

# The Structure of the Web

Machine Learning with Graphs

Department of Computer Science  
University of Massachusetts, Lowell  
Spring 2021

Hadi Amiri  
[hadi@cs.uml.edu](mailto:hadi@cs.uml.edu)



# Lecture Topics

- The Web

# Information Networks

- Information Network
  - Nodes are pieces of information and edges join the related ones!
- Examples of information networks:
  - The Web
  - Citation networks
  - Encyclopedia References
  - Wireless communication
  - Etc.

# Information Networks- Cnt.

- Sample Citation Net

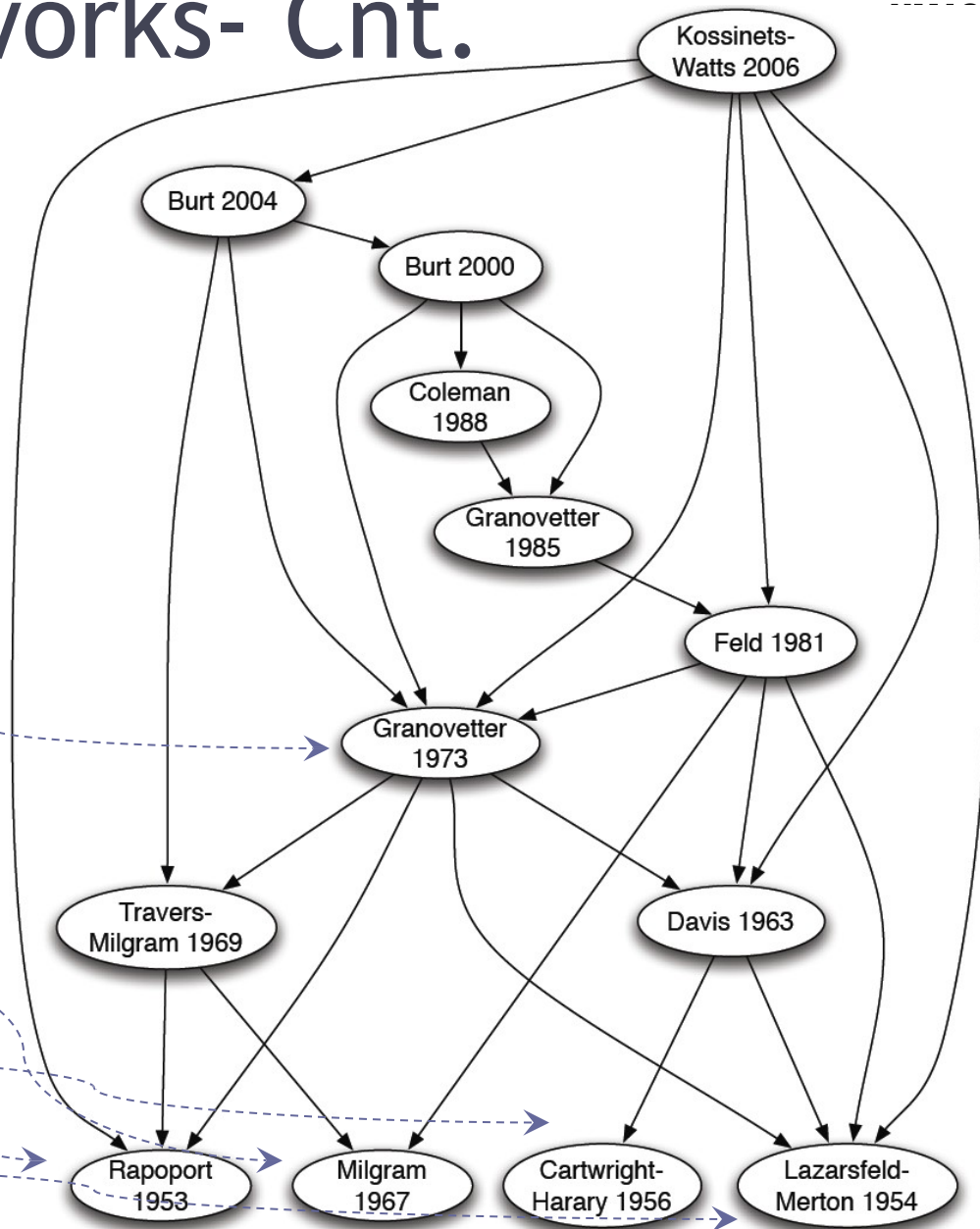
Strength of weak ties

Triadic closure

Small-world phenomenon

Structural balance

Homophily



# The World Wide Web

- Created by Tim Berners-Lee & his colleagues during 1989-1991 in CERN:
  - CERN is in 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.

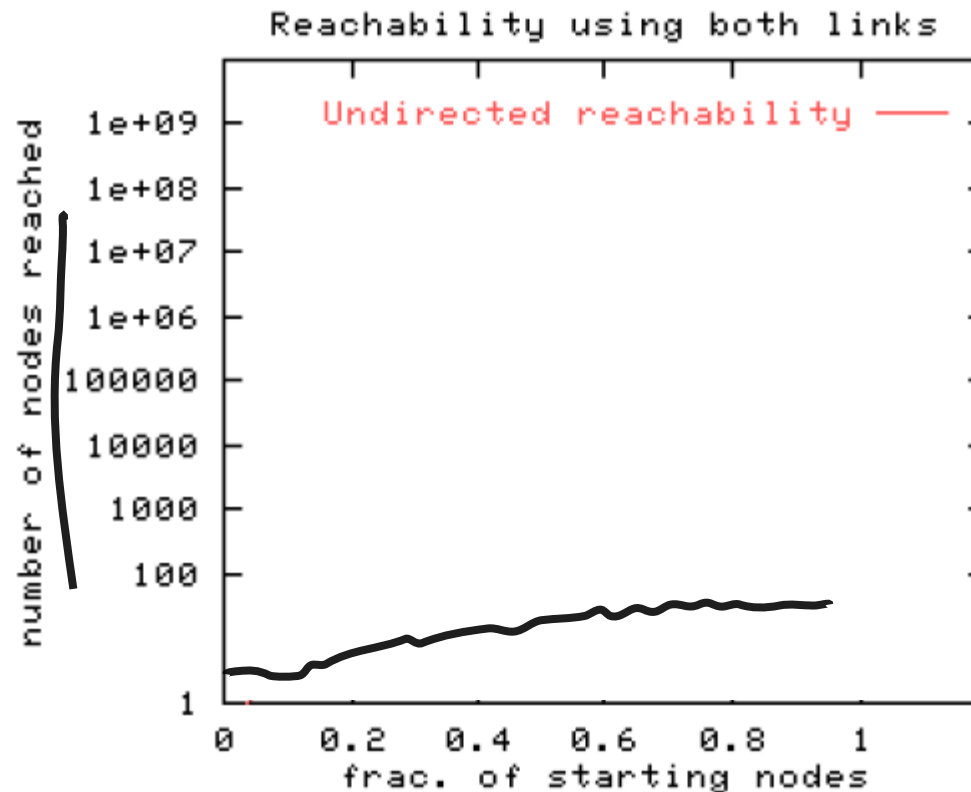


# Web Structure

- How does the Web look like?
  - Broder et al., Graph structure in the Web. WWW 2000:
    - Altavista data
    - Crawl from October, 1999 containing
      - 203 million URLs
      - 1,466 million links.

# Web Structure- Cnt.

- Running BFS starting from random nodes
  - undirected.

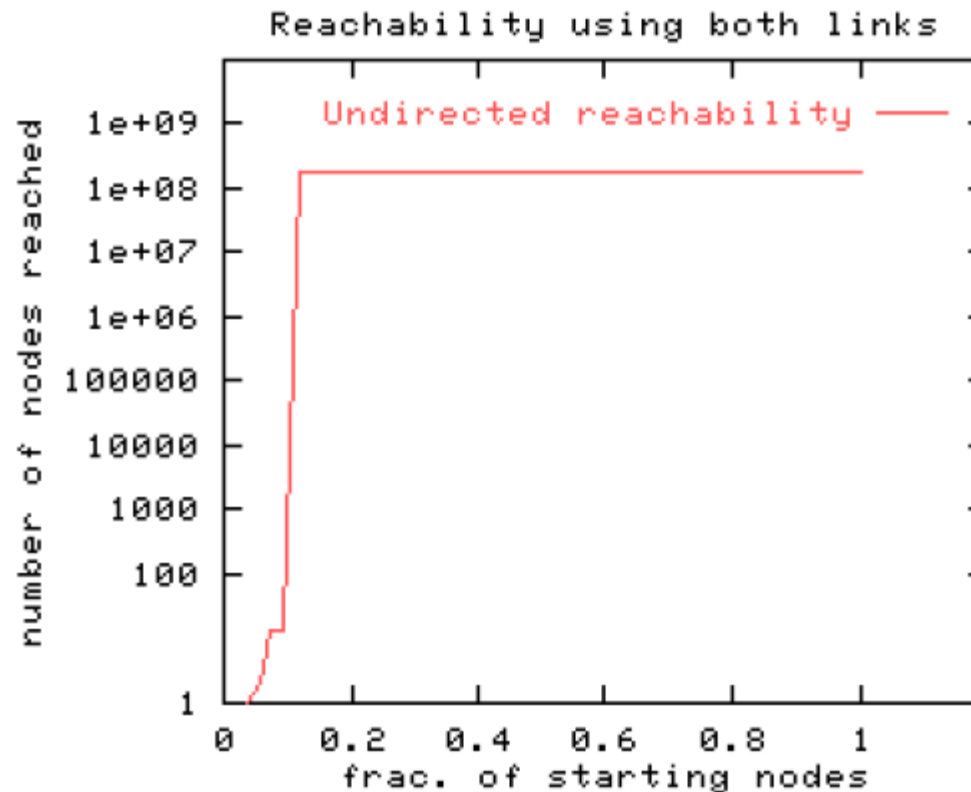


$\frac{841 \times N}{N}$

500

# Web Structure- Cnt.

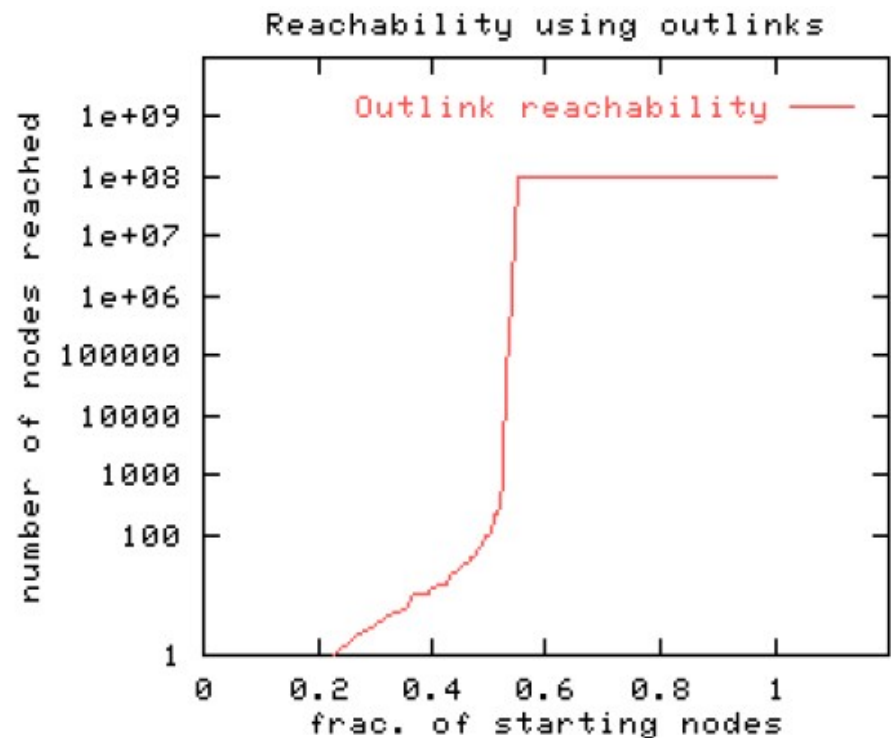
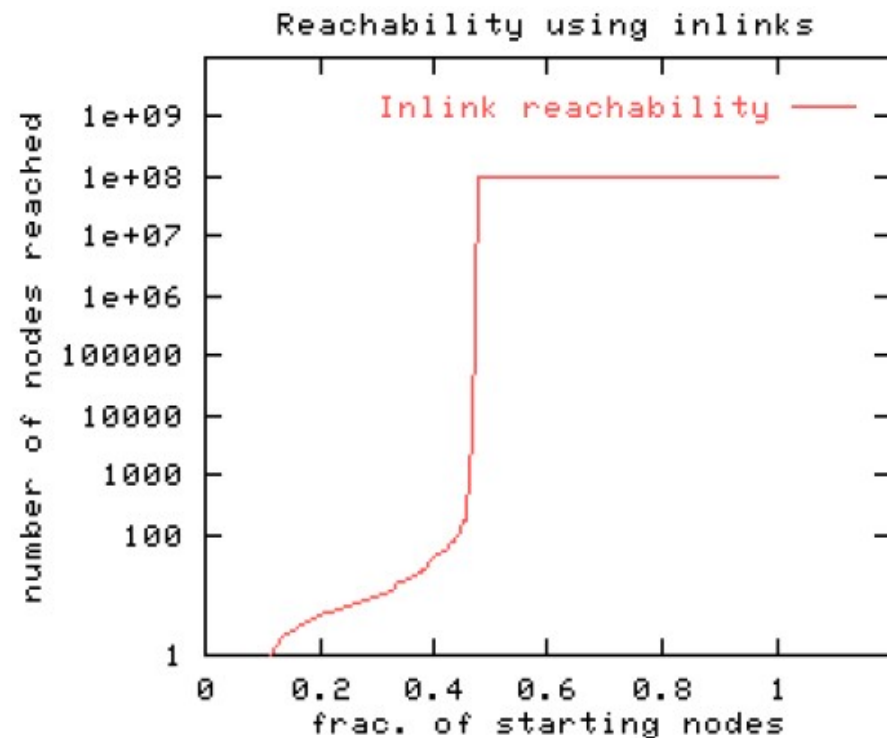
- Running BFS starting from random nodes
  - undirected.





# Web Structure- Cnt.

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



# Web Structure- Cnt.

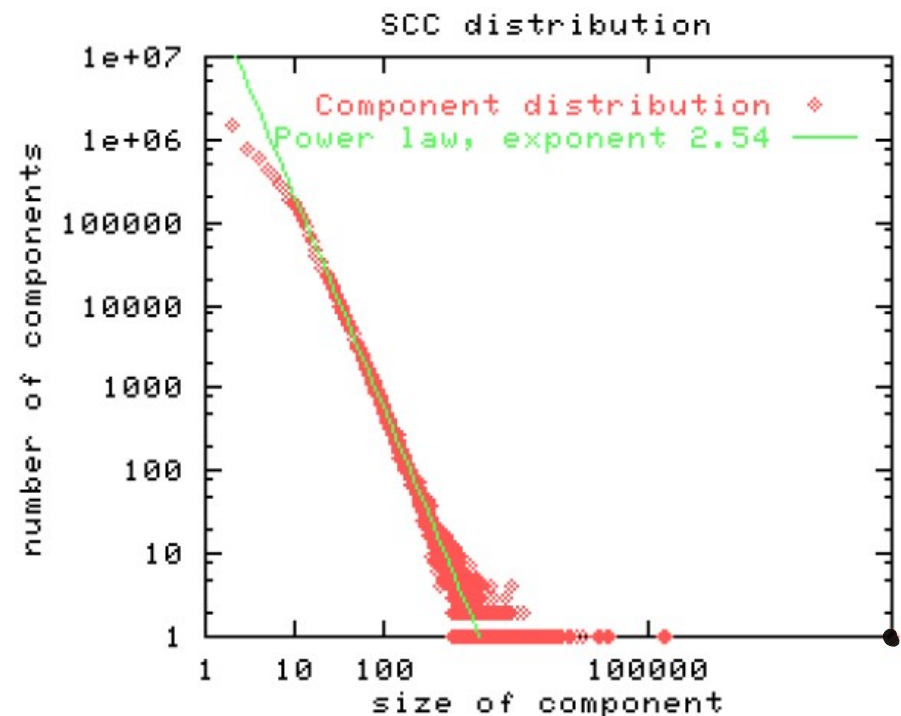
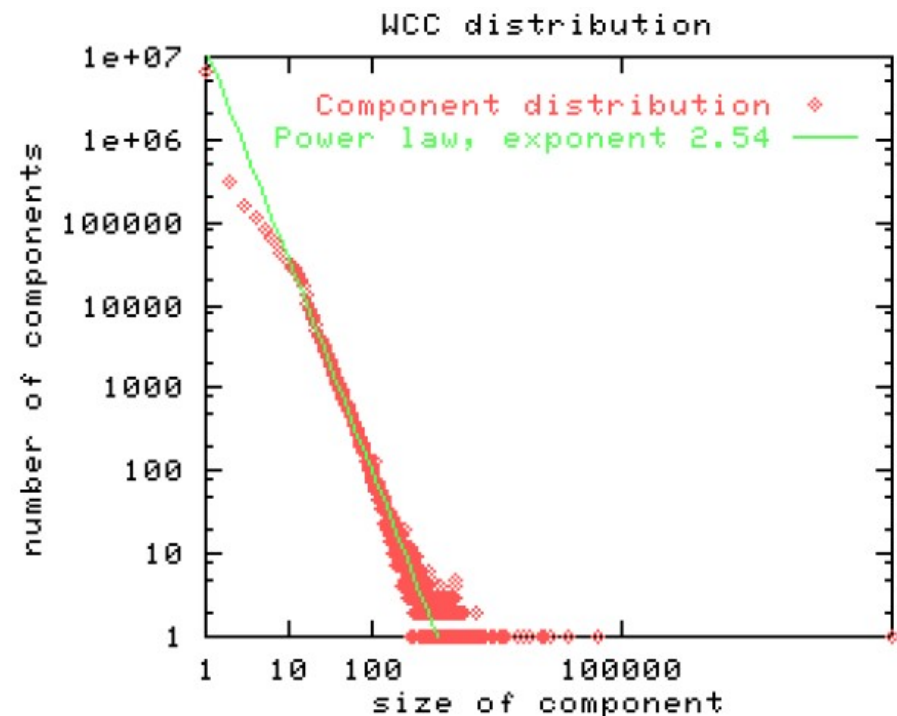
- Distribution of SCCs and WCCs on the web.
  - SCC of  $G$  is a **maximal** set of nodes  $\mathbf{C}$  such that for all  $u, v$  in  $\mathbf{C}$ , both  $u$  and  $v$  are **reachable from each other**.

# 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.

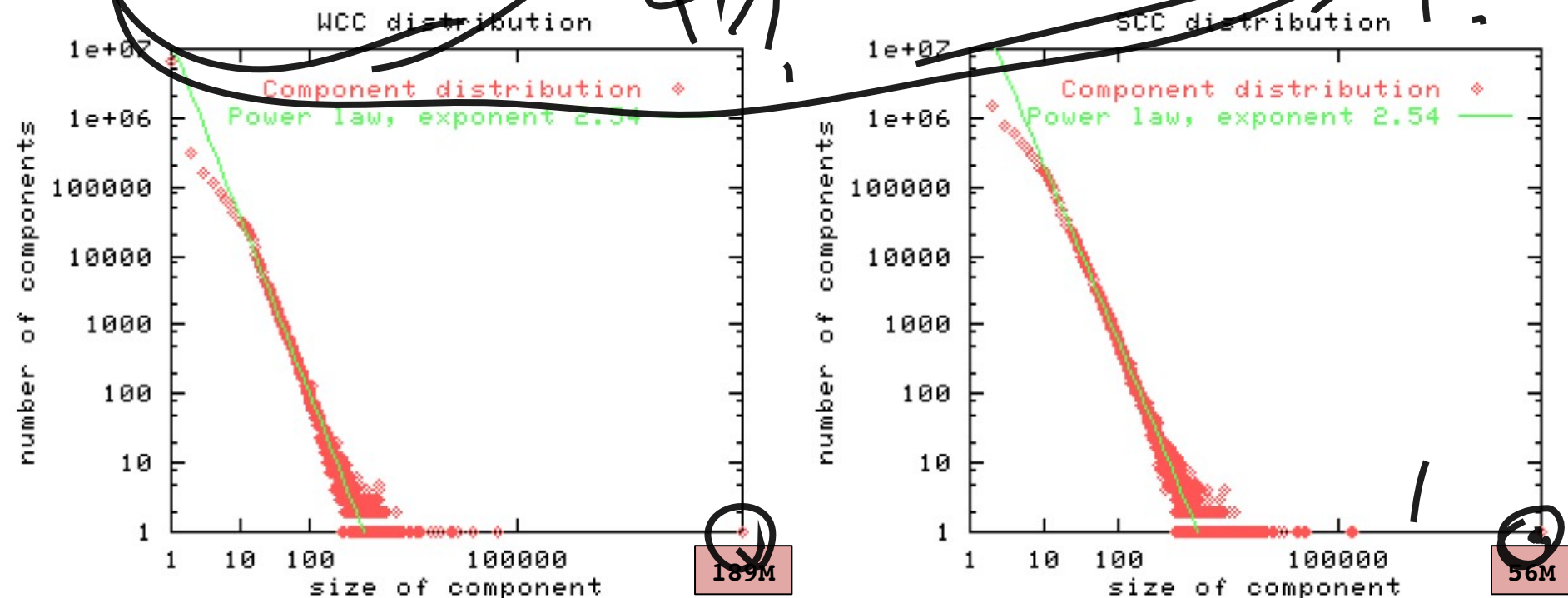


# Web Structure- Cnt.

- Distribution of SCCs and WCCs on the web.

4/17/11

16/11



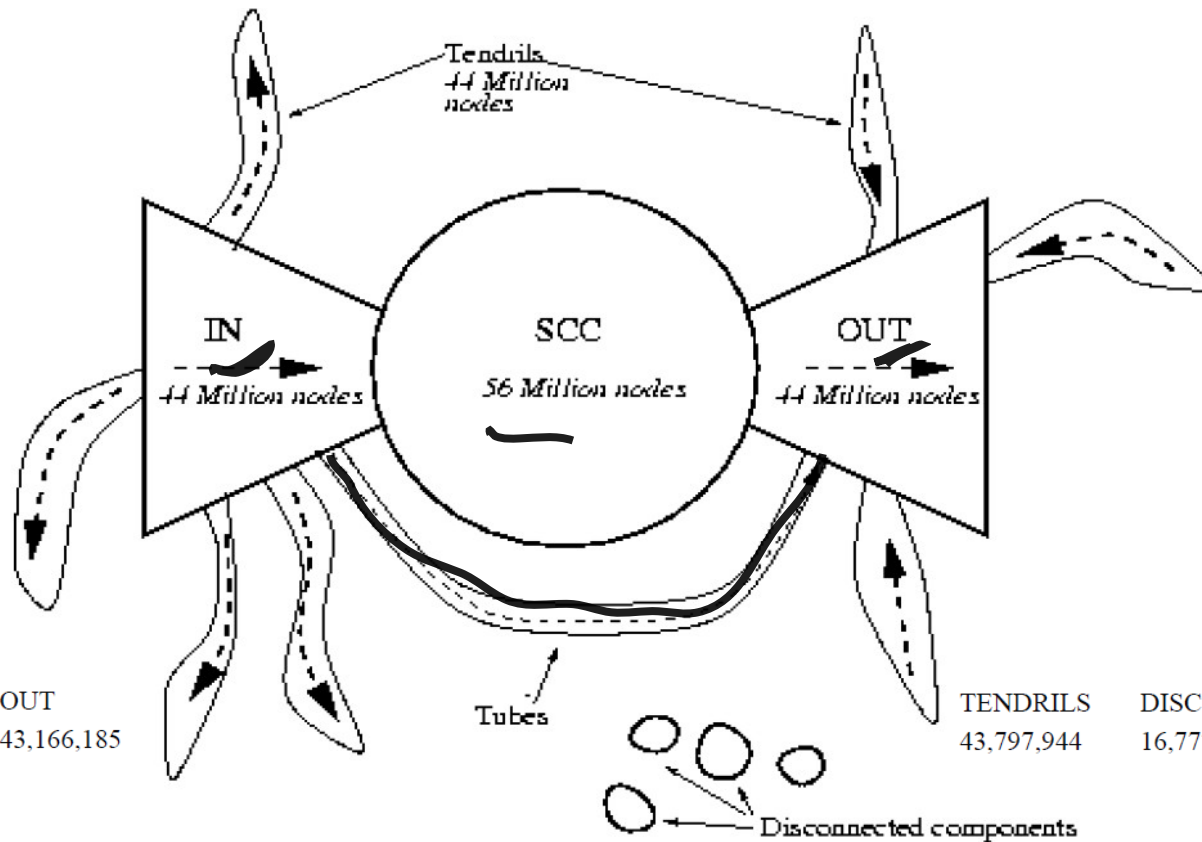
WCC: A giant component of 186 million nodes in which fully 91% of the nodes in our crawl are reachable from one another by following either forward or backward links.

# Web Structure- Cnt.

- The Web contains a 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.



SCC	IN	OUT
56,463,993	43,343,168	43,166,185

TENDRILS	DISC.	Total
43,797,944	16,777,756	203,549,046

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

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

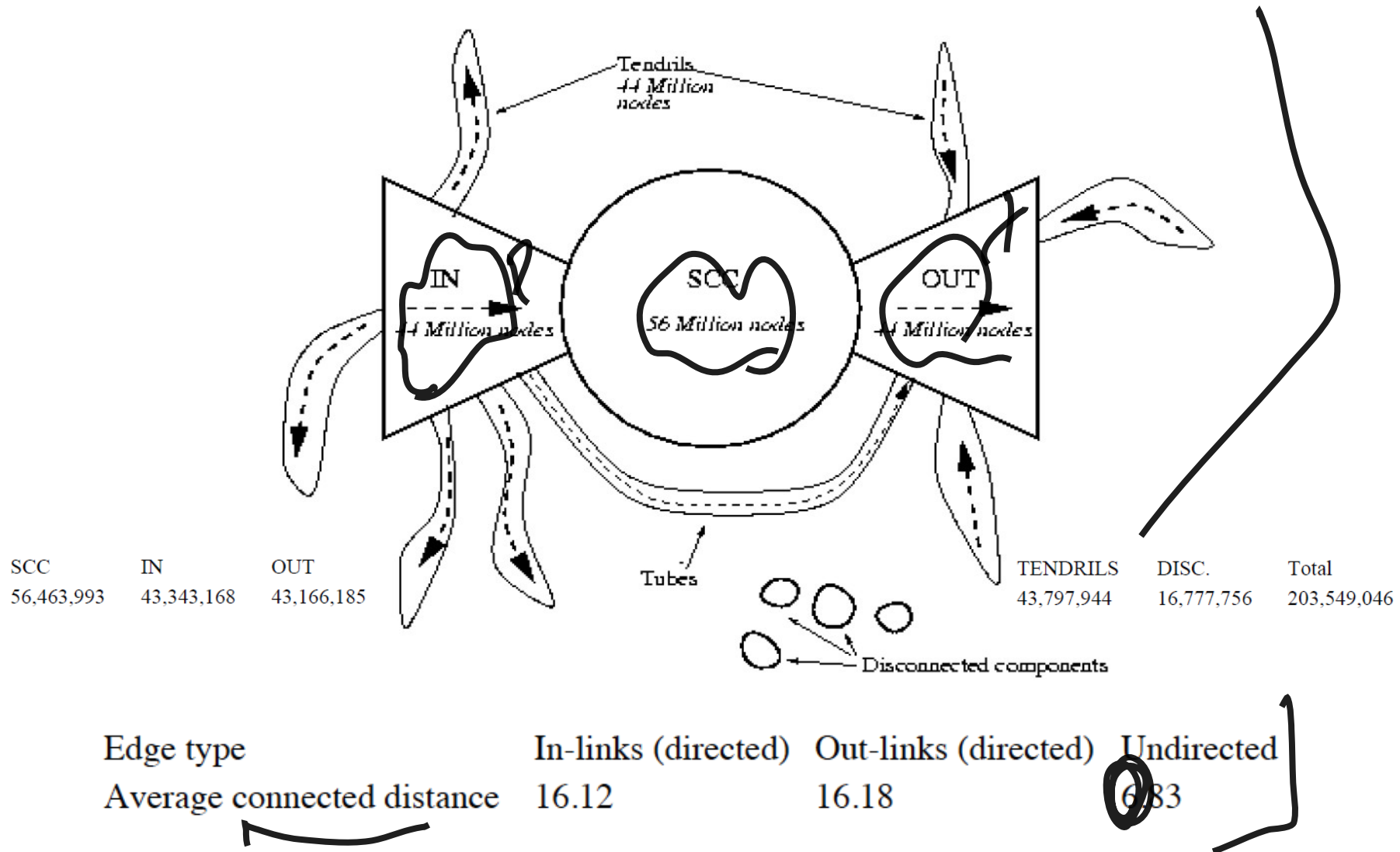
**Tendrils nodes:** (a) reachable from IN but cannot reach SCC, (b) can reach OUT but cannot be reached from SCC.

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.

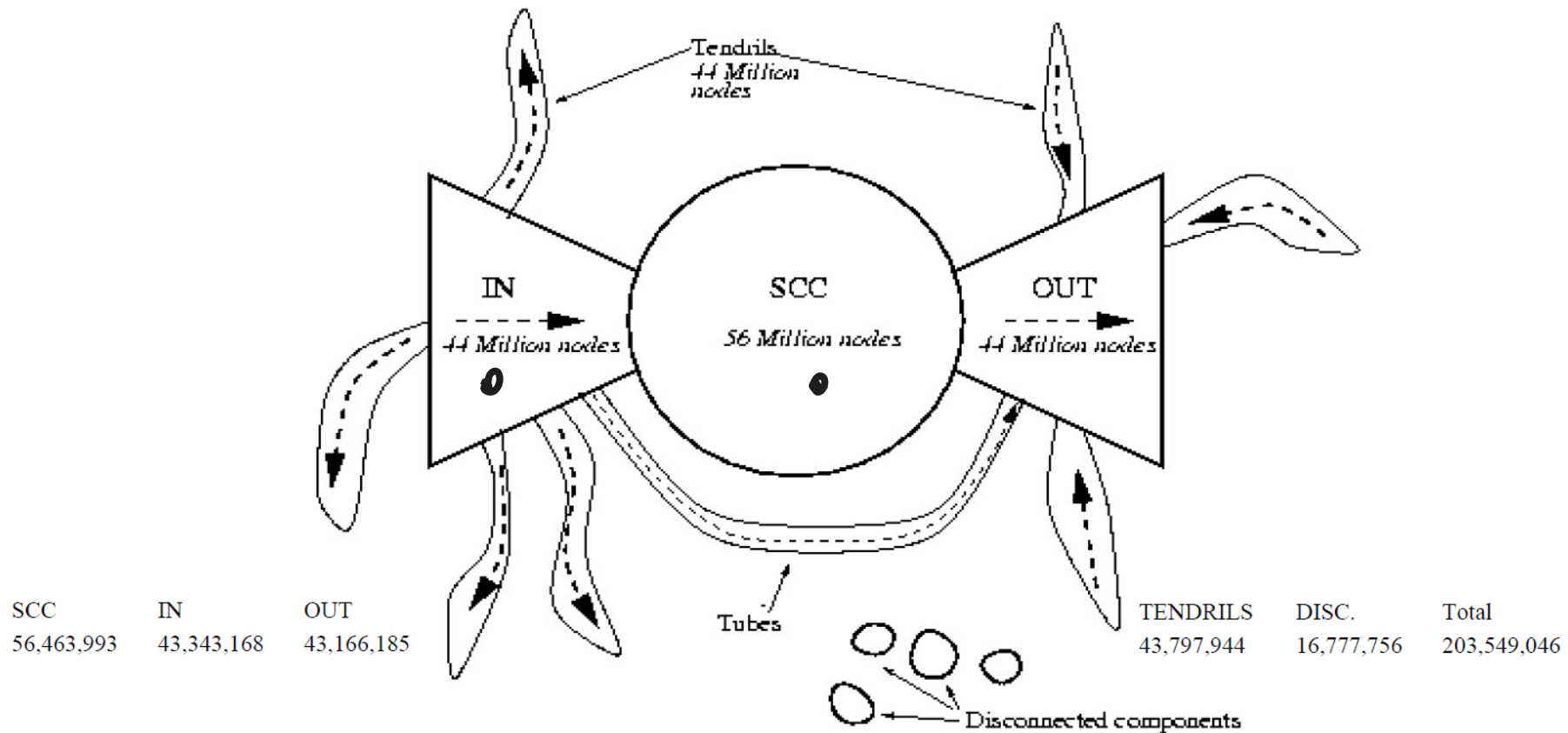
## Bow-Tie Structure of the Web.





# Web Structure- Cnt.

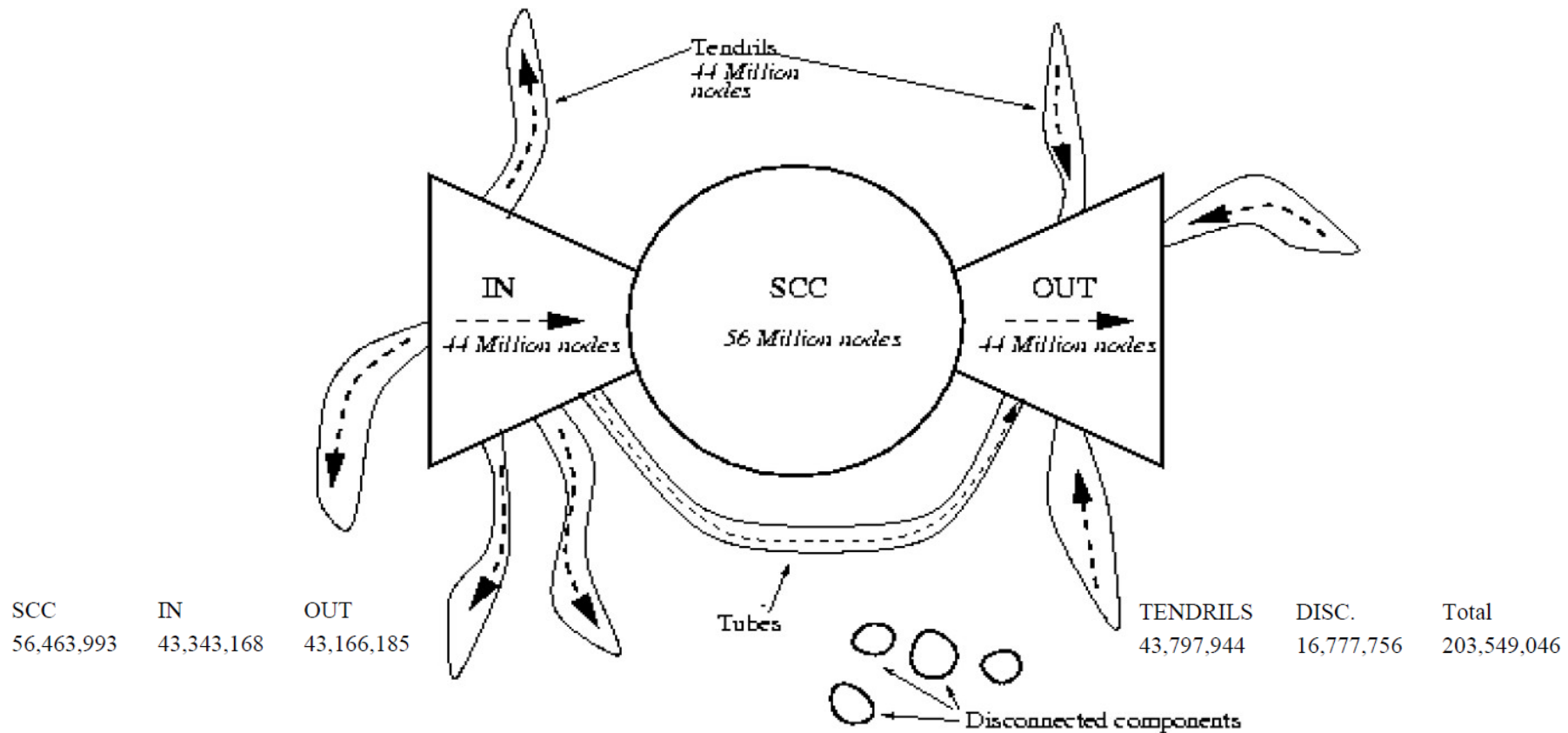
## Bow-Tie Structure of the Web.



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

# Web Structure- Cnt.

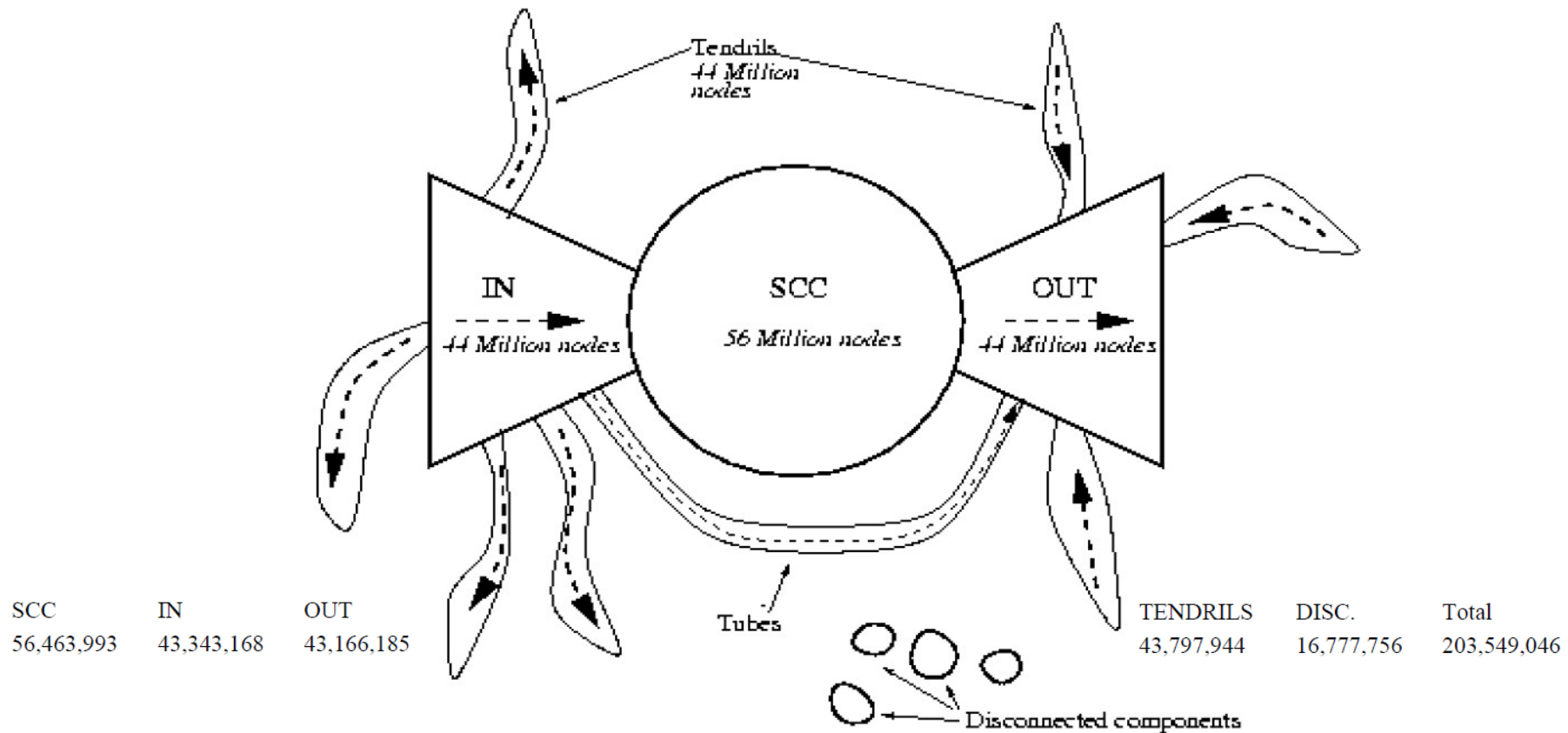
## Bow-Tie Structure of the Web.



The Web structure is relatively stable despite the fact that the constituent pieces of the bow-tie are constantly shifting their boundaries, with nodes entering (and also leaving) the giant 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.

$k$	1000	100	10	5	4	3
Size (millions)	177	167	105	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.

- Further work can be divided into three directions:
  1. Would this basic structure, and the relative fractions of the components, **remain stable over time**?
  2. Mathematical **models for evolving graphs**, motivated in part by the structure of the web.
  3. What **notions of connectivity** might be appropriate for the web graph?
    - weak and strong, co-citation relation, bibliographic coupling, etc.

# Reading

- Ch.13 The Structure of the Web [NCM]