Random Walks 2

Sarunas Girdzijauskas

ID2211

March 2019

Random Walk

$$\mathbf{P_t = P_0 M^t}$$

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

$$\begin{pmatrix}0 & 1/2 & 1/2 & 0 & 0\end{pmatrix}\begin{pmatrix}0 & 1/2 & 1/2 & 0 & 0\\ 1/2 & 0 & 0 & 1/2 & 0\\ 1/3 & 0 & 0 & 1/3 & 1/3\\ 0 & 1/3 & 1/3 & 0 & 1/3\\ 0 & 0 & 1/2 & 1/2 & 0\end{pmatrix}=\begin{pmatrix}0.42 & 0 & 0 & 0.42 & 0.17\end{pmatrix}$$

- When $P_{t+1} = P_t = \pi$, we have reached stationary distribution, i.e. **$\pi M = \pi$**
- Recall: that v is **eigenvector** of matrix M and λ its eigenvalue if **$vM=\lambda v$**
  - so ***π is eigenvector of M with eigenvalue λ=1***

More intuition on Spectra of matrix A

- Example: d-regular graph (connected & unweighted for now…)
- Recap: What is the meaning of *Ax* ?

Consider **x** as a vector repersenting values for each node in the graph

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$y_j = \sum_{i=1}^{n} A_{ij} x_i = \sum_{(j,i)\in E} x_i$$

Entry yi is a sum of labels xj of neighbors of i

- And what is an eigenvector of *A* ?

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$\boldsymbol{A \cdot x = \lambda \cdot x}$$

Set of eigenvalues and eigenvectors

Graph Spectra

- **$vA=\lambda v$** ($Av= \lambda v$ for column vector)
- If A is a real symmetric matrix then it has n eigenvectors and associated n eigenvalues. All n eigenvalues are real $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{n}$, eigenvectors orthogonal
  - If G is a d-regular graph then $\lambda_1=d$
- For random walk matrix M, i.e., normalized adjacency matrix $\lambda_1=1$ **($\pi M= \pi$)**

What about weighted graphs?

Graph Laplacian

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

Adjacency Matrix

$$L = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 2 & 0 & -1 & 0 \\ -1 & 0 & 3 & -1 & -1 \\ 0 & -1 & -1 & 3 & -1 \\ 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$

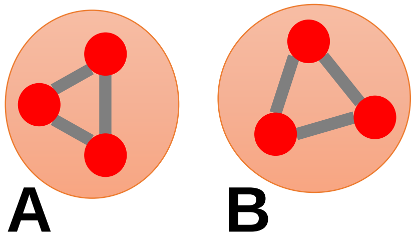
Laplacian Matrix

- **Laplacian Matrix L=D-A** where **D** is a degree matrix
  - $\lambda_1 = 0$
- If $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ eigenvalues of L then:
  - G has k connected components if $\lambda_k = 0$
- **(optional) Check ID2222 lectures on Canvas to understand why.**
  - **Normalized Laplacian =** $D^{-½} A\ D^{-½}$
    - **Sometimes also known as** $I - D^{-½} A\ D^{-½}$

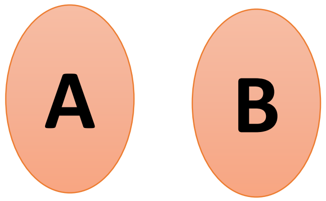
Graph Spectra (cont.)

- We call λ_1, λ_2,…, λ_n the spectra of graph G
  - from matrices A, M or L
- We call λ_1-λ_2 eigengap (or **spectral gap**)
  - 1-λ_2 for M
- So what if the graph is disconnected?
  - Think of convergence of Random Walk on M…
  - $\lambda_1=\lambda_2$

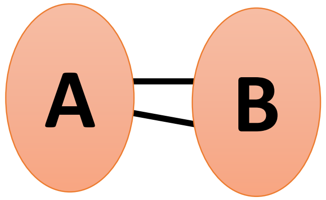
Intuition: Disconnected Graph



- **What if G is not connected?**
  - G has **2** components, each d-regular
- **What are some eigenvectors?**
  - $x =$ Put all 1s on A and 0s on B or vice versa
    - $x' = (\underbrace{1, \dots, 1}_{|A|}, 0, \dots, 0)$ **then** $\mathbf{A} \cdot x' = (d, \dots, d, 0, \dots, 0)$
    - $x'' = (0, \dots, 0, \overbrace{1, \dots, 1}^{|B|})$ **then** $A \cdot x'' = (0, \dots, 0, d, \dots, d)$
    - And so in both cases the corresponding $\lambda = d$
- **A bit of intuition:**



$\lambda_1 = \lambda_2$



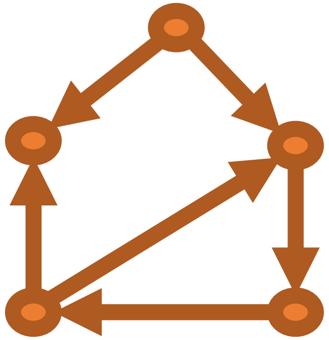
$\lambda_1 - \lambda_2 \approx 0$

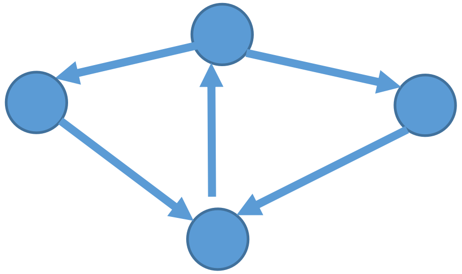
Convergence time on Expanders

- There is a connection between expansion of the graph and the spectral gap
  - Large gap (that is small λ_2) implies good expansion and vice versa.
- If G is a connected, d-regular, non-bipartite graph on n vertices, then λ_2 of M is < 1 and G has mixing time $O(\frac{\log N}{1-\lambda_2})$
  - Practically mixing time is ~ O(logN)

Convergence of Random Walk

- For every ***connected non-bipartite*** undirected graph G, the distribution Pi converges to a limit and unique stationary distribution **π**.
- Moreover, if G is regular then this distribution is the uniform distribution on V.
- Intuition:
  - Why connected?
  - Why non-bipartite?
- What about Directed graphs?
  - Has to be ***strongly connected***! (otherwise the "walks will leak").
  - Has to be ***aperiodic,*** i.e., visits to some state (node) S should never be a multiple of k (k>1)
    - if the greatest common divisor of the lengths of its cycles is one

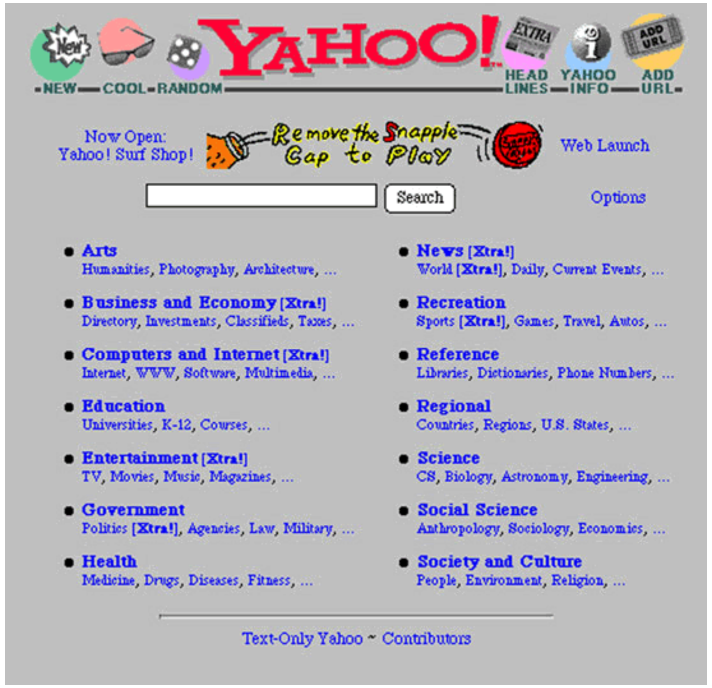




Relation to the web Search?

- **1st generation (Directories):**
  - Manual curation of **web directories** (e.g., Yahoo)
  - Web was growing too quickly to catch up.

- **2nd generation (Information Retrieval)**
  - – Altavista
  - – Classical Invormation retrieval, processing text in the pages
  - – Term Spam
- **3rd generation (Google)**
  - – Google page rank
    - Very hard to fake in-links.



- **Arts**
  Humanities, Photography, Architecture, ...
- **Business and Economy** [Xtra!]
  Directory, Investments, Classifieds, Taxes, ...
- **Computers and Internet** [Xtra!]
  Internet, WWW, Software, Multimedia, ...
- **Education**
  Universities, K-12, Courses, ...
- **Entertainment** [Xtra!]
  TV, Movies, Music, Magazines, ...
- **Government**
  Politics [**Xtra!**], Agencies, Law, Military, ...
- **Health**
  Medicine, Drugs, Diseases, Fitness, ...
- **News** [Xtra!]
  World [**Xtra!**], Daily, Current Events, ...
- **Recreation**
  Sports [**Xtra!**], Games, Travel, Autos, ...
- **Reference**
  Libraries, Dictionaries, Phone Numbers, ...
- **Regional**
  Countries, Regions, U.S. States, ...
- **Science**
  CS, Biology, Astronomy, Engineering, ...
- **Social Science**
  Anthropology, Sociology, Economics, ...
- **Society and Culture**
  People, Environment, Religion, ...

Text-Only Yahoo ~ Contributors

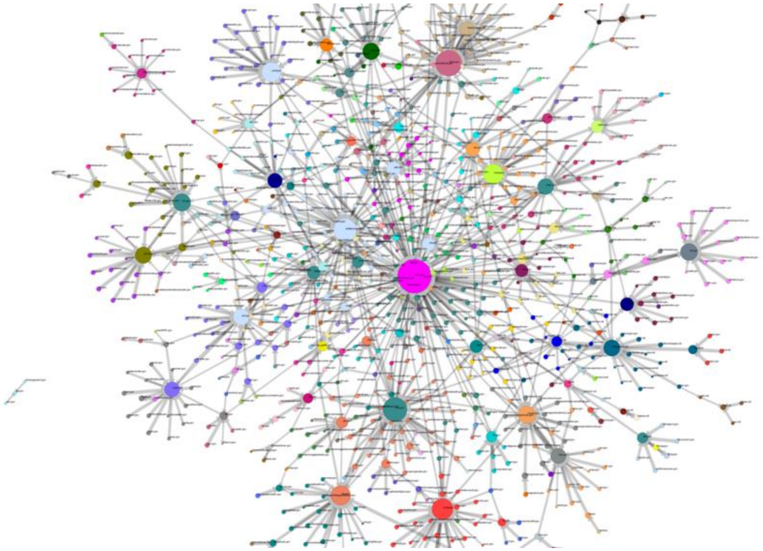
PageRank ”Voting” formulation

- Each page has a budget of ”votes” and distributes them evenly to all the outgoing links
  - E.g., if page j has r_j budget of votes, and n out-links, each link gets r_j/n votes.
  - INSIGHT: **A vote from an important page is worth more.**
- Node’s j own importance is the **sum of the votes on its in-links**.
- Did we see this before??
  - Notice the similarity with the **random walk**

Kahoot?

Web Graph (cont.)

- **Web as a graph:**
  - Nodes = web pages
  - Edges = hyperlinks
- Directed or undirected?
- Connected or Disconnected?
- Cyclic or Acyclic?
- **Seminal paper (test of time award)**
  - A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener. *Graph structure in the Web*. Computer Networks, 33, 2000.
    - Web crawl is based on a large set of starting points accumulated over time from various sources, including voluntary submissions.
    - 203 million URLS and 1.5 billion links
  - **Computer:** Server with 12GB of memory

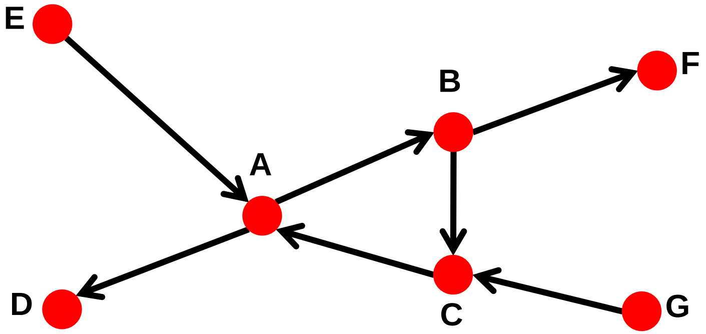


What Does the Web Look Like?

- **How is the Web linked?**
- **What is the "map" of the Web?**

Web as a directed graph [Broder et al. 2000]:

- Given node ***v***, what can ***v*** reach?
- What other nodes can reach ***v***?



$In(v) = \{w \mid w \text{ can reach } v\}$

$Out(v) = \{w \mid v \text{ can reach } w\}$

For example:

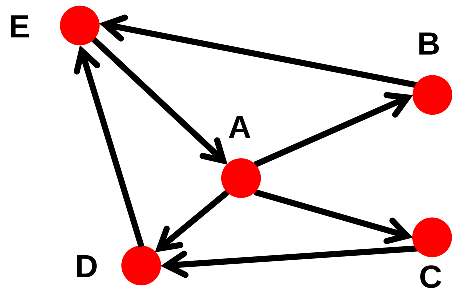
In(A) = {A,B,C,E,G}

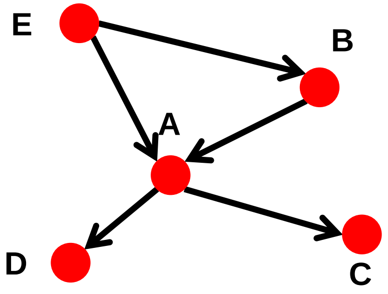
Out(A)={A,B,C,D,F}

Reasoning about Directed Graphs

- **Two types of directed graphs:**
  - **Strongly connected:**
    - Any node can reach any node via a directed path

    In(A)=Out(A)={A,B,C,D,E}
  - **Directed Acyclic Graph (DAG):**
    - Has no cycles: if ***u*** can reach ***v***, then ***v*** cannot reach ***u***



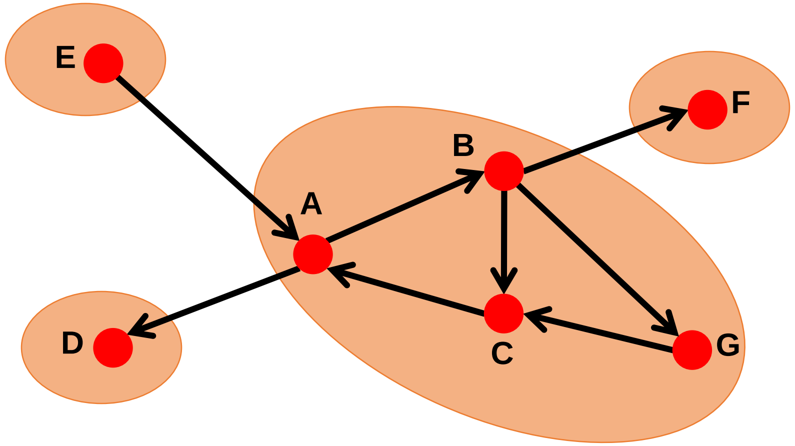


- **Any directed graph (the Web) can be expressed in terms of these two types!**
  - Is the Web a big strongly connected graph or a DAG?

Strongly Connected Component

- **A Strongly Connected Component (SCC)** is a set of nodes $\boldsymbol{S}$ so that:
  - Every pair of nodes in $\boldsymbol{S}$ can reach each other
  - There is no larger set containing $\boldsymbol{S}$ with this property

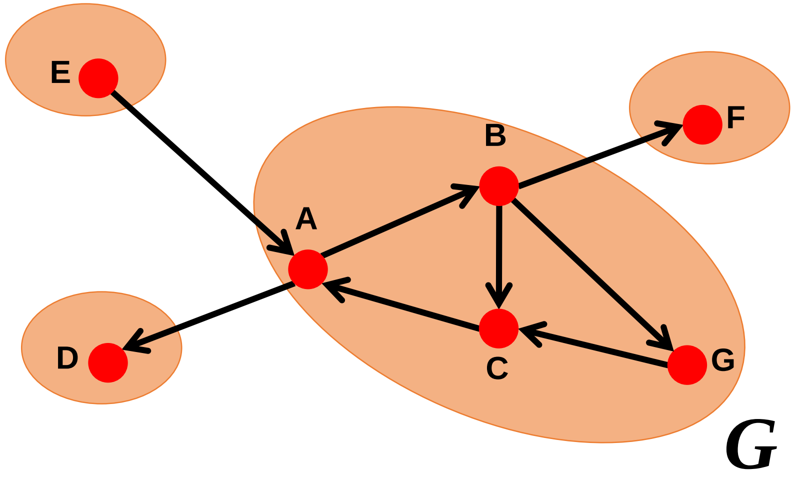
Can you identify a strongly connected component?



Strongly connected components of the graph: {A,B,C,G}, {D}, {E}, {F}

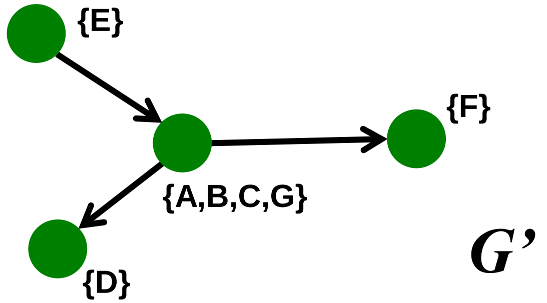
Strongly Connected Component

- **<u>Fact:</u> Every directed graph is a DAG on its SCCs**
  - **(1)** SCCs partitions the nodes of ***G***
    - That is, each node is in exactly one SCC
  - **(2)** If we build a graph ***G'*** whose nodes are SCCs, and with an edge between nodes of ***G'*** if there is an edge between corresponding SCCs in ***G***, then ***G'*** is a DAG



(1) Strongly connected components of graph G: {A,B,C,G}, {D}, {E}, {F}

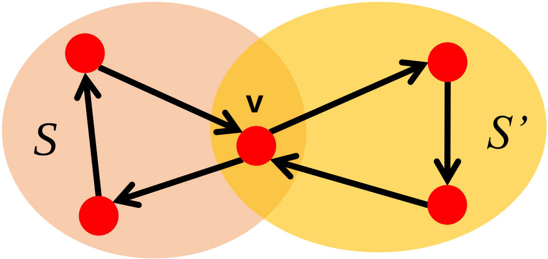
(2) G' is a DAG:



Proof of (1)

- **Claim: SCCs partition nodes of G.**
  - This means: Each node is member of exactly 1 SCC

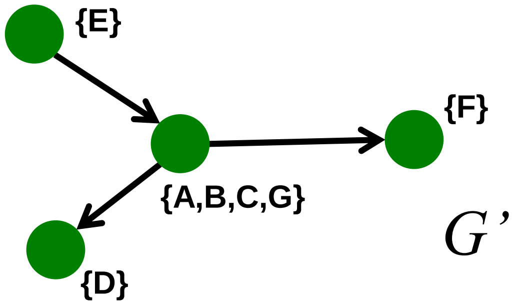
- Proof by contradiction:
  - Suppose there exists a node $\boldsymbol{v}$ which is a member of two SCCs $\boldsymbol{S}$ and $\boldsymbol{S'}$

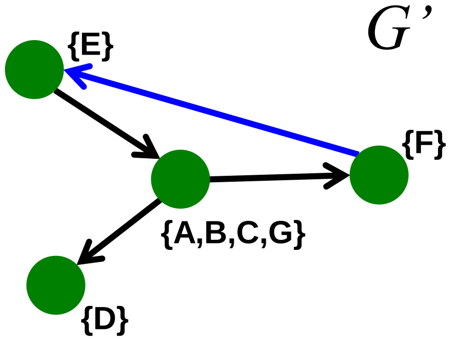


  - But then $\boldsymbol{S} \cup \boldsymbol{S'}$ is one large SCC!
    - **Contradiction:** By definition SCC is a maximal set with the SCC property, so $\boldsymbol{S}$ and $\boldsymbol{S'}$ were not two SCCs.

Proof of (2)

- **Claim: *G'* (graph of SCCs) is a DAG.**
  - This means: ***G'*** has no cycles
- Proof by contradiction:
  - Assume ***G'*** is <u>not</u> a DAG
  - Then ***G'*** has a directed cycle
  - Now all nodes on the cycle are mutually reachable, and all are part of the same SCC
  - But then ***G'*** is not a graph of connections between SCCs (SCCs are defined as maximal sets)
    - **Contradiction!**





Now {A,B,C,G,E,F} is a SCC!

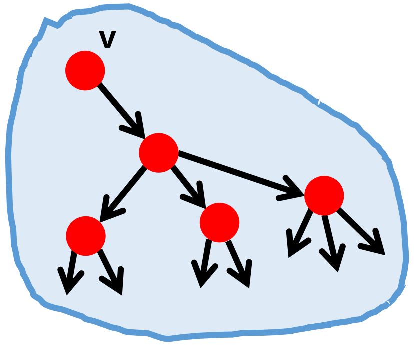
Back to…

How is the Web linked?

<u>Goal:</u> Take a large snapshot of the Web and try to understand how its SCCs “fit together” as a DAG

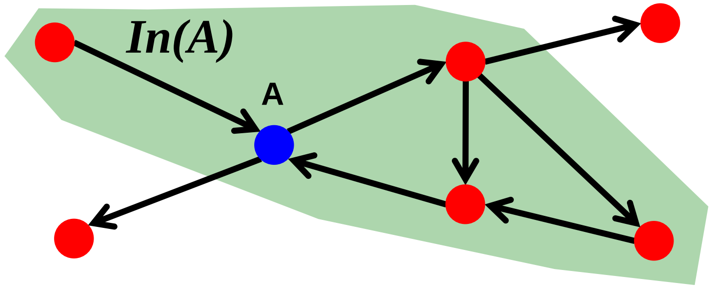
Graph Structure of the Web

- **Computational issue:**
  - Want to find a SCC containing node ***v***?



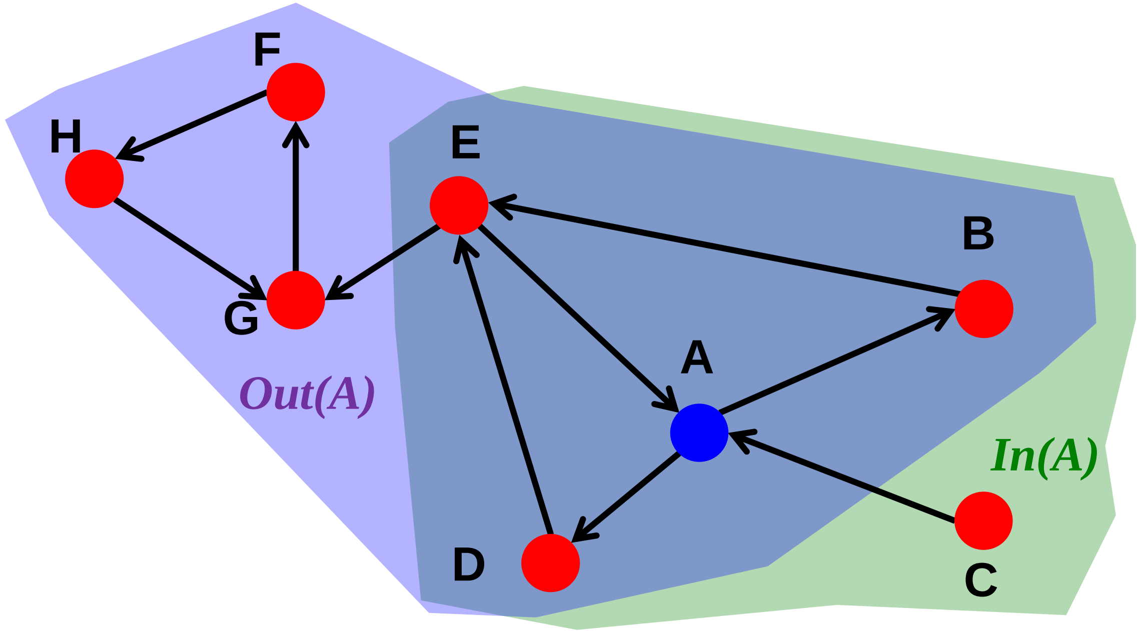
Out(v)

- **Observation:**
  - ***Out(v)*** … nodes that can be reached from *v*
  - **SCC containing *v* is:** *Out(v)* ∩ *In(v)*
    = *Out(v,G)* ∩ *Out(v,G')*, where *G'* is *G* with all edge directions flipped



Out(A) ∩ In(A) = SCC

- **Example:**

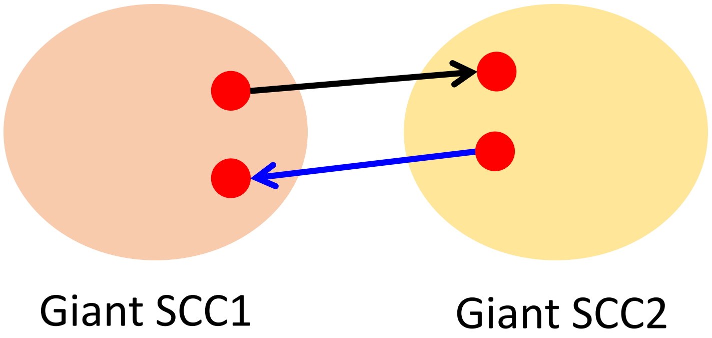


- Out(A) = {A, B, D, E, F, G, H}
- In(A) = {A, B, C, D, E}
- So, SCC(A) = Out(A) ∩ In(A) = {A, B, D, E}

Graph Structure of the Web

- **How many giant SCCs?**

- **Why only 1 big SCC? Heuristic argument:**
  - Assume two equally big SCCs.
  - It just takes 1 page from one SCC to link to the other SCC.
  - If the two SCCs have millions of pages the likelihood of this not happening is very very small.

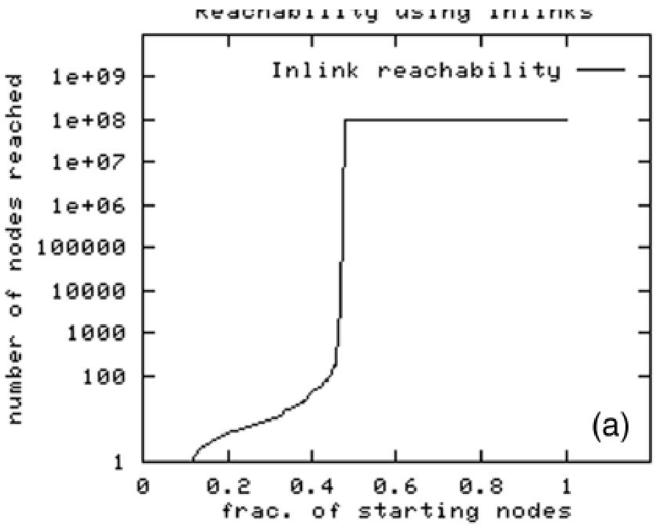


Structure of the Web

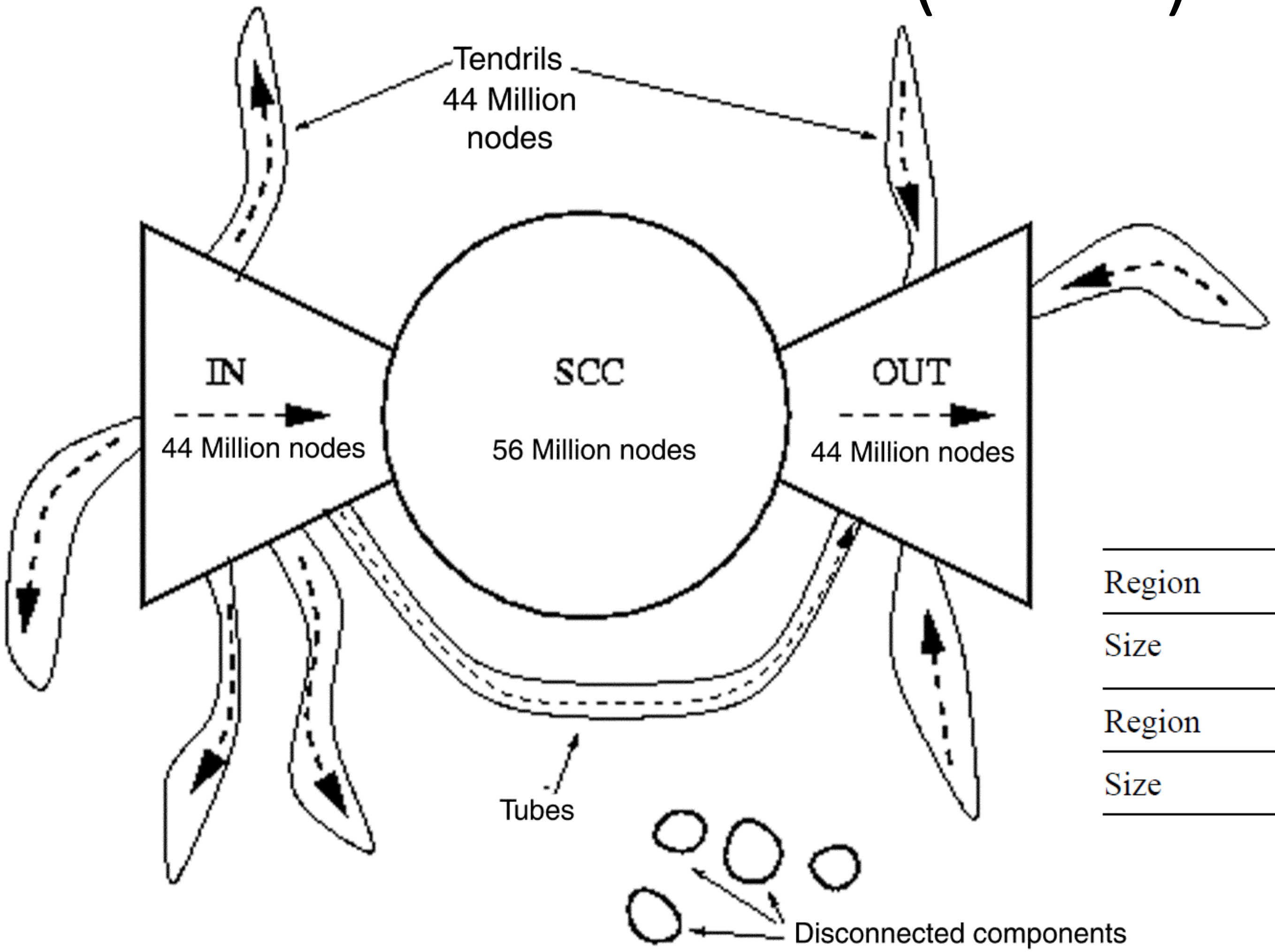
- **Directed version of the Web graph:**
  - Altavista crawl from October 1999
    - 203 million URLs, 1.5 billion links

Computation:

- Compute IN(v) and OUT(v) by starting at random nodes.
- **Observation:** The BFS either visits many nodes or gets quickly stuck.



Structure of the Web (cont.)



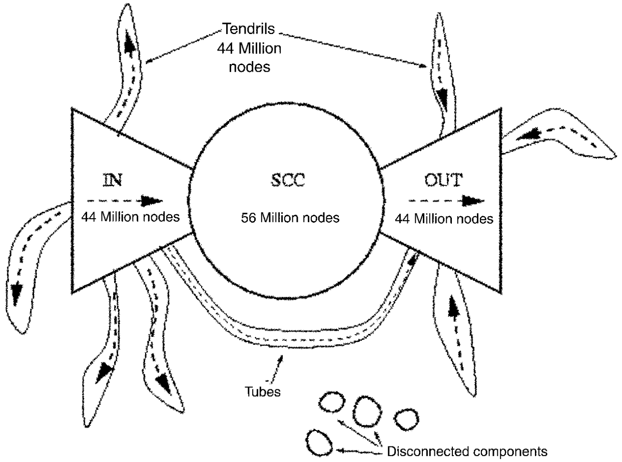
Region	SCC	IN	OUT
Size	56,463,993	43,343,168	43,166,185
Region	TENDRILS	DISC.	Total
Size	43,797,944	16,777,756	203,549,046

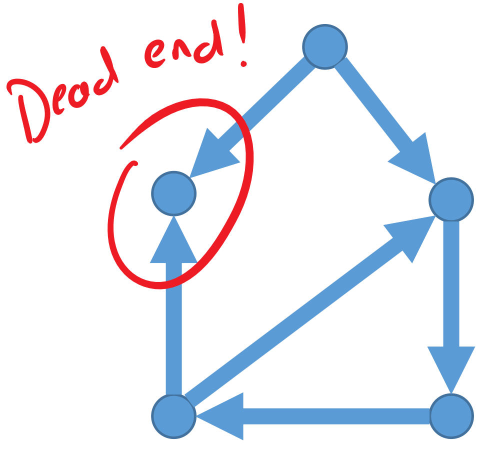
Fig. 9. Connectivity of the Web: one can pass from any node of IN through SCC to any node of OUT. Hanging off IN and OUT are TENDRILS containing nodes that are reachable from portions of IN, or that can reach portions of OUT, without passage through SCC. It is possible for a TENDRIL hanging off from IN to be hooked into a TENDRIL leading into OUT, forming a TUBE: i.e., a passage from a portion of IN to a portion of OUT without touching SCC.

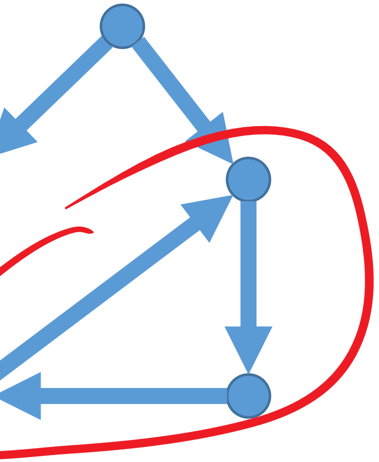
Google Page Rank

Does this Graph has unique stationary distribution of the random walk?

- Google page rank: **principal eigen vector** on the transition matrix of the web graph!
  - How to compute?
    - Power iteration.
  - Any issues?
    - Undirectional vs directional graph?
      - **Dead ends**
        - Nodes with no out-degree
        - votes leak out
      - **Spider traps**
- WWW is not strongly connected!

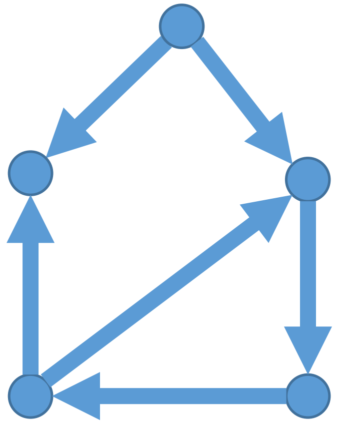


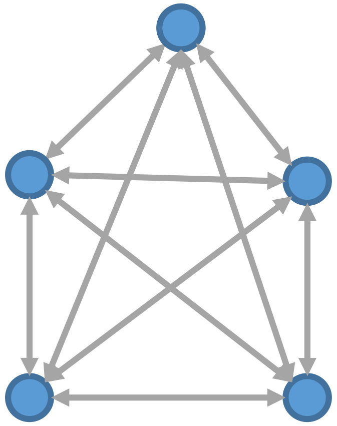




How do we fix PageRank?

- Ideas?
  - Make the graph **strongly connected** and **aperiodic**
  - Google solution
    - Make ”**tiny tiny**” links from each node to every other node
    - Keep **the core** of the initial graph

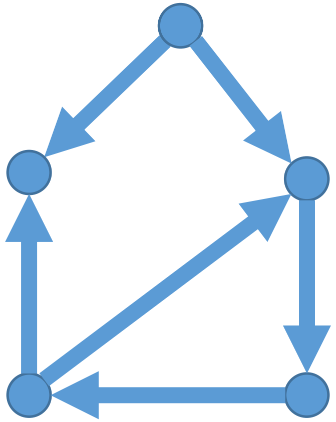


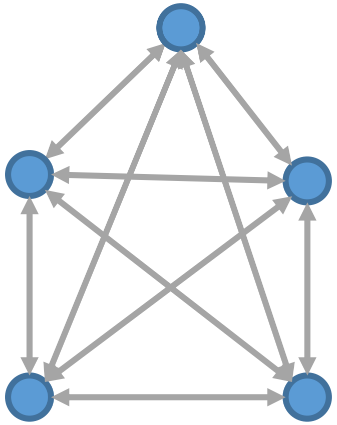


How do we fix PageRank? (cont.)

- Google random walker will
  - With **prob β** follow ”the real” link at random.
  - With **prob 1- β** jump to some random page.
  - Usually β is in the range of 0,8 to 0,9
    - I.e.,Radom walker will ”teleport” from any spider trap after 5-10 steps
  - For **”dead ends” 1-β=1**, i.e., once in the dead end, random walker always teleports to a random node.

Radom Walk Matrix with teleportation



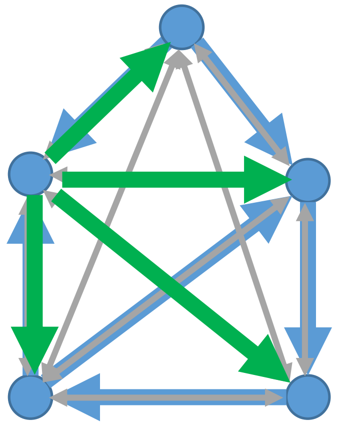


$$M=\begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \beta \quad + \quad T=\begin{pmatrix} 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 0 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 0 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 0 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 \end{pmatrix} (1-\beta)$$

Transition (random walk) Matrix

Teleportation Matrix

Radom Walk Matrix with teleportation



$$M_{\text{pageRank}} = \begin{pmatrix} 0 & 0.45 & 0.45 & 0.05 & 0.05 \\ 0.25 & 0 & 0.25 & 0.25 & 0.25 \\ 0.05 & 0.05 & 0 & 0.05 & 0.85 \\ 0.05 & 0.45 & 0.45 & 0 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.85 & 0 \end{pmatrix}$$

$$M_{\text{pageRank}} = \beta \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} + \begin{pmatrix} 1-\beta & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1-\beta & 0 & 0 \\ 0 & 0 & 0 & 1-\beta & 0 \\ 0 & 0 & 0 & 0 & 1-\beta \end{pmatrix} \begin{pmatrix} 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 0 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 0 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 0 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 \end{pmatrix}$$

Is it fixed now?

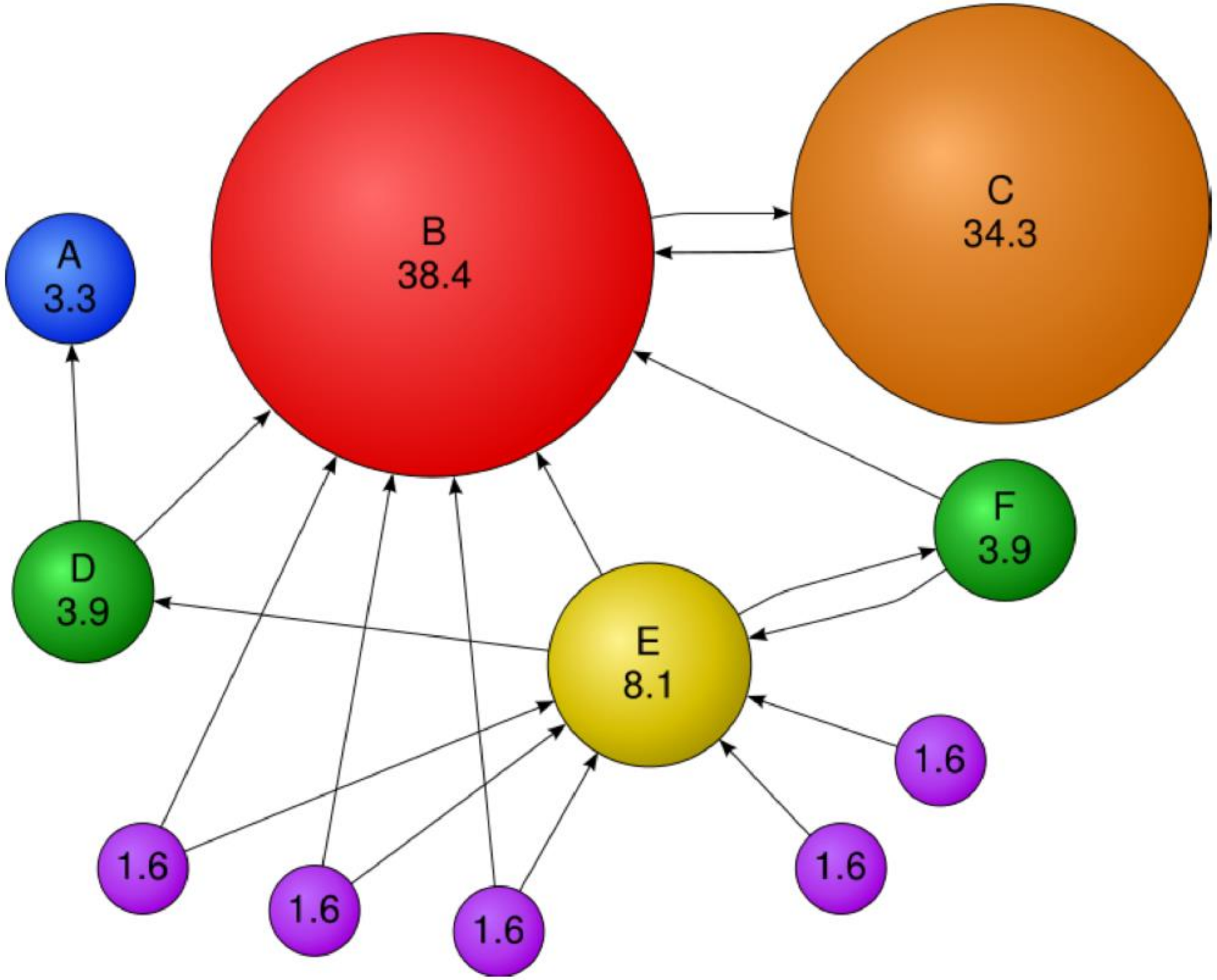
- What if **β=0** for all the nodes?
- What are the problems with teleportation?
  - How does $M_{pageRank}$ look for 10bn pages?
    - Dense random walk matrix!
    - N^2 non-zero elements! (instead of O(N*d))
  - Can you even store it in memory?
  - Insight for the Fix:
    - Interpret teleportation as **fixed tax** (always the same),
    - At every power iteration instead of computing rank vector $\mathbf{r}^{new}$

      $\mathbf{r}^{new}$=$\mathbf{r}^{old}$ $M_{pageRank}$, we compute $\mathbf{r}^{new}$= β ($\mathbf{r}^{old}$*M*)+c,

      where c = (1- β)/N (i.e., a tax)
    - Notice $M_{pageRank}$ is dense and M is sparse matrix!
    - If M contains dead-ends then r^{new} has to be made stochastic again i.e., so that it sums up to 1.

Example

Node size proportional to the PageRank score

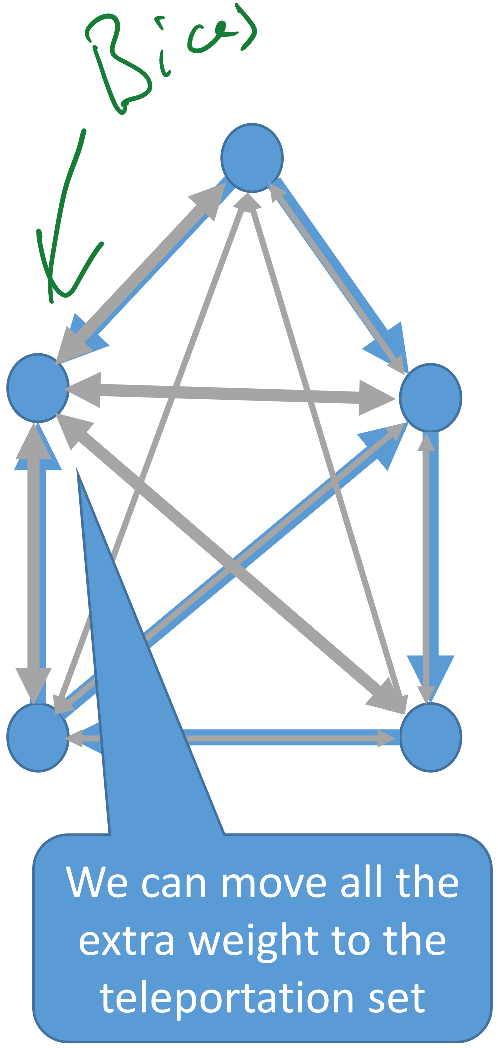


Problems with PageRank?

- **Measures ”generic popularity” of a page**
  - Might miss topic-specific authorities
  - Will solve it by **Topic-Specific** PageRank
- **Susecptible to Link spam**
  - Artificial link topologies created to boost page rank
  - We will solve it by **TrustRank (ID2222)**
- **Uses a single measure of importance**
  - We will address it by **Hubs-and-Authorities (ID2222)**

Topic Specific Page Rank

- Insight: **Bias the random walk** towards ”relevant set nodes”
  - Give **more influence** to the webpages that are close to a particular topic (given by a query), e.g., ”sports”, ”travel” etc
- Instead of teleporting to ”any node” - teleport to ”relevant pages” (**teleport set**) for topic-specific PageRank
  - Could also assign different weights to pages within the teleport set
- Once we have a biased (green) weights we recalculate PageRank in a regular fashion



We can move all the extra weight to the teleportation set

Topic Specific Page Rank (cont)

- One can precalculate PageRanks for each page for different topics
  - E.g., arts, business, sports etc.
- Where from to get the **teleport set?**
- Which topic ranking to use?
  - User can pick from a menu
  - Classify query into a topic
  - Exploit the **”context”** of the query:
    - Where the query is launched from (a topic specific webpage?)
    - User browsing history (e.g., could be a difference if one queries “Manchester” after “football” or after “travel” queries)
    - User bookmarks, cookies etc.

Take the sets from the ”previous generation” serach engines