A Graph based Methodology for Web Structure Mining - with a Case Study on the Webs of UK Universities

Tahani Alqurashi and Wenjia Wang
School of Computing Sciences,
University of East Anglia,
Norwich, UK

Outline

1. Introduction.
2. Web as a Graph.
3. Graph-based Web Structure Mining Framework.
4. A Case Study on UK Universities Webs.
5. Discussion and Conclusion.

1. Introduction: Background

- World wide webs have become a must-have thing for an organisation in public, private and business sections.
- Different domains have their different nature of their business or service and that is how the URLs are organised, such as,
  - Education: ac (for UK). edu. (for North America)
  - Companies: co. com.
  - Governments: gov.
  - General organisations: org. etc.

Introduction: Research Questions

- Questions:
  - What kind of web structure would be more suitable for a particular domain to serve their business better?
  - What the structures of current websites look like and where they could be improved?
- This paper presents a graph-based framework for mining web structures and takes UK university websites as a case study to demonstrate the concept.

Related Work

- Some tools for web structure analysis:
  - Webometrics
  - Pajek (Btageli and Mrvar 2001)
  - Visual Analytics(Analytics 2011)
- Websites in the UK academic domain have been studied through Webometrics and Web Impact Factors, in 2001, focusing on website performance and health.

2. Graph of Web Structure

- The structure of a website can be naturally represented by a graph *G*
  - web pages => graph nodes
  - hyperlinks between pages and other websites => links
- Evaluate a given graph in two levels:
  - Global level: external structures and
  - Local level: internal structures.

Web Structure Graph

- Given a graph *G(N, L) contains N* nodes and *L* links, we want to know:
  - What is the distribution of in-degrees and out-degrees?
  - What is its connectivity structure?
  - What is the diameter of the Web?
- These structural properties can be measured by:
  1. Size and density of G: total number of nodes, average # of links
  2. The degree of each node in G,
  3. The path between the nodes in G,
  4. The size of the giant connective component (GCC),
  5. The cluster coefficient (CC),
  6. The closeness centralisation (C) of G.

The degree of a node

- The degree of the node represents a measure of the node activity in the graph.
  - It is actually the number of links, or neighbours, that a particular node has.
- The average degree for the nodes provides an informative summarization of the graph.
  - It indicates the average number of the links pointing to or from a node in the graph.

The path in the graph

- A path is a sequence of connected nodes (web pages) in a graph,
  - indicates a possible navigation route between pages.
- The shortest path between any two nodes represents:
  - the geodesic distance or the optimal path.
  - the minimum number of links users needed to click to navigate the site easily.

Giant Connective Components

- GCC: pages are heavily linked to each other.
- There are four different types of GCC in web structures:
  - Strongly connected component,
  - IN component,
  - OUT component
  - Tendrils,
- Analysing the GCCs in the web graph aided in assessing navigability, and in identifying the parts of the website that increased the ease of navigation.

Four Types of GCC

- The strongly connected component (SCC) refers to those pages (nodes) that are bi-directly inter-connected.
  - Usually represents central core pages of a web
  - allows users to navigate more easily than in other components because it provides a link back to the core of the website.
- The IN components are the pages with direct links into the SCC but not back.
- The OUT refers to the pages with links from the SCC but not into it.
- The Tendrils are the pages without link to or from the SCC at all.

Four Types of GCC

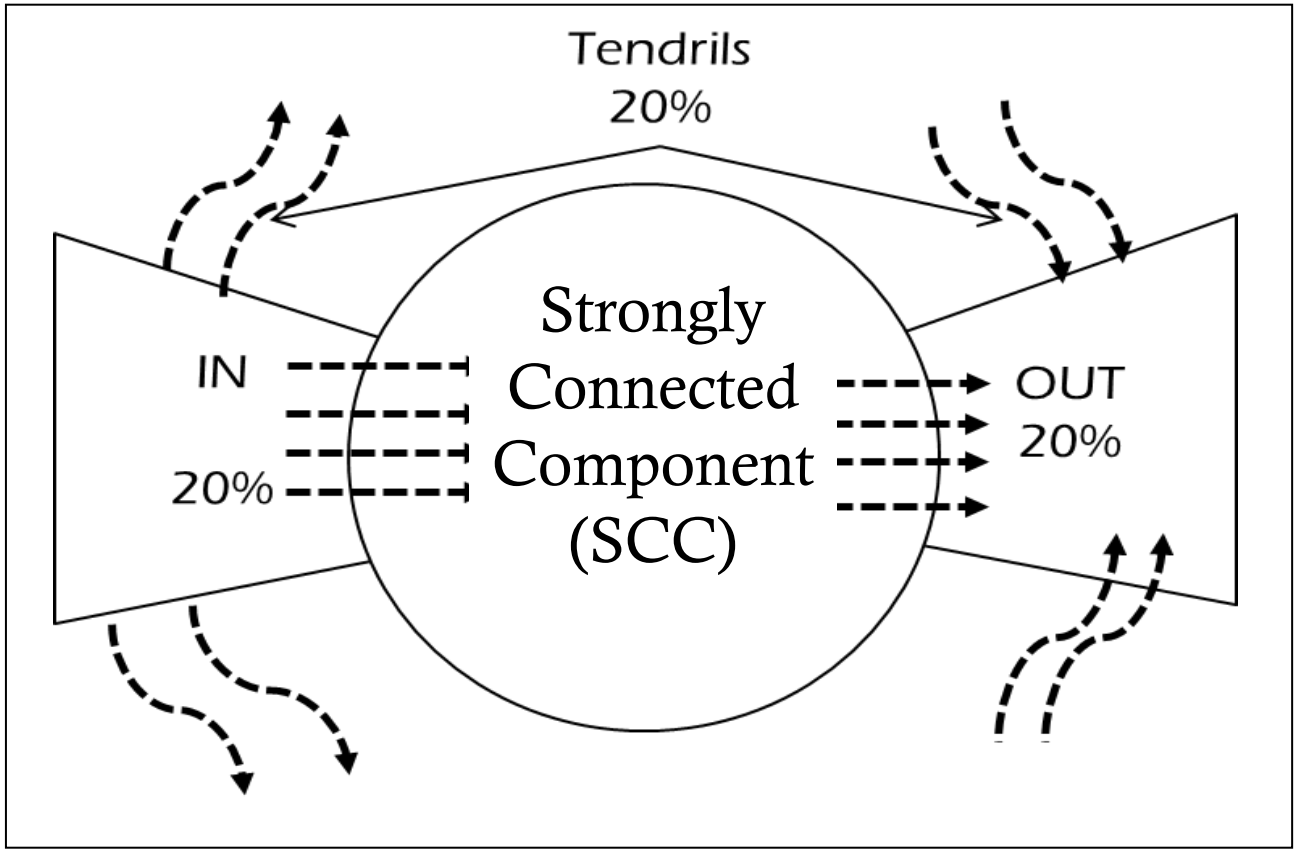


Figure1: The bow tie structure of the Web.

The Cluster Coefficient (CC)

- The cluster coefficient (CC) measures how closely the nearest neighbours to a node are interconnected.
  - for a node *e* that has *k* nearest neighbours with *s* links between them, it is defined as follows:

$$CC(e) = \frac{s}{k(k - 1)}$$

Average cluster coefficient (ACC)

- The average cluster coefficient(ACC) measures the strength of the graph cohesion and the densities of the neighbourhood of all the nodes.

- It reveals the overall local structures of the graphs, which then facilitates comparisons between the different graphs.

The Network closeness centralisation (C)

- Closeness Centralisation is a measure of the distance between the nodes according to 'closeness'.
  - The closeness centrality of node *e* is calculated:

$$C(e) = \frac{1}{\sum_{i \in N} D(e,i)}$$

Where D is the shortest distance between nodes *e* and *i*.

3. Graph-based Web Structure Mining Framework

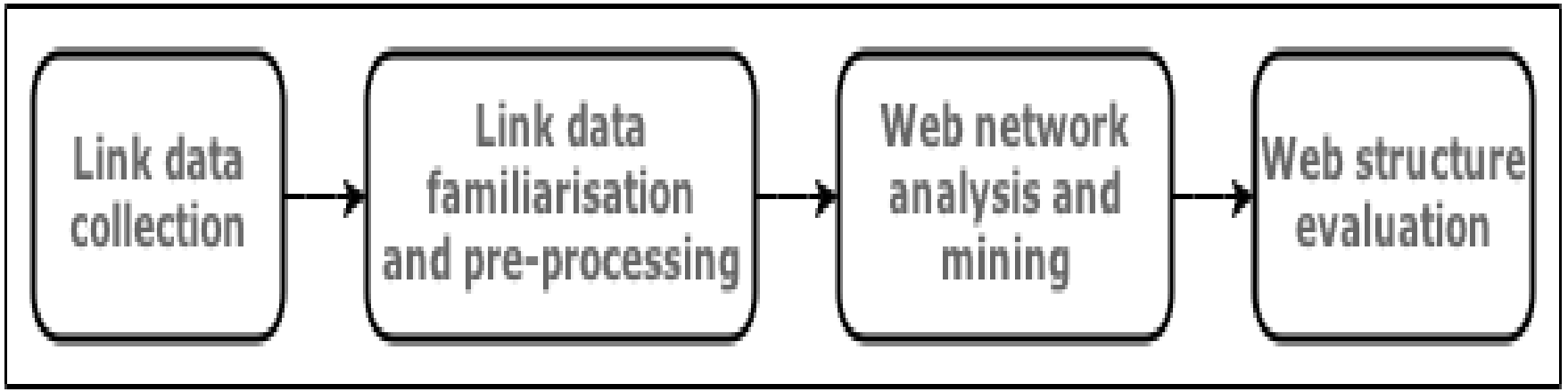


Figure 2: Web structure mining roadmap.

Link Data Collection

The first stage is to gather the structural data of websites, following the steps shown in Figure 3.

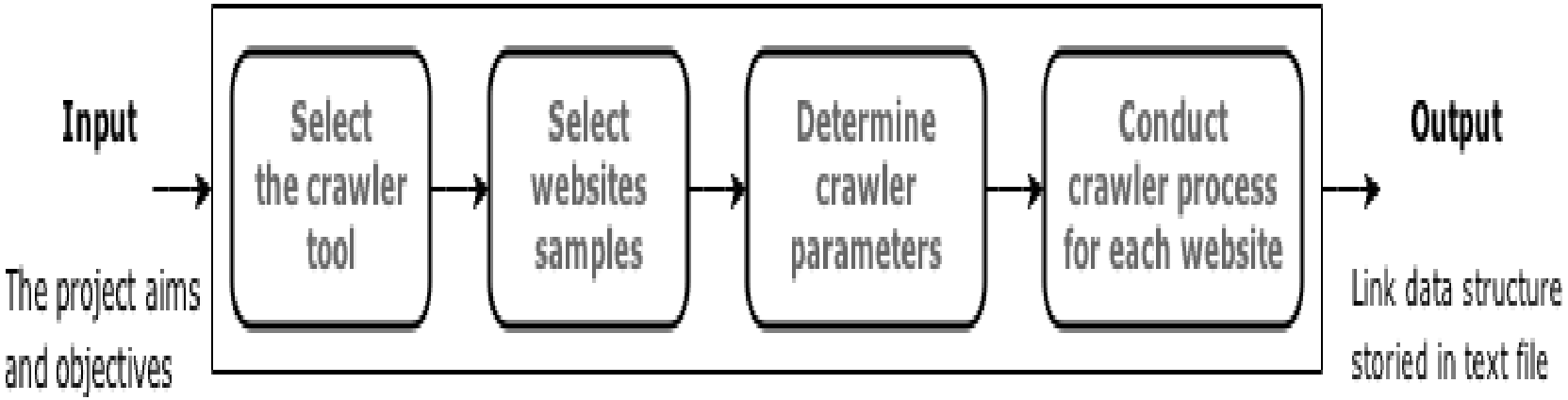


Figure 3: Link data collection.

Link Data Familiarisation

- It aims to understand the nature of the collected data statistically in order to decide how we carry out the subsequent phases.
- A number of characteristics need to be considered for each website:
  - The number of pages that have been collected for each website by the crawler.
  - The total number of URL links on each website.
  - The exploration of all out-links on each website.

Data Cleansing

- Duplicated data.
  - Two different types of duplicated data can exist in the link data: pages with identical content with different URL addresses, and duplicated hyperlinks (at the page or whole-site level).
- Undesirable data:
  - Typically, web page documents with file extensions, such as html, asp, jsp and xml.
  - Other file documents with different extensions, such as pdf, WS and audio documents.

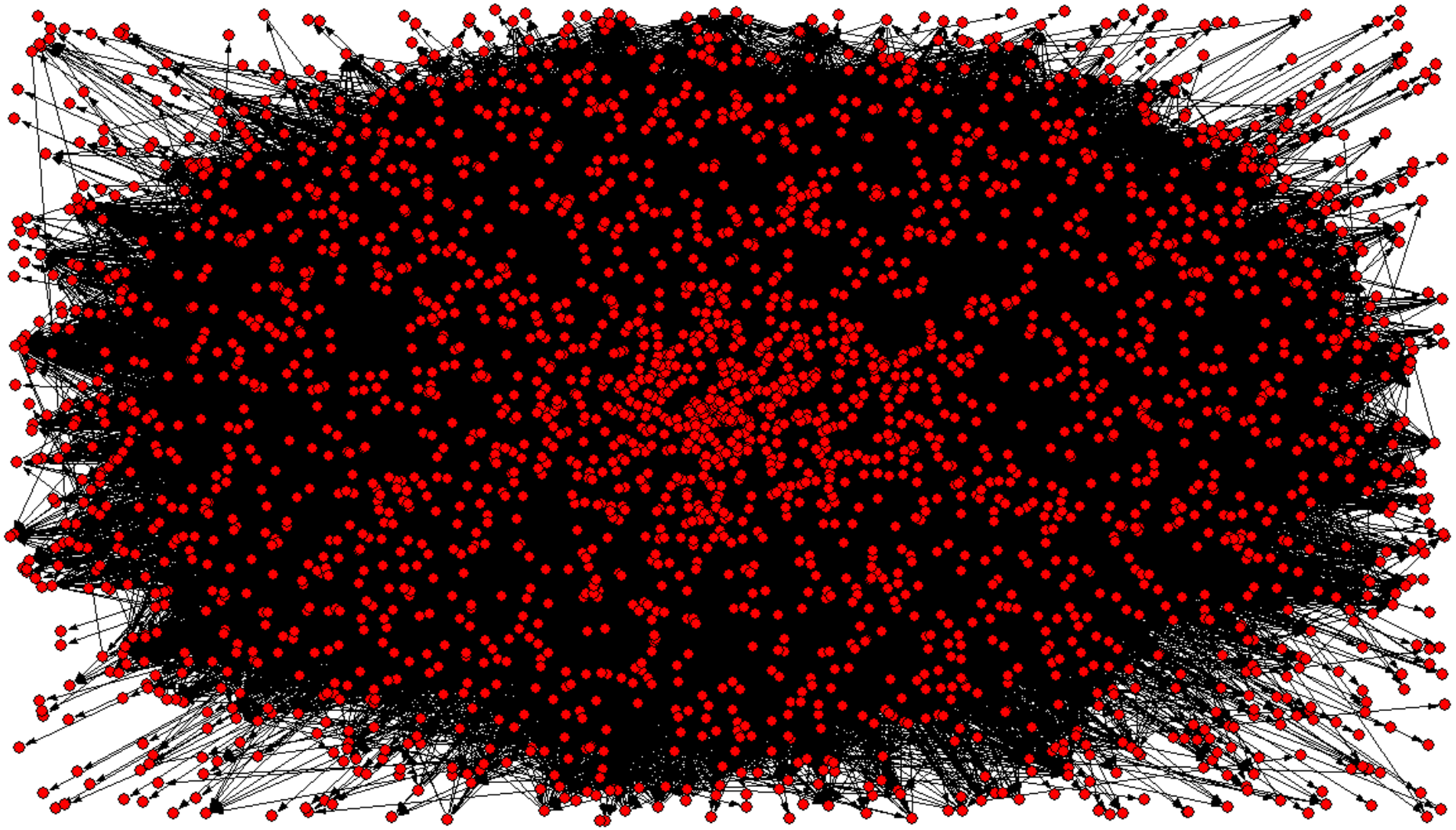
Web Graph Mining

- Graphs of webs are ready for mining
- Measure properties of web structures
  1. Web density
  2. Distribution of IN and OUT links
  3. ACC
  4. C
  5. Etc.

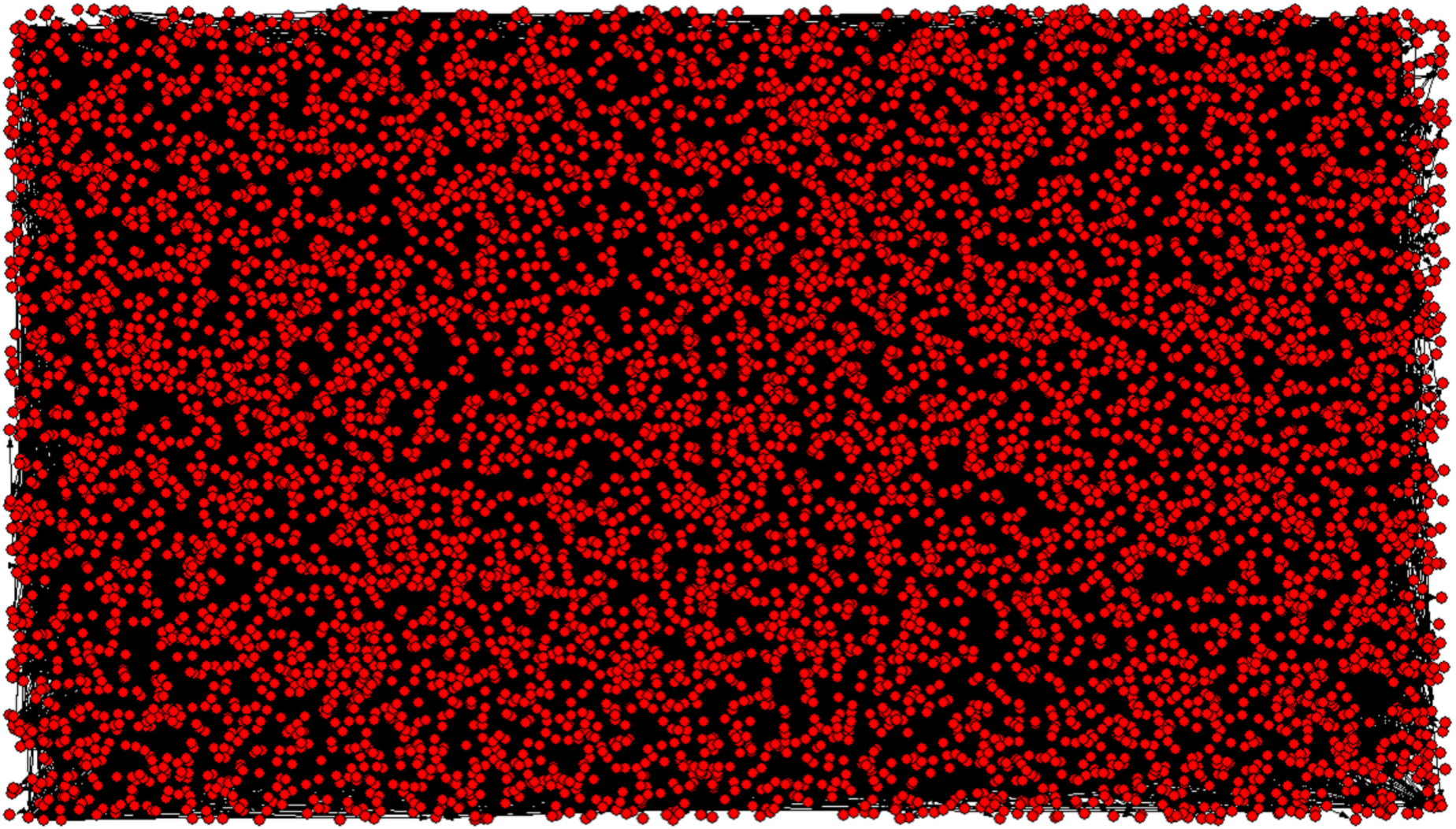
4. A Case Study on UK Universities Webs.

- 110 universities in the UK websites were selected in this study.
  - The Russell group: 16 websites chosen,
  - The 1994 group: 18 websites chosen
  - The Universities Alliance: 21 websites
  - The Million Plus group: 28 websites.
  - Unaffiliated: 27 websites
- The data were collected in 2012 with web crawler SocSciBot4

Web Internal Structure: Example



Web Internal Structure: UEA

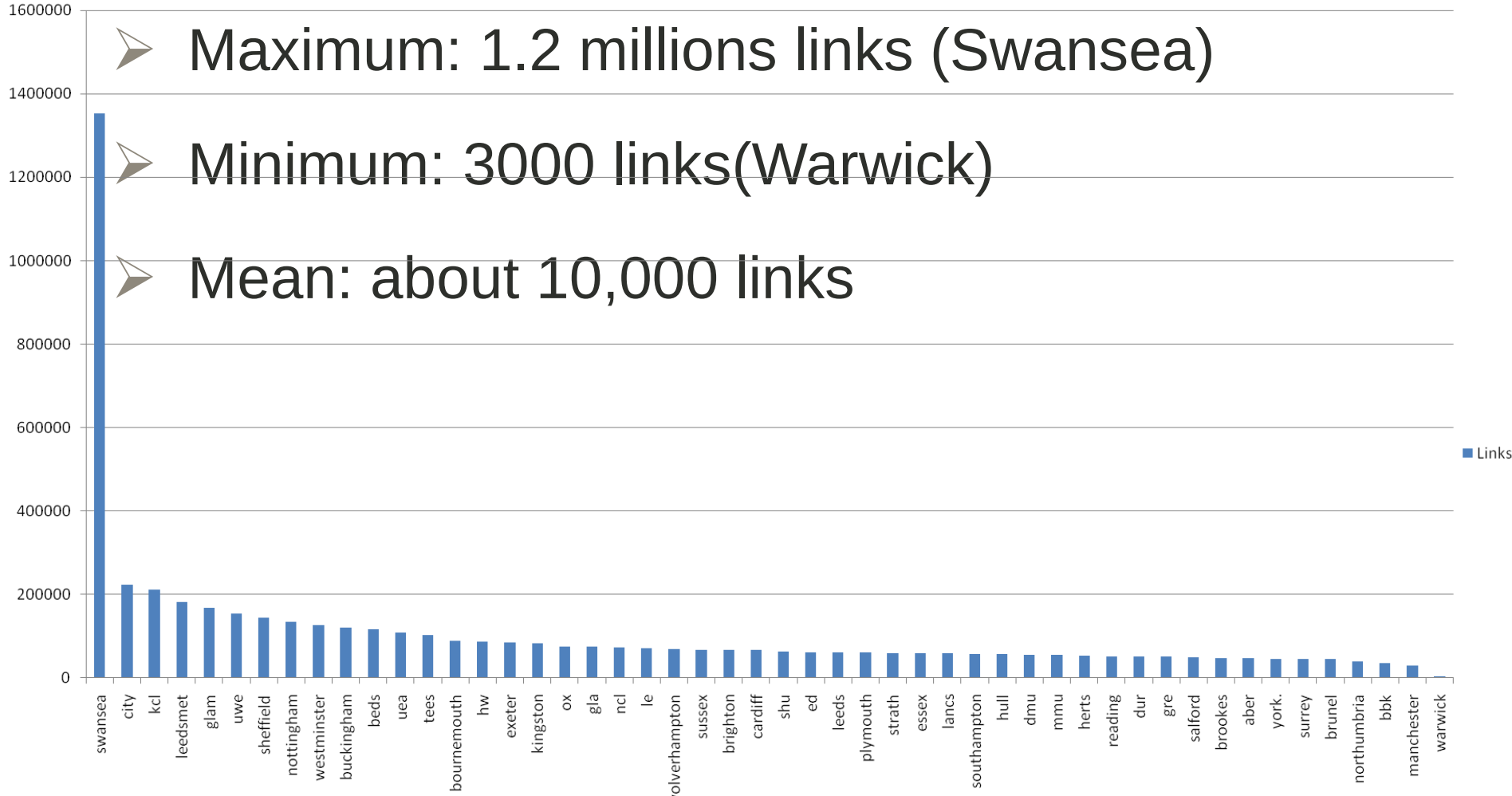


UK Universities: Web Size

- The nodes of the webs
  - max: around 46,129 nodes (Bedfordshire)
  - Min: 1,067 (Warwick) nodes
  - Most: around 5,000
  - (UEA: 7,865, Buckingham: 2,763)

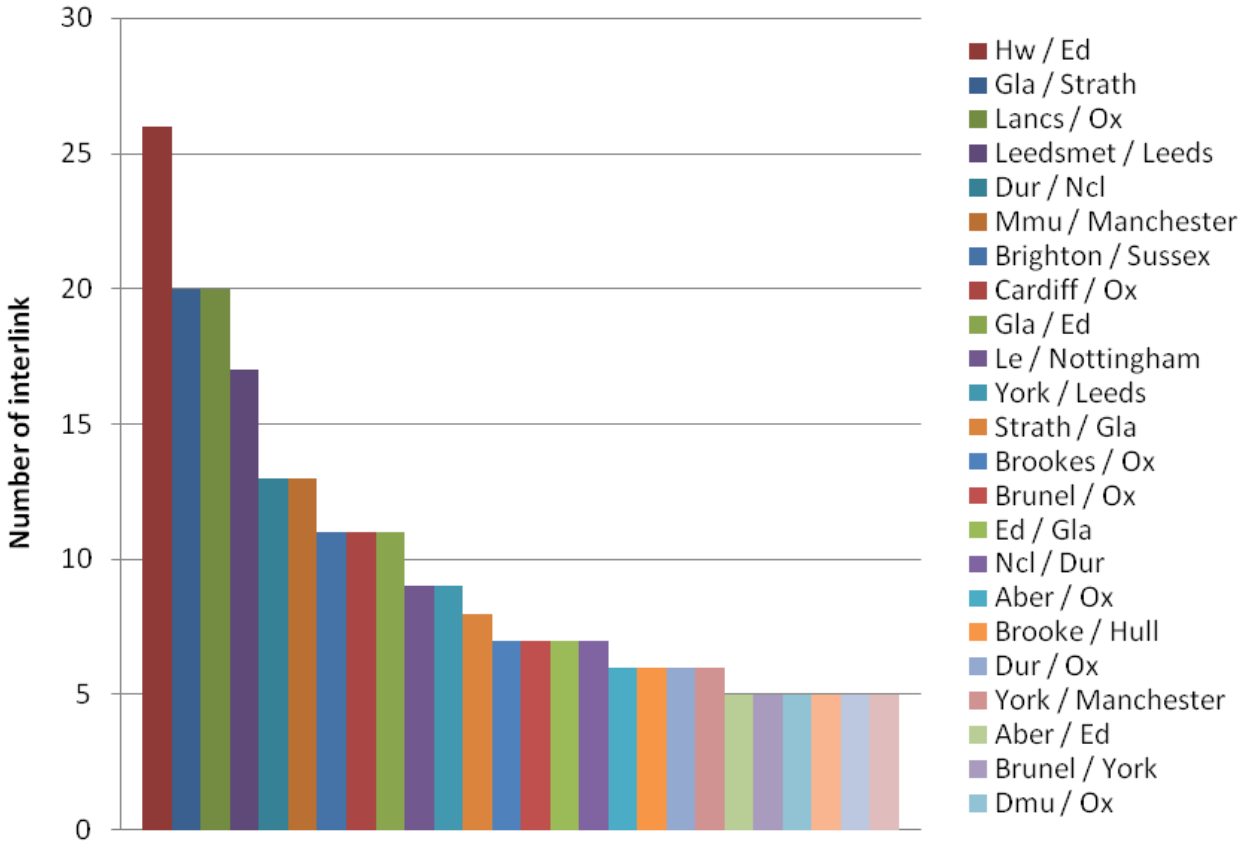
The Number of Links

- Maximum: 1.2 millions links (Swansea)
- Minimum: 3000 links(Warwick)
- Mean: about 10,000 links

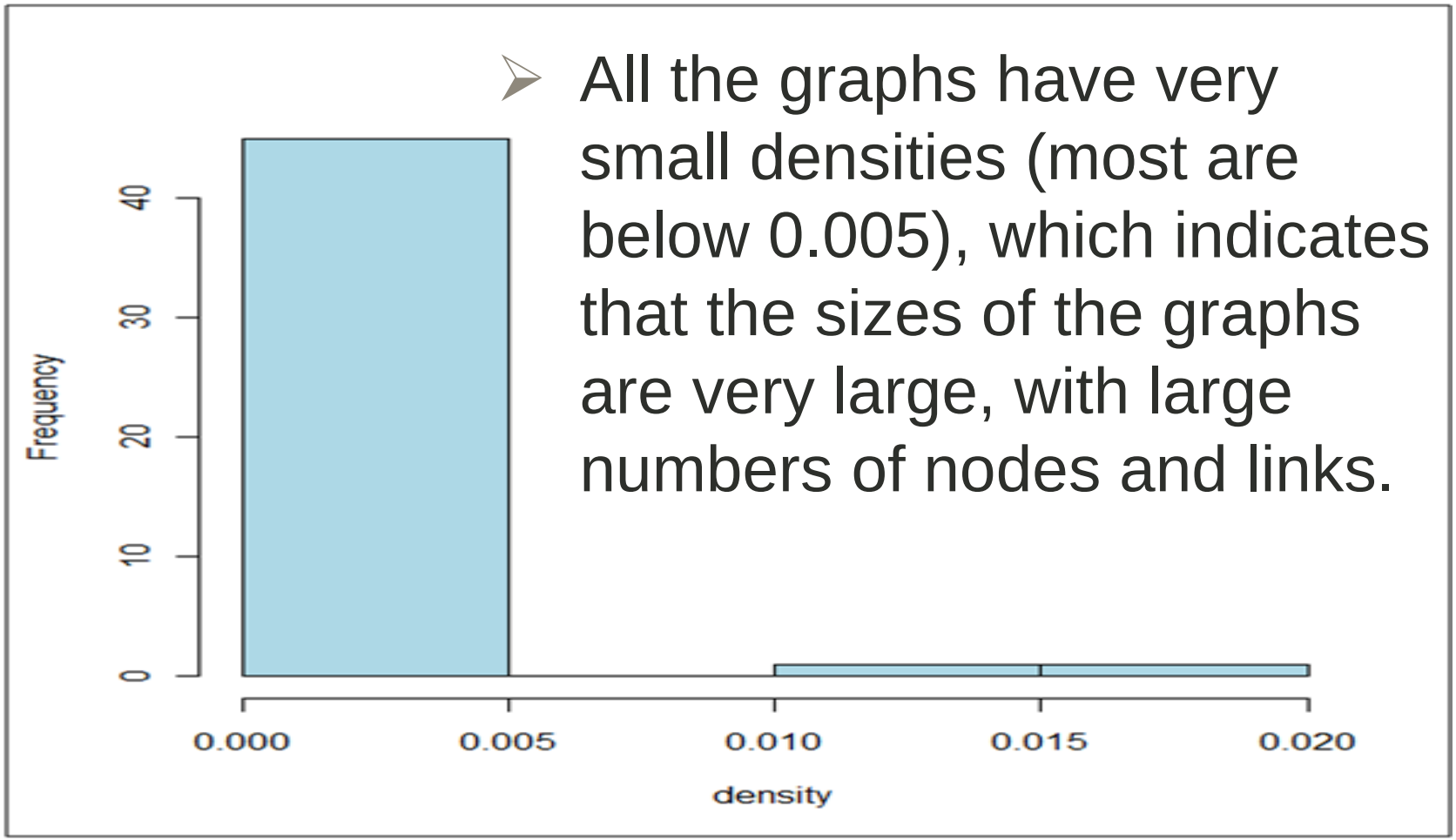


Interlinks between Universities

- Max: 26 links between Heriot-Watt to Edinburgh



UK Uni. Web: Density



- All the graphs have very small densities (most are below 0.005), which indicates that the sizes of the graphs are very large, with large numbers of nodes and links.

Strongly Connected Component(SCC)

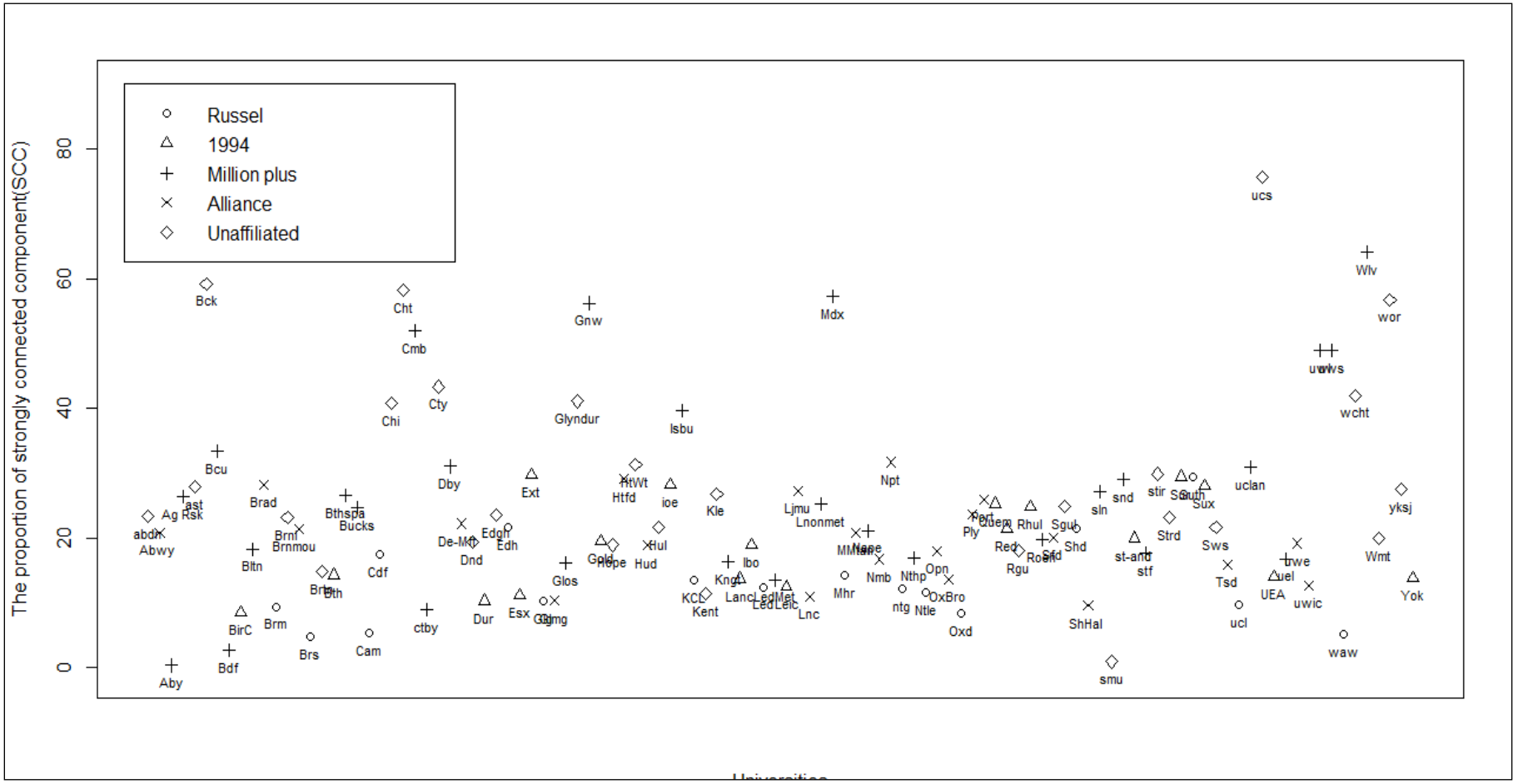


Figure 4: The proportion of the SCC components in the university websites.

Giant OUT Components

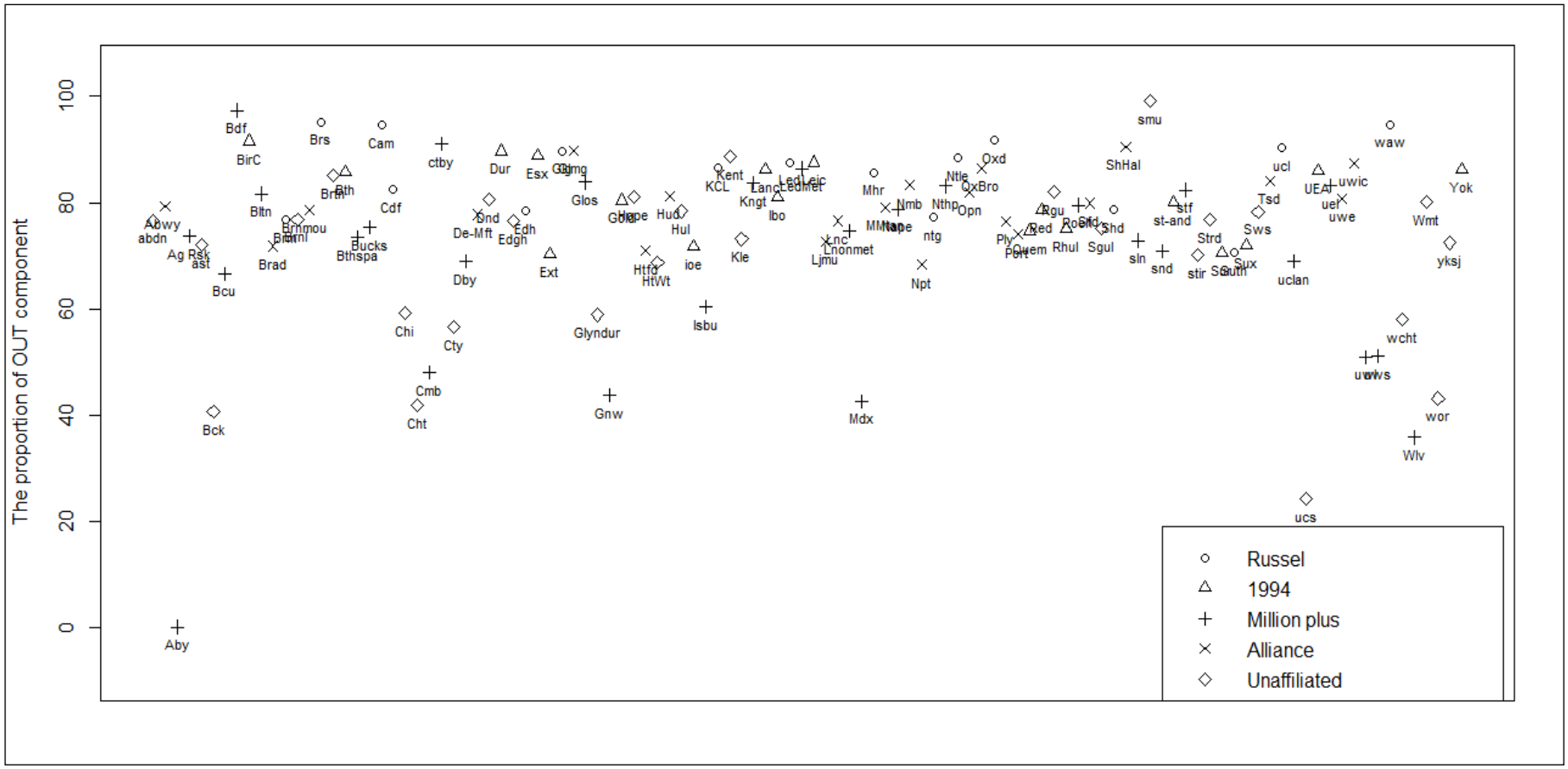
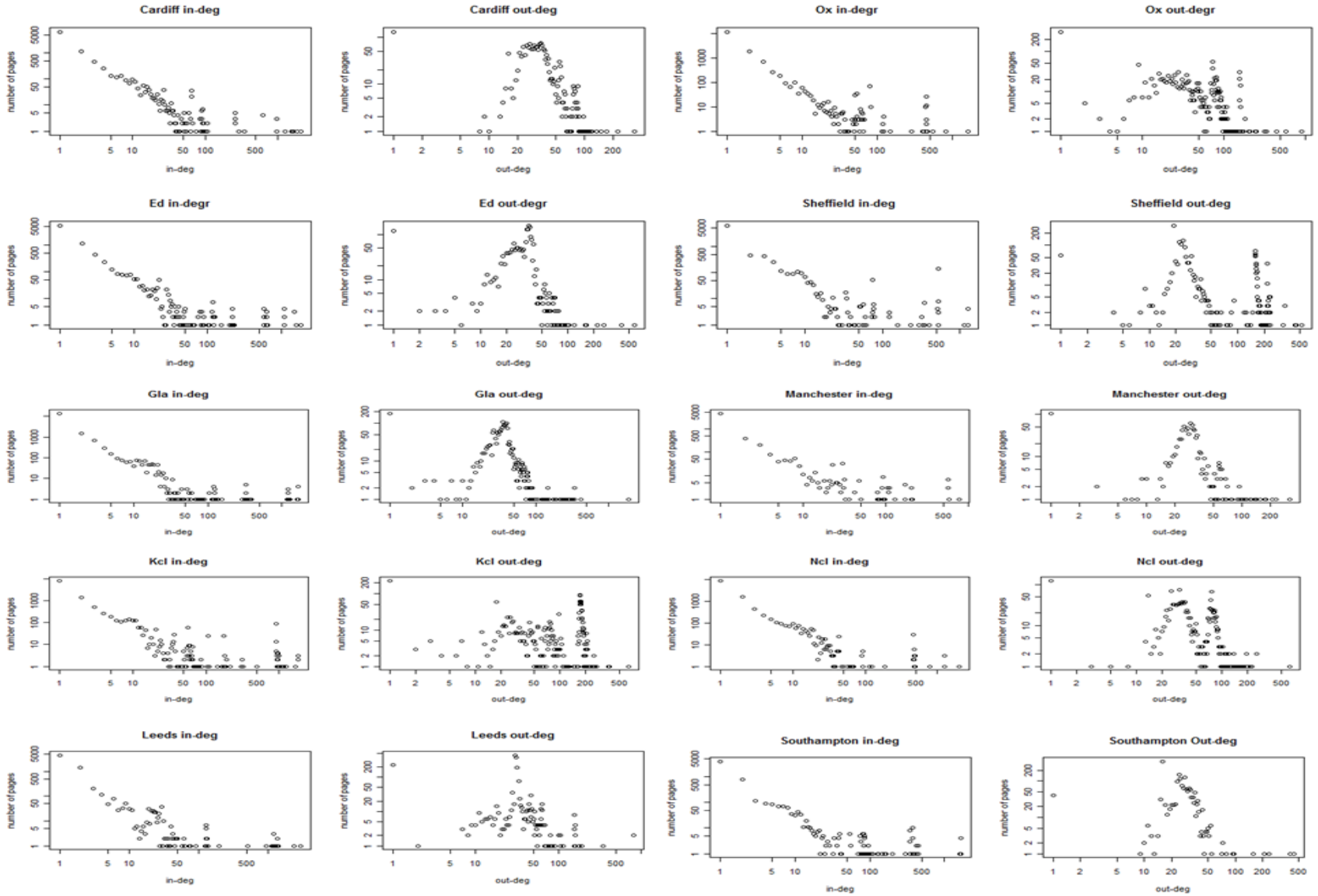
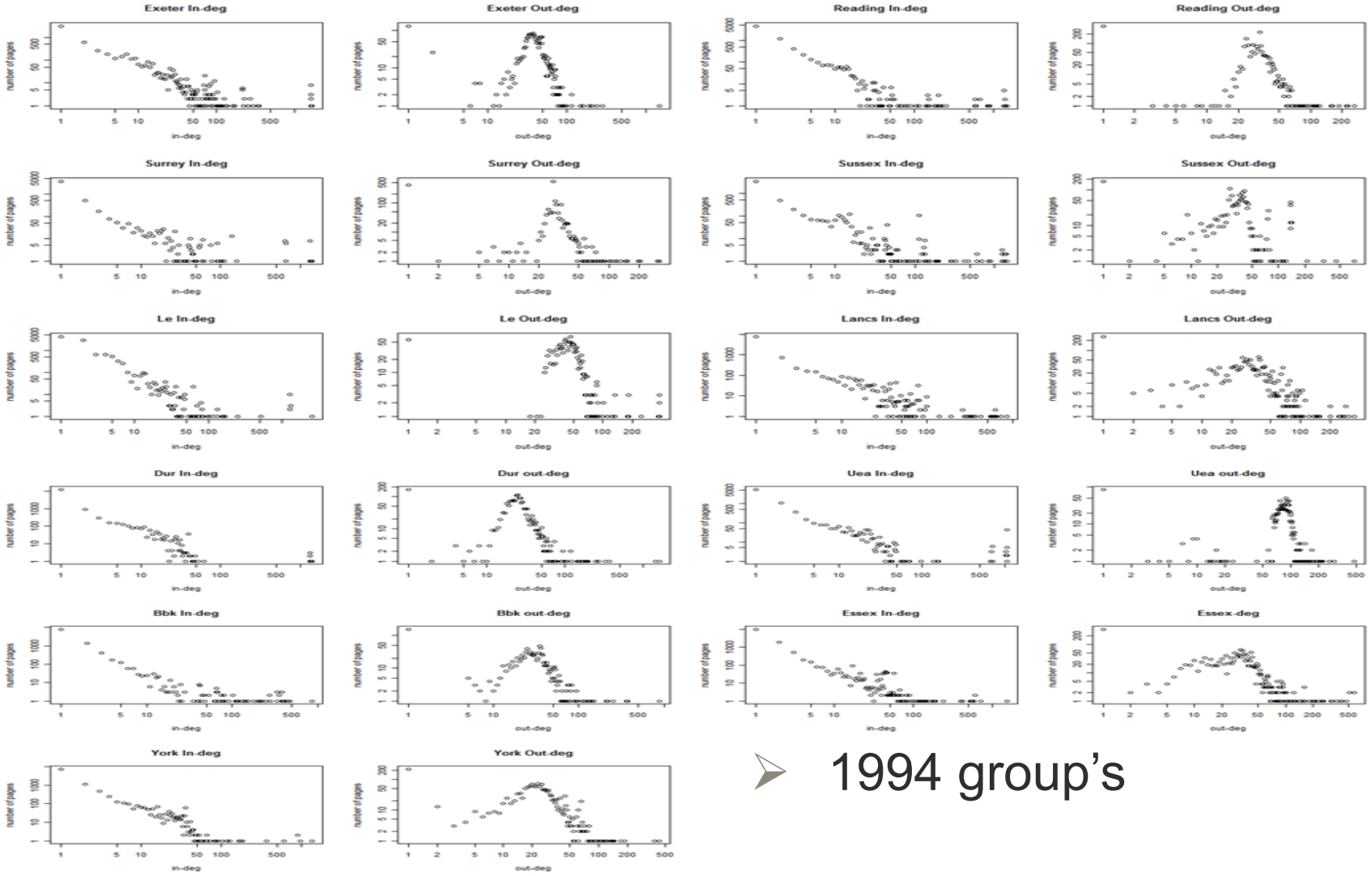


Figure 5: The proportion of OUT components in each university website.

The IN/OUT-Degree Distributions: Russell Group

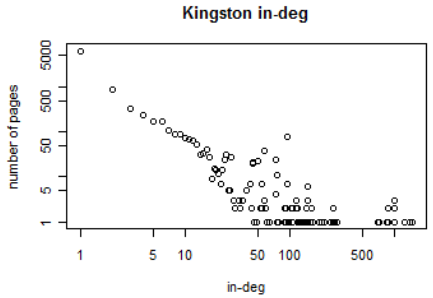


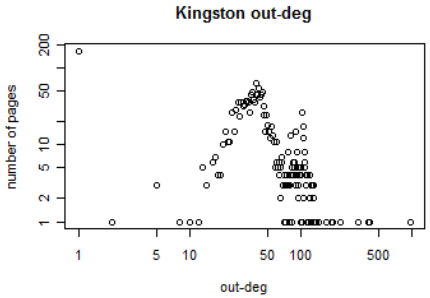
The IN/OUT- Degree Distributions

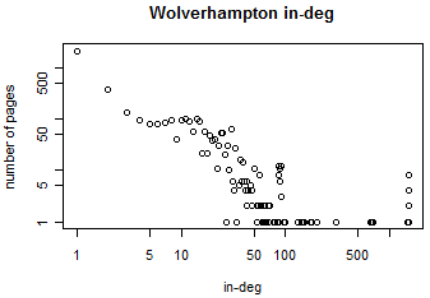


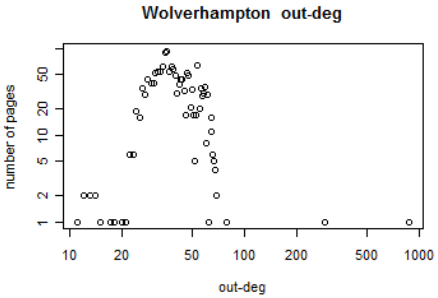
- 1994 group's

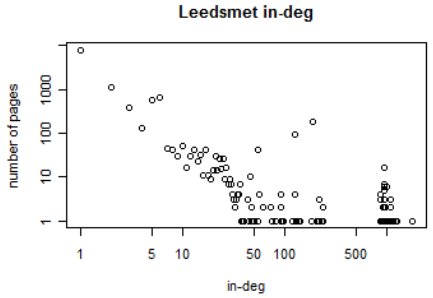
The IN/OUT- Degree Distributions

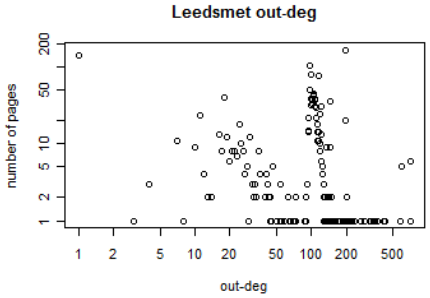


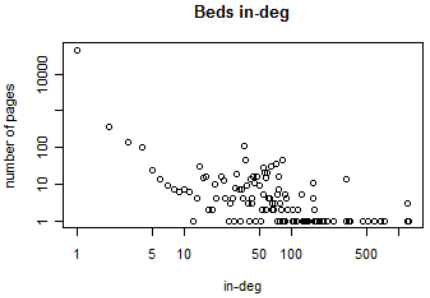


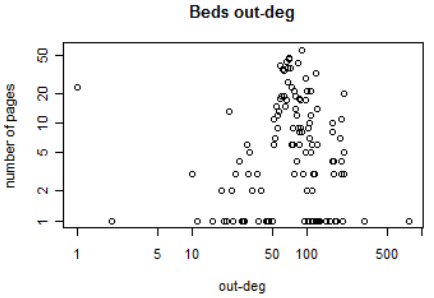


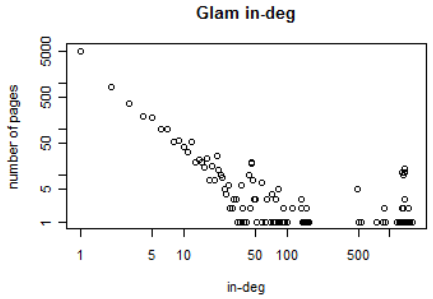


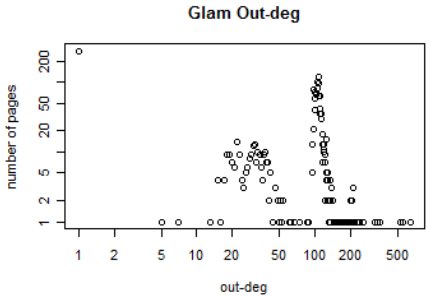








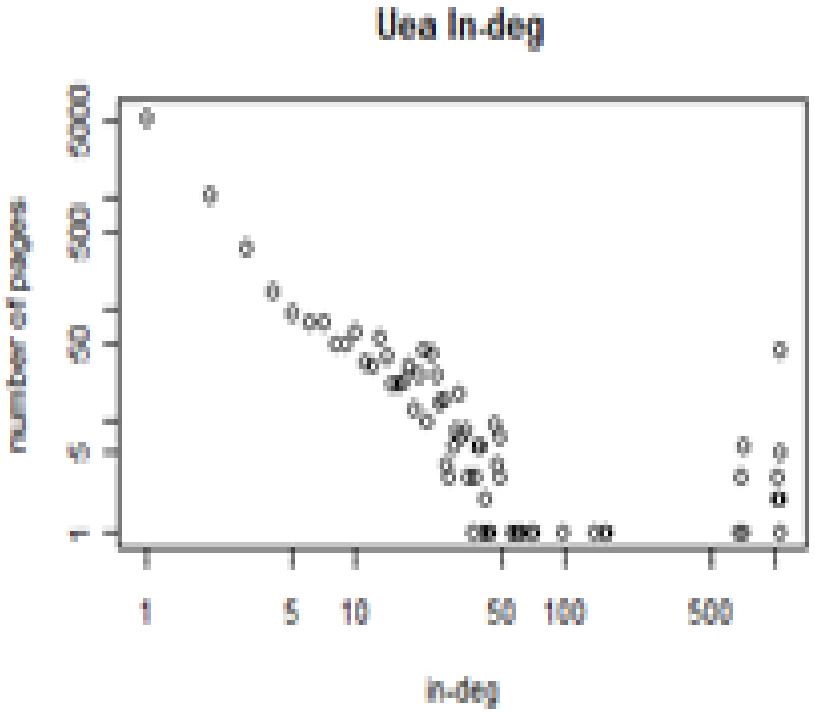


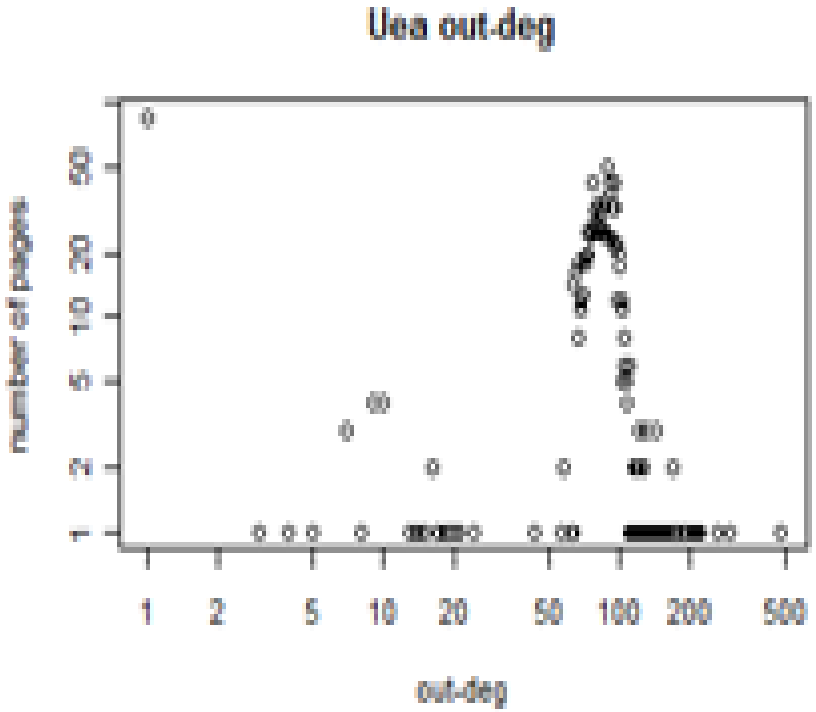


millions group's

The IN- and OUT-Degree: UEA

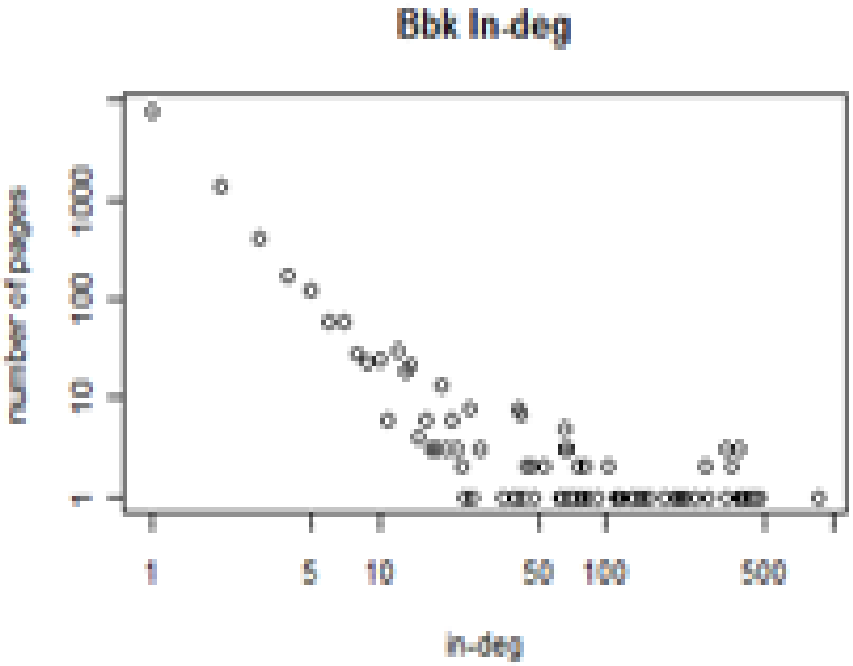
- The degrees of IN links and OUT links.

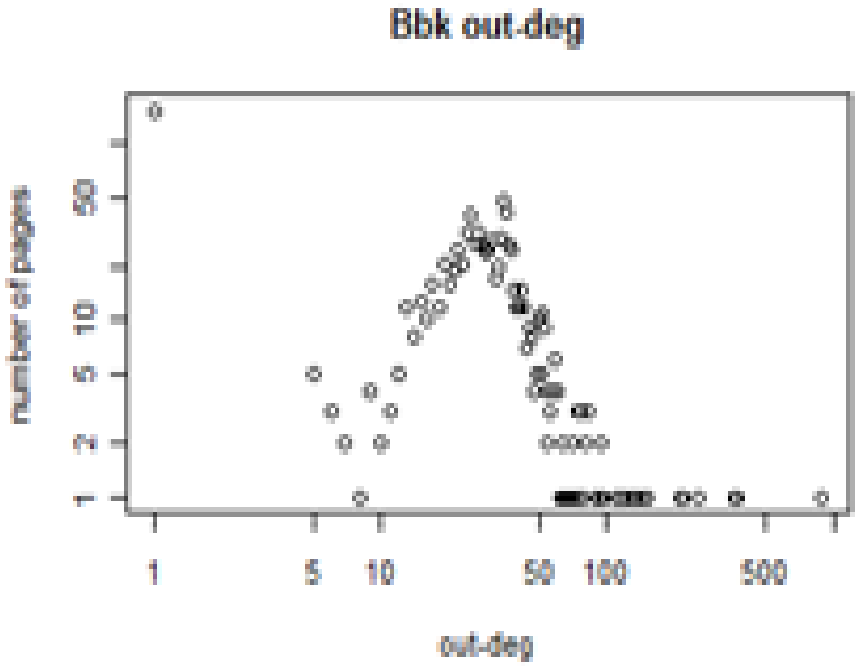




The IN- and OUT- Degree: Buck

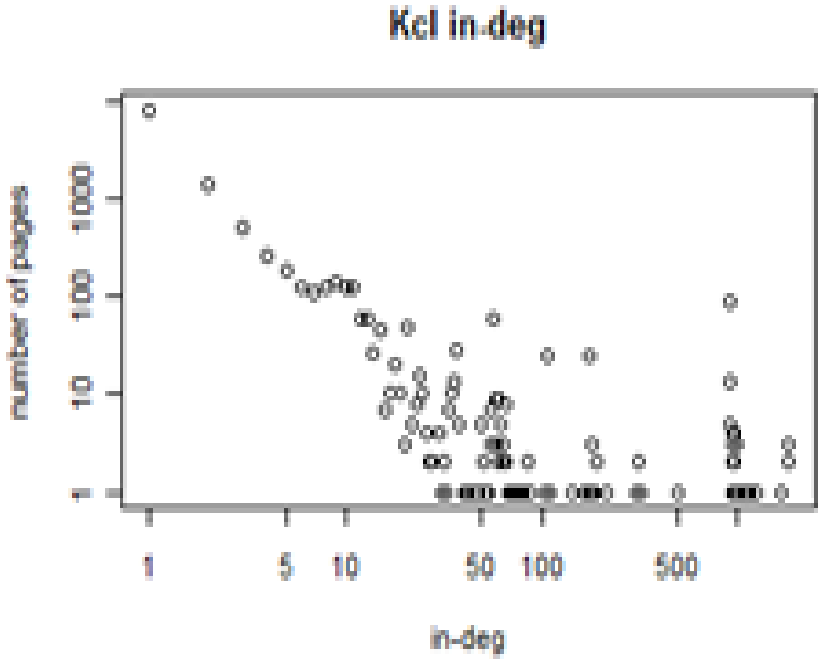
- The degrees of IN links and OUT links.

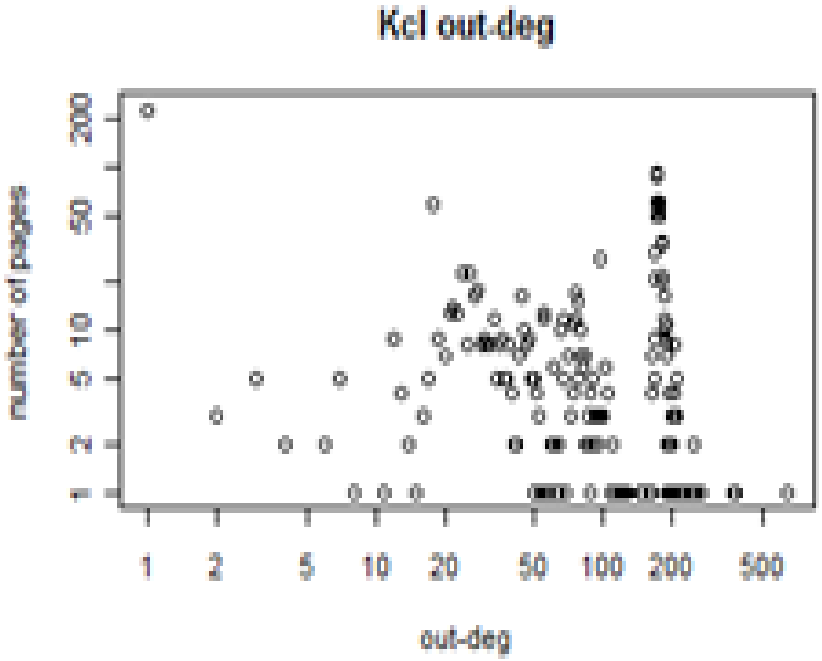




The IN/OUT-Degree: “Bad”

- The distributions of IN/OUT links: unbalanced.



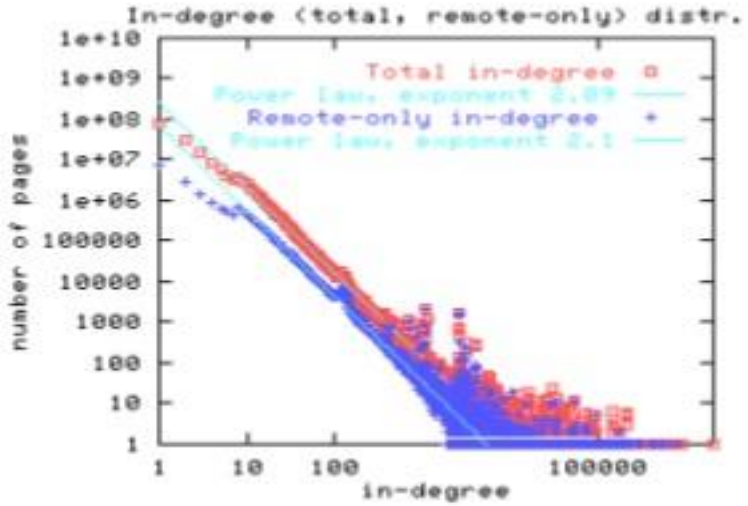


The IN/OUT-Degree: “Bad”

In-Degree Distribution

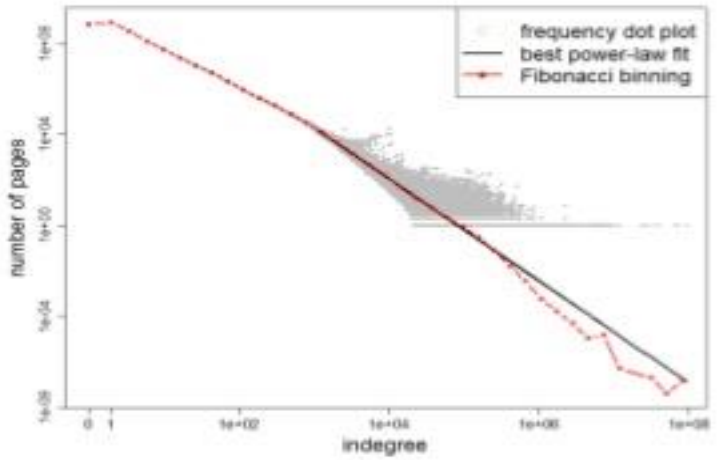
Broder et al. (2000)

Power law with exponent 2.1

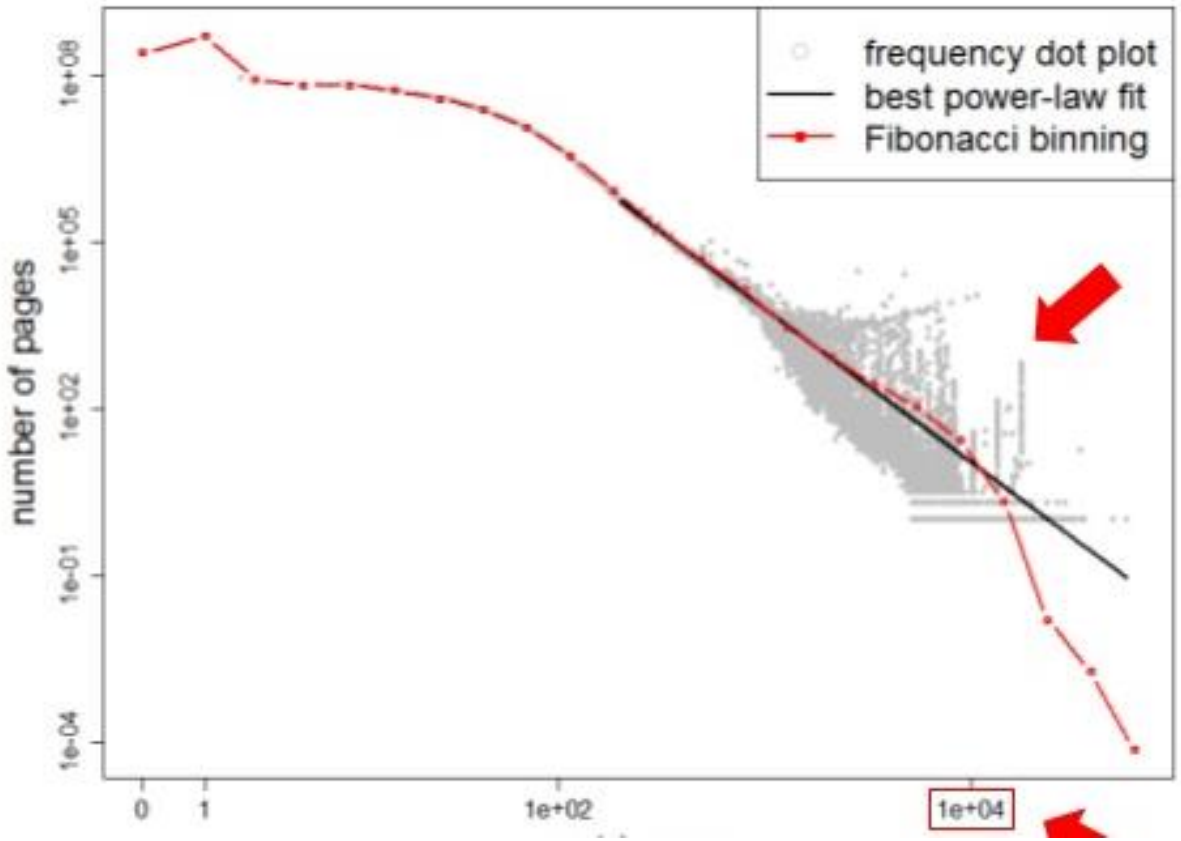


WDC Hyperlink Graph (2012)

Best power law exponent 2.24



Compare with other Out degree distributions



Broder et al.:
Power law exponent 2.78

WDC:
Best power law exponent 2.77

p-value = 0

The Length of the Path

- The length of the shortest path between two pairs of reachable nodes was calculated for each graph.
- The number of reachable nodes differed markedly among the universities.
  - The minimum distance for all the university groups was 1 to 3 hyperlinks,
  - the maximum was 5 to 9 for the Russell group, 5 to 7 for the Million Plus group, and up to 9 for the other groups.

The Length of the Path

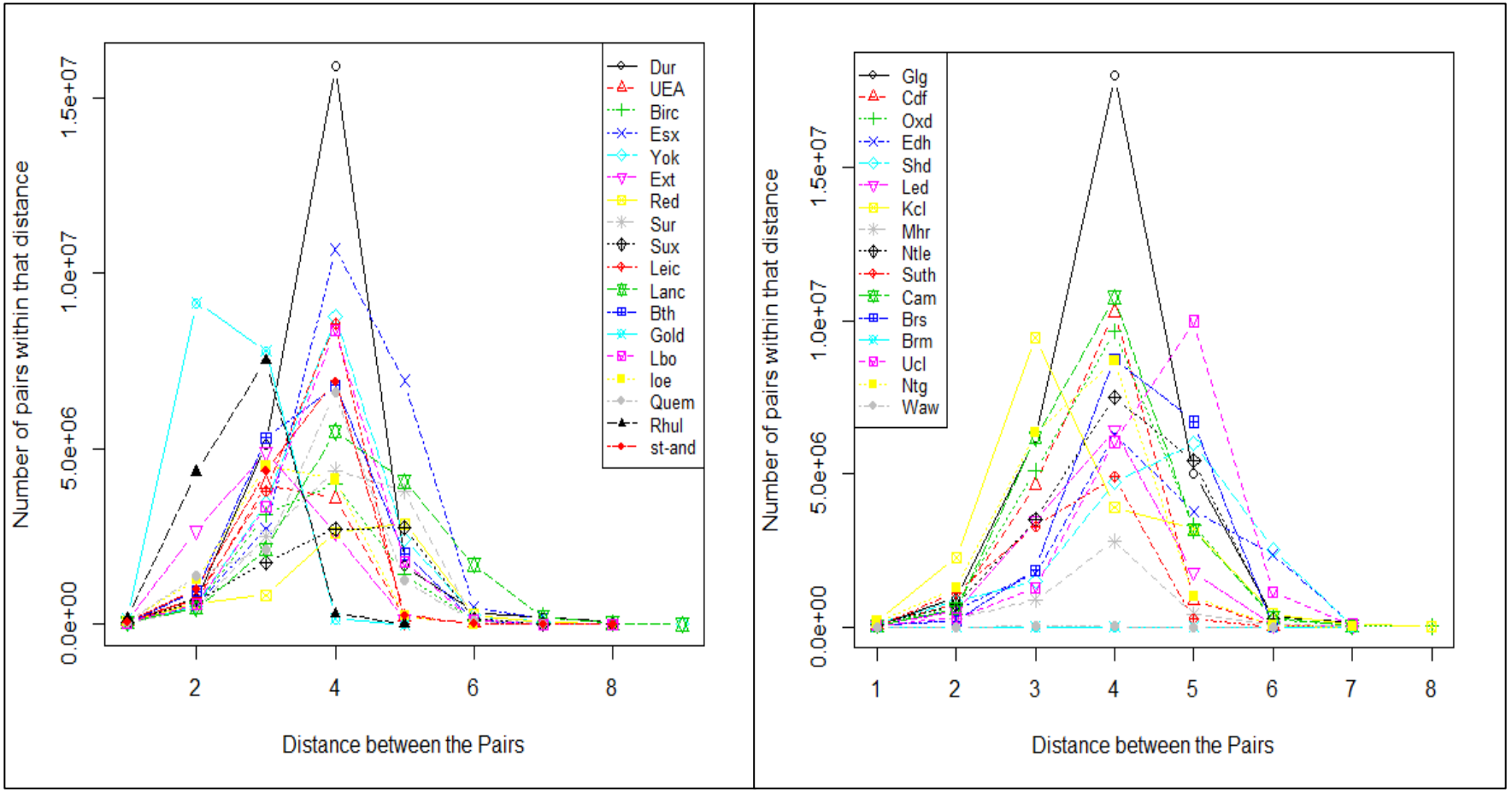


Figure 6: The distribution of the path length between reachable pairs in each university in the 1994 group (left) and the Russell (right).

The Average Directed Distance

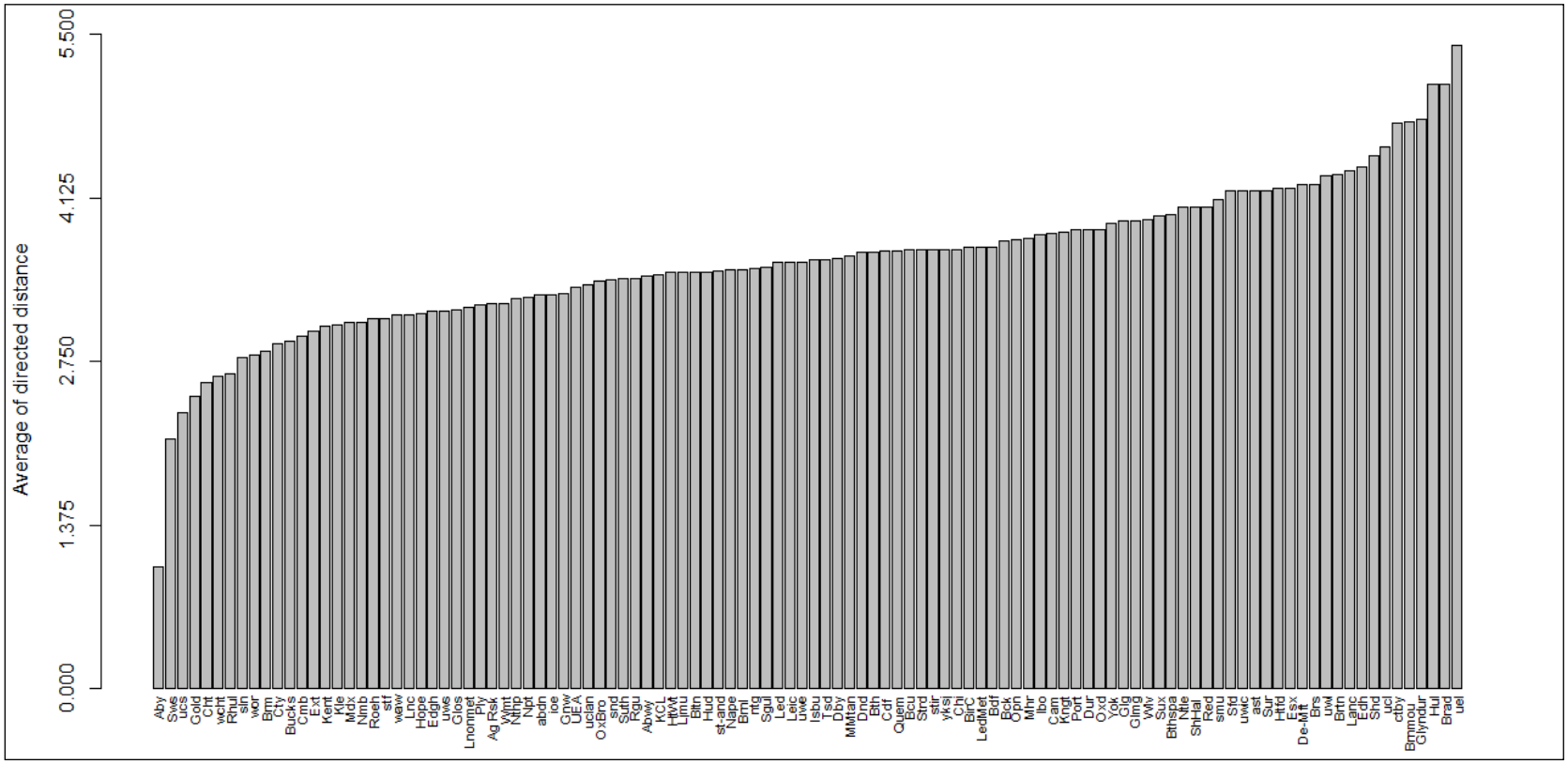
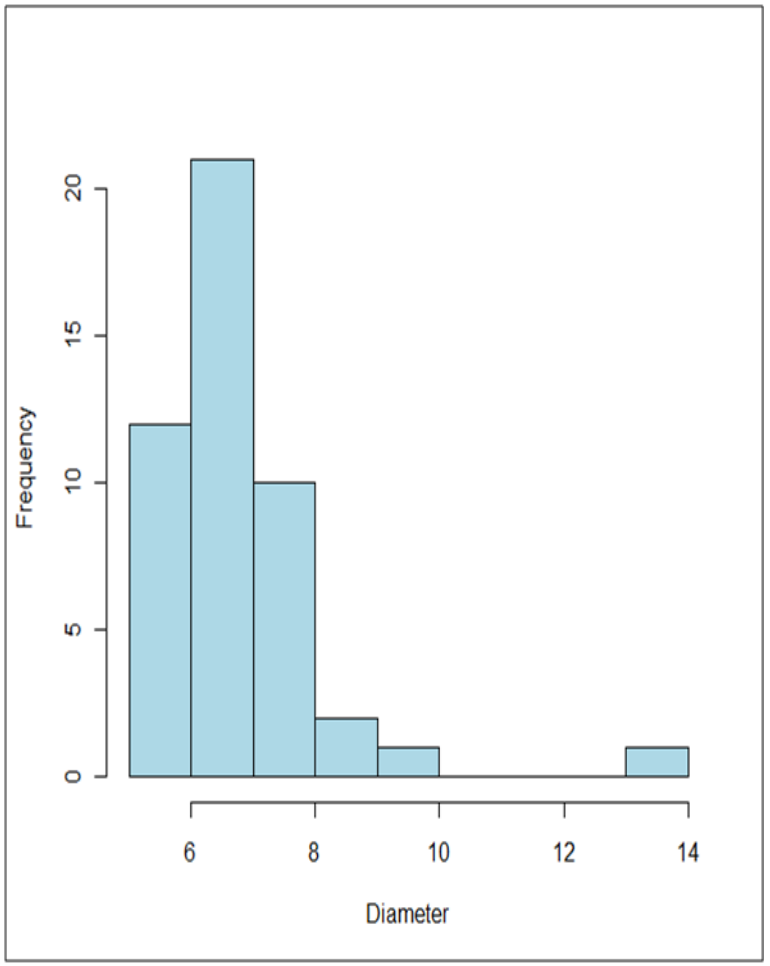


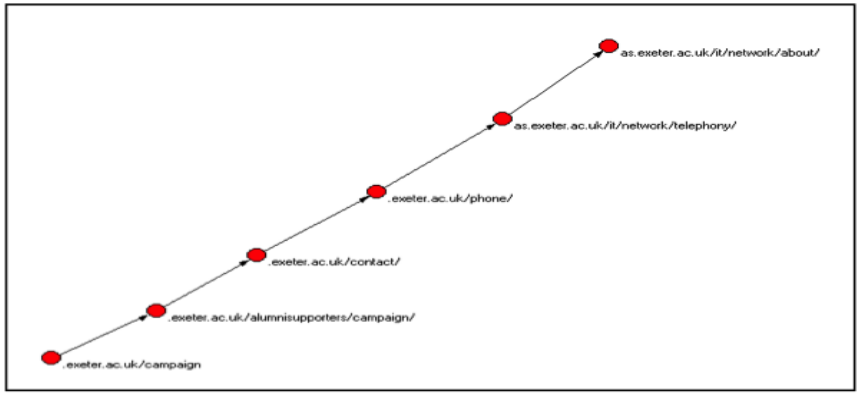
Figure 7: The average directed distance for all the universities

Diameter of Web Graphs

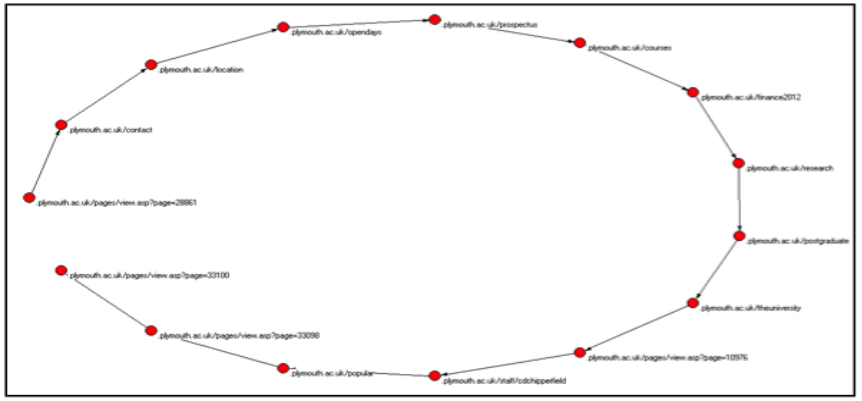
- The diameter of a web graph is the largest shortest path for reachable pairs of nodes in the graph.
  - Represents the maximum number of links in the shortest path between two pages that users need to click on to access the pages.
  - Is the worst case of the optimal path.



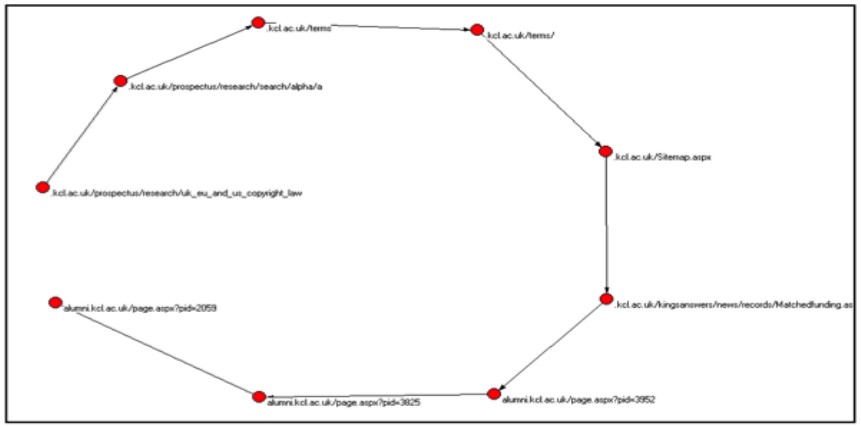
Diameter of the Graph



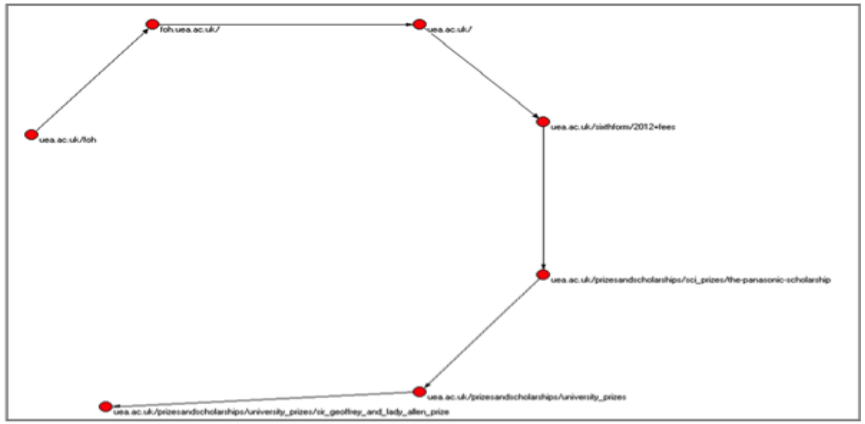
(a) The Diameter University of Exeter which is not through the sitemap or the home page.



(c) The longest diameter for University of Plymouth.



(b) The diameter of King's College London which is through the sitemap



(d) The diameter of University of East Anglia which is through the home page.

The Average Degree of a Node

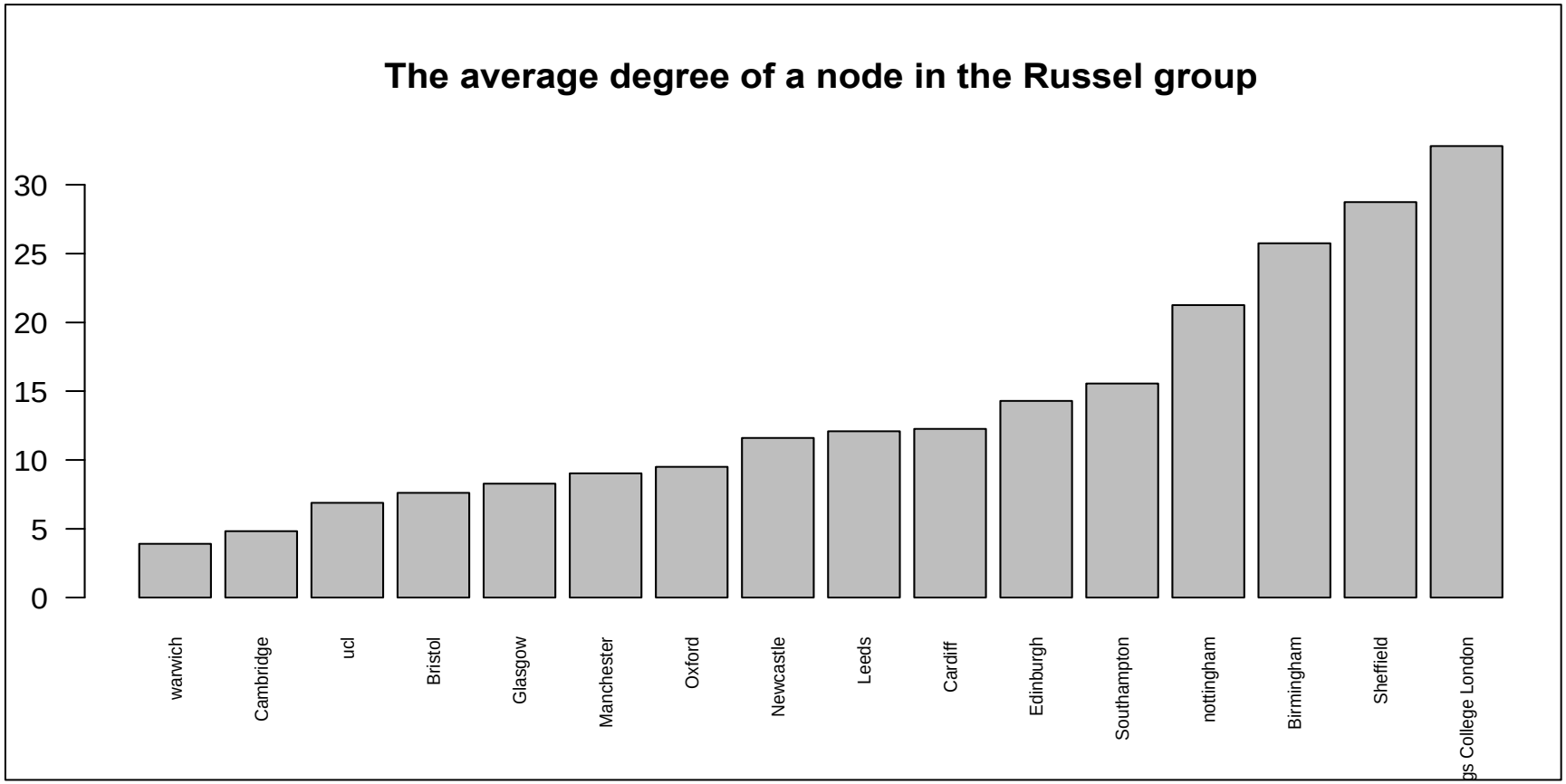


Figure 8: The average degree for each university in the Russell group.

The Average Degree of a Node

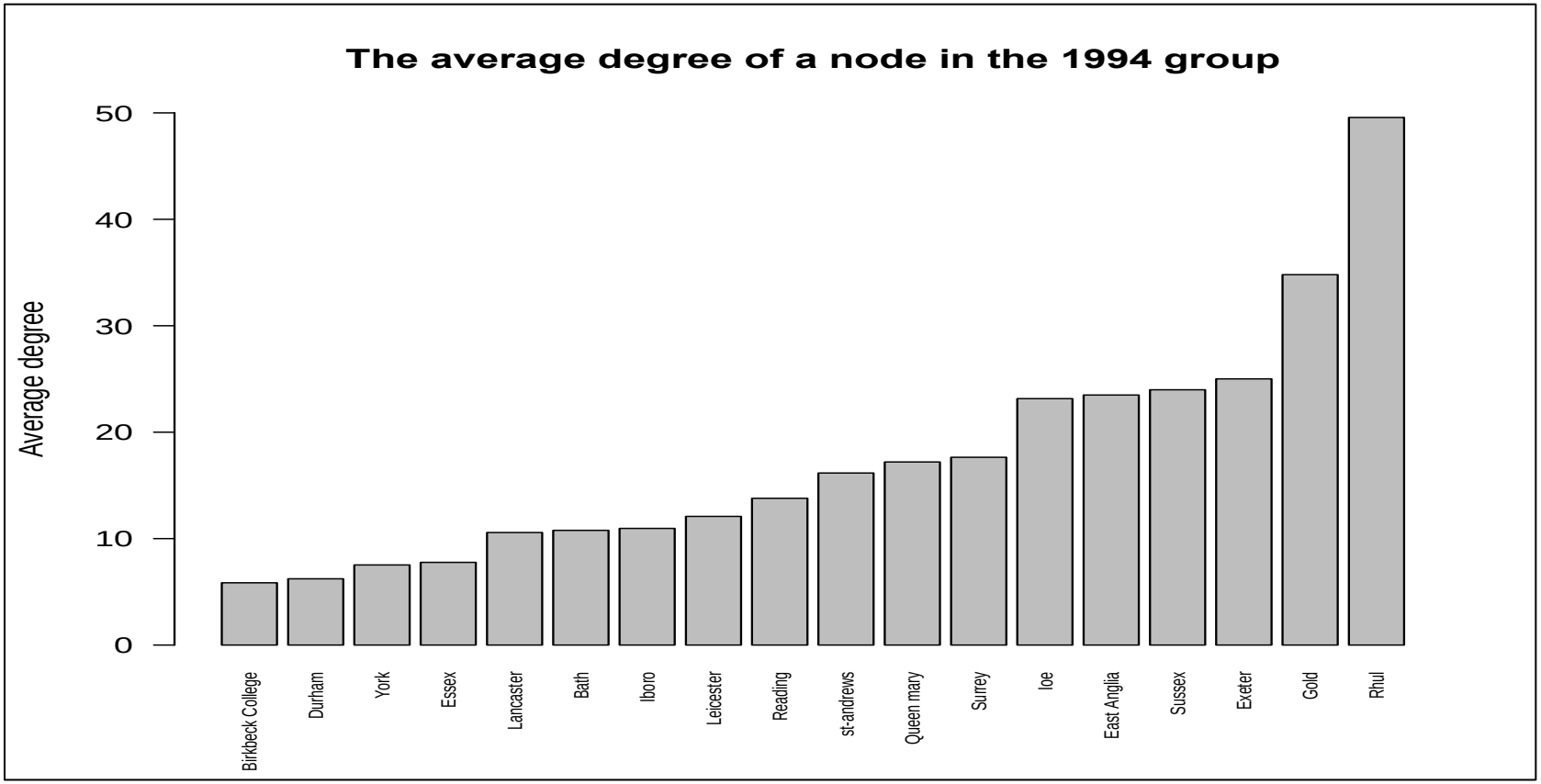


Figure 9: The average degree for each university in the 1994 group.

The Average Degree of a Node

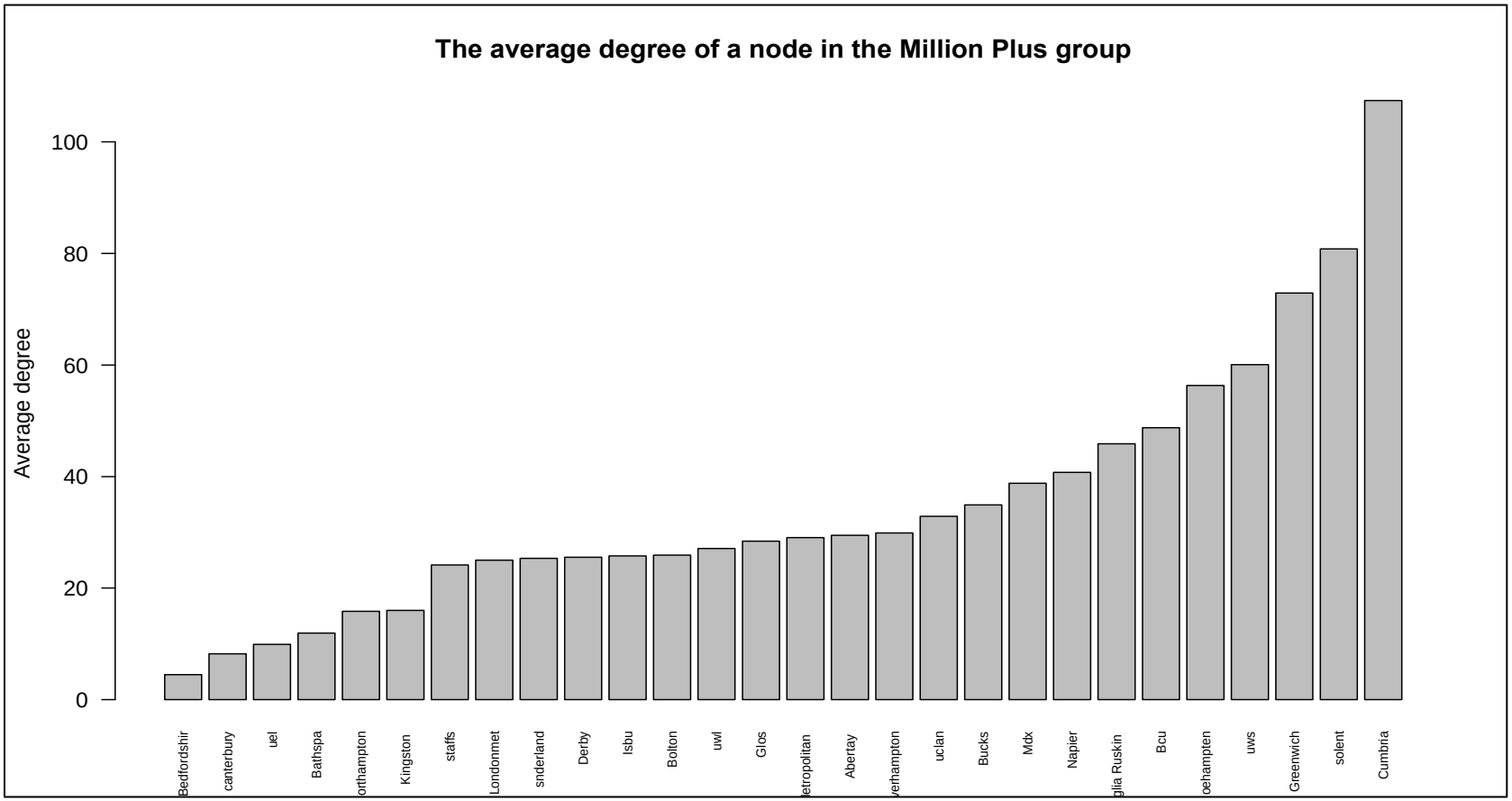


Figure 10: The average degree for each university in the Million Plus group.

The Average Degree of a Node

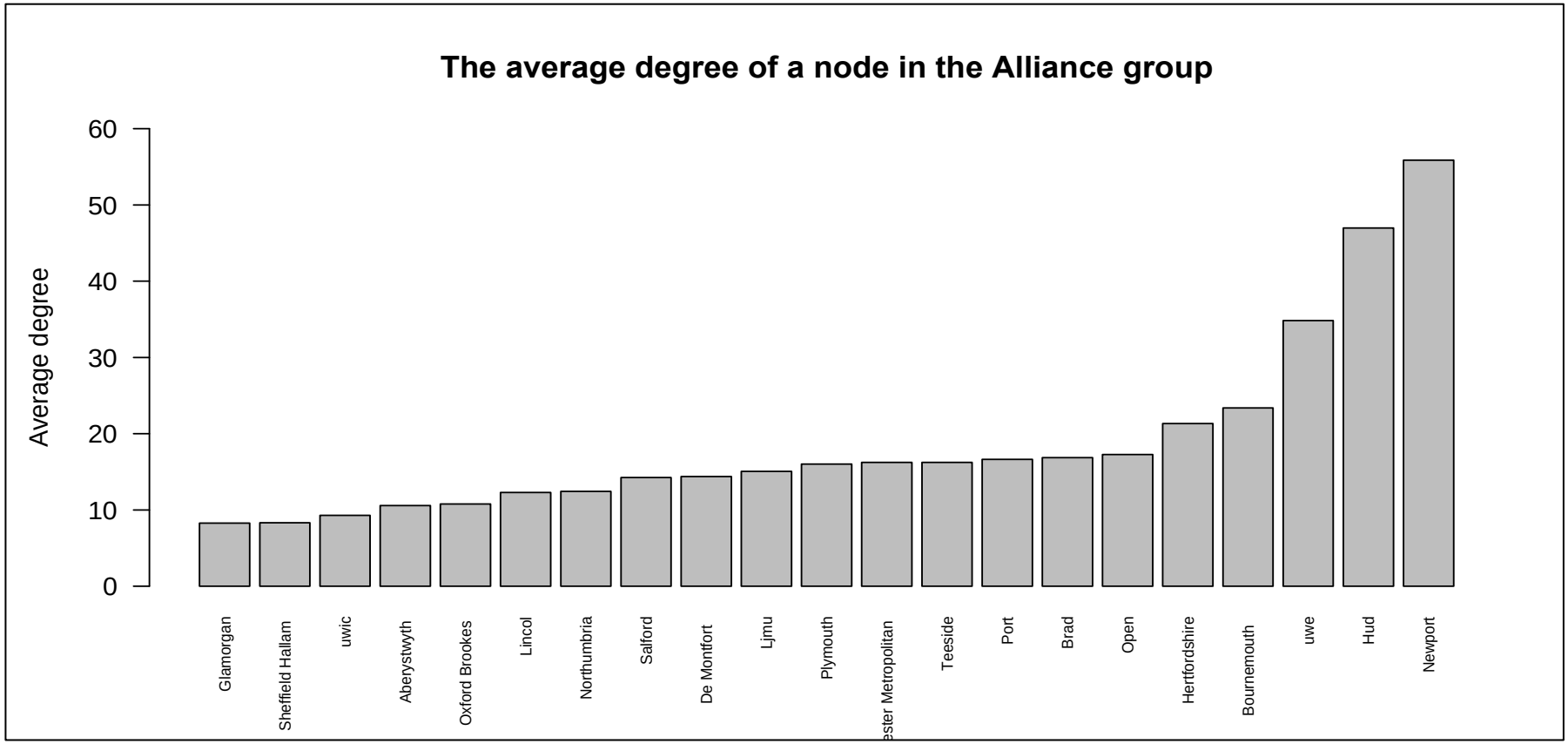


Figure 11: The average degree for each university in the Alliance group.

The Average Degree of a Node

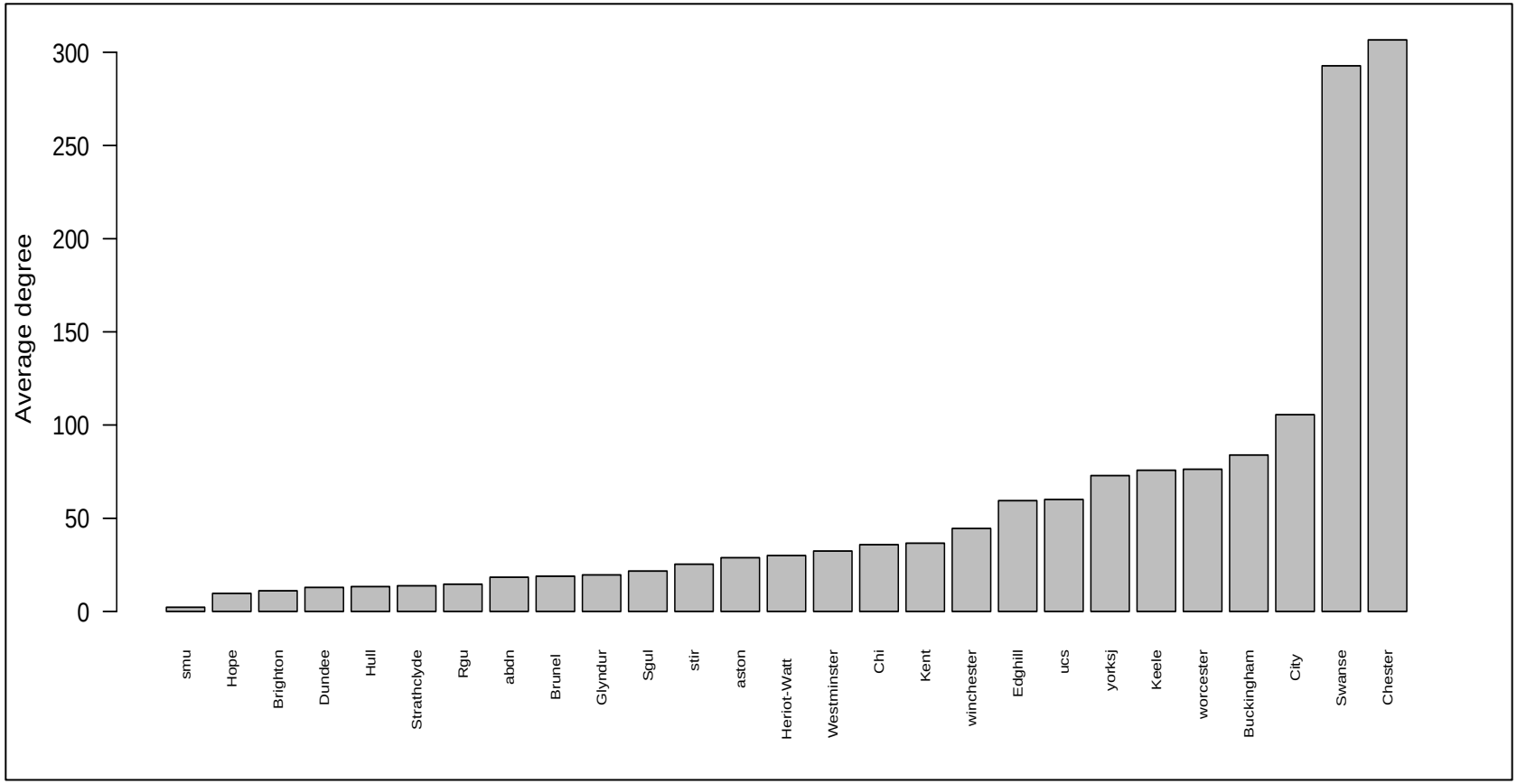


Figure 12: The average degree for each university in the Unaffiliated group.

The Average of Cluster Coefficient (ACC)

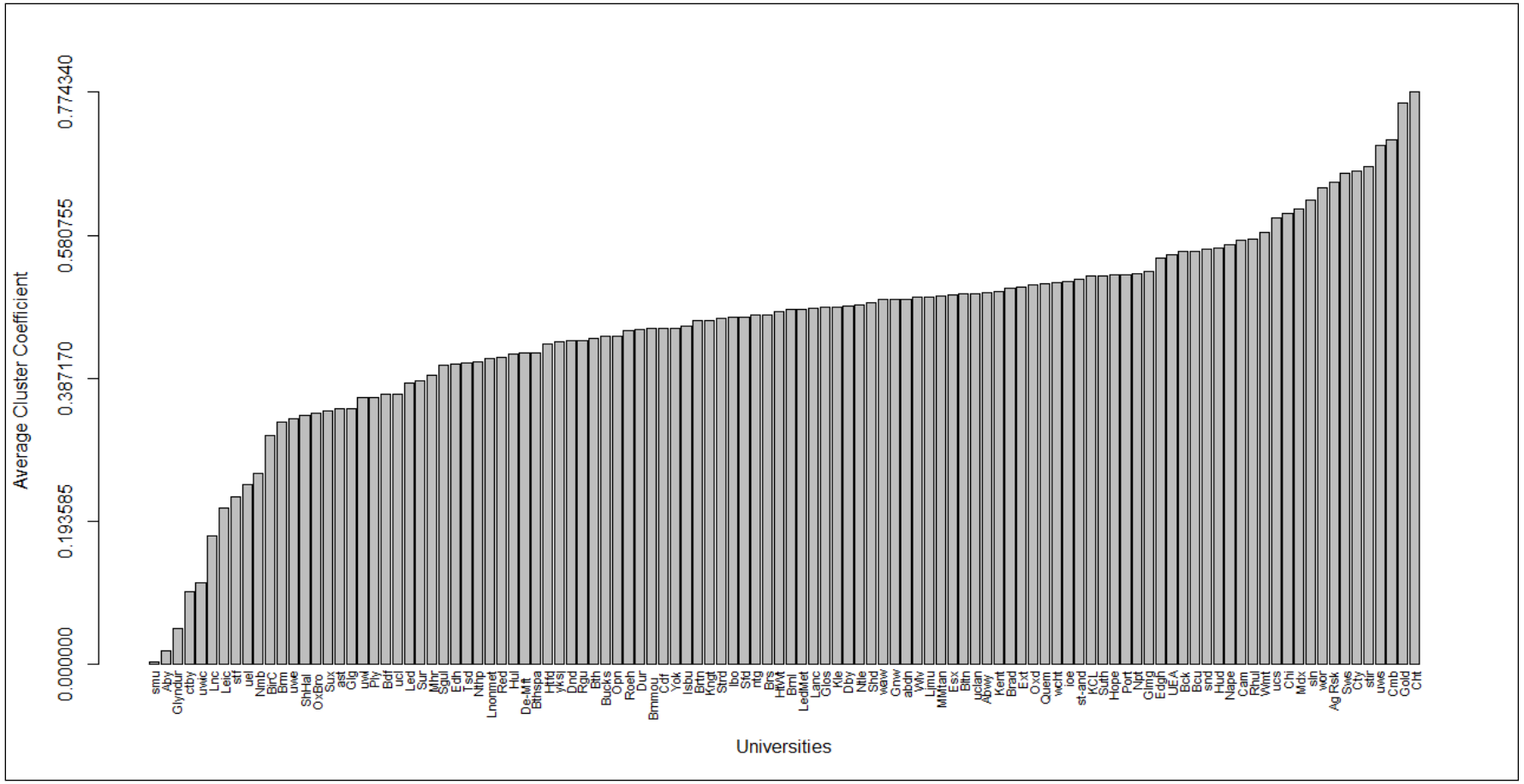


Figure 9: The average of the cluster coefficient ACC in each university graph sorted from the highest to the lowest average.

Network Closeness Centralisation

- The closeness centrality of the node measures the path from it to its closest neighbour.
- The highest score was Swansea University (Sws) (0.75), followed by University Campus Suffolk (0.74); two others with the same scores were 0.65 for Worcester and Middlesex.
- The lowest score was recorded for Hull (0.39), followed by Bradford (Brad) (0.37) and Lancaster (Lanc) (0.34).

Web Structure Evaluation

Table 1: Correlation coefficients, t-tests and p-value.

Correlation	Correlation coefficient	t-test	p-value
Average degree and average distance	-0.45	-5.31	5.955e-07
Average degree and website size	-0.25	-2.69	0.01
Average distance and website size	0.26	2.81	0.00
Website size and SCC	-0.57	-7.20	8.59e-11
Website size and OUT	0.59	7.56	1.423e-11
Average degree and SCC	0.45	5.32	5.715e-07
Average degree and OUT	-0.40	-4.60	1.14e-05

Web Structure Evaluation

Rank the Universities on a Scale

- The ACC is the best measure for clearly identifying the proportion of the number of pages to hyperlinks in a website.

Category	ACC	Rank
Very poor structure	From 0.003 to 0.29	1
Poor structure	From 0.3 to 0.39	2
Reasonable structure	From 0.4 to 0.49	3
Good structure	From 0.5 to 0.59	4
Very good structure	From 0.6 to 0.77	5

Web Structure Evaluation

Defining Criteria for Good and Bad Structure

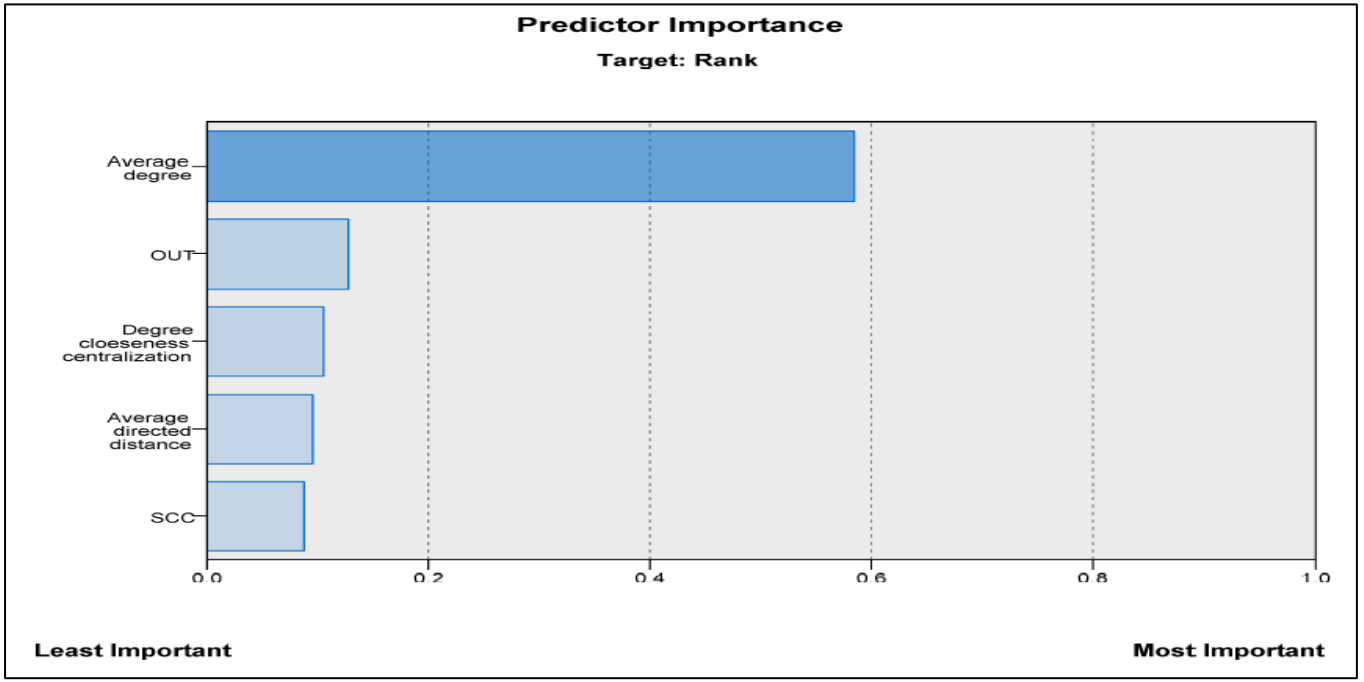


Figure 10: Important graph properties classifying good and bad internal structures.

Conclusion

- The results have revealed the rules and criteria for determining whether the internal structure of an academic website is good or bad in terms of navigation.
  - that the average degree and the percentage of SCC pages play an important role in determining good and bad structure.
- As a result of analysing the graph properties and studying the correlations between them, we found that to design an easily navigable website, the number of pages must be balanced with the number of links in the graph, which will make the distance between pages shorter and easier to navigate. This correlation can be achieved using ACC.

References

- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. 2000. Graph structure in the web. Computer Networks. 33, 1-6, 309–320.

- Holmberg, K. 2009. Webometric network analysis : mapping cooperation and geopolitical connections between local government administration on the web. Ph.D. dissertation, *Åbo Akademi University, Department of information Science,* Finland, (July. 2011). DOI= http://doria17kk.lib.helsinki.fi/bitstream/handle/10024/52528/HolmbergKim.pdf?sequence=2

- Noruzi, A. 2006. The web impact factors a critical review. *The Electronic Library.* 24, 4, 490–500. DOI= http://eprints.rclis.org/archive/00005543/

- Petricek, V., Escher, T., Cox, I. and Margetts, H. 2006. The web structure of e-government-developing a methodology for quantitative evaluation. *In Proceeding of the 15th international conference on World Wide Web*. 669–678.

- Meusel, R et al. Graph Structure of the Web Revised. Int. conference on WWW, April 4-7, 2014, Seoul, Korea.