Link analysis: HITS

Web search results: desired

- List of webpages / websites ranked according to
  - Relevance to query – we have already studied in detail
  - Importance / trustworthiness of websites - centrality
  - Location / time of query
  - Recency of page
  - … and many other factors

Node centrality

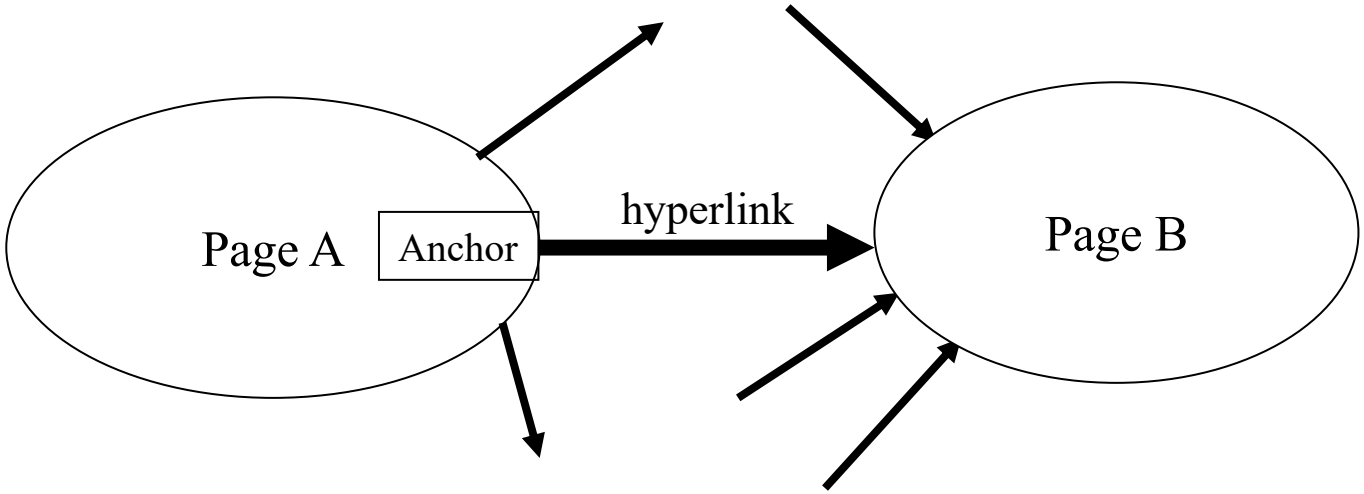
- Relative importance of a node in a network

- How influential a person is within a social network
- How important a webpage is in the Web

Node centrality in Web

- Web graph:
  - Nodes are webpages
  - Edges are hyperlinks (directed)

The Web as a Directed Graph



Hypothesis 1: A hyperlink between pages denotes a conferral of authority (quality signal)

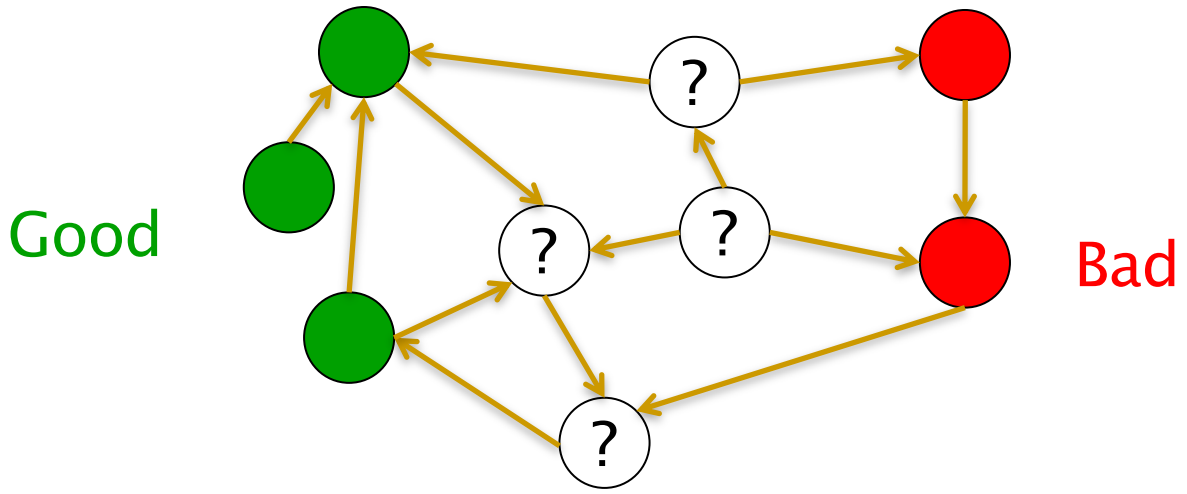
Hypothesis 2: The text in the anchor of a hyperlink on page A describes the target page B

Importance of node centrality in Web

- If only relevance used to rank webpages, ranking algorithm can be easily spammed
- Previously, indegree of webpages used to rank pages according to importance
- Easily gamed by spammers creating their own webpages

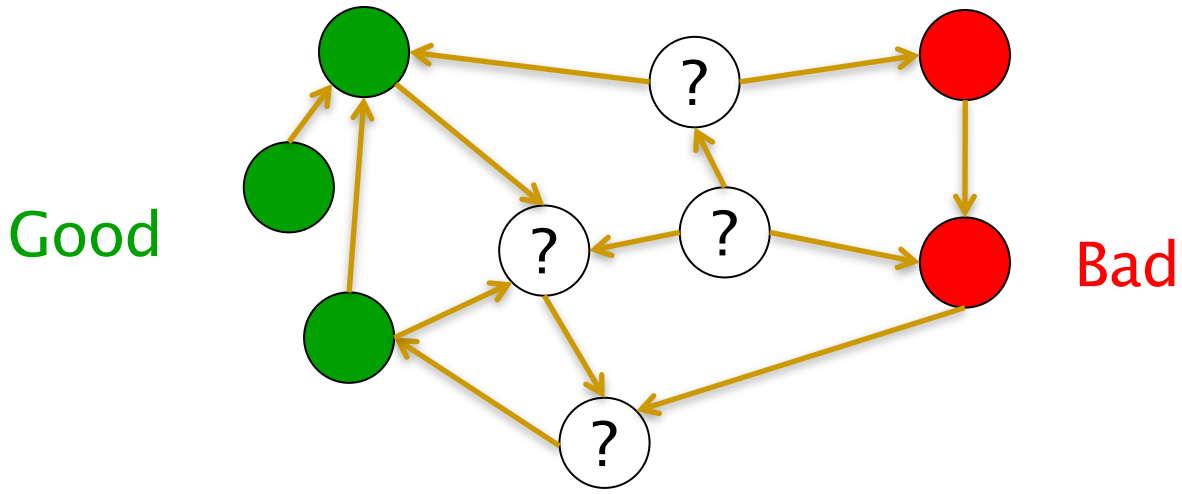
A better idea

- Nodes of three types: The Good, The Bad and The Unknown
  - Assumption: Good nodes won't point to Bad nodes
  - All other combinations plausible



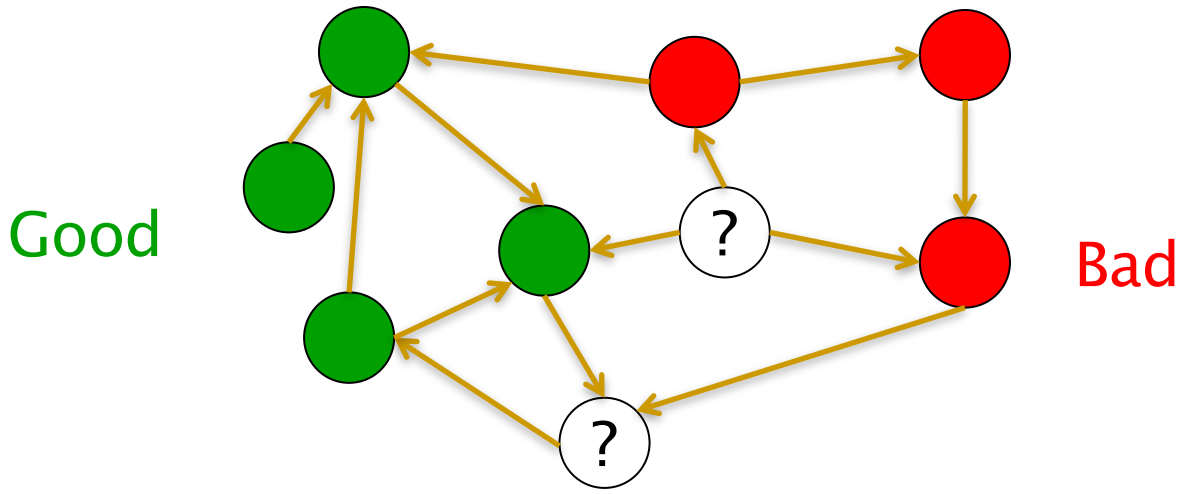
Simple iterative logic

- Good nodes won't point to Bad nodes
  - If you point to a Bad node, you're Bad
  - If a Good node points to you, you're Good



Simple iterative logic

- Good nodes won't point to Bad nodes
  - If you point to a Bad node, you're Bad
  - If a Good node points to you, you're Good



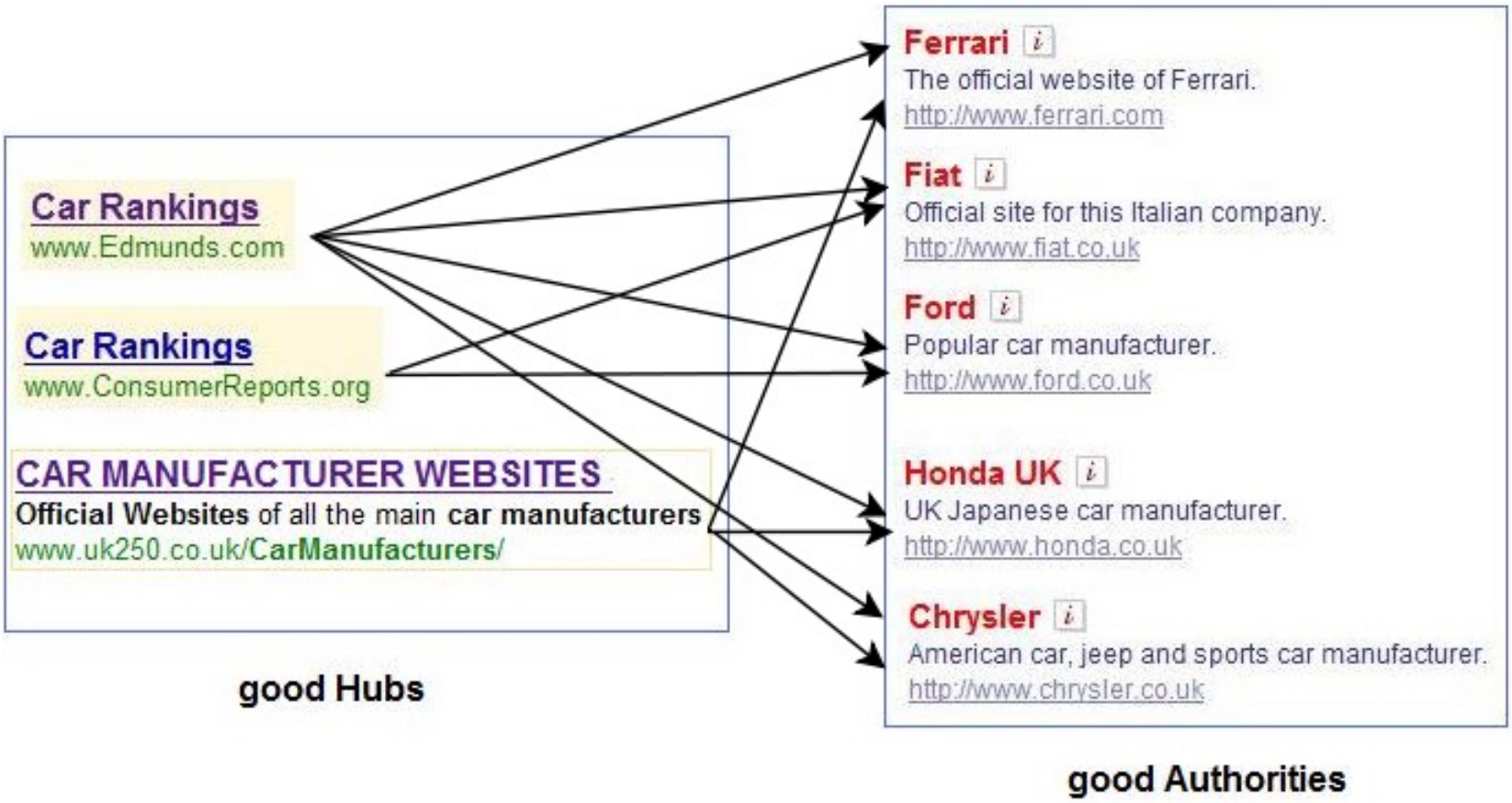
HITS ALGORITHM

HITS algorithm

- Hyperlink-Induced Topic Search, by Kleinberg

- Two types of important pages on the Web
  - Authority: has authoritative content on a topic
  - Hub: pages which link to many authoritative pages, e.g., a directory or catalog
  - A good hub is one which links to many good authorities
  - A good authority is one which is linked to by many good hubs

The hope



Query: Top automobile makers

HITS

- HITS computes two scores for each page *p*
  - Authority score: sum of hub scores of all pages which point to *p*
  - Hub score: sum of authority scores of all pages which *p* points to

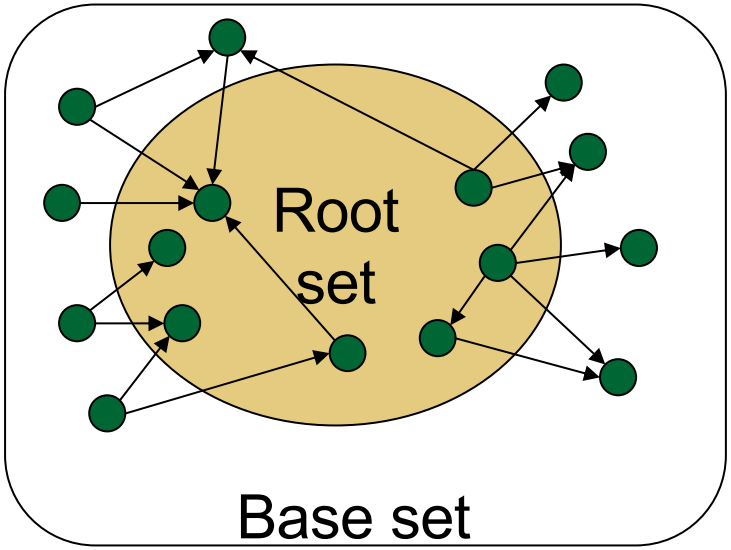
- Iterative algorithm
  - A series of iterations run, until the scores of all pages converge

HITS run on a query-dependent sub-graph

- Meant to run on a (sub)set of pages that are relevant to a given query
  - Top N pages relevant to query retrieved based on content → called the root set
  - Add to the root set all pages that are linked from it or that links to it → base set
  - Sub-graph of all nodes in base set → focused sub-graph

- Motivation of building base set
  - A good authority page may not contain the query term
  - Hubs describe authorities through the anchor text / text surrounding hyperlinks

Visualization



HITS Algorithm

Find focused sub-graph G of pages relevant to given query

for each page p in G:

p.auth ← 1, p.hub ← 1

do until convergence

for each page p in G

p.auth ← Σ q.hub for all pages q which link to p

p.hub ← Σ r.auth for all pages r which p links to

Normalize hub and auth scores for all pages

Check convergence of scores

Normalization of scores

- Scores need to be normalized after each iteration
  - To prevent the *hub* and *auth* values from getting too big
  - Scaling factor does not really matter; we are only concerned with the relative values of the scores

- Different normalization schemes proposed
  - Normalize so that score vectors sum to 1

  - Normalization factor F: square root of sum of squares of current scores of all pages; divide score of each page by F at the end of each iteration

Checking for convergence

- Various convergence criteria used
  - Fixed number of iterations
  - Iterate until scores do not change appreciably from one iteration to the next (compute difference of score vectors from previous and current iterations)
  - Iterate until rankings of pages do not change

Matrix version of HITS

- Matrices / vectors
  - A: adjacency matrix of web graph. (u, v)-th element is 1 if page u links to page v
  - h: vector of hub scores of all pages
  - a: vector of authority scores of all pages

- $h \leftarrow A.a$
- $a \leftarrow A^T.h$

HITS not used commonly

- Topic Drift: Off-topic pages can cause off-topic "authorities" to be returned
- Hubs often transit to authorities
- Search engines themselves become hubs