The Graph Structure in the Web

6.886

Joana M. F. da Trindade

Feb 14th 2018

Why study the structure of the web?

Authors:

- Characterize social forces and mechanisms that explain its growth
- Devise better crawling algorithms
- Model the web’s structure with more accuracy

Me:

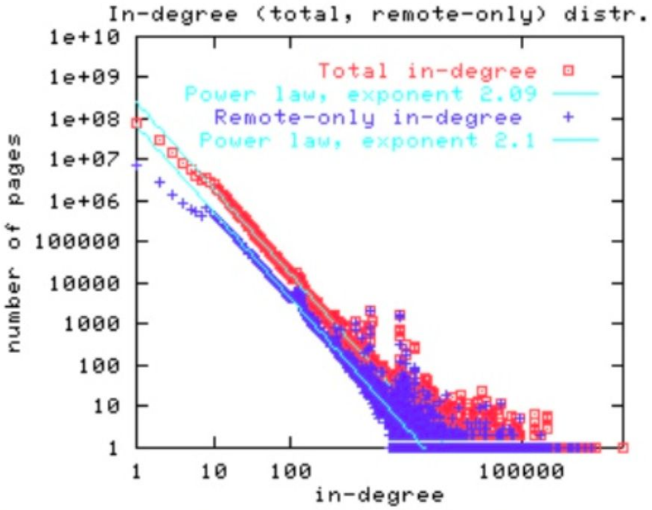
- If we can characterize what a “normal” structure looks like, that may help detect anomalies, e.g., spam-bots, fake news dissemination nets, etc

How was the web structure viewed before?

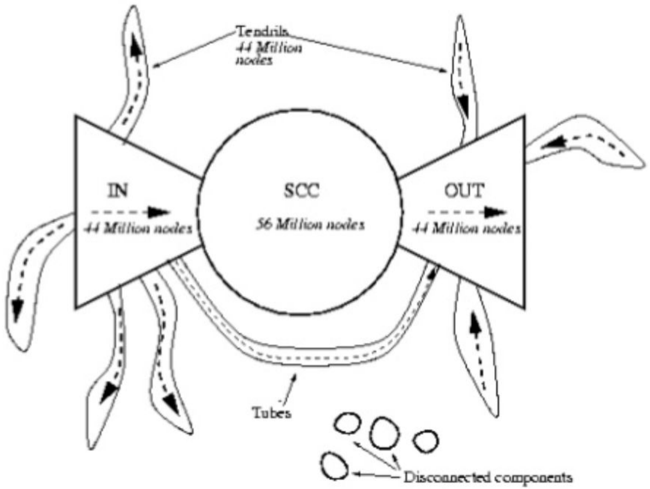
Broder et al: Graph structure in the Web (WWW2000)

- Two AltaVista crawls (200 mi pages, 1.5 links)

Power Laws



Bow-Tie



How this paper differs from previous work

Largest web structure graph studied at the time (2015)

Shows that web structure depends on crawling process

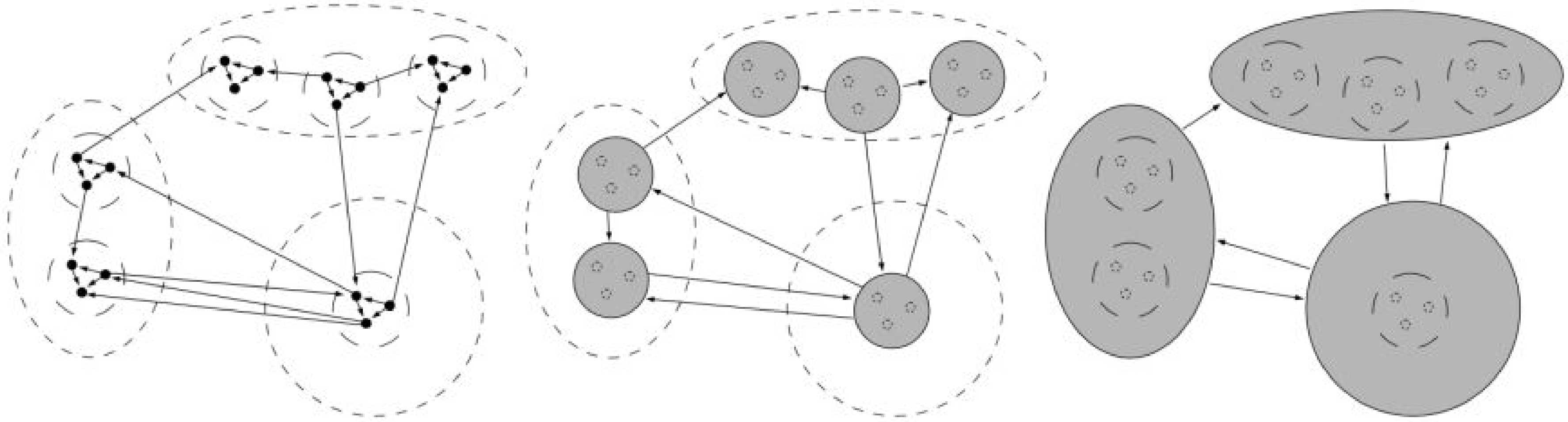
- Initial seed pages significantly affect how much of the graph is discovered, and its “bow-tie” structure

Power-law not always present in node degree distributions

- Long tails, but not necessarily power law at page level
- Power law still present at other levels of aggregation

Aggregation levels

host.domain.com/page.html　host.domain.com　domain.com



(a) page graph　(b) host graph　(c) PLD graph

Figure 1: Different aggregation levels of the graph

Aggregation levels

Graph	#Nodes	#Arcs	Size (zipped)
Page graph	3.56 billion	128.73 billion	376 GB
Subdomain graph	101 million	2,043 million	10 GB
1st level subdomain graph	95 million	1,937 million	9.5 GB
PLD graph	43 million	623 million	3.1 GB

Crawling process affects bow-tie LSCC size

AltaVista Crawl (2002)

- Size: 1.4 bi pages, LSCC is 4% of graph

ClueWeb (2009)

- Size: 1 bi pages, LSCC is 3% of graph

ClueWeb (2012)

- Size: 733 mi pages, LSCC is 76% of graph

Crawling process

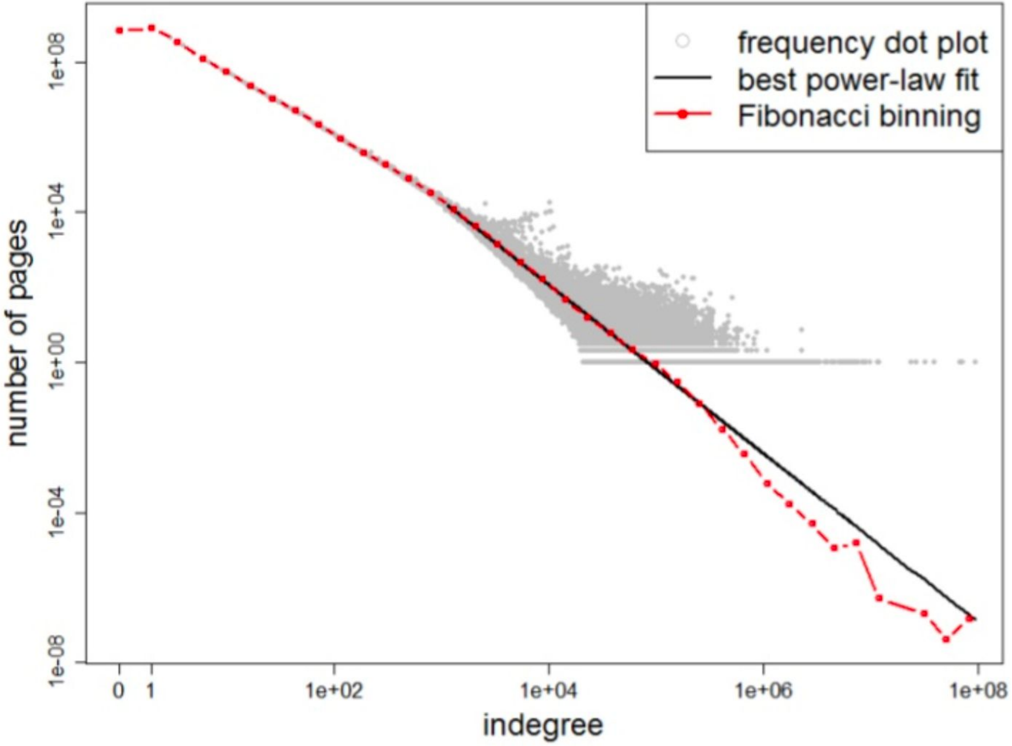
Used Common Crawl project: largest publicly available crawl at the time

- 3.5 billion pages, 128 billion links, 43 million pay-level domains (PLDs)

Crawling strategy

- Traversal: breadth-first search
- 71 million seeds from previous crawls and from Wikipedia

In-Degree Distribution



Fails goodness of fit for power-law (p-value not sufficient

Authors conclude:

- In-degree does not follow power law
- In-degree has non-fat heavy-tailed distribution
- Potentially log-normal

Divergences in average node-degree and LSCC

Previous study (Broder et al. 2000)

- Average node degree was 7.5
- Largest SCC was 27.7% of the graph

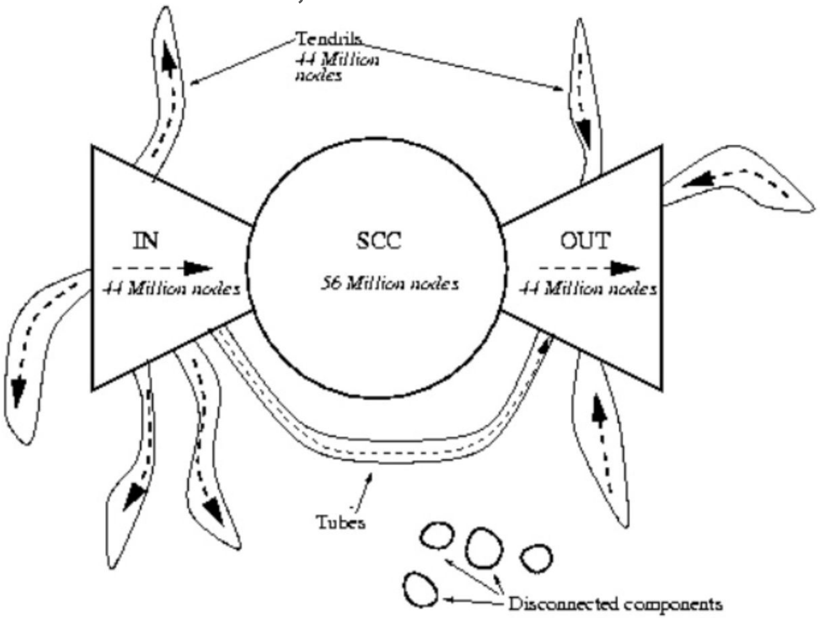
This paper

- Average node degree was 36.8
- Largest SCC was 51.8% of the graph

Divergences in average node-degree and LSCC

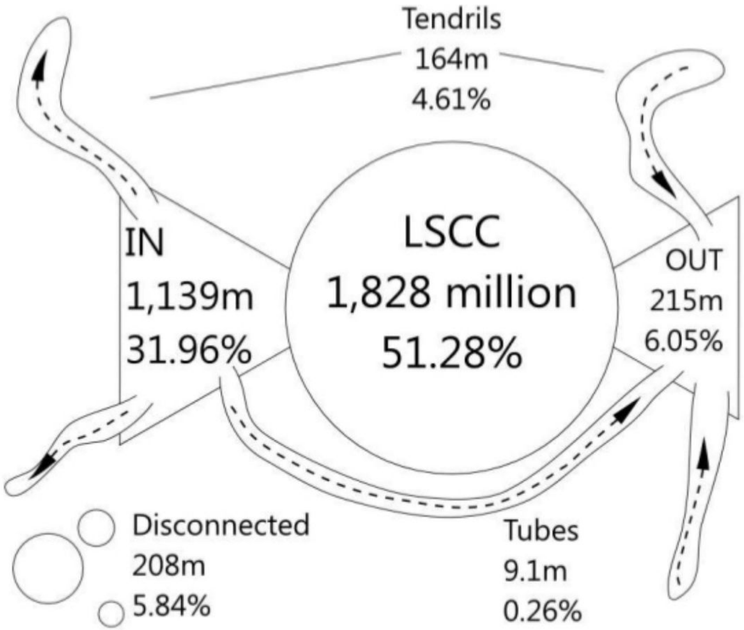
Previous study (Broder et al. 2000):

LSCC smaller; balanced IN and OUT



This paper: LSCC much larger

IN much larger than OUT



How do aggregation levels change bow-tie?

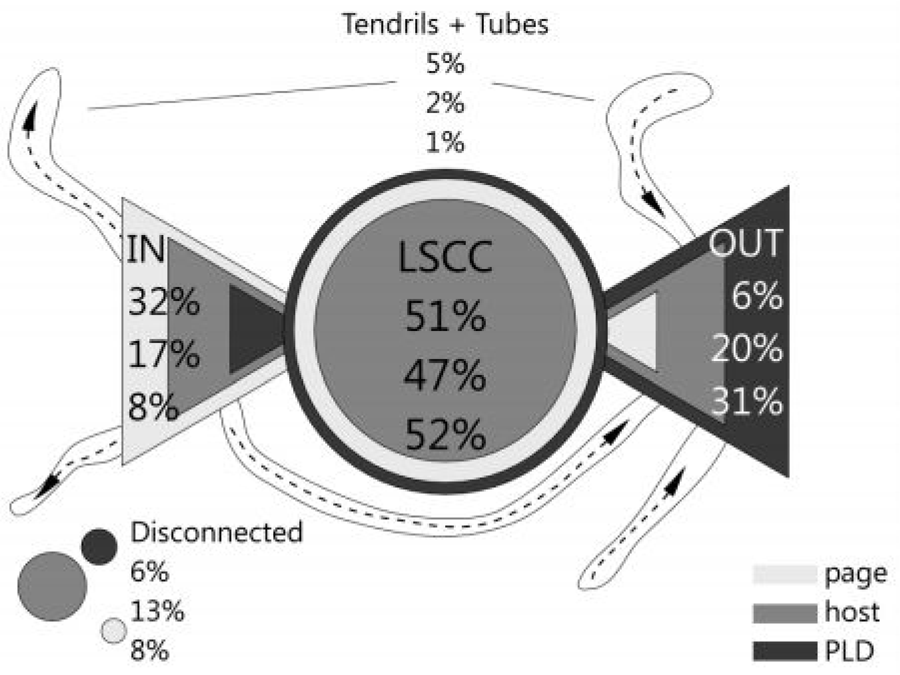


Figure 15: The bow tie on different aggregation levels

IN decreases over aggregation levels

OUT grows over aggregation levels at a similar rate

Both confirm previous findings by Zhu et al., 2008

Conclusion

Authors show that web had become much more dense and more connected:

- Much larger average degree than previous studies

Crawling process influences structure:

- Different bow-tie components proportions depending on which crawl you use

Directions for future work

- What is the actual underlying distribution of degree and components in the web -> power-law or log normal?
- More principled way to characterize web structure other than bow-tie?

Questions for discussion

This paper is almost entirely experimental: what would you change in their methodology?

Why is the LSCC important? Any other ways to characterize the web other than bow-tie?

Crawling process seems to introduce biased sampling not easy to characterize: how would different graph sampling strategies affect the observed graph structure?

Why can the authors fit a power-law only at PLD level of aggregation?