A Scalable Approach to Probabilistic Latent Space Inference of Large-Scale Networks

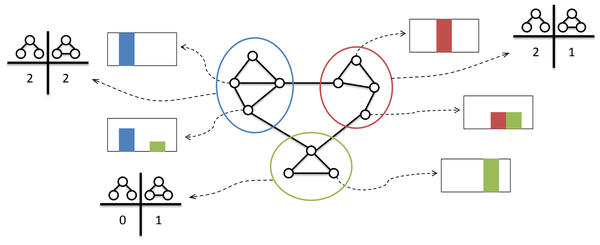
Junming Yin, Qirong Ho, Eric P. Xing

Carnegie Mellon University

Overview

In latent space modeling of networks, **we aim to**

- reduce the high-dimensional network data to a K-dimensional "feature space" – each feature corresponds to a role or a community



For large-scale networks, **we need**

- inferential mechanisms that scale in both the number of vertices N and the number of roles K

However, popular statistical network models and inference algorithms do not scale **linearly**

- MMSB batch variational [Airoldi et. al (2008)]: $O(N^2K^2)$
- MMTM Gibbs sampling [Ho et. al (2012)]: $O(NK^3)$

We present

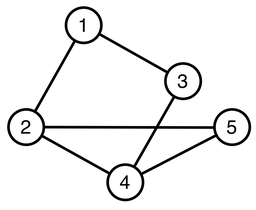
- an $O(NK)$ scalable approach, with competitive or improved accuracy for latent space recovery and link prediction
- an implementation that allows network analysis with millions of nodes and hundreds of roles in hours on a single multi-core machine

Triangular Representation of Networks

Represent networks as a bag of triangular motifs

- record each vertex triple containing 2 or 3 edges
- ignore triples with 0 or 1 edge

Original network



A bag of triangular motifs

2 – 1 – 3
1 – 3 – 4
3 – 4 – 2
3 – 4 – 5
4 – 2 – 1
5 – 2 – 1

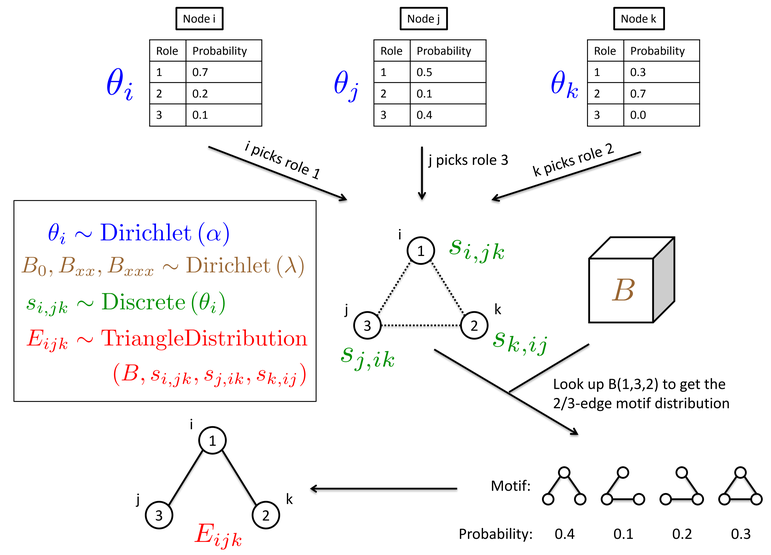
2 – 4 – 5 – 2

Why Triangular Representation?

- Well studied in social science, data mining and biology
- Basis for network clustering coefficient, i.e., ratio of 3-edge triangles to 2/3-edge triangles
- Preserves network information – 2/3-edge triangles cover almost all the edges except isolated edges
- A succinct representation – the number of triangular motifs is bounded by $O(ND^2)$, where D is the maximum vertex degree
- For high-degree networks, use δ-subsampling to maintain $O(N\delta^2)$ triangular motifs
  - for each vertex with degree higher than δ, uniformly sample $\delta(\delta-1)/2$ triangles from the set composed of (a) its adjacent 3-edge triangles and (b) its adjacent 2-edge triangles with that vertex in the center

Parsimonious Triangular Model

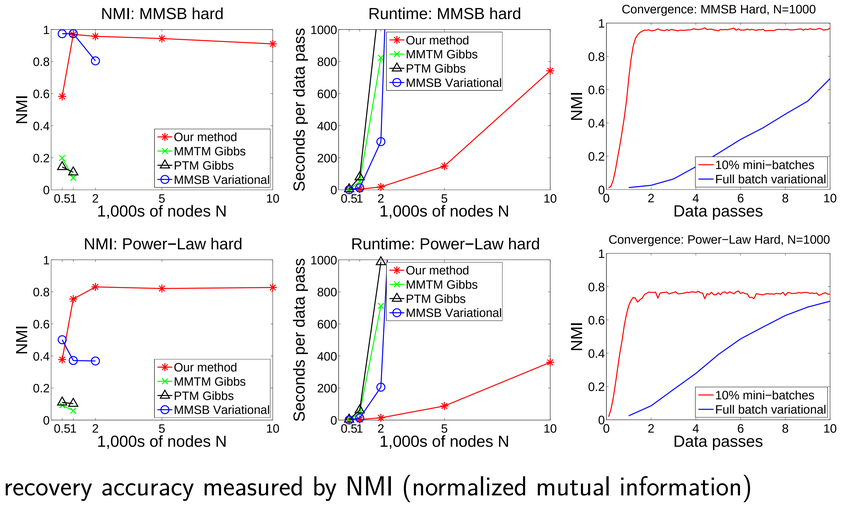
- A generative model for a bag of triangular motifs, not for network simulation



- $O(K)$ parameters for triangle-generating probabilities
  - B_{xxx}: if all three role indices are in the same state x
  - B_{xx}: if only two role indices exhibit the same state x (called majority role), shared across different minority roles
  - B_0: if the three role indices are all distinct, independent of the role configurations
- Posterior inference of (s, θ, B) by stochastic variational inference
  - randomly sample a mini-batch of triangular motifs, and only update their local variational parameters for roles $s_{i,jk}$ in parallel
  - accumulate sufficient statistics for the natural gradients of global variational parameters for θ_i and B
  - optimize the global variational parameters by a stochastic natural gradient ascent rule

Experimental Results

Latent space recovery on MMSB and power-law graphs with increasing network size N and K

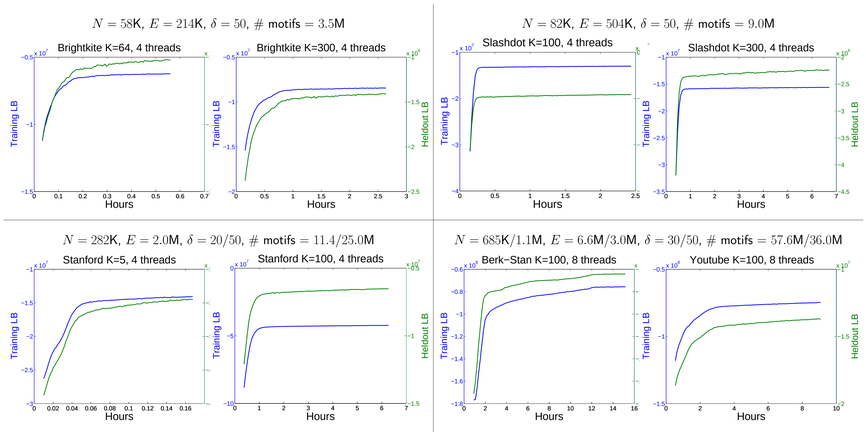


- recovery accuracy measured by NMI (normalized mutual information)
- our method is faster and more accurate on large networks
- on $N = 1000$ networks, the algorithm converges within 1-2 data passes

Link prediction on synthetic and real networks

Link Prediction on Synthetic and Real Networks									
Network Type	Synthetic		Dictionary		Biological	arXiv Collaboration		Internet	Social
Name	MMSB	Power-law	Roget	Odlis	Yeast	GrQc	AstroPh	Stanford	Youtube
Nodes N	2.0K	2.0K	1.0K	2.9K	2.4K	5.2K	18.7K	282K	1.1M
Edges	40K	40K	3.6K	16K	6.6K	14K	200K	2.0M	3.0M
Our Method AUC	**0.93**	**0.97**	0.65	0.81	0.75	**0.82**	**0.86**	**0.94**	**0.71**
MMSB Variational AUC	0.91	0.94	**0.72**	**0.88**	**0.81**	0.77	—	—	—

Real world networks – convergence on heldout data



- our algorithm is able to infer a 100-role latent space on a 1M-node Youtube social network in 4 hours, using a single machine with 8 threads