Identification in a Stochastic PageRank Model

Cody E. Clifton* Bozenna Pasik-Duncan†

Abstract

A key feature of modern web search engines is the ability to display relevant and reputable pages near the top of the list of query results. The PageRank algorithm, developed by Google co-founders Brin and Page in the late 1990s, provides one way of achieving such a useful hierarchical indexing, by assigning a measure of relative importance, called the PageRank value, to each web page. PageRank is motivated by the inherently hypertextual structure of the World Wide Web; specifically, the idea that pages with more incoming hyperlinks should be considered more popular and that popular pages should rank highly in search results, all other factors being equal.

In this paper, we present a model for PageRank whose dynamics are described by a controlled stochastic system that depends on an unknown parameter. The fact that the value of the parameter is unknown implies that the system is unknown. We establish strong consistency of a least squares estimator for the parameter. Furthermore, motivated by the recent work of Ishii and Tempo on distributed randomized methods for PageRank computation, we show that the least squares estimator remains strongly consistent within a distributed framework.

Keywords – stochastic system, linear system identification, least squares estimation, PageRank algorithm

*Department of Mathematics, The University of Kansas, Lawrence, KS, 66045, cclifton@ku.edu
†Department of Mathematics, The University of Kansas, Lawrence, KS, 66045, bozenna@ku.edu