On Analyzing Estimation Errors due to Constrained Connections in Online Review Systems

Junzhou Zhao Xiaohong Guan Jing Tao
Xi'an Jiaotong University
CrowdRec, HK 2013

Background and the Problem

- Online Review Systems
  - Reviewers: spammer/non-spammer
  - Items: good/bad
- The *labels* of reviewer/item can be co-inferred using rating data.
- However, most of the works ignore the effects of online review system *topologies*,
  - i.e., the bipartite graph representing which reviewers can review which items.
- **Constrained Connections**: A reviewer usually only reviews a few items due to limited interest scope, attention capacity etc.

In this work

How constrained connections can affect the estimation performance in online review systems?

Background and the Problem

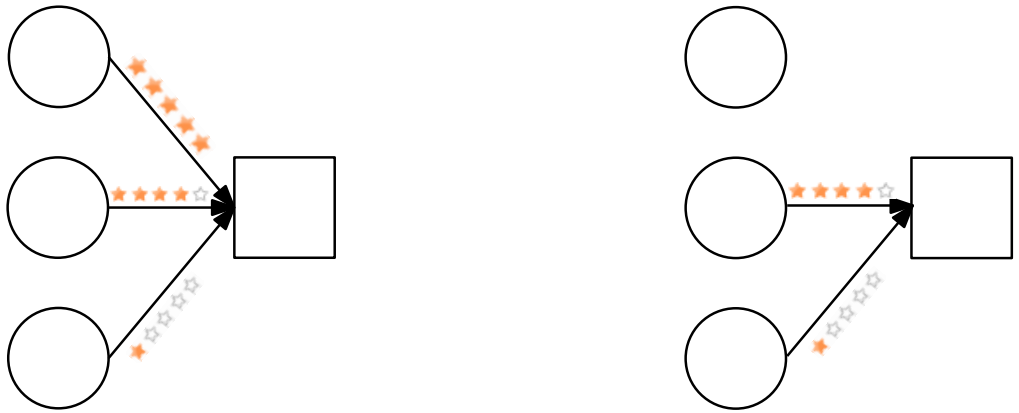
- Online Review Systems
  - Reviewers: spammer/non-spammer
  - Items: good/bad
- The *labels* of reviewer/item can be co-inferred using rating data.
- However, most of the works ignore the effects of online review system *topologies*,
  - i.e., the bipartite graph representing which reviewers can review which items.
- **Constrained Connections**: A reviewer usually only reviews a few items due to limited interest scope, attention capacity etc.

In this work

How constrained connections can affect the estimation performance in online review systems?

Why Constrained Connections can Affect Estimation?

An Intuitive Example



- In Figure (a), we can infer with high confidence that the item is probably good, and the bottom reviewer is likely to be a spammer.
- In Figure (b), due to constrained connections, we cannot give high confidence conclusions.

Data Model

- A set of reviewers V and a set of items I.
- Each item $i \in I$ has an unknown binary label $z_i \in \{\pm 1\}$.
- A review r_{ui} represents reviewer u's attitude to item i, e.g., $r_{ui} = +1$ if u thinks i to be good.
- Each reviewer u makes correct reviews with probability θ_u, and $P(\theta_u) \propto \theta_u^{\alpha-1}(1-\theta_u)^{\beta-1}$, where $\alpha > \beta$.
- Reviewer u can review item i if there is an edge $(u, i) \in E$ in bipartite graph $G(V, I, E)$.
- Given a collection of n reviews $R_n = \{r_1, \cdots, r_n\}$, our goal is to study how G affects the estimation of $\theta = \{\theta_u\}_{u \in V}$ and $z = \{z_i\}_{i \in I}$.

Maximum A Posteriori Estimator (MAPE)

- We treat θ as parameters and z as hidden variables, and use EM algorithm to estimate them:

$$\textbf{Objective: } \max_{\theta} \log P(\theta|R_n) = \max_{\theta} \log \sum_{z} P(\theta, z|R_n)$$

$$\textbf{E-Step: } \mu_i(z_i) \equiv P(z_i|R_{\cdot i}, \theta) \propto P(z_i) \prod_{u \in V_i} \theta_u^{n_{ui}^{z_i}} (1-\theta_{ui})^{n_{ui}^{-z_i}},$$

$$\textbf{M-Step: } \theta_u^{(t+1)} = \frac{\sum_{i \in I_u} \sum_{x \in \{\pm 1\}} n_{ui}^{x} \mu_i(x) + \alpha - 1}{|R_{u\cdot}| + \alpha + \beta - 2}.$$

- The EM algorithm implicitly defines an estimator of θ, i.e., $\hat{\theta}_{\text{MAP}} = \text{EM}(R_n)$.

Estimation Errors due to G

- To understand how G affects $\hat{\theta}_{\text{MAP}}$, we can calculate MSE of $\hat{\theta}_{\text{MAP}}$ over different graphs. ($\text{MSE}(\hat{\theta}_u) = \mathbb{E}\left[\hat{\theta}_u - \theta_u\right]^2$)
- MSE of $\hat{\theta}_{\text{MAP}}$ is lower bounded by the Bayes Cramer-Rao bound requiring that the estimator is weakly unbiased.
- The posterior distribution of θ can be approximated by normal distribution, i.e.,

$$P(\theta|R_n) \to \mathcal{N}(\hat{\theta}_{\text{MAP}}, \mathcal{I}(\hat{\theta}_{\text{MAP}})^{-1}) \quad \text{as } n \to \infty,$$

where $\mathcal{I}(\hat{\theta}_{\text{MAP}})$ is the *observed Fisher information matrix*, and each element is

$$\left[\mathcal{I}(\hat{\theta}_{\text{MAP}})\right]_{uv} = -\frac{\partial^2 \log P(\theta|R_n)}{\partial \theta_u \partial \theta_v}\Big|_{\theta=\hat{\theta}_{\text{MAP}}}.$$

Obtaining the Observed Fisher Information Matrix

Theorem

The observed Fisher information matrix is a diagonal matrix, with each diagonal element given by

$$\mathcal{I}_{uu} = \frac{\alpha - 1}{\hat{\theta}_u^2} + \frac{\beta - 1}{(1 - \hat{\theta}_u)^2}.$$

- Given hyper-parameters α and β, θ_u is most uncertain at $\frac{1}{1+\sqrt[4]{(\beta-1)/(\alpha-1)}}$, and most certain at 0 or 1.

Experiments and Observations I

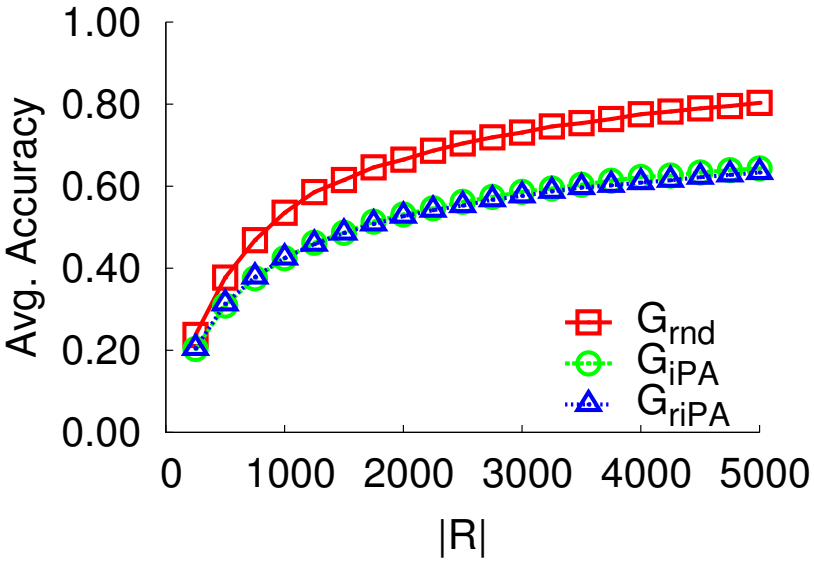
Bipartite Graph Models

- Random Model G_{rnd}:
  - Each edge (u, i) is formed by randomly connecting a reviewer u and item i.
- Item Preferential Attachment Model G_{iPA}:
  - For edge (u, i), u is chosen randomly, and i is chosen with probability proportion to i's degree.
  - This mimics real-world situation that popular items are more likely to be reviewed.
- Reviewer and Item Preferential Attachment Model G_{riPA}:
  - For edge (u, i), u and i are chosen with probability proportion to their degrees in the graph.

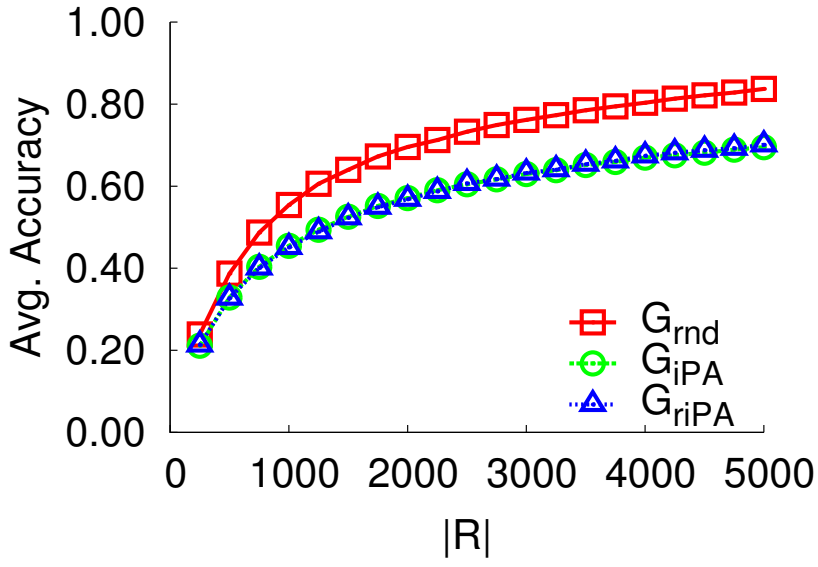
Remarks:

- From G_{rnd} to G_{iPA} and G_{riPA}, we put more constraints on the topology.
- How do constraints affect estimation errors?

Experiments and Observations II



(a) $|E| = 2000$

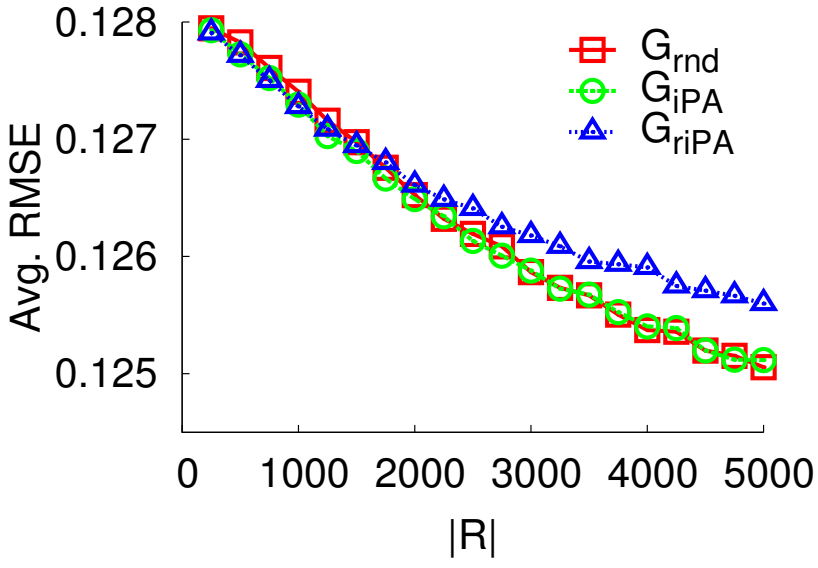


(b) $|E| = 3000$

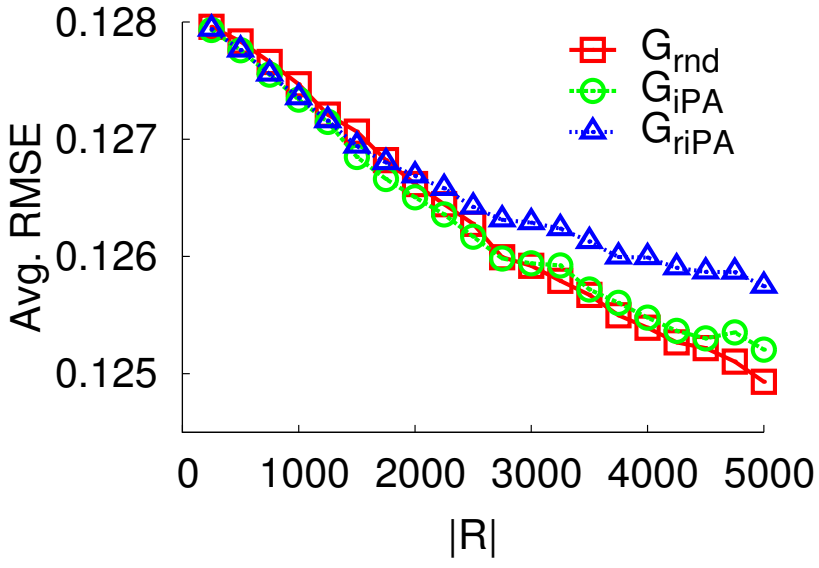
Estimation Accuracy Curve:

- Accuracy is the fraction of items can be correctly inferred.
- MAPE is asymptotically unbiased.
- Accuracy on G_{rnd} is larger than the others: **more constraints make estimation accuracy poor**.

Experiments and Observations III



(c) $|E| = 2000$



(d) $|E| = 3000$

RMSE (RMSE $= \sqrt{\text{MSE}}$) Lower Bound:

- The RMSE decreases approximately with rate $\frac{1}{n}$.
- When n is large, RMSE on G_{riPA} is largest, then comes G_{iPA} and RMSE on G_{rnd} is the lowest: **more constraints implies harder to be estimated.**

Conclusion and Future Work

- More constraints on the topology usually cause worse estimation performance.
- Suppose the constrains are **fixed**, is there an optimal topology that has the lowest RMSE? How can we find this topology?

Thank you!

Feel free to contact me if you have any questions:
junzhouzhao@gmail.com