Date · · No.

NETWORK SEARCH

I/ Web search

- The Web : A network such that { Nodes = Web pages; Edges = Hyperlinks }

(Hyperlinks run in one direction ⇒ The Web is a directed network)

- Web crawler : A computer program that automatically surfs the Web looking for pages by performing a breadth-first search on the network from a given initial page.

- The overall process of modern search engines :
  - The web crawler finds web pages
  - Text from those pages is processed to create an annotated index while the structure of hyperlinks is used to calculate centrality scores for each page
  - When a query is entered, the search engine extracts a broad set of matching pages from the index and scores them based on various query-specific measures
  - Those scores combine with the pre-computed centrality scores and some other pre-computed quantities to give each page in the set an overall score
  - The pages are sorted in order of their overall scores and those with the highest scores are transmitted to the user

⊕ Example: Google search engine makes use of the eigenvector centrality measure known as PageRank. Pages are given a high score if they receive hyperlinks from many other pages, but a link from those hyperlinks is weighted higher if it is from a highly ranked page itself. (More details can be found in Section 7.1, [Ne])

LOOSE-LEAF 7mm ruled × 35lines

No. Date

II/ Searching distributed databases

* Mainly on peer-to-peer file-sharing network

- Peer-to-peer file-sharing network (P2P):
  + A network such that { Nodes = Computers containing files
    Edges = Virtual links established for file-sharing purpose
  + P2P is used to form a direct transfer of data between two users' computers in the network.
    ⇒ Contrast to server-client model (e.g. World Wide Web) in which central server computers supply the requested data to a large number of client machines

⚠ World Wide Web is also a distributed database but web search works in a different way than searches in other distributed database.

- The simplest form of distributed search is a version of the breadth-first search algorithm
  + Advantage: The number of steps is small even when the network is large (See Section 11.7, [Ne])
  + Disadvantage: The network can be overloaded due to the size of the network and the variety in the nodes' bandwidth capabilities

⇒ Solution: Use supernodes — high-bandwidth nodes chosen from the network and connected to each other to form a supernode network

⇒ A supernode acts as a link between its clients — normal users attached to the supernode — and the rest of the network.

⇒ In modern peer-to-peer networks, the search is mostly a breadth-first search in the supernode network.

Date · · No.

III / Sending messages

- The problem of delivering a message to a particular node in the network is a different variation of the distributed search problem.

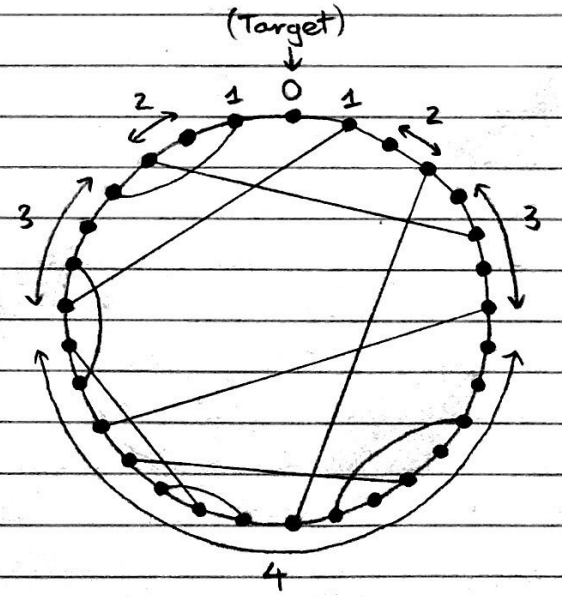
⊛ The classic Stanley Milgram's "small-world" experiment (see Section 4.6, [Ne])

- The participants were able to find short paths without knowing the structure of the whole network.

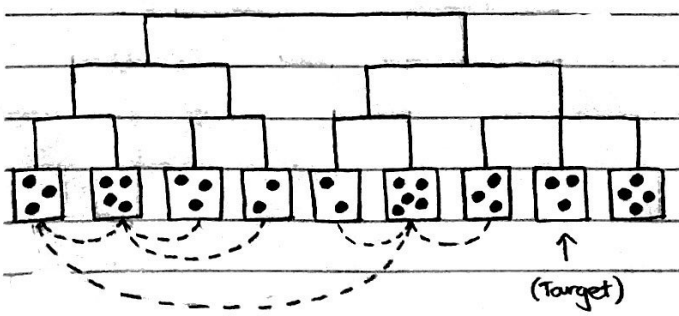
⇒ The social (or other) networks must have a specific structure so that one can find short paths without knowledge about the whole network.

1) The Kleinberg's model:

+ The model consists of a ring of nodes and a number of "shortcut" edges.
+ Nodes are divided into classes based on their "closeness" to the target.
+ More shortcuts can be found connecting nearby nodes than distant nodes.



2) A hierarchical model:



+ Nodes are grouped into "boxes", which form a hierarchical structure represented by a tree.
+ The structure might correspond to the division of geographical space into countries, regions, cities, and so forth.
+ Social connections (dotted curves) are more likely to be found between nodes that are closer to each other in the tree.

LOOSE-LEAF 7mm ruled × 35 lines