A Practical Algorithm for Topic Modling with Provable Guarantees

Sanjeev Arora
Rong Ge
Yoni Halpern
David Mimno
Ankur Moitra
David Sontag
Yihcen Wu
Michael Zhu

Presented by: Vanush Vaswani and Kristy Hughes

Topic modeling

- Statistical modeling
- Discovers hidden thematic structure (topics) in a collection of documents
- Help to develop new ways to:
  - Search
  - Browse
  - Summarize

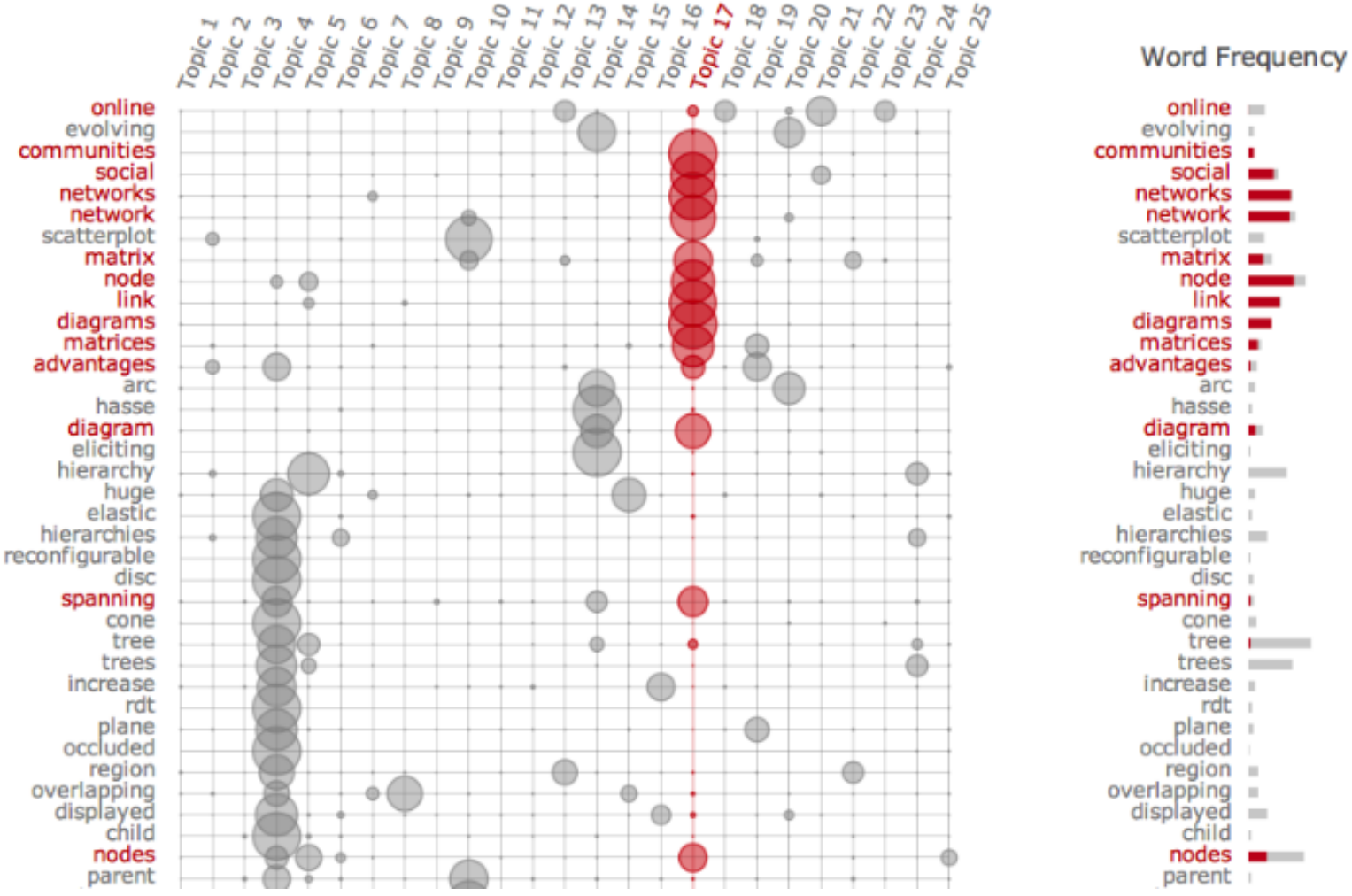
Recent Work

- Posterior inference is NP-hard (worst case)
- Approximate techniques used (SVD, Variational Inference, MCMC)
- Provably polynomial time algorithms: Statistical recovery problem
- Anandkumar et al. (2012)
  - Third-order moments
  - Assumes topics are not correlated
- Arpra et al.
  - Second-order moments
  - Assumes topics are separable
  - i.e. There exists an anchor word for every topic
  - Steps: find anchor words, reconstruct topic distributions

Contributions

- Combinatorial anchor selection algorithm
  - Assumes separability
  - Stable in presence of noise
  - Polynomial sample complexity
- Simple probabilistic interpretation of the recovery step
  - Arora et al. (2012) use matrix inversions $\rightarrow$ sensitive to noise
  - Replace matrix inversion with gradient-based inference
- Empirical comparison between recovery-based algorithms and existing likelihood-based algorithms

Word-topic matrix



Original recovery method

Bayes' Rule

New Algorithm

Finding Anchor Words

Efficient Algorithm

Efficient Algorithm

Related Work

Methodology

Efficiency

Semi-synthetic documents

Real Documents

Conclusion