PageRank (by Google founder Larry Page)

Google Rankings
Which webpages are “more important?”

Model of internet:
• Users click random link on a page.
• Occasionally start over.

A page is “more important” if viewed a large fraction of time

Random Walk on the Web
View the entire web as digraph
• vertices are webpages
• edge \((V,W)\) exists if link from page \(V\) to page \(W\)
• edges out of \(V\) equally likely
  \(\Pr[(V,W)] = \frac{1}{\text{outdeg}(V)}\)

Random Walk on the Web
To model starting over:
* add a “super-node” to the graph
* an edge from super-node to each other node
* edges from each other node back to super-node
  - may get customized probabilities
Compute stationary distribution \( \tilde{s} \)

PageRank(V) ::= \( s_V \)

Rank V above W when \( s_V > s_W \)

Resistance to scamming

* Creating fake nodes pointing to self
* Adding links to important nodes

won’t improve PageRank

Importance of Super-node ensures

* unique stable distribution \( \tilde{s} \)
* every initial distribution \( \tilde{p} \)
converges to \( \tilde{s} \)

\[ \lim_{\mathcal{t} \to \infty} \tilde{p} \cdot \mathbb{M}^\dagger = \tilde{s} \]

* convergence is rapid:
\( \mathcal{t} \) is small so \( \tilde{s} \) easy to compute
Actual Google Rank

Google rank rules are a closely held trade secret using text, location, payment, and other criteria that have evolved for 15 years. But PageRank continues to play a significant role.