

## Link-based Classification of Blogs

Smriti Bhagat, Graham Cormode, S. Muthukrishnan, Irina Rozenbaum, Hongyi Xue  
Rutgers University  
Primary Center of Excellence Researcher: S. Muthukrishnan (Rutgers)

**Project Scope:** Information networks like blogs, the web, social networks and so on are rich collections of open source data. One of the chief concerns with user generated data such as blogs is that of trustworthiness. We are developing methods to understand and predict the trustworthiness of blogs. We model this as a multigraph labeling problem: each blog is a node which connects in different ways to other blogs, web pages etc., and the labels capture different notions of trustworthiness. Learning involves propagation of the labels based on features such as the link structure, labels of adjacent nodes, and features such as links to the web, in order to infer the reliability of the unlabeled blog nodes. We propose two general classes of methods for propagating these labels in a multigraph namely, local (which rely on neighborhood for inferring the label) and global (based on features occurring across the whole graph). We apply these methods using the blog author's *age* as a label, on large scale collections of blog data.

**Recent progress:** Our work over recent months has included the following results:

- We have collected and cleaned data from several million blogs and many millions of links across multiple blog networks such as LiveJournal, Xanga, and Blogger.
- We have conducted detailed analysis of the structure and features of the blog networks, described feature distributions, identified language and linking patterns, and identified the prevalence of one individual operating across multiple networks via identifier matching.
- We have developed and tested methods based on the local and global paradigms, and shown that these can be highly accurate on the problem of labeling age and similar labels.

**Future plans:** Our ongoing work is extending our study to include richer feature information, such as the intricate conversations carried out via commenting, with detailed timestamp information, as well as greater use of textual and language features. We are also looking at analysis of personas across information networks. Specifically, identification of personas within a particular geographic region, language or a set defined by interests and further exploring the neighborhood around the personas of interest. We are working on understanding the challenges of maintaining privacy in the light of such analysis.

**Publications:** *No Blog is an Island—Analyzing Connections Across Information Networks* (Under Submission)

*Link-based classification of blogs, and its applications* (In Preparation)