CMSC724: Project Proposal
Uncertain Data and Uncertain Graphs

Greg Benjamin, Samet Ayhan, Kishan Sudusinghe
University of Maryland, College Park

March 5, 2012

1 Proposal

The past decade or so has seen significant growth in the field of large-scale graph data. Social networking sites have become tremendously popular, and these sites must efficiently maintain and analyze graphs of friend relationships, circles, and groups. Internet measurement and experimentation requires the collection and manipulation of data modeling a set of real-world nodes and links spanning the entire globe. Large-scale graphs have even found a place in learning algorithms, where edges can be used to represent dependence relations between variables and events, and edge weights can be tuned as the algorithm progresses in order to make efficient decisions.

In addition to this trend of high scalability in graph data, much recent work in this area has explored the idea of uncertainty in graphs. In brief, edges or nodes in a graph often have weights associated with them. If these weights are normalized between 0 and 1, they may be interpreted as the probability of a particular edge or node exisiting in the graph. This has particularly useful applications when one is trying to model and analyze real-world networks, both between hardware components and between people, when links are only semi-permanent at best and there is high churn. There are also many useful applications involving causal links with probabilistic weights in fields like Machine Learning and Natural Language Processing.

Uncertain graphs become particularly interesting from a research perspective when one tries to apply standard graph-theoretic algorithms such as nearest-neighbor, clustering, and clique/subgraph existence to the probabilistic framework. This has been a topic of much work in recent years (see [1], [2], [3]).

We propose to explore this area of uncertain graphs and graph algorithms for our class project this semester. In particular, we believe there is a sort of duality between this area and the broader area of uncertain data management in databases; more precisely, uncertain graphs should be a subset of uncertain data, since graphs with probabilistic edge and node weights are modelled relationally as lists of tuples representing edges and nodes, each with its associated probability. There are many open questions regarding how to operate and query on these probabilities, as well as how to infer the tuple probabilities, or infer how good they are, and we hope to be able to shed some light on some of these open questions.

At the moment, our precise direction of research is a little undefined. We intend to begin with the listed references and perform a literature survey of the area, before identifying a precise problem to work on. The databases group here at Maryland has been a major research participant in this field for many years, and we expect their work to be of particular interest. In addition, we are certainly open to suggestions and advice pertaining to problems of particular interest, and so we would welcome any input offered.

References

[1] Lei Chen, Changliang Wang. Continuous Subgraph Pattern Search over Certain and Uncertain Graph Streams. In *IEEE Transactions on Knowledge and Data Engineering*, pp. 1093-1109, August, 2010. http://www.computer.org/portal/web/csdl/doi/10.1109/TKDE.2010.67

[2] M. Potamias, F. Bonchi, A. Gionis, G. Kollios. 2010. k-Nearest Neighbors in Uncertain Graphs. In *The Proceedings of the VLDB Endowment* (PVLDB), Volume 3 (2010). http://research.yahoo.com/node/3236

[3] Zhaonian Zou, Hong Gao, and Jianzhong Li. 2010. Discovering frequent subgraphs over uncertain graph databases under probabilistic semantics. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '10). ACM, New York, NY, USA, 633-642. DOI=10.1145/1835804.1835885 http://doi.acm.org/10.1145/1835804.1835885