Chameleon: Towards the Preservation of Privacy and Reliability in Anonymized Uncertain Graphs

Dongqing Xiao, Xiangnan Kong, Mohamed Y. Eltabakh

Computer Science Department, Worcester Polytechnic Institute
Worcester, United States of America
{dxiao, meltabakh, xkong} @wpi.edu

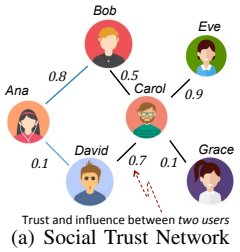
***Abstract*—Many real-world applications involve uncertain graphs, where the existence of the edges is not deterministic but probabilistic. Uncertain graph data are often of great value to researchers. However, such data are rarely released to the public for research due to privacy and security concerns. Conventional approaches are mainly focused on deterministic graphs. In this paper, we study the problem of privacy preserving uncertain graph publishing problem.**

I. Introduction

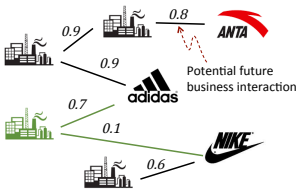
In many prevalent application domains, such as business to business (B2B) [**?**], social networks [**?**], [**?**], and sensor networks [**?**], graphs serve as powerful models to capture the complex relationships inherent in these applications. Most graphs in these applications are uncertain by nature, where each edge carries a degree of uncertainty (probability) representing the probability of its presence in the real world. This uncertainty can be due to various reasons ranging from the use of prediction models to predict the edges (as in social media and B2B networks) to physical properties that affect the edges' reliabilities (as in sensor and communication networks).

These uncertain graphs are of significant importance to support various data mining tasks *e.g.*, understanding graph structures [**?**], [**?**], social interactions [**?**], information discovery and propagation [**?**], advertising and marketing [**?**], among many others. When compared to sharing the results of data mining, data publishing gives greater flexibility because recipients can perform unlimited analysis, data explosion and novel methods. However, the publishing of these uncertain graphs might violate participants' privacy due to the existence of sensitive information.

Motivation Scenario I (Social Trust Networks): *In social networks, the trust and influence relationships among users—which may greatly impact users' behaviors—are usually probabilistic and uncertain [**?**] (See Figure **??**). The existence of the trust relationship depends on many factors, such as the area of expertise and emotional connections. Researchers are very interested in studying the structure of social trust networks, in order to promote products, or choose strategies for a campaign. However, the release of such uncertain graphs with simple anonymization may cause serious privacy issues. The attackers can re-identify private and sensitive information, such as the identity of the users and their trustiness relationship, from the released data.*



(a) Social Trust Network



(b) B2B Network

Figure 1: Examples of real-world uncertain graphs with privacy concerns.

Motivation Scenario II (B2B Networks): *Another uncertain graph example comes from Businesses to Businesses network (See Figure **??**). In these networks, e.g., "Alibaba", nodes represent companies (or businesses in general) while edges represent the trust and the potential of future transactions among them [**?**]. Such future interactions are uncertain since they are obtained by prediction models based on historical data [**?**]. B2B networks can be analyzed and mined for various applications including advertisement targeting [**?**] and customer segmenting [**?**]. Certainly, information about a company's interactions with other companies is considered sensitive data since any leak can be used to infer the company's financial conditions.*

These scenarios show the immediate need for efficient methods privacy preserving uncertain graph publishing, where sanitation or anonymization is applied to the input uncertain graph before publishing. Despite the number of graph anonymization techniques have been proposed [**?**], [**?**], [**?**], [**?**], [**?**], [**?**], [**?**], [**?**], the ignorance of edge uncertainty makes them inefficient for uncertain graph sanitation task. The inefficiency can be due to various reasons ranging from wrong assumption of privacy attacks to improper utility loss metrics. More specifically, the key new challenges in the context of uncertain graphs include:

- **Privacy Attacks leveraging Edge Uncertainty:** In Fig-

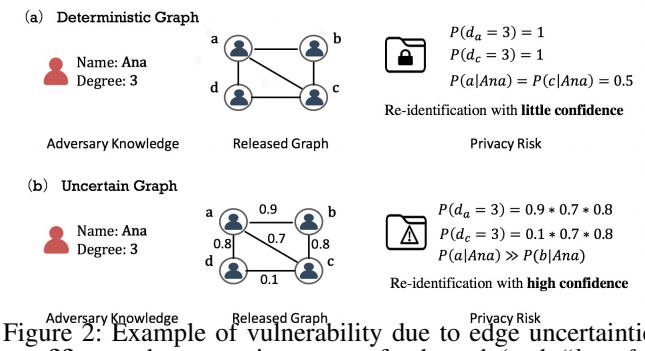


Figure 2: Example of vulnerability due to edge uncertainties.

ure **??**, we show two instances of released (and *"hopefully anonymized"*) graphs with the same exact topology. In Figure **??**(a), the graph is deterministic, while in Figure **??**(b), the graph is uncertain with probabilities associated to its edges. The goal of anonymization is to make nodes indistinguishable in spite of external information. We also assume that the adversary has the same knowledge in both cases, i.e., user `Ana` has a degree of 3 in the original graph. As indicated in the figure, in the case of the deterministic graph, the adversary can *not* re-identify Ana with high confidence since there are two nodes {`a`, `c`} each with the matching degree 3. However, in the uncertain case, the revealed edge uncertainties make the two nodes follow different degree distributions, and thus the adversary has more confidence (around 90%) that `Ana` maps to Node `a`.

This example shows that the release of the associated edge uncertainty increases the potential privacy risk. Therefore, uncertain graph anonymization must take into consideration these edge uncertainties, otherwise, it will fail to protect the privacy correctly. Evidently, ignoring the probabilities altogether and not adding them to the released graph in Figure **??**(b) is not a practical solution as it severely destroys the structure and the utility of the original graph.

• **Appropriate Utility Loss Metric for Uncertain Graphs:** Ideally, the anonymized graph should preserve the privacy with the smallest utility loss for permitting meaningful analysis tasks. Thus, it is crucial to understand and model the utility loss through well-defined metrics. The utility loss metric acts as a safeguard in anonymization process. Various metrics have been proposed for deterministic graphs such as the total number of edge modification [**?**], [**?**], spectrum discrepancy [**?**], community reconstruction error [**?**], [**?**], and shortest path length discrepancy [**?**]. These utility loss metrics have a clear and precise definition in the context of deterministic graphs, which is not the case for uncertain graphs. It is important to understand and model the key properties of *uncertain graph* to be preserved for analysis tasks. Then, the corresponding derived utility loss metric should be incorporated into uncertain graph anonymization as replacement of the aforementioned classical ones.

• **Increased Exponential Complexity of Uncertain Graph Anonymization:** The problem of k-anonymizing a given deterministic graph by as few graph contractions (edge addition, edge deletion, vertex addition and vertex deletion) as possible is shown to be NP-hard [**?**]. Existing techniques usually rely on heuristics to avoid combinatorial intractability. In uncertain graphs, the problem is even harder since an edge operation is no longer a binary operation (addition or deletion), but there can be infinite probability values that can be assigned to each edge. Therefore, efficient *uncertainty-aware* heuristics need to be developed to bring the solution to the realm of feasibility.

In this paper, we present the *"Chameleon"* framework for addressing the aforementioned challenges. Chameleon incorporates edge uncertainties into the core of the anonymization processing such as evaluations of privacy gain and utility loss. In contrast to the classical deterministic graph utility metrics, we propose a new utility metric based on the *reliability* measure—which is a core metric in numerous uncertain graph applications [**?**], [**?**], [**?**]. The anonymization process need to change the graph structure by modifying the edge probabilities of a subset of the edges, which is an exponential search space. Therefore, we propose a ranking algorithm that ranks the edges w.r.t the impact of a change on the graph structure—which we refer to as *"reliability Relevance"*—and that ranking will guide the edge selection process. Moreover, we propose a theoretically-founded probability-alteration strategy based on the entropy of graph degree sequence, which enables achieving maximum privacy gain for an added amount of perturbation.

In summary, the key contributions of this paper are the following:

- Identifying the new and important problem of uncertain graph anonymization where edge uncertainties need to be seamless integrated into the core of the anonymization process. Otherwise, either the privacy will not be protected or the utility will be severely damaged.
- Proposing a new utility-loss metric based on the solid connectivity-based graph model under the possible world semantics, namely the *reliability discrepancy* (Section **??**).
- Introducing a theoretically-founded criterion, called *reliability relevance*, that encodes the sensitivity of the graph edges and vertices to the possible injected perturbation. The criterion will guide the edges' selection during the anonymization process (Section **??**).
- Proposing uncertainty-aware heuristics for efficient edge selection and noise injection over the input uncertain graph to achieve anonymization at a slight cost of reliability (Section **??**).
- Building the Chameleon framework that integrates the aforementioned contributions. Chameleon is experimentally evaluated using several real-world datasets to evaluate its effectiveness and efficiency. The results demonstrate a significant advantage over the conventional methods that do not directly consider edge uncertainties (Section **??**).