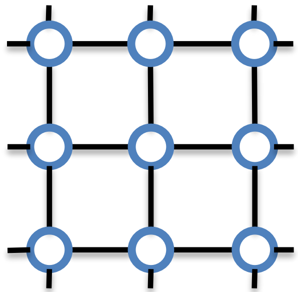
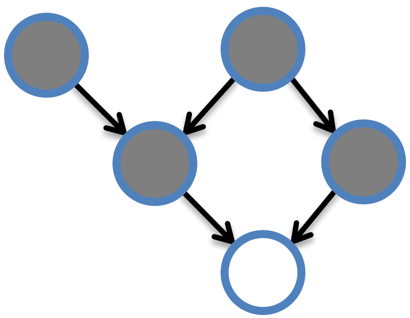
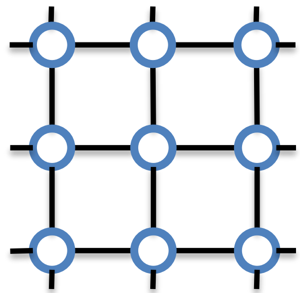
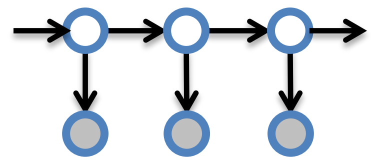
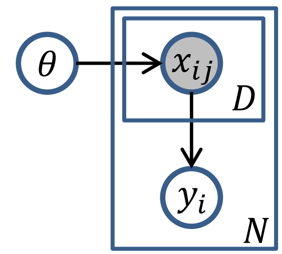
Scaling Up
Graphical Model Inference



Graphical Models

- View observed data and unobserved properties as *random variables*
- Graphical Models: compact graph-based encoding of probability distributions (high dimensional, with complex dependencies)

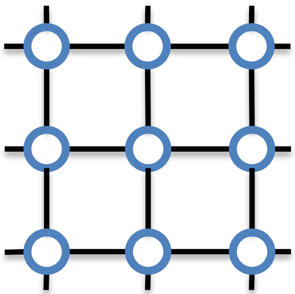
- Generative/discriminative/hybrid, un-,semi- and supervised learning
  - Bayesian Networks (directed), Markov Random Fields (undirected), hybrids, extensions, etc. HMM, CRF, RBM, M^{3}N, HMRF, etc.
- Enormous research area with a number of excellent tutorials
  - [J98], [M01], [M04], [W08], [KF10], [S11]

Graphical Model Inference

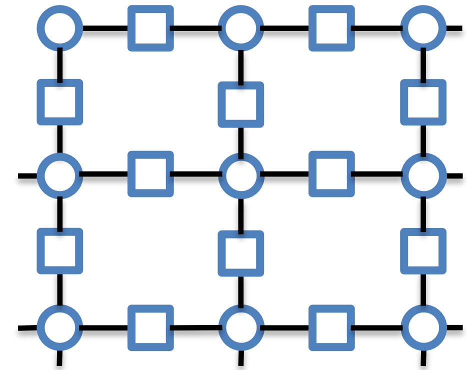
- Key issues:
  - **Representation**: syntax and semantics (directed/undirected,variables/factors,..)
  - **Inference**: computing probabilities and most likely assignments/explanations
  - **Learning**: of model parameters based on observed data. *Relies on inference!*
- Inference is NP-hard (numerous results, incl. approximation hardness)
- Exact inference: works for very limited subset of models/structures
  - E.g., chains or low-treewidth trees
- Approximate inference: highly computationally intensive
  - Deterministic: variational, loopy belief propagation, expectation propagation
  - Numerical sampling (Monte Carlo): Gibbs sampling

Inference in Undirected Graphical Models

- Factor graph representation



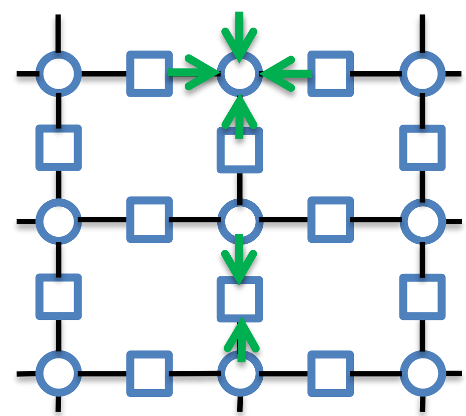
$$p(x_1,..,x_n) = \frac{1}{Z}\prod_{x_j \in N(x_i)} \psi_{ij}(x_1,x_2)$$



- Potentials capture compatibility of related observations
  - e.g., $\psi(x_i,x_j) = \exp(-b|x_i - x_j|)$
- Loopy belief propagation = message passing
  - iterate (read, update, send)

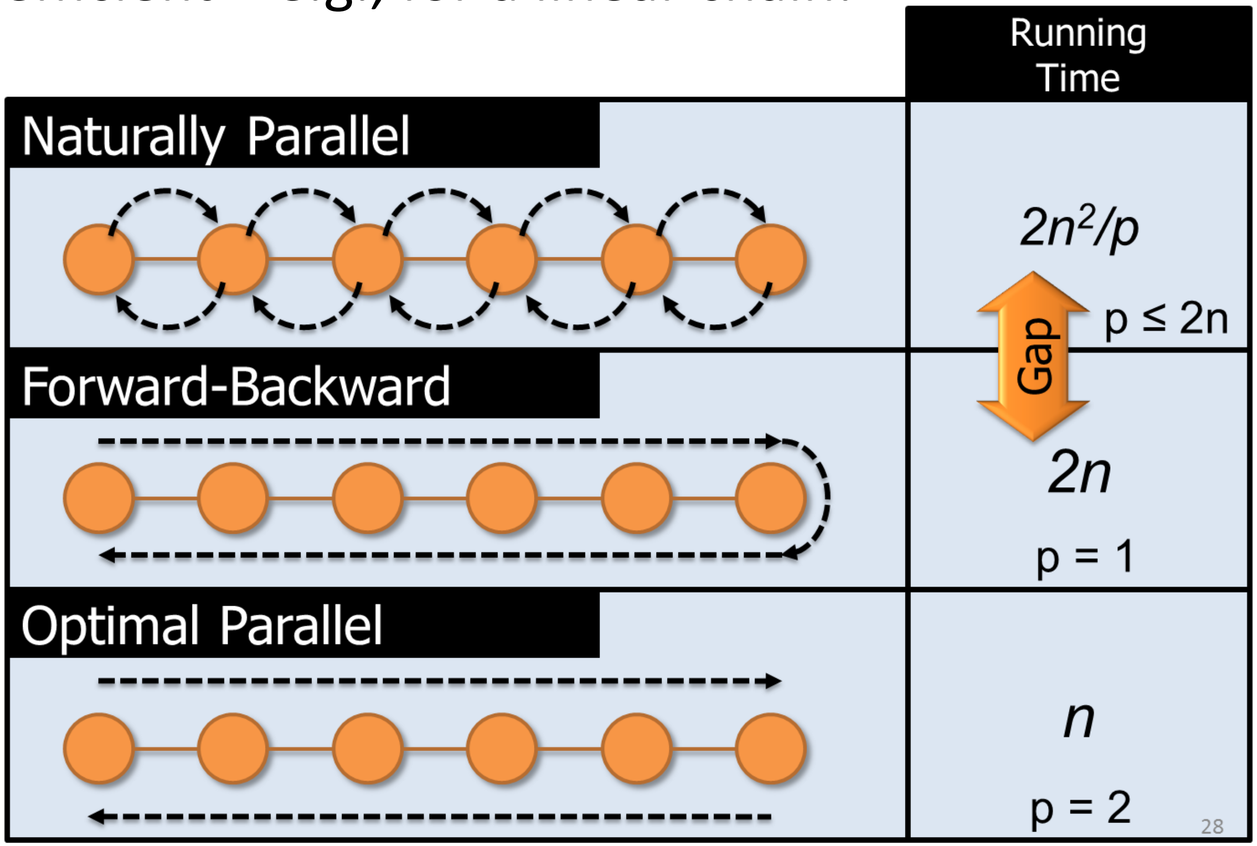
$$m_{X_i \to \psi_{\mathbf{A}}}(x_i) \propto \prod_{\psi_{\mathbf{B}} \in \mathbf{N}[X_i] \setminus \psi_{\mathbf{A}}} m_{\psi_{\mathbf{B}} \to X_i}(x_i)$$

$$m_{\psi_{\mathbf{A}} \to X_i}(x_i) \propto \sum_{\mathbf{x}_{\mathbf{A}} \setminus x_i} \psi_{\mathbf{A}}(\mathbf{x}_{\mathbf{A}}) \prod_{X_k \in \mathbf{N}[\psi_{\mathbf{A}}] \setminus X_i} m_{X_k \to \psi_{\mathbf{A}}}(x_k)$$



Synchronous Loopy BP

- Natural parallelization: associate a processor to every node
  - Simultaneous receive, update, send
- Inefficient – e.g., for a linear chain:

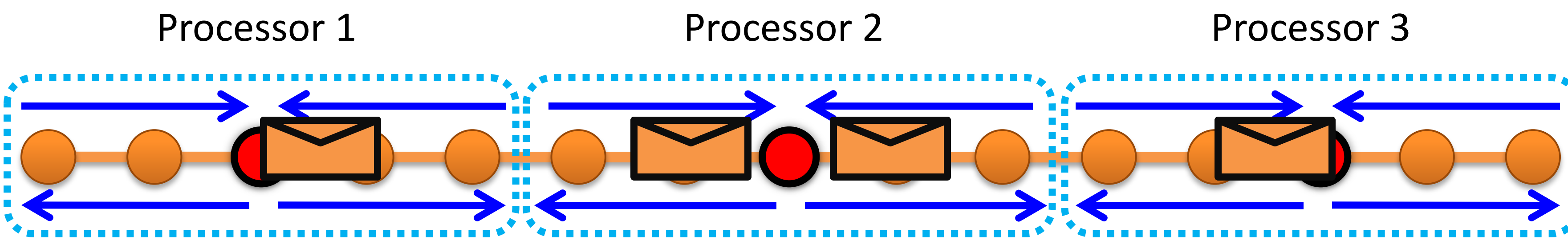


$2n/p$ time per iteration
n iterations to converge

[SUML-Ch10]

Optimal Parallel Scheduling

- Partition, local forward-backward for center, then cross-boundary



Synchronous

$$O\left(\frac{n\tau_\epsilon}{p}\right)$$

Optimal Schedule

$$O\left(\frac{n}{p} + \tau_\epsilon\right)$$

llel Component

Gap

equential Component

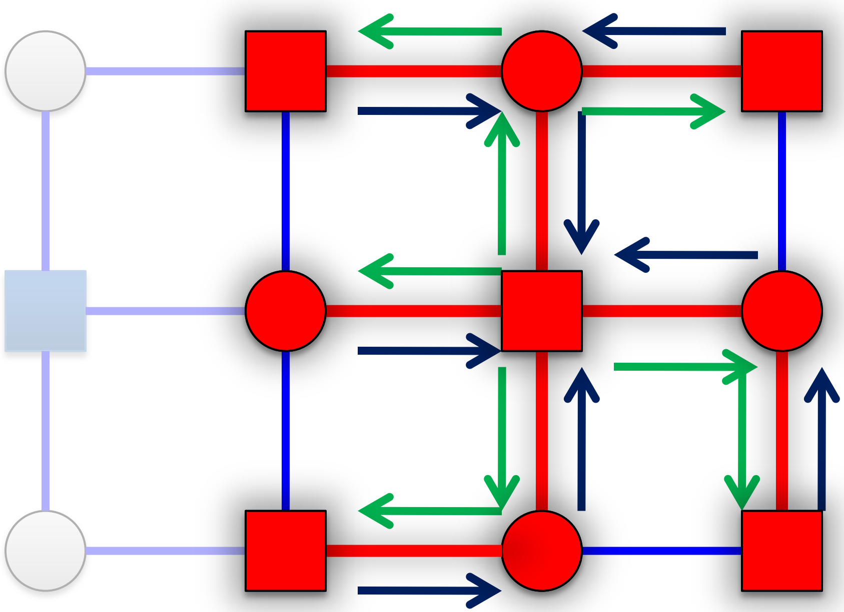
Splash: Generalizing Optimal Chains

1) Select root, grow fixed-size BFS Spanning tree

2) Forward Pass computing all messages at each vertex

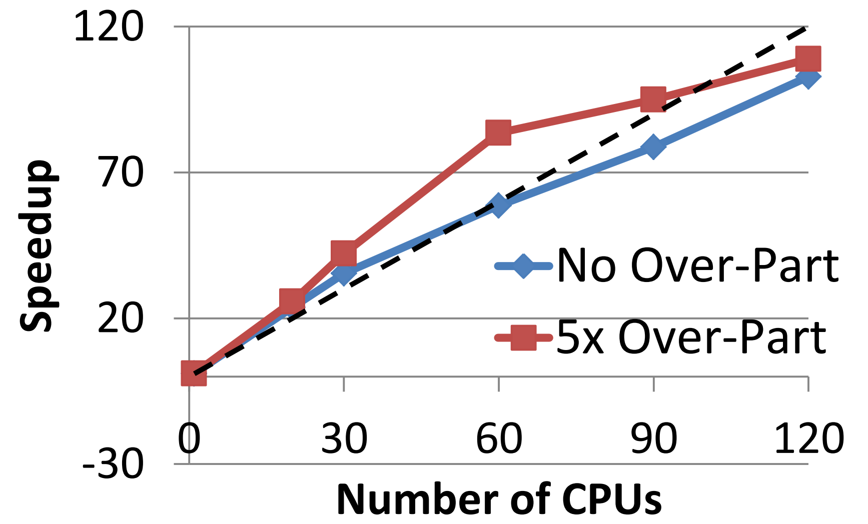
3) Backward Pass computing all messages at each vertex

- Parallelization:
  - Partition graph
    - Maximize computation, minimize communication
    - Over-partition and randomly assign
  - Schedule multiple Splashes
    - Priority queue for selecting root
    - Belief residual: cumulative change from inbound messages
    - Dynamic tree pruning

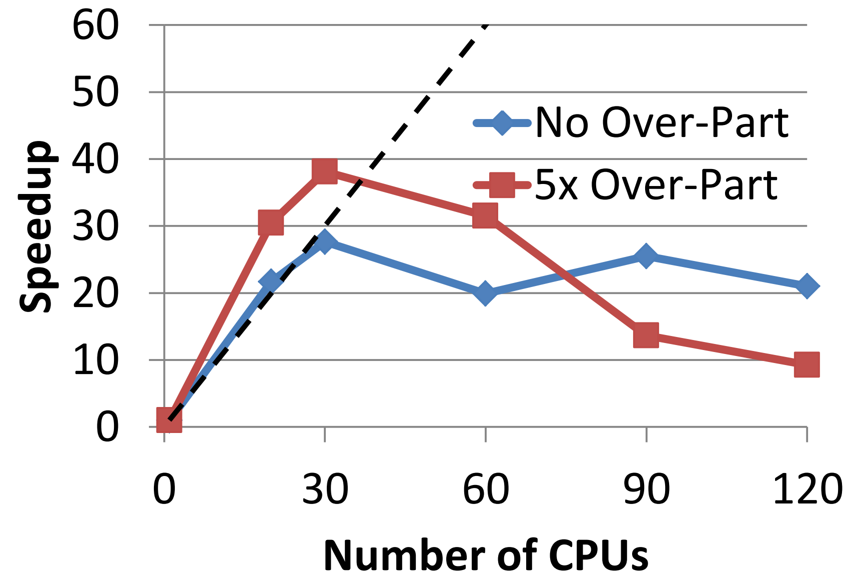


DBRSplash: MLN Inference Experiments

- Experiments: MLN Inference
- 8K variables, 406K factors
- Single-CPU runtime: 1 hour
- Cache efficiency critical



- 1K variables, 27K factors
- Single-CPU runtime: 1.5 minutes
- Network costs limit speedups



Topic Models

- Goal: unsupervised detection of topics in corpora
  - Desired result: topic mixtures, per-word and per-document topic assignments

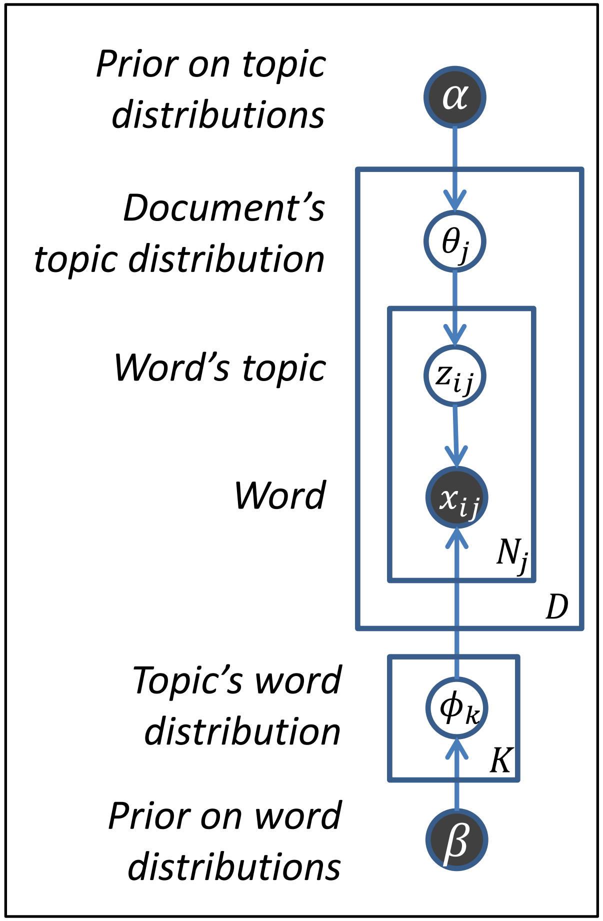
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

[B+03]

Directed Graphical Models: Latent Dirichlet Allocation [B+03, SUML-Ch11]

- Generative model for document collections
  - K topics, topic k: $\text{Multinomial}(\phi_k)$ over words
  - D documents, document j:
    - Topic distribution $\theta_j \sim \text{Dirichlet}(\alpha)$
    - N_j words, word x_{ij}:
      - Sample topic $z_{ij} \sim \text{Multinomial}(\theta_j)$
      - Sample word $x_{ij} \sim \text{Multinomial}\left(\phi_{z_{ij}}\right)$
- Goal: infer posterior distributions
  - Topic word mixtures $\{\phi_k\}$
  - Document mixtures $\{\theta_j\}$
  - Word-topic assignments $\{z_{ij}\}$



Gibbs Sampling

- Full joint probability

$$p(\theta, z, \phi, x|\alpha, \beta) = \prod_{k=1..K} p(\phi_k|\beta) \prod_{j=1..D} p(\theta_j|\alpha) \prod_{j=1..N_j} p\left(z_{ij}\middle|\theta_j\right) p(x_{ij}|\phi_{z_{ij}})$$

- Gibbs sampling: sample ϕ, θ, z independently
- Problem: slow convergence (a.k.a. mixing)
- Collapsed Gibbs sampling
  - Integrate out ϕ and θ analytically

$$p(z|x, d, \alpha, \beta) \propto \frac{N'_{xz} + \beta}{\sum_x (N'_{xz} + \beta)} \frac{N'_{dz} + \alpha}{\sum_z (N'_{dz} + \alpha)}$$

  - Until convergence:
    - resample $p\left(z_{ij}\middle|x_{ij}, \alpha, \beta\right)$,
    - update counts: N_z, N_{zd}, N_{xz}

Parallel Collapsed Gibbs Sampling [SUML-Ch11]

- Synchronous version (MPI-based):
  - Distribute documents among p machines
  - Global topic and word-topic counts N_z, N_{wz}
  - Local document-topic counts N_{dz}
  - After each local iteration, AllReduce N_z, N_{wz}

- Asynchronous version: gossip (P2P)
  - Random pairs of processors exchange statistics upon pass completion
  - Approximate global posterior distribution (experimentally not a problem)
  - Additional estimation to properly account for previous counts from neighbor

Parallel Collapsed Gibbs Sampling [SN10,S11]

- Multithreading to maximize concurrency
  - Parallelize both *local* and *global* updates of N_{xz} counts
  - Key trick: N_z and N_{xz} are effectively constant for a given document
    - No need to update continuously: update once per-document in a separate thread
    - Enables multithreading the samplers
  - Global updates are asynchronous -> no blocking

	Google LDA	Mallet	Irvine'08	Irvine'09	Yahoo LDA
Multicore	no	yes	yes	yes	yes
Cluster	MPI	no	MPI	point 2 point	memcached
State table	dictionary split	separate sparse	separate	separate	joint sparse
Schedule	synchronous exact	synchronous exact	synchronous exact	asynchronous approximate messages	asynchronous exact

[S11]

Scaling Up Graphical Models: Conclusions

- Extremely high parallelism is achievable, but variance is high
  - Strongly data dependent
- Network and synchronization costs can be explicitly accounted for in algorithms
- Approximations are essential to removing barriers
- Multi-level parallelism allows maximizing utilization
- Multiple caches allow super-linear speedups

References

[SUML-Ch11] Arthur Asuncion, Padhraic Smyth, Max Welling, David Newman, Ian Porteous, and Scott Triglia. Distributed Gibbs Sampling for Latent Variable Models. In “Scaling Up Machine Learning”, Cambridge U. Press, 2011.

[B+03] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003.

[B11] D. Blei. Introduction to Probabilistic Topic Models. Communications of the ACM, 2011.

[SUML-Ch10] J. Gonzalez, Y. Low, C. Guestrin. Parallel Belief Propagation in Factor Graphs. In “Scaling Up Machine Learning”, Cambridge U. Press, 2011.

[KF10] D. Koller and N. Friedman Probabilistic graphical models. MIT Press, 2010.

[M01] K. Murphy. An introduction to graphical models, 2001.

[M04] K. Murphy. Approximate inference in graphical models. AAAI Tutorial, 2004.

[S11] A.J. Smola. Graphical models for the Internet. MLSS Tutorial, 2011.

[SN10] A.J. Smola, S. Narayanamurthy. An Architecture for Parallel Topic Models. VLDB 2010.

[W08] M. Wainwright. Graphical models and variational methods. ICML Tutorial, 2008.