REPORT ON "Sampling and Inference in Complex Networks"
By Jithin KAZHUTHUVEETTIL SREEDHARAN

This thesis focuses on problems related to sampling large graphs and making ststistically meaningful inferences of a variety of properties related to the graph. Broadly speaking the paper focuses on three specific problems. The first is that of computing or inferring eigenvalues and eigcenvectors associated with an undirected graph. The second problem regards the design of efficient random walk based estimators that provide unbiased estimates. The third regards the study of the extremal properties of samples collected using some sampling algorithm for the purpose of characterizing dependencies among the samples. Throughout, the treatment of these three topics is thorough and rigorous. Furthermore, all theoretical results and their impliications on the design of practical sampling and inference algorithms are solidly backed by experiments executed on synthetic datasets as well as on real world datasets. This is strong thesis; the results on eigen-decomposition and on the construction of unbiased random walk sampling algorithms constitute significant advances in this area of network sampling and inference. They will find application in the future development of such techniques. In the remainder of this report, I will comment in more detail on each of these problems. I have also annotated the thesis, pointing out grammatical errors and needed clarifications. I will make this available after the defense.

Chapter 1 provides an introduction to the topic of sampling, estimation, and inference of large networks. It is followed by Chapter 2, which provides a thorough review of sampling and inference methods based on random walks. The rest of the thesis consist of our chapters, one on the problem of eigen-decomposition, two on random walk based sampling, and one on the application of extreme value theory to sampling. The thesis concludes with a chapter summarizing the work along with future research directions. I will write more about the technical chapters.

Chapter 3 treats the problem of characterizing the spectrum of a network using distributed computations. By spectrum is meant the eigenvalues and eigenvectors associated with either the network adjacency matrix or the network Laplacian. The chapter presents a number of different algorithms for computing or estimating either the largest k or smallest k eigenvalues for an undirected graph. The key to the different presented algorithms is the use of complex power iterations, which appears serve the following purpose. Ordinarily the eigen-structure of an undirected graph lives in the real space. The proposed complex power iteration concept transforms this decomposition into the complex space where the eigenvalues now appear as frequencies allowing for the application of a rich set of techniques for identifying these "frequencies". Jithin's innovation is in recognizing this and then designing several distributed algorithms that can run on the network under study to compute the set of k eigenvalues. One of these algorithms is based on a quantum random walk; Jithin shows how this algorithmcan be simulated through the use of many classical random walks.

In Chapter 4, Jithin considers networks where the nodes have labels and focuses on the problem of estimating averages of functions of these labels using random walks. Random walks suffer the problem that they provide biased estimates due to large mixing times of the underlying Markov chain. Jithin addresses this problem in an elegant way. He recognizes that the return of a random

walk to a specific node constitutes a renewal point and that the behaviors of the walk before and after the return are independent and statistically identical. Recognizing that there exists no node in the network such that the time between visits will be small, Jithin proposes to transform the network into a smaller one where a subset of nodes have been replaced with one "super" node, connected to the rest of the network in a way to represent how the individual nodes making up the super node were connected with in the original network. Using this super node as a renewal point in the new network, Jithin presents an unbiased estimator for the average of functions of labels in the original network. This includes derivations of confidence intervals, mean square errors, etc. The last part of the chapter shows this to be a powerful and effective approach to estimation.

Chapter 5 presents some prelimnary ideas on how to apply reinforcement learning to the problem of estimating averages of functions.

Chapter 6 shifts gears and focuses on the application of extreme value theory to the study of dependence in sequences of observations taken from a crawl of the network. Suppose the label is real valued (degree), and one sets a threshold and look at the statistics of the number of observations that either exceed or lie below that threshold. They will depend on dependencies within the sequence itself. Moreover they relate to a metric known as the extremal index (EI). In Chapter 6, Jithin computes the extremal index associated with a sequence of degree observations for a random graph model that includes dependencies for several random walk based samplers. In addition, he develops estimators for EI and shows their utility on several real-world datasets.

The thesis concludes with a summary of the results in Chapter 7 along with a number of interesting open problems.

In summary, this thesis constitutes a very strong and innovative piece of work. The strongest and likely to have significant impact are the results on eigen=decomposition (Ch 3) and on the use of super nodes for the purpose of developing efficient unbiased estimators (Ch 4). These results are likely to have high impact in the area of network sampling and inference Last, I have some annotations in the thesis itself, primarily correcting grammar and spelling, which I will provide to Jithin separately.

In summary I recommend that Jithin be granted a PhD based on this thesis. It is ready to be defended.