UNIVERSITÄT DES SAARLANDES
MAX-PLANCK-INSTITUT INFORMATIK
Dr.-Ing. Ralf Schenkel
Dr.-Ing. Marc Spaniol

Web Dynamics (SS 10)
Assignment 1

Handout on: April 29, 2010

Due on: May 6, 2010

Exercise 1.1: Estimating the size of a search engine

Consider the sets A, B and their intersection on Figure 1. Suppose that we do not know any of their sizes but that we can sample uniformly from any set and check membership in any set. For example, sampling uniformly from A, we may find that about $1/2$ of the samples are in B as well. Similarly, $1/6$ of the samples from B may be in A. We now want to estimate the size of the overlap $A \cap B$ and the relative size of each of the sets. In our example, B is three times as large as A.

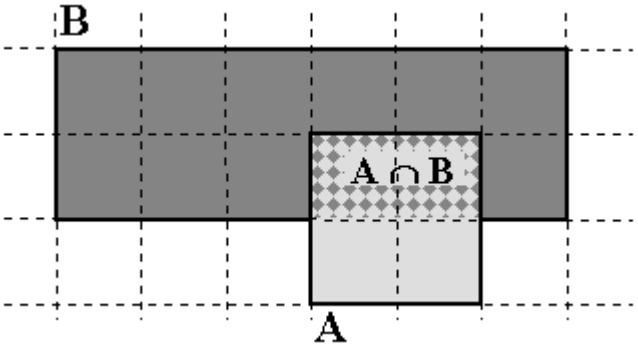

Figure 1: Two sets A and B

Let $e \in A \cup B$. Let $P(A)$ be the probability that $e \in A$ and $P(A \cap B|A)$ be the conditional probability that $e \in A \cap B$ given $e \in A$. We define $P(B)$ and $P(A \cap B|B)$ analogically.

Develop an estimator for the ratio $S(A)/S(B)$ between the size $S(A)$ of A and the size $S(B)$ of B using only $P(A)$, $P(B)$, $P(A \cap B|A)$, and $P(A \cap B|B)$.

Assume now that we have two search engines instead of A and B, and that we know the size of one of them. Estimate the size of the other search engine in a similar fashion. Suggest a sampling procedure for the search engines using only the query interface. Discuss what the problems can be expected during a sampling procedure.

Further reading: K. Bharat, A. Broder: A technique for measuring the relative size and overlap of public web search engines, WWW Conference, 1998

Exercise 1.2: Evolving graph models

Recall the preferential attachment algorithm:

- Start with a set of M_0 nodes.
- When a new node is added, add $m \leq M_0$ random edges, where the probability of adding an edge to the node v is $\frac{deg(v)}{\sum deg(v')}$, with $\sum deg(v')$ the sum of degrees of all the nodes in the graph.

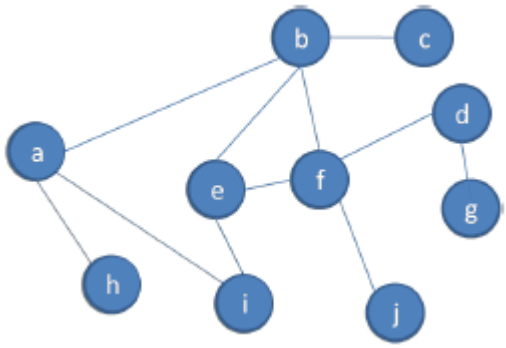

Figure 2: Example graph G

Now consider the undirected graph in Figure 2. Compute the probabilities of the nodes getting an edge during a process of preferential attachment with $m = 4$.

Exercise 1.3: Centrality Measures

Consider again the undirected graph $G = (V, E)$ in Figure 2. Compute the closeness centrality and the betweenness centrality of the nodes e and f of the graph.

The closeness centrality $C_C(c)$ of a node v is defined as

$$C_C(v) = \frac{1}{\sum_{w \in V} d(v, w)}$$

The betweenness centrality $C_B(v)$ of a node v is calculated as:

$$C_B(v) = \sum_{s \neq v \neq t \neq s \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the number of shortest paths between the nodes s and t and $\sigma_{st}(v)$ is the number of shortest paths between s and t that pass through v.

Give examples when the betweenness centrality is more important than closeness centrality and vice versa.

Exercise 1.4: Benford's Law

Benford's Law states that the leading digit of a large set of numbers often is not uniformly distributed, but follows a different distribution. More specifically, the probability that the leading digit of a base b number is d $(1 \leq d \leq b - 1)$ can be expressed as

$$P(d) = \log_b(d + 1) - \log_b(d) = \log_b(1 + \frac{1}{d})$$

Find an explanation for this law (hint: assume that numbers are generated from an exponential growth process). Can you find examples for large sets of numbers where Benford's law does not hold, and can you characterize these sets? Could you think of applications of Benford's law?