Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

# Highly Scalable Inference Techniques for Mix-Membership Block Models

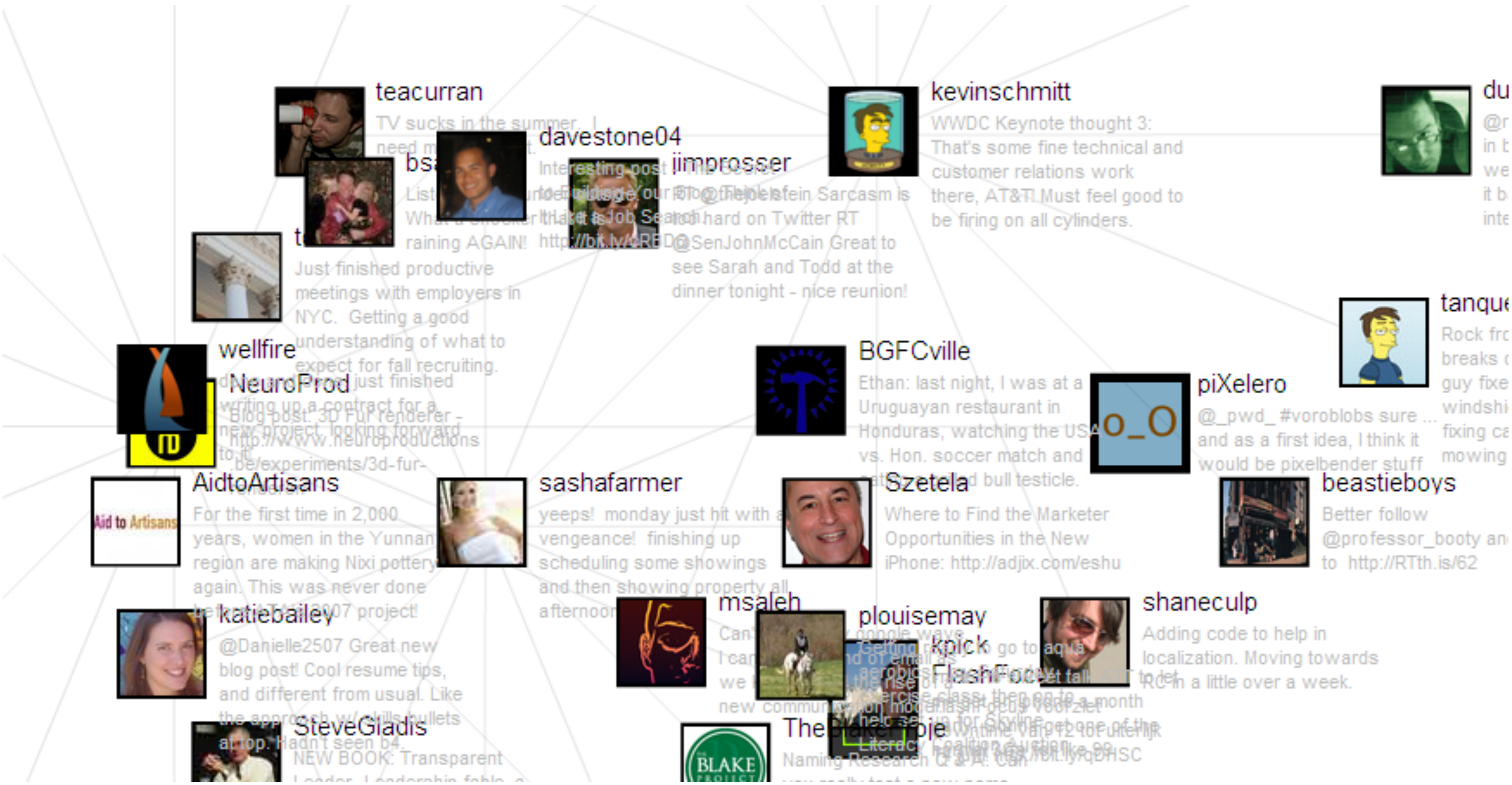
-

Eric Xing

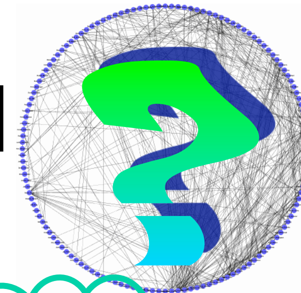
epxing@cs.cmu.edu

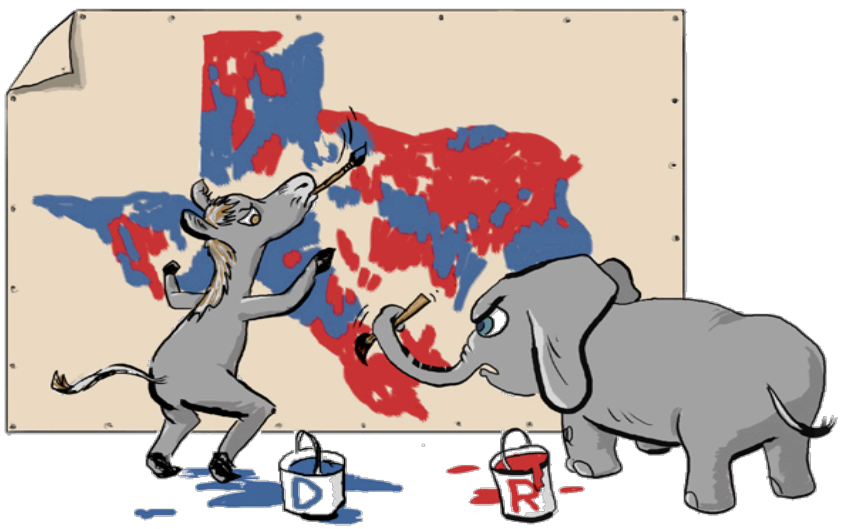
Machine Learning Dept./Language Technology Inst./Computer Science Dept.
Carnegie Mellon University

# Social networks



# Estimating Evolving "Latent" Social Networks





**Can I get his vote?**

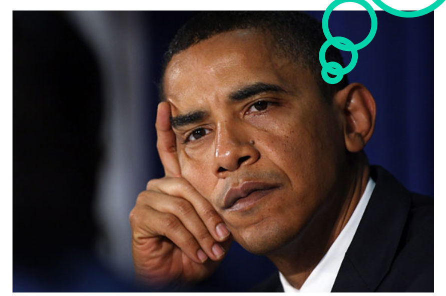
Corporativity,

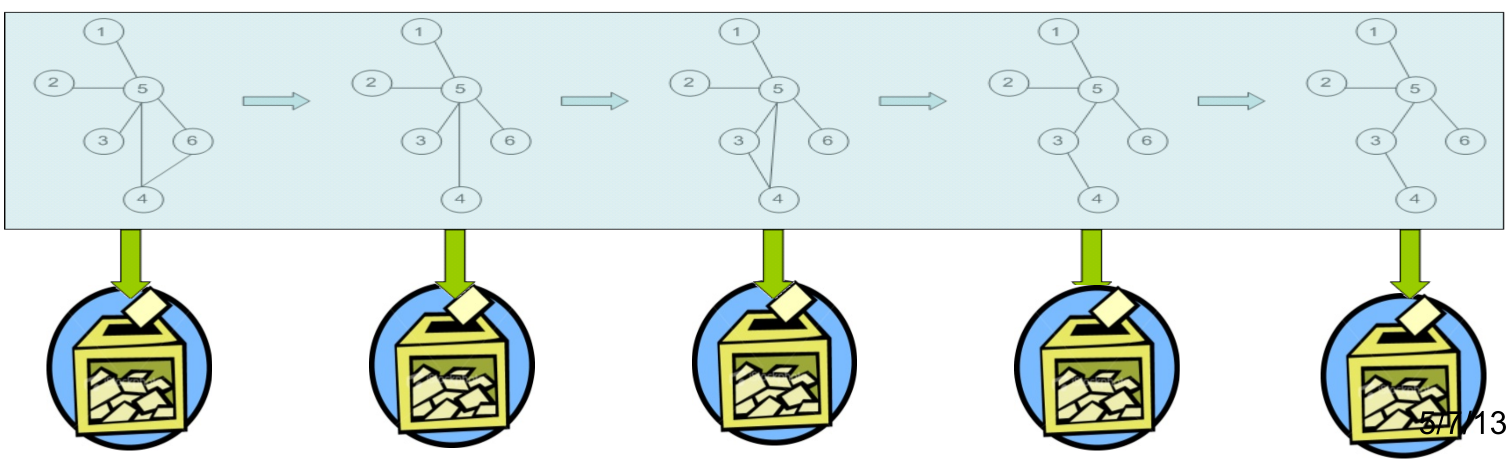
Antagonism,

Cliques,

…

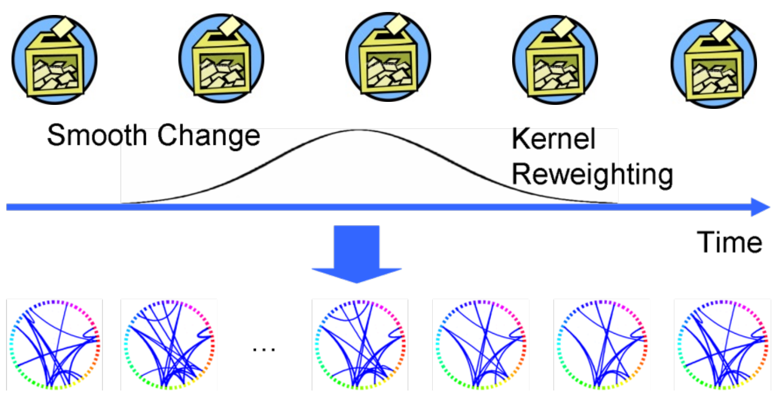
over time?





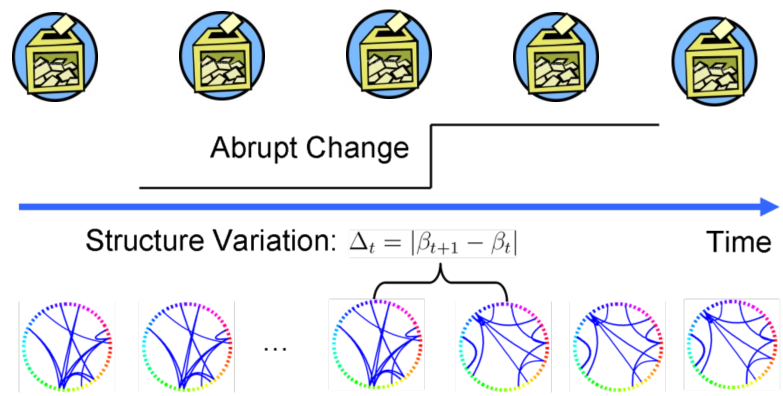
# Asymptotic-consistent graph estimation algorithms [Kolar and Xing, 09,12]

## KELLER



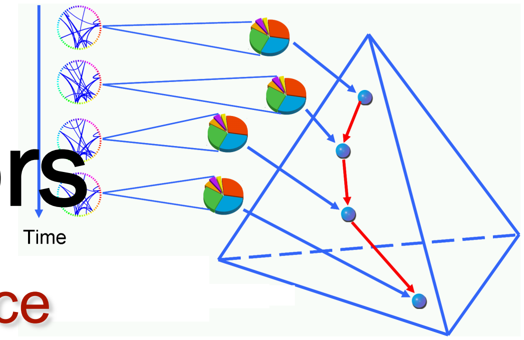

**Smoothly evolving graphs**

## TESLA



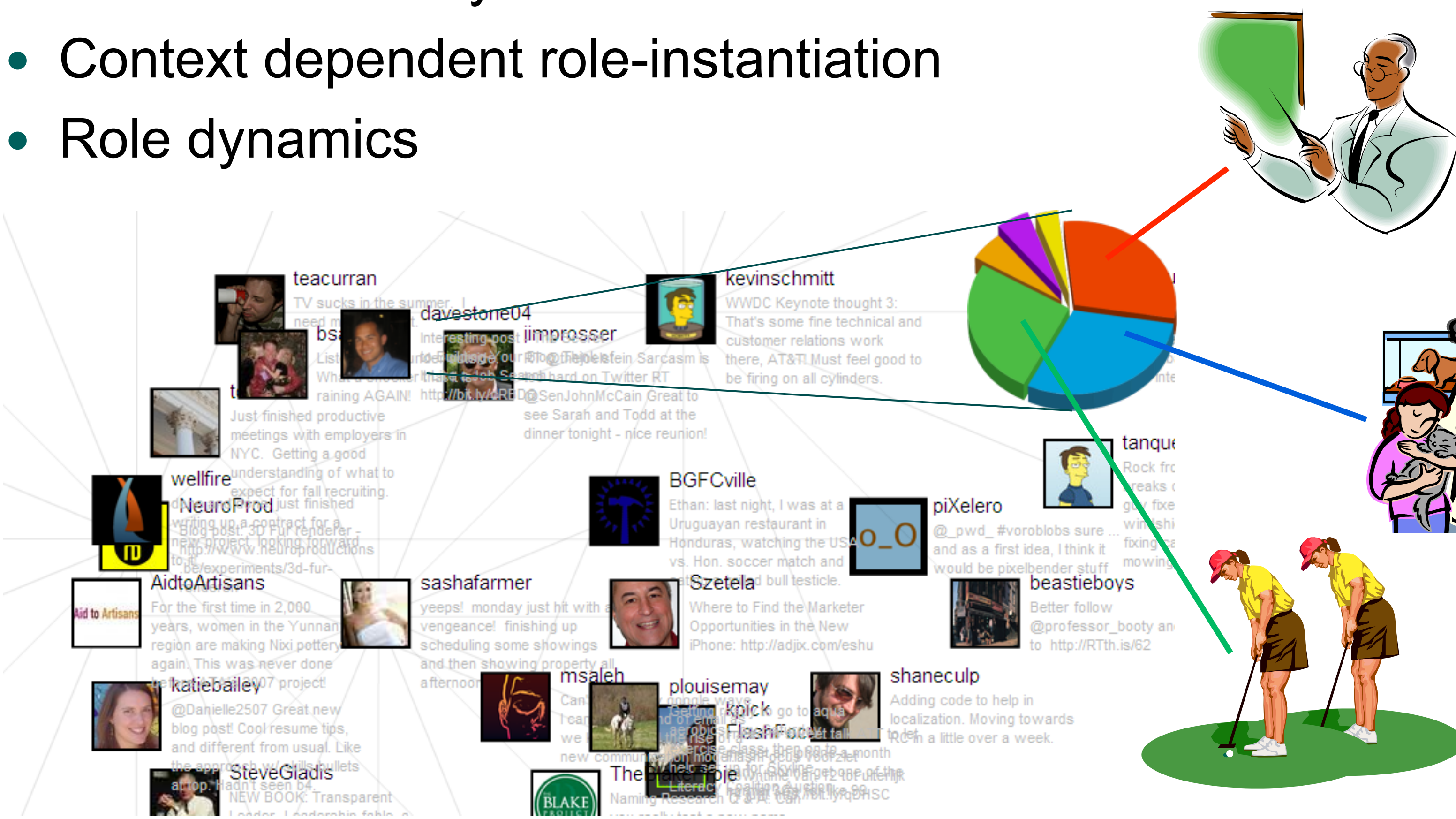

**Abruptly evolving graphs**

$$\mathbb{P}\left[\hat{G}(\lambda_n) \neq G\right] = \mathcal{O}\left(\exp\left(-Cn^{\epsilon}\right)\right) \to 0$$



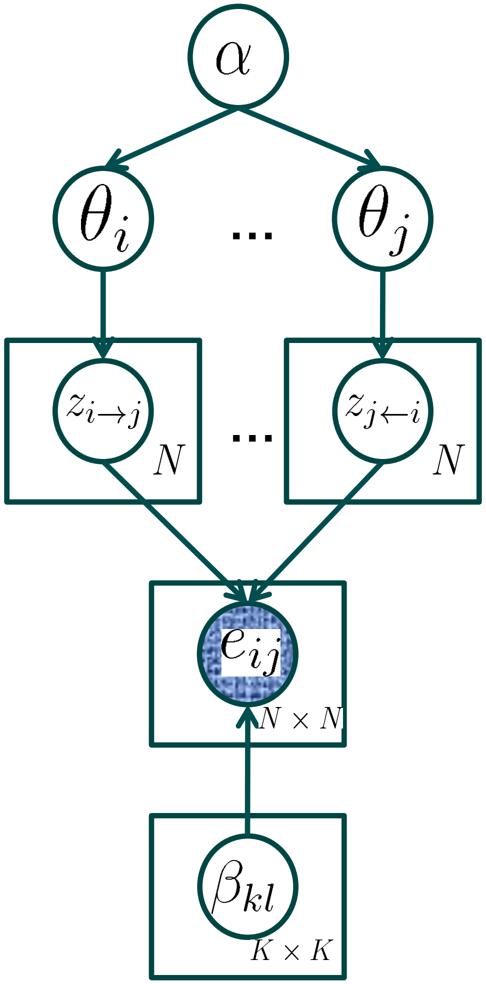

# Mixed Membership of Actors

- Micro-inference vs. Meso- or Macro-inference
- Multi-role of every node
- Context dependent role-instantiation
- Role dynamics



SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

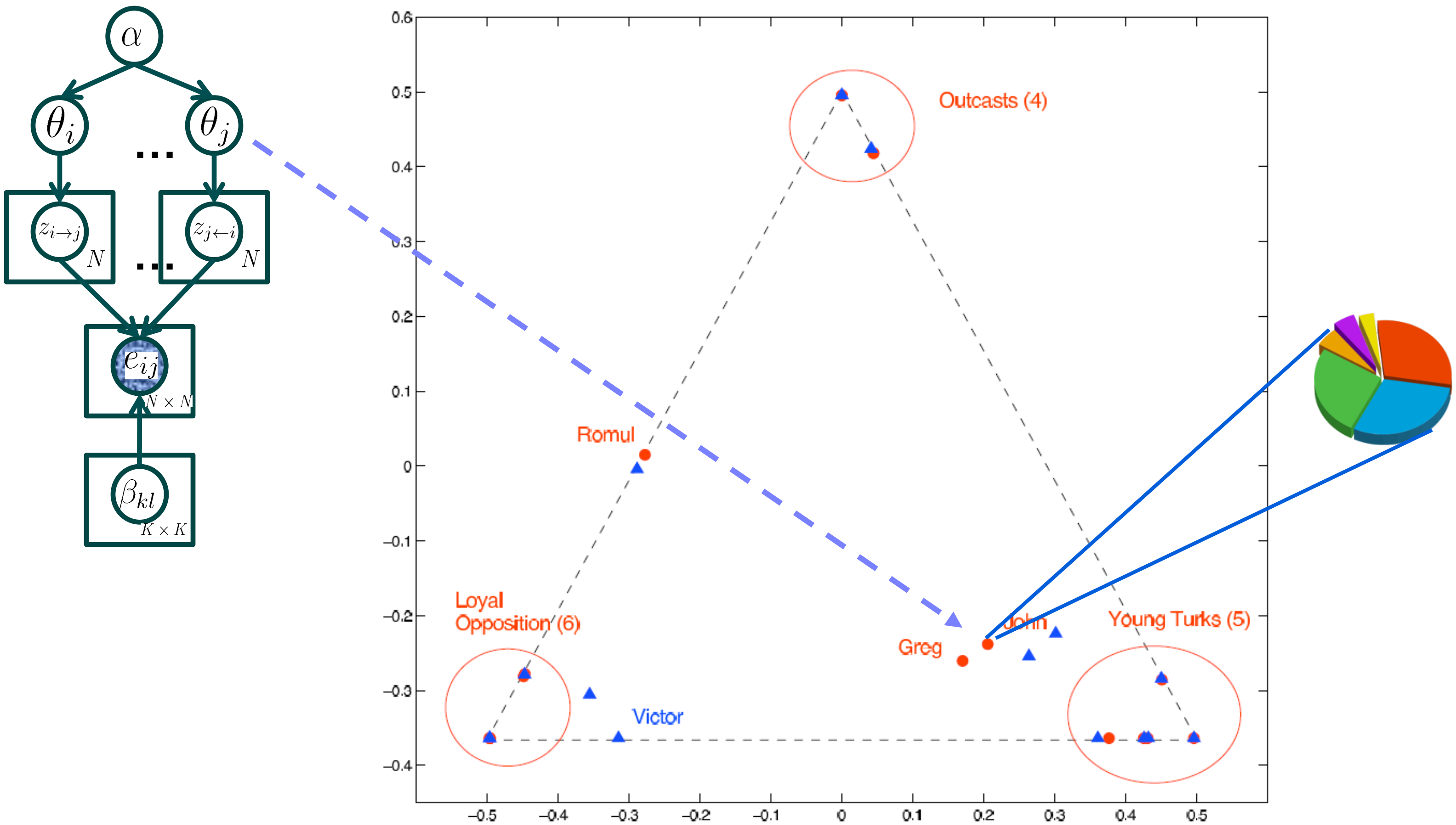
# Mixed Membership Stochastic Blockmodel [Airoldi, Blei, Fienberg and Xing, JMLR 2008]




1. $\{\theta_i\}_{i=1}^N \sim p(\theta|\alpha) \equiv \text{Dirichlet}(\theta;\alpha)$
   sample mixed membership vectors.
2. For each actor $v_j$ that actor $v_i$ possibly interacts with:
   - $z_{i\to j} \sim \text{Multinomial}(z|\theta_i)$
     sample an indicator for $v_i$;
   - $z_{i\leftarrow j} \sim \text{Multinomial}(z|\theta_j)$
     sample an indicator for $v_j$;
   - $e_{ij} \sim \text{Bernoulli}(e|z_{i\to j}^\top B z_{i\leftarrow j})$
     sample a link.

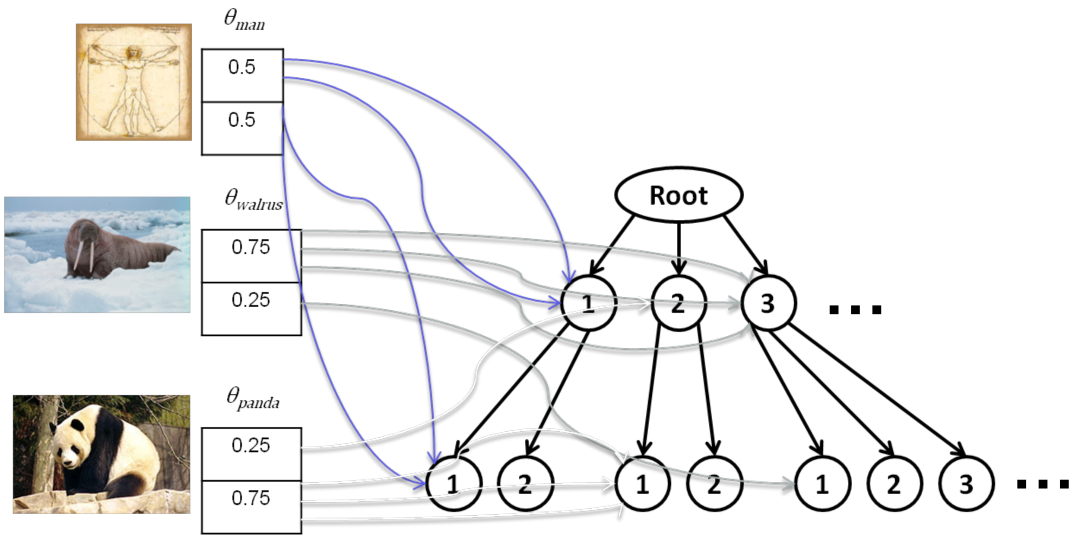
# In the mixed-membership simplex

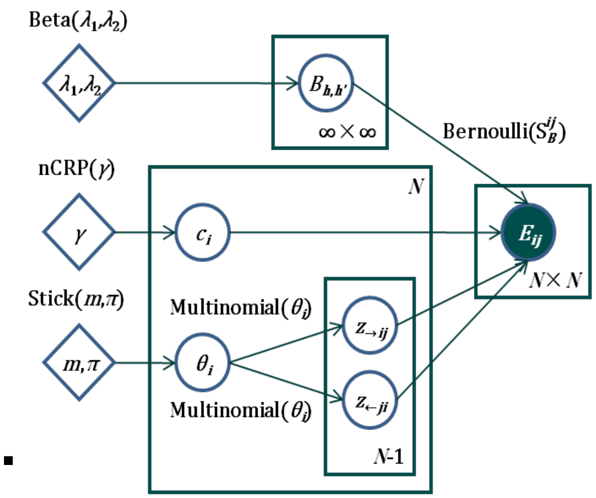
[Airoldi, Blei, Fienberg and Xing, JMLR 2008]

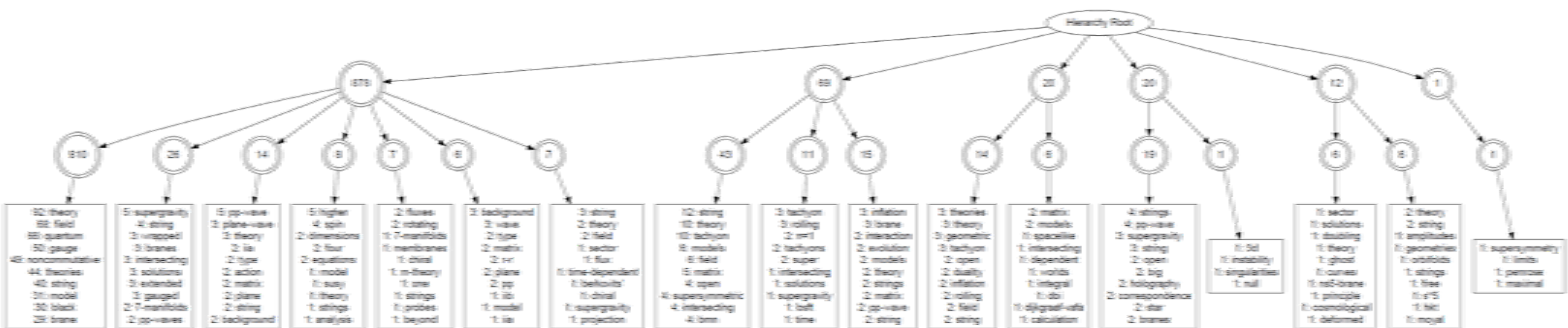


# Multi-Scale Community Blockmodel

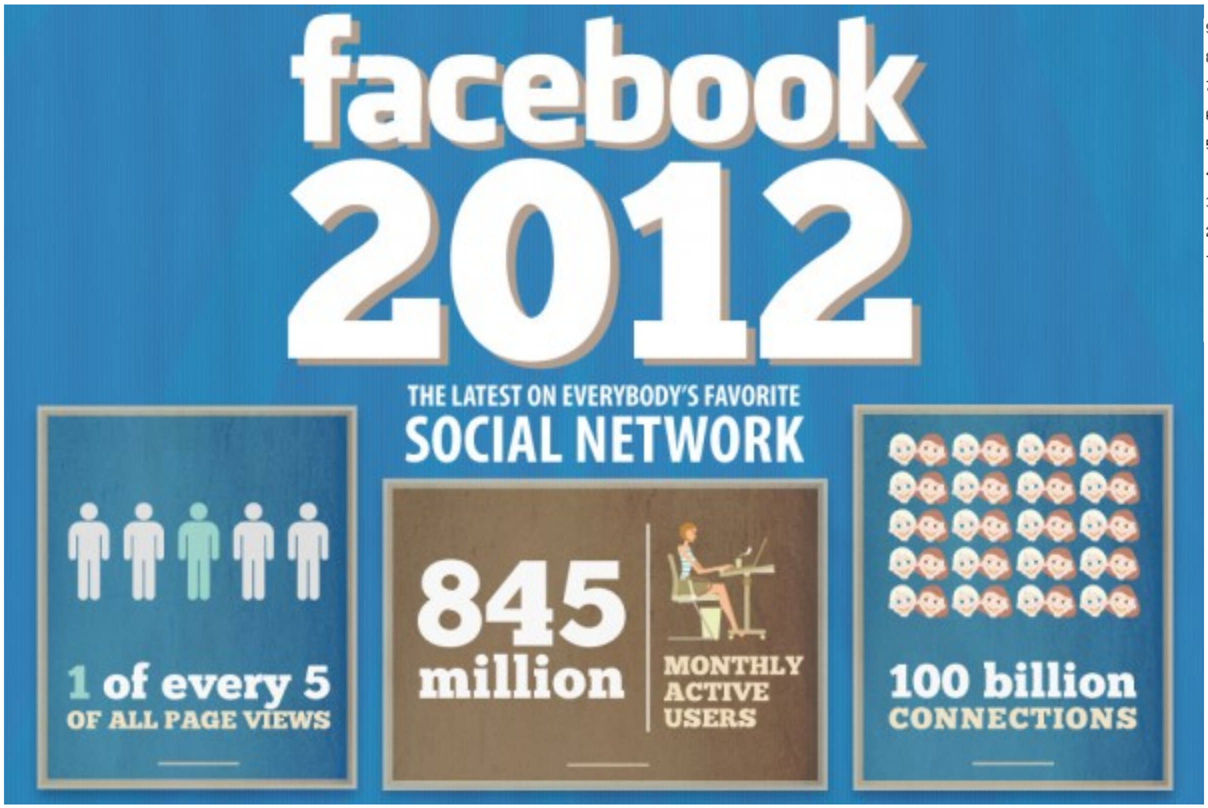
[Ho, Parikh, and Xing, JASA 2012]

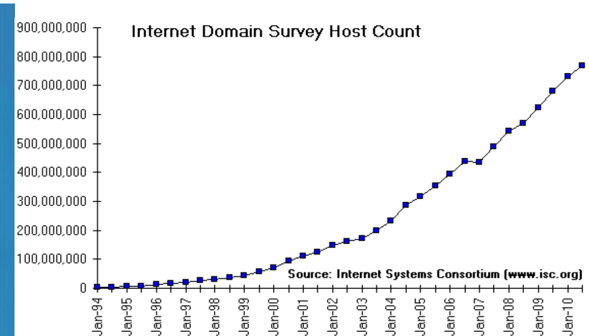


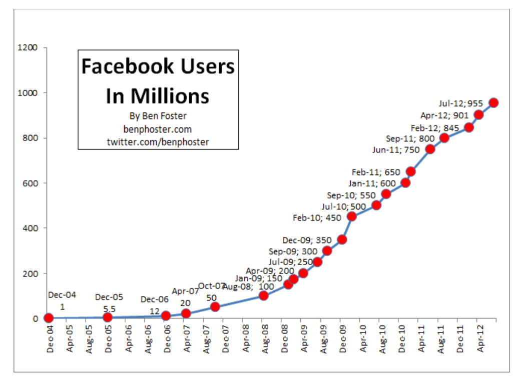



# Challenge – Massive Data Scale




**Does not fitting into memory, nor a single machine, a familiar problem!**

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

# Popular Statistical Network Models don't scale well

- Mixed-Membership Stochastic Blockmodel
  - Models every element A(i,j) of the adjacency matrix, which has size $\Theta(N^2)$
  - Hence $\Theta(N^2)$ latent variables
  - Hence $\Omega(N^2)$ time per iteration of approximate inference
- Latent Factor models
  - Only $\Theta(N)$ latent variables, but Markov Blanket of each variable is $\Theta(N)$ in size
  - Thus $\Omega(N^2)$ time per iteration of approximate inference
- Exponential Random Graph Models
  - Estimated via MCMC-MLE, which samples the adjacency matrix
  - So $\Omega(N^2)$ time for approximate inference
- Fundamental problem: the above models all represent the network by its adjacency matrix, i.e. the matrix of all relationships A(i,j)
  - Adjacency matrix has size $\Theta(N^2)$, so inference will take $\Theta(N^2)$ time as well!
  - The more compact adjacency list representation is NOT a solution, because the above models statistically depend on "missing edges" A(i,j) = 0 as well

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

# ML with scalable, adaptive, online, parallelizable, and confident …

- **Representation:** simpler, but as informative
- **Model:** Generic building blocks: loss functions, structures, constraints, priors ..
- **Algorithm:** sequential, batch → parallel, stochastic
- **System:** multi-core, distributed file system, shared memory, cloud
- **Theory:** linear or sub-linear convergence rate, sample complexity, etc

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

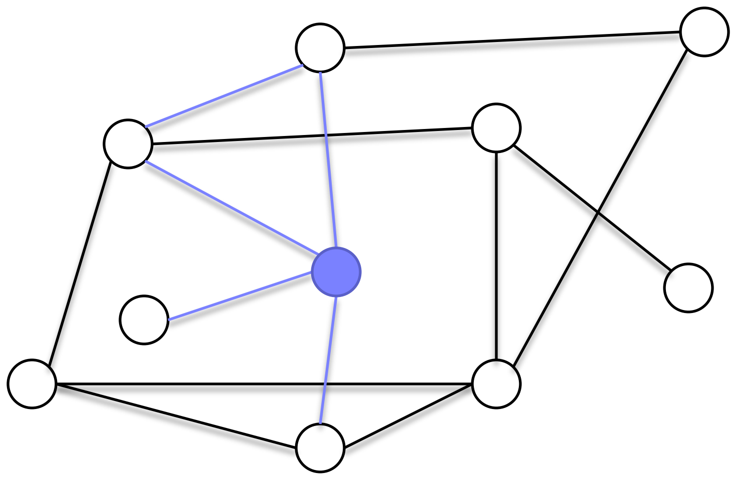
# Scalable Representation

[Ho and Xing, NIPS 2013]

**Node 0**
**Adj. Matrix Features**

| Dest. Node | Edge Status |
|---|---|
| 1 | No |
| 2 | Yes |
| 3 | Yes |
| 4 | Yes |
| 5 | No |
| 6 | No |
| 7 | No |
| 8 | No |
| 9 | Yes |

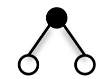
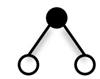
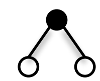
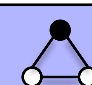
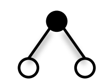
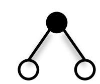
four edges, five non-edges



**Length O(N) vs. Length O(Degree$^2$)**

Triangle features more compact for low node degree!

**Node 0**
**2/3-Triangle Features**

| Dest. Nodes | Triangle Status |
|---|---|
| (2,3) |  |
| (2,4) |  |
| (2,9) |  |
| (3,4) |  |
| (3,9) |  |
| (4,9) |  |

one 3-triangle,
five 2-triangles

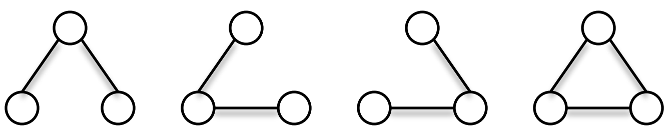
# Why 2/3-edge triangular motifs?

- Well studied in many fields:
  - Biology
  - Social science (transitivity)
  - Data mining (clustering coefficients)
- Basis for network clustering coefficient (CC)
  - Ratio of 2-edge motifs to 2-edge + 3-edge motifs
  - High CC implies stronger, more well-connected clusters
- 2/3-edge motifs contain almost all edges from the adjacency matrix
  - Exception: isolated components with exactly 1 edge
  - Thus, the triangular representation preserves almost all network information!

# Triangular Model Intuition

- Adjacency matrix models (MMSB, Latent Factor) are concerned with edge probabilities
  - i.e. the distribution over events { A(i,j) = 0, A(i,j) = 1 }

- Our triangular motif model is concerned with probabilities over 2/3-edge motifs



  - i.e. the probability of a triple (i,j,k) exhibiting one of the three possible 2-edge motifs, or the sole 3-edge motif

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

# MMTM Generative Process

**Node i**

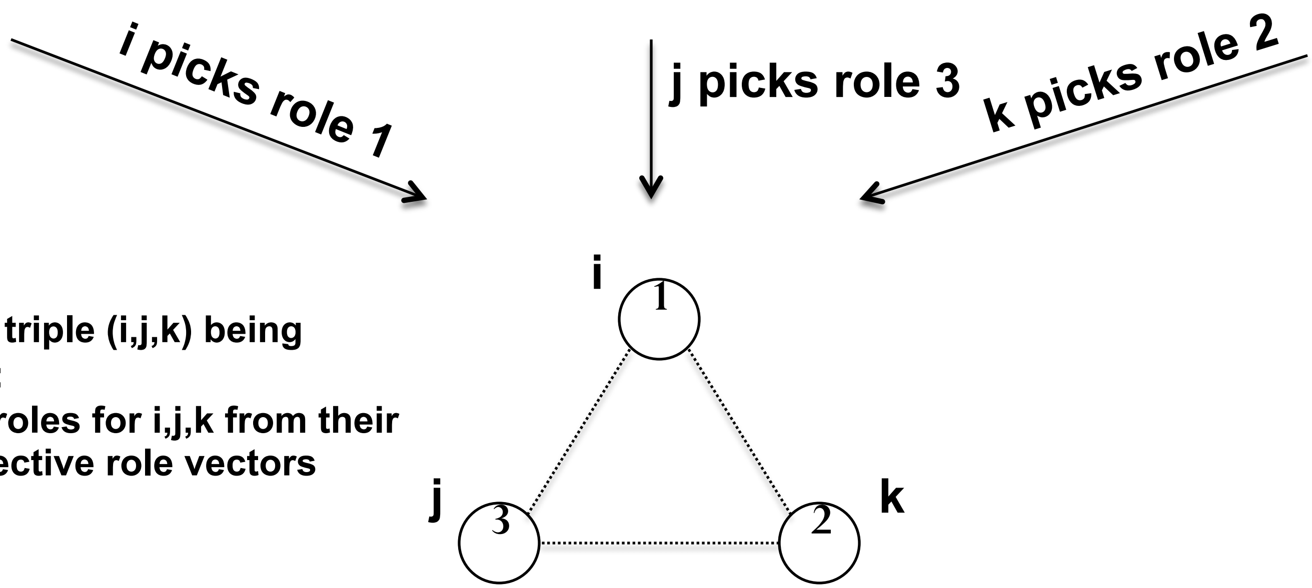
| Role | Probability |
|---|---|
| 1 | 0.7 |
| 2 | 0.2 |
| 3 | 0.1 |

**Node j**

| Role | Probability |
|---|---|
| 1 | 0.5 |
| 2 | 0.1 |
| 3 | 0.4 |

**Node k**

| Role | Probability |
|---|---|
| 1 | 0.3 |
| 2 | 0.7 |
| 3 | 0.0 |



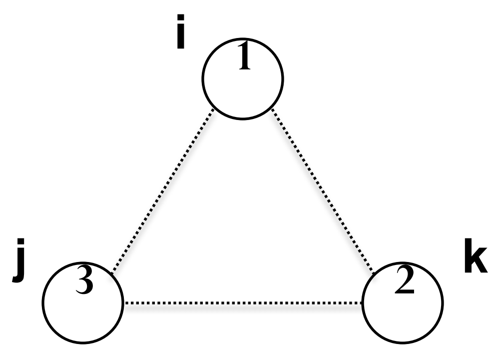

**For each triple (i,j,k) being modeled:**

1. **Pick roles for i,j,k from their respective role vectors**

# MMTM Generative Process

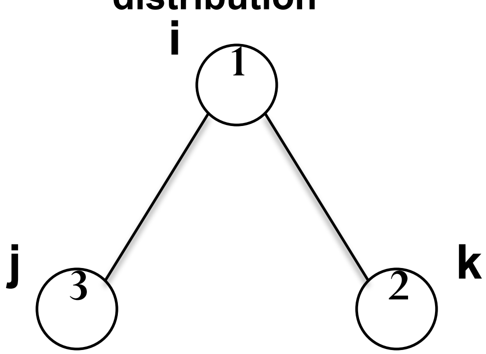
For each triple (i,j,k) being modeled:

1. Pick roles for i,j,k from their respective role vectors
2. Given the combination of roles (in this case 1,3,2), we look up a tensor of parameters B to get that role combination's 2/3-edge motif distribution
3. Generate the motif from the distribution



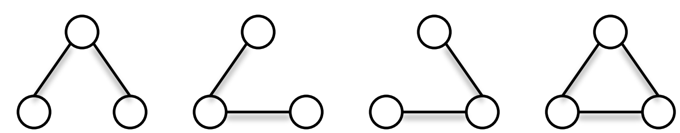

Note: we permit adjacent node triples to generate "incompatible" 2/3-edge motifs. This is in line with the "bag-of-motifs" assumption!

Look up B(1,3,2) to get the 2/3-edge motif distribution




Picks motif 1

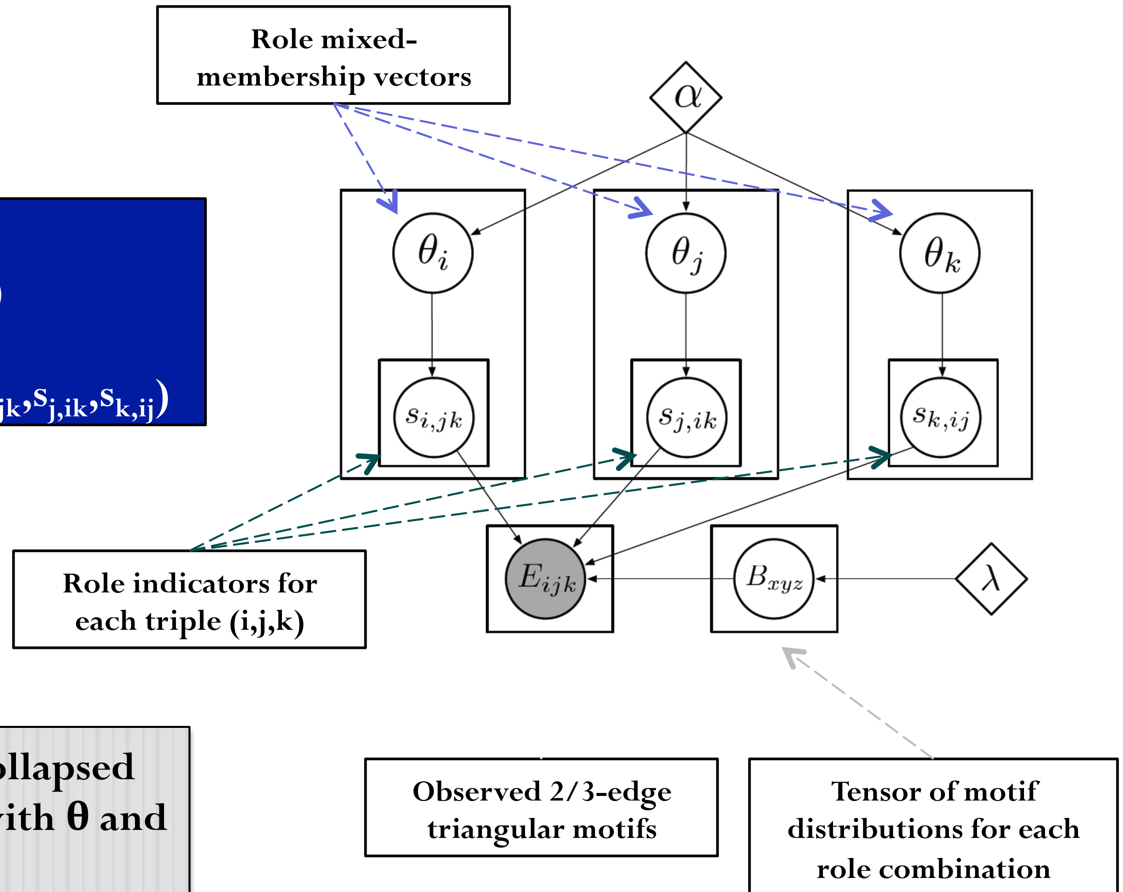
Motif:



Probability: 0.4 0.1 0.2 0.3

# MMTM Graphical Model

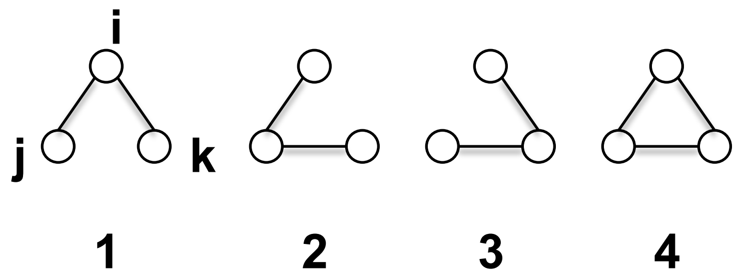
$\theta_i \sim \text{Dirichlet}(\alpha)$
$s_{i,jk} \sim \text{Multinomial}(\theta_i)$
$B_{xyz} \sim \text{Dirichlet}(\lambda)$
$E_{ijk} \sim \text{TriangleDistribution}(B, s_{i,jk}, s_{j,ik}, s_{k,ij})$




We use Rao-Blackwellized/Collapsed Gibbs Sampling for inference, with **θ** and B integrated out

# Additional modeling and scaling technologies

- Isomorphism




- δ-subsampling
  - we pick a constant δ and subsample δ(δ -1)/2 motifs from every node with degree > δ
  - A possible theory of projection invariance
- O(K) triangle probability parameters (instead of $O(K^3)$)
- Stochastic variation inference
- Parallel inference with parameter server architecture and bounded staleness

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

# Simulations

- Statistics for N=4,000 simulation networks:

| | #0,1-edges | #1-edges | $\max(D_i)$ | $\#\Delta_3, \Delta_2$ | $\delta = 20$ | $\delta = 15$ | $\delta = 10$ | $\delta = 5$ |
|---|---|---|---|---|---|---|---|---|
| MMSB | 7,998,000 | 55,696 | 51 | 1,541,085 | 749,018 | 418,764 | 179,841 | 39,996 |
| Latent position | = | 56,077 | 51 | 1,562,710 | 746,979 | 418,448 | 179,757 | 39,988 |
| Biased scale-free | = | 60,000 | 231 | 3,176,927 | 497,737 | 304,866 | 144,206 | 35,470 |
| Pure membership | = | 55,651 | 44 | 1,533,365 | 746,796 | 418,222 | 179,693 | 39,986 |

Table 2: Number of edges, maximum degree, and number of 3- and 2-edge triangles $\Delta_3, \Delta_2$ for each $N = 4,000$ synthetic network, as well as #triangles when subsampling at various degree thresholds $\delta$. MMSB inference is linear in #0,1-edges, while our MMTM's inference is linear in $\#\Delta_3, \Delta_2$.

# Simulations

- MMTM with δ-subsampling is not only much faster, but also more accurate

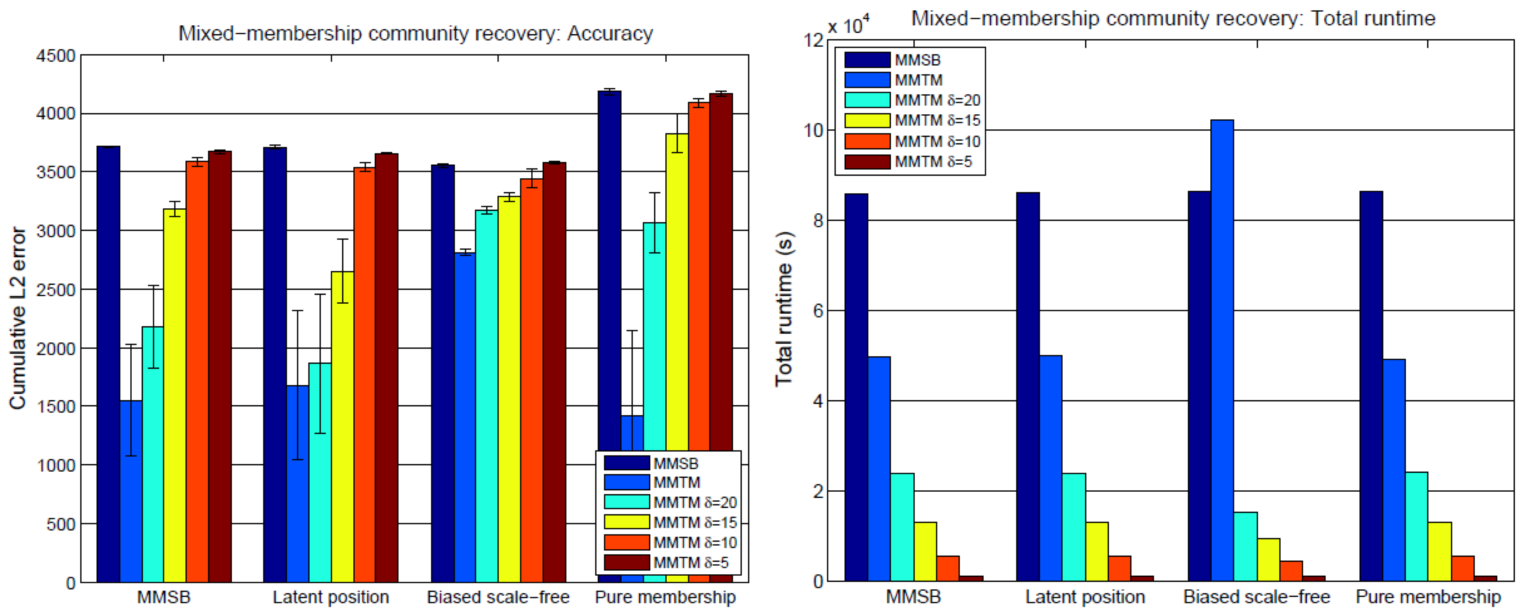
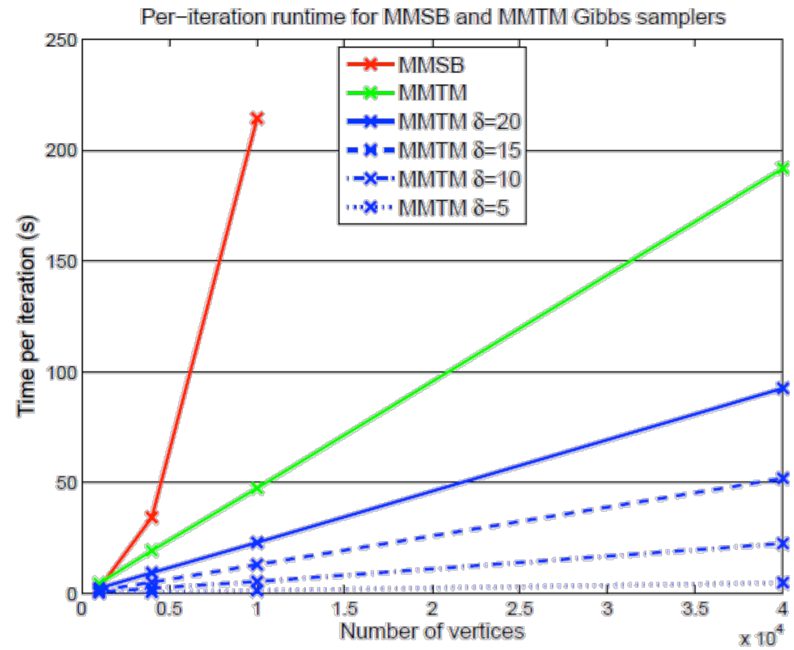



Figure 3: Mixed-membership community recovery task: Cumulative $\ell_2$ errors and runtime per trial for MMSB, MMTM and MMTM with $\delta$-subsampling, on $N = 4,000$ synthetic networks.

# Improvement over state of the art

[Ho, Yin and Xing, NIPS 2012, UAI 2013]

- As the number of nodes increases:
  - MMSB (edge-based representation) runtime increases quadratically
  - MMTM (triangle-based) runtime increases linearly when δ is held constant
  - Stochastic variational inference further improves speed




| Name | Nodes $N$ | Edges | Roles $K$ | Threads | Runtime |
|---|---|---|---|---|---|
| Brightkite | 58K | 214K | 64 | 4 | 35 min[1] |
| Brightkite | \|\| | \|\| | 300 | 4 | 2.7 h |
| Slashdot Feb 2009 | 82K | 504K | 100 | 4 | 2.3 h |
| Slashdot Feb 2009 | \|\| | \|\| | 300 | 4 | 6.8 h |
| Stanford Web | 282K | 2.0M | 5 | 4 | 12 min[2] |
| Stanford Web | \|\| | \|\| | 100 | 4 | 6.3 h |
| Berkeley-Stanford Web | 685K | 6.6M | 100 | 8 | 20.7 h |
| Youtube | 1.1M | 3.0M | 100 | 8 | 10.7 h |

**Competing methods**

8 days (Blei, NIPS 2012)

18 hrs (Ho et al, NIPS 2012

# A Larger-scale Demonstration

- Stanford web graph, N≈280,000
  - Ran for 2,000 sampling iterations, convergence observed by 500 iterations
  - Total runtime: 74 hours on a single computational thread

Every circle represents a node in the network

Circle sizes are proportional to node degrees

Colors and positions represent inferred role MM vectors

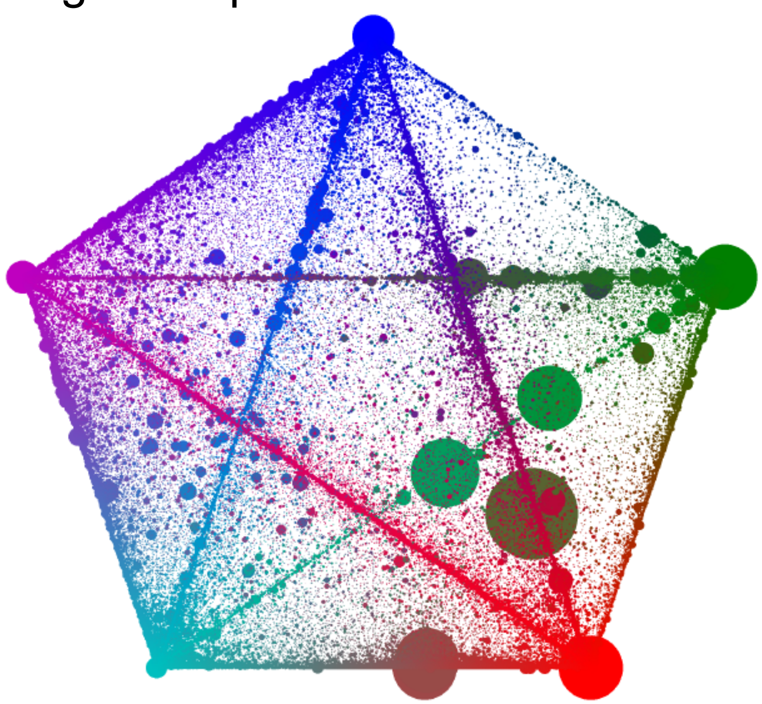


Figure 5: $N = 281,903$ Stanford web graph, MMTM mixed-membership visualization.

# MMTM Stochastic Variational

- Variational EM on randomly chosen triangles (data points)
  - Similar to stochastic variational for LDA and MMSB
  - Only need to touch every triangle 2-3 times to converge

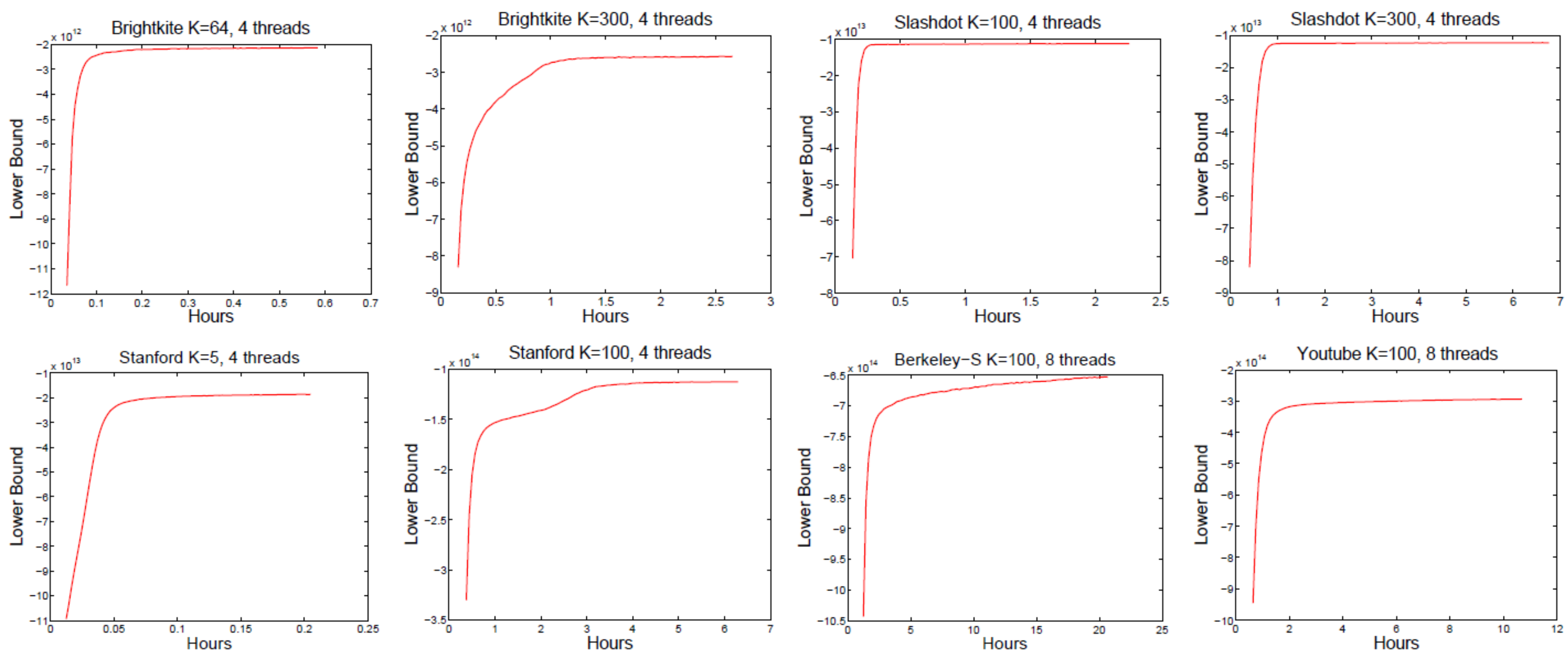
- O(K) triangle probability parameters B
  - Old model:
    - B(a,b,c) for all $K^3$ choices of roles a,b,c → $K^3$ parameters
  - New model:
    - **3-roles-same:** B(a,a,a) for each of the K choices of a → K parameters
    - **2-roles-same:** B(a,a,·) for each of the K choices of a, where · ≠ a → K parameters
    - **All-roles-different:** B(·,·,·) where all three · are different → 1 parameter
    - Total 2K + 1 parameters

- Parallelization
  - Alternate inference between
    - Node topic vectors Θ
    - Triangle role assignments s
    - Triangle probability parameters B
  - Each variable type can be parallelized given the other 2 types

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

# Running time on real networks

| Real Networks — Statistics, Experimental Settings and Runtime | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Name | Nodes $N$ | Edges | $\delta$ | 2,3-Tris (for $\delta$) | Frac. of 3-Tris | Roles $K$ | Threads | Runtime |
| Brightkite | 58K | 214K | 50 | 3.5M | 0.11 | 64 | 4 | 35 min[1] |
| Brightkite | \|\| | \|\| | \|\| | \|\| | \|\| | 300 | 4 | 2.7 h |
| Slashdot Feb 2009 | 82K | 504K | 50 | 9.0M | 0.030 | 100 | 4 | 2.3 h |
| Slashdot Feb 2009 | \|\| | \|\| | \|\| | \|\| | \|\| | 300 | 4 | 6.8 h |
| Stanford Web | 282K | 2.0M | 20 | 11.4M | 0.57 | 5 | 4 | 12 min[2] |
| Stanford Web | \|\| | \|\| | 50 | 25.0M | 0.42 | 100 | 4 | 6.3 h |
| Berkeley-Stanford Web | 685K | 6.6M | 35 | 67.1M | 0.55 | 100 | 8 | 20.7 h |
| Youtube | 1.1M | 3.0M | 50 | 36.0M | 0.053 | 100 | 8 | 10.7 h |

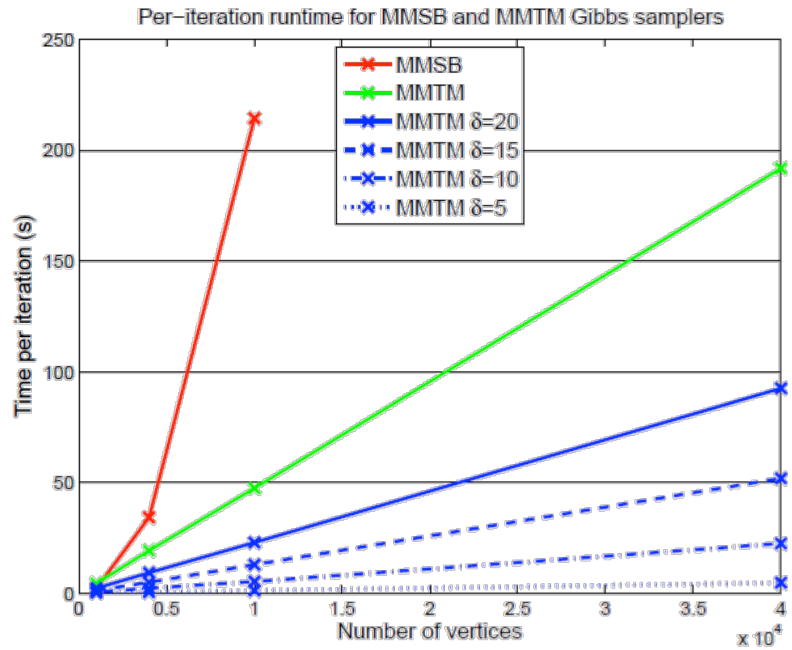



**Graphs/runtime are for 10 passes per data point (triangle)**
**Convergence occurs in 2-3 data passes**
**Youtube network (1.1M nodes, K = 100) in <2h using 8 threads**

# Further Improvement over state of the art

- As the number of nodes increases:
  - MMSB (edge-based representation) runtime increases quadratically
  - MMTM (triangle-based) runtime increases linearly when δ is held constant
  - Stochastic variational inference further improves speed

[Ho, Yin and Xing, NIPS 2012, UAI 2013]




| Name | Nodes $N$ | Edges | Roles $K$ | Threads | Runtime |
|---|---|---|---|---|---|
| Brightkite | 58K | 214K | 64 | 4 | 35 min[1] |
| Brightkite | \|\| | \|\| | 300 | 4 | 2.7 h |
| Slashdot Feb 2009 | 82K | 504K | 100 | 4 | 2.3 h |
| Slashdot Feb 2009 | \|\| | \|\| | 300 | 4 | 6.8 h |
| Stanford Web | 282K | 2.0M | 5 | 4 | 12 min[2] |
| Stanford Web | \|\| | \|\| | 100 | 4 | 6.3 h |
| Berkeley-Stanford Web | 685K | 6.6M | 100 | 8 | 20.7 h |
| Youtube | 1.1M | 3.0M | 100 | 8 | 10.7 h |

**Competing methods**

8 days (Blei, NIPS 2012)

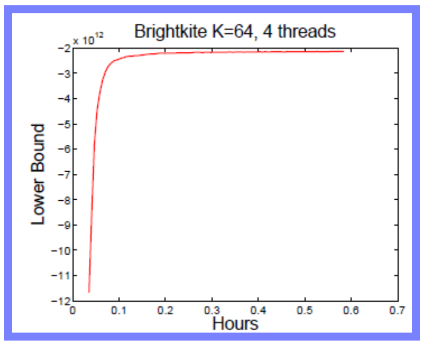
18 hrs (Ho et al, NIPS 2012

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

# And the improvement continues ...

| Real Networks — Statistics, Experimental Settings and Runtime | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Name | Nodes $N$ | Edges | $\delta$ | 2,3-Tris (for $\delta$) | Frac. of 3-Tris | Roles $K$ | Threads | Runtime |
| Brightkite | 58K | 214K | 50 | 3.5M | 0.11 | 64 | 4 | 35 min[1] |
| Brightkite | \|\| | \|\| | \|\| | \|\| | \|\| | 300 | 4 | 2.7 h |
| Slashdot Feb 2009 | 82K | 504K | 50 | 9.0M | 0.030 | 100 | 4 | 2.3 h |
| Slashdot Feb 2009 | \|\| | \|\| | \|\| | \|\| | \|\| | 300 | 4 | 6.8 h |
| Stanford Web | 282K | 2.0M | 20 | 11.4M | 0.57 | 5 | 4 | 12 min[2] |
| Stanford Web | \|\| | \|\| | 50 | 25.0M | 0.42 | 100 | 4 | 6.3 h |
| Berkeley-Stanford Web | 685K | 6.6M | 35 | 67.1M | 0.55 | 100 | 8 | 20.7 h |
| Youtube | 1.1M | 3.0M | 50 | 36.0M | 0.053 | 100 | 8 | 10.7 h |

**Brightkite 58K nodes**

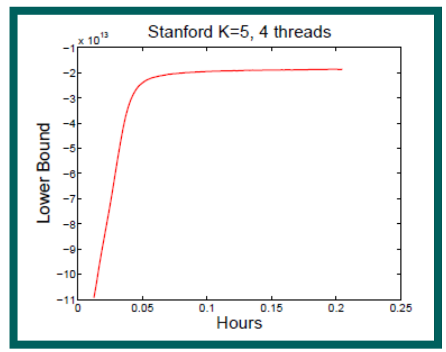



**New MMTM converges in 12 min**

**Stochastic Variational MMSB (Gopalan et al, NIPS 2012) took 8 days using 4 threads**

**1000x speedup!**

**Stanford 282K nodes**




**New MMTM converges in 6 min**

**Gibbs MMTM (Ho et al, NIPS 2012) took 18.5 hours using 1 thread**

**200x speedup!**

Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

# Conclusion

- MMTM exploits a "bag-of-triangular-motifs" network representation
  - Specifically, MMTM models triangular motifs with 2 or 3 edges
  - Parsimonious alternative to edge-based adjacency matrix representation, which has size $N^2$

- MMTM scales to much larger networks than adjacency matrix models such as MMSB, ERGMs or latent position models
  - With δ-subsampling, # 2/3-edge triangular motifs << $N^2$
  - 100K node networks are feasible with MMTM (single thread)
  - Whereas 10K node networks are already impractical for the a/m models

- MMTM inference yields better role MM vector recovery than MMSB inference on a variety of models
  - Even on the MMSB model itself!
  - This is partly because MMTM's state space is much smaller (fewer latent variables), thus MMTM approximate inference converges much faster

Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

# A note on scalable ML

- Our New MMTM is built on 3 principles:
  - Compact **data representation** (triangles rather than edges)
  - Parsimonious **model** with linear O(K) number of role parameters
  - Fast, scalable, distributable **inference algorithm** (stochastic variational EM)

- These principles are the building blocks for truly scalable Big ML

- Next step: distributed general ML inference engine for large clusters

# Future Work

- Parallelization
  - One thread can perform inference on N=280K nodes in 3 days
  - We aim to parallelize to 1,000 threads, so as to perform inference on networks with N=100M nodes

- Subsampling strategies
  - What are the theoretical properties of δ-subsampling?
  - Are there better subsampling strategies?