Concentration Inequalities and its Application to Ranking Problem

Muskaan

MS14072

A dissertation submitted for the partial fulfillment of BS-MS dual degree in Science

Indian Institute of Science Education and Research, Mohali

Certificate of Examination

This is to certify that the dissertation titled "Concentration Inequalities and its Application to Ranking Problem" submitted by Ms. Muskaan (Reg. No. MS14072) for the partial fulfilment of BS-MS dual degree programme of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr. Amit Kulshrestha Dr. Lingaraj Sahu Dr. Neeraja Sahasrabudhe

(Supervisor)

Dated: April 25, 2019

Declaration

The work in this dissertation has been carried out by me under the guidance of Dr. Neeraja Sahasrabudhe at the Indian Institute of Science Education and Research Mohali.

This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.

Muskaan

(Candidate)

Dated: April 25, 2019

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr. Neeraja Sahasrabudhe
(Supervisor)

Acknowledgement

I would like to express my gratitude to Dr. Neeraja Sahasrabudhe for giving me the opportunity to work on a project of my interest. I thank her for her constant guidance and invaluable support.

I would like to thank all my friend for making 5 years wonderful experience and encouraging me to perform better. I also thank my family for the constant motivation, having an unwavering belief in me, and bearing my expenses.

Muskaan

Contents

Abstract

In this thesis, we study some basic concentration inequalities and their applications to a ranking problem. Concentration inequalities refer to the phenomenon of concentration of a function of independent random variables around the mean. In this thesis, we mainly study how the sum of independent random variables concentrate around the mean. These inequalities are used to study error bounds for estimated ranks in the BTL model [SSD17], which gives a framework to determine the ranks for n objects based on k pair-wise comparisons between pairs of objects. We then study the effect of perturbing the transition matrix of a defined Markov chain on the errors in the estimated rank. In some cases, we obtain an explicit lower bound on the number of comparisons k, in terms of the perturbations, needed to obtain a "good" estimation for the underlying rank. Finally, through simulations, we study what kind of perturbation matrices lead to larger errors.

Introduction

Concentration inequalities provide bounds on how independent or "weakly dependent" random variables or their functions deviate from a particular value. Such bounds are extremely useful in applied probability and statistics. The law of large numbers states that the sum of independent and identically distributed random variables converges to the common expectation as number of random variables increases. We also know, from the Central Limit Theorem, that appropriately scaled sample mean converges to standard normal random variable in distribution. Concentration inequalities for such sums essentially provide a bound on how close does the mean of n random variables get to the common expectation μ. For example, the simplest inequality in probability theory - the Markov's inequality, deals with the probability that a non-negative random variable is greater than a particular constant $a > 0$. Markov's inequality in turn gives us a bound on the deviation of montonically increasing non negative functions of random variables from a positive constant. The Chernoff bound is a concentration inequality that gives exponentially decreasing bounds on the tails of sums of independent random variables. Such bounds on sums of random variables are most elementary examples of concentration of random variables. Over the last few decades, work done in this area demonstrates that such structured behaviour is exhibited by a much larger category of general functions of independent random variables. In general, the idea of concentration is to find appropriate (and hopefully sharp) bounds on deviation of a random function $f(X_1, \ldots, X_n)$ from its expectation $E[f(X_1, \ldots, X_n)]$. This turns out to be very useful in cases where one wants to study the behaviour of a very general function that might be very complicated to compute.

While this project started as a study of various concentration inequalities including bounds on variance of certain functions of random variables and functions with bounded differences, we have chosen to omit most of these inequalities in interest of space. An interested reader can refer to [SPG12] for details. We have only stated the concentration results that are subsequently used in estimating errors in the ranking problem. The ranking problem studied in this thesis is the BTL model described in [SSD17]. Suppose we have n items and k pairwise comparisons between pair of items. It is assumed throughout that k is same for each pair. This can be generalized to number of comparisons being different for different pair of items. Based on these k pairwise comparisons, we want to determine (upto certain error) the underlying ranking of items. In [SSD17], this estimate is obtained

by defining a Markov chain on the graph with nodes corresponding to each item and edges corresponding to the pair of nodes for which the pairwise comparisons are available. The stationary distribution of this Markov chain provides a "good" estimate for the ranks of items. The advantage of using Markov chains is that they are well studied in literature, and their convergence properties can be explicitly characterized. The error bounds for rank the estimate are obtained using the tools from concentration inequalities. The crucial step is to provide a "good" bound on the spectral radius of a matrix. While we have only discussed concentration for random variables, a lot of results can be extended to sum of random matrices [J11].

Our aim is to look at this ranking problem from an adversarial point of view and determine how much error should be introduced in the transition matrix corresponding to Markov chain discussed above to produce significant and specific error in the final estimated rank. We use the same methods as in [SSD17], and come up with new bounds for the error in the estimated rank in presence of perturbations. The larger problem is to come up with strategies for the adversary to perturb the rank in a specific pre-defined manner to obtain a target rank that differs from the actual rank in a way desired by the adversary. For example, an adversary might be interested in permuting the ranks of a subset of nodes. With this aim in mind, we study some examples of perturbation matrices and explicitly compute the final rank to illustrate how perturbation affects the rank.

This thesis is organized as follows: In chapter 1, we discuss some very basic and very well-known concentration inequalities. An mentioned earlier, we have chosen to mention only a few basic ones in this chapter. This has been done keeping in mind the usefulness of these results in subsequent chapters and to maintain a more legible flow. More details on the topic can be found in [SPG12]. The second chapter deals with the study of the BTL model [SSD17], the estimation of pair-wise comparison ranks and comparing the bounds on spectral radius of error matrix by using three different concentration inequalities. In the third chapter, we discuss the effects of perturbation of the transition matrix of the Markov chain defined to estimate the underlying rank of the objects. The final chapter details some examples and simulations done to disturb the actual rank, and illustrate what kind of perturbation matrices lead to what kind of effects on the final estimated rank.

Chapter 1

Concentration Inequalities

Consider independent and identically distributed random variable $X_1, X_2, \ldots$ such that $E[X_i] = \mu$ and $var(X_i) = \sigma^2 < \infty$. Define $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. We want to know near close does $\overline{X_n}$ get to μ . Central limit theorem and Law of large numbers provide an asymptotic understanding. Strong law of large numbers says that $\overline{X_n}$ converges almost surely to μ as $n \to \infty$. Central limit theorem tells us that with appropriatly scaling the behaviour (in distance) is like a standard Gaussian. However, the asymptotic results are not so in real life problems since in an actual problem we would want to know how close does the sum of $X_i's$ get to the mean. In other words, we would like to have explicit bounds on the $|\overline{X_n} - \mu|$ in terms of n. The class of such bounds is known as "Concentration Inequalities". In this chapter, we discuss some basic concentration inequalities and reproduce the proofs of some well-known results.

1.1 Markov's Inequality

Theorem 1.1.1. *(***p.19, [SPG12]***) Let X be a non-negative random variable and* $a \geq 0$*, then*

$$P(X \geqslant a) \leqslant \frac{E[X]}{a}.$$

Proof. Let $\mathbb{1}$ be an indicator function. It is clear that for all $a \geq 0$, $E[a\mathbb{1}_{X \geq a}] \leq E[X]$. Also,

$$\begin{aligned} aE[\mathbb{1}_{X \geq a}] &= a(1.P(X \geq a) + 0.P(X \leq a)) \\ &= aP(X \geq a), \end{aligned}$$

which implies that

$$P(X \geqslant a) \leqslant \frac{E[X]}{a}.$$

□

1.2 Chernoff Bound

Theorem 1.2.1. *(**p.21, [SPG12]**) Let X be a non-negative and sum of independent random variables. Then,*

$$P(X \geq a) \leq min\{e^{ta} \prod_i E[e^{-tX_i}]\}_{t>0}.$$

Proof. Consider e^{tX}. It is a nondecreasing and nonnegative function therefore, by Markov's inequality we get that for every $t > 0$

$$\begin{aligned} P(X \geqslant a) &= P(e^{tX} \geqslant e^{ta}) \\ &\leqslant \frac{E[e^{tX}]}{e^{ta}} \\ &\leqslant \frac{E\left[\prod_i e^{tX_i}\right]}{e^{ta}} \\ &\leqslant \frac{\prod_i E[e^{tX_i}]}{e^{ta}}. \end{aligned}$$

We get for any t,

$$P(X \geqslant a) \leqslant \frac{\prod_i E[e^{tX_i}]}{e^{ta}}.$$

This implies that

$$P(X \geqslant a) \leqslant min_{t>o} \frac{\prod_i E[e^{tX_i}]}{e^{ta}}.$$

□

1.3 Hoeffding's Inequality

Theorem 1.3.1. *(**Theorem 2.8, [SPG12]**) Let $X_0, X_1, \ldots$ be independent random variables bounded by the interval $[a_i, b_i]$. Let $Y_N = \sum_{i=1}^{n} X_i$. Then,*

$$P(Y_N - E[Y_N] \geq t) \leq 2exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Proof. Let $Y_N = X_1 + X_2 + \cdots + X_n$.

Then for all $s,t > 0$, the independence of X_i and Markov's inequality implies that

$$\begin{aligned} P(Y_N - E[Y_N] \geq t) &= P(Y_N - E[Y_N] \geq t) \\ &= P(e^{s(Y_N - E[Y_N])} \geq e^{st}) \\ &\leq e^{-st} E\left[e^{s(Y_N - E[Y_N])}\right] \\ &= e^{-st} \Pi_{i=1}^{n} E\left[e^{s(X_i - E[X_i])}\right] \\ &\leq e^{-st} \Pi_{i=1}^{n} e^{\frac{s^2(b_i-a_i)^2}{8}} \quad \textit{(by Hoeffding's Lemma)} \\ &= exp\left(-st + \frac{1}{8}s^2 \sum_{i=1}^{n} (b_i - a_i)^2 \right). \end{aligned}$$

Now we will find the minimum of the $k(s) = -st + \frac{1}{8}s^2 \sum_{i=1}^{n} (b_i - a_i)^2$ in order to get the best upper bound. As k achieves its minimum at $s = \frac{4t}{\sum_{i=1}^{n} (b_i - a_i)^2}$, we get

$$P(Y_N - E[Y_N] \geq t) \leq 2exp\left(\frac{-2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right).$$

□

1.4 Martingales Concentration

So far we have assumed independence of random variable. In this section, we consider a special class of random processes known as Martingales and state concentration results for them.

Definition 1.4.1. *A sequence $X_0, X_1, \ldots$ is said to be a martingale if for every n,*

$$E[X_{n+1} | X_1, X_2, \ldots, X_n] = X_n.$$

Definition 1.4.2. *A sequence $X_0, X_1, \ldots$ is said to be a supermartingale if for every n,*

$$E[X_{n+1} | X_1, X_2, \ldots, X_n] \leq X_n.$$

Definition 1.4.3. *A sequence $X_0, X_1, \ldots$ is said to be a submartingale if for every n,*

$$E[X_{n+1} | X_1, X_2, \ldots, X_n] \geq X_n.$$

1.4.1 Bounded difference property

Theorem 1.4.1. *(***Lemma 5.1, [DA09]***) Let $X_0, X_1, \ldots$ be a Martingale. The X_i's satisfy the bounded difference condition with parameters a_i and b_i if $a_i \leq X_i - X_{i-1} \leq b_i$ for some reals $a_i, b_i, i > 0$. Consider a random variable Y with $E[Y] = 0$ and $a \leq Y \leq b$ for some reals a and b. Then,*

$$E[e^{\lambda Y}] \leqslant e^{\frac{\lambda^2 (b-a)^2}{8}}.$$

Proof. For any $\lambda > 0$, $e^{\lambda x}$ is convex in the interval (a, b) , therefore its graph lies entirely at a lower position than the line joining $(a, e^{\lambda a})$ and $(b, e^{\lambda b})$. Thus for $a \leq Y \leq b$,

$$e^{\lambda Y} \leqslant \frac{Y-a}{b-a} e^{\lambda a} + \frac{b-Y}{b-a} e^{\lambda b},$$

taking the expectation on both sides, we get

$$E[e^{\lambda Y}] \leqslant E\left[\frac{Y-a}{b-a} e^{\lambda a} + \frac{b-Y}{b-a} e^{\lambda b}\right].$$

This implies that

$$E[e^{\lambda Y}] \leqslant \frac{b}{b-a} e^{\lambda b} - \frac{a}{b-a} e^{\lambda a}.$$

Put $s = \frac{-a}{b-a}$ and $y = \lambda(b-a)$. We get

$$E[e^{\lambda Y}] = (1-s)e^{sy} + se^{(1-s)y}.$$

Let $K(x) = -sx - ln(1-s+se^x) = e^{K(y)}$. Then

$$K^{'} = -s + \frac{s}{s+(1-s)e^{-x}},$$

$$K^{''} = \frac{s+(1-s)e^{-x}}{s+(1-s)e^{-x}}^2 \leqslant \frac{1}{4},$$

also $K(0) = 0 = K^{'}(0)$.
At last by Taylor's theorem, we get

$$K(y) = K(0) + K^{'}(0) + K^{''}(k)\frac{t^2}{2} \leqslant 0 + 0 + \frac{1}{8}y^2 = \frac{\lambda^2 (b-a)^2}{8},$$

where $0 < k < y$. □

1.4.2 Azuma Hoeffding's Inequality

Theorem 1.4.2. *(***Theorem 5.1, [DA09]***) Let $X_0, X_1, \ldots$ be a sequence of random variables that is martingale and have the bounded difference property with parameters $a_i, b_i, i \geqslant 1$.*

Then

$$p(X_n > X_0 + t), P(X_n < X_0 - t) \leq exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Proof. We will use Theorem 1.4.1 to prove this inequality. Consider a random variable K, where

$$K = (Z_n | X_{n-1}),$$

where $Z_n = X_n - X_{n-1}$. Now a rapid calculation shown below shows that $E[K] = 0$.

$$\begin{aligned} E[K] &= E[Z_n | X_{n-1}] \\ &= E[X_n - X_{n-1} | X_{n-1}] \\ &= E[X_n | X_{n-1}] - E[X_{n-1} | X_{n-1}] \\ &= X_{n-1} - X_{n-1} = 0. \end{aligned}$$

By bounded difference property ,

$$a_n \leq K \leq b_n.$$

Hence K satisfies the hypothesis of the Theorem 1.4.1 and we get,

$$E[K] \leq e^{\frac{\lambda^2 (b_n - a_n)^2}{8}}.$$

Which implies that

$$E[e^{\lambda Z_n} | X_{n-1}] \leq e^{\frac{\lambda^2 (b_n - a_n)^2}{8}}.$$

By total law of expectation we get,

$$\begin{aligned} E[e^{\lambda X_n}] &= E[e^{\lambda X_{n-1}} E[e^{\lambda X_n} | X_{n-1}] \\ &\leq E[e^{\lambda X_{n-1}}] e^{\frac{\lambda^2 (b_n - a_n)^2}{8}} \\ &= \prod_{i=1}^{n} e^{\frac{\lambda^2 (b_n - a_n)^2}{8}} \\ &= e^{\frac{\lambda^2 \sum_{i=1}^{n} (b_n - a_n)^2}{8}}. \end{aligned}$$

Since $\frac{e^{\frac{\lambda^2 \sum_{i=1}^{n} (b_n - a_n)^2}{8}}}{e^{\lambda t}}$ gain its minimum value at $\lambda = \frac{4t}{\sum_{i=1}^{n} (b_n - a_n)^2}$,

we get

$$
\begin{aligned}
P(X_n > t) &\leq \min_{\lambda \geq 0} \frac{E[e^{\lambda X_n}]}{e^{\lambda t}} \\
&\leq \min_{\lambda > 0} exp \left(\frac{\lambda^2 \sum_{i=1}^{n} (b_n - a_n)^2}{8} - \lambda t \right) \\
&= exp \left(\frac{-2t^2}{\sum_{i=1}^{n} (b_n - a_n)^2} \right).
\end{aligned}
$$

□

There has been a alot of work in this area and concentration results have been obtained for a large class of functions of independent random variables. We refer intrested reader to [SPG12].

Chapter 2

Ranking through pair-wise comparisons

Suppose we have n items and k pairwise comparisons between the pair of items. Based on those pairwise comparisons, we want to estimate the underlying ranking of the items. The ranking problem is a fairly well-known problem in applied mathematics and computer science. We study the method used in [SSD17] to obtain the ranking from pair-wise comparisons. In [SSD17], the authors start with the assumption that there is an unknown underlying distribution over the ranking of objects and the outcome of pairwise comparison between items is generated as per this underlying distribution. The idea introduced in [SSD17] is to define a Markov chain on the graph of nodes corresponding to each item with edges connecting the nodes for which the pairwise comparisons are available. The stationary distribution of this Markov chain provides a "good" estimate for the ranks of n items. One of the crucial steps in the proof of obtaining an error bound on the estimated rank and the actual rank involves using concentration inequalities. The authors used Hoeffding's inequality to obtain this bound. In this chapter, we use three different concentration inequalities for obtaining a bound on the same object as in the proof of Lemma 4 in [SSD17], and compare the results.

2.1 Model

In this section, we describe a model to estimate ranks using pair-wise comparisons between n objects. Ranking problems, in particular, ranking via pair-wise comparisons have been widely studied.

BTL Model.[SSD17] This model assumes that there is weight $w_i \in \mathbf{R}_+ \equiv \{x \in \mathbf{R} : x > 0\}$ Corresponding to each item $i \in [n]$, while comparing pair of items. The outcome of pairwise comparison is ascertain by these wieghts ie. w_i and w_j decides the outcome of a comparison for pair of items i and j. Let the outcome of the l-th comparison of the pair i

and j, denoted by Y_{ij}^l in a way such that $Y_{ij}^l = 1$ if j is wins over i and 0 otherwise. Then,

$$Y_{ij}^l = \begin{cases} 1 & \text{with probability } \frac{w_j}{w_i+w_j} \\ 0 & \text{otherwise.} \end{cases}$$

In [SSD17], Authors assume that there are fixed k number of pairwise comparisons for all those pairs that are considered. Consider a Markov chain on a weighted directed graph G=([n],E,A), where a pair $(i, j) \in E$ for which the pairwise comparison are available and let d_i be the out degree of the i^{th} node and $d = max_i\{d_i\}$
Let P be the transition matrix corresponding to above defined Markov chain, where

$$P_{ij} = \begin{cases} \frac{1}{d}\frac{1}{k}\sum_{i=1}^{k} {Y_{ij}}^l & \text{if } i \neq j \\ 1 - \frac{1}{d}\sum_{s\neq i}\frac{1}{k}\sum_{i=1}^{k} {Y_{is}}^l & \text{if } i = j \end{cases}$$

for all $(i, j) \in E$ and $P_{ij} = 0$ otherwise.
Let π be the stationary distribution corresponding to P.
$\tilde{P}$ is the ideal matrix (Expectation of P) which is defined as

$$\tilde{P_{ij}} = \begin{cases} \frac{1}{d}\frac{w_j}{w_i+w_j} & \text{if } i \neq j \\ 1 - \frac{1}{d}\sum_{l\neq i}\frac{w_l}{w_i+w_l} & \text{if } i = j \end{cases}$$

for all $(i, j) \in E$ and $\tilde{P_{ij}} = 0$ otherwise.
Let $\tilde{\pi}$ be the stationary distribution corresponding to $\tilde{P}$. $\tilde{\pi}$ gives the actual ranks whereas π gives the estimated ranks.

2.2 Estimating Pair-wise comparison ranks via Markov Chain Convergence

In this section, we discuss the method from [SSD17] and state their main theorem.

Theorem 2.2.1. *(Theorem 2,***[SSD17]***) Given n objects and a connected comparison graph $G = ([n], E)$, let each pair $(i, j) \in E$ be compared for k times with outcomes produced as per a BTL model with parameters $w_1, \ldots, w_n$. Then, for some positive constant $C \geq 8$ and when $k \geq 4C^2\left(\frac{b^2k^5}{d\xi^2}\right)\log n$, the following bound on the normalized error holds with probability at least $1 - 4n^{-C/8}$:*

$$\frac{||\pi - \tilde{\pi}||}{||\tilde{\pi}||} \leq \frac{b^{1/5}\kappa}{\xi}\delta(\triangle) \tag{2.1}$$

where $\delta(\triangle) \leq C\sqrt{\frac{\log n}{kd}}$, $\tilde{\pi}(i) = \frac{w_i}{\sum_l w_l}$, $\quad b = \max_{i,j}\{w_i/w_j\}, \quad$ *and* $\quad \kappa = d/d_{min}$

Definition 2.2.1. *Let* $\lambda_1, \ldots, \lambda_n$ *be the eigenvalues of a matrix* $M \in C_{n\times n}$*. Then its spectral radius* $\delta(M)$ *is defined as:* $\delta(M) = \max\{|\lambda_1|, \ldots, |\lambda_n|\}$.

The crucial step in providing the bound between the stationary distributions is obtaining a bound on spectral radius of Δ. In [SSD17], the bound is obtained by Azuma hoeffding's inequality. In the next section, we use different concentration inequalities to obtain a bound on a $\delta(\triangle)$ and we compare these bounds.

2.3 Comparison of bounds on $\delta(\triangle)$

In this section, we shall compare various inequalities and understand the resulting bounds corresponding to each inequality with respect to the model discussed above.
To bound $\delta(\triangle)$, where $\triangle = P - \widetilde{P}$ so that for $1 \leq i, j \leq n$,

$$\begin{aligned}\triangle_{ij} &= P_{ij} - \widetilde{P}_{ij} \\ &= \frac{1}{kd}\sum_{i=1}^{k}(Y_{ij}^{l} - kp_{ij}) \\ &= \frac{1}{kd}C_{ij}, \end{aligned} \tag{2.2}$$

where $C_{ij} = \sum_{i=1}^{k} {Y_{ij}}^{l} - kp_{ij}$ and $C_{ij} = 0$ for $(i,j) \notin E$ for $1 \leq i \leq n$,

$$\begin{aligned}\triangle_{ii} &= P_{ii} - \widetilde{P}_{ii} \\ &= \left(1 - \sum_{j\neq i} P_{ij}\right) - \left(1 - \sum_{j\neq i}\widetilde{P}_{ij}\right) \\ &= \sum_{j\neq i}(\widetilde{P}_{ij} - P_{ij}) \\ &= -\sum_{j\neq i}\triangle_{ij}. \end{aligned}$$

Let D be the diagonal matrix with $D_{ii} = \triangle_{ii}$ for $1 \leq i \leq n$ and $\overline{\triangle} = \triangle - D$. Then

$$\delta(\triangle) \leq \delta(D + \overline{\triangle}) \leq \delta(D) + \delta(\overline{\triangle}).$$

Bounding $\delta(D)$:

$$\delta(D) = \max_i |D_{ii}| = \max_i |\triangle_{ii}|$$

The aim of Lemma 4 in [SSD17] is to provide a bound on the spectral radius of Δ. This is done by splitting the analysis into different cases: $d \geq \log n$ and $d < \log n$, part of which

involves giving a bound on $\delta(D)$. In the following sections, we provide three different bounds for $\delta(D)$ using three different concentration inequalities in case of $d < \log n$. As mentioned before, the aim of this chapter is to compare these bounds and conclude which inequality works better in this case.

2.3.1 Hoeffding Inequality

Let $X_1, \ldots, X_k$ be random variables such that

$$\begin{aligned} X_1 &= Y_{ij}^1 - p_{ij} \\ &\vdots \\ X_k &= Y_{ij}^k - p_{ij}. \end{aligned}$$

Therefore,

$$\sum_{i=1}^{k} X_i = \sum_{l=1}^{k} Y_{ij}^l - p_{ij} \qquad \text{which implies}$$

$$\sum_{i \neq j} \sum_{i=1}^{k} X_i = \sum_{i \neq j} C_{ij}.$$

Then we have

$$\begin{aligned} P(|\sum_{i \neq j} C_{ij}| > t) &< 2exp\left(\frac{-2t^2}{\sum_{i=1}^{kd}(1^2)}\right) \\ &\leq 2exp\left(\frac{-t^2}{2kd}\right), \end{aligned}$$

which implies

$$P(kd|\triangle_{ii}| \geq t) \leq 2exp\left(\frac{-t^2}{2kd}\right).$$

Put $t = C(kd\log n)^{\frac{1}{2}}$ where C is appropriately large constant,

$$\begin{aligned} P(\delta(D) \geq C\left(\frac{\log n}{kd}\right)^{\frac{1}{2}}) &\leq \sum_{i=1}^{n} P\left(\triangle_{ii} > C\left(\frac{\log n}{kd}\right)^{\frac{1}{2}}\right) \\ &\leq 2nn^{\frac{-C^2}{2kd}} \\ &\leq 2nn^{\frac{-C^2}{2}} \\ &= 2n^{\frac{-C^2}{2}+1}. \end{aligned}$$

2.3.2 Chernoff's Inequality

Let Z be the sum of independent random variable. Then by Chernoff's inequality we have

$$P(Z \geq t) \leq min\left\{ \frac{E[e^{\lambda Z}]}{e^{\lambda t}} \right\}_{t>0} \quad \text{for all} \quad \lambda \geq 0.$$

Let $Z = \sum_{i \neq j} C_{ij}$. Then

$$P\left(\sum_{i \neq j} C_{ij} \geq t \right) \leq \frac{E\left[exp(\lambda \sum_{i \neq j} C_{ij}) \right]}{e^{\lambda t}}, \quad \text{for all} \quad \lambda \geq 0.$$

To compute $E\left[e^{\lambda \sum_{i \neq j} C_{ij}} \right]$:

$$\begin{aligned}
E\left[e^{\lambda \sum_{i \neq j} C_{ij}} \right] &= E\left[e^{\lambda \sum_{i \neq j} \sum (Y_{ij}^l - k p_{ij})} \right] \\
&= e^{-\lambda k p_{ij}} E\left[e^{\lambda (\sum_{i \neq j} \sum_{l=1}^{k} Y_{ij}^l)} \right] \\
&= e^{-\lambda k p_{ij}} \prod_{j \neq i} \sum_{s=1}^{s=k} e^{\lambda s} \binom{k}{s} (p_{ij}^s)(1-p_{ij})^{k-s} \qquad \left(\sum_{l=1}^{k} Y_{ij}^l \approx Bin(k,s) \right) \\
&= e^{-\lambda k p_{ij}} \prod_{j \neq i} \sum \binom{k}{s} e^{\lambda} (p_{ij})^s (1-p_{ij})^{k-s} \\
&\leq e^{-\lambda k p_{ij}} \prod_{j \neq i} (p_{ij} e^{\lambda} + 1 - p_{ij})^k \quad \text{by Binomial Theorem}
\end{aligned}$$

Now we have

$$\begin{aligned}
P\left(\sum_{i \neq j} C_{ij} \geq t \right) &\leq \prod_{j \neq i} \frac{(p_{ij} e^{\lambda} + 1 - p_{ij})^k}{e^{\lambda k p_{ij}} e^{\lambda t}} \\
&\leq \prod_{j \neq i} \frac{(e^{\lambda} + 1)^k}{e^{\lambda t}} \\
&\leq \frac{(e^{\lambda} + 1)^{kd}}{e^{\lambda t}}.
\end{aligned}$$

Let

$$t = \frac{kd \log(e^{\lambda} + 1)}{\lambda} + \frac{\log(n^C)}{\lambda},$$

then,

$$P(|\triangle_{ii}| \geq \frac{t}{d}) = P\left(|\triangle_{ii}| \geq \frac{\log(e^{\lambda}+1)^{kd}}{\lambda kd} + \frac{\log(n^{C})}{\lambda kd} \right) \leq n^{-C} \quad \text{for all } \lambda \geq 0.$$

2.3.3 Azuma Hoeffding's Inequality

For a fixed i , construct a sequence of random variables $X_0, X_1, X_2, \ldots, X_k, \ldots$ as follows:
$X_0 = 0$
$X_1 = Y^1_{i,j_1} - p_{i,j_1}$
$X_2 = (Y^1_{i,j_1} - p_{i,j_1}) + (Y^2_{i,j_1} - p_{i,j_1})$
$X_3 = (Y^1_{i,j_1} - p_{i,j_1}) + (Y^2_{i,j_1} - p_{i,j_1}) + (Y^3_{i,j_1} - p_{i,j_1})$
$\vdots$
$X_k = (Y^1_{i,j_1} - p_{i,j_1}) + (Y^2_{i,j_1} - p_{i,j_1}) + (Y^3_{i,j_1} - p_{i,j_1}) + \cdots + (Y^k_{i,j_1} - p_{i,j_1})$
$X_{k+1} = X_k + (Y^1_{i,j_2} - p_{i,j_2}) + (Y^2_{i,j_2} - p_{i,j_2})$
$X_{2k+1} = X_{2k} + (Y^1_{i,j_3} - p_{i,j_3})$
so on up to X_{dk} and let $X_{dk+n} = X_{dk}$ for all $n \geq 1$.
By Azuma Hoeffding's inequality, it follows that

$$\begin{aligned} P\left(\left| \sum_{i \neq j} C_{i,j} \right| \geq t \right) &\leq P(|X_{dk}| \geq t) \\ &\leq 2exp\left(\frac{-t^2}{2 \sum\limits_{i \neq j} {c_n}^2} \right) \\ &\leq 2exp\left(\frac{-t^2}{2kd} \right), \end{aligned}$$

which implies

$$P(kd|\triangle_{ii}| \geq t) \leq 2exp\left(\frac{-t^2}{2kd} \right).$$

Put $t = C(kd\log n)^{\frac{1}{2}}$ where C is appropriately large constant,

$$\begin{aligned} P\left(\delta(D) \geq C\left(\frac{\log n}{kd} \right)^{\frac{1}{2}} \right) &\leq \sum_{i=1}^{n} P\left(\triangle_{ii} > C\left(\frac{\log n}{kd} \right)^{\frac{1}{2}} \right) \\ &\leq 2nn^{\frac{-C^2}{2kd}} \\ &\leq 2nn^{\frac{-C^2}{2}} \\ &= 2n^{\frac{-C^2}{2}+1}. \end{aligned}$$

From above results we conclude that Azuma Hoeffding's gives better results than Chernoff bound and Hoeffding's gives the same bound as Azuma Hoeffding's.

Chapter 3

Effects of Perturbation

In this chapter, we look at the problem of ranking n objects from an adversarial point of view. As an adversary, we want to introduce perturbations in the transition matrix for the Markov chain whose stationary distribution provides the estimate for the rank. For convenience, we assume throughout that the perturbations are made such that the resulting matrix remains stochastic and therefore, we get a new (perturbed) stationary distribution that provides a new estimate for the rank. The larger idea is to obtain strategies in terms of perturbation matrices that produce specific kind of perturbations. For example, if one wants to only flip the ranks of the l^{th} and k^{th} nodes, what kind of perturbation matrix is required? Unfortunately, as of now, we are unable to provide any general answers to these questions. In the next chapter, we illustrate some examples of perturbation matrices that lead to such specific perturbations in the rank. In this chapter, following the results and tools in [SSD17], we obtain a new lower bound for number of pairwise comparisons k in terms of the perturbations introduced in the transition matrix P. Subsequently, for $d < \log n$, we also obtained a new version of Theorem 1 in [SSD17].

3.1 Setup

Define perturbation matrix S,

$$S_{ij} = \varepsilon_{ij} \quad \text{for} \quad 1 \leq i, j \leq n,$$

where $\varepsilon_{ij} = -\varepsilon_{ij}$ and for a fixed i, $\sum_{j=1}^{n} \varepsilon_{ij} = 0$

In this chapter, we study the consequences of perturbation of P matrix, say P^* where, P^* is

defined as $P^* = P + S$ ie.

$$P^*_{ij} = \begin{cases} \frac{1}{d}\frac{1}{k}\sum_{i=1}^{k} {Y_{ij}}^l + \varepsilon_{ij} & \text{if } i \neq j \\ 1 - \sum_{s \neq i}\left(\frac{1}{d}\frac{1}{k}\sum_{i=1}^{k} {Y_{is}}^l + \varepsilon_{is}\right) & \text{if } i = j \end{cases}$$

for all $(i, j) \in E$ and $P^*_{ij} = 0$ otherwise. Note that because of the conditions on S, P^* is still a stochastic matrix. We denote the corresponding stationary distribution by π^*. In the next section, we will determine a lower bound (in terms of ε_{ij}) on k to obtain "good" estimated rank despite the perturbation.

3.2 Main results

Lemma 3.2.1. *For some constant $C \geq 2$, when $d < \log n$, the error matrix $\triangle^* = P^* - \widetilde{P}$ satisfies $\delta(\triangle^*) \leq 2C\sqrt{\frac{\log n}{kd}} + 2d|\varepsilon_{ij}|$ with probability at least $1 - 4n^{-\frac{C^2}{2}+1}$.*

Proof. Our interest is in bounding $\delta(\triangle^*)$. Now $\triangle^* = P^* - \widetilde{P}$ so that for $1 \leq i, j \leq n$, we have

$$\begin{aligned} \triangle^*_{ij} &= P^*_{ij} - \widetilde{P}_{ij} \\ &= \frac{1}{d}\sum_{i=1}^{k}({Y_{ij}}^l + d\varepsilon_{ij} - p_{ij}) \\ &= \frac{1}{kd}\left(\sum_{i=1}^{k} {Y_{ij}}^l + kd\varepsilon_{ij} - kp_{ij}\right) \\ &= \frac{1}{kd}C^*_{ij}, \end{aligned} \tag{3.1}$$

where $C^*_{ij} = \sum_{i=1}^{k} {Y_{ij}}^l + kd\varepsilon_{ij} - kp_{ij}$ and $C^*_{ij} = 0$ for $(i, j) \notin E$ and let $C_{ij} = \sum_{i=1}^{k} {Y_{ij}}^l - kp_{ij}$. For $1 \leq i \leq n$,

$$\begin{aligned} \triangle^*_{ii} &= P^*_{ii} - \widetilde{P}_{ii} \\ &= (1 - \sum_{j \neq i} P^*_{ij}) - (1 - \sum_{j \neq i} \widetilde{P}_{ij}) \\ &= \sum_{j \neq i}(\widetilde{P}_{ij} - P^*_{ij}) \\ &= -\sum_{j \neq i} \triangle^*_{ij}. \end{aligned}$$

Let D^* be the diagonal matrix with $D^*_{ii} = \triangle^*_{ii}$ for $1 \leq i \leq n$ and $\overline{\triangle}^* = \triangle^* - D^*$ Then

$$\delta(\triangle^*) \leq \delta(D^* + \overline{\triangle}^*) \leq \delta(D^*) + \delta(\overline{\triangle}^*).$$

Bounding $\delta(D^*)$:

$$\delta(D^*) = \max_i |D^*_{ii}| = \max_i |\triangle^*_{ii}|.$$

For a fixed i, construct a sequence of random variables $X_0, X_1, X_2, \ldots, X_k, \ldots$ as follows:

$$\begin{aligned}
&X_0 = 0 \\
&X_1 = Y^1_{i,j_1} - p_{i,j_1} \\
&X_2 = (Y^1_{i,j_1} - p_{i,j_1}) + (Y^2_{i,j_1} - p_{i,j_1}) \\
&X_3 = (Y^1_{i,j_1} - p_{i,j_1}) + (Y^2_{i,j_1} - p_{i,j_1}) + (Y^3_{i,j_1} - p_{i,j_1}) \\
&\vdots \\
&X_k = (Y^1_{i,j_1} - p_{i,j_1}) + (Y^2_{i,j_1} - p_{i,j_1}) + (Y^3_{i,j_1} - p_{i,j_1}) + \cdots + (Y^k_{i,j_1} - p_{i,j_1}) \\
&X_{k+1} = (Y^1_{i,j_1} - p_{i,j_1}) + (Y^2_{i,j_1} - p_{i,j_1}) + (Y^3_{i,j_1} - p_{i,j_1}) + \cdots + (Y^k_{i,j_1} - p_{i,j_1}) + (Y^1_{i,j_2} - p_{i,j_2}) \\
&X_{k+1} = X_k + (Y^1_{i,j_2} - p_{i,j_2}) + (Y^2_{i,j_2} - p_{i,j_2}) \\
&X_{2k+1} = X_{2k} + (Y^1_{i,j_3} - p_{i,j_3})
\end{aligned}$$

so on up to X_{dk} and let $X_{dk+n} = X_{dk}$ for all $n \geq 1$

Lemma 3.2.2. *The sequence of random variable defined above is a Martingale with bounded difference property.*

Proof. $E[X_n] = 0$ as $Y^l_{i,j}$ is a Bernoulli random variable and $E[Y^l_{i,j}] = p_{ij}$

$$\begin{aligned}
E[X_n|X_{n-1}] &= E[(Y^l_{i,j} - p_{ij}) + X_{n-1}|X_{n-1}] \quad \text{for some j and l} \\
&= E[(Y^l_{i,j} - p_{ij})|X_{n-1}] + E[X_{n-1}|X_{n-1}] \\
&= E[(Y^l_{i,j} - p_{ij})] + E[X_{n-1}|X_{n-1}] \\
&= X_{n-1}
\end{aligned}$$

Moreover, $X_n - X_{n-1} = Y^l_{i,j} - p_{ij}$, for some j and l.
Therfore, $|X_n - X_{n-1}| \leq c_n = 1$ for all n.
Also, $X_{dk} = \sum_{i \neq j,l} (Y^l_{i,j} - p_{ij}) = \sum_{i \neq j} C_{i,j}$. □

Therefore by an application of Azuma Hoeffding's inequality, it follows that:

$$\begin{aligned}
P(kd|\triangle^*_{ii}| \geq t) &\leq P\left(\left|\sum_{i \neq j} C^*_{ij}\right| \geq t\right) \\
&\leq P\left(\left|\sum_{i \neq j} C_{ij}\right| + \left|\sum_{i \neq j} kd\varepsilon_{ij}\right| \geq t\right).
\end{aligned}$$

Moreover,

$$P\left(\left|\sum_{i\neq j} C_{ij}\right| + \left|\sum_{i\neq j} kd\varepsilon_{ij}\right| \geq t\right) \leq P\left(\left|\sum_{i\neq j} C_{ij}\right| + kd^2|\varepsilon_{ij}| \geq t\right)$$
$$\leq P(|X_{dk}| \geq t - kd^2|\varepsilon_{ij}|)$$
$$\leq 2exp\left(\frac{-(t-kd^2|\varepsilon_{ij}|)^2}{2\sum\limits_{i\neq j} {c_n}^2}\right)$$
$$\leq 2exp\left(\frac{-(t-kd^2|\varepsilon_{ij}|)^2}{2kd}\right),$$

which implies

$$P(kd|\triangle^*_{ii}| \geq t) \leq 2exp\left(\frac{-(t-kd^2|\varepsilon_{ij}|)^2}{2kd}\right).$$

Put t=$C(kd\log n)^{\frac{1}{2}} + kd^2|\varepsilon_{ij}|$ where C is appropriately large constant,

$$P\left(\delta(D^*) \geq C\left(\frac{\log n}{kd}\right)^{\frac{1}{2}} + d|\varepsilon_{ij}|\right)$$
$$\leq \sum_{i=1}^{n} P\left(|\triangle^*_{ii}| \geq C\left(\frac{\log n}{kd}\right)^{\frac{1}{2}} + d|\varepsilon_{ij}|\right)$$
$$\leq 2nexp\left(\frac{-(C(kd\log n)^{\frac{1}{2}} + kd^2|\varepsilon_{ij}| - kd^2|\varepsilon_{ij}|)^2}{2kd}\right)$$
$$= 2nn^{\frac{-C^2}{2}}$$
$$= 2n^{\frac{-C^2}{2}+1}.$$

To Bound $\delta(\overline{\triangle}^*)$:
As we know $\|M\|_2 \leq \sqrt{\|M\|_1\|M\|_\infty}$ for any square matrix M, where $\|M\|_1 = max_i \sum\limits_{ij} |M_{ij}|$ and $\|M\|_\infty = \left\|M^T\right\|_1$. Clearly, It is suffices to get a bound for maximal row-sum of absolute values of $\overline{\triangle}^*$. We know

$$\overline{\triangle}^* = \triangle^* - D^*.$$

Let R_i be the sum of the absolute of the ith row-sum of $\overline{\triangle}^*$.

$$\text{then}, R_i = \frac{1}{kd}\sum_{j\neq i} |C^*_{i,j}|$$

For a fixed i and $\xi_{j_i} \in \{-1,1\}$, construct a sequence of random variables $X_0, X_1, X_2, \ldots, X_k, \ldots$

as follows:

$X_0 = 0$

$X_1 = \xi_{j_1}(Y^1_{i,j_1} - p_{i,j_1})$

$X_2 = \xi_{j_1}(Y^1_{i,j_1} - p_{i,j_1}) + \xi_{j_1}(Y^2_{i,j_1} - p_{i,j_1})$

$X_3 = \xi_{j_1}(Y^1_{i,j_1} - p_{i,j_1}) + \xi_{j_1}(Y^2_{i,j_1} - p_{i,j_1}) + \xi_{j_1}(Y^3_{i,j_1} - p_{i,j_1})$

$\vdots$

$X_k = \xi_{j_1}(Y^1_{i,j_1} - p_{i,j_1}) + \xi_{j_1}(Y^2_{i,j_1} - p_{i,j_1}) + \xi_{j_1}(Y^3_{i,j_1} - p_{i,j_1}) + \cdots + \xi_{j_1}(Y^k_{i,j_1} - p_{i,j_1})$

$X_{k+1} = X_k + \xi_{j_2}(Y^1_{i,j_2} - p_{i,j_2}) + \xi_{j_2}(Y^2_{i,j_2} - p_{i,j_2})$

$\vdots$

$X_{2k+1} = X_{2k} + \xi_{j_3}(Y^1_{i,j_3} - p_{i,j_3})$

$\vdots$

so on up to X_{dk} and let $X_{dk+n} = X_{dk}$ for all $n \geq 1$

Lemma 3.2.3. *The sequence of random variable defined above is a Martingale with bounded difference property.*

Proof. $E[X_n] = 0$ as $Y^l_{i,j}$ is a Bernoulli random variable and $E[Y^l_{i,j}] = p_{ij}$.

$$\begin{aligned} E[X_n|X_{n-1}] &= E[\xi_j(Y^l_{i,j} - p_{ij}) + X_{n-1}|X_{n-1}] \quad \text{for some j and l.} \\ &= E[\xi_j(Y^l_{i,j} - p_{ij})|X_{n-1}] + E[X_{n-1}|X_{n-1}] \\ &= E[\xi_j(Y^l_{i,j} - p_{ij})] + E[X_{n-1}|X_{n-1}] \\ &= X_{n-1}. \end{aligned}$$

Hence the above sequence is a martingale.
Moreover, $X_n - X_{n-1} = \xi_j(Y^l_{i,j} - p_{ij})$ for some j and l. Therefore, $|X_n - X_{n-1}| \leq c_n = 1$ for all n. Also

$$X_{dk} \geq \sum_{i \neq j,l} \xi_j(Y^l_{i,j} - p_{ij}) = \sum_{i \neq j} \xi_j C_{i,j}.$$

□

Therefore by an application of Azuma Hoeffding's inequality, it follows that :

$$\begin{aligned} P(R_i > s) &= P\left(\sum_{j \neq i} |C^*_{ij}| > kds\right) \\ &\leq P\left(\sum_{j \neq i} |C_{ij}| + kd|\varepsilon_{ij}| > kds\right) \\ &\leq P\left(\sum_{j \neq i} |C_{ij}| + dk|\varepsilon_{ij}| > kds\right) \end{aligned}$$

Also,

$$P\left(\sum_{j\neq i}|C_{ij}|+dk|\varepsilon_{ij}|>kds\right)=P\left(\sum_{j\neq i}|C_{ij}|>kd(s-d|\varepsilon_{ij}|)\right)$$
$$\leq\sum_{j\in\delta_i}\sum_{\xi\in\{-1,1\}}P\left(\sum_j\xi_jC_{ij}>kd(s-d|\varepsilon_{ij}|)\right)\quad\text{by union bound,}$$
$$\leq\sum_{j\in\delta_i}\sum_{\xi\in\{-1,1\}}exp\left(\frac{-2k^2d(s-d|\varepsilon_{ij}|^2)}{2kd}\right)$$

Now, as the number of terms in the above summation is 2^{d_i}, and also, $d_i\leq d$. Thus we get

$$\sum_{j\in\delta_i}\sum_{\xi\in\{-1,1\}}exp\left(\frac{-k^2d(s-d|\varepsilon_{ij}|^2)}{kd}\right)$$
$$\leq exp(-kd(s-d|\varepsilon_{ij}|)^2+d\log 2).$$

By union bound, we get

$$P(\delta(\overline{\triangle})\geq s)\leq 2nP(R_i\geq s)$$
$$\leq 2nexp(-kd(s-d|\varepsilon_{ij}|)^2+d\log 2).$$

If we put

$$s=d|\varepsilon_{ij}|+\sqrt{\frac{d\log 2+C\log n}{kd}},$$

then,

$$P\left(\delta(\overline{\triangle}^*)\geq d|\varepsilon_{ij}|+\sqrt{\frac{C\log n+d\log 2}{kd}}\right)\leq 2n^{-(C^2/2-1)}.$$

As we assumed $d<\log n$, so we get

$$\delta(\overline{\triangle}^*)\leq C\sqrt{\frac{\log n}{kd}}+d|\varepsilon_{ij}|,$$

with probability atleast $1-2n^{-\frac{C^2}{2}+1}$. Also

$$\delta(\triangle^*)\leq\delta(D^*+\overline{\triangle}^*)\leq\delta(D^*)+\delta(\overline{\triangle}^*),$$

implies that

$$
\begin{aligned}
P\left(\frac{\delta(\triangle^*)}{2} \geq C\sqrt{\frac{\log n}{kd}} + d|\varepsilon_{ij}|\right) &\leq P\left(\frac{\delta(\overline{\triangle}^*) + \delta(D^*)}{2} \geq C\sqrt{\frac{\log n}{kd}} + d|\varepsilon_{ij}|\right) \\
&=\leq P\left(\delta(\overline{\triangle}^*) + \delta(D^*) \geq 2C\sqrt{\frac{\log n}{kd}} + 2d|\varepsilon_{ij}|\right) \\
&\leq P\left(\left\{\delta(\overline{\triangle}^*) \geq C\sqrt{\frac{\log n}{kd}} + d|\varepsilon_{ij}|\right\}\right. \\
&\qquad \left.\bigcup\left\{\delta(D^*) \geq C\sqrt{\frac{\log n}{kd}} + d|\varepsilon_{ij}|\right\}\right) \\
&\leq P\left(\delta(\overline{\triangle}^*) \geq C\sqrt{\frac{\log n}{kd}} + d|\varepsilon_{ij}|\right) \\
&\qquad + P\left(\delta(D^*) \geq C\sqrt{\frac{\log n}{kd}} + d|\varepsilon_{ij}|\right) \\
&\leq 2n^{-\frac{C^2}{2}+1} + 2n^{-\frac{C^2}{2}+1} \\
&= 4n^{-\frac{C^2}{2}+1}.
\end{aligned}
$$

Therefore we get

$$
P\left(\delta(\triangle^*) \leq 2C\sqrt{\frac{\log n}{kd}} + 2d|\varepsilon_{ij}|\right) \geq 1 - 4n^{-\frac{C^2}{2}+1}.
$$

□

We recall the notation from [SSD17]:
$\rho = \lambda_{max}(\tilde{P}) + \delta(\triangle^*)\sqrt{\frac{\tilde{\pi}_{max}}{\tilde{\pi}_{min}}}$ and $\xi \equiv 1 - \lambda_{max}(Q)$ where $Q_{ij} = \frac{1}{d_i}$ for $(i,j) \in E$.

Theorem 3.2.4. *Given n objects and a connected comparison graph $G = ([n], E)$, let each pair $(i,j) \in E$ be compared for k times with outcomes produced as per a BTL model with parameters $w_1, \ldots, w_n$. Then, for some positive constant $C \geq 2$, when $d < \log n$ and $k \geq \frac{4C^2 b^5 d \log n}{(\xi d_{min} + |\varepsilon_{ij}| d(4b^2 d\sqrt{b}))^2}$, the following bound on the normalized error holds with probability at least $1 - 4n^{-\frac{C^2}{2}+1}$:*

$$
\frac{||\pi^* - \tilde{\pi}||}{||\tilde{\pi}||} \leq \frac{4b^{1/5}\kappa}{\xi}\left(C\sqrt{\frac{\log n}{kd}} + |\varepsilon_{ij}|d\right)
$$

where $\tilde{\pi}(i) = \frac{w_i}{\sum_l w_l}$, $b = max_{i,j} w_i/w_j$, and $\kappa = d/d_{min}$

Proof. By Lemma 3.2.1, we have that for some $C \geq 2$ and $d < \log n$,

$$\begin{aligned} 1-\rho &= 1-\lambda_{max}(\tilde{P}) - \delta(\triangle^*)\sqrt{b} \\ &\geqslant 1-\lambda_{max}(\tilde{P}) - \left(C\sqrt{\frac{\log n}{kd}} + |\varepsilon_{ij}|d\right)2\sqrt{b}, \end{aligned}$$

with probability atleast $1-4n^{-\frac{C^2}{2}+1}$. Lemma 6 in [SSD17] says

$$1-\lambda_{max}(\tilde{P}) \geq \frac{\xi d_{min}}{b^2 d}.$$

Also for $k \geq \frac{4C^2b^5 d\log n}{(\xi d_{min}+|\varepsilon_{ij}|d(4b^2d\sqrt{b}))^2}$, we have

$$2C\sqrt{\frac{b\log n}{kd}} + 2|\varepsilon_{ij}|d\sqrt{b} \leq \frac{\xi d_{min}}{2b^2 d}$$

which implies that

$$\begin{aligned} 1-\rho &\geq 1-\lambda_{max}(\tilde{P}) - 2C\sqrt{\frac{b\log n}{kd}} - 2|\varepsilon_{ij}|d\sqrt{b} \\ &\geq \frac{\xi d_{min}}{b^2 d} - \frac{\xi d_{min}}{2b^2 d} \\ &= \frac{\xi d_{min}}{2b^2 d} \end{aligned}$$

By Lemma 2 [SSD17], we get

$$\begin{aligned} \frac{\|\pi^* - \tilde{\pi}\|}{\|\tilde{\pi}\|} &\leq \frac{1}{1-\rho}\delta(\triangle)\frac{\tilde{\pi}_{max}}{\tilde{\pi}_{min}} \\ &\leq \frac{2b^2 d}{\xi d_{min}}\left(2C\sqrt{\frac{\log n}{kd}} + 2|\varepsilon_{ij}|d\right)b^{1/2} \\ &= \frac{4b^{\frac{5}{2}}d}{\xi d_{min}}\left(C\sqrt{\frac{\log n}{kd}} + |\varepsilon_{ij}|d\right). \end{aligned}$$

with probability atleast $1-4n^{-\frac{C^2}{2}+1}$. □

Lemma 3.2.5. *For some constant $C \geq \sqrt{\log nkd}+1$, when $d \geq \log n$ the matrix $\overline{\triangle}^* = \triangle^* - D^*$ satisfies $\delta(\overline{\triangle}^*) \leq C\sqrt{\frac{logn}{kd}} + d|\varepsilon_{ij}|$ with probability at least $1-2n^{\frac{-C}{\sqrt{\log nkd}+1}}$.*

Proof. We will use Corollary 3.7 in [J11] to prove concentration results on $\overline{\triangle}^* = \triangle^* - D^* = \sum_{i<j} Z^{ij*}$ where

$$Z^{ij*} = (e_i e_j^T - e_j e_i^T)(P^*_{ij} - \tilde{P})$$

for $(i,j) \in E$, and $Z^{ij*} = 0$ if i and j are not connected. Above defined Z^{ij*} are independent,

but they are not self adjoint. First we symmetrize it to get self adjoint random matrix.

$$\tilde{Z^{ij*}} = \begin{bmatrix} 0 & Z^{ij*} \\ Z^{ij*} & 0 \end{bmatrix}$$

As hypothesis of the Corollary 3.7 [J11] is satisfied we can apply it to these self-adjoint and independent random matrices.

Let $A^{ij} = \begin{bmatrix} 0 & e_i e_j^T - e_j e_i^T \\ e_j e_i^T - e_i e_j^T & 0 \end{bmatrix}$
if $(i,j) \in E$ and zero otherwise. Then $\tilde{Z}^{ij*} = \triangle^* A^{ij}$. In the following, we showed that

$$Ee^{\theta \tilde{Z}^{ij*}} \leqslant e^{(\theta|\varepsilon_{ij}|)A^{ij2}}$$

for $0.5 < |\theta| < 1$.

$$\begin{aligned} Ee^{(\theta \tilde{Z}^{ij*})} &= I + \theta E(\tilde{Z}^{ij*}) + \sum_{p=2}^{\infty} \frac{\theta^p E[(\tilde{Z}^{ij*})^p]}{p!} \\ &\leqslant I + \theta|\varepsilon_{ij}|A^{ij} + \sum_{p=2}^{\infty} \frac{\theta^p E[(\triangle_{ij}^* A^{ij})^p]}{p!} \\ &\leqslant I + \theta|\varepsilon_{ij}|A^{ij2} + \sum_{p=2}^{\infty} \frac{\theta^p E[|\triangle_{ij}^*|^p](A^{ij})^2}{p!} \end{aligned}$$

To Solve $E(|\triangle_{ij}^*|^p)$:

$$\begin{aligned} E[|\triangle_{ij}^*|^p] &= \int_0^{\infty} P(|\triangle_{ij}^*|^p > x)dx \\ &= \int_0^{\infty} P\left(\frac{|C_{ij}^*|^p}{(kd)^p} > x\right)dx \\ &= \int_0^{\infty} px^{p-1} P\left(\frac{|C_{ij}^*|}{kd} > x\right)dx \\ &\leqslant \int_0^{\infty} px^{p-1} 2exp\left(\frac{-(x-|\varepsilon_{ij}|)^2 kd^2}{2}\right)dx \end{aligned}$$

Let $x = t + |\varepsilon_{ij}|$. Then we have,

$$\begin{aligned}
E(|\triangle_{ij}^{*}|^{p}) &\leqslant 2p\int_{-c}^{\infty}(t+|\varepsilon_{ij}|)^{p-1}exp\left(-\frac{t^2kd^2}{2}\right)dt \\
&\leqslant 2p\int_{-|\varepsilon_{ij}|}^{\infty}\left(\binom{p-1}{0}t^{p-1}+\binom{p-1}{1}t^{p-2}|\varepsilon_{ij}|+\cdots\right. \\
&\quad \left.+\binom{p-1}{0}|\varepsilon_{ij}|^{p-1}\right)e^{\left(-\frac{t^2kd^2}{2}\right)}dt + 2p\int_{1}^{\infty}\left(\binom{p-1}{0}t^{p-1}\right. \\
&\quad \left.+\binom{p-1}{1}t^{p-2}|\varepsilon_{ij}|+\cdots+\binom{p-1}{0}|\varepsilon_{ij}|^{p-1}\right)e^{\left(-\frac{t^2kd^2}{2}\right)}dt \\
&\leqslant 2p\int_{-|\varepsilon_{ij}|}^{1}2^{p-1}exp\left(\frac{t^2kd^2}{2}\right)dt + 2p\int_{1}^{\infty}2^{p-1}t^{p-1}exp\left(\frac{t^2kd^2}{2}\right)dt.
\end{aligned}$$

First we will solve first part of the above integral.

$$\begin{aligned}
2p\int_{-|\varepsilon_{ij}|}^{1}2^{p-1}exp\left(\frac{t^2kd^2}{2}\right)dt &\leq 2^p p\frac{3}{\sqrt{\pi}}\frac{\sqrt{\pi}+1}{2\sqrt{kd/2}} \\
&= 2^p p\frac{3}{\sqrt{\pi}}\frac{\sqrt{\pi}+1}{\sqrt{2kd}}.
\end{aligned}$$

Now we will solve second part of the above integral.
Let $u = \frac{t^2(kd^2)}{2}$. Then,

$$\begin{aligned}
2p\int_{1}^{\infty}2^{p-1}t^{p-1}exp\left(\frac{t^2kd^2}{2}\right)dt &= 2^p p\int_{1}^{\infty}e^{-u}\left(\sqrt{\frac{2u}{kd^2}}\right)^{p-1}\frac{1}{\sqrt{2ukd}}du \\
&= 2^p p\int_{1}^{\infty}e^{-u}\frac{(\sqrt{2u})^{p-2}}{(\sqrt{k}d)^p}du \\
&\leq \frac{2^p p(\sqrt{2})^{p-2}}{(\sqrt{k}d)^p}\int_{1}^{\infty}e^{-u}(\sqrt{u})^{p-2} \\
&\leq \frac{2^p p(\sqrt{2})^{p-2}}{(\sqrt{k}d)^p}\left(\frac{p}{2}\right)! \\
&\leq \frac{2^p p(\sqrt{2})^{p-2}}{(\sqrt{k}d)^p}\left(\frac{p!}{2}\right) \\
&\leq \frac{2^{p+p/2-2}}{(\sqrt{k}d)^p}p(p!).
\end{aligned}$$

Therefore we get,

$$\begin{aligned}
E[|\triangle_{ij}^{*}|^{p}] &\leq 2^{p} p \frac{3}{\sqrt{\pi}} \frac{\sqrt{\pi}+1}{\sqrt{2kd}} + \frac{2^{p+p/2-2}}{(\sqrt{kd})^{p}} p(p!) \\
&\leq 2^{p} p \frac{5}{\sqrt{2kd}} + \frac{2^{p+p/2}}{(\sqrt{kd})^{p}} p(p!) \\
&\leq 2^{p+p/2} p(p!) \left(\frac{5}{\sqrt{2kd}} + \frac{1}{(\sqrt{kd})^{p}} \right) \\
&\leq \frac{2^{p+p/2} 6p(p!)}{\sqrt{kd}} \\
&\leq \frac{2^{2p} 6p(p!)}{\sqrt{kd}}.
\end{aligned}$$

Now we have

$$\begin{aligned}
Ee^{(\theta\tilde{Z}^{ij*})} &\leqslant I + \theta|\varepsilon_{ij}|(A^{ij})^{2} + \left(\sum_{p=2}^{\infty} \frac{\theta^{p} 2^{2p} 6p(p!)}{(p!)\sqrt{kd}} \right) (A^{ij})^{2} \\
&= I + \theta|\varepsilon_{ij}|(A^{ij})^{2} + \left(\sum_{p=2}^{\infty} \frac{6p(4\theta)^{p}}{\sqrt{kd}} \right) (A^{ij})^{2} \\
&= I + \theta|\varepsilon_{ij}|(A^{ij})^{2} + \frac{6(A^{ij})^{2}}{\sqrt{kd}} \left(-4\theta + \frac{4\theta}{(1-4\theta)^{2}} \right) \\
&= I + \left(\theta|\varepsilon_{ij}| - \frac{24\theta}{\sqrt{kd}} + \frac{24\theta}{(1-4\theta)^{2}\sqrt{kd}} \right) (A^{ij})^{2}.
\end{aligned}$$

Let

$$\theta|\varepsilon_{ij}| - \frac{24\theta}{\sqrt{kd}} + \frac{24\theta}{(1-4\theta)^{2}\sqrt{kd}} = g(\theta).$$

A quick calculation shows that $g(\theta) \leqslant \theta|\varepsilon_{ij}|$ for $0.5 < |\theta| < 1$. we get

$$Ee^{\theta\tilde{Z}^{ij*}} \leqslant e^{(\theta|\varepsilon_{ij}|)(A^{ij})^{2}}, \text{for } 0.5 < |\theta| < 1.$$

for $0.5 < |\theta| < 1$.

Also

$$\begin{aligned}\sum_{i<j}(A^{ij})^2 &= \sum_{i<j}\mathbb{1}_{(i,j)\in E}\begin{bmatrix} e_ie_i^T+e_je_j^T & 0 \\ 0 & e_ie_i^T+e_je_j^T \end{bmatrix} \\ &\leqslant \sum_{i=1}^{n} d_i \begin{bmatrix} e_ie_i^T & 0 \\ 0 & e_ie_i^T \end{bmatrix} \\ &\leqslant d\sum_{i=1}^{n}\begin{bmatrix} e_ie_i^T & 0 \\ 0 & e_ie_i^T \end{bmatrix}\end{aligned}$$

Therefore, $\delta\left(\sum_{i<j}(A^{ij})^2\right) \leqslant d.$

By Lemma 3.2.5 we get,

$$P(\delta(\tilde{Z}^{ij*}) \geqslant t) \leqslant 2n.exp(-\theta t + \theta|\varepsilon_{ij}|d).$$

Let $\theta = \frac{1}{1+\frac{1}{\sqrt{\log nkd}}}$. we get

$$\begin{aligned}P(\delta(\tilde{Z}^{ij*}) \geqslant t) &\leqslant 2n.exp\left(-\theta t + \theta|\varepsilon_{ij}|d\right) \\ &\leqslant 2n.exp\left(-\frac{1}{1+\frac{1}{\sqrt{\log nkd}}}(t-|\varepsilon_{ij}|d)\right).\end{aligned}$$

Let

$$t = C\sqrt{\frac{\log n}{kd}} + |\varepsilon_{ij}|d.$$

Then

$$\begin{aligned}P\left(\delta(\tilde{Z}^{ij*}) \geqslant C\sqrt{\frac{\log n}{kd}} + |\varepsilon_{ij}|d\right) &\leqslant 2n.exp\left(-\frac{1}{1+\frac{1}{\sqrt{\log nkd}}}\left(C\sqrt{\frac{\log n}{kd}}\right)\right) \\ &= 2n.exp\left(-\frac{\sqrt{\log nkd}}{\sqrt{\log nkd}+1}\left(C\sqrt{\frac{\log n}{kd}}\right)\right) \\ &= 2n.exp\left(-\frac{C}{\sqrt{\log nkd}+1}\log n\right) \\ &= 2nn^{\frac{C}{\sqrt{\log nkd}+1}} \\ &= 2n^{\frac{C}{\sqrt{\log nkd}+1}+1}.\end{aligned}$$

which implies

$$P\left(\delta(\overline{\triangle}^*) \leqslant C\sqrt{\frac{\log n}{kd}} + |\varepsilon_{ij}|d\right) \geqslant 1 - 2n^{\frac{C}{\sqrt{\log nkd}+1}+1}$$

□

It turns out that the bound on $\delta(\overline{\triangle}^*)$ obtained for $d \geq \log n$ above is not very useful when it comes to obtaining a "good" error bound for $||\pi^* - \pi||$. We hope to improve this bound by using alternate techniques in future.

Chapter 4

Effects of Perturbation: Simulations

In this chapter, we consider some specific example of perturabation matrix S and try to understand how the estimated rank depends on the perturbation. We do not have explicit theoretical results supporting our observation.
We define the following notations:
Let r^* denote the stationary distribution of P^* (To recall P^*, see chapter 3)
We use the following norms to ascertain the error between the actual ranks (denoted by r) and the estimated rank (denoted by r^*).

- $p-$norm: $\|r-r^*\|_p = \left(\sum_i \left|r(i)-r^*(i)\right|^p\right)^{\frac{1}{p}}$,
- *sup*-norm: $\|r-r^*\|_m = \max_i\{|r(i)-r^*(i)|\}$
- The following norm is used in [SSD17]:
  $D'_r(r^*) = \sqrt{\frac{1}{2n||r||^2}\sum_{i<j}(r_i-r_j)^2 \mathbb{1}_{||(r_i-r_j)|-|(r_i^*-r_j^*)||>\delta}}$.

  All the examples in this chapter have the following parameter fixed.

- Total number of objects, $n = 8$
- Actual rank r=[2 3 4 1 5 7 6 8]
- For D'_r norm, δ is fixed to be 0.5
- We assume that we have a complete graph and d=7.

Given below is the algorithm to find the stationary distribution of the perturbed matrix.

Algorithm 1: Stationary distribution of perturbed matrix P^*

Inputs: r = [2,3,4,1,5,7,6,8], d = 7, $0 \leq i, j \leq 8$, k , S;

generate $Y_{ij}^l = \begin{cases} 1 & \text{with probability } \frac{r(j)}{r(i)+r(j)} \\ 0 & \text{with probability } 1-\frac{r(j)}{r(i)+r(j)} \end{cases}$;

For $i < j$, $P_{ij} = \frac{1}{kd}\sum_{i=1}^{k} Y_{ij}^l$ and $P_{ji} = \frac{1}{d} - P_{ij}$;

For $i = j$, $P_{ii} = 1 - \sum_{i \neq j} P_{ij}$;

Compute perturbed probability matrix Q:

$Q = P + S$

Compute stationary distribution, r^*

r^* = Left eigen vector of Q with respect to eigen value 1;

Compute p-norm: $\|r - r^*\|_p = (\sum_i |r(i) - r^*(i)|^p)^{\frac{1}{p}}$;

Compute max-norm: $\|r - r^*\|_m = \max_i\{|r(i) - r^*(i)|\}$;

Compute D_r-norm:

$D_r(r^*) = \sqrt{\frac{1}{2n\|r\|^2}\sum_{i<j}(r(i)-r(j))^2 \mathbb{1}_{\|(r(i)-r(j))|-|(r^*(i)-r^*(j))\|>0.5}}$;

Output: r*, $\|r - r^*\|_1$, $\|r - r^*\|_2$, $\|r - r^*\|_m$, $\|r - r^*\|_{D_r}$;

Now, we illustrate the behaviour of r^* and the error in various cases.

I. As expected, error increases with increased perturbation and decreases with increasing k. This is illustrated in figure 4.1 and figure 4.2 . The obtained rank r^* and corresponding errors are listed/tabulated below:

Case 1: Fix k=1000 and S such that:

$$S_{ij} = \begin{cases} \varepsilon & \textit{for i=4, j=8} \\ -\varepsilon & \textit{for i=8, j=4} \\ 0 & \textit{otherwise} \end{cases}$$

We have:

ε	r^*
0.003	[1.8824 2.9507 3.8607 1.0232 4.9613 7.2273 6.0074 7.9337]
0.002	[1.8800 2.9477 3.8570 1.0116 4.9569 7.2214 6.0022 7.9506]
0.001428	[1.8786 2.9459 3.8549 1.0050 4.9544 7.2180 5.9993 7.9603]
0.00028	[1.8758 2.9423 3.8506 0.9917 4.9493 7.2112 5.9933 7.9798]
0.00014	[1.8755 2.9419 3.8501 0.9901 4.9487 7.2104 5.9926 7.9822]
0.00009	[1.8753 2.9417 3.8499 0.9895 4.9485 7.2101 5.9924 7.9830]
0.0000001	[1.8751 2.9414 3.8496 0.9885 4.9481 7.2096 5.9919 7.9846]

Table 4.1: Estimated rank table

ε	$\|\|r-r^*\|\|_1$	$\|\|r-r^*\|\|_2$	$\|\|r-r^*\|\|_m$	$\|\|r-r^*\|\|_{D'_w}$
0.003	0.6690	0.3062	0.2273	0
0.002	0.6430	0.3017	0.2214	0
0.001428	0.6297	0.2999	0.2180	0
0.00028	0.6283	0.2980	0.2112	0
0.00014	0.6294	0.2979	0.2104	0
0.00009	0.6298	0.2979	0.2101	0
0.0000001	0.6304	0.2978	0.2096	0

Table 4.2: Error table

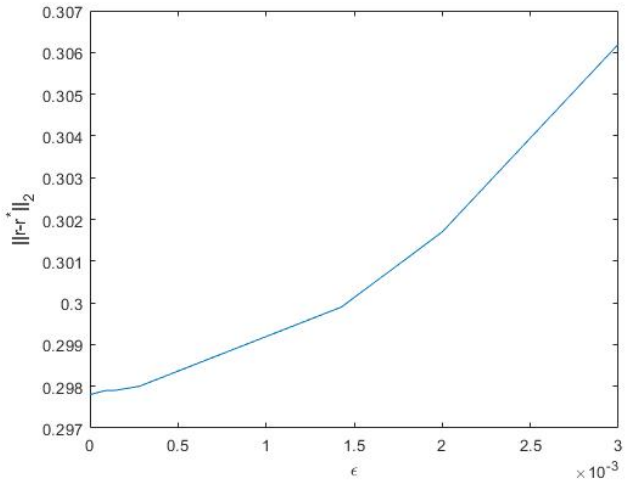

Figure 4.1: Effect of ε on error ($||r-r^*||_2$)

Case 2: For S given by

$$S_{ij} = \begin{cases} 0.001428 & \textit{for i=4, j=8} \\ -0.001428 & \textit{for i=8, j=4} \\ 0 & \textit{otherwise} \end{cases},$$

We have:

k	r^*
100	[1.7918 2.9763 4.3360 1.1675 4.6508 6.8869 5.3924 8.5855]
1000	[1.9956 3.0614 3.9340 1.0196 5.1867 7.1017 6.0297 7.7794]
10000	[1.9769 2.9952 4.0156 0.9946 5.0689 6.9826 5.9601 8.0072]

Table 4.3: estimated rank table

k	$\|\|r-r^*\|\|_1$	$\|\|r-r^*\|\|_2$	$\|\|r-r^*\|\|_m$	$\|\|r-r^*\|\|_{D'_w}$
100	2.3909	1.0157	0.6076	0.1892
1000	0.6901	0.3214	0.2206	0.2996
10000	0.1823	0.0867	0.0689	0

Table 4.4: Error table

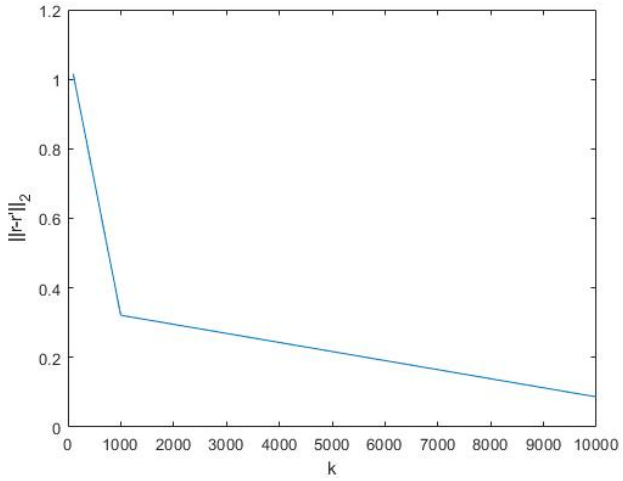

Figure 4.2: Effect of k on error($||r - r^*||_2$)

II. Here we compare the error in estimation when the perturbation introduced is concentrated at one pair (i,j) and when it is spread out across the matrix P.
Fix $\varepsilon = 0.000028$, $k = 10000$ and the perturbation matrices S and S' given by:

$$S_{ij} = \begin{cases} \varepsilon & for\ (i,j) = (a,b) \\ -\varepsilon & for\ (i,j) = (b,a) \\ 0 & otherwise \end{cases}$$

and

$$S'_{ij} = \begin{cases} \varepsilon/2 & for\ (i,j) \in \{(a,b),(c,d)\} \\ -\varepsilon/2 & for\ (i,j) \in \{(b,a),(d,c)\} \\ 0 & otherwise \end{cases},$$

We make following observations:

Entry perturbed	$\|\|r-r^*\|\|_1$	$\|\|r-r^*\|\|_2$	$\|\|r-r^*\|\|_m$
(1,5) and (4,8)	0.1525	0.0614	0.0404
(1,5)	0.1520	0.0627	0.0425
(4,8)	0.1531	0.0604	0.0383

Table 4.5: Error table

Entry perturbed	$\|\|r-r^*\|\|_1$	$\|\|r-r^*\|\|_2$	$\|\|r-r^*\|\|_m$
(3,4) and (5,6)	0.1572	0.0631	0.0427
(3,4)	0.1571	0.0638	0.0433
(5,6)	0.1574	0.0629	0.0422

Table 4.6: Error table

Entry perturbed	$\|\|r-r^*\|\|_1$	$\|\|r-r^*\|\|_2$	$\|\|r-r^*\|\|_m$
(1,2) and (4,8)	0.1536	0.0618	0.0406
(1,2)	0.1541	0.0633	0.0429
(4,8)	0.1531	0.0604	0.0383

Table 4.7: Error table

We observe that the error obtained when two entries are perturbed is very close to the average of the error obtained when those entries are perturbed independently in different experiments.

III. Now, we consider matrix S that perturb only one comparison. The question we want to answer is with full knowledge of actual rank, should the adversary perturb (i, j) such that ranks of i and j are far apart or the entry (k,l) such that ranks of k and l are close.

Fix k=100000 and S is as follows:

$$S_{ij} = \begin{cases} 0.002 & for\ (i,j) = (a,b) \\ -0.002 & for\ (i,j) = (b,a) \\ 0 & otherwise. \end{cases}$$

We want to observe the dependence of estimated rank r^* on the difference of the ranks corresponding to the entry perturbed, ie. on $r(a) - r(b)$, when (a,b) is perturbed.

1. $|r(a) - r(b)| = 7$

(a,b)	r^*
(4,8)	[2.0080 3.0030 4.0065 0.9979 4.9868 7.0039 5.9943 8.0081]

Table 4.8: Estimated rank table table

(a,b)	$\|\|r-r^*\|\|_1$	$\|\|r-r^*\|\|_2$	$\|\|r-r^*\|\|_m$	$\|\|r-r^*\|\|_{D'_w}$
(4,8)	0.0504	0.0202	0.0132	0

Table 4.9: Error table

2. $|r(a) - r(b)| = 6$

(a,b)	r^*
(1,8)	[2.0055 3.0021 4.0054 0.9961 4.9856 7.0023 5.9928 8.0132]
(4,6)	[2.0074 3.0022 4.0055 0.9949 4.9858 7.0052 5.9931 8.0098]

Table 4.10: Estimated rank table

(a,b)	$\|\|r-r^*\|\|_1$	$\|\|r-r^*\|\|_2$	$\|\|r-r^*\|\|_m$	$\|\|r-r^*\|\|_{D'_w}$
(1,8)	0.0539	0.0227	0.0144	0
(4,6)	0.0565	0.0222	0.0142	0

Table 4.11: Error table

3. $|r(a) - r(b)| = 5$

(a,b)	r^*
(2,8)	[2.0074 2.9997 4.0054 0.9961 4.9856 7.0023 5.9928 8.0135]
(1,6)	[2.0074 2.9997 4.0054 0.9961 4.9856 7.0023 5.9928 8.0135]
(4,7)	[2.0075 3.0023 4.0057 0.9950 4.9859 7.0027 5.9954 8.0100]

Table 4.12: Estimated rank table

(a,b)	$\|\|r-r^*\|\|_1$	$\|\|r-r^*\|\|_2$	$\|\|r-r^*\|\|_m$	$\|\|r-r^*\|\|_{D'_w}$
(2,8)	0.0544	0.0234	0.0144	0
(1,6)	0.0539	0.0215	0.0142	0
(4,7)	0.0519	0.0211	0.0141	0

Table 4.13: Error table

4. $|r(a) - r(b)| = 4$

(a,b)	r^*
(3,8)	[2.0074 3.0021 4.0024 0.9961 4.9856 7.0023 5.9929 8.0140]
(2,6)	[2.0075 3.0000 4.0056 0.9962 4.9858 7.0059 5.9931 8.0098]
(1,7)	[2.0060 3.0023 4.0057 0.9962 4.9859 7.0027 5.9957 8.0100]
(4,5)	[2.0076 3.0024 4.0057 0.9952 4.9877 7.0029 5.9934 8.0102]

Table 4.14: Estimated rank table

(a,b)	$\|\|r-r^*\|\|_1$	$\|\|r-r^*\|\|_2$	$\|\|r-r^*\|\|_m$	$\|\|r-r^*\|\|_{D'_w}$
(3,8)	0.0536	0.0232	0.0144	0
(2,6)	0.0538	0.0220	0.0142	0
(1,7)	0.0490	0.0203	0.0141	0
(4,5)	0.0525	0.0207	0.0123	0

Table 4.15: Error table

5. $|r(a) - r(b)| = 3$

(a,b)	r^*
(5,8)	[2.0075 3.0022 4.0055 0.9962 4.9821 7.0024 5.9930 8.0144]
(4,7)	[2.0075 3.0023 4.0028 0.9962 4.9858 7.0063 5.9931 8.0099]
(3,6)	[2.0075 3.0003 4.0057 0.9962 4.9859 7.0027 5.9960 8.0101]
(2,5)	[2.0076 3.0016 4.0060 0.9960 4.9880 7.0029 5.9934 8.0102]
(1,4)	[2.0076 3.0024 4.0071 0.9954 4.9861 7.0030 5.9934 8.0103]

Table 4.16: Estimated rank table

(a,b)	$\|\|r-r^*\|\|_1$	$\|\|r-r^*\|\|_2$	$\|\|r-r^*\|\|_m$	$\|\|r-r^*\|\|_{D'_w}$
(5,8)	0.0608	0.0263	0.0179	0
(4,7)	0.0537	0.0217	0.0142	0
(3,6)	0.0483	0.0207	0.0141	0
(2,5)	0.0471	0.0198	0.0140	0
(1,4)	0.0554	0.0220	0.0139	0

Table 4.17: Error table

6. $|r(a) - r(b)| = 2$

(a,b)	r^*
(7,8)	[2.0075 3.0023 4.0056 0.9962 4.9858 7.0026 5.9888 8.0150]
(5,6)	[2.0075 3.0023 4.0057 0.9962 4.9826 7.0067 5.9932 8.0100]
(3,7)	[2.0076 3.0024 4.0032 0.9962 4.9860 7.0028 5.9964 8.0101]
(2,5)	[2.0071 3.0024 4.0032 0.9960 4.9883 7.0029 5.9934 8.0102]
(1,4)	[2.0071 3.0025 4.0059 0.9968 4.9862 7.0031 5.9936 8.0105]
(4,2)	[2.0076 3.0034 4.0059 0.9956 4.9862 7.0030 5.9935 8.0104]

Table 4.18: Estimated rank table

(a,b)	$\|\|r-r^*\|\|_1$	$\|\|r-r^*\|\|_2$	$\|\|r-r^*\|\|_m$	$\|\|r-r^*\|\|_{D'_w}$
(7,8)	0.0621	0.0258	0.0150	0
(5,6)	0.0603	0.0246	0.0174	0
(3,7)	0.0476	0.0202	0.0140	0
(2,5)	0.0540	0.0200	0.0139	0
(1,4)	0.0526	0.0213	0.0138	0
(4,2)	0.0550	0.0218	0.0138	0

Table 4.19: Error table

7. $|r(a) - r(b)| = 1$

(a,b)	r^*
(4,1)	[2.0082 3.0025 4.0059 0.9958 4.9862 7.0031 5.9935 8.0104]
(1,2)	[2.0067 3.0036 4.0059 0.9962 4.9862 7.0031 5.9935 8.0104]
(2,3)	[2.0076 3.0009 4.0076 0.9962 4.9861 7.0030 5.9935 8.0104]
(3,5)	[2.0076 3.0025 4.0036 0.9962 4.9886 7.0030 5.9935 8.0103]
(5,7)	[2.0076 3.0024 4.0058 0.9962 4.9830 7.0029 5.9968 8.0103]
(6,7)	[2.0078 3.0026 4.0061 0.9963 4.9864 6.9990 5.9978 8.0107]
(6,8)	[2.0076 3.0024 4.0057 0.9962 4.9860 6.9978 5.9933 8.0156]

Table 4.20: Estimated rank table

(a,b)	$\|\|r-r^*\|\|_1$	$\|\|r-r^*\|\|_2$	$\|\|r-r^*\|\|_m$	$\|\|r-r^*\|\|_{D'_w}$
(4,1)	0.0546	0.0219	0.0138	0
(1,2)	0.0537	0.0214	0.0138	0
(2,3)	0.0537	0.0220	0.0139	0
(3,5)	0.0487	0.0195	0.0114	0
(5,7)	0.0530	0.0229	0.0170	0
(6,7)	0.0478	0.0206	0.0136	0
(6,8)	0.0580	0.0245	0.0156	0

Table 4.21: Error table

Diff of ranks	average of $\|\|r-r^*\|\|_2$
7	0.202
6	0.02245
5	0.022
4	0.02155
3	0.0221
2	0.022
1	0.021828

Table 4.22: Average error table

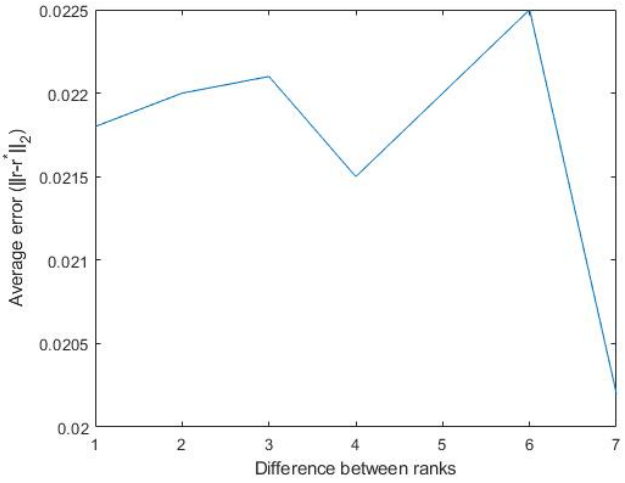

Figure 4.3

From the above table 4.22 and figure 4.3 we observe that the average error $||r - r^*||_2$ is maximum when the difference of ranks of the pair (i,j) that is perturbed is 6.

IV. In this case we tried to construct a perturbation matrix P that results in flipping the rank 5 and 8. In other words, for r=[2 3 4 1 5 7 6 8], We want to flip the rank of 5^{th} and 8^{th} item. We consider three different perturbation matrices, where we perturbed more and more entries. Three such perturbation matrices are listed below followed by the estimated ranks r^*.

- For

$$S_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0.0074 & 0 & 0 & -0.0074 \\ 0 & 0 & 0 & 0 & 0.0153 & 0 & 0 & -0.0153 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.0074 & -0.0153 & 0 & 0 & 0 & 0 & 0 & -0.0351 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.0074 & 0.0153 & 0 & 0 & 0.0351 & 0 & 0 & 0 \end{bmatrix},$$

  r_1^* = [2.0703 3.0451 4.2446 1.0374 6.3079 7.3111 6.1429 6.3602].

- For

$$S_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0.0074 & 0 & 0 & -0.0074 \\ 0 & 0 & 0 & 0 & 0.0153 & 0 & 0 & -0.0153 \\ 0 & 0 & 0 & 0 & 0.0120 & 0 & 0 & -0.0120 \\ 0 & 0 & 0 & 0 & 0.0074 & 0 & 0 & -0.0074 \\ -0.0074 & -0.0153 & -0.0120- & 0.0074 & 0 & 0 & 0 & -0.0351 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.0074 & 0.0153 & 0.0120 & 0.0074 & 0.0351 & 0 & 0 & 0 \end{bmatrix},$$

r_2^* = [2.0703 3.0451 4.2446 1.0374 6.3079 7.3111 6.1429 6.3602].

- For

$$S_3 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0.0074 & 0 & 0 & -0.0074 \\ 0 & 0 & 0 & 0 & 0.0153 & 0 & 0 & -0.0153 \\ 0 & 0 & 0 & 0 & 0.0120 & 0 & 0 & -0.0120 \\ 0 & 0 & 0 & 0 & 0.0074 & 0 & 0 & -0.0074 \\ -0.0074 & -0.0153 & -0.0120 & -0.0074 & -0.4777 & -0.0220 & -0.0154 & -0.0351 \\ 0 & 0 & 0 & 0 & 0.0220 & 0 & 0 & -0.0220 \\ 0 & 0 & 0 & 0 & 0.0154 & 0 & 0 & -0.0154 \\ 0.0074 & 0.0153 & 0.0120 & 0.0074 & 0.0351 & 0.0220 & 0.0154 & 0 \end{bmatrix},$$

r_3^* = [2.0257 2.9663 4.1553 1.0087 7.8598 7.1379 6.0125 4.9009].

As expected, perturbing more entries (in a specifically chosen way) leads to flipping of rank. Thus, an adversary with complete knowledge of actual rank can construct a perturbation matrix that leads to flipping of ranks in his/her favour.

Concluding Remarks and future directions

The aim of the thesis was to study the ranking problem from [SSD17] under perturbation using tools from concentration inequalities. As mentioned earlier, one could generalise the work in [SSD17] to a setup where number of comparisons of each pair of items is distinct. For the perturbation part, we would like to get some theoretical results that help the adversary to choose a suitable perturbation matrix. Another direction to explore is to look at the case of random perturbations.

Bibliography

[SSD17] Sahand Negahban, Sewoong Oh, Devavrat Shah,*Rank Centrality: Ranking from Pairwise Comparisons*, Operations Research 65(1):266-287, 2017.

[J11] Joel A. Tropp,*User-friendly tail bounds for sums of random matrices*,Found. Comput. Math., Vol. 12, num. 4, pp. 389-434, 2011

[SPG12] Stephane Bouchdron, Pascal Massart, Gabor Lugosi,*Concentration Inequalities: a known asymptotic theory of independence*,Clarendon press .Oxford, 2012

[DA09] Devdatt P. Dubhashi, Alessandro Panconesi,*Concentration of measure for the analysis of randomized algorithms*,Cambridge Press, 2009