Abstract Submitted
for the MAR14 Meeting of
The American Physical Society

Distinguishing fiction from non-fiction with complex networks DAVID M. LARUE, LINCOLN D. CARR, LINNEA K. JONES, JOE T. STEVANAK, Colorado School of Mines — Complex Network Measures are applied to networks constructed from texts in English to demonstrate an initial viability in textual analysis. Texts from novels and short stories obtained from Project Gutenberg and news stories obtained from NPR are selected. Unique word stems in a text are used as nodes in an associated unweighted undirected network, with edges connecting words occurring within a certain number of words somewhere in the text. Various combinations of complex network measures are computed for each text's network. Fisher's Linear Discriminant analysis is used to build a parameter optimizing the ability to separate the texts according to their genre. Success rates in the 70% range for correctly distinguishing fiction from non-fiction were obtained using edges defined as within four words, using 400 word samples from 400 texts from each of the two genres with some combinations of measures such as the power-law exponents of degree distributions and clustering coefficients.

David M. Larue
Colorado School of Mines

Date submitted: 14 Nov 2013 Electronic form version 1.4