# Composite Likelihood and Particle Filtering Methods for Network Estimation

Arthur Asuncion
5/25/2010

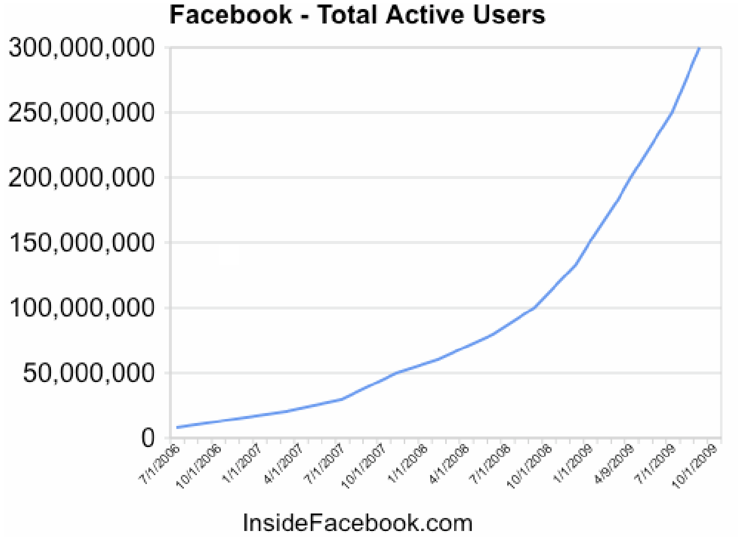
Joint work with:
Qiang Liu, Alex Ihler, Padhraic Smyth

# Roadmap

- **Exponential random graph models (ERGMs)**
- **Previous approximate inference techniques:**
  - MCMC maximum likelihood estimation (MCMC-MLE)
  - Maximum pseudolikelihood estimation (MPLE)
  - Contrastive divergence (CD)
- **Our new techniques:**
  - Composite likelihoods and blocked contrastive divergence
  - Particle-filtered MCMC-MLE

# Why approximate inference?

- Online social networks can have hundreds of millions of users:




- Even moderately-sized networks can be difficult to model
  - e.g. email networks for a corporation with thousands of employees

- Models themselves are becoming more complex
  - Curved ERGMs, hierarchical ERGMs
  - Dynamic social network models

# Exponential Random Graph Models

- Exponential Random Graph Model (ERGM):

Parameters to learn

Network statistics
(e.g. # edges, triangles, etc.)

$$P(Y = y|\theta) = \frac{\exp\{\theta^t s(y)\}}{Z(\theta)}$$

A particular graph configuration

Partition function
(intractable to compute)

$$Z(\theta) = \sum_y \exp\{\theta^t s(y)\}$$

- **Task:** Estimate the set of parameters θ under which the observed network, Y, is most likely.

- **Our goal:** Perform this parameter estimation in a computationally efficient and scalable manner.

# A Spectrum of Techniques

MCMC-MLE — ?? — MPLE

Accurate but Slow

Composite Likelihood, Contrastive Divergence

Inaccurate but Fast

Also see Ruth Hummel's work on partial stepping for ERGMs:
http://www.ics.uci.edu/~duboisc/muri/spring2009/Ruth.pdf

# MCMC-MLE

[Geyer, 1991]

- Maximum likelihood estimation: $\theta_{ML} \equiv \arg\max_{\theta} \mathcal{L}(\theta|y)$
- MLE has nice properties: asymptotically unbiased, efficient
- **Problem:** Evaluating the partition function. **Solution:** Markov Chain Monte Carlo.

$$\mathcal{L}(\theta|y) = \log \prod_{i}^{N} p(y^i|\theta)$$

$$= \sum_{i}^{N} \theta s(y^i) - N \log Z(\theta)$$

$$= \sum_{i}^{N} \theta s(y^i) - N \log \left[ Z(\theta_0) \sum_{y} \exp\{(\theta - \theta_0) s(y)\} p(y|\theta_0) \right]$$

// Equation to transform partition function

$$\propto \frac{1}{N} \sum_{i}^{N} \theta s(y^i) - \log \sum_{y} \exp\{(\theta - \theta_0) s(y)\} p(y|\theta_0)$$

$$\approx \frac{1}{N} \sum_{i}^{N} \theta s(y^i) - \log \frac{1}{S} \sum_{s} \exp\{(\theta - \theta_0) s(y^s)\}$$

// Markov Chain Monte Carlo approximation: $y^s \sim p(y \mid \theta_0)$

$$P(Y = y|\theta) = \frac{\exp\{\theta^t s(y)\}}{Z(\theta)}$$

# Gibbs sampling for ERGMs

Since

$$\log \frac{P(Y_j = 1 | y_{\neg j}, \theta)}{P(Y_j = 0 | y_{\neg j}, \theta)} = \theta^t \Delta s(y)_j$$

Change statistics

then

$$P(Y_j = 1 | y_{\neg j}, \theta) = \sigma(\theta^t \Delta s(y)_j)$$

Use this conditional probability to perform Gibbs sampling scans until the chain converges.

# MPLE

[Besag, 1974]

- Maximum pseudolikelihood estimation:

$$\theta_{PL} \equiv \arg\max_{\theta} \mathcal{PL}(\theta|y)$$

where

$$\mathcal{PL}(\theta|y) = \log \prod_{i}^{N} \prod_{j}^{M} p(y_j^i | y_{\neg j}^i, \theta)$$

- Computationally efficient (for ERGMs, reduces to logistic regression)
- Can be inaccurate

# Composite Likelihoods (CL)
[Lindsay, 1988]

- Composite Likelihood (generalization of PL):

$$\mathcal{CL}(\theta|y) = \log \prod_{i}^{N} \prod_{c}^{C} p(y^i_{A_c}|y^i_{B_c}, \theta)$$

Only restriction: $A_c \cap B_c$ is null

- Consider 3 variables $Y_1$, $Y_2$, $Y_3$. Here are some possible CL's:

$$P(Y_1, Y_2|Y_3, \theta)P(Y_2|Y_1, \theta)$$

$$P(Y_2, Y_3|\theta)P(Y_1|\theta)$$

$$P(Y_1, Y_3|\theta)P(Y_1|Y_2, \theta)$$

- **MCLE:** Optimize CL with respect to θ

# Contrastive Divergence (CD)

[Hinton, 2002]

- A popular machine learning technique, used to learn deep belief networks and other models
- (Approximately) optimizes the difference between two KL divergences through gradient descent.

| | |
|---|---|
| **CD-∞** | = MLE |
| **CD-n** | = A technique between MLE and MPLE |
| **CD-1** | = MPLE |
| **BCD** | = MCLE (also between MLE and MPLE) |

# Contrastive Divergence (CD-∞)

$$\mathcal{L}(\theta|y) = \log \prod_i^N p(y^i|\theta)$$

$$\propto \frac{1}{N} \sum_i^N \theta s(y^i) - \log Z(\theta)$$

$$\frac{d\mathcal{L}(\theta|y)}{d\theta} = \langle s(y) \rangle_0 - \frac{1}{Z(\theta)} \frac{dZ(\theta)}{d\theta}$$

$$= \langle s(y) \rangle_0 - \frac{1}{Z(\theta)} \sum_y \frac{d}{d\theta} \exp\{\theta s(y)\}$$

$$= \langle s(y) \rangle_0 - \frac{1}{Z(\theta)} \sum_y s(y) \exp\{\theta s(y)\}$$

$$= \langle s(y) \rangle_0 - \sum_y s(y) p(y|\theta)$$

$$= \langle s(y) \rangle_0 - \langle s(y) \rangle_\infty$$

$$\approx \langle s(y) \rangle_0 - \frac{1}{S} \sum_s s(y^s)$$

$$P(Y = y|\theta) = \frac{\exp\{\theta^t s(y)\}}{Z(\theta)}$$

$$\langle s(y) \rangle_0 = \frac{1}{N} \sum_i^N s(y^i)$$

$$Z(\theta) = \sum_y \exp\{\theta^t s(y)\}$$

CD-∞ -- MCMC is run for an "infinite" # of steps

Monte Carlo approximation: $y^s \sim p(y \mid \theta)$

# Contrastive Divergence (CD-n)

- Run MCMC chains for *n* steps only (e.g. n=10):

$$\frac{d\mathcal{L}(\theta|y)}{d\theta} \approx \langle s(y) \rangle_0 - \langle s(y) \rangle_n$$

- **Intuition:** We don't need to fully burn in the chain to get a good rough estimate of the gradient.

- Initialize the chains from the data distribution to stay close to the true modes.

# Contrastive Divergence (CD-1) and connection to MPLE

[Hyvärinen, 2006]

$$P(Y=y|\theta) = \frac{\exp\{\theta^t s(y)\}}{Z(\theta)}$$

$$\mathcal{PL}(\theta|y) = \log \prod_i^N \prod_j^M p(y_j^i | y_{\neg j}^i, \theta)$$

$$= M \sum_i^N \log p(y^i|\theta) - \sum_i^N \sum_j^M \log p(y_{\neg j}^i|\theta)$$

Use definition of conditional probability

$$= M \sum_i^N \log \frac{\exp\{\theta s(y^i)\}}{Z(\theta)} - \sum_i^N \sum_j^M \log \sum_{y_j} \frac{\exp\{\theta s(y_{\neg j}^i, y_j)\}}{Z(\theta)}$$

Z(θ) will cancel

$$\propto \frac{1}{N} \sum_i^N \theta s(y^i) - \frac{1}{N} \sum_i^N \frac{1}{M} \sum_j^M \log \sum_{y_j} \exp\{\theta s(y_{\neg j}^i, y_j)\}$$

$$\frac{d\mathcal{PL}(\theta|y)}{d\theta} = \langle s(y) \rangle_0 - \frac{1}{N} \sum_i^N \frac{1}{M} \sum_j^M \frac{1}{\sum_{y_j} \exp\{\theta s(y_{\neg j}^i, y_j)\}} \sum_{y_j} s(y_{\neg j}^i, y_j) \exp\{\theta s(y_{\neg j}^i, y_j)\}$$

$$= \langle s(y) \rangle_0 - \frac{1}{N} \sum_i^N \frac{1}{M} \sum_j^M \frac{1}{p(y_{\neg j}^i|\theta)} \sum_{y_j} s(y_{\neg j}^i, y_j) p(y_{\neg j}^i, y_j|\theta)$$

$$= \langle s(y) \rangle_0 - \frac{1}{N} \sum_i^N \frac{1}{M} \sum_j^M \sum_{y_j} s(y_{\neg j}^i, y_j) p(y_j | y_{\neg j}^i, \theta)$$

$$= \langle s(y) \rangle_0 - \langle s(y) \rangle_1$$

$$\approx \langle s(y) \rangle_0 - \frac{1}{S} \sum_s s(y^s)$$

Monte Carlo approximation:

1. Sample y from data distribution
2. Pick an index j at random
3. Sample $y_j$ from $p(y_j \mid y_{\neg j}, \theta)$

This is random-scan Gibbs sampling.

CD-1 with random scan Gibbs sampling is stochastically performing MPLE!

# Blocked Contrastive Divergence (BCD) and connections to MCLE

- Derivation is very similar to previous slide (simply change j→c, $y_j \to y_{Ac}$):

$$\mathcal{CL}(\theta|y) = \log \prod_i^N \prod_c^C p(y^i_{A_c}|y^i_{\neg A_c}, \theta)$$

We focus on "conditional" composite likelihoods

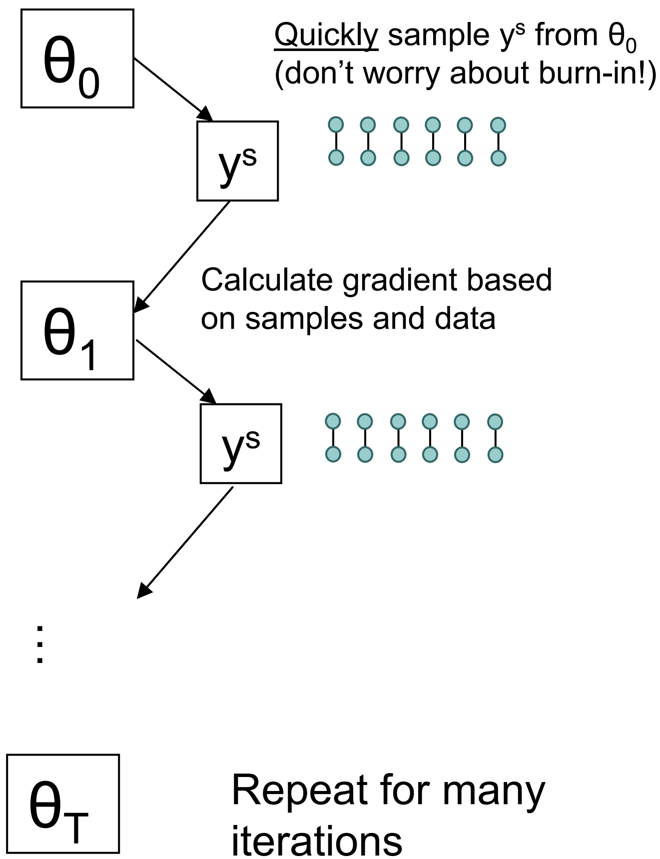
$$\frac{d\mathcal{CL}(\theta|y)}{d\theta} = \langle s(y) \rangle_0 - \frac{1}{N}\sum_i^N \frac{1}{C}\sum_c^C \sum_{y_{A_c}} s(y^i_{\neg A_c}, y_{A_c}) p(y_{A_c}|y^i_{\neg A_c}, \theta)$$

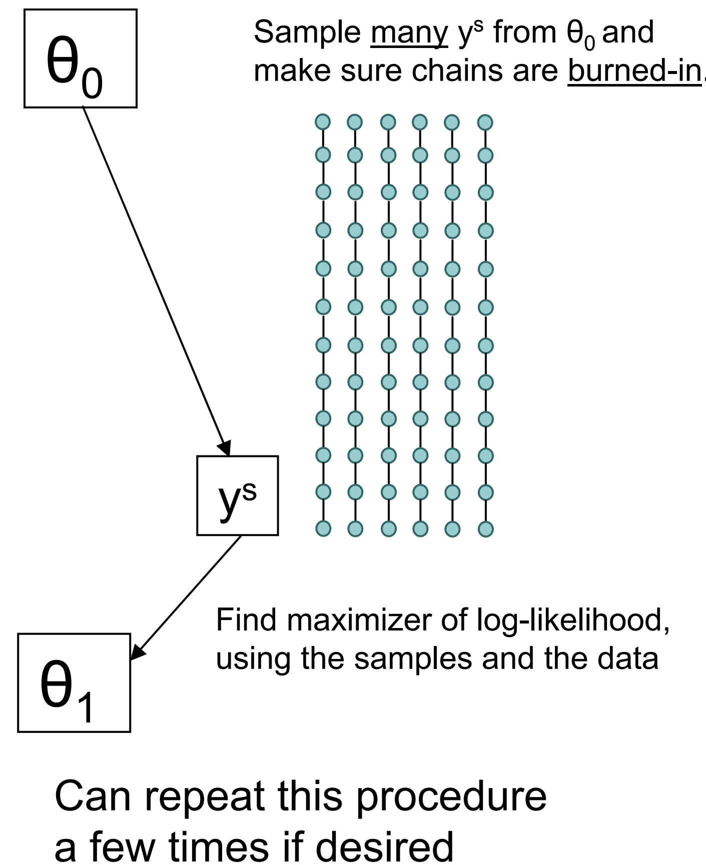
Monte Carlo approximation:

1. Sample y from data distribution
2. Pick an index c at random
3. Sample $y_{Ac}$ from $p(y_{Ac} \mid y_{\neg Ac}, \theta)$

CD with random-scan blocked Gibbs sampling corresponds to MCLE!

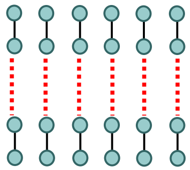
# CD vs. MCMC-MLE



# Some CD tricks

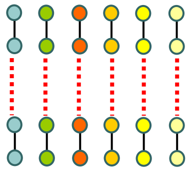
- **Persistent CD** [Younes, 2000; Tieleman & Hinton, 2008]



Use samples at the ends of the chains at the previous iteration to initialize the chains at the next CD iteration.

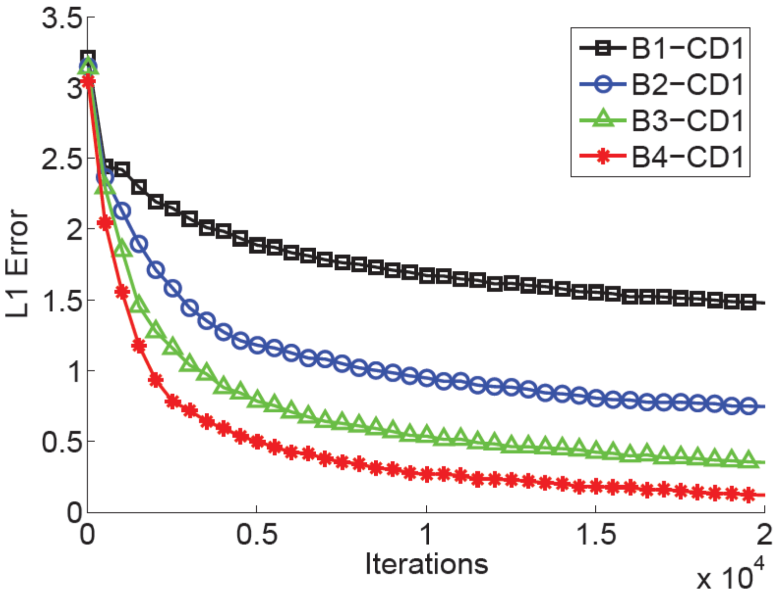
- **Herding** [Welling, 2009]. Instead of performing Gibbs sampling, perform iterated conditional modes (ICM).

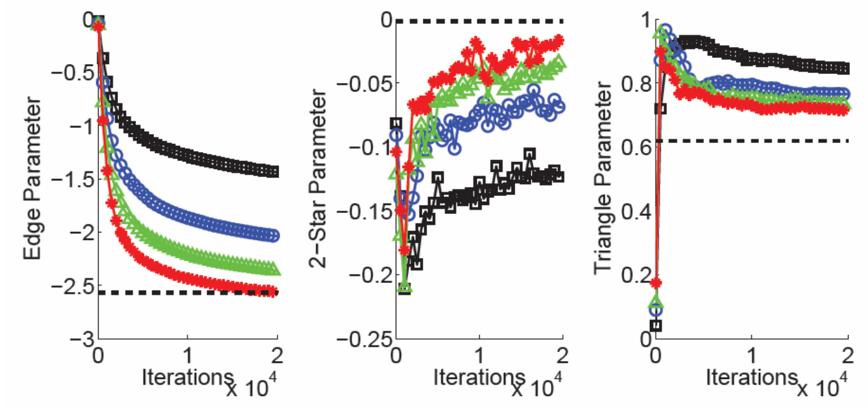
- **Persistent CD with tempered transitions** ("parallel tempering") [Desjardins, Courville, Bengio, Vincent, Delalleau, 2009].



Run persistent chains at different temperatures and allow them to communicate (to improve mixing)

# Blocked CD (BCD) on ERGMs




Lazega subset (36 nodes; 630 edges)
Triad model: edges + 2-stars + triangles

“Ground truth” parameters were obtained by running MCMC-MLE using statnet.

# Particle Filtered MCMC-MLE

- MCMC-MLE uses importance sampling to estimate the log-likelihood gradient:

Data — Sample from $P(y|\theta_0)$

$$\frac{d\mathcal{L}(\theta|y)}{d\theta} \approx \frac{1}{N}\sum_{i=1}^{N} s(y^i) - \frac{1}{S}\sum_{s=1}^{S} w^s s(y_0^s)$$

Importance weight: $P(y_0|\theta) / P(y_0|\theta_0)$

- **Main Idea:** Replace importance sampling with sequential importance resampling (SIR), also known as particle filtering

# MCMC-MLE vs. PF-MCMC-MLE

Obtain samples from $\theta_0$

**Algorithm 1** MCMC-MLE

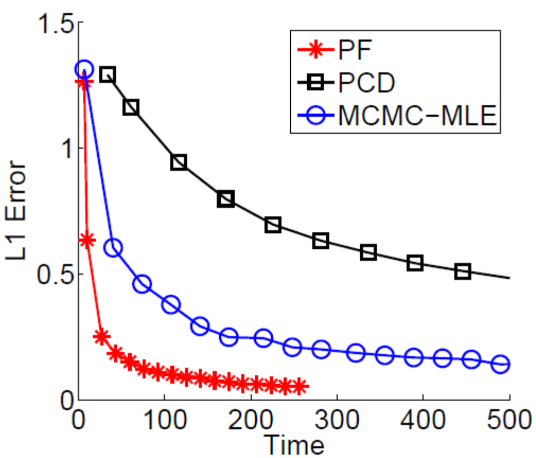
Initialize $\theta_0$
Sample $\{x^s\} \sim p(x|\theta_0)$
$\theta_1 \leftarrow \theta_0$
**for** $i = 1$ to max-iterations (or convergence) **do**
  Calculate $\{w^s\}$ via eq. 6, using $\theta_i, \theta_0, \{x^s\}$
  Calculate $\nabla\tilde{L}$ via eq. 5, using $\{w^s\}, \{x^s\}$
  $\theta_{i+1} \leftarrow \theta_i + \eta\nabla\tilde{L}$
**end for**

**Algorithm 2** Particle Filtered MCMC-MLE

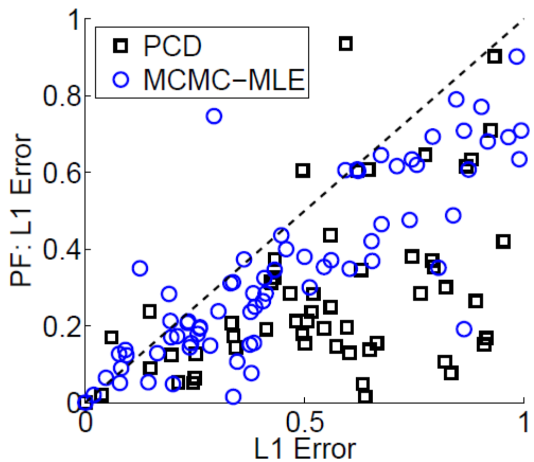
Initialize $\theta_0$
Sample $\{x^s\} \sim p(x|\theta_0)$
$\theta_1 \leftarrow \theta_0$
**for** $i = 1$ to max-iterations (or convergence) **do**
  Calculate $\{w^s\}$ via eq. 10, using $\theta_i, \theta_{i-1}, \{x^s\}$
  **if** ESS($\{w^s\}$) $<$ threshold **then**
    Resample $\{x^s\}$ in proportion to $\{w^s\}$
    $\{w^s\} \leftarrow 1$
    Rejuvenate $\{x^s\}$ for $n$ MCMC steps, using $\theta_i$
  **end if**
  Calculate $\nabla\tilde{L}$ via eq. 5, using $\{w^s\}, \{x^s\}$
  $\theta_{i+1} \leftarrow \theta_i + \eta\nabla\tilde{L}$
**end for**

PF-MCMC-MLE:
- calculate ESS to monitor "health" of particles.
- resample and rejuvenate particles to prevent weight degeneracy.

# Some ERGM experiments




(a) L1 error over time.




(b) L1 errors, for 100 models.

Particle filtered MCMC-MLE is faster than MCMC-MLE and persistent CD, without sacrificing accuracy.

Synthetic data used (randomly generated).
Network statistics: # edges, # 2-stars, # triangles.

# Conclusions

- A unified picture of these estimation techniques exists:
  - MLE, MCLE, MPLE
  - CD-$\infty$, BCD, CD-1
  - MCMC-MLE, PF-MCMC-MLE, PCD
- Some algorithms are more efficient/accurate than others:
  - Composite likelihoods allow for a principled tradeoff.
  - Particle filtering can be used to improve MCMC-MLE.
- These methods can be applied to network models (ERGMs) and more generally to exponential family models.

# References


- "Learning with Blocks: Composite Likelihood and Contrastive Divergence." Asuncion, Liu, Ihler, Smyth. AI & Statistics, 2010.
- "Particle Filtered MCMC-MLE with Connections to Contrastive Divergence." Asuncion, Liu, Ihler, Smyth. Intl Conference on Machine Learning, 2010.