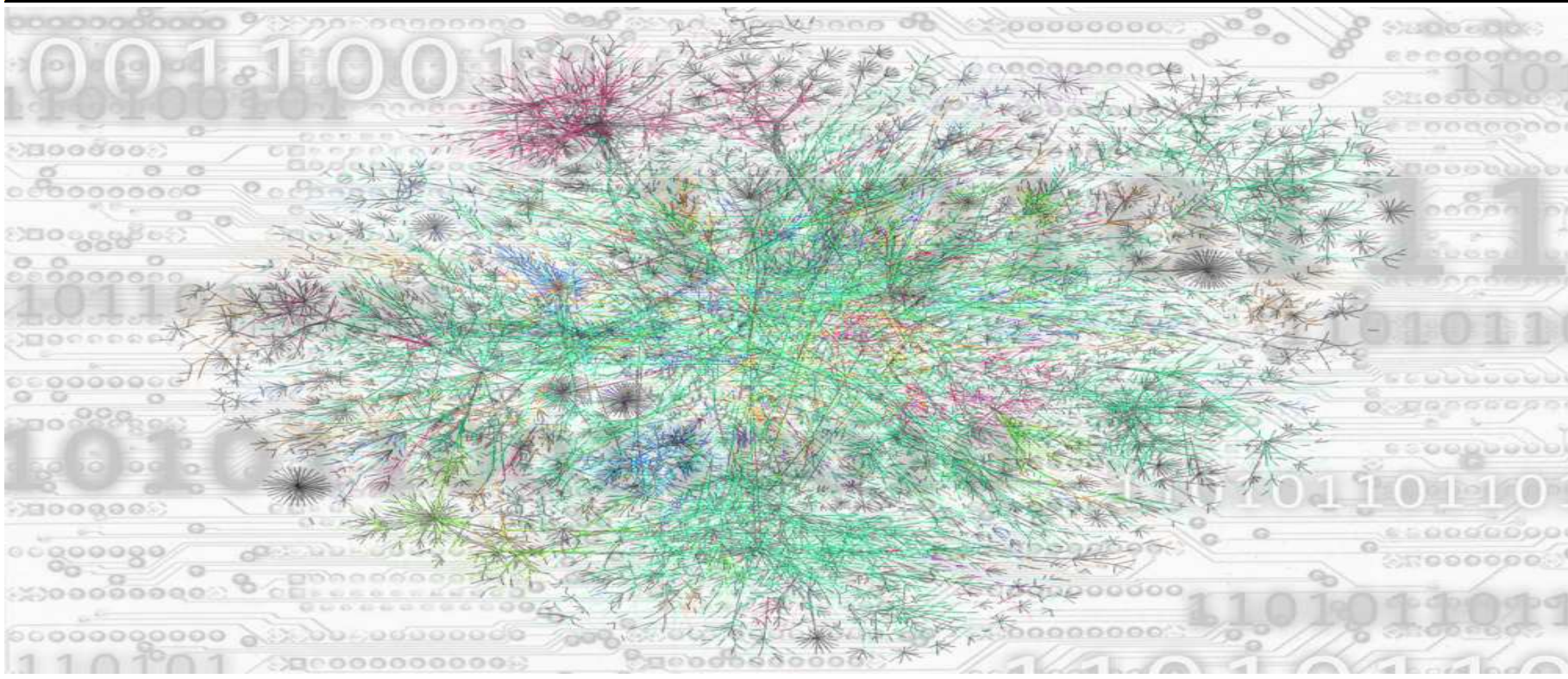


Power Laws And Rich Get Richer Phenomena



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Seminar Machinelles Lernen
2011-01-12, Tomek Grubba



Outline



-
- The Structure of the Web
 - History of the Web
 - Graph structure
 - Bow-Tie Structure
 - Web 2.0
 - Power Laws and the Rich-Get-Richer Phenomena
 - Popularity of web pages
 - Degree distribution
 - Effects of search engines

History



-
- World Wide Web
 - 89-91: Tim Berners-Lee
 - original conception:
 - create a place to...
 - ...publish information/documents
 - ...access information/documents
 - today: sequence of webpages rendered in a browser

Hypertext



- Web forms an information **network**
- Idea of Hypertext:
 - replace traditional linear structure of text with a network structure
- Precursors of hypertext
 - concept of citations
 - books
 - patents
 - legal decisions

Memex



- Hypothetical prototype by Vannevar Bush (1945)
- Contains digitized versions of all human knowledge
- Connected by associative links
 - inspired by the „associative memory“ of humans

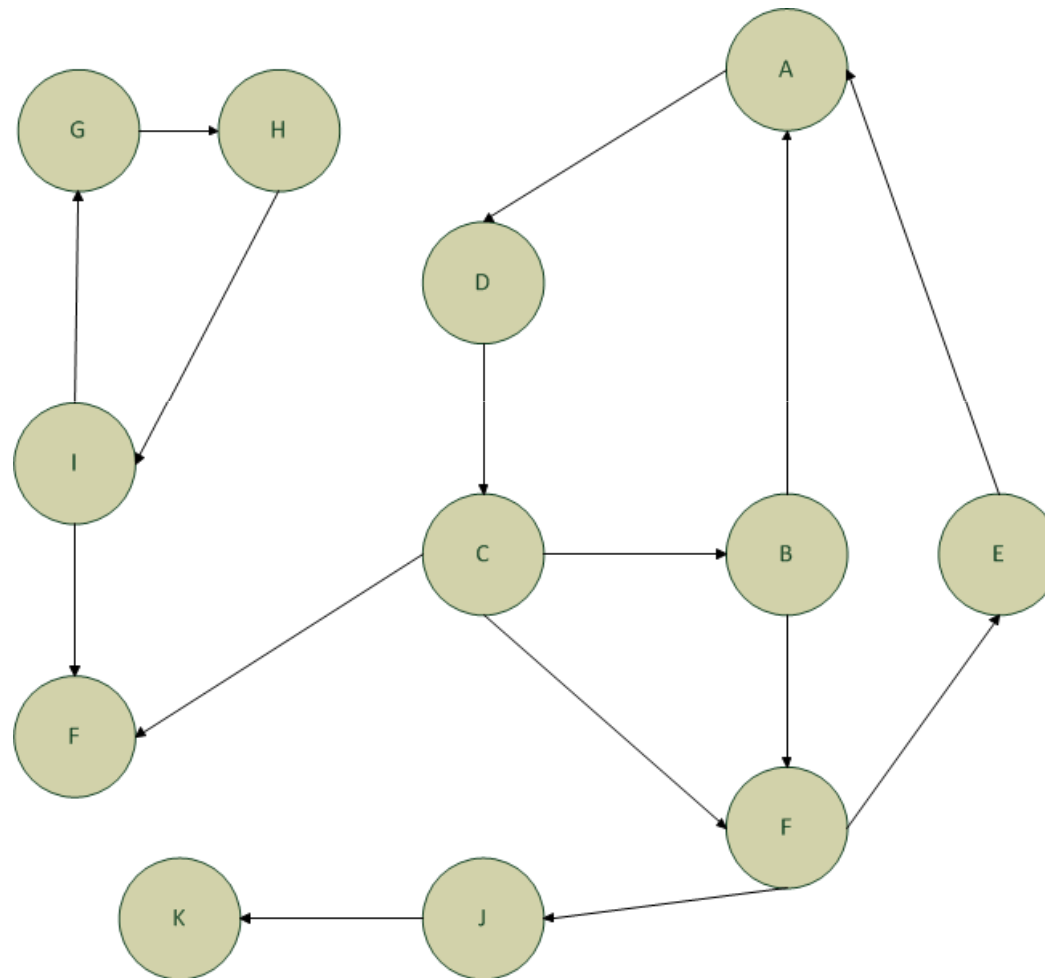
The Web as a Graph



- Web pages considered as nodes
- Links between web pages considered as edges/vertices
 - links are unidirectional!

- Distinction between navigational and transactional links
 - only navigational links considered

The Web as a Directed Graph



The Web as a Directed Graph

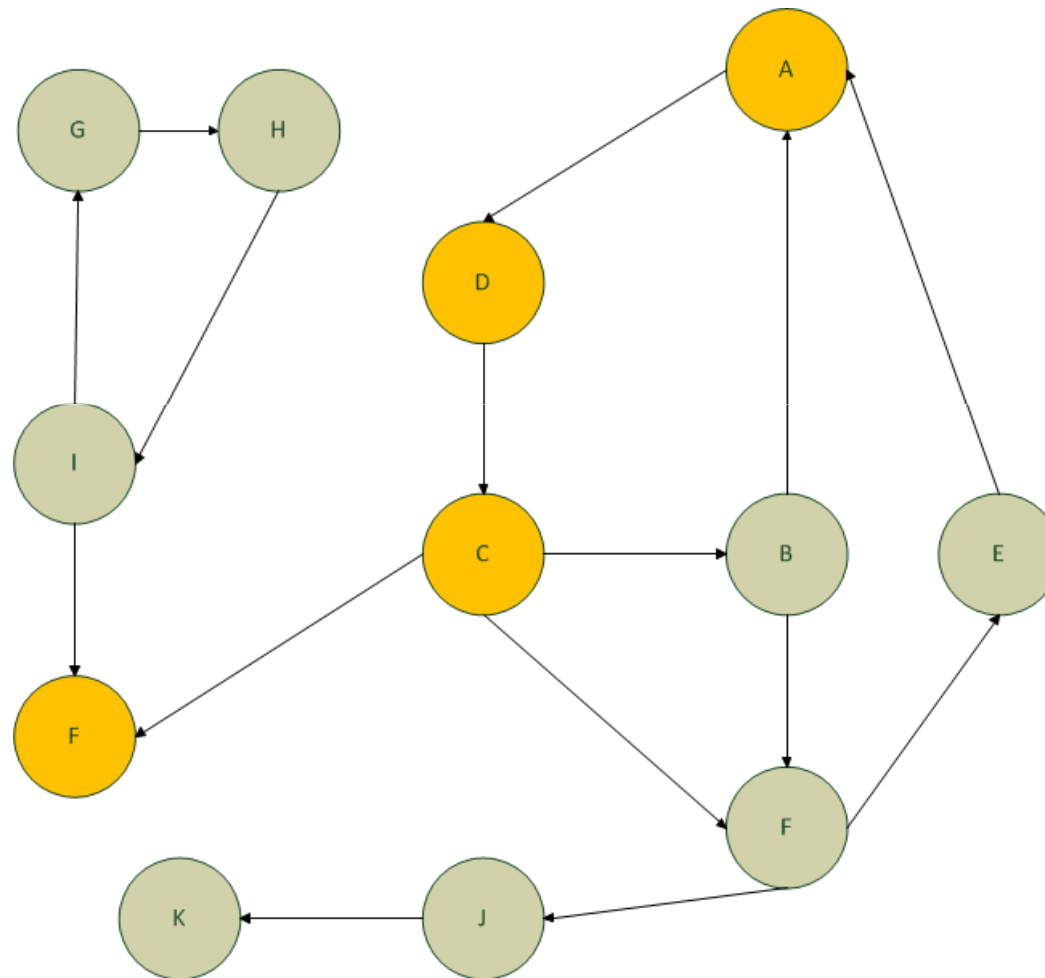


- Path
 - sequence of nodes from A to B, with the property that each consecutive pair of nodes in the sequence is connected by an edge pointing in the forward direction

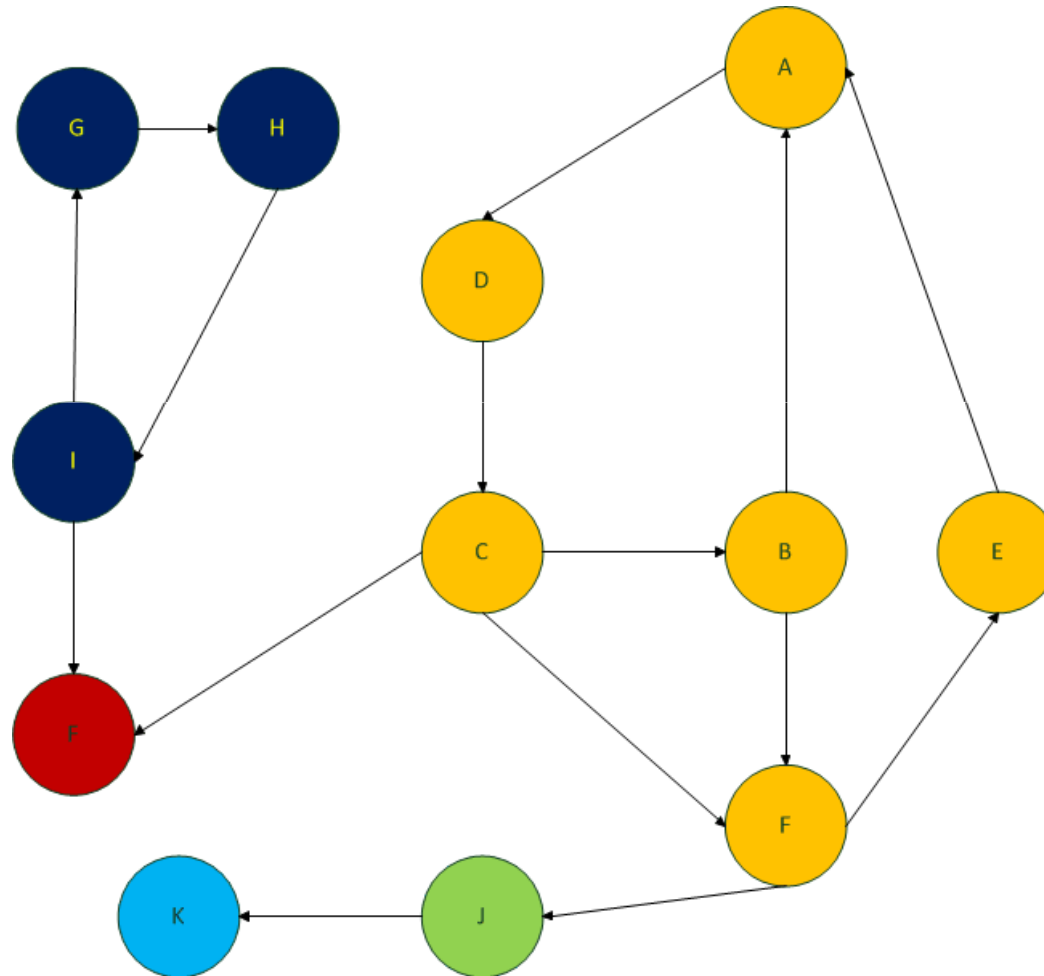
 - Strong Connectivity
 - a path exists from every node to every other node

 - Strongly Connected Component (SCC)
 - subset of nodes such that
 - i. every node in the subset has a path to every other node
 - ii. the subset is not part of some larger set with the property of i
-

The Web as a Directed Graph



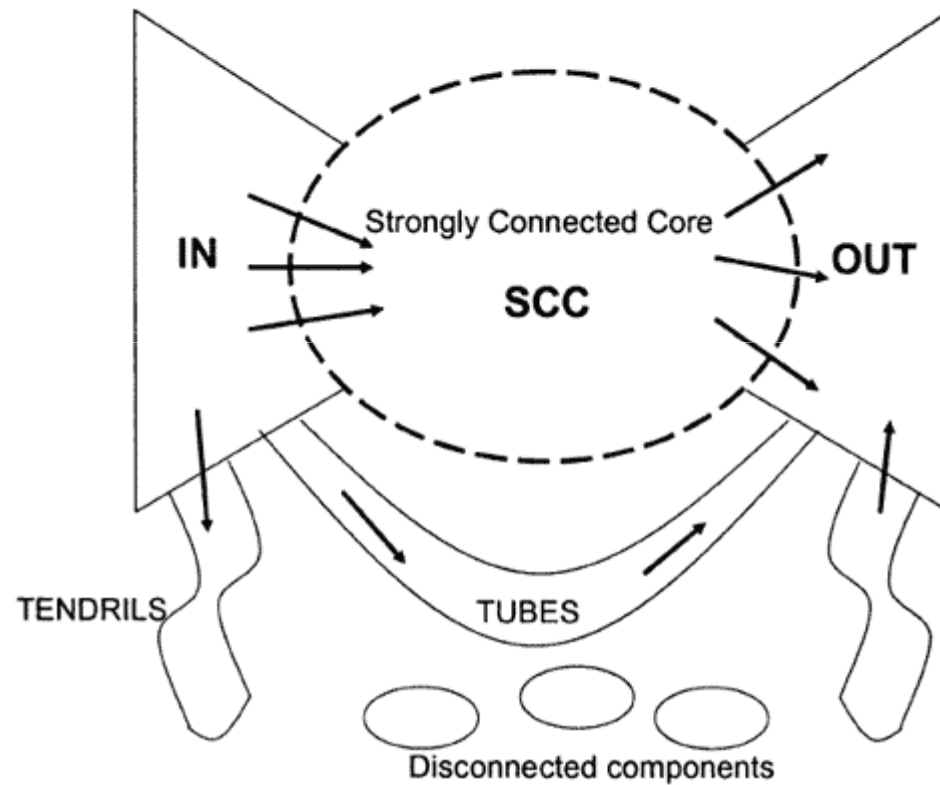
The Web as a Directed Graph



Bow-Tie Structure

- Provides a high-level view of the structure
- „Components“ of the web
 - One giant Strongly Connected Component
 - IN-component
 - OUT-component
 - Tendrils
 - Disconnected components

Bow-Tie Structure



Web 2.0



-
- Changes between (roughly) 2000 and 2009
 - No exact definition

 - Major forces
 - i. growth of authoring styles to allow people to create content
 - ii. movement of people's personal data to the internet („cloud“)
 - iii. on-line connections between people

Web 2.0



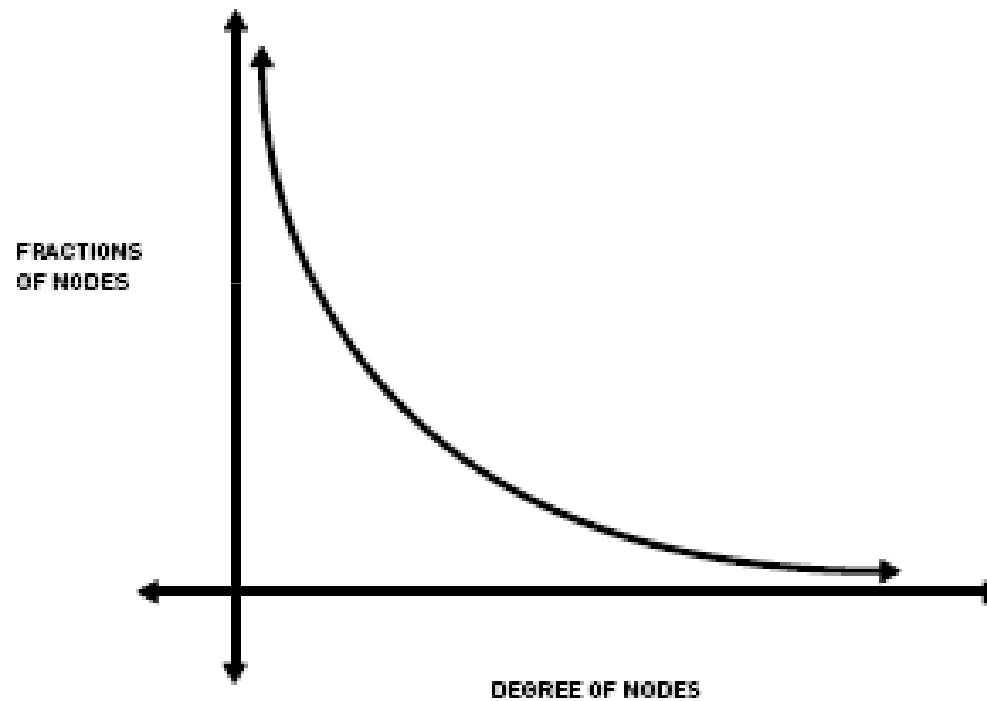
-
- Social phenomena of Web 2.0
 - „Software gets better the more people use it“
 - „The wisdom of crowds“
 - „The Long Tail“

Popularity



- The popularity of a web page is measured by the in-degree (number of in-links)
- As a function of k , what fraction of pages on the web have k in-links?
- First assumption: The Normal Distribution
 - turned out to be wrong
- Observation: $1/k^2$
 - such a function is called Power Law

Power Law



<http://www.learner.org/courses/mathilluminated/units/11/textbook/05.php>

Rich Get Richer Models



- Barabasi-Albert Model
 - makes use of „Preferential Attachment“
 - the probability, that node A links to node B is proportional to the in-degree of B

- Analogy from the real world
 - the more people know a person A (e.g. a celebrity), the higher the probability, that person B has heard of it

Unpredictability



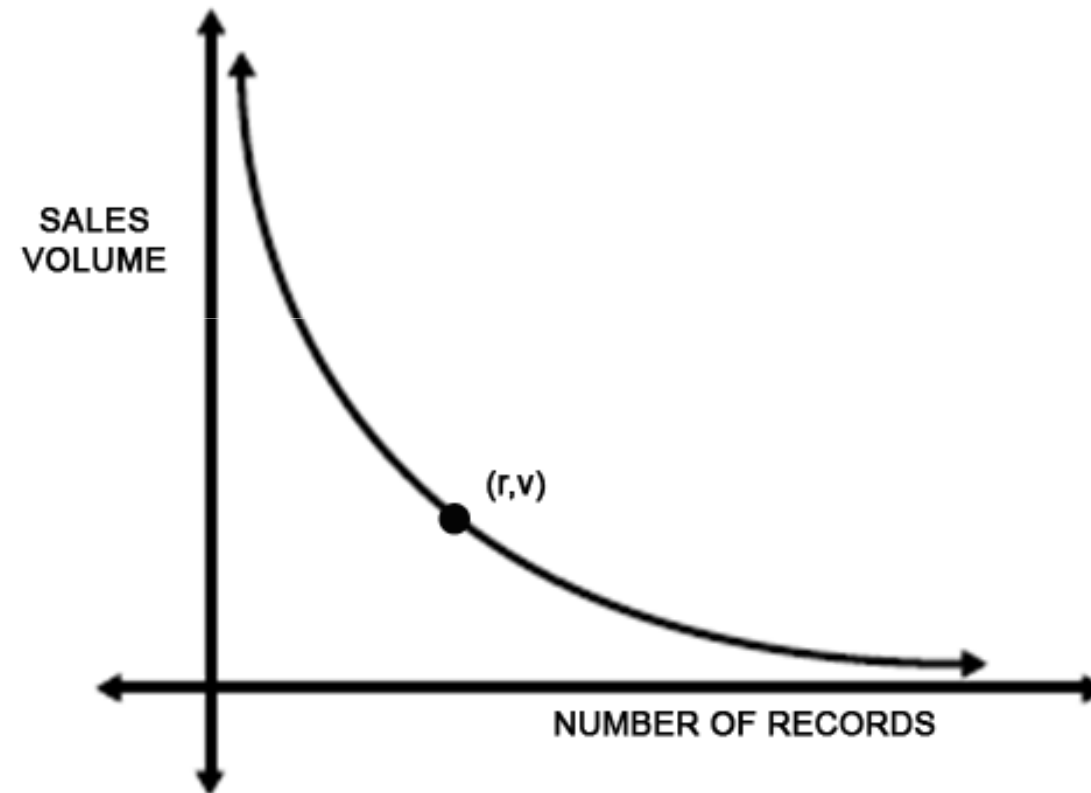
- Information cascades can depend on the outcome of a small number of initial decisions
 - Salgankik et al: experiments on a music download site
- May lead to such an effect, that a worse technology can win, because it reaches a certain critical audience earlier

The Long Tail



- Popular retailing concept
 - e.g. music business
 - selling records by many less known artists can make a bigger bunch of the revenue than selling records by fewer bigger artist

The Long Tail



Effect of Search Tools



- Do internet search tools, such as Google, make the rich-get-richer dynamics of popularity more extreme or less extreme?

- What about recommendation tools?

Thank You!