Abstract

The recent emergence of large evolving networks, mainly due to the rise of Online Social Networks (OSNs), brought out the difficulty to gather a complete picture of a network and it opened up the development of new distributed techniques. Due to the constraints imposed by large networks, it is realistic to assume that only local information is available at each node: the list of neighboring nodes that the node connects to. Random walk based and diffusion techniques are notably suitable in this frame work. However, despite many recent studies, several open issues remain to be addressed for such algorithms. This thesis proposes some novel distributed algorithms for sampling, estimation and inference of network functions, and for approximating the spectrum of graph matrices.

The thesis begins by tackling the problem of sampling in spectral domain: the classical problem of finding the dominant eigenvalues and their corresponding eigenvectors of symmetric graph matrices such as adjacency or Laplacian of undirected graphs. By relating the spectrum to a Schrödinger-type differential equation, we develop a scalable technique called "complex power iterations", a variant of power iterations, which gives a simple interpretation of spectrum in terms of peaks at the eigenvalue points. Distributed implementation is then formed with diffusion over the graph and with gossiping algorithms. The relation of quantum random walks with our formulation leads us to a simple algorithm based on quantum computing. Higher order approximations and symplectic numerical integrators are also proposed to solve the underlying differential equation.

Next, we consider sampling and estimation of network functions (aggregate and average) using random walks on graph. In order to avoid the burn-in time of Markov chain sampling, we use the idea of regeneration at the renewal epochs when the random walk revisits a fixed node. This help us to develop an estimator for the aggregate function, which is non-asymptotically unbiased and can be implemented in a massively distributed way. We introduce the idea of a "super-node" as the anchoring node for the renewals, to tackle disconnected or "weakly-knit" graphs. We derive an approximation to the Bayesian posterior of the estimate and it provides a real-time assessment of estimation accuracy. As a cross between the deterministic iteration and Markov sampling, an estimator based on reinforcement learning is also developed making use of the idea of regeneration.

The final part of the thesis deals with the use of extreme value theory to make inference from the stationary samples of the random walks. Extremal events like the first hitting time of a large degree node, order statistics and mean cluster size are well captured in the parameter "extremal index" from extreme value theory. We theoretically study and estimate the extremal indices of different random walk sampling techniques.

The techniques and tools developed in this thesis are tested on real-world networks and show promising results.