On Uncertain Graphs Modeling and Queries

Arijit Khan* Lei Chen†

*Systems Group, ETH Zurich, Switzerland
† The Hong Kong University of Science and Technology, China
arijit.khan@inf.ethz.ch leichen@cse.ust.hk

ABSTRACT

Large-scale, highly-interconnected networks pervade both our society and the natural world around us. Uncertainty, on the other hand, is inherent in the underlying data due to a variety of reasons, such as noisy measurements, lack of precise information needs, inference and prediction models, or explicit manipulation, e.g., for privacy purposes. Therefore, uncertain, or probabilistic, graphs are increasingly used to represent noisy linked data in many emerging application scenarios, and they have recently become a hot topic in the database research community. While many classical graph algorithms such as reachability and shortest path queries become $\#\mathbf{P}$-complete, and hence, more expensive in uncertain graphs; various complex queries are also emerging over uncertain networks, such as pattern matching, information diffusion, and influence maximization queries. In this tutorial, we discuss the sources of uncertain graphs and their applications, uncertainty modeling, as well as the complexities and algorithmic advances on uncertain graphs processing in the context of both classical and emerging graph queries. We emphasize the current challenges and highlight some future research directions.

1. INTRODUCTION

With the advent of the Internet and the mobile technology, availability of network data have increased dramatically, including the World-Wide Web, social networks, information networks, traffic networks, genome databases, knowledge graphs, medical and government records. Such data are often represented as attributed graphs, where nodes are entities and edges represent relations among these entities. However, uncertainty is evident in graph data due to a variety of reasons, such as noisy measurements, inconsistent, incorrect, and possibly ambiguous information sources, lack of precise information needs, inference and prediction models, or explicit manipulation, e.g., for privacy purposes [2, 1, 3]. In these cases, data is represented as an uncertain graph, that is, a graph whose nodes, edges, and attributes are accompanied with a probability of existence. With the popularity of uncertain data, uncertain graphs are increasingly becoming important in many emerging application domains including biological networks [16], knowledge bases [4], social networks [22], viral marketing [9], road networks [6], crowd sourcing [18], among many others.

Given such popularity of graphs and uncertain data in database research, coupled with the wide variety of works that have been done in the domain of uncertain graphs, our tutorial on uncertain graphs modeling and queries is very relevant. While there are many surveys in the domain of graphs [10, 8] and also in uncertain data management [2, 14], the number of such tutorials and surveys in the areas of uncertain graphs is very few [1, 5], and they were often targeted towards specific sub-problems (e.g., influence maximization [5]), or towards specific sub-areas (e.g., social networks [1]). Hence, our tutorial would be of great interest to the broader database community.

1.1 Tutorial Outline

```
1 Introduction
    1.1 Data as uncertain graphs
              -Sources of uncertain graphs
              -What is uncertain
    1.2 Applications of uncertain graphs
2 Modeling of uncertain graphs
    2.1 Independent probabilities
    2.2 Conditional probabilities
3 Challenges in uncertain graphs processing
4 Queries over uncertain graphs
    4.1 Classical and emerging graph queries
    4.2 Reliability queries: reachability,
    shortest path, and nearest neighbor
    4.2 Pattern matching queries
    4.3 Similarity queries
    4.3 Influence maximization
5 Open Problems
    5.1 Good possible worlds
    5.2 Scalability vs. accuracy
    5.3 Novel queries: centrality,
    partitioning of uncertain graphs
    5.4 Datasets and benchmarks
```

Time. The intended length of our tutorial is 1.5 hours.

2. MODELING OF UNCERTAIN GRAPHS

In an uncertain, or probabilistic, graph, uncertainty can be associated with any one or multiple of the following components, i.e., edge uncertainty, node uncertainty, and attribute uncertainty.

2.1 Uncertainty Models

Independent Probabilities. The bulk of the literature on uncertain graphs assumes the existence of the components in the graph

Proceedings of the VLDB Endowment, Vol. 8, No. 12

independent from one another, and interprets uncertain graphs according to the well-known *possible-world semantics* [7, 15]. For example, an uncertain graph with m edges yields 2^m possible deterministic graphs, which are derived by sampling independently each edge with its corresponding probability.

Correlated Probabilities. While the independent model discussed above is one of the simplest possible way to deal with uncertainty in graph databases, it naturally ignores the correlations among various graph components. For example, in a traffic network, if a road is crowded at a certain point of time, most likely the road in the next intersection would also be crowded. The independent model fails to consider these relationships. There are a few works that model such correlations with conditional probabilities, e.g., [15, 12]; however, this also incurs additional complexity in the problem.

2.2 Challenges

The challenges in uncertain graph processing are both semantics and computation driven. From the perspective of the semantics, there is no uniform model of uncertain graphs; rather assignment and interpretation of the probabilities must be application specific. For example, how can we define the shortest path between two nodes in an uncertain graph? The definition could depend on the application and the specific uncertainty semantic [15]; and therefore, the techniques for processing of uncertain graphs would also vary. From the computation perspective, while many graph algorithms such as subgraph isomorphism are intrinsically hard problems, even the simplest graph algorithms such as reachability and shortest path queries become $\#\mathbf{P}$-complete; and hence, more expensive over uncertain graphs. Therefore, exact computation is almost infeasible with today's large-scale graph data and focus now-a-days is towards designing approximation algorithms with efficient sampling, indexing, and filtering strategies.

3. QUERIES OVER UNCERTAIN GRAPHS

3.1 Reliability Queries

A fundamental problem on uncertain graphs is reliability, which deals with the probability of nodes being reachable one from another. Due to $\#\mathbf{P}$-completeness of the problem [17], we shall discuss both Monte-Carlo sampling and RHT[7] sampling in this tutorial. In addition, we shall discuss several novel distance metrics over uncertain graphs introduced in [15], as well as efficient algorithms for answering shortest path queries over uncertain graphs.

3.2 Pattern Matching Queries

Pattern matching queries over uncertain networks identify all occurrences of a query graph in an uncertain graph with probability of existence more than a predefined threshold. We shall present state-of-the-art indexing and pruning techniques [12, 11, 21, 19] for pattern matching queries over uncertain graphs.

3.3 Similarity-based Search

Other than exact match, similarity search over uncertain graphs is widely used in many real application fields, such as RDF data analysis and predication in biological interaction graphs. We will present the solution framework together with the detailed techniques [20] for subgraph similarity search on uncertain graphs.

3.4 Influence Maximization

Influence maximization aims at finding a set of seed nodes that generates the largest expected information cascade in a social network. We shall discuss various information diffusion models, approximation algorithms with provable performance guarantees to solve the influence maximization problem [9], as well as information diffusion in the presence of multiple campaigners.

4. MAJOR OPEN PROBLEMS

We conclude by discussing the current challenges and some interesting future research directions.

- Is it possible to pre-compute a *good possible world* [13] for various kinds of uncertain graph queries?
- Since an exact computation is often infeasible over large-scale uncertain graphs, it is important to identify the application areas and their specific requirements, e.g. efficiency vs. effectiveness, false positive vs. false negative rates etc., and the algorithm-specific parameters to tune these results.
- With the emergence of uncertain graph applications, it is also important to re-define the semantics of many classical graph operations, e.g., centrality measure and graph partitioning.

5. REFERENCES

[1] E. Adar and C. Re. Managing Uncertainty in Social Networks. *IEEE Data Eng. Bull.*, 30(2):15–22, 2007.

[2] C. C. Aggarwal. *Managing and Mining Uncertain Data*. Springer, 2009.

[3] P. Boldi, F. Bonchi, A. Gionis, and T. Tassa. Injecting Uncertainty in Graphs for Identity Obfuscation. *PVLDB*, 2012.

[4] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *SIGMOD*, 2008.

[5] C. Castillo, W. Chen, and L. V. S. Lakshmanan. Information and Influence Spread in Social Networks. In *KDD*, 2012.

[6] M. Hua and J. Pei. Probabilistic Path Queries in Road Networks: Traffic Uncertainty aware Path Selection. In *EDBT*, 2010.

[7] R. Jin, L. Liu, B. Ding, and H. Wang. Distance-Constraint Reachability Computation in Uncertain Graphs. *PVLDB*, 2011.

[8] Z. Kaoudi and I. Manolescu. Cloud-based RDF Data Management. In *SIGMOD*, 2014.

[9] D. Kempe, J. M. Kleinberg, and E. Tardos. Maximizing the Spread of Influence through a Social Network. In *KDD*, 2003.

[10] A. Khan, Y. Wu, and X. Yan. Emerging Graph Queries in Linked Data. In *ICDE*, 2012.

[11] X. Lian and L. Chen. Efficient Query Answering in Probabilistic RDF Graphs. In *SIGMOD*, 2011.

[12] W. E. Moustafa, A. Kimmig, A. Deshpande, and L. Getoor. Subgraph Pattern Matching over Uncertain Graphs with Identity Linkage Uncertainty. In *ICDE*, 2014.

[13] P. Parchas, F. Gullo, D. Papadias, and F. Bonchi. The Pursuit of a Good Possible World: Extracting Representative Instances of Uncertain Graphs. In *SIGMOD*, 2014.

[14] J. Pei, M. Hua, Y. Tao, and X. Lin. Query Answering Techniques on Uncertain and Probabilistic Data: Tutorial Summary. In *SIGMO*, 2008.

[15] M. Potamias, F. Bonchi, A. Gionis, and G. Kollios. k-Nearest Neighbors in Uncertain Graphs. *PVLDB*, 2010.

[16] P. Sevon, L. Eronen, P. Hintsanen, K. Kulovesi, and H. Toivonen. Link Discovery in Graphs Derived from Biological Databases. In *DILS*, 2006.

[17] L. G. Valiant. The Complexity of Enumeration and Reliability Problems. *SIAM J. on Computing*, 1979.

[18] J. Wang, T. Kraska, M. J. Franklin, and J. Feng. CrowdER: Crowdsourcing Entity Resolution. In *VLDB*, 2012.

[19] Y. Yuan, G. Wang, and L. Chen. Pattern match query in a large uncertain graph. In *CIKM*, pages 519–528, 2014.

[20] Y. Yuan, G. Wang, L. Chen, and H. Wang. Efficient subgraph similarity search on large probabilistic graph databases. *PVLDB*, 5(9):800–811, 2012.

[21] Y. Yuan, G. Wang, H. Wang, and L. Chen. Efficient Subgraph Search over Large Uncertain Graphs. *PVLDB*, 4(11), 2011.

[22] Z. Zou, H. Gao, and J. Li. Discovering Frequent Subgraphs over Uncertain Graph Databases under Probabilistic Semantics. In *KDD*, 2010.