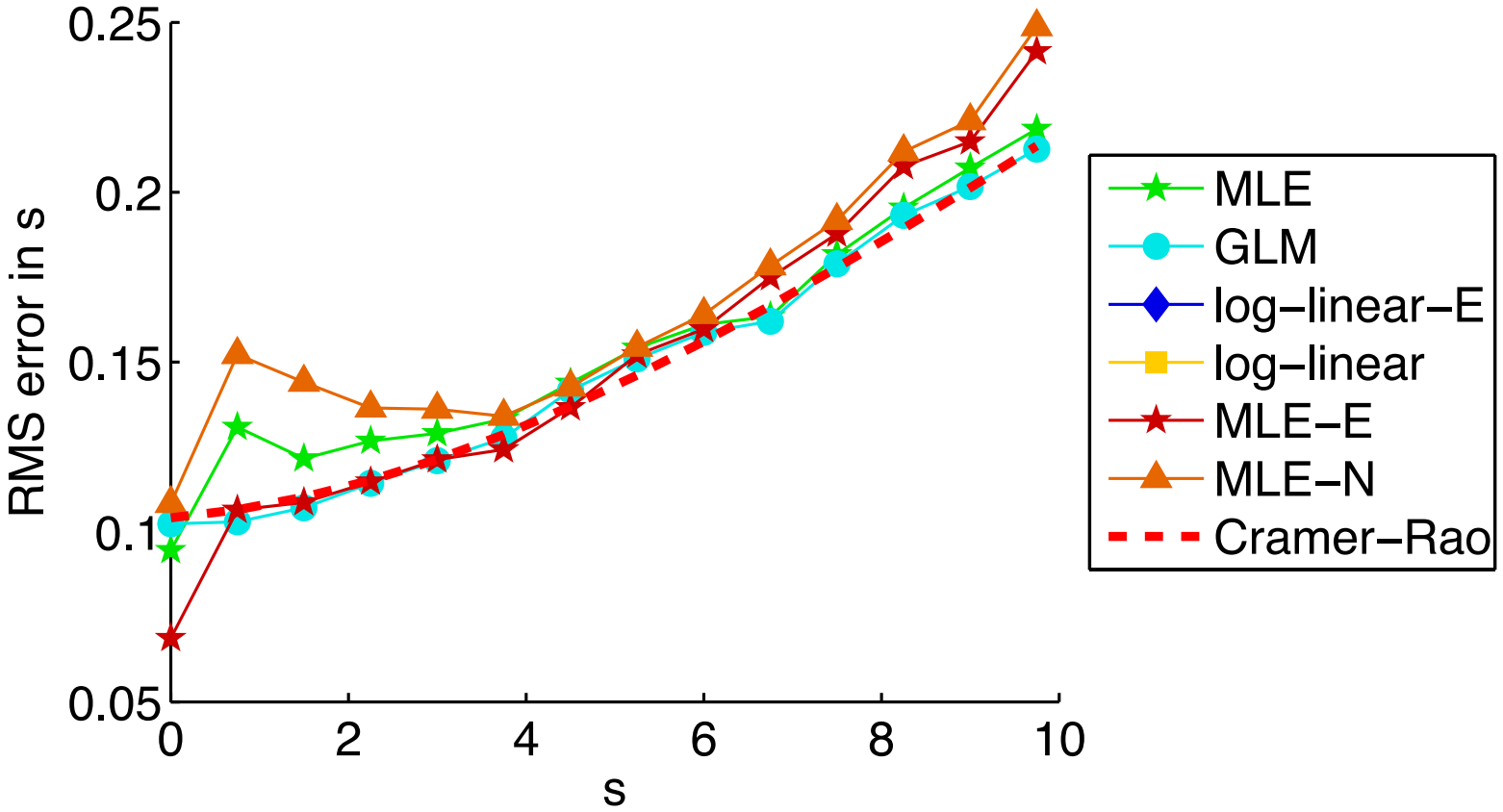
- Pros
  - Simple, fairly fast (histogram is $\Omega(n^2)$ but fast)
  - Regression diagnostics

- Cons
  - All distances needed for the scaling *g(d)*
  - Arbitrary bin size must be chosen
  - Poor accuracy

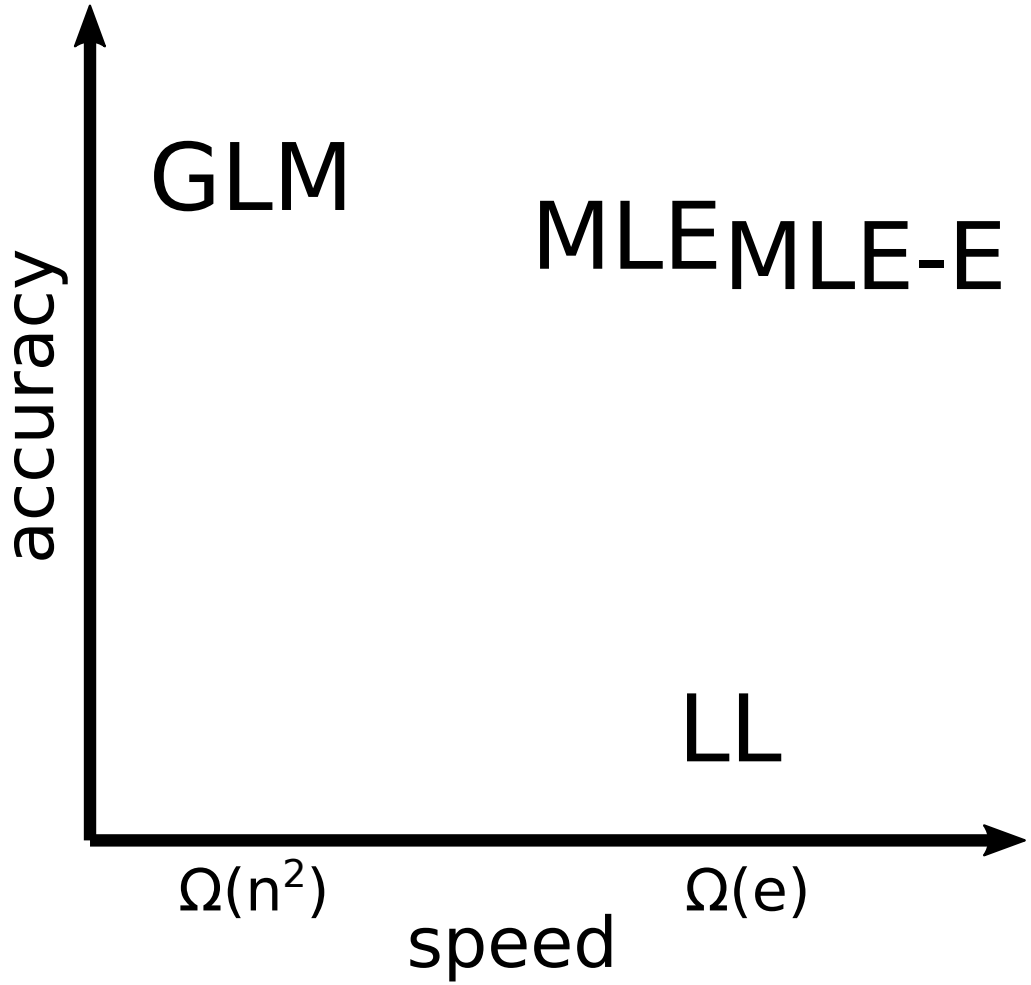
Performance



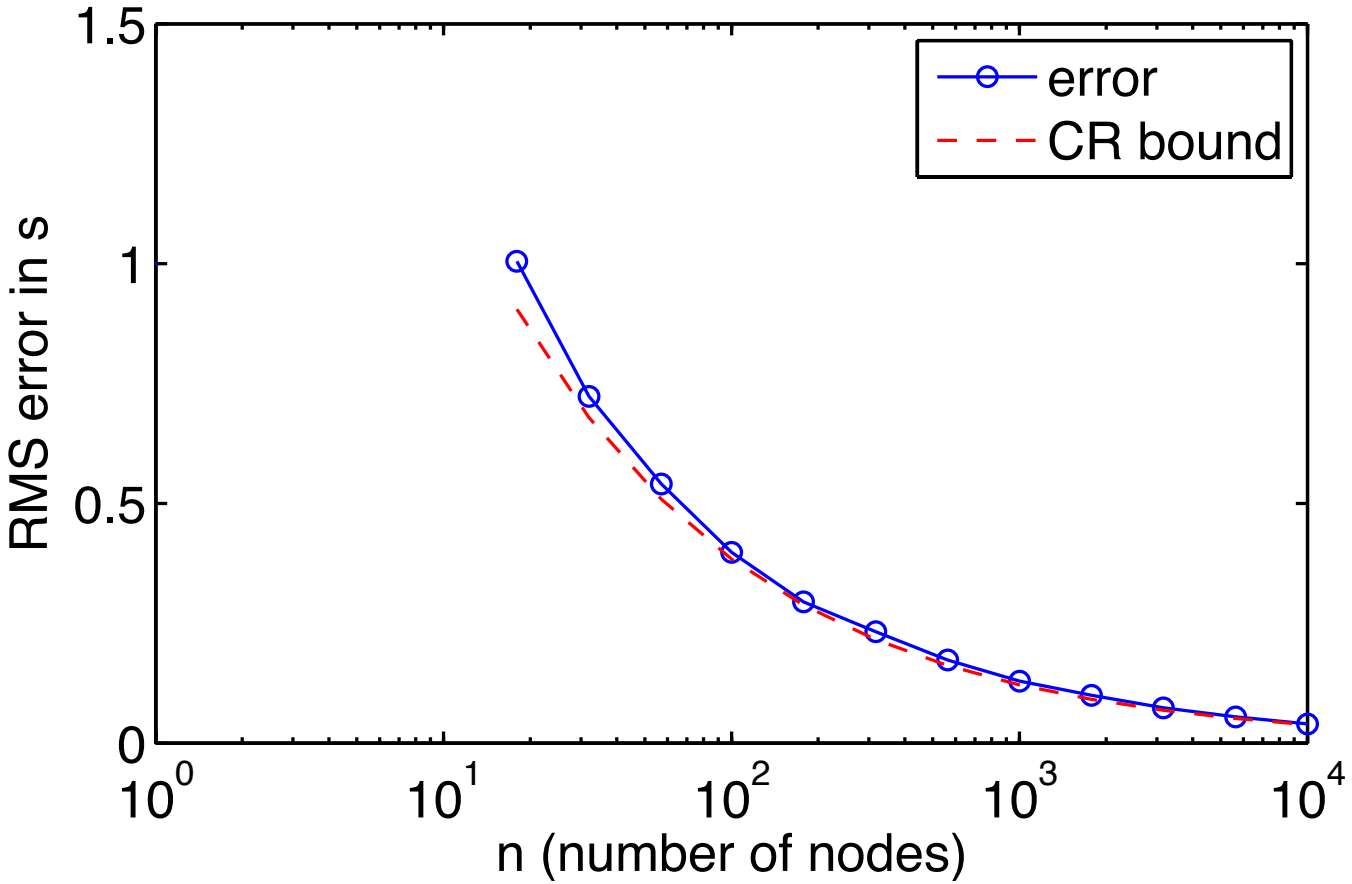
Results (zoom)



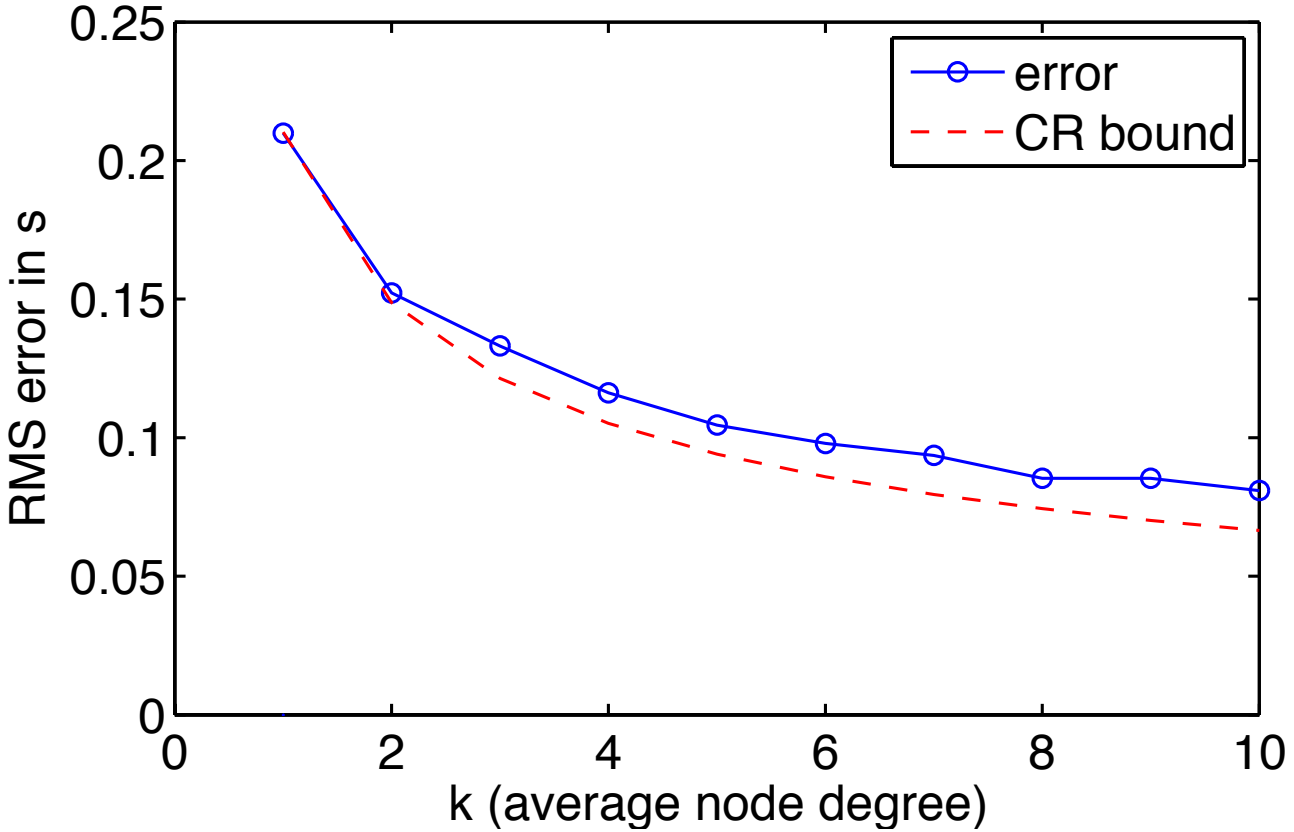
Computation-cost tradeoff



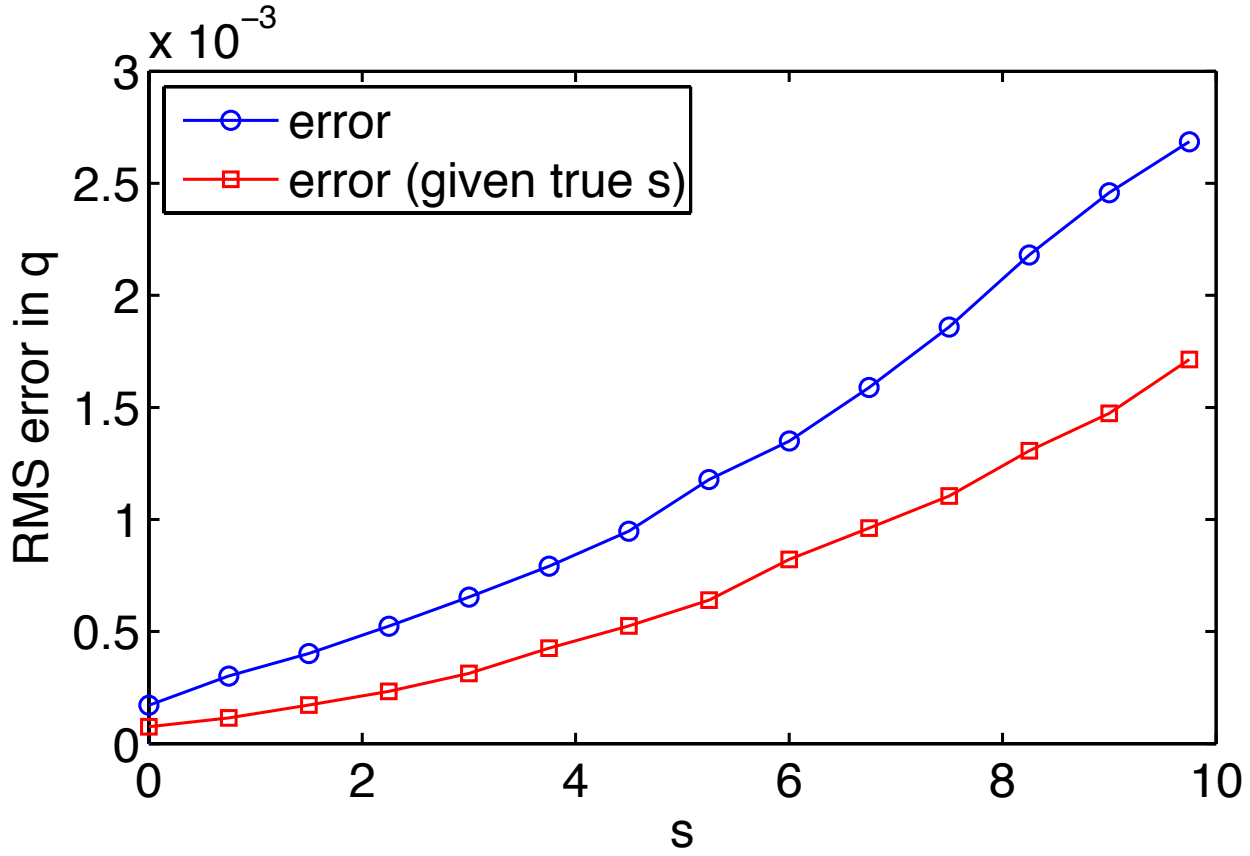
Extra plots: MLE error v size



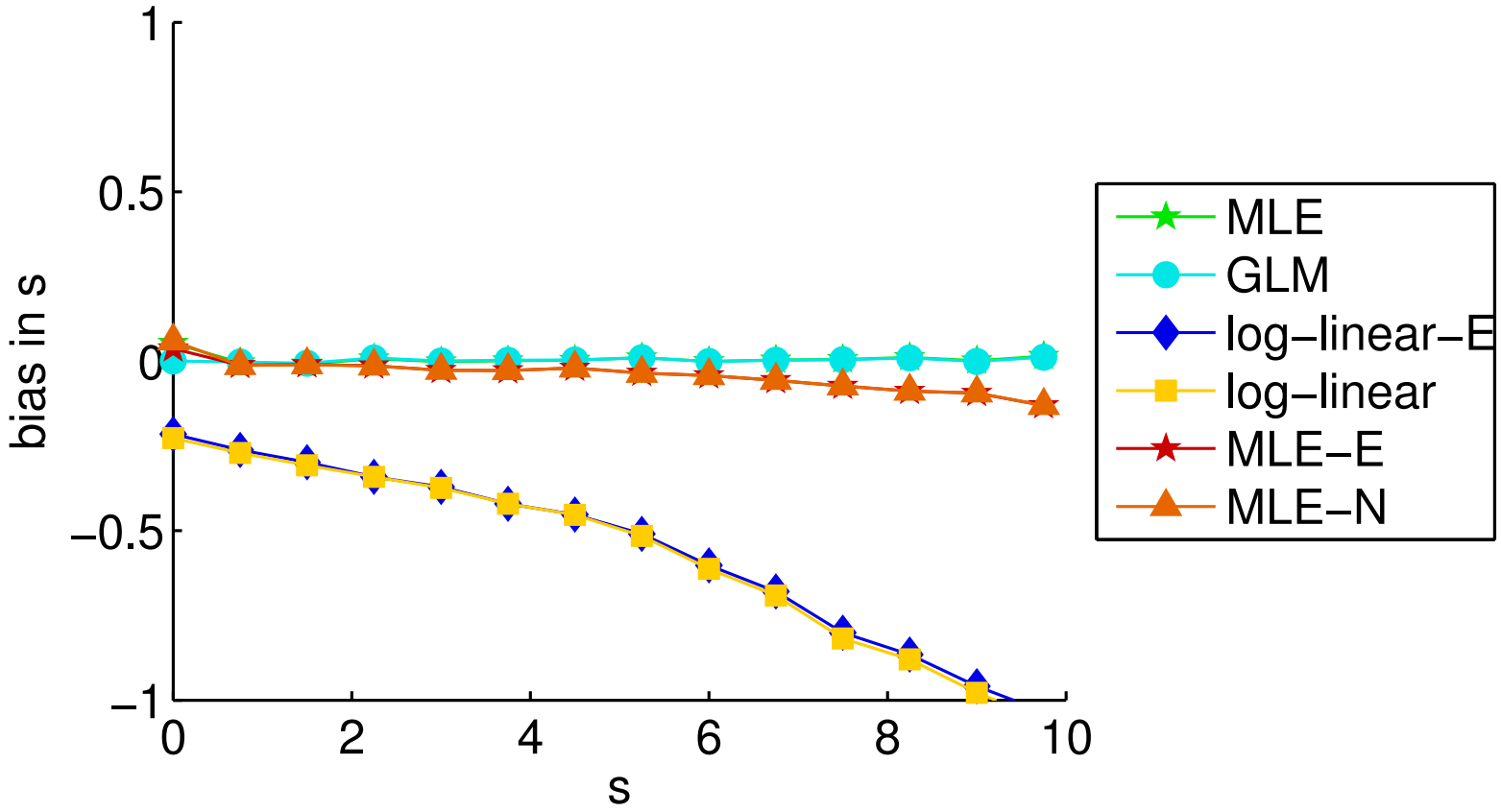
Extra plots: MLE error v density



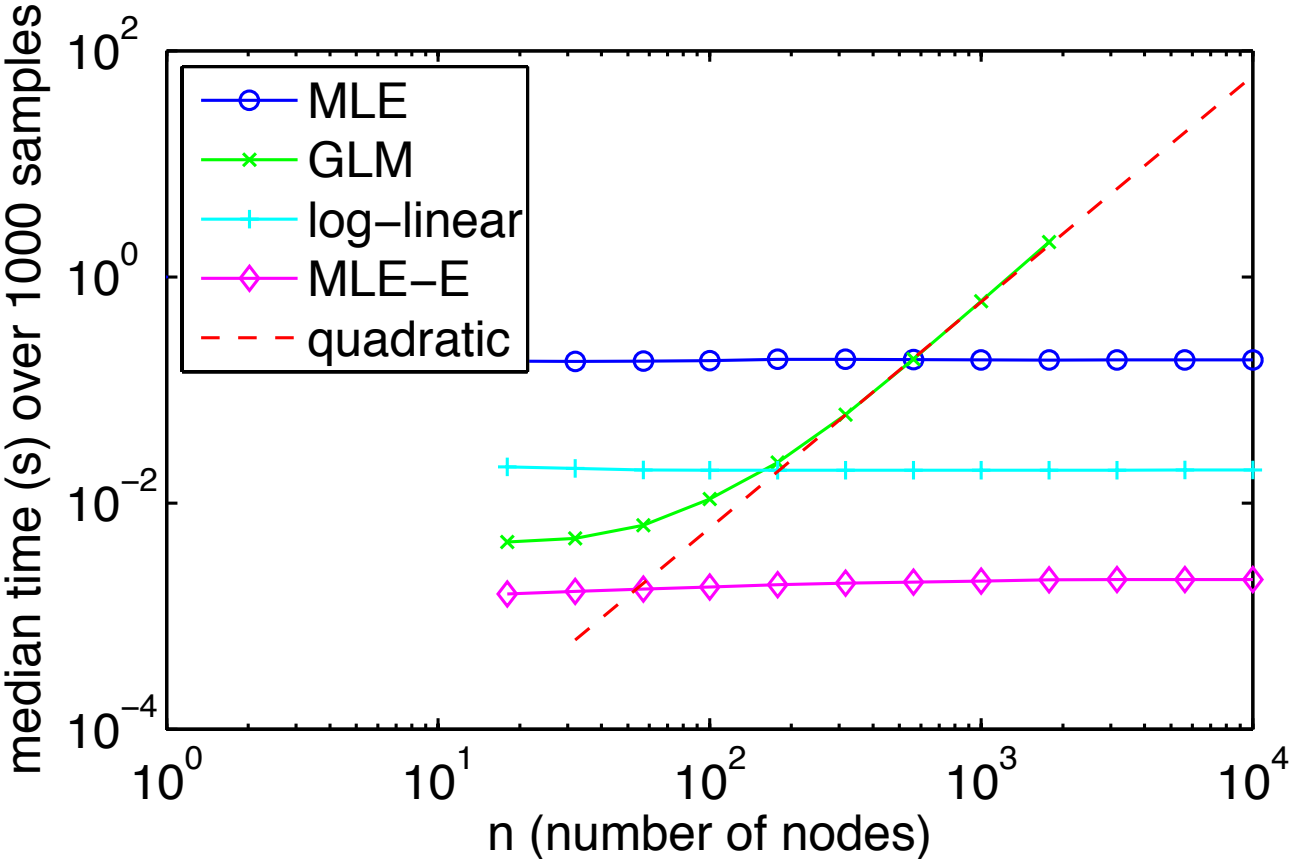
Extra plots: estimating q



Extra plots: bias



Extra plots: time



Extra plots: region mismatch