How do we find patterns and anomalies, on graphs with billions of nodes and edges, which do not fit in memory? How to use parallelism for such Tera- or Peta-scale graphs? In this thesis, we propose Pegasus, a large scale graph mining system implemented on the top of the Hadoop platform, the open source version of Mapreduce. Pegasus includes algorithms which help us spot patterns and anomalous behaviors in large graphs.

Pegasus enables the structure analysis on large graphs. We unify many different structure analysis algorithms, including the analysis on connected components, PageRank, and radius/diameter, into a general primitive called GIM-V. GIM-V is highly optimized, achieving good scale-up on the number of edges and available machines. We discover surprising patterns using GIM-V, including the 7-degrees of separation in one of the largest publicly available Web graphs, with ~7 billion edges.

Pegasus also enables the inference and the spectral analysis on large graphs. We design an efficient distributed belief propagation algorithm which infer the states of unlabeled nodes given a set of labeled nodes. We also develop an eigensolver for computing top k eigenvalues and eigenvectors of the adjacency matrices of very large graphs. We use the eigensolver to discover anomalous adult advertisers in the who-follows-whom Twitter graph with 3 billion edges. In addition, we develop an efficient tensor decomposition algorithm and use it to analyze a large knowledge base tensor.

Finally, Pegasus allows the management of large graphs. We propose efficient graph storage and indexing methods to answer graph mining queries quickly. We also develop an edge layout algorithm for better compressing graphs.

Thesis Committee:
Christos Faloutsos, Chair
Tom Mitchell
Garth Gibson
Robert Grossman, University of Chicago