Lesson 6: Modeling the Web as a graph

Unit 3: Descriptive statistics for the Web graph

Rene Pickhardt

Introduction to Web Science Part 2
Emerging Web Properties

WeST
People and Knowledge Networks

Institute for Web Science and Technologies · University of Koblenz-Landau, Germany
Completing this unit you should

• Know terms like Size and (unique) volume

• Be able to count the in and out degree of web pages

• Have an idea what kind of law (in & out) degree distributions follow

• Know that degree is not distributed in a fair way

• Know that the Gini coefficient can be used to measure fairness
Some basic statistics about Simple English

- **Size** (Number of vertices)
  - 100‘312

- **Volume** (Number of edges)
  - 1‘627‘472

- **Unique volume** (Number of unique edges)
  - 746‘086
Empirical connection between Size and volume in over 100 networks

• There seems to be a connection

• The larger the size, the larger the volume

• Beware the log log plot
Neighbors, in degree and out degree

\[ In(v) = \{ u \in V | \exists e \in E : e = (u, v) \} \]
\[ \text{indeg}(v) = |In(v)| \]

\[ Out(v) = \{ u \in V | \exists e \in E : e = (v, u) \} \]
\[ \text{outdeg}(v) = |out(v)| \]

\[ Neighbours(v) = In(v) \cup Out(v) \]
\[ \text{deg}(v) = |Neighbours(v)| \]
\[ \text{deg}(v) \leq \text{indeg}(v) + \text{outdeg}(v) \]
Degree distribution for Simple English Wiki

Most nodes have a degree of 1
Degree distribution for Simple English Wiki

Few nodes have high degree
Out degree distribution

Highest outdegree is ~ 1000
In degree distribution

Highest indegree is ~ 10’000
Comparing in and out degree distributions

Which one is more even?
Use the Gini coefficient to describe Fairness

\[ G = \frac{\sum_i \sum_j |\text{deg}(v_i) - \text{deg}(v_j)|}{2 \sum_i \sum_j \text{deg}(v_i)} \]

\[ G = \frac{A}{A + B} \]

\[ G \in [0, 1] \]
- 0 = fair
- 1 = unfair
Out degree lorenz curve

Share of edges

Share of nodes with smallest degrees

G = 77.7%
P = 19.5%
In degree lorenz curve

\[ G = 80.1\% \]
\[ P = 17.6\% \]
Thank you for your attention!

Contact:
Rene Pickhardt
Institute for Web Science and Technologies
Universität Koblenz-Landau
rpickhardt@uni-koblenz.de