Infinite Edge Partition Models for Overlapping Community Detection and Link Prediction

Mingyuan Zhou

Department of Information, Risk, and Operations Management
The University of Texas at Austin, Austin, TX, USA

Introduction

A hierarchical gamma process infinite edge partition model is proposed to factorize the binary adjacency matrix of an unweighted undirected relational network under a Bernoulli-Poisson link:

- The Bernoulli-Poisson link connects each edge to a latent count that is further partitioned. Each node is assigned to one or multiple latent communities depending on how its edges are partitioned.
- The model describes both homophily and stochastic equivalence, and is scalable to big sparse networks by focusing its computation on pairs of linked nodes.
- It can not only discover overlapping communities and inter-community interactions, but also predict missing edges.
- The number of communities is automatically inferred in a nonparametric Bayesian manner, and efficient inference via Gibbs sampling is derived using novel data augmentation techniques.

Model and Inference

❑ Modeling Components

Poisson Factor Analysis:

Modeling Assortativity: $m_{ij} \sim \mathrm{Po}\left(\sum_{k=1}^{K} r_k \phi_{ik}\phi_{jk}\right)$

Both assortativity and dissortativity: $m_{ij} \sim \mathrm{Po}\left(\sum_{k_1=1}^{K}\sum_{k_2=1}^{K} \phi_{ik_1}\lambda_{k_1k_2}\phi_{jk_2}\right)$

Bernoulli-Poisson Link:

Link binary to count: $b = \mathbf{1}(m \geq 1),\ m \sim \mathrm{Po}(\lambda)$

Marginal distribution: $b \sim \mathrm{Ber}\left(1 - e^{-\lambda}\right)$

Conditional posterior: $(m|b,\lambda) \sim b \cdot \mathrm{Po}_{+}(\lambda)$

Overlapping community structure:

$$m_{ij} = \sum_{k_1}\sum_{k_2} m_{ik_1k_2j},\ m_{ik_1k_2j} \sim \mathrm{Po}\left(\phi_{ik_1}\lambda_{k_1k_2}\phi_{jk_2}\right)$$

The count $m_{ik_1k_2j}$ represents how often nodes i and j interact due to their affiliations with communities k1 and k2, respectively.

❑ Hierarchical Gamma Process

$$\mathbf{B}|\Lambda \sim \mathrm{Ber}\left[1 - \prod_{k_1=1}^{\infty}\prod_{k_2=1}^{\infty}\exp\left(-\phi_{k_1}\lambda_{k_1k_2}\phi_{k_2}^{T}\right)\right]$$

$$\Lambda|G \sim \mathrm{r}\Gamma\mathrm{P}(G,\xi,1/\beta)$$

$$G \sim \Gamma\mathrm{P}(G_0, 1/c_0)$$

❑ Hierarchical Gamma Process Edge Partition Model

$$b_{ij} = \mathbf{1}(m_{ij} \geq 1),\ m_{ij} = \sum_{k_1=1}^{K}\sum_{k_2=1}^{K} m_{ik_1k_2j},$$

$$m_{ik_1k_2j} \sim \mathrm{Po}\left(\phi_{ik_1}\lambda_{k_1k_2}\phi_{jk_2}\right),$$

$$\phi_{ik} \sim \mathrm{Gam}(a_i, 1/c_i),\ a_i \sim \mathrm{Gam}(e_0, 1/f_0),$$

$$\lambda_{k_1k_2} \sim \begin{cases}\mathrm{Gam}(\xi r_{k_1}, 1/\beta), & \text{if } k_2 = k_1,\\ \mathrm{Gam}(r_{k_1}r_{k_2}, 1/\beta), & \text{if } k_2 > k_1,\end{cases}$$

$$r_k \sim \mathrm{Gam}(\gamma_0/K, 1/c_0),$$

❑ Scalability for Big Sparse Networks

Computation is mainly spent on pairs of linked nodes, as if b_ij=0, then all $m_{ik_1k_2j}$ are equal to zeros almost surely.

O(dN) instead of O(N^2), where d is the average node degrees.

❑ Gamma Process Edge Partition Model

$$b_{ij} = \mathbf{1}(m_{ij} \geq 1),$$

$$m_{ij} = \sum_{k=1}^{K} m_{ijk},\ m_{ijk} \sim \mathrm{Po}\left(r_k\phi_{ik}\phi_{jk}\right),$$

$$\phi_{ik} \sim \mathrm{Gam}(a_i, 1/c_i),\ a_i \sim \mathrm{Gam}(e_0, 1/f_0),$$

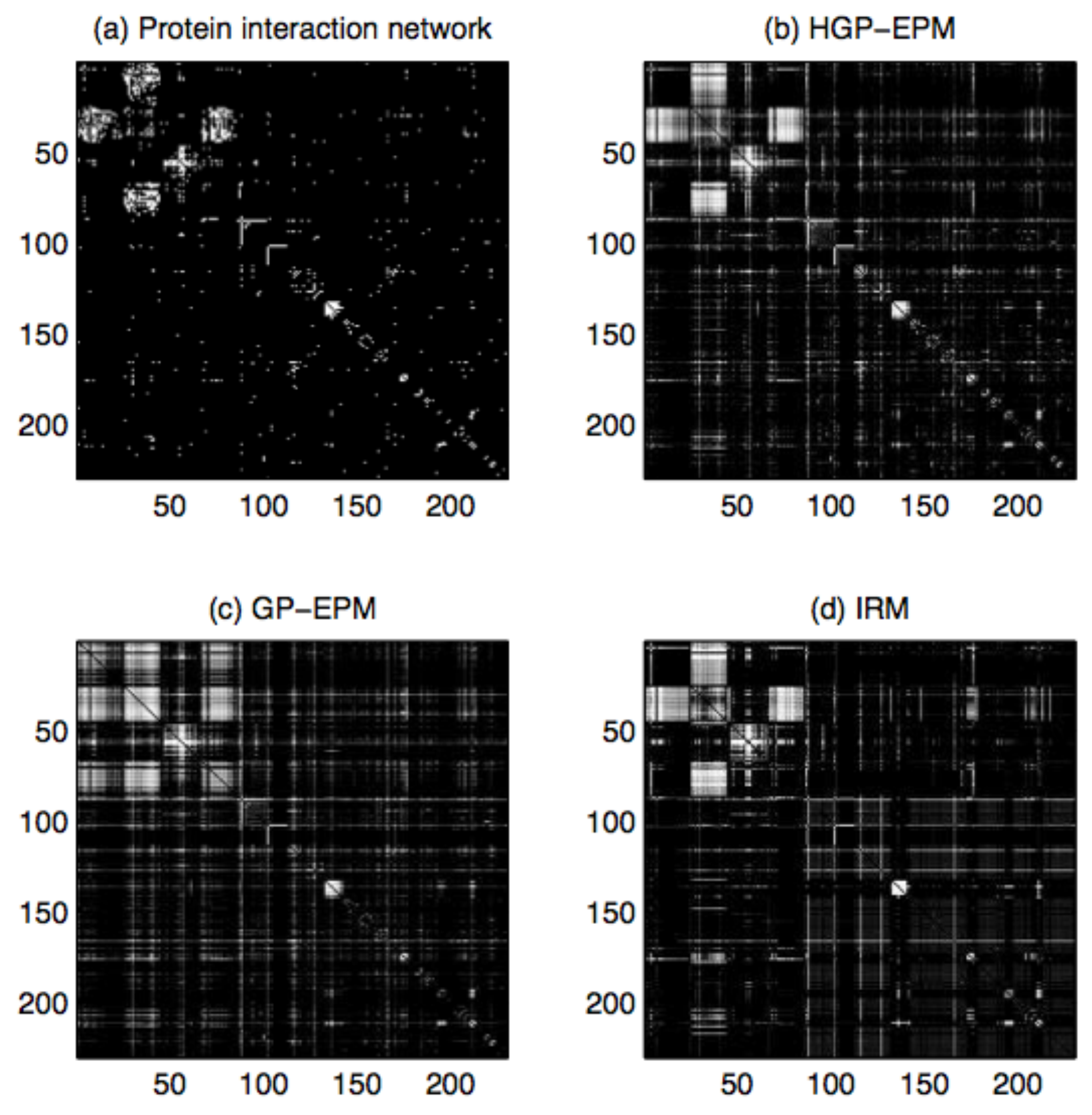
$$r_k \sim \mathrm{Gam}(\gamma_0/K, 1/c_0),\ \gamma_0 \sim \mathrm{Gam}(e_1, 1/f_1)$$

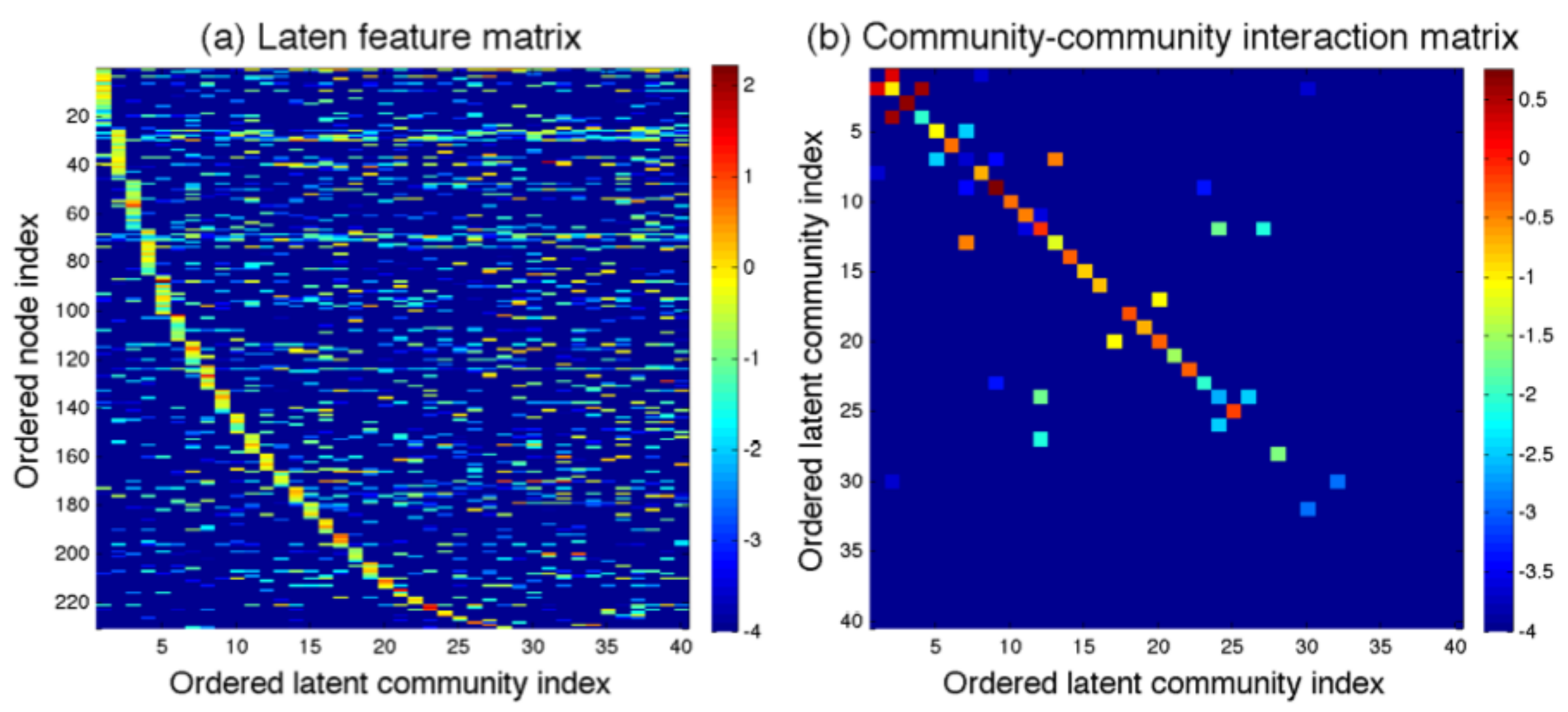
The community-affiliation graph model of Yang & Leskovec (2012) can be considered as a special case if we restrict $\phi_{ik} \in \{0,1\}$.

❑ Gibbs Sampling via Data Augmentation and Marginalization

Using inference techniques developed for the Bernoulli-Poisson link, and the Poisson, multinomial, and negative binomial distributions.

❑ Protein-Protein interaction network





Example Results

Table 1: Comparison of six algorithms on predicting missing edges of the Protein230 network. The Eigenmodel achieves its best performance at $K = 10$.

Model	AUC-ROC	AUC-PR
IRM	0.9338 ± 0.0128	0.5026 ± 0.0676
Eigenmodel	0.9314 ± 0.0188	**0.5468** ± 0.0500
ILA	0.8971 ± 0.0297	0.3693 ± 0.0234
AGM	0.9145 ± 0.0160	0.3339 ± 0.0359
GP-EPM	0.9335 ± 0.0110	0.4011 ± 0.0452
HGP-EPM	**0.9519** ± 0.0100	**0.5655** ± 0.0505

Table 2: Comparison of six algorithms on predicting missing edges of the NIPS234 coauthor network. The Eigenmodel achieves its best performance at $K = 10$.

Model	AUC-ROC	AUC-PR
IRM	**0.9476** ± 0.0114	0.6677 ± 0.0201
Eigenmodel	0.9269 ± 0.0177	0.6784 ± 0.0364
ILA	0.9171 ± 0.0222	0.6793 ± 0.0295
AGM	0.8906 ± 0.0164	0.5842 ± 0.0357
GP-EPM	**0.9501** ± 0.0123	**0.7415** ± 0.0319
HGP-EPM	**0.9469** ± 0.0163	**0.7289** ± 0.0540

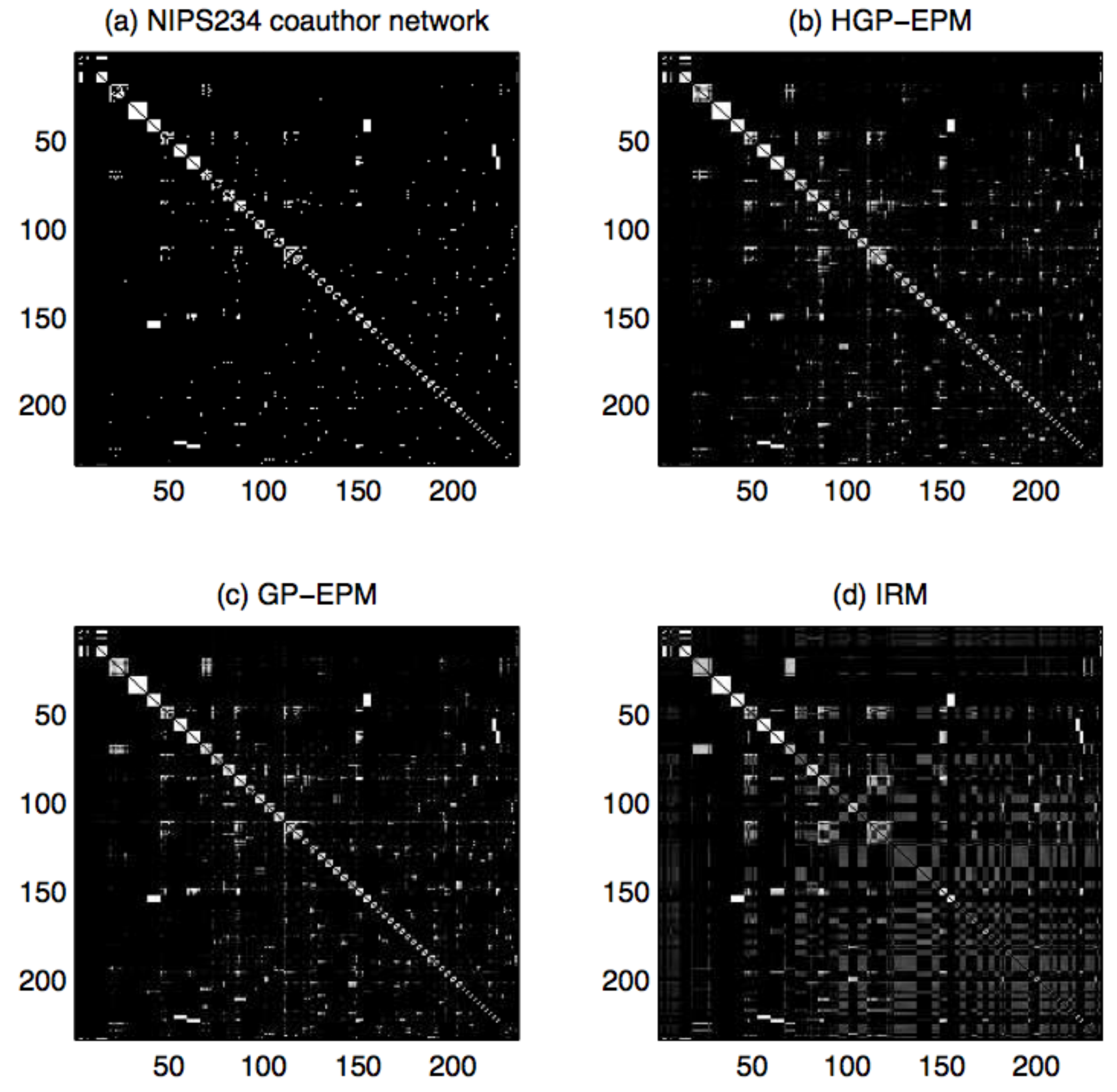
Table 3: Comparison of four algorithms on predicting missing edges of the Yeast protein interaction network.

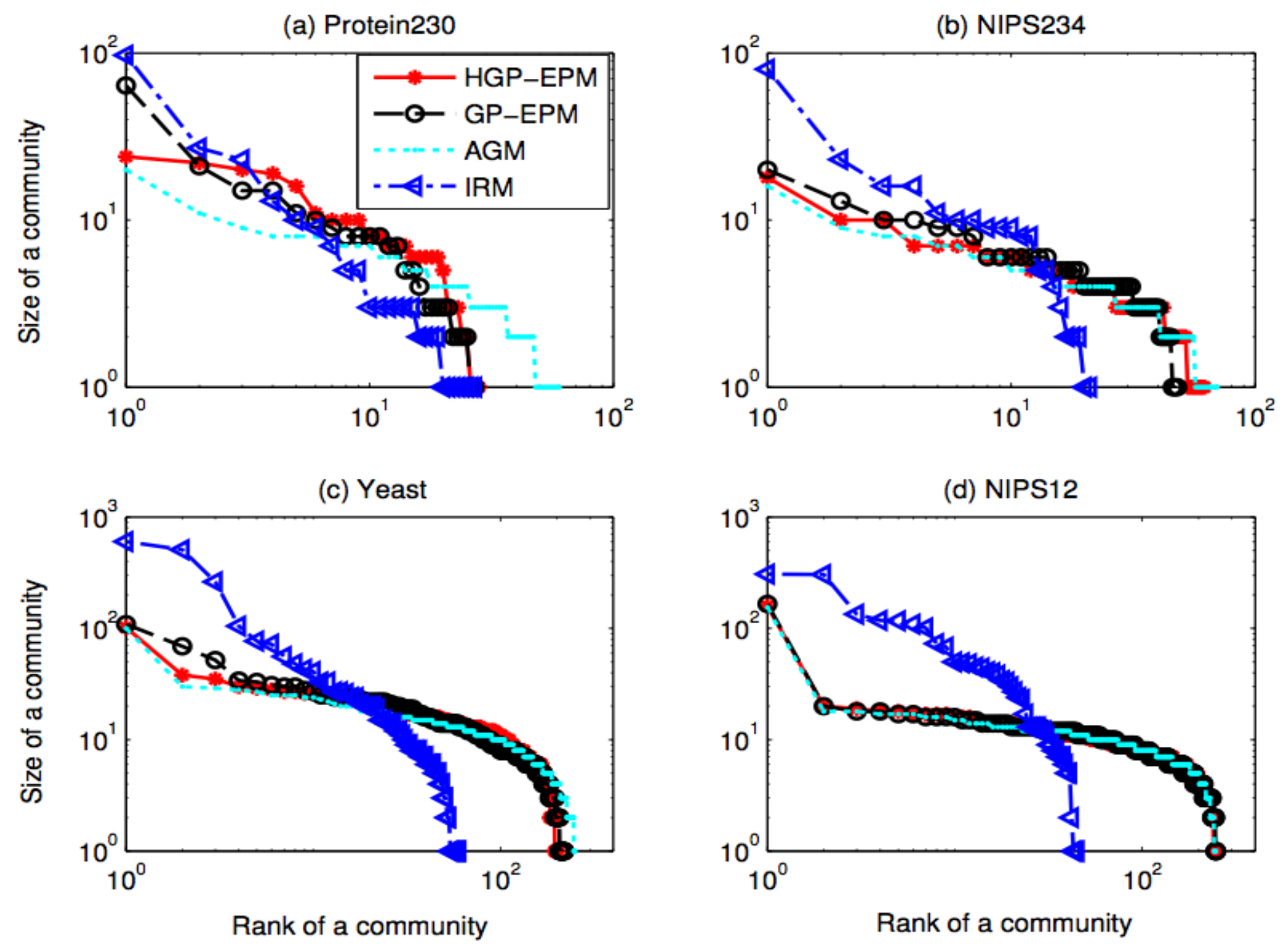
Model	AUC-ROC	AUC-PR
IRM	0.9093 ± 0.0059	0.1878 ± 0.0142
AGM	0.9009 ± 0.0025	0.1225 ± 0.0129
GP-EPM	0.9331 ± 0.0014	**0.2486** ± 0.0149
HGP-EPM	**0.9367** ± 0.0012	**0.2628** ± 0.0184

Table 4: Comparison of four algorithms on predicting missing edges of the NIPS12 coauthor network.

Model	AUC-ROC	AUC-PR
IRM	0.9427 ± 0.0121	0.2066 ± 0.0331
AGM	0.9328 ± 0.0049	0.2350 ± 0.0177
GP-EPM	**0.9768** ± 0.0079	**0.4705** ± 0.0362
HGP-EPM	**0.9762** ± 0.0081	**0.4493** ± 0.0229

❑ NIPS234 Coauthor network





2015