

The Web Click Network

Filippo Menczer

Indiana University, USA, and CNLL, ISI Foundation, Turin, Italy

Abstract:

We analyze the traffic-weighted Web host graph obtained from a large sample of real Web users over about seven months. A number of interesting structural properties are revealed by this complex dynamic network, some in line with the well-studied boolean link host graph and others pointing to important differences. We find that while search is directly involved in a surprisingly small fraction of user clicks, it leads to a much larger fraction of all sites visited. The temporal traffic patterns display strong regularities, with a large portion of future requests being statistically predictable by past ones. Given the importance of topological measures such as PageRank in modeling user navigation, as well as their role in ranking sites for Web search, we use the traffic data to validate the PageRank random surfing model.

The ranking obtained by the actual frequency with which a site is visited by users differs significantly from that approximated by the uniform surfing/teleportation behavior modeled by PageRank, especially for the most important sites. To interpret this finding, we consider each of the fundamental assumptions underlying PageRank and show that each is violated by actual user behavior.

Joint work with Mark Meiss, Santo Fortunato, Alessandro Flammini, and Alessandro Vespignani.