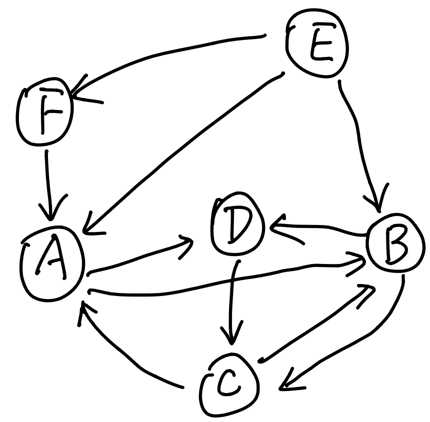
Case Studies A

PageRank

Problem: Rank the webpages?

Data: The linkages of webpages, e.g.



A, B, ⋯, F are webpages.

"Ⓔ→Ⓕ" means webpage Ⓔ has a hyperlink to Ⓕ, and so on.

Model: Let π_i denote the score of node i.

In the simplest form, Google uses the following model

$$\pi_i = (1-p)\frac{1}{n} + p\sum_{j\in N(i)}\frac{\pi_j}{L_j}, \quad \cdots\cdots (PR)$$

where n — total number of nodes

$N(i)$ — number of nodes pointing to node i.

L_j — the number of outbound links of node j.

p — a parameter satisfying $0<p<1$.

Actually, π_i is the probability representing the likelihood that a preson randomly clicking on links will arrive at webpage i. There are two terms in π_i:

$(1-p)\frac{1}{n}$ — the person will, with probability $1-p$, open an arbitrary new webpage with equal probability, so he/she opens webpage i with prob. $(1-p)\frac{1}{n}$.

$p\sum_{j\in N(i)}\frac{\pi_j}{L_j}$ — the preson will, with probability p, click links an arbitrary link on the current webpage with equal probability. So, the probability that the person arrives webpage i is

$$p \sum_{j \in N(i)} \frac{\pi_j}{L_j}$$

π_j: the probability the current webpage is j.

L_j: there are L_j outbounds of j.

So the conditional probability from $j \to i$ is $\frac{\pi_j}{L_j}$

Example: Consider the example above. The equation (PR) is

$$\begin{bmatrix} \pi_A \\ \pi_B \\ \pi_C \\ \pi_D \\ \pi_E \\ \pi_F \end{bmatrix} = (1-p) \begin{bmatrix} 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \end{bmatrix} + p \begin{bmatrix} 0 & 0 & 1/2 & 0 & 1/3 & 1 \\ 1/2 & 0 & 1/2 & 0 & 1/3 & 0 \\ 0 & 1/2 & 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 0 \end{bmatrix} \begin{bmatrix} \pi_A \\ \pi_B \\ \pi_C \\ \pi_D \\ \pi_E \\ \pi_F \end{bmatrix}$$

e.g., C has two outbound links (to A and B)

normalized adjacency matri~~x~~ (s.t. column sum is 1)

This is a linear equation on vector $\pi \in \mathbb{R}^6$. We can use, e.g., LU decomposition, to solve it. If $p = 0.85$, then the solution is

$$\pi = \begin{bmatrix} 0.189 \\ 0.242 \\ 0.305 \\ 0.208 \\ 0.025 \\ 0.032 \end{bmatrix}.$$

So the ranking is

rank	node	score	in	out
1	C	0.305	2	2
2	B	0.242	3	2
3	D	0.208	2	1
4	A	0.189	3	2
5	F	0.032	1	1
6	E	0.025	0	3

- Why C ranks higher than B?
- Why B ranks higher than A?

In general, Equation (PR) can be written in matrix form as

$$\pi = \frac{1-p}{n} \cdot \mathbb{1} + pA\pi,$$

where $\pi \in \mathbb{R}^n$ is the score vector, $\mathbb{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n$, and A is the normalized adjancency matrix with each column sum 1. We obtain

$$(I - pA)\pi = \frac{1-p}{n}\mathbb{1}, \qquad \cdots\cdots$$

which can be solved by LU decomposition

Question: Prove that the solution $\mathbb{1}^T \pi = 1$ and $\pi_i \geq 0 \ \forall i$. (i.e., prove that the solution π is indeed a probability distribution).

Open questions: ① How to improve the model (PR)?

② Any other possible applications of (PR)?

Image Deconvolution

Problem: Recover the clear image from a blurred one.

Model: For simlicity, we consider 1D example, i.e. the unknown clear image is a vector $x \in \mathbb{R}^n$, where x_i, $i = 1, \cdots, n$ is gray level at pixel i.

We also assume the blurring is a moving average with weights $\frac{1}{3}[1 \ 1 \ 1]$ then the observed blurred image $b \in \mathbb{R}^n$ satisfyies

(other weights are possible, depending on physics model)

$$b_i = (x_{i-1} + x_i + x_{i+1})/3 \qquad i = 1, 2, \cdots, n.$$

For $i = 1$, we have $b_1 = (x_0 + x_1 + x_2)/3$

For $i = n$, we have $b_n = (x_{n-1} + x_n + x_{n+1})/3$

where x_0, x_{n+1} are not defined yet. we assume $x \in \mathbb{R}^n$ is periodic, i.e.,

$$x_1 \cdots x_n \,|\, x_1 \ x_2 \ \cdots \ x_n \,|\, x_1 \ x_2 \cdots$$

$$x_0 = x_n \ \text{ and } \ x_{n+1} = x_1$$ — Called boundary condition.

Therefore, the model is

$$
\begin{cases}
b_1 = (x_1 + x_2 + x_n)/3 \\
b_i = (x_{i-1} + x_i + x_{i+1})/3, & i = 2, \cdots, n-1 \\
b_n = (x_1 + x_{n-1} + x_n)/3
\end{cases}
$$

In matrix form,

$$A x = b,$$

where

$$
A = \frac{1}{3}\begin{bmatrix}
1 & 1 & & & 1 \\
1 & 1 & 1 & & \\
& \ddots & \ddots & \ddots & \\
& & \ddots & \ddots & 1 \\
1 & & & 1 & 1
\end{bmatrix}
$$

The de-blur is to solve x from $Ax=b$.

In general, many blurring can be modeled as $Ax=b$.

Solution: However, A is singular (i.e., not invertible), and there are infinitely many solutions of $Ax=b$.

To have a unique solution, we use Tikhonov regularization, where we solve

$$\min_x \|Ax-b\|_2^2 + \alpha \|x\|_2^2, \quad \cdots\cdots \quad (TK)$$

Here $\alpha > 0$ is called regularization parameter.

In other words, we find an $x \in \mathbb{R}^n$ s.t.

① $\|Ax-b\|_2^2$ is small.

② $\|x\|_2^2$ is small.

α balances the relative importance of ① and ②.

Larger α $\Longrightarrow$ $\|Ax-b\|_2^2$ is larger, and $\|x\|_2^2$ is smaller.

smaller α $\Longrightarrow$ $\|Ax-b\|_2^2$ is smaller, and $\|x\|_2^2$ larger.

To solve the minimization (TK), we take gradient of the objective function and set it to 0. We obtain (TK) is equivalent to solve

$$2A^T(Ax-b) + 2\alpha x = 0$$

$$\Updownarrow$$

$$(A^TA + \alpha I)\, x = A^Tb.$$

Fact: $A^TA + \alpha I$ is symmetric positive definite.

proof. i) $(A^TA+\alpha I)^T = (A^TA)^T + \alpha I^T = A^TA + \alpha I$

ii) $\forall\, y \neq 0,\quad y^T(A^TA+\alpha I)y = y^TA^TAy + \alpha y^Ty$

$$= \|Ay\|_2^2 + \alpha\|y\|_2^2 > 0 \quad (\text{since } \|y\|_2 > 0,\ \alpha > 0)\ \boxtimes$$

Therefore, we can use Cholesky decomposition to solve

$$(A^TA + \alpha I)\, x = A^Tb.$$