Introduction to Data Management
CSE 344

HW6 Extra: Statistical Properties of the Web

Based on: Kumar, Raghavan, Rajagopalan, Sivakumar, Tomkins, Upfal: The Web as a Graph, PODS 2000

Exponential Distribution
- n(d) \propto c2^{-d} \quad (generally, cx^d, for some x < 1)
- A random graph has exponential distribution
- Best seen when n is on a log scale

Exponential Distribution

Zipf Distribution
- n(d) \propto 1/d^x, for some value x>0
- Human-generated data has Zipf distribution: letters in alphabet, words in vocabulary, etc.
- Best seen in a log-log scale

The Histogram of the Web

The Bowtie Structure of the Web

Figure 2: In-degree distribution.
HW6

- In HW6, we study a subset of the web: Semantic web.
- “The Semantic Web is a collaborative movement led by the World Wide Web Consortium (W3C) that promotes common formats for data on the World Wide Web. By encouraging the inclusion of semantic content in web pages, the Semantic Web aims at converting the current web of unstructured documents into a “web of data”. It builds on the W3C’s Resource Description Framework (RDF)” (from Wikipedia)

RDF

- Graph is a series of RDF triples: (subject-predicate-object)
- Example: “Paris is capital of France” in RDF: (Paris-isCapitalOf-France)