HITS (Hyperlink Induced Topic Search) Centrality
What have we done to this point?

- Identified vertices that are important because of their number of connections (degree centrality), and
- vertices important because they are adjacent to important vertices (eigenvector centrality), and
- generalized these ideas to directed networks (Katz centrality), and
- identified a mechanism to control the “flow of centrality” from high centrality vertices in an appropriate manner (PageRank centrality).

- What we’ve not done is classify the important vertices by what makes them important.
Authority and Hub Centralities

- Authority centrality: $i$’s rank score $x_i$ is proportional to the sum of the hub centralities that point to it:
  \[ x_i = \alpha \sum_j A_{ij} y_j, \]
  where $\alpha$ is a constant and $y_j$ are hub centralities of those vertices pointing to $x_i$.

- Hub centrality: $i$’s rank score $y_i$ is proportional to the sum of the authority centralities it points to:
  \[ y_i = \beta \sum_j A_{ij} x_j, \]
  where $\beta$ is another constant and $x_j$ are authority centralities of those vertices pointing to $y_i$. 
Matrix notation (1)

We re-write authority and hub centrality in matrix-vector form:

\[ x = \alpha A y, \quad y = \beta A^T x, \] and combine the two:

\[ AA^T x = \lambda x, A^T A y = \lambda y, \] where \( \lambda = (\alpha \beta)^{-1}. \)

Authority and hub centralities given by leading eigenvectors of \( AA^T \) and \( A^T A \), respectively. This works because all eigenvalues are the same for both matrices.

Now, how do we solve?
Matrix notation (2)

Since $AA^T x = \lambda x$, multiply both sides by $A^T$:

$$A^T A A^T x = A^T \lambda x = \lambda A^T x$$

And we note that $A^T x$ is an eigenvector of $A^T A$ with eigenvalue $\lambda$.

Recalling the equation for calculating hub centrality: $A^T A y = \lambda y$, we see calculating hub centralities is easy when we know authority centrality:

$$y = A^T x$$
## Overview

<table>
<thead>
<tr>
<th>Quality: what makes a node important (central)</th>
<th>Mathematical Description</th>
<th>Appropriate Usage</th>
<th>Identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lots of one-hop connections to high out-degree vertices</td>
<td>A weighted degree centrality based on the out degree of the neighbors</td>
<td>Directed graphs that are not strongly connected</td>
<td>Katz $C_i = \alpha \sum_j A_{ij} C_j + \beta$ Where $\beta$ is some small weight for each node</td>
</tr>
<tr>
<td>As above but distribute the weight that a node has to the nodes it points to</td>
<td>$\frac{C_j}{\text{out deg } j}$</td>
<td>As above but distributing the wealth of a node to the ones it points to</td>
<td>Page Rank: $C_i = \alpha \sum_j A_{ij} \frac{C_j}{\text{out deg } j} + \beta$ or $x = \alpha AD^{-1}x + \beta 1$</td>
</tr>
<tr>
<td>Authority: important if hubs point to it; Hub: important if it points to authorities</td>
<td>Authority: $x = \alpha Ay$, Hub: $y = \beta A^T x$ $\alpha$, $\beta$ constants</td>
<td>ID entities that have information, as in cocitation and bibliographic coupling</td>
<td>Hub Centrality: $y = A^T x$</td>
</tr>
</tbody>
</table>
Why is HITS not used as much as, say, PageRank?

PageRank
• Computed for all web pages stored prior to query
• Computes authorities
• Computationally fast

HITS
• Performed on a query-based subset
• Computes authorities and hubs
• Easy to compute, but hard to compute real time