Episode 3 exercise: All pairs shortest path length distribution

Rasmus Pagh

August 21, 2013

Background. It is a famous fact that social networks are highly connected, and that two arbitrary people are most often connected by a rather short path of acquaintances. Suppose you would like to figure out how well connected you are on, say, Facebook, you could run BFS to figure out the number of people at distance 1, 2, 3, etc. However, if such a computation is needed for every node it would be prohibitively expensive on a Facebook-size graph.

Estimation based on min-wise hashing. Following Cohen (FOCS 1994) and Backstrom et al (Web Science 2012) we consider an efficient solution to this problem based on min-wise hashing. Let $S_\ell(v)$ denote the number of nodes reachable from v by a path of length at most ℓ.

a) Suppose that a node v has neighbors $w_1, \ldots, w_t$, and we have computed minhashes for $S_\ell(w_1), \ldots, S_\ell(w_t)$. Argue that based on these we can construct a minhash for $S_{\ell+1}(v)$.

b) Based on the above, argue that it is possible to construct a minhash of size k of each set $S_i(v)$, $v \in V$ and $i = 1, \ldots, \ell$, in time $O(k\ell|E|)$.

As applications we can efficiently estimate the number of nodes at different distances (e.g. in the Facebook graph there seems to be average distance less than 5), and estimate the overlap of the "social neighborhood" of two users.