Graph Structure in the Web — Revisited

or A Trick of the Heavy Tail

Author: Robert Meusel, Sebastiano Vigna, Oliver Lehmberg, Christian Bizer

September 30, 2017

Overall

- Crawled in 2012
- Containing 3.5 billion web pages and 128.7 billion links
- Analyzed features of the Web graph, including
  - degrees (indegree, outdegree)
  - components (weekly connected, strongly connected)
  - diameter and distances

Overall

Intuition

It is natural to treat the web as graph.

- Nodes correspond to static pages on the Web
- Arcs correspond to links between pages

Overall

Intuition

It is natural to treat the web as graph.

- Nodes correspond to static pages on the Web
- Arcs correspond to links between pages

By studying Web graph, we can

- design crawl strategies on the web
- improve PageRank algorithms
- understand the sociology of content creation on the web
- predict the evolution of the web

Bow-Tie Structure

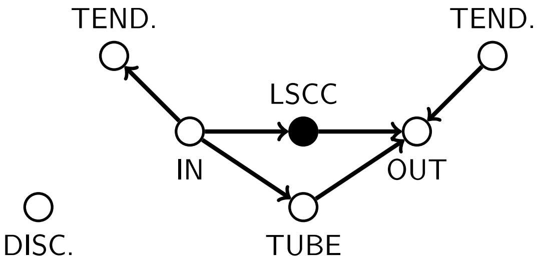
Components of the web graph

Several studies confirm the existence of a large strongly connected components, significantly larger than any other components.

Bow-Tie Structure

Components of the web graph

Several studies confirm the existence of a large strongly connected components, significantly larger than any other components.



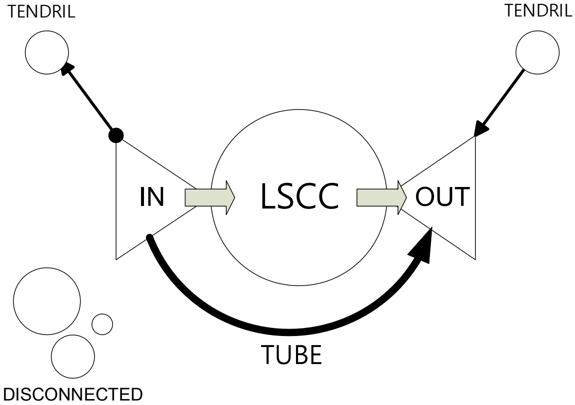
Bow-Tie Structure

Components of Bow-Tie Structure

- LSCC: large strongly connected components
- IN: nodes that can reach LSCC
- OUT: nodes that can be reached from LSCC
- TENDRILS: nodes that can either be reached from IN, or can reach OUT
- TUBES: nodes that lie on paths from IN to OUT, without passing LSCC
- DISCONNECTED: nodes that are not weakly connected to LSCC

Bow-Tie Structure

A Typical Bow-Tie Structure



Bow-Tie Structure

Comparison of Sizes of Bow-Tie Components

	Common Crawl 2012		Broder *et al.* (2000)	
Component	# nodes (k)	% nodes	# nodes (k)	% nodes
LSCC	1 827 543	51.28	56 464	27.74
IN	1 138 869	31.96	43 343	21.29
OUT	215 409	6.05	43 166	21.21
TENDRILS	164 465	4.61	43 798	21.52
TUBES	9 099	0.26	-	-
DISC.	208 217	5.84	16 778	8.24

Table: Comparison of sizes of bow-tie components

Bow-Tie Structure

Phenomena & Analysis

1 The size of LSCC has almost doubled.

Bow-Tie Structure

Phenomena & Analysis

1. The size of LSCC has almost doubled.
   - The web has become more dense and connected.

Bow-Tie Structure

Phenomena & Analysis

1. The size of LSCC has almost doubled.
   - The web has become more dense and connected.
2. The IN component has become much larger than OUT component in size.

Bow-Tie Structure

Phenomena & Analysis

1. The size of LSCC has almost doubled.
   - The web has become more dense and connected.
2. The IN component has become much larger than OUT component in size.
   - Crawl methodology (esp. crawl seeds)
   - Small websites?

Bow-Tie Structure

Comparison between Page Graph and PLD Graph

	page graph		PLD graph	
Component	# nodes (M)	% nodes	# nodes (M)	% nodes
LSCC	1 828	51.28	22.3	51.94
IN	1 139	31.96	3.3	7.65
OUT	215	6.05	13.3	30.98
TENDRILS	164	4.61	0.5	1.20
TUBES	9	0.26	0.2	0.04
DISC.	208	5.84	3.5	8.20

Table: Comparison between Page Graph and PLD Graph

Degree Distribution

the Power Law

Broder *et al.* claimed that the degree distribution follows the power law, both indegree and outdegree.

Degree Distribution

the Power Law

Broder *et al.* claimed that the degree distribution follows the power law, both indegree and outdegree.

Power Law

$$f(x) = ax^{-k}, k > 1$$

Degree Distribution

the Power Law

Broder *et al.* claimed that the degree distribution follows the power law, both indegree and outdegree.

Power Law

$$f(x) = ax^{-k}, k > 1$$

- Scale invariance: $f(cx) = a(cx)^{-k} = c^{-k} f(x) \propto f(x)$

Degree Distribution

the Power Law

Broder *et al.* claimed that the degree distribution follows the power law, both indegree and outdegree.

Power Law

$$f(x) = ax^{-k}, k > 1$$

- Scale invariance: $f(cx) = a(cx)^{-k} = c^{-k} f(x) \propto f(x)$
- Heavy-tailed, 80–20 rule

Degree Distribution

the Power Law

Broder *et al.* claimed that the degree distribution follows the power law, both indegree and outdegree.

Power Law

$$f(x) = ax^{-k}, k > 1$$

- Scale invariance: $f(cx) = a(cx)^{-k} = c^{-k}f(x) \propto f(x)$
- Heavy-tailed, 80–20 rule
- Well-defined mean exists only when $k > 2$

Degree Distribution

the Power Law

Broder *et al.* claimed that the degree distribution follows the power law, both indegree and outdegree.

Power Law

$$f(x) = ax^{-k}, k > 1$$

- Scale invariance: $f(cx) = a(cx)^{-k} = c^{-k}f(x) \propto f(x)$
- Heavy-tailed, 80–20 rule
- Well-defined mean exists only when $k > 2$
- Linear in log-log plot

Degree Distribution

Why not Power Law?

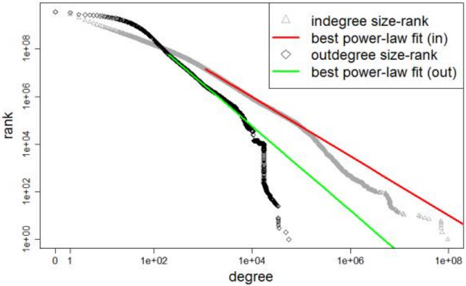


Figure: Log-log plot of degree distributions

Degree Distribution

Why not Power Law?

Problems:

Degree Distribution

Why not Power Law?

Problems:

- The conclusion was drawn just by the approximate linear shape in log-log plot.

Degree Distribution

Why not Power Law?

Problems:

- The conclusion was drawn just by the approximate linear shape in log-log plot.
- The concavity in the left part cannot be explained.
  - There are not so much pages with few hyperlinks as expected.

Degree Distribution

Why not Power Law?

Problems:

- The conclusion was drawn just by the approximate linear shape in log-log plot.
- The concavity in the left part cannot be explained.
  - There are not so much pages with few hyperlinks as expected.
- The data points in the right part deviate the line.
  - The number of pages with huge number of hyperlinks decreases rapidly as the number of links increases. (hyperpolynomial decrease)

Degree Distribution

Other Conclusions

In fact, indegree distribution fits the power law better than outdegree.

Degree Distribution

Other Conclusions

In fact, indegree distribution fits the power law better than outdegree.

- The concavity is more obvious in indegree distribution.

Degree Distribution

Other Conclusions

In fact, indegree distribution fits the power law better than outdegree.

- The concavity is more obvious in indegree distribution.
- The indegree distribution curve drops much faster than outdegree , when degree grows large.

Degree Distribution

Other Conclusions

In fact, indegree distribution fits the power law better than outdegree.

- The concavity is more obvious in indegree distribution.
- The indegree distribution curve drops much faster than outdegree , when degree grows large.

Degree Distribution

Other Conclusions

In fact, indegree distribution fits the power law better than outdegree.

- The concavity is more obvious in indegree distribution.
- The indegree distribution curve drops much faster than outdegree , when degree grows large.

Why?

Degree Distribution

Other Conclusions

In fact, indegree distribution fits the power law better than outdegree.

- The concavity is more obvious in indegree distribution.
- The indegree distribution curve drops much faster than outdegree , when degree grows large.

Why?

- technical limitations

Degree Distribution

Other Conclusions

In fact, indegree distribution fits the power law better than outdegree.

- The concavity is more obvious in indegree distribution.
- The indegree distribution curve drops much faster than outdegree , when degree grows large.

Why?

- technical limitations
- although the average degree has significantly increased by 5

Diameter and Distances

- The average distance is 12.84.

Diameter and Distances

- The average distance is 12.84.
- The harmonic diameter is 24.43.

Diameter and Distances

- The average distance is 12.84.
- The harmonic diameter is 24.43.
- The average distance was 16.12 in 2000, reported by Broter *et al.*

Diameter and Distances

- The average distance is 12.84.
- The harmonic diameter is 24.43.
- The average distance was 16.12 in 2000, reported by Broter *et al.*
- "Small-world network"

References

[1] R. Meusel, S. Vigna, O. Lehmberg, and C. Bizer. Graph structure in the Web — Revisited: A trick of the heavy tail. *Proceedings of WWW Companion '14*, 427–432, 2014.

[2] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener. Graph structure in the Web: experiments and models. *Computer Networks*, 33(1–6):309-320, 2000.

[3] O. Lehmberg, R. Meusel and C. Bizer. Graph Structure in the Web — Aggregated by Pay-Level Domain. *Proceedings of the 1024 ACM conference on Web Science*, 119–128, 2014.

[4] D. Donato, S. Leonardi, S. Millozzi, and P. Tsaparas. Mining the inner structure of the web graph. *WebDB*, 145-150, 2005.