Reliability Maximization in Uncertain Graphs

(Extended Abstract)

Xiangyu Ke
NTU Singapore
xiangyu001@e.ntu.edu.sg

Arijit Khan
NTU Singapore
arijit.khan@ntu.edu.sg

Mohammad Al Hasan
Indiana University-Purdue University
alhasan@cs.iupui.edu

Rojin Rezvansangsari
NTU Singapore
N1600122F@ntu.edu.sg

***Abstract*—Network reliability measures the probability that a target node is reachable from a source node in an uncertain graph, i.e., a graph where every edge is associated with a probability of existence. In this paper, we investigate the novel and fundamental problem of adding a small number of edges in the uncertain network for maximizing the reliability between a given pair of nodes. We study the $\mathbf{NP}$-hardness and the approximation hardness of our problem, and design effective, scalable solutions. Furthermore, we consider extended versions of our problem (e.g., multiple source and target nodes can be provided as input) to support and demonstrate a wider family of queries and applications, including sensor network reliability maximization and social influence maximization.**

I. Introduction

RICH expressiveness of probabilistic graphs and their utility to model the inherent uncertainty in a wide range of applications have prompted a large number of research works on probabilistic graphs by the data management research communities [1]. Uncertainty in a graph arises due to many reasons, including noisy measurements of an edge metric, edge imputation using inference and prediction models, and explicit manipulation of edges, e.g., for privacy purposes.

In an uncertain graph setting, *Network Reliability* is a well-studied problem [2], which requires to measure the probability that a target node is reachable from a source node. In this paper, we investigate the novel problem of adding a small number of edges in an uncertain network for maximizing the reliability between a given pair of nodes. We refer to such edges as *shortcut edges* and the problem of identifying the best set of k edges as the *budgeted reliability maximization* problem. Our problem falls under the broad category of uncertain networks design, optimization, and modification problems [3], yet surprisingly this specific problem has not been studied in the past. The budgeted reliability maximization problem is critical in the context of many real networks, such as transportation and communication networks. An example application is identifying k new connections in road networks or ad-hoc networks (e.g., building new roads, flyovers, adding Ethernet cables) such that the reliability between a pair of important nodes is maximized, where budget k is decided based on resource constraints.

The budgeted reliability maximization problem is a non-trivial one. Our thorough investigation of the budgeted reliability maximization problem have yielded the following theoretical results: (1) We prove that, even assuming polynomial-time sampling methods to estimate reliability, our problem of computing a set of k shortcut edges that maximizes the reliability between two nodes remains $\mathbf{NP}$-hard; (2) the budgeted reliability maximization problem is hard to approximate, as **(i)** it does not admit any $\mathbf{PTAS}$, and **(ii)** the underlying objective function is neither submodular, nor supermodular. Finally, we have the following observations for our problem: (1) The optimal solution may vary with different input probability threshold; (2) the optimal solution may vary when the edge probabilities in the input graph change; (3) for $k_1 < k_2$, the optimal solution with k_1 may not be a subset of that with k_2.

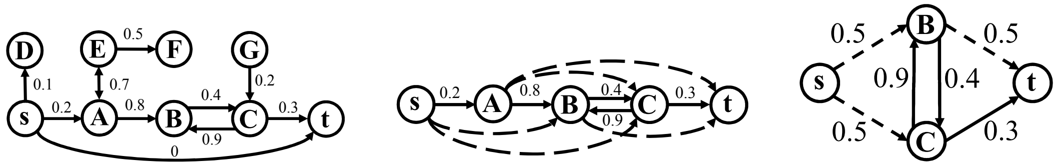
As the optimal solution to our problem varies based on most input parameters, even if the other set of input parameters remains the same, thereby making it non-trivial to utilize pre-existing solutions of past queries, as well as indexing-based or incremental methods. In this paper, we, therefore, propose a practical online algorithm for budgeted reliability maximization problem. We further consider a restricted and one extended version of our problem to support a wider family of queries. In the restricted version, the reliability is estimated only by the most reliable path, thus it can be solved in polynomial-time. In the extended version, multiple sources and targets can be provided as input. The proposed algorithms are generalized to multiple-source-target case.

The research of Arijit Khan is supported by MOE Tier-1 and Tier-2 grants RG117/19 and MOE2019-T2-2-042.

II. Problem and Solution Overview

Problem Statement: An uncertain graph $\mathcal{G}$ is a triple (V, E, p), where V is a set of n nodes, $E \subseteq V \times V$ is a set of m directed edges, and $p(e) \in [0, 1]$ is the probability that the edge $e \in E$ exists. We assume that edge probabilities are independent of each other. Given a source node $s \in V$, a target node $t \in V$, the *reliability* $R(s, t, \mathcal{G})$, also known as the s-t reliability, is defined as the probability that t is reachable from s in $\mathcal{G}$. Suppose we also have a probability threshold $\zeta \in (0, 1]$, and a small positive integer k as budget, our goal is to find the top-k edges to add in $\mathcal{G}$, each with probability $p(e) = \zeta$, so that the reliability from s to t is maximized.

Our proposed solution consists of three steps: (1) reliability-based search space elimination; (2) highly reliable paths extraction; and (3) top-k edges selection. We only provide a brief overview for each part here due to space limitation. Figure 1 provides a run-through example.



(a) Input graph (b) Adding relevant candidate edges (c) Select top-l reliable paths

Fig. 1. Run-through example for the proposed algorithm

TABLE I

SINGLE-SOURCE-TARGET RELIABILITY MAXIMIZATION ON DIFFERENT REAL DATASETS. $k = 10$, $\zeta = 0.5$, $r = 100$, $l = 30$.

Dataset	Reliability Gain				Running Time (sec)				Memory Usage (GB)			
	HC	MRP	IP	BE	HC	MRP	IP	BE	HC	MRP	IP	BE
lastFM (social, $\|v\| = 6.9K$, $\|E\| = 23.7K$)	0.31	0.25	0.30	**0.33**	717	24	**14**	25	0.06	**0.04**	**0.04**	**0.04**
AS_Topology (device, $\|v\| = 45.5K$, $\|E\| = 172.3K$)	**0.42**	0.40	0.41	**0.42**	785	30	**26**	32	0.30	**0.28**	**0.28**	0.29
DBLP (social, $\|v\| = 1.3M$, $\|E\| = 7.1M$)	**0.24**	0.19	0.22	**0.24**	1105	125	**118**	129	6.9	**6.2**	6.5	6.5
Twitter (social, $\|v\| = 6.3M$, $\|E\| = 11.1M$)	0.13	0.11	0.15	**0.19**	1063	141	**126**	148	11.0	**9.4**	9.8	9.8

Reliability-based Search Space Elimination: In a sparse input graph $\mathcal{G}$, one can have as many as $\mathcal{O}(n^2)$ candidate edges. However, given a specific s-t query, all candidate edges may not be equally relevant. If two nodes u and v both have low reliability either from source s, or to target t, then adding an edge between u and v will not improve the s-t reliability significantly. Therefore, we find a set of top-r nodes with highest reliability from s, and another set of top-r nodes with highest reliability to t. We only consider to add an edge from each node in the first set to each node in the second set if it does not exist in the input graph. For the input graph in Figure 1(a), if r is set to be 3, we only keep two sets of nodes $\{s, A, B\}$ and $\{B, C, t\}$, and the dotted edges in Figure 1(b) are those candidate edges to add.

Highly Reliable Paths Extraction: Recent research has shown that what really matters in computing the reliability between two nodes is the set of highly reliable paths between them [4], [5]. After adding the relevant candidate edges found previously, we identify the top-l most reliable paths from s to t with the Eppstein's algorithm [6]. If $l = 3$, the graph in Figure 1(b) will further be reduced to that in Figure 1(c).

Top-k Edges Selection: Our next objective is to find the top-k edges among the candidate set of edges in the reduced graph. The reduced graph consists of top-l most reliable paths from s to t, and we observe that (1) different paths can share same set of candidate edges; (2) the candidate edge set of a path can be a subset of that for another path; and (3) different paths may have different number of candidate edges to be included. Therefore, we design a path batch-based (instead of individual path-based) edge selection algorithm. First, we go through all the paths found in the previous step. If two paths share same set of candidate edges, they will be assigned to the same "path batch". Each path batch is labeled by its candidate edge set, and in our algorithm we include a path batch with highest marginal gain in reliability in every round. For the example in Figure 1, suppose $k = 2$, our method is able to find the optimal solution $\{sC, Bt\}$ (with 0.31 reliability gain), while the baseline method based on individual path selection will return the solution $\{sB, Bt\}$ (with 0.28 reliability gain).

Various algorithms and more details can be found in [7]. In the full paper, we also consider a *restricted* version of our problem, which approximates the reliability by considering *only* the most reliable path between the source and the target node. We prove that improving the probability of the most reliable path can be solved exactly in polynomial time, which yields an efficient algorithm for the restricted version of our problem. Finally, we focus on *generalizations* where multiple source and target nodes can be provided as input, thus opening the stage to a wider family of queries and applications, e.g., network modification for targeted influence maximization.

III. EVALUATION AND CONCLUSION

We conduct thorough experimental evaluation using a variety of settings, parameters, and variables with real-world social networks, device network, and synthetic graphs. Table I depicts the reliability gain, running time, and memory usage performance of our proposed solution (path batches-based edge selection, or in short, BE) against many other baselines: Hill Climbing (HC), most reliable path (MRP), and individual path-based edge selection (IP). This validates the effectiveness and the efficiency of our technique. The details about the baseline methods and more experimental results can be found in [7], where we also illustrate the usefulness of our problem in critical applications such as sensor network reliability maximization and influence maximization in social networks. In future, a total reliability budget on new edges, instead of a fixed/individual budget on each new edge, can be considered. This will add more complexity on selecting proper candidate edges and allocating reliability budget to them.

REFERENCES

[1] A. Khan, Y. Ye, and L. Chen, *On Uncertain Graphs*, Synthesis Lectures on Data Management. Morgan & Claypool Pub., 2018.

[2] M. O. Ball, "Computational Complexity of Network Reliability Analysis: An Overview," *IEEE Tran. Rel.*, vol. 35, no. 3, pp. 230–239, 1986.

[3] J. Peng, B. Zhang, and S. Li, "Towards Uncertain Network Optimization," *J. Uncert. Anal. and Applications*, vol. 3, no. 1, 2015.

[4] W. Chen, C. Wang, and Y. Wang, "Scalable Influence Maximization for Prevalent Viral Marketing in Large-Scale Social Networks," in *KDD*, 2010.

[5] A. Khan, B. Zehnder, and D. Kossmann, "Revenue Maximization by Viral Marketing: A Social Network Host's Perspective," in *ICDE*, 2016.

[6] D. Eppstein, "Finding the k Shortest Paths," *SIAM J. Comput*, vol. 28, no. 2, pp. 652–673, 1998.

[7] X. Ke, A. Khan, M. A. Hasan, and R. Rezvansangsari, "Reliability Maximization in Uncertain Graphs," *IEEE Trans. Knowl. Data Eng.*, 2020 (Accepted).