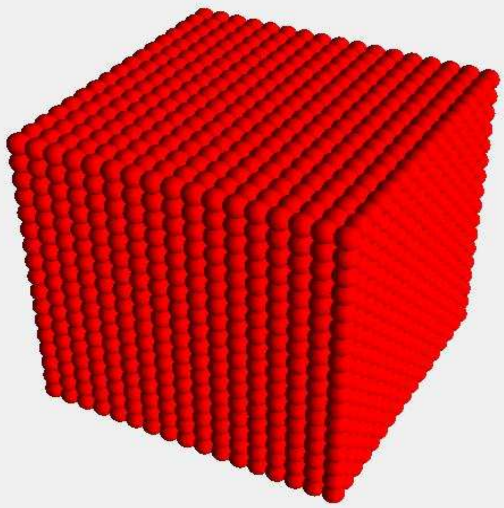
MathCon
The Mathematics Firm

Page Rank



Contenido

1
PageRank

1.1. Introducción

[redacted]

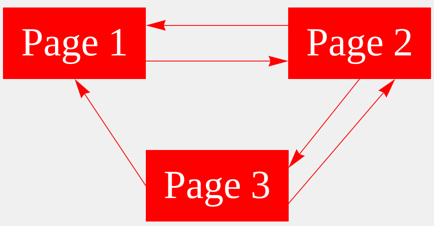
En esta ocasión nos dedicamos a estudiar el problema de “PageRank”, que es un algoritmo que lista las páginas web en orden de importancia. Obviamente pueden existir diferentes criterios para dar una medida de importancia a una página web, sin embargo aquí explicaremos el que se usa en el algoritmo usado por Google.

1.2. Matriz de conectividad

[redacted]

Un ejemplo

Para comenzar nuestro estudio pensemos en un baby-Internet que consiste solo de tres páginas web y que están conectadas como muestra la figura, es decir, en la página 1 existen una liga para que se acceda a la página 2, de manera similar en la página 2 existe una liga para poder acceder a la página 1. Ocurre lo mismo para las páginas 2 y 3. Para las páginas 1 y 3, solo existe un botón en 3 para ir a 1.



Si x_i es el número de ligas que hacen referencia a la página i. Del ejemplo anterior tenemos que:

$$
\begin{aligned}
x_1 &= 2 \\
x_2 &= 2 \\
x_3 &= 1
\end{aligned}
$$

Entonces, de acuerdo a qué páginas tienen una liga de la página i:

$$\begin{aligned} x_1 &= x_2 + x_3 \\ x_2 &= x_1 + x_3 \\ x_3 &= x_2 \end{aligned}$$

Si añadimos la información de la proporción de ligas que tiene cada página, queda como:

$$\begin{array}{lcllcl} x_1 & = & & x_2/2 & + & x_3/2 \\ x_2 & = & x_1 & & + & x_3/2 \\ x_3 & = & & x_2/2 & & \end{array}$$

Escrita en forma matricial:

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

Matriz de conectividad: es una matriz C_{ij} donde $c_{ij} = 0$ si no existe una liga de la página j a la página i y $c_{ij} = \frac{1}{k}$ donde k es el número total de ligas que salen de la página j siempre y cuando exista una liga a la página i.

Para nuestro caso tenemos una matriz de conectividad 3×3.

$$M_t = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix}$$

Calculo del PageRank

Para la página 1 Veamos que ocurre con el pagerank de la página 1,

- **De la página 2** Recibe $1/2$ del pagerank de la 2, debido a que la página 2 tiene 2 salidas.
- **De la página 3** Recibe también $1/2$ de la página 3 ya que tiene dos salidas.

Por lo tanto la página 1 tiene un pagerank de $(1/2)(1/3)$ de 3 y $(1/2)(1/3)$ de 2, después de la primera transición, es decir un total de $(1/2)(1/3) + (1/2)(1/3) = 1/3$.

Para la página 2 Veamos que ocurre con el pagerank de la página 2,

- **De la página 1** Recibe 1 de esta página, debido a que está página solo tiene una salida.
- **De la página 3** Recibe 1/2 de esta página, debido a que está página tiene 2 salidas.

Por lo tanto la página 2 tiene un pagerank de $(1/2)(1/3)$ de 1 y $(1)(1/3)$ de 3, después de la primera transición. Es decir un total de $(1)(1/3) + (1/2)(1/3) = 1/2$.

Para la página 3 Veamos que ocurre con el pagerank de la página 3,

- **De la página 1** No recibe nada de esta página.
- **De la página 2** Recibe $(1/2)$, debido a que la página 2 tiene dos ligas de salida.

Por lo tanto la página 3 tiene un pagerank de $(0)(1/3)$ de 1 y $(1/2)(1/3)$ de 3, después de la primera transición. Es decir un total de $(1/2)(1/3) = 1/6$.

En resumen el nuevo vector Pagerank es:

$$PR_1 = \begin{pmatrix} 1/3 \\ 1/2 \\ 1/6 \end{pmatrix}$$

Ahora cómo podemos medir la importancia de nuestras páginas. Para iniciar el estudio supongamos que las tres tienen la misma importancia definiendo al vector PageRank inicial, que para nuestro caso es:

$$x_0 = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}$$

De manera matricial tenemos que:

$$Mx_0 = x_1$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/2 \\ 1/6 \end{pmatrix} = \begin{pmatrix} 0{,}3333 \\ 0{,}5000 \\ 0{,}1666 \end{pmatrix}$$

Sin embargo no hay razón para pensar que x_2 sea mejor medida que x_1. Entonces ampliamos el número de transiciones, y veamos que sucede.

$$Mx_1 = x_2$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 1/2 \\ 1/6 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 5/12 \\ 1/4 \end{pmatrix} = \begin{pmatrix} 0{,}3333 \\ 0{,}4166 \\ 0{,}2500 \end{pmatrix}$$

$$Mx_2 = x_3$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 5/12 \\ 1/4 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 11/24 \\ 5/24 \end{pmatrix} = \begin{pmatrix} 0{,}3333 \\ 0{,}4583 \\ 0{,}2083 \end{pmatrix}$$

$$Mx_3 = x_4$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 11/24 \\ 5/24 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 7/16 \\ 11/48 \end{pmatrix} = \begin{pmatrix} 0{,}3333 \\ 0{,}4375 \\ 0{,}2291 \end{pmatrix}$$

$$Mx_4 = x_5$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 7/16 \\ 11/48 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 43/96 \\ 7/32 \end{pmatrix} = \begin{pmatrix} 0{,}3333 \\ 0{,}4479 \\ 0{,}2187 \end{pmatrix}$$

$$Mx_5 = x_6$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 43/96 \\ 7/32 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 85/192 \\ 43/192 \end{pmatrix} = \begin{pmatrix} 0{,}3333 \\ 0{,}4427 \\ 0{,}2239 \end{pmatrix}$$

$$Mx_6 = x_7$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 85/192 \\ 43/192 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 57/128 \\ 85/384 \end{pmatrix} = \begin{pmatrix} 0{,}3333 \\ 0{,}4453 \\ 0{,}2213 \end{pmatrix}$$

$$Mx_7 = x_8$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 57/128 \\ 85/384 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 341/768 \\ 57/256 \end{pmatrix} = \begin{pmatrix} 0{,}3333 \\ 0{,}4440 \\ 0{,}2226 \end{pmatrix}$$

$$Mx_8 = x_9$$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 341/768 \\ 57/256 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 683/1536 \\ 341/1536 \end{pmatrix} = \begin{pmatrix} 0{,}3333 \\ 0{,}4446 \\ 0{,}2220 \end{pmatrix}$$

Lo anterior nos permite inferir que esta sucesión converge al vector:

$$x = \begin{pmatrix} 0{,}333 \\ 0{,}444 \\ 0{,}222 \end{pmatrix}$$

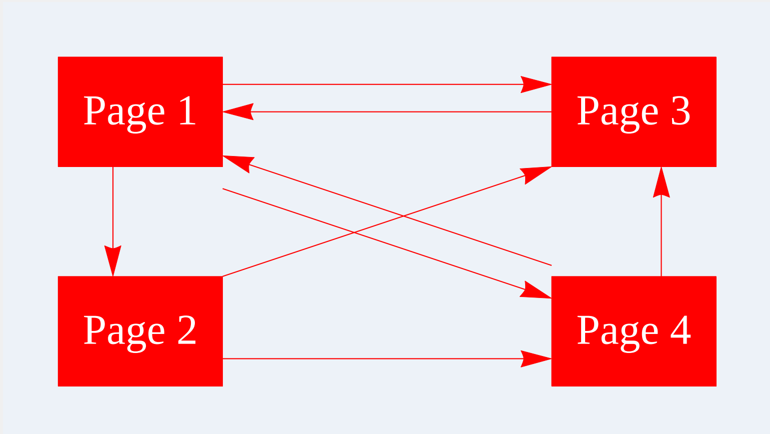
Además observamos que:

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 0{,}333 \\ 0{,}444 \\ 0{,}222 \end{pmatrix} = \begin{pmatrix} 0{,}333 \\ 0{,}444 \\ 0{,}222 \end{pmatrix}$$

Es decir x es un vector propio de la matriz de conectividad M asociado al valor propio 1.

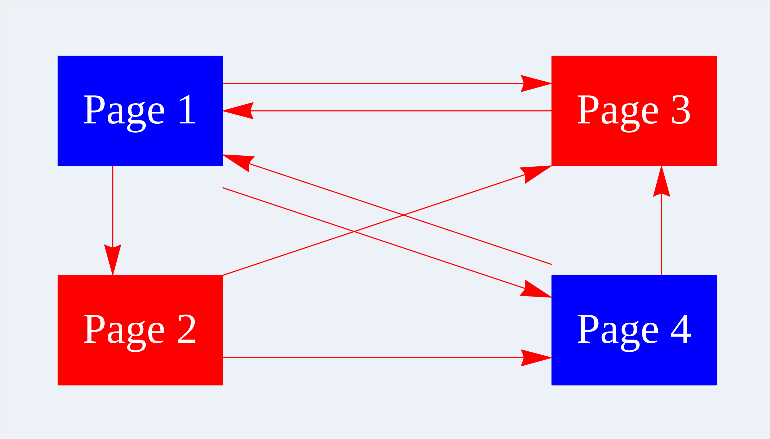
$$Mx = (1)x$$

Otro ejemplo



- Puntuación de la página k, es x_k.
- x_k = el número de ligas dirigidas a la página k.

$$\begin{aligned} x_1 &= 2 \\ x_2 &= 1 \\ x_3 &= 3 \\ x_4 &= 2 \end{aligned}$$



- Si la página j contiene n_j ligas, una de las cuales va a la página k, la puntuación de la página k se calcula por x_j/n_j.
- $x_k = \sum_{j \in L_k} \frac{x_j}{n_j}$

$$\begin{array}{lclclclcl} x_1 & = & & & & & x_3/1 & + & x_4/2 \\ x_2 & = & x_1/3 & & & & & & \\ x_3 & = & x_1/3 & + & x_2/2 & & & + & x_4/2 \\ x_4 & = & x_1/3 & + & x_2/2 & & & & \end{array}$$

En forma matricial queda como:

$$\begin{pmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

Obtenemos la ecuación matricial $Ax = x$, es decir, el vector x deberá ser también el vector propio del valor propio 1.

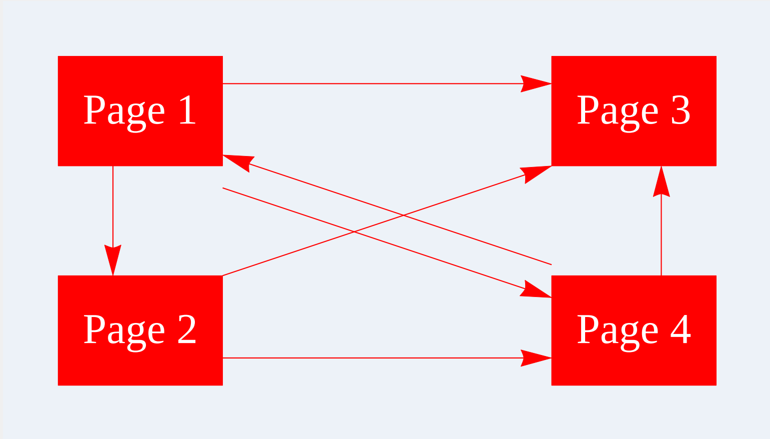
Si calculamos el vector PageRank de manera similar que antes, obtenemos:

$$\begin{array}{lcrcl} x_1 & = & (1/4)+(1/4)(1/2) & = & 0.375 \\ x_2 & = & (1/3)(1/4) & = & 0.083 \\ x_3 & = & (1/3)(1/4) + (1/2)(1/4) + (1/2)(1/4) & = & 0.333 \\ x_4 & = & (1/3)(1/4) + (1/2)\ (1/4) & = & 0.2 \end{array}$$

SORPRESA!!, no es el orden que se esperaba de PageRank, es decir, que el orden fuera (3,1,4,2) o (3,4,1,2).

Esto es debido a que la página 3 solo tiene un voto a la página 1.

Nodos Colgantes.



Matriz de conectividad de una red con un nodo colgante.

$$\begin{pmatrix} 0 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix}$$

Esta matriz no tiene como valor propio el 1. Es llamada una matriz sub estocastica.

1.3. Algo de definiciones

[redacted]

Ahora estamos en condiciones de formalizar algunos conceptos anteriores.

Definición 1 *Una red Internet es una gráfica dirigida, donde en cada nodo sale al menos una dirección y no existen direcciones que salen y entran al mismo nodo.*

Definición 2 *Sea una red Internet con n vértices y sea k_j el número de aristas que tienen como origen el vértice P_j. La matriz de conectividad de la red Internet es la matriz $n \times n$ $A = (a_{ij})$, donde:*

$$a_{ij} = \begin{cases} \dfrac{1}{k_j} & \text{si existe una arista de } P_j \text{ a } P_i \\ 0 & \text{en otro caso} \end{cases}$$

Teorema 1 *Si A es una matriz de conectividad de una red Internet, entonces:*

1. $a_{ii} = 0$.

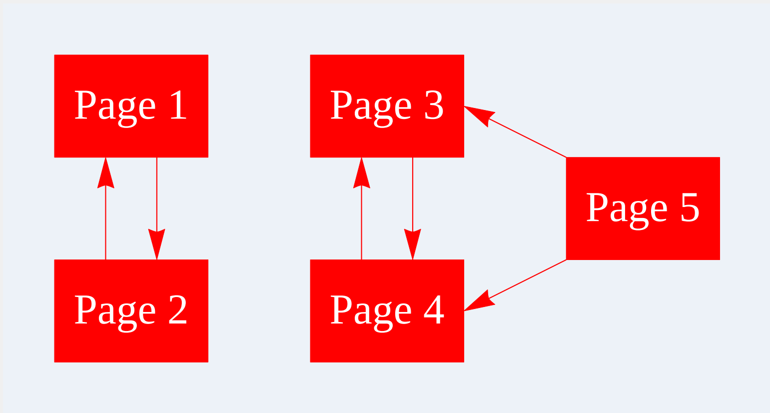
2. *Las entradas son no negativas, y de cada columna suman 1.*

3. *El número $\lambda = 1$ es un valor propio de A.*

Estas matrices reciben el nombre de matrices escotasticas.

Definición 3 *Si A es la matriz de conectividad de una red Internet, sea $X = (x_1, x_2, ..., x_n)$ un valor propio de A asociado con el valor propio $\lambda = 1$, y sea $s = x_1 + x_2 + \cdots x_n$. Si $s \neq 0$, entonces un simple PageRank de la red Internet es el vector $(1/s)X$.*

Redes disconexas.

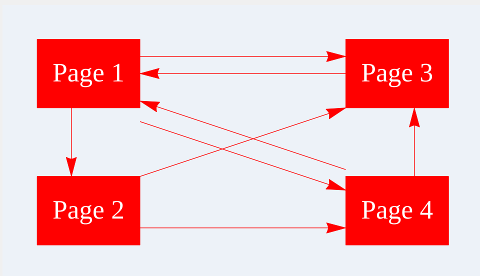


Las redes disconexas suelen tener dos valores propios 1, por lo tanto, dos vectores propios diferentes.

1.4. Matriz de conectividad modificada

Sea S una matriz $n \times n$ con las entradas todas $1/n$. Definimos la matriz $M = (1-m)A + mS$, donde $0 \leq m \leq 1$. El valor reportado por Google es $m = 0{,}15$

Ejemplo:



$$A = \begin{pmatrix} 0 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix}$$

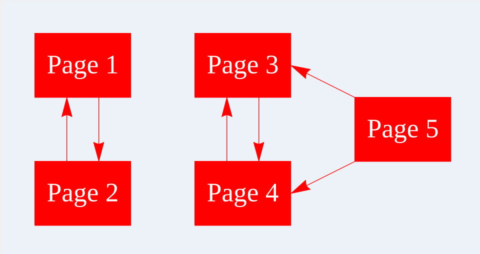
con $n = 4$, y $m = 0{,}15$, obtenemos:

$$S = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$

$$\begin{pmatrix} 0{,}0375 & 0{,}0375 & 0{,}8875 & 0{,}4625 \\ 0{,}320833 & 0{,}0375 & 0{,}0375 & 0{,}0375 \\ 0{,}320833 & 0{,}4625 & 0{,}0375 & 0{,}4625 \\ 0{,}320833 & 0{,}4625 & 0{,}0375 & 0{,}0375 \end{pmatrix}$$

Donde el vector **PageRank** es: $(0{,}6964, 0{,}2682, 0{,}5447, 0{,}3823)$

Ejemplo:



$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

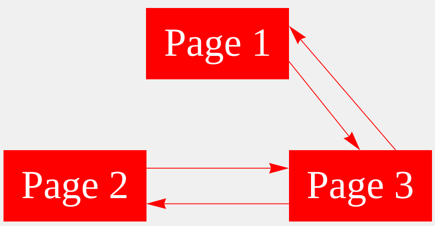
$$S = \begin{pmatrix} \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix}$$

$$X = \begin{pmatrix} 0{,}03 & 0{,}88 & 0{,}03 & 0{,}03 & 0{,}03 \\ 0{,}88 & 0{,}03 & 0{,}03 & 0{,}03 & 0{,}03 \\ 0{,}03 & 0{,}03 & 0{,}03 & 0{,}88 & 0{,}455 \\ 0{,}03 & 0{,}03 & 0{,}88 & 0{,}03 & 0{,}455 \\ 0{,}03 & 0{,}03 & 0{,}03 & 0{,}03 & 0{,}03 \end{pmatrix}$$

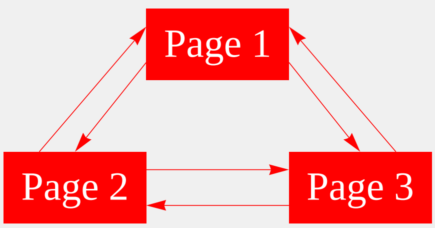
$$PR = (0{,}405429, 0{,}405429, 0{,}577736, 0{,}577736, 0{,}0608143)$$

1.5. Ejercicios

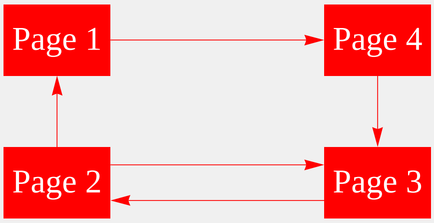
Encontrar la matriz de conectividad y el vector PageRank de las siguientes redes Internet.



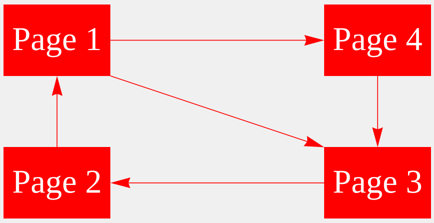
1.



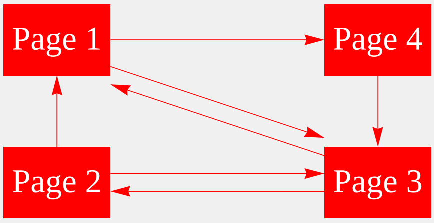
2.


Page 1
Page 4
Page 2
Page 3

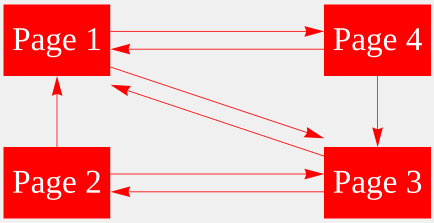
3.


Page 1
Page 4
Page 2
Page 3

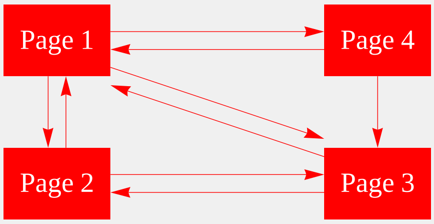
4.


Page 1
Page 4
Page 2
Page 3

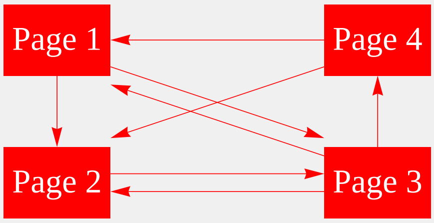
5.


Page 1
Page 4
Page 2
Page 3

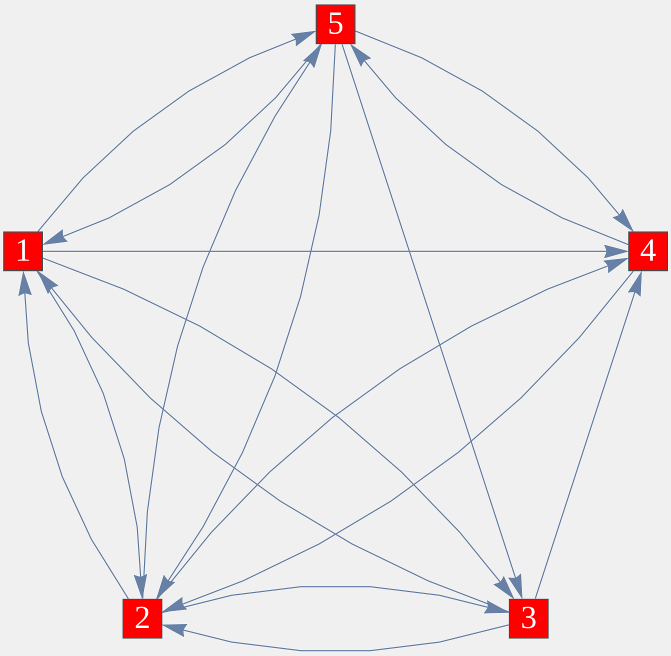
6.


Page 1
Page 4
Page 2
Page 3

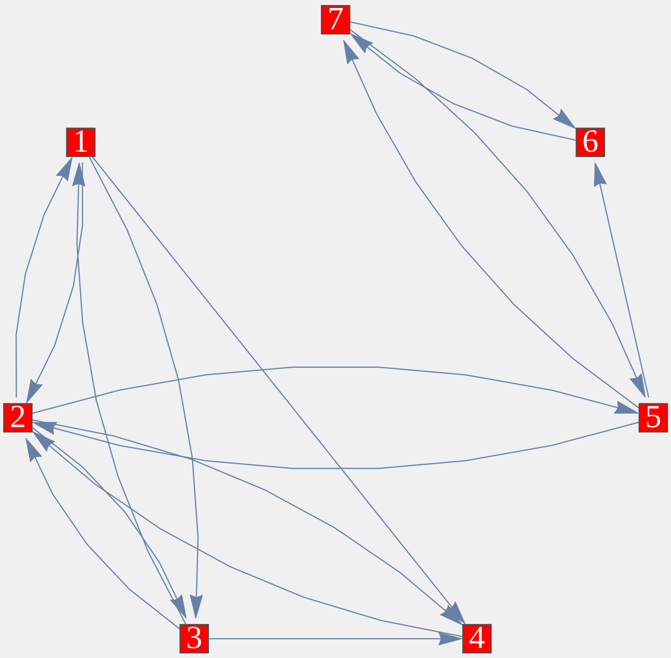
7.



8.



9.



10.