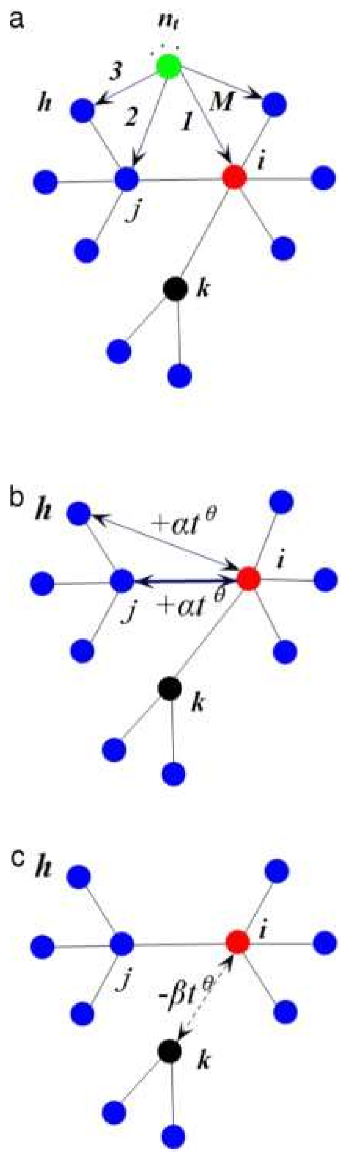
Dinámica de Redes

Inteligencia Colectiva y Sistemas de Recomendación

Master Propio en Data Science y Big Data

Redes Dinámicas

- Las redes complejas evolucionan en el tiempo:
  - Evolución en nodos
    - Nacimiento/Muerte de nodos
    - Variación de características internas
  - Evolución en aristas
    - Nacimiento/Muerte de aristas
    - Variación de características internas
  - Evolución simultánea en nodos y aristas
    - Suele ser la más común



Nacimiento/Muerte de Nodos/Aristas

- El ejemplo más paradigmático es el propio algoritmo de **Enlace Preferencial** visto en el bloque anterior (**Modelos Dinámicos**).
- Analicemos el algoritmo en el modelo…
- Pero hay muchos otros que intentan modelar comportamientos observados en sistemas reales.

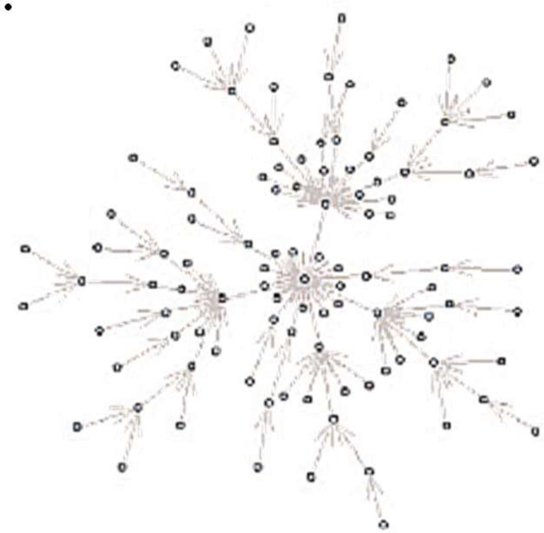
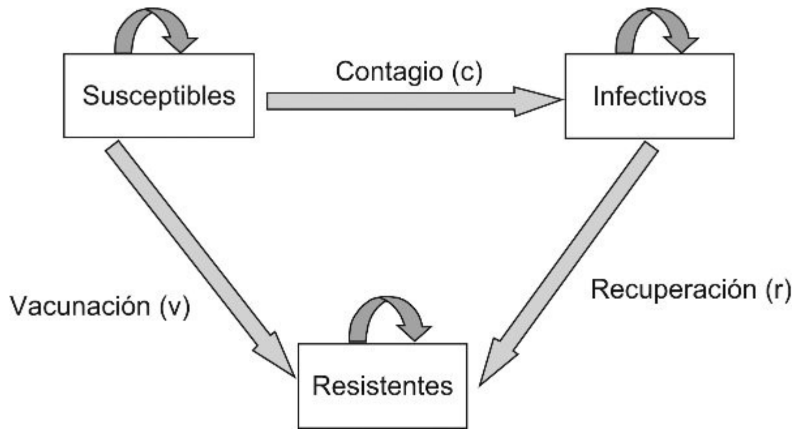


Fig. 14. Modelo de Enlace Preferencial. Los nodos más conectados tiene mayor probabilidad de atrapar a los recién llegados. Las flechas indican estos enlaces.

Variación de Características Internas (nodos)

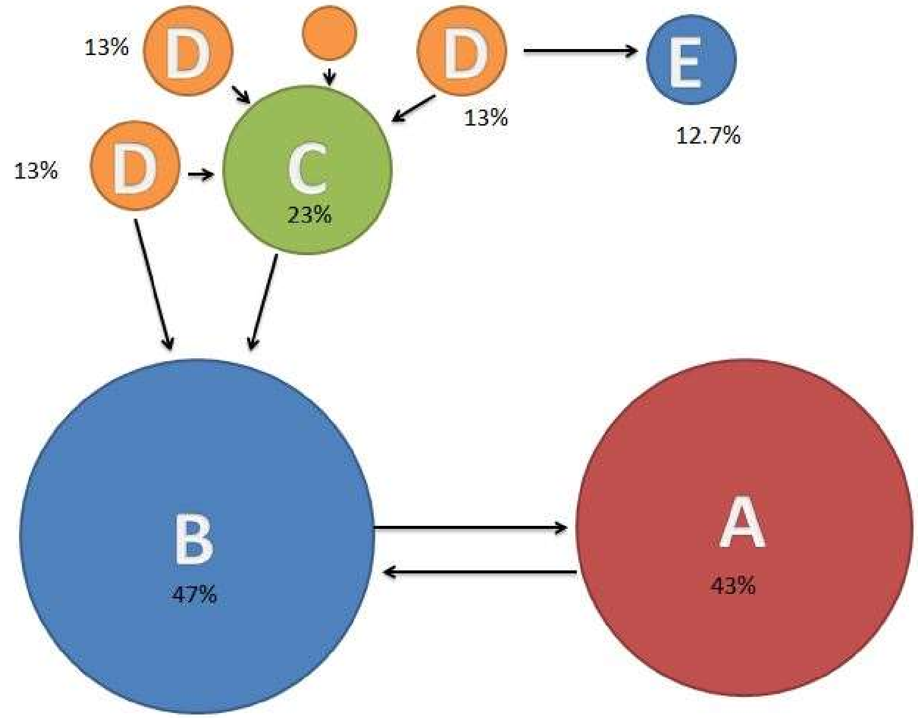
- El ejemplo más paradigmático es el de **SIR** de modelos epidemiológicos:
  - **S:** Susceptible
  - **I:** Infectado
  - **R:** Recuperado



- En determinadas condiciones se puede modelar con EDOs.
- Diversas variantes: SIS, SIRS, SEIS (E: Exposed), SEIR, MSIR (M: Inmunes pasivos), MSEIR,…

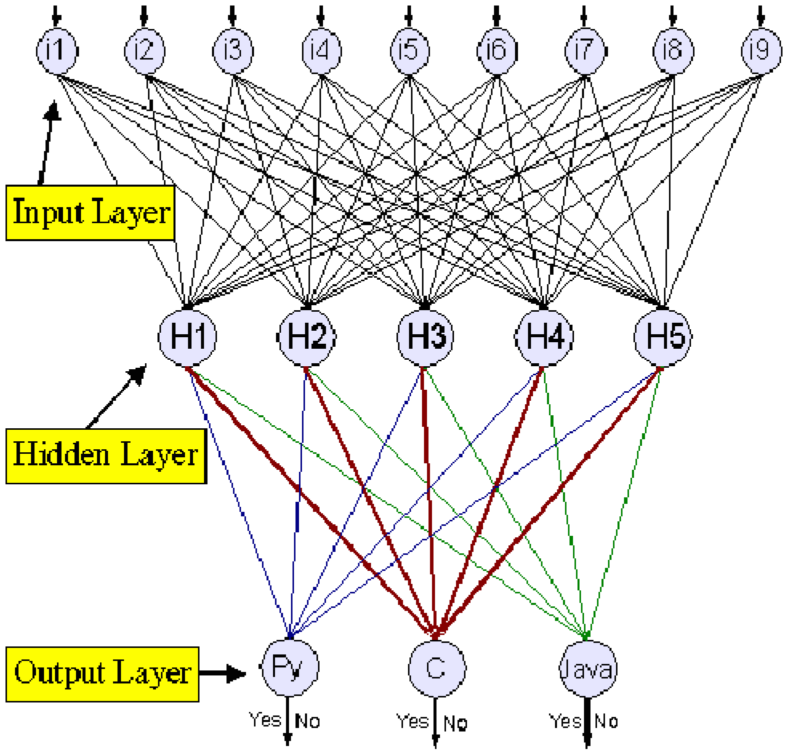
Variación de Características Internas (nodos)

- Dentro de este tipo de procesos dinámicos también encontramos los **procesos de propagación/difusión**.
- Más adelante veremos un caso particular de este tipo de procesos con especial relevancia en Data Science: **Algoritmo Page-Rank**.



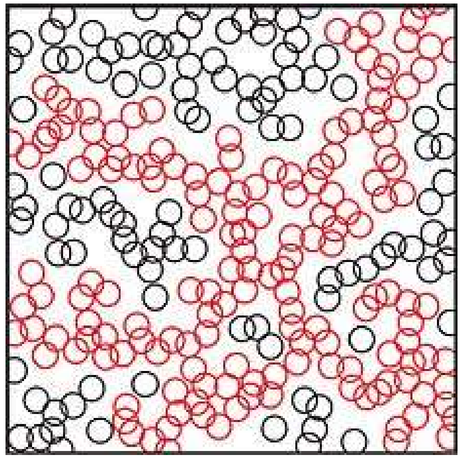
Variación de Características Internas (aristas)

- Un ejemplo de interés para Data Science: **Entrenamiento de Redes Neuronales**
  - Los pesos sinápticos se ajustan dinámicamente hasta conseguir que la red "aproxime" adecuadamente la función objetivo.



Procesos de Percolación

- Describe el comportamiento de clusters conectados en una red aleatoria.
- Idea: Supongamos que se vierte líquido sobre un material poroso. ¿Será el líquido capaz de ir de cavidad en cavidad hasta llegar al fondo?



- Si fijamos la probabilidad, p, de que dos cavidades (nodos) estén conectadas, se puede expresar la probabilidad de que haya percolación en función de p. Y hay un p_c, **crítico**, que presenta un cambio de fase de tipo 0/1.
- Fuertemente relacionado con muchos procesos: físicos (modelo de Ising), químicos (formación de superestructuras), y sociales (opinión, transmisión de rumores,…).

Page Rank

- Algoritmo que aproxima una determinada dinámica en las características internas de los nodos de una red.
- Famoso porque dio origen a **Google**, optimizado por **Larry Page** y **Sergey Brin** en la Universidad de Stanford.

- En 1997 ofrecieron el algoritmo a **Yahoo** por 1 millón US$, Yahoo declinó la oferta. En 2002, Yahoo ofreció 3.000 millones US$ por él, Google la rechazó.

Page Rank

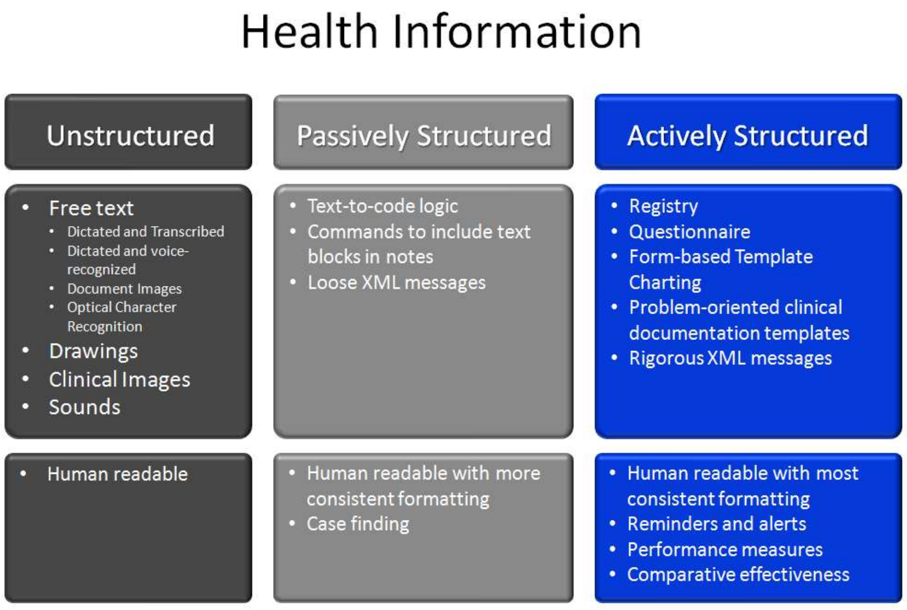
- Quizás la primera aplicación práctica conocida del algoritmo se debe al economista: **Wassily Leontief** (premio Nobel en 1971), de la Universidad de Harvard. Lo utilizó para representar el funcionamiento de una economía mediante un modelo de red.
- Hoy en día ya no se usa este algoritmo (al menos, no la versión inicial que veremos).

Page Rank

“La importancia de una página web es un problema inherentemente subjetivo que depende del interés de los lectores, de su conocimiento y de sus inclinaciones. Aun así, se puede decir objetivamente mucho sobre la importancia relativa de las páginas web. Este artículo describe PageRank, un método para valorar las páginas web de forma objetiva y mecánica, midiendo de forma efectiva la atención e interés humanos dirigidos hacia cada página. Comparamos PageRank con un "web surfer” aleatorio idealizado. Mostramos cómo calcular de forma eficiente el PageRank para un número grande de páginas y mostramos cómo utilizar el PageRank para la búsqueda y navegación de los usuarios.”

Page Rank

- Tras la revolución del acceso a la información (internet) viene la de ser capaz de poner orden en ese mar de información:
  - **Información Estructurada** vs **Información No Estructurada**
- Problema de Ingeniería Matemática:
  - Modelado Matemático
  - Resultados Matemáticos para su resolución efectiva
  - Técnicas Computacionales para su resolución eficiente



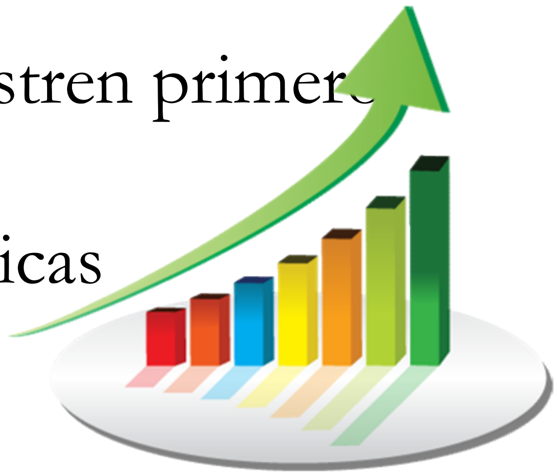
Page Rank

Supongamos que, como resultado de una consulta en un buscador hemos determinado que un conjunto de páginas contienen información que, de una manera u otra, puede resultar relevante para el usuario.

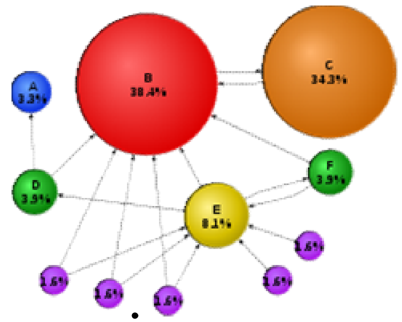
¿En qué orden mostramos esos resultados para que al usuario le resulte más sencillo revisarlos?

Lo lógico sería usar un **ranking** de forma que se muestren primero las páginas con más probabilidad de ser las adecuadas.

Para resolver este problema solo requerimos Matemáticas de Licenciatura **“básicas”**.

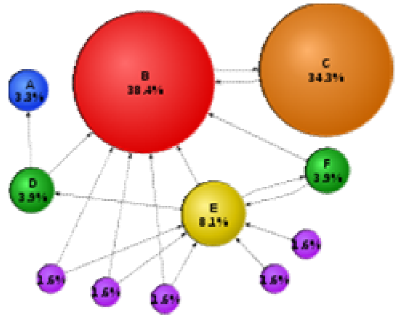


Page Rank



- Si denotamos las páginas por $P_1,\ldots,P_n$, nuestro objetivo es asignar a cada P_i una importancia numérica x_i.
- Como solo nos importa la importancia relativa entre ellas, podemos suponer que x_i está en [0,1].
- El modelo que usaremos para ver cómo calcular estas importancias hará uso únicamente de las etiqueta asignadas a las páginas y los enlaces entre ellas: **Red dirigida de páginas**.
- Notaremos por $M=(m_{ij})$ a la matriz de adyacencias asociada a la red anterior.

Page Rank



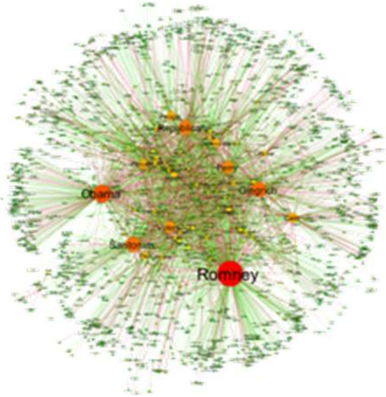
Vamos a postular que la importancia de una cierta página P_j "tiene que ver" con las páginas desde las que hay enlaces a ella:

Si muchas páginas enlazan con P_j será porque la información que ésta contiene ha sido considerada por muchos participantes de la red como "recomendable".

Page Rank

1er intento: x_j es proporcional al grado (entrante, saliente, o total) del nodo P_j . (Se puede obtener de la suma de la fila y/o columna)

- No recoge adecuadamente el hecho de que una página que es citada desde pocas, pero importantes, páginas, debe ser también importante.
- Necesitamos que el modelo asigne importancia alta a páginas que sean citadas desde páginas importantes… lo que nos lleva a una definición recursiva.



Page Rank

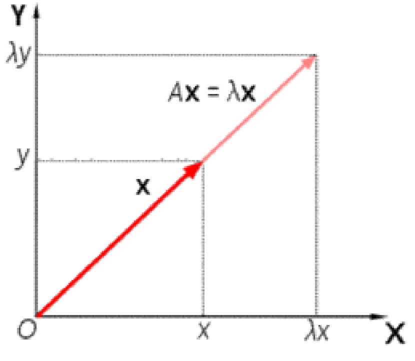
2º intento: la importancia x_j de cada página P_j es proporcional a la suma de las importancias de las páginas que enlazan con P_j .

- Nos lleva a un enorme sistema de ecuaciones:

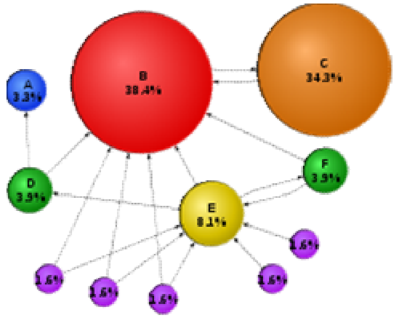
$$Mx = \lambda\, x$$

Que expresa la importancia de cada página a partir de las importancias de las demás páginas que están conectadas a ella.

- El problema se ha transformado en un problema de **autovalores** y **autovectores**: x es un **autovector** de M.



Page Rank



- Problemas:
  - Puede haber muchos autovectores (impondremos que todas sus coordenadas sean positivas, y podríamos ver en qué condiciones hay un único autovector).
  - La matriz es enorme (millones de páginas), así que los métodos habituales (exactos) para calcular autovectores pueden ser completamente inadecuados.
  - Propondremos una solución alternativa… aproximada, pero mucho más rápida.

Page Rank

- De hecho, se puede comprobar que, bajo determinadas condiciones, siempre se puede calcular este autovector, y que se corresponde con el autovalor de mayor tamaño.
- En este caso, sea cual sea el vector y considerado, se verifica que:

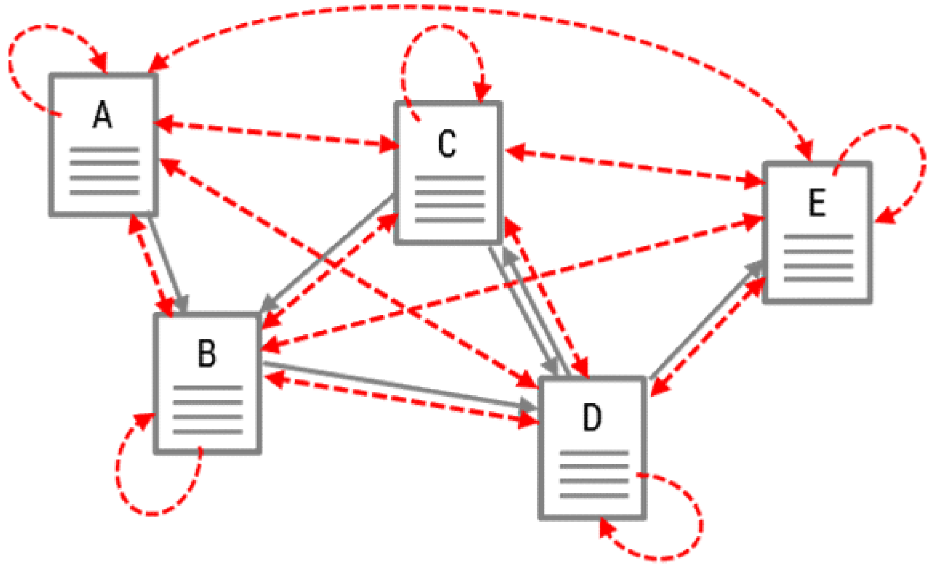
  $M^n\, y \to K\, v$, (cuando $n \to \infty$) donde v es el autovector

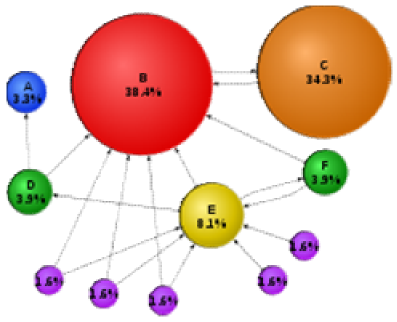
- Hemos "reducido" el problema de calcular las importancias (el autovector) en el problema de calcular las potencias de la matriz de conexiones, M… ¿Es una reducción?

Page Rank

El **surfista Aleatorio**:

- Supongamos que un surfista navega por la red.
- En un cierto instante está en la página *P*. Al instante siguiente salta a una de las páginas conectadas con *P*, seleccionada al azar y con probabilidad uniforme.
- Repite este proceso siguiendo lo que se denomina un **paseo aleatorio**.



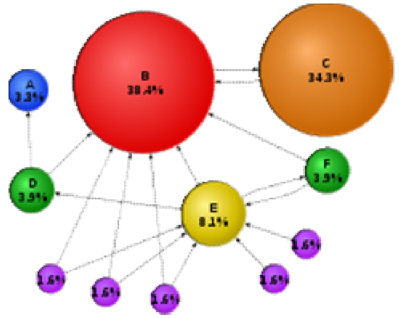


Page Rank

El **surfista Aleatorio**:

- ¿Qué probabilidad hay de que partiendo de P_i esté en P_j tras k pasos?
- Si notamos N_j el número de enlaces en P_j (la suma de la columna).
- Denotamos por M' la nueva matriz (m_{ij} / N_j). Ahora, la suma de cada columna de M' es 1.
- M' es **estocástica**.

$$\mathbf{M} = \begin{array}{c} \text{página } P_j \\ \downarrow \\ \left(\begin{array}{ccc} & m_{1j} & \\ & \vdots & \\ & m_{ij} & \\ & \vdots & \\ & m_{nj} & \end{array} \right) \\ \downarrow \\ N_j \end{array} \qquad \mathbf{M}' = \begin{array}{c} \text{página } P_j \\ \downarrow \\ \left(\begin{array}{ccc} & \frac{m_{1j}}{N_j} & \\ & \vdots & \\ & \frac{m_{ij}}{N_j} & \\ & \vdots & \\ & \frac{m_{nj}}{N_j} & \end{array} \right) \\ \downarrow \\ 1 \end{array}$$



Page Rank

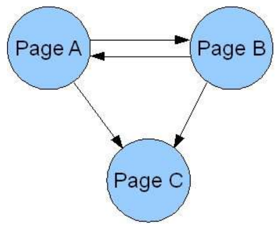
El **surfista Aleatorio**:

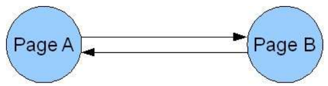
- En consecuencia, calcular el elemento m_{ij} de la matriz potencia se puede aproximar por mover aleatoriamente un surfista desde la página P_i y ver con qué probabilidad alcanza la página P_j.
- Este método tiene algunas ventajas, y es que es robusto a cambios en las conexiones, porque la mayoría de las probabilidades se verán poco afectadas si los cambios son locales.

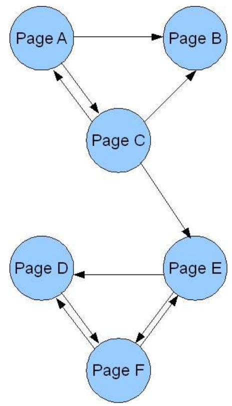
Page Rank

El **surfista Aleatorio**… limitaciones:

- – Sumideros (páginas sin salida).
  - Individuales
  - Hordas
- – Referencias circulares







- Soluciones:
  - – Reemplazar filas de 0's por 1's (en realidad, por 1/n)
  - – Saltos aleatorios del surfista de vez en cuando (sin considerar los enlaces)
- Veamos el modelo…