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

Junzhou Zhao   Xiaohong Guan   Jing Tao
MOE KLINNS Lab, Xi'an Jiaotong University
{jzzhao,xhguan,jtao}@sei.xjtu.edu.cn

ABSTRACT

Constrained connection is the phenomenon that a reviewer can only review a subset of products/services due to narrow range of interests or limited attention capacity. In this work, we study how constrained connections can affect estimation performance in online review systems (ORS). We find that reviewers' constrained connections usually cause poor estimation performance, both from the measurements of estimation accuracy and Bayesian Cramér Rao lower bound.

1. INTRODUCTION

Online reviews are more and more important factors for customers to decide whether to buy a product or service in online markets. However, "Internet Water Mercenaries" which are also known as paid spammers, write fake reviews to disturb customers' judgments on quality of products and damage company's reputation. Hence, an important problem is how to obtain the *truths* of both reviewers (e.g., the reviewer is a spammer or non-spammer) and items (e.g., the product is good or bad) according to unreliable reviews.

Existing studies ignore the function of the underlying topology of ORS. The topology of an ORS is a bipartite graph representing which reviewers can review which items. Other than that reviewers can review all the items (e.g., the example in Fig. 1(a)), a reviewer can only review a subset of items in real-world, which results in *constrained connections* for each reviewer in the topology. The constrained connections may be because of either the reviewer's narrow range of interests or the reviewer's limited attention capacity. The topology of ORS can affect the performance of jointly estimating the truths of reviewers and items. For example, let us consider a simplest ORS consisting of three reviewers and one item. If we assume that the majority of reviewers are non-spammers, then in case Fig. 1(b), from this topology and reviews by reviewers we can infer with high confidence that the item is probably good and the bottom reviewer is likely to be a spammer. However, in the case of Fig. 1(c), we cannot obtain a high confidence conclusion because we do not know the reviews of the top reviewer.

This simple example tells us that different topologies of ORS along with unreliable reviews contain different amounts of information for jointly estimating the truths of reviewers and items. Actually, connections between reviewers and items act as constraints in such systems. They constrain the joint probability distribution of the truths of reviewer-item pairs they connect. For example, a non-spammer usually gives good (bad) items good (bad) reviews with high probability, which indicates that the truth of a reviewer and the truth of an item he reviewed are related. Hence the topology of the ORS yields a set of constraints that the truths of reviewers and items must obey, and these constraints help to reduce the *uncertainty* of the system, with varying abilities.

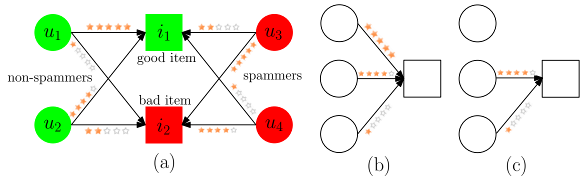


Figure 1: Illustrative examples.

In order to compare the amounts of information contained in different topologies (and reviews), we calculate the Bayesian Cramér Rao lower bound (BCRLB) of maximum a posteriori estimator (MAPE) in such systems for different bipartite graphs. We find that estimation errors vary for different topologies. This indicates that for some topologies the truths become difficult to be estimated by any MAPEs.

2. DATA MODEL AND ESTIMATOR

2.1 Data Model

Following the existing works, we assume that an ORS consists of a set of reviewers V and a set of items I. Each item $i \in I$ has an unknown binary label $z_i \in \{\pm 1\}$ representing the quality of i, e.g., $z_i = +1$ if i is *good*; $z_i = -1$ if i is *bad*. Each reviewer can review items. A review, denoted by $r_{ui} \in \{\pm 1\}$, represents the reviewer's attitude to an item, i.e., if u considers i to be good (or bad), then $r_{ui} = +1$ (or $r_{ui} = -1$). Since reviewers' attitudes are not always correct, we use $\theta_u \in [0, 1]$ to represent the probability that u can give correct reviews, i.e., $\theta_u = P(r_{ui} = z_i)$. In practice, it is reasonable to assume that the majority of reviewers have $\theta_u > 0.5$. This is achieved by putting a prior distribution on θ_u. A convenient choice of such a prior is the beta distribution, i.e., $P(\theta_u) \propto \theta_u^{\alpha-1}(1-\theta_u)^{\beta-1}$, where $\alpha > \beta$.

Different from previous works, we assume that a user u can only review item i if there exists an edge $(u, i) \in E$ in bipartite graph $G(V, I, E)$, where E is the set of edges. To make the model more general, we assume that items are chosen independently *with replacement* by reviewers and constrained by G. This forms a collection of n review samples $R = \{r_1, r_2, \cdots, r_n\}$ where r_k denotes the k-th sample representing some reviewer u gives some item i a review r_{ui}. Since an item can be reviewed many times by a reviewer, we use n_{ui}^x to represent the number of times u gives i a review x in R. Our goal is to study *how G can affect the estimation when using R to estimate $\theta = \{\theta_u\}_{u \in V}$ and $z = \{z_i\}_{i \in I}$.*

2.2 Maximum A Posteriori Estimator (MAPE)

A convenient way to estimate θ and z is that we treat θ as parameters and z as hidden variables. David and Skene[1] presented an expectation maximization (EM) approach to maximize the likelihood. Here we propose a different approach to maximize the posteriori of θ which has the benefit of including priori information of θ. That is,

$$\text{Objective: } \max_{\theta} \log P(\theta|R) = \max_{\theta} \log \sum_{z} P(\theta, z|R).$$

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

$$\text{M-Step: } \theta_u^{(t+1)} = \frac{\sum_{i\in I_u} \sum_{x\in\{\pm1\}} n_{ui}^{x} \mu_i(x) + \alpha - 1}{|R_{u\cdot}| + \alpha + \beta - 2}.$$

Here, $R_{u\cdot}$ is the set of reviews given by reviewer u.

The EM algorithm implicitly defines an estimator of θ, i.e., $\hat{\theta}_{\text{MAP}} = \text{EM}(R)$, which is also related to G. To understand how G can affect MAPE, we need to analyze the Mean Squared Errors (MSE) of $\hat{\theta}_{\text{MAP}} = \{\hat{\theta}_u\}_{u\in V}$.

3. ESTIMATION ERROR ANALYSIS

3.1 Lower Bound on MSE

MSE of $\hat{\theta}_u$ is defined as $\text{MSE}(\hat{\theta}_u) = \mathbb{E}[\hat{\theta}_u - \theta_u]^2$, which is lower bounded by the Bayesian Cramér Rao lower bound (BCRLB) requiring that $\hat{\theta}_{\text{MAP}}$ is *weakly unbiased*[2, Chapter 2], which is unknown for the above MAPE. However, it is well known that under general conditions, for large n, the posterior distribution of θ can be approximated by normal distribution, $P(\theta|R) \to \mathcal{N}(\hat{\theta}_{\text{MAP}}, \mathcal{I}(\hat{\theta}_{\text{MAP}})^{-1})$ as $n \to \infty$, where $\mathcal{I}(\hat{\theta}_{\text{MAP}})$ is the *observed Fisher information matrix*, and each element is $[\mathcal{I}(\hat{\theta}_{\text{MAP}})]_{uv} = -\frac{\partial^2 \log P(\theta|R)}{\partial\theta_u \partial\theta_v}\big|_{\theta=\hat{\theta}_{\text{MAP}}}$.

The usefulness of this result is that $\hat{\theta}_{\text{MAP}}$ is a consistent estimator of θ with large-sample covariance matrix $\mathcal{I}^{-1}$. Hence, to study G's effect on estimation errors, we only need to evaluate $\mathcal{I}^{-1}$, which tells the variance of estimator (for large n). Next, we present how to obtain $\mathcal{I}$.

3.2 Obtaining the Observed Fisher Information Matrix

Theorem 1 below presents how to calculate the observed Fisher information matrix. The proof is included in [3].

Theorem 1. *The observed Fisher information matrix $\mathcal{I}$ is a diagonal matrix, with each diagonal element*

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

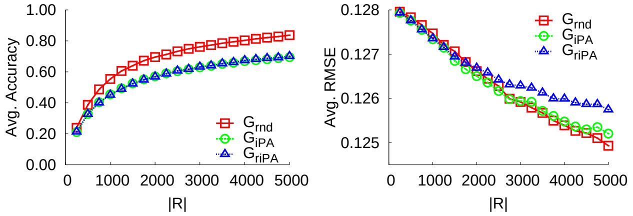
Note that Eq. (1) is convex, $\mathcal{I}_{uu}$ gets the minimum value at $\hat{\theta}_u^* = \frac{1}{1+\sqrt[4]{(\beta-1)/(\alpha-1)}}$ and $\mathcal{I}_{uu}$ gets the maximum value at $\hat{\theta}_u = 0$ or 1. This tells us that $\hat{\theta}_u$ is most *uncertain* when $\hat{\theta}_u = \hat{\theta}_u^*$ and most *certain* at $\hat{\theta}_u = 0$ or 1.

4. EMPIRICAL RESULTS

We present three bipartite graph models, and study the effects of topologies to estimator performance.

4.1 Bipartite Graph Models

Random Model G_{rnd}: Each edge (u, i) is formed by randomly connecting a reviewer u and an item i.



(a) Accuracy (b) RMSE

Figure 2: Estimation performance ($|E| = 3000$).

Item Preferential Attachment Model G_{iPA}: For edge (u, i), u is randomly chosen, i is chosen with probability proportion to i's degree in G_{iPA}. 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 G_{riPA}. This also mimics that active reviewers are more likely to review items.

4.2 Comparing Estimation Accuracy

In the first experiment, we compare the accuracy of classifying items under different graph models. We set an item i with $z_i = +1$ (or -1) if $\mu_i(+1) > 0.5$ (or < 0.5). The accuracy is the fraction of items that can be correctly inferred. We first generated graphs with number of nodes $|V| = 500$ and varying number of edges (we only show $|E| = 3000$ here) using different models. For each graph, we generated review samples of different sizes ($500 \le |R| \le 5000$)[3], and show the accuracy averaged over 100 experiments in Fig. 2(a). We observe that when $|R|$ increases, the accuracy increases and approaches 1. This confirms that MAPE is asymptotically unbiased. We also observe that the accuracy on G_{rnd} is larger than the others. This indicates that constrained connections will make the estimation accuracy poor.

4.3 Comparing Estimation Errors

In the second experiment, we study how different graphs affect estimation errors. The settings are same as in the previous experiment. We compare the average Rooted Mean Squared Error (RMSE, defined as $\text{RMSE} = \sqrt{\text{MSE}}$) over different graphs in Fig. 2(b). The RMSE decreases approximately with rate $1/n$ over all the graphs. For different graphs, when n is large, RMSE on G_{riPA} is largest, then comes G_{iPA} and RMSE on G_{rnd} is lowest. This indicates that when more constraints are added to graphs, the RMSE becomes larger, which means the parameters are more difficult to be estimated by any MAPEs.

5. CONCLUSION

The constrained connections are common in real world. We find that it will cause poor inference performance, both from the measurements of estimation accuracy and estimation error. Hence, it is necessary to find out the optimal topology in such systems, which will be left as future works.

6. REFERENCES

[1] A. P. Dawid and A. M. Skene. MLE of observe error-rates using the EM Alg. *JSTOR*, 1979.

[2] H. Van Trees. *Detection, Estimation, and Modulation Theory.* Wiley, 1st edition, 1968.

[3] J. Zhao. Supplementary information. Arxiv 1307.3687, 2013.