Research Summary

Alexander Tsiatas

I am examining World Wide Web graph structure in order to eventually develop classifications for nodes based on the local graph structure around them. Eventually, I hope to be able to give good properties of a local "neighborhood" around a node that could potential determine a class of spam websites based on the local graph structure around them. Currently, I am using a dataset of the .uk domain with all nodes in one host condensed into one node for the host, with all edges between two hosts condensed into one weighted edge between the two host nodes. The dataset has labels for many of the nodes as spam, not spam, or borderline. I hope to be able to classify those nodes using clustering algorithms and/or machine learning to identify labeled nodes correctly as well as give predictions for the unlabeled ones.

Dataset source:
Web Spam Challenge: http://webspam.lip6.fr/wiki/pmwiki.php