

The Web

Social Computing

Department of Computer Science
University of Massachusetts, Lowell

Hadi Amiri
hadi@cs.uml.edu

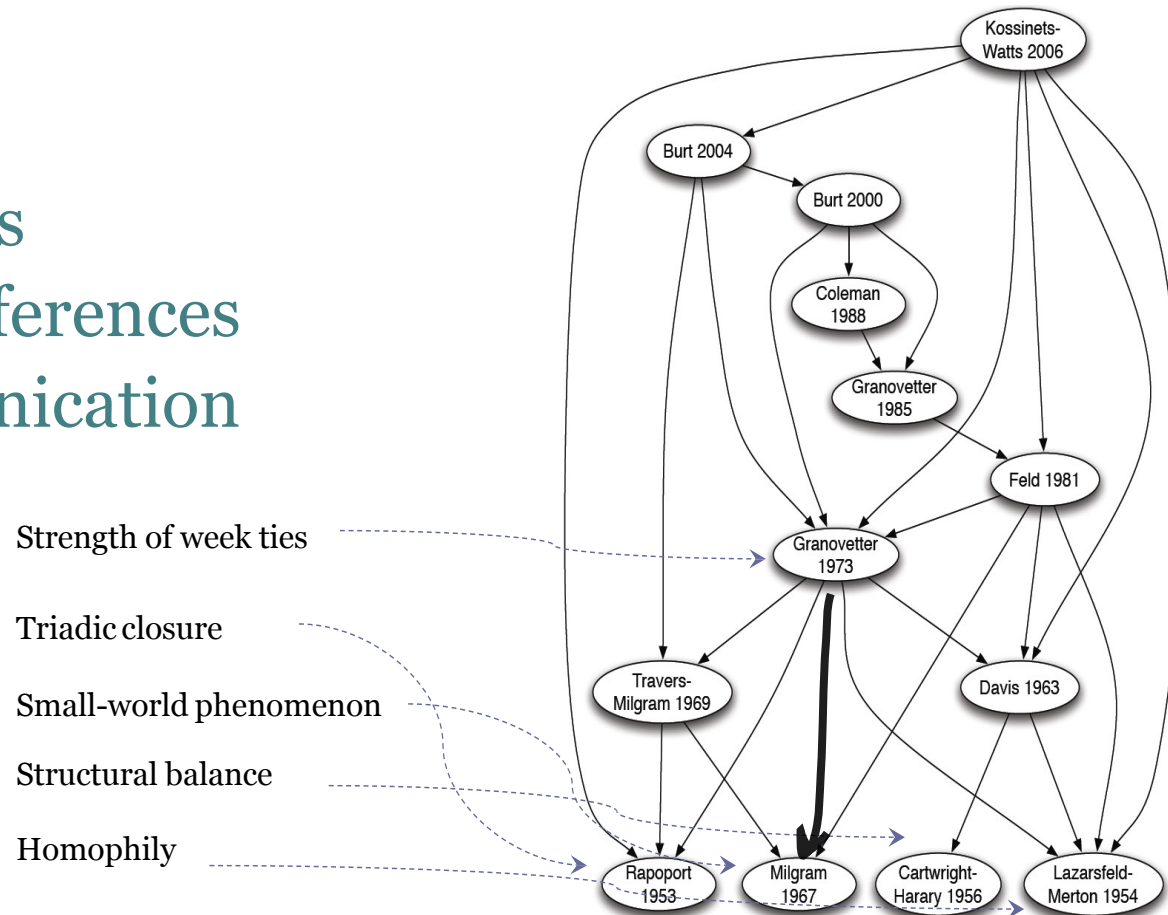


Lecture Topics

- The structure of the Web
- Power Law in SCC and WCC

Information Networks

- Information Network
 - Nodes carry content and Edges join related nodes!
- Examples:
 - The Web
 - Citation networks
 - Encyclopedia References
 - Wireless communication
 - etc.



The World Wide Web

- Created by Tim Berners-Lee & his colleagues during 1989-1991 in CERN:
 - CERN (Geneva, Switzerland)



Q: Did you invent the internet?

A:

No, no, no!

When I was doing the WWW, most of the bits I needed were already done.

Vint Cerf and people he worked with had figured out the Internet Protocol, and also the Transmission Control Protocol.

Paul Mockapetris and friends had figured out the Domain Name System.

People had already used TCP/IP and DNS to make email, and other cool things. So I could email other people who maybe would like to help work on making the WWW.

I didn't invent the hypertext link either. The idea of jumping from one document to another had been thought about lots of people, including Vanevar Bush in 1945, and by Ted Nelson (who actually invented the word hypertext). Bush did it before computers really existed. Ted thought of a system but didn't use the internet. Doug Engelbart in the 1960's made a great system just like WWW except that it just ran on one [big] computer, as the internet hadn't been invented yet. Lots of hypertext systems had been made which just worked on one computer, and didn't link all the way across the world.

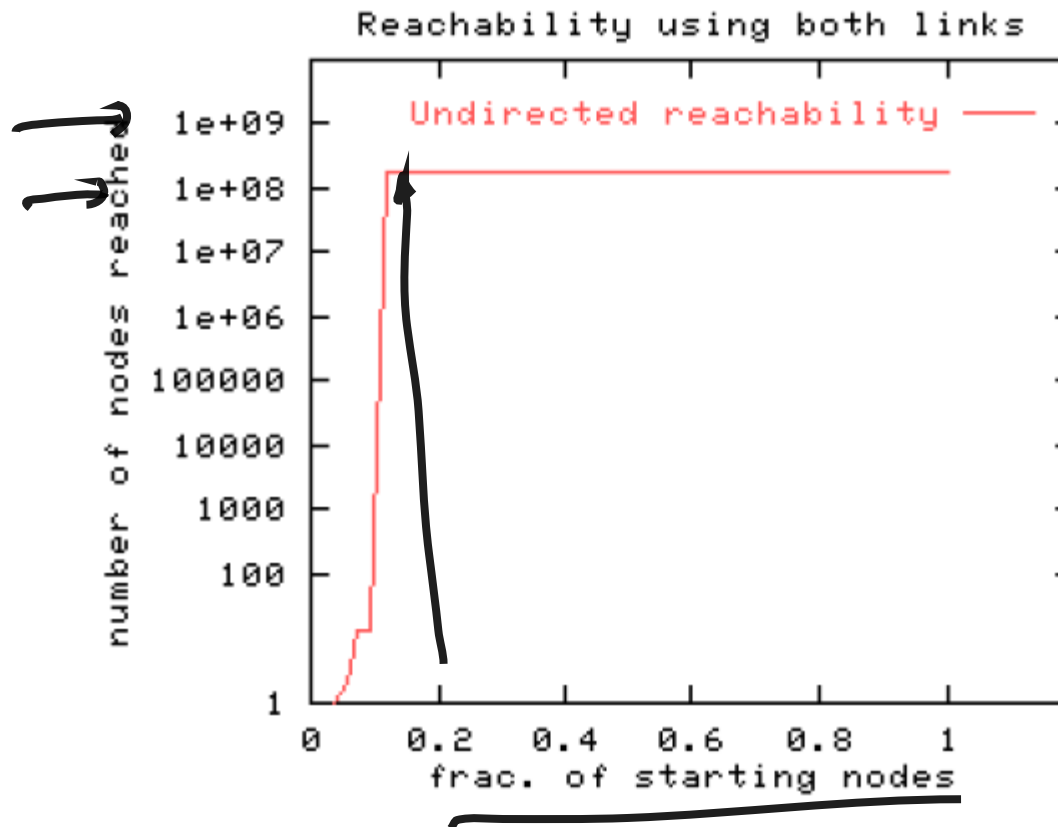
I just had to take the hypertext idea and connect it to the TCP and DNS ideas and -- ta-da! -- the World Wide Web.

Web Structure

- How does the Web look like?
 - Broder et al., Graph structure in the Web. WWW'00:
 - Altavista data
 - Crawl from October, 1999 containing
 - 203 million URLs
 - 1,466 million links.

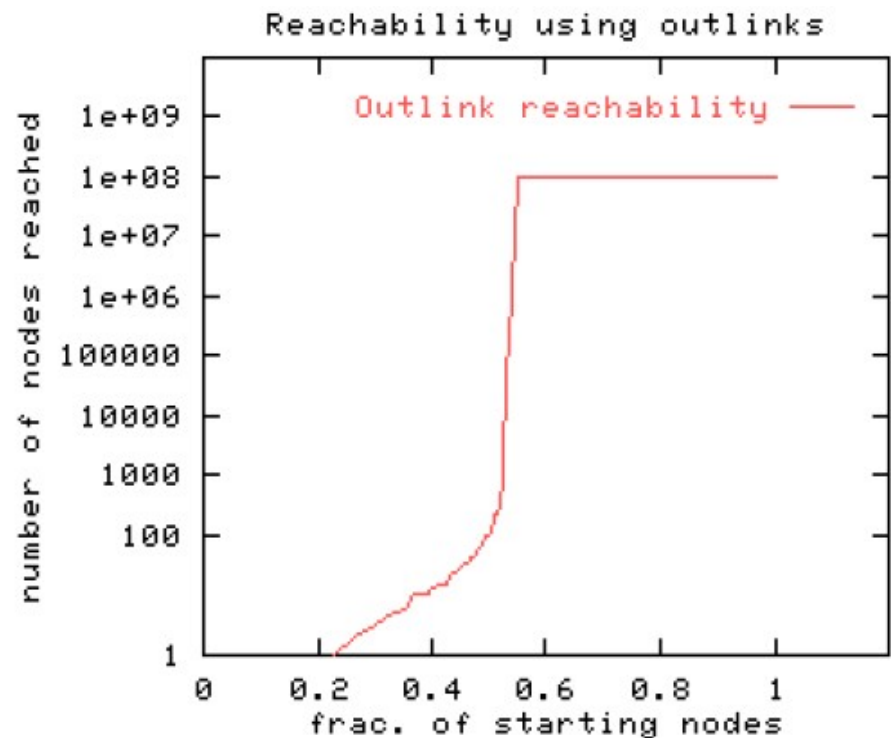
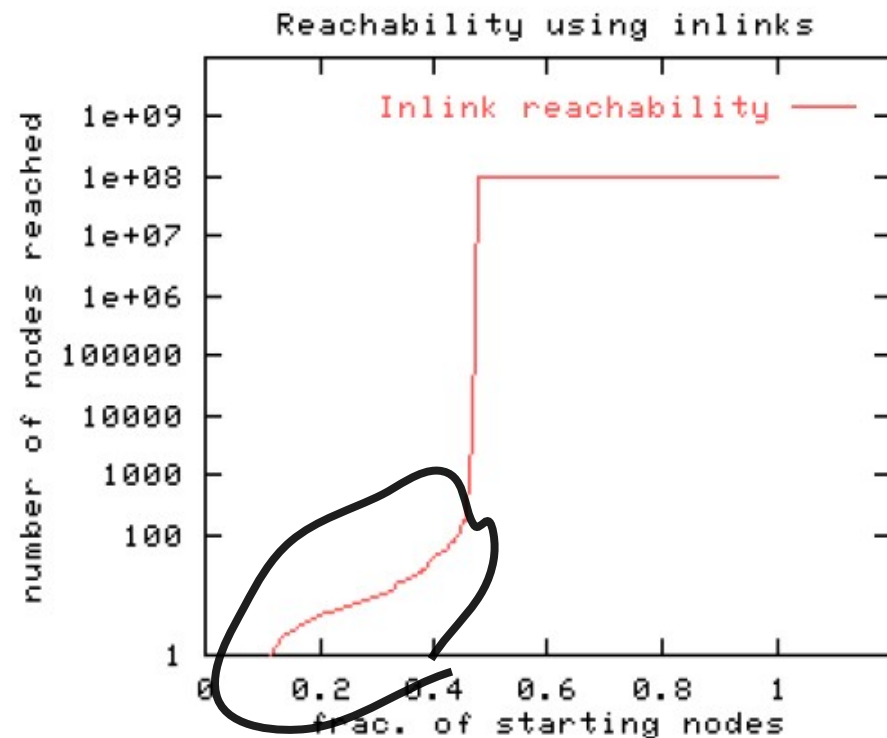
Web Structure- Cnt.

- Running BFS starting from random nodes
 - undirected.



Web Structure- Cnt.

- Running BFS starting from random nodes
 - in-links & out-links.



Web Structure- Cnt.

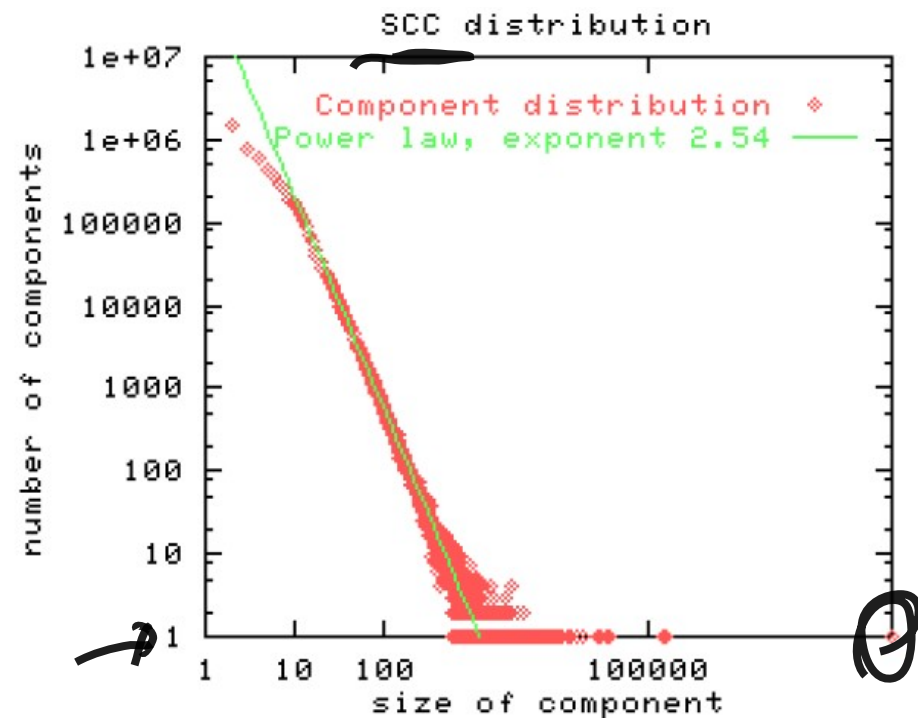
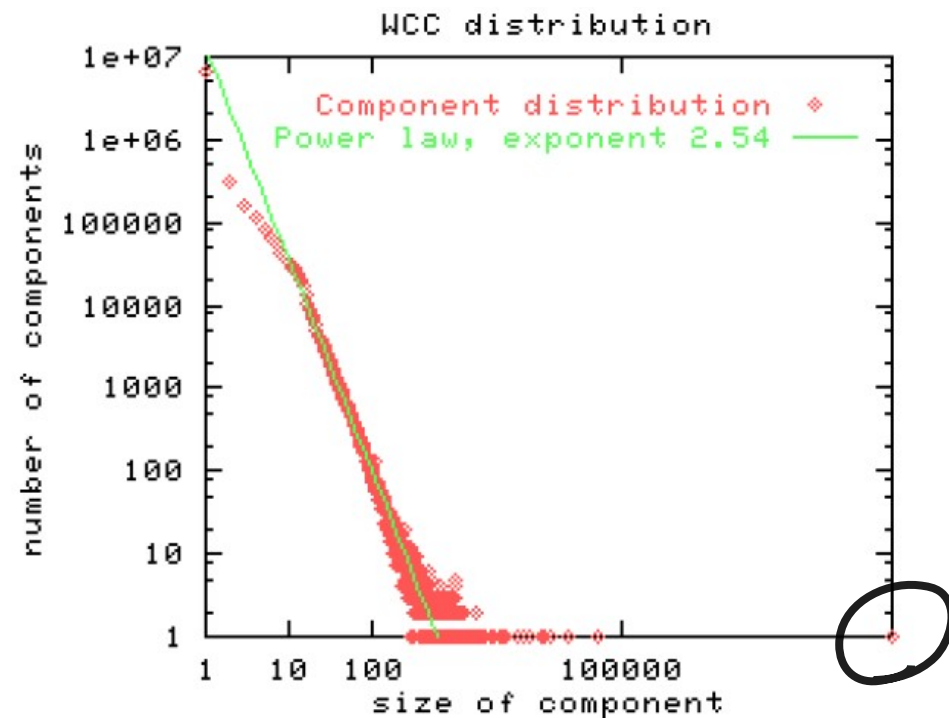
- Distribution of SCCs and WCCs on the web.
 - SCC of G is a **maximal** set of nodes \mathbf{C} such that for all u, v in \mathbf{C} , both u and v are **reachable from each other**.

Web Structure- Cnt.

- Distribution of SCCs and WCCs on the web.
 - SCC of G is a **maximal** set of nodes \mathbf{C} such that for all u, v in \mathbf{C} , both u and v are **reachable from each other**.
 - WCC of G is a **maximal** set of nodes \mathbf{C} such that for all u, v in \mathbf{C} , there is an **undirected path between them**.

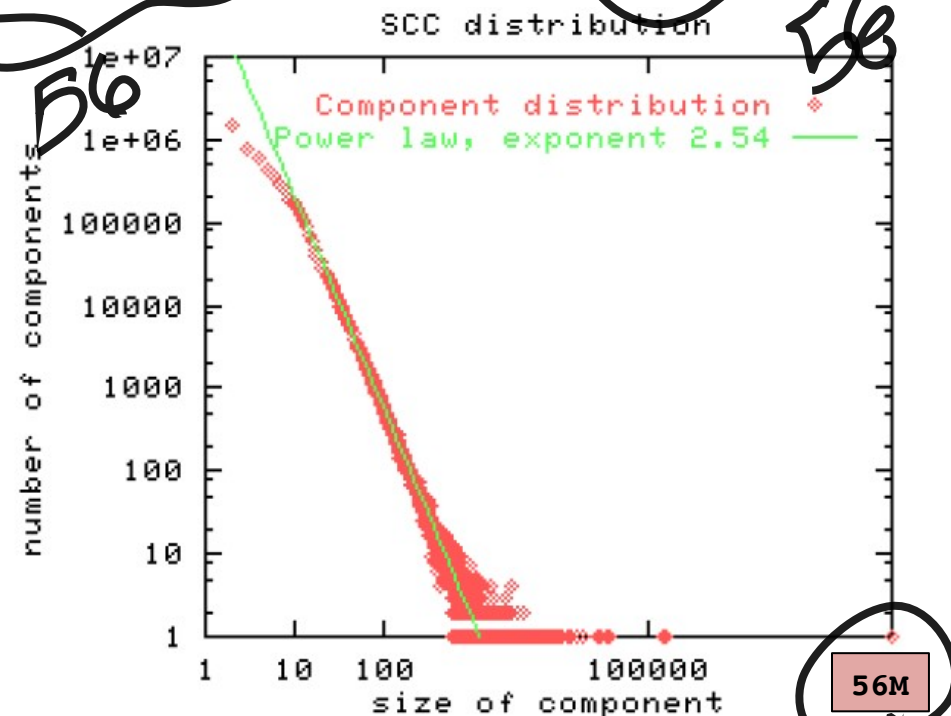
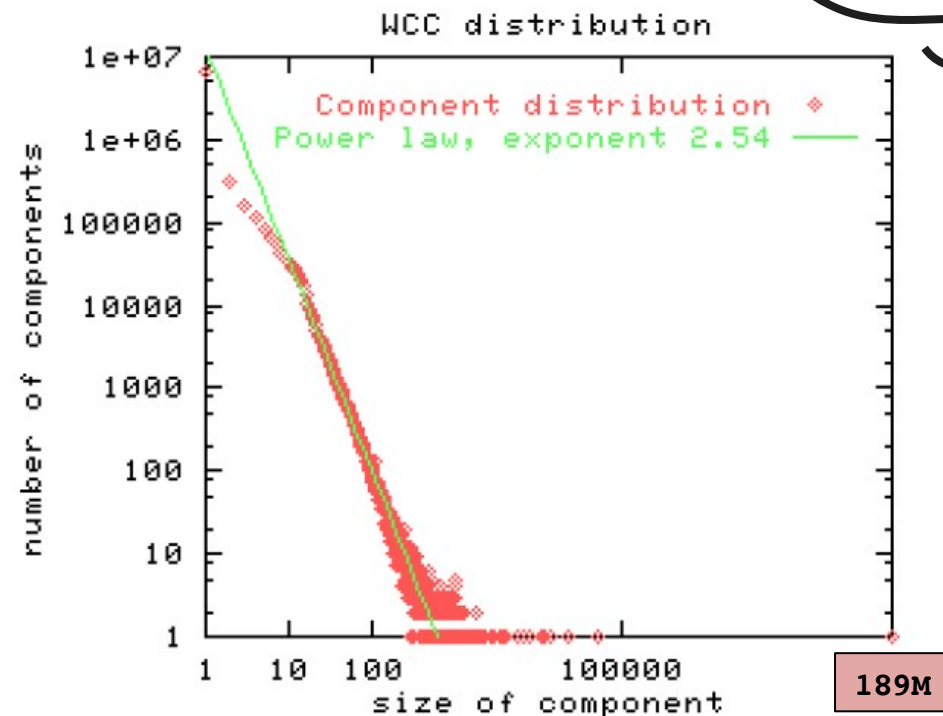
Web Structure- Cnt.

- Distribution of SCCs and WCCs on the web.



Web Structure- Cnt.

- Distribution of SCCs and WCCs on the web.



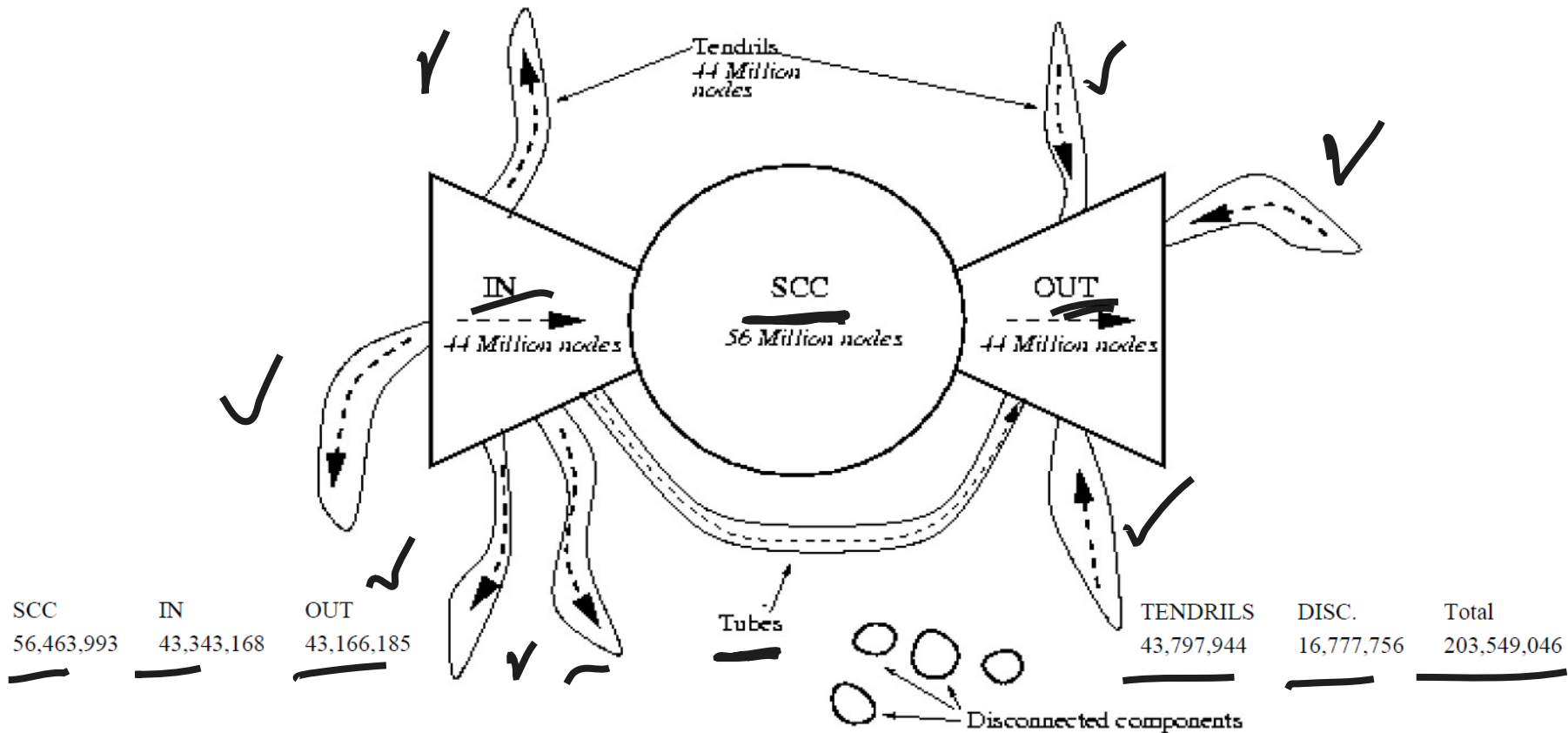
WCC: A giant component of 186 million nodes, 91% of the nodes in our crawl are reachable from one another through a path.

Web Structure- Cnt.

- The Web contains a SINGLE giant SCC.
 - If there were 2 giant SCCs, X and Y
 - a single link from any node in X to any node Y, and another link from any node in Y to any node in X is enough to merge X and Y to become part of a single SCC.

Web Structure- Cnt.

Bow-Tie Structure of the Web.



IN nodes: can reach SCC but cannot be reached from it.

OUT nodes: can be reached from SCC but cannot reach it.

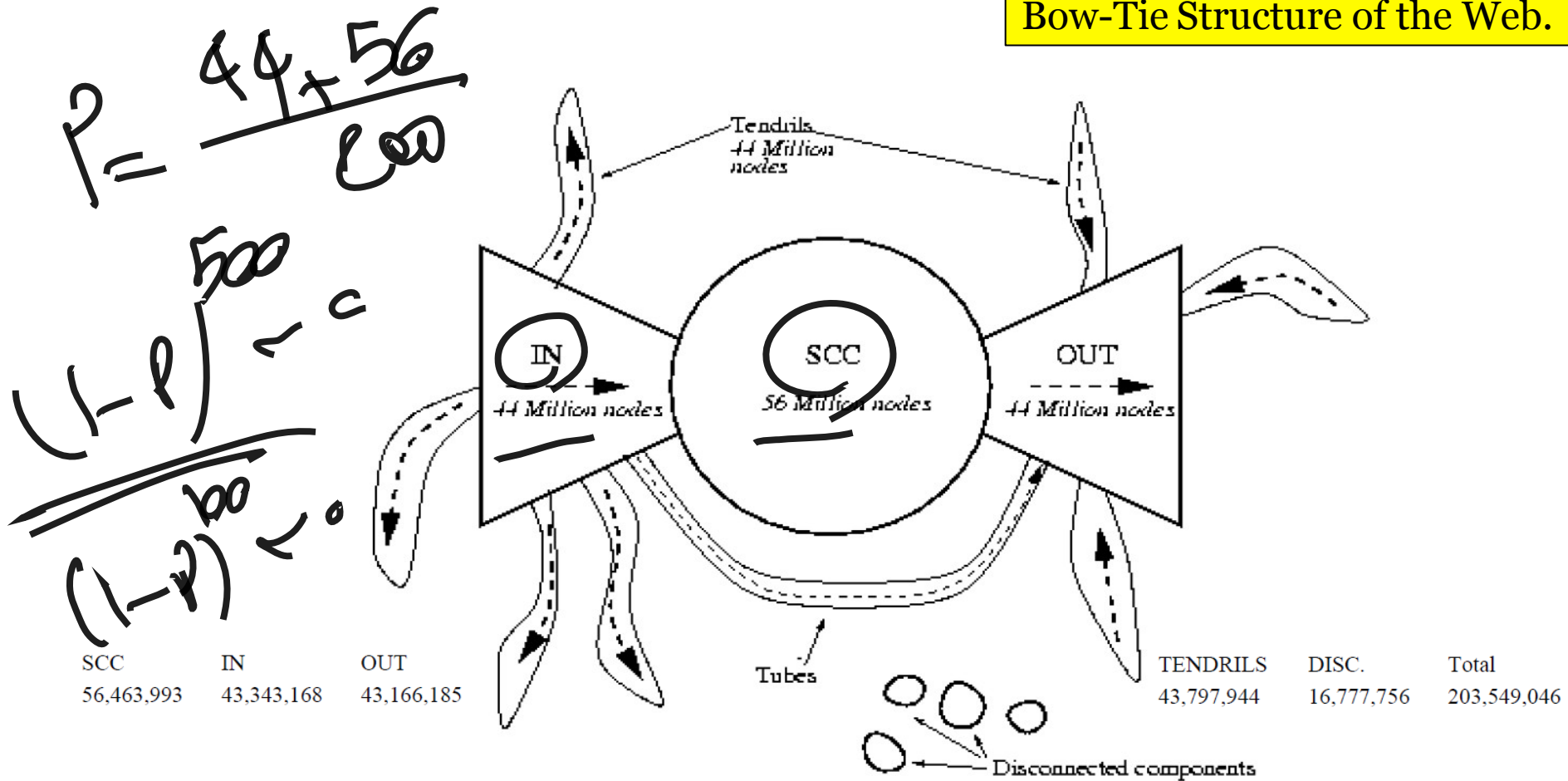
Tendrils nodes: (a) reachable from IN but cannot reach SCC, (b) can reach OUT but cannot be reached from SCC.

Tendrils nodes satisfying both a & b, travel in **tube** from IN to OUT without touching SCC.

Disconnected nodes: have no path to SCC ignoring directions

Web Structure- Cnt.

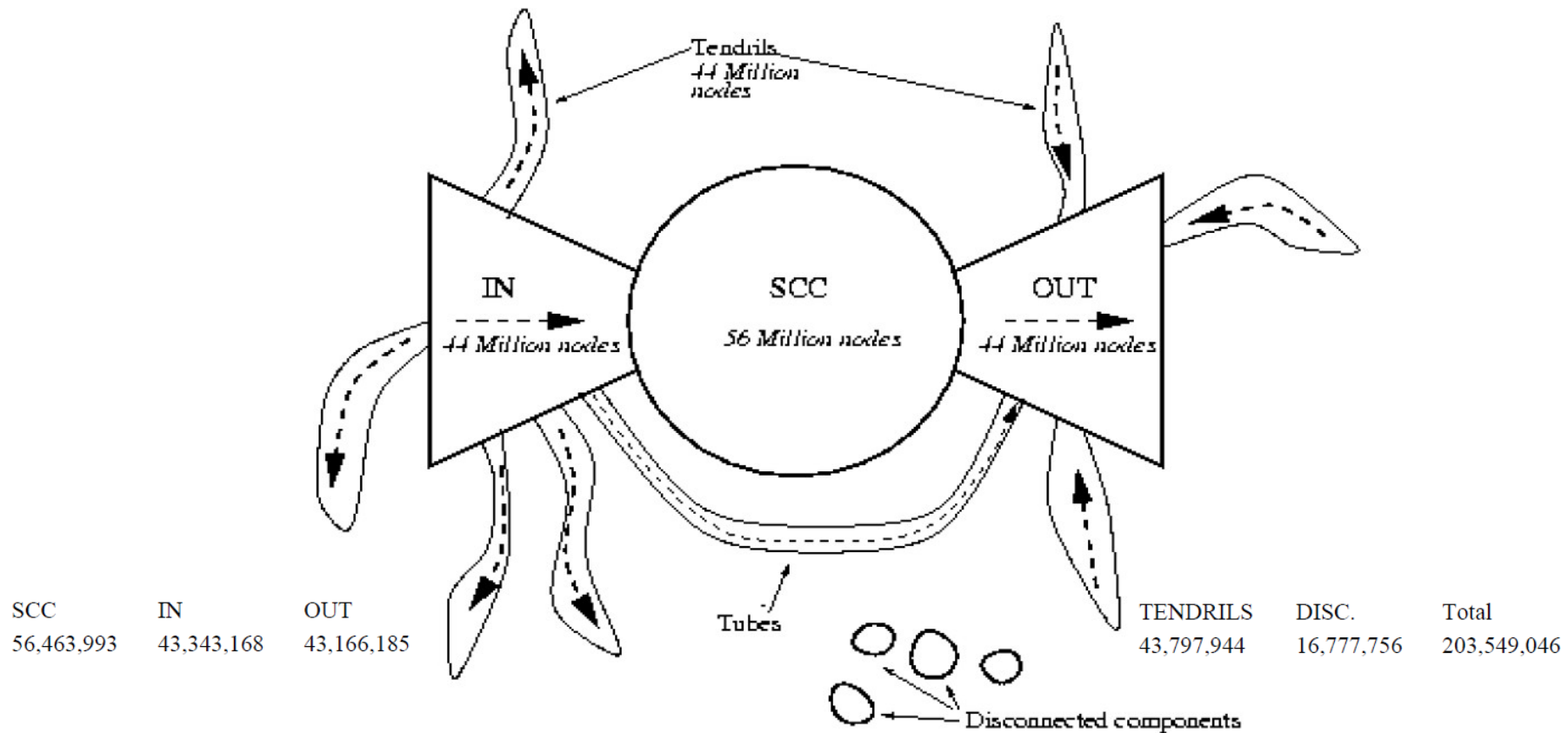
Bow-Tie Structure of the Web.



A forward BFS from any node in the SCC or IN will explode (reaches many other nodes), as will a backward BFS from any node in either the SCC or OUT.

Web Structure- Cnt.

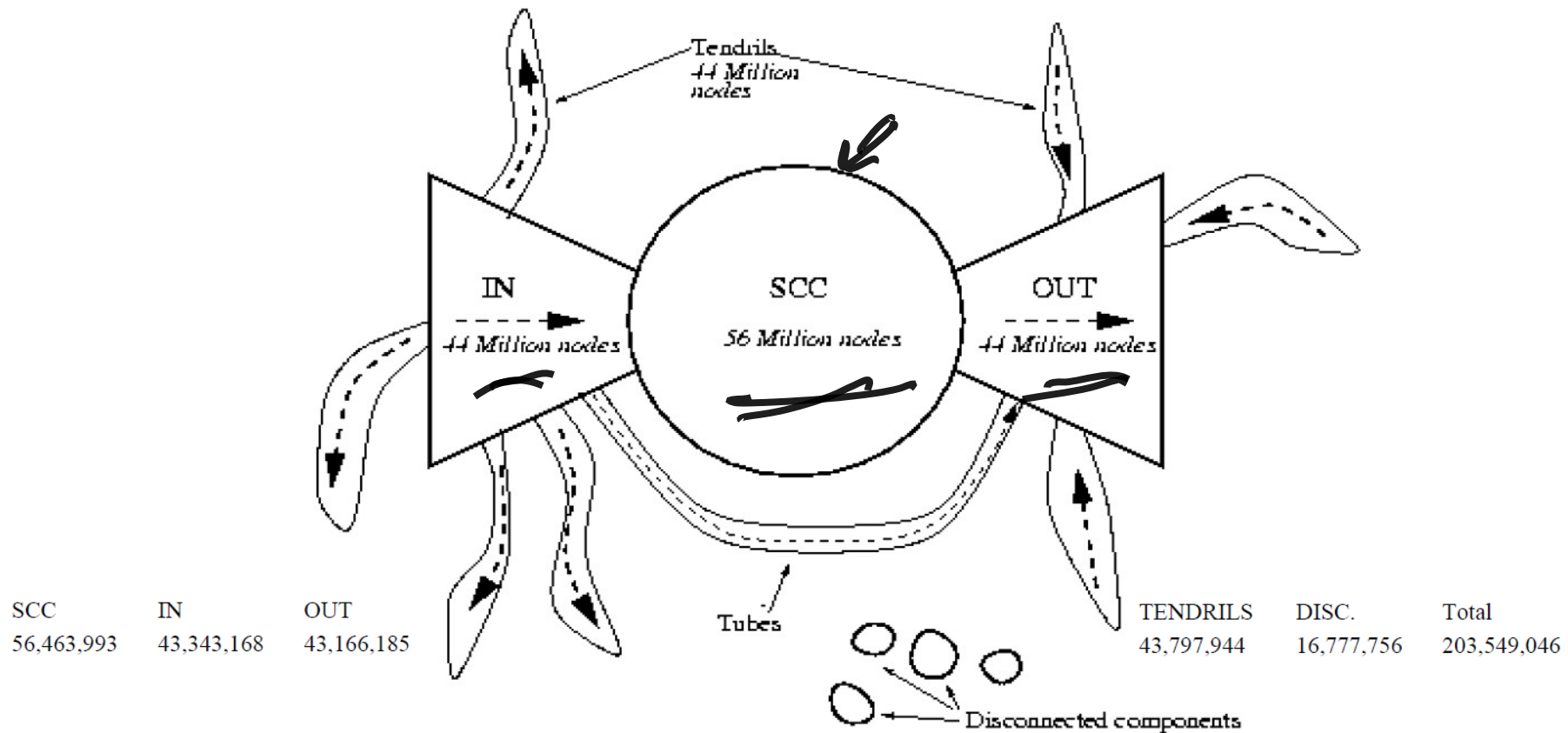
Bow-Tie Structure of the Web.



The Web structure is relatively stable despite the fact that nodes are constantly shifting their boundaries by entering and leaving the SCC over time.

Web Structure- Cnt.


Bow-Tie Structure of the Web.



Bow-tie structure provides a global view of the Web, but it doesn't provide insight into patterns of connections within the parts.

Web Structure- Cnt.

- Distribution of WCCs on the web.



k	<u>1000</u>	100	<u>10</u>	5	4	3
Size (millions)	<u>177</u>	167	<u>105</u>	59	41	15

Table 1: Size of the largest surviving weak component when links to pages with in-degree at least k are removed from the graph.

WCC: The graph is still connected:

1. The connectivity is extremely resilient and doesn't depend on the nodes with high in-degree.
2. High in-degree nodes are embedded in a graph that is well connected without them.

Web Structure- Cnt.

Bow-Tie Structure of the Web.

- Interesting directions:
 1. Does the structure **remain stable over time**?
 2. Mathematical **models for evolving graphs**?
 3. What's a good **notions of connectivity** for the web graph?
 - weak and strong
 - co-citation relation
 - bibliographic coupling (two nodes citing one or more nodes in common)
 - etc.

Reading

- Ch.13 The Structure of the Web [NCM]
- Strongly Connected Components
 - <http://www.personal.kent.edu/~rmuhamma/Algorithms/MyAlgorithms/GraphAlgor/strongComponent.htm>