Goodness of Fit in Social Networks

Beau Dabbs

February 18, 2014

Motivation

Model Selection

- Detect a Misspecified Model
- Select the Best Model from Candidate Models

Model Comparison:

- How Well Can a Mis-Specified Model Recover
- What Features of a Model Are Unique

Overview

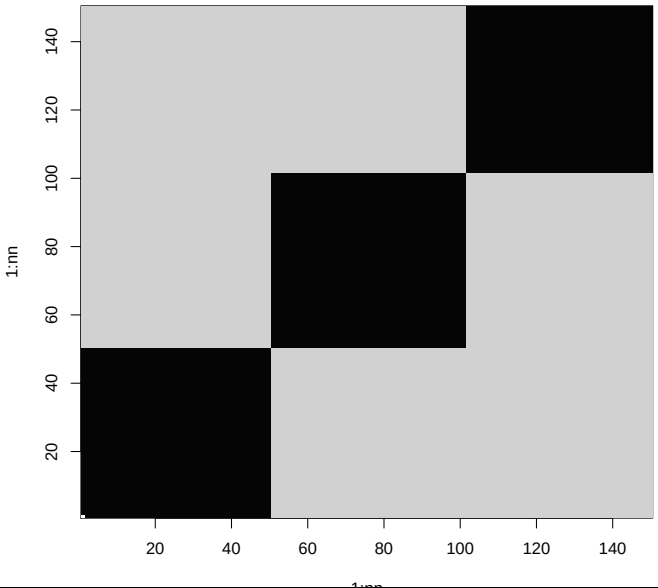
- Three Similar Generative Models:
  - Stochastic Block Model
  - Mixed Membership Stochastic Block Model
  - Latent Space Model
- Goodness of Fit Tests
  - Posterior Predictive Checking
  - Out of Sample Validation

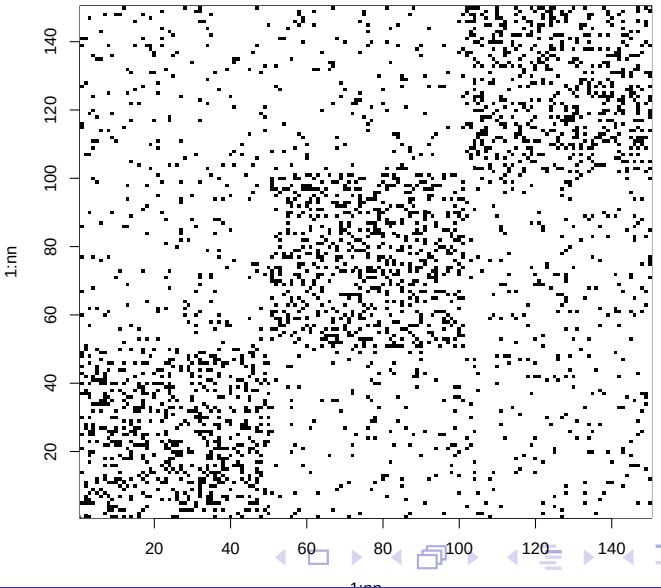
Stochastic Block Model

$$Y_{ij}|\pi_i = k, \pi_j = l \sim Bernoulli(b_{kl})$$

$$b_{kl} \sim Beta(a, b) \qquad \pi_i \sim multinom(\frac{1}{m}, ..., \frac{1}{m})$$

- Simple Community Detection Model
- Group Membership is Strict



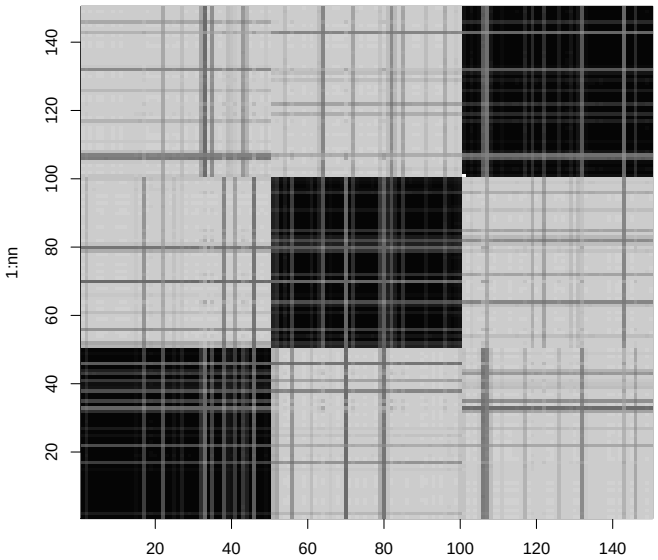


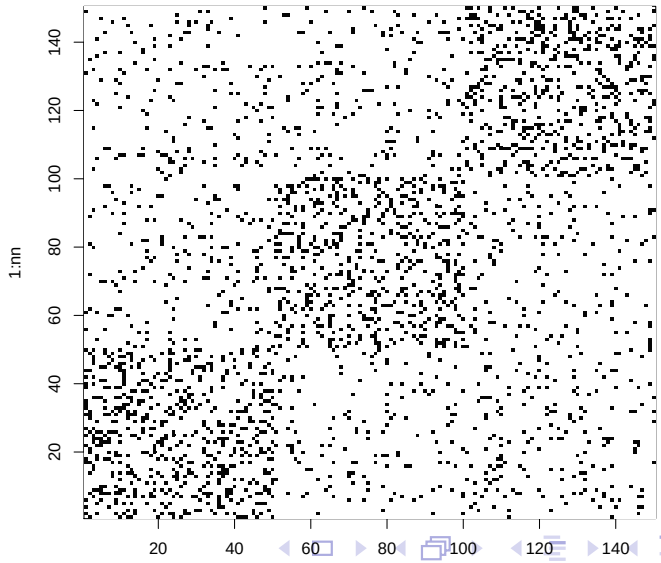
Mixed Membership Stochastic Block Model

$$Y_{ij}|\pi_i, \pi_j \sim Bernoulli(\pi_i^T B \pi_j)$$

$$b_{kl} \sim Beta(a, b) \qquad \pi_i \sim Dirichlet(\vec{\alpha})$$

- Group Membership is Partial
- Within Block Ties Decrease; Between Block Ties Increase





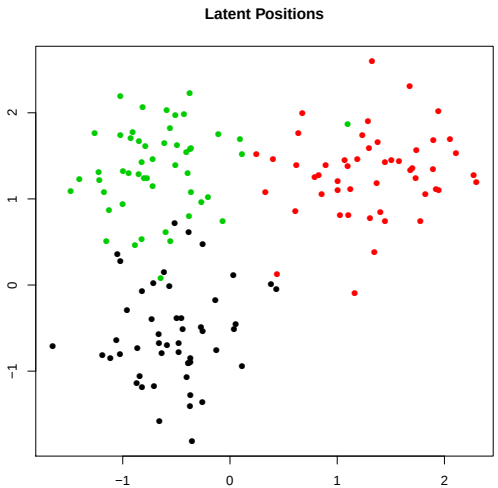
Latent Space Model

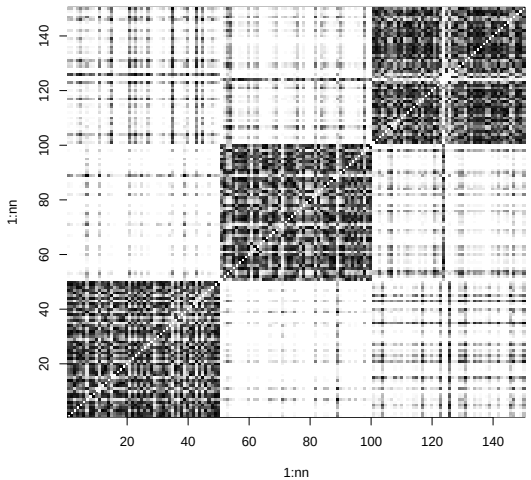
$$Y_{ij} \sim Bernoulli(p_{ij})$$

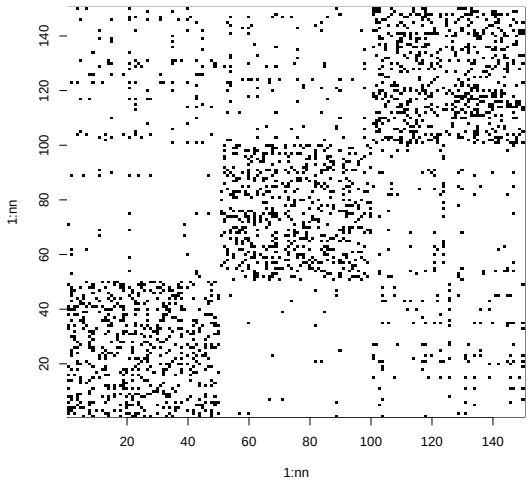
$$logit(p_{ij}) = a - ||z_i - z_j||_2$$

$$z_i \sim N((0,0), \sigma^2 I_2) \qquad a \sim N(0, \tau^2)$$

- Clusters in Latent Space Can Mimic Communities
- Number of Communities Not Specified







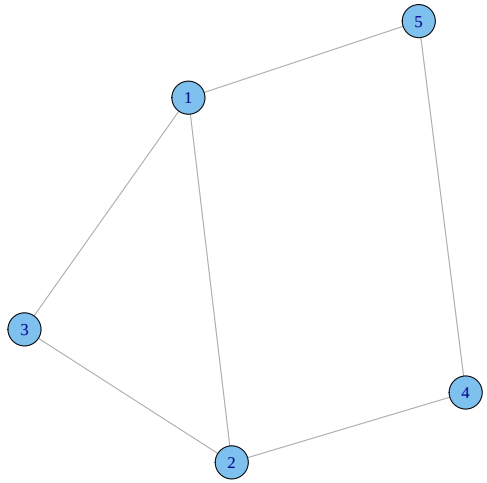
Goodness of Fit Tests

- Posterior Predictive Checking
  - Given Posterior Distribution, Sample New Networks and Compare Statistics
  - Questions:
    - Which Statistics Should we Use?
    - How Extreme is Too Extreme?
- Out of Sample Tests
  - Leave-Out Validation
  - K-Fold Cross Validation
  - Questions:
    - What Percentage of Data Should be Left Out
    - What Metric Should be Used

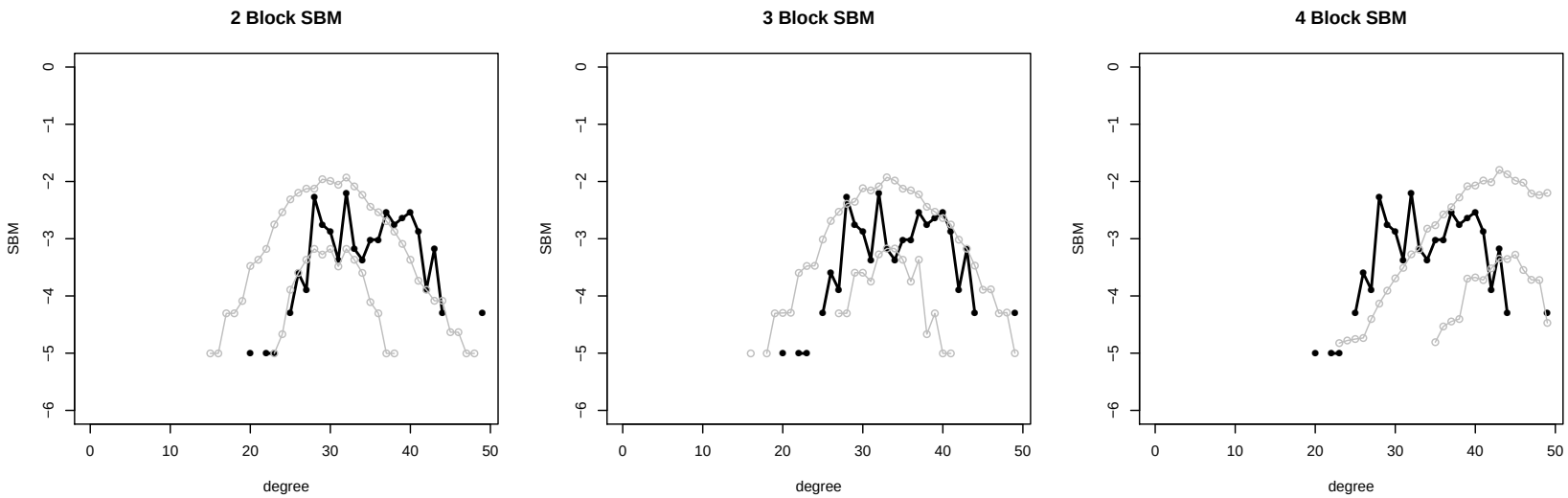
Posterior Predictive Checks

Network Statistics:

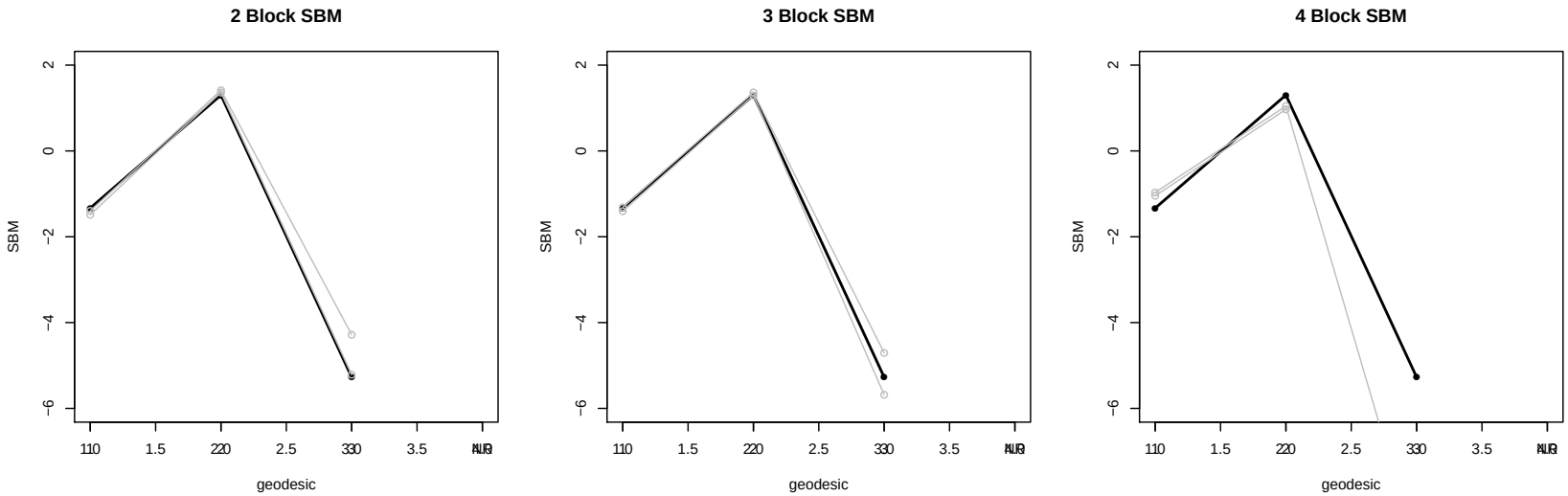
- Degree Distribution **[0.0, 0.0, 0.6, 0.4]**
- Geodesic Distance Distribution - Distribution of shortest paths between pairs of nodes **[0.6, 0.4]**
- Shared Partner Distribution - Distribution of shared partners between adjacent nodes **[0.5, 0.5]**



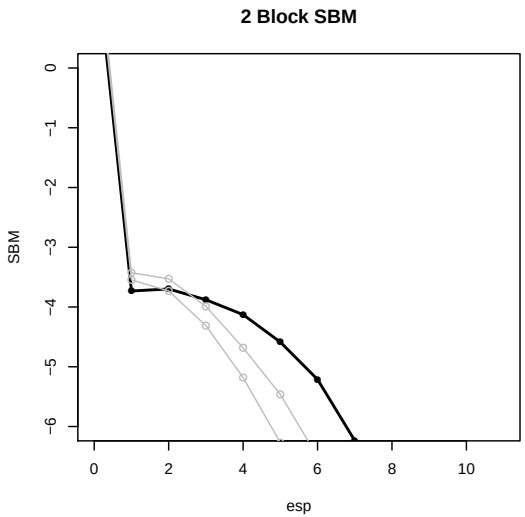
3 Block SBM - Degree Distribution

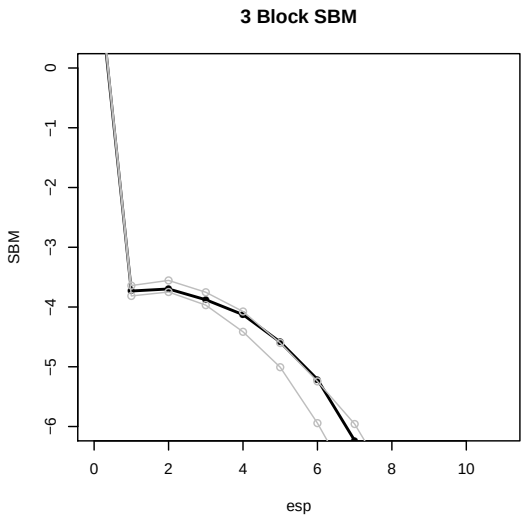


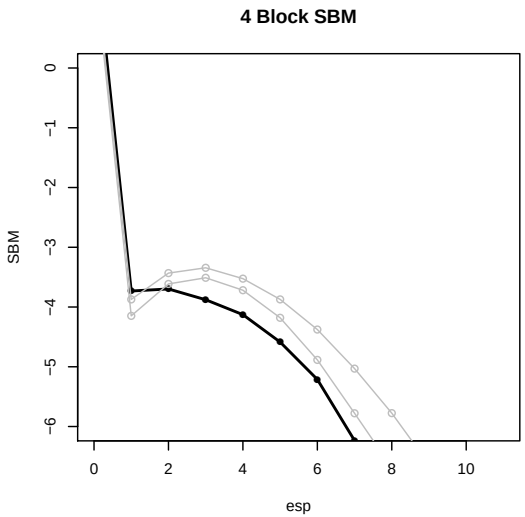
3 Block SBM - Geodesic Distance Distribution



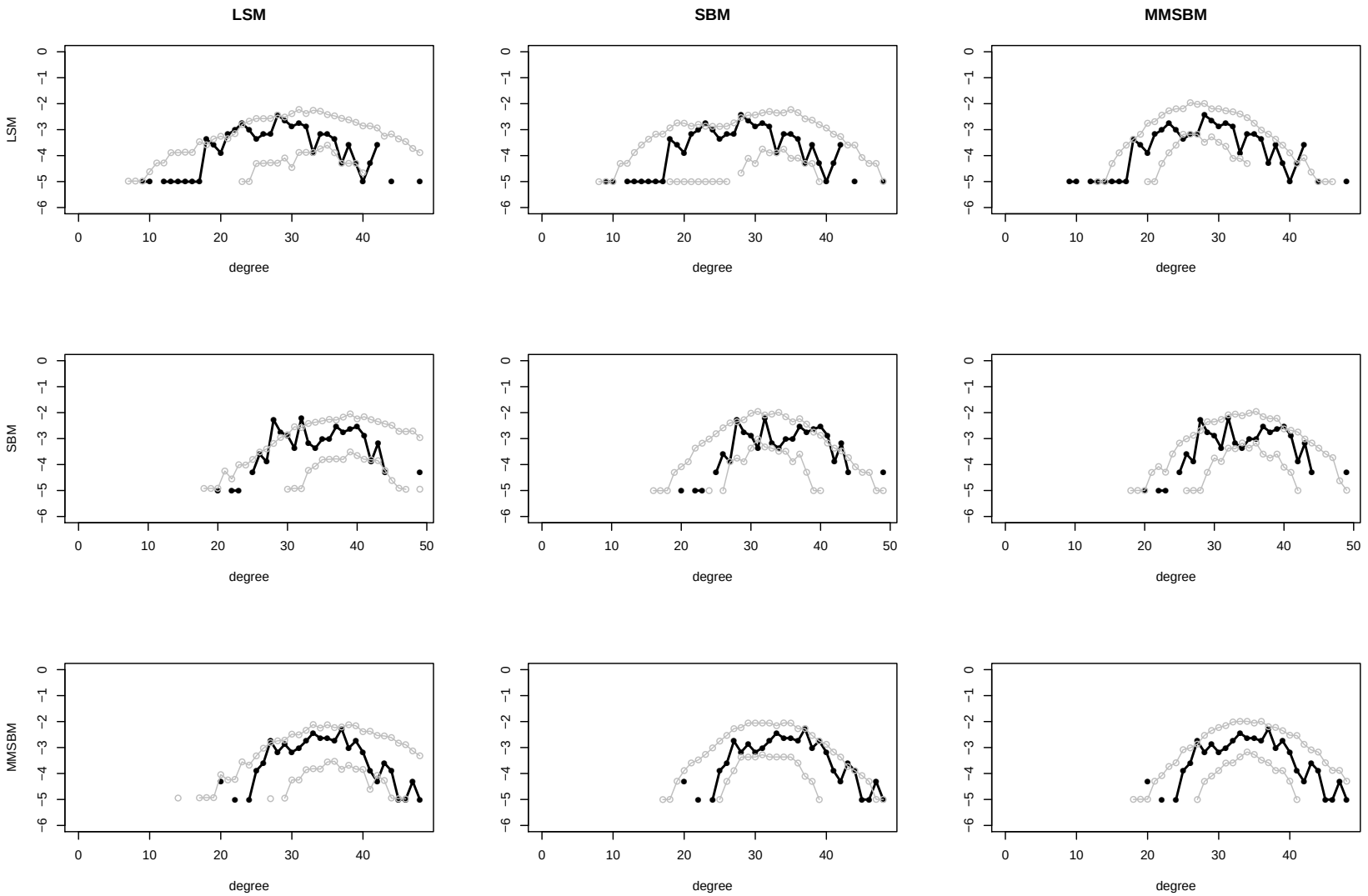
3 Block SBM - Shared Partner Distribution



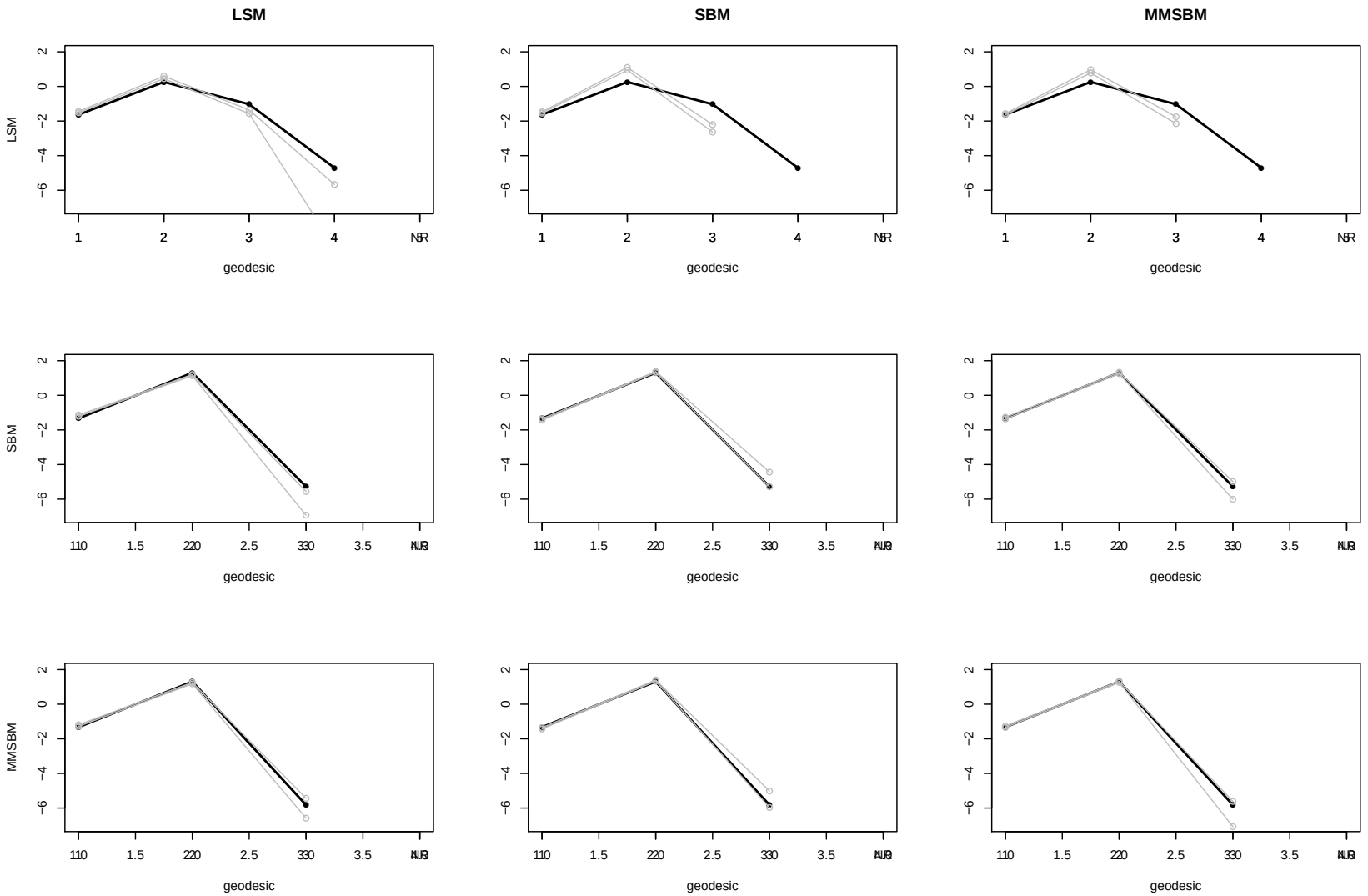




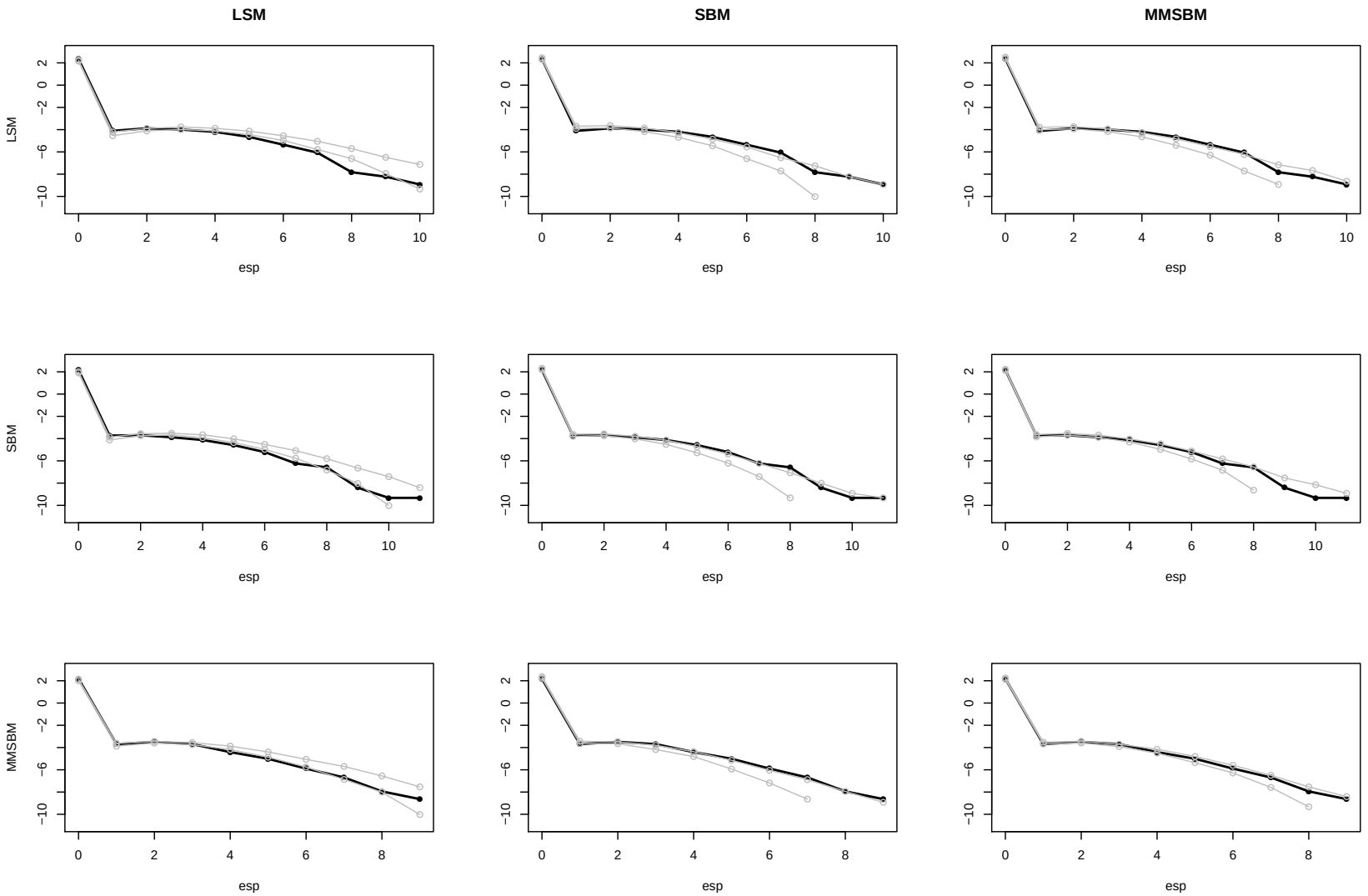
Degree Distribution



Geodesic Distance Distribution



Shared Partner Distribution



Out of Sample Metrics

Leave Out Validation Method:

1. Observe complete graph G
2. Create a new graph H by removing p percent of the edges at random.
3. Fit a model M on H to predict tie probabilities
4. Use a loss function L to measure accuracy of prediction

Cross-Validation Method

1. Divide data into K sets.
2. Iteratively leave out one of the K sets, and compute Leave Out Validation for each

Sample Sizes in Cross-Validation

Risk: $\hat{f}$ trained on full dataset; $\mathbf{X}$ and Y new points

$$\mathbb{E}_P\left[(\hat{f}(\mathbf{X}) - Y)^2\right] = \mathbb{V}_P[\hat{f}(\mathbf{X})] + \mathbb{E}_P[(\hat{f}(\mathbf{X}) - Y)]^2$$

Leave One Out Risk (LOO): $\hat{f}_{(i)}$ trained without ith point;

$$\mathbb{E}_P\left[(\hat{f}_{(i)}(\mathbf{X}_i) - Y_i)^2\right] = \mathbb{V}_P[\hat{f}_{(i)}(\mathbf{X}_i)] + \mathbb{E}_P[(\hat{f}_{(i)}(\mathbf{X}_i) - Y_i)]^2$$

- Variance of $\hat{f}_{(i)}$ is generally larger
- LOO Risk is a conservative, biased estimate of true Risk
- Leaving out more data increases this bias, but lowers the variance of our risk estimation.

Commonly Used Loss Functions

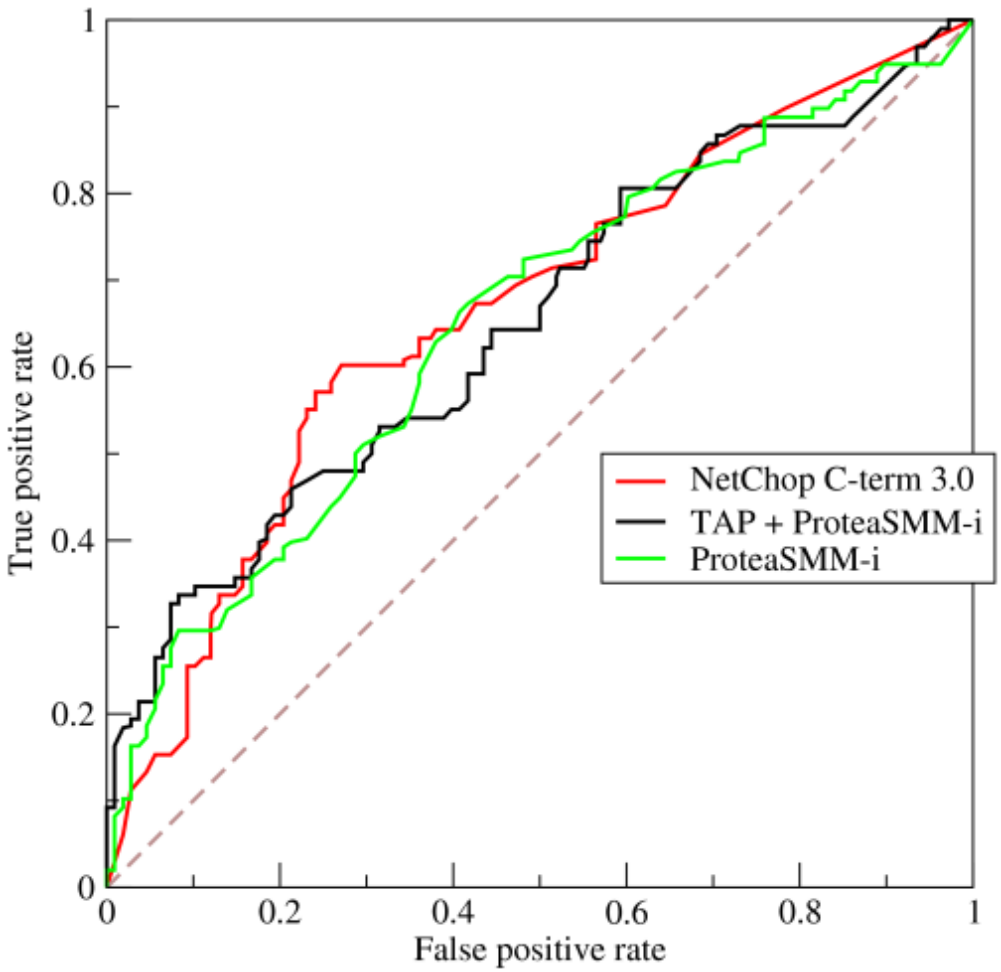
Loss Function Definitions:

- RMSE $= \sqrt{\frac{1}{T}\sum_{i=1}^{T}(Y_i - p_i)^2}$
- 1 - AUC $= \frac{1}{N_0 N_1}\sum_{i:Y_i=0}\sum_{j:Y_j=1}\mathbf{1}\{p_i > p_j\}$

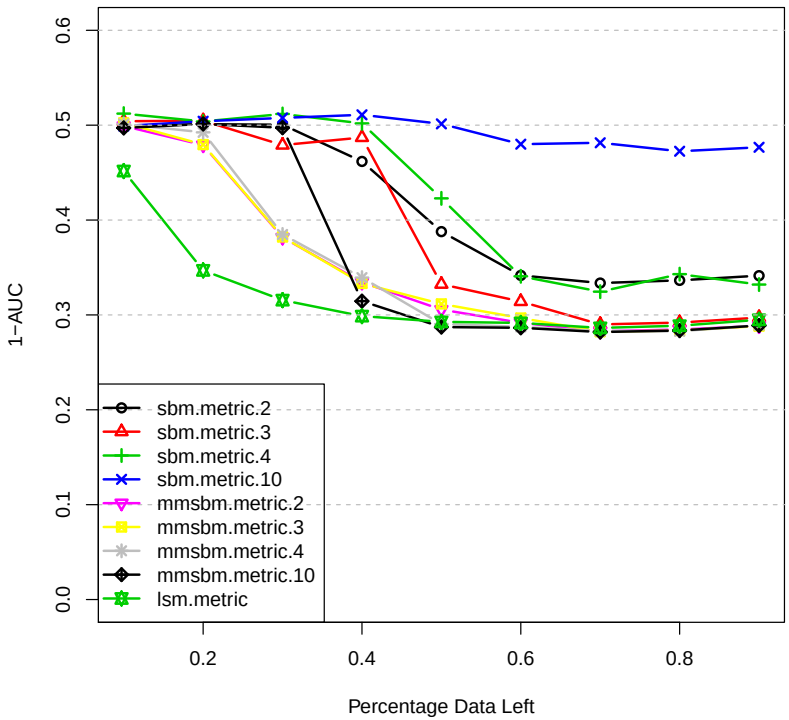
Details

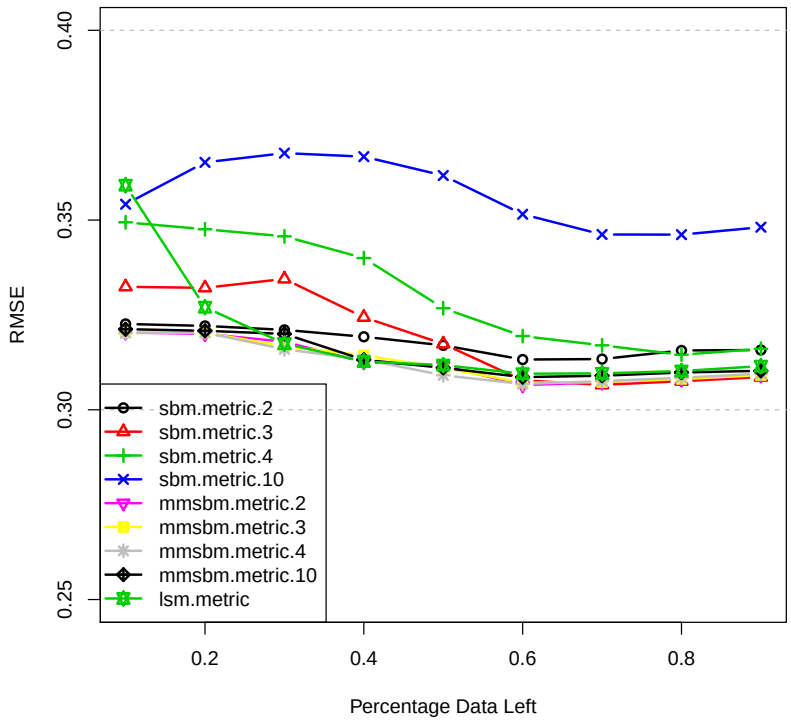
- RMSE Is a Standard Metric in Statistics
- AUC - Area Under ROC Curve
- AUC Cares About Ordering of Points

ROC Curve


True positive rate
False positive rate
0
0.2
0.4
0.6
0.8
1
NetChop C-term 3.0
TAP + ProteaSMM-i
ProteaSMM-i

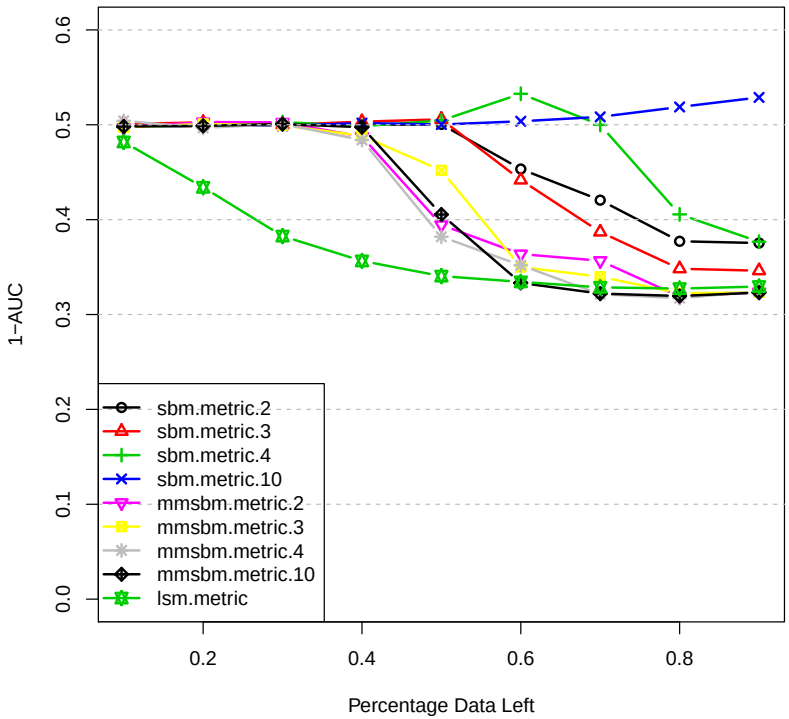
SBM - 3 Blocks - Leave-out Validation

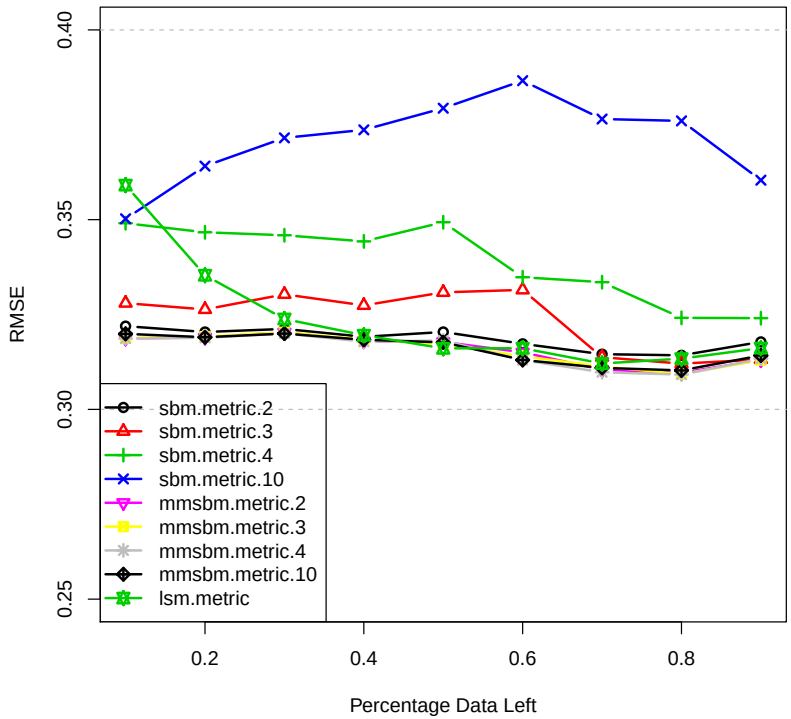




- All MMSBM fits and LSM achieve optimal risk
- SBM with two blocks underfits
- SBM with four blocks overfits

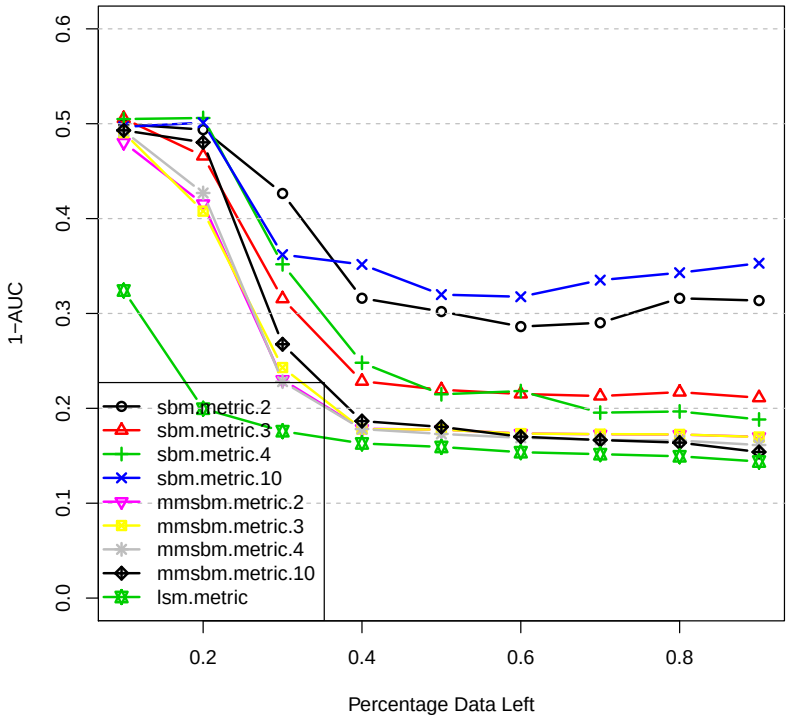
MMSBM - 3 Blocks - Leave-out Validation

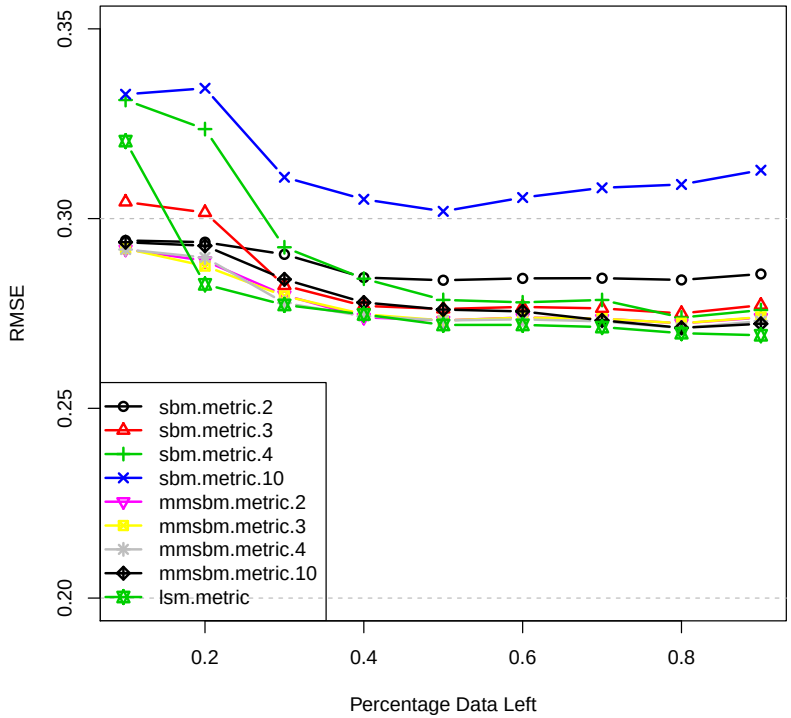




- Only Stochastic Block Model has poor fit
- Latent Space Model out-performs with low data

LSM - 3 Clusters - Leave-out Validation





- Latent Space Model is a relatively clear winner
- Four block SBM outperforms three block SBM
- Adding blocks to MMSBM does not increase risk

Conclusions

Goodness of Fit Techniques:

- Posterior Predictive checks are hard to interpret
- Posterior Predictive plots can detect model misspecification
- Out-of-Sample Validation techniques avoid some overfitting and underfitting

Comparison of Models:

- Latent Space Models can recover SBM and MMSBM properties well
- LSM communities have properties different from SBM and MMSBM communities
- SBM is most sensitive to under and overfitting
- MMSBM does not appear to overfit