Proceedings of the 6th International Conference on Complex Networks and Their Applications, Lyon, France, 2017

Core Decomposition of Uncertain Graphs Using Representative Instances

Damien Seux[1], Fragkiskos D. Malliaros[2], Apostolos N. Papadopoulos[3], and Michalis Vazirgiannis[1]

[1] Computer Science Laboratory, École Polytechnique, France
{damien.seux, michalis.vazirgiannis}@polytechnique.edu
[2] Department of Computer Science and Engineering, University of California San Diego, USA
fmalliaros@eng.ucsd.edu
[3] Department of Informatics, Aristotle University of Thessaloniki, Greece
apostol@csd.auth.gr

1 Introduction and Problem Statement

In many real-world applications, the corresponding graphs are inherently associated with *uncertainty*, which can be due to various reasons, such as uncertainty introduced by the data-collection process or for privacy-preserving reasons. For example, in the case of protein-protein interaction networks (PPI) in the domain of biology, each node corresponds to a specific protein and the edges capture information about the interaction of two proteins. Since in many cases those interactions are indicated either by noisy laboratory experiments or by prediction algorithms based on features of the proteins (instead of being actually observed), a level of uncertainty is introduced in the edges of the graph. This uncertainty can be captured by the model of *uncertain* or *probabilistic* graphs, where each edge is associated with a probability of existence.

In this work, we are interested in a widely applied graph analytics tool, namely the one of *k-core decomposition* [4]. Let H be a subgraph of graph G. Subgraph H is defined to be a *k-core* of G, denoted by G_k, if it is a maximal connected subgraph of G in which all vertices have degree at least k. Based on that, vertex i has *core number* $\text{core}_G(i) = k$, if it belongs to a k-core but not to any $(k+1)$-core. Due to its simplicity and computational efficiency, the k-core decomposition has been applied in many domains, including community detection and identification of influential spreaders in social networks. Then, the following questions arise, which also describe the goals of this work: *how to define the concept of core decomposition in uncertain graphs and how to efficiently compute it?* Bonchi et al. [2] proposed an extension of the k-core decomposition to uncertain graphs which requires that the probability that each vertex v within the core subgraph H has degree at least k, is greater than or equal to a parameter η. Nevertheless, this definition has two main weaknesses: (i) an extra probability threshold η is required in order to define the core structure – making the resulting decomposition dependent on this user-defined parameter; (ii) the increased computational cost for performing the decomposition. Based on that, our goal is to define a simple-yet-effective core decomposition of uncertain graphs. To do so, we consider the *expected degree* of each vertex in the uncertain graph, and in particular the concept of *representative instance*.

2 Cores in Probabilistic Graphs

Let $\mathscr{G} = (V, E, p)$ be an uncertain graph, where $p : E \to (0, 1]$ is a function that assigns probabilities to the edges of the graph. A widely used approach to analyze uncertain graphs is the one of *possible worlds*, where each possible world constitutes a deterministic realization of $\mathscr{G}$. According to this model, an uncertain graph $\mathscr{G}$ is interpreted as a set $\{G = (V, E_G)\}_{E_G \subseteq E}$ of $2^{|E|}$ possible deterministic graphs [3]. Let $G \sqsubseteq \mathscr{G}$ indicates that G is a possible world of $\mathscr{G}$. Then, the probability that $G = (V, E_G)$ is observed as a possible world of $\mathscr{G}$ is given by $\Pr(G|\mathscr{G}) = \prod_{e \in E_G} p(e) \ \prod_{e \in E \setminus E_G} (1 - p(e))$. In our approach, we are using this general framework to derive an analogous of the k-core decomposition in uncertain graphs. In particular, we use the property of *expected degree* $[d](v)$ of each node $v \in \mathscr{G}$, leading to the concept of uncertain $[k]$-core decomposition.

Definition 1 (Uncertain $[k]$-core). *Given an uncertain graph $\mathscr{G} = (V, E, p)$, the uncertain $[k]$-core of $\mathscr{G}$ is the maximal subgraph $\mathscr{H} = (C, E|C, p)$ such that each vertex $v \in C$ has expected degree at least k in $\mathscr{H}$, where $k \in \mathbb{R}^+$.*

According to this definition, in order to compute the decomposition, we can extract a *deterministic representative instance* $G \sqsubseteq \mathscr{G}$ that preserves the expected degree, i.e., the degree of any vertex $v \in G$ to be as close as possible to the expected degree of $v \in \mathscr{G}$ – therefore, casting the problem to a weighted version of the k-core decomposition on deterministic graphs. That way, the proposed algorithm comprises of two phases: (i) extraction of a representative instance of the uncertain graph; (ii) apply a modified version of the k-core decomposition, suitable for fractional degree values.

For the first step of the algorithm that converts the uncertain graph to a deterministic one by preserving the expected degree, we rely on conversion algorithms that aim at minimizing the *discrepancy* of each vertex of the graph [3]. In particular, the discrepancy $\text{dis}_G(v)$ of a vertex v in the representative instance $G \sqsubseteq \mathscr{G}$, is defined as the difference between the degree in the representative instance and expected degree in the uncertain graph, i.e., $\text{dis}_G(v) = d(v) - [d](v)$. As the existence of the edges of the graph are independent of each other, the expected degree $[d](v)$ of a vertex $v \in V$ is the sum of the probabilities of the incident edges, i.e., $[d](v) = \sum_{e=(v,u) \in E} p(e)$. The overall discrepancy of the representative instance $G \sqsubseteq \mathscr{G}$ is defined as $\Delta(G) = \sum_{v \in V} |\text{dis}_G(v)|$. Then, the problem of finding a "good" representative instance G^* can be expressed as a minimization optimization problem: $G^* = \arg\min_{G \sqsubseteq \mathscr{G}} \Delta(G)$.

After extracting an *average degree-preserving* representative instance of the uncertain graph, the core number of a vertex is not an integer anymore but a real number. Thus, the second phase of the proposed technique consists of a modified k-core decomposition algorithm that operates on a *deterministic graph* with fractional node degrees.

3 Experimental Results and Discussion

We have performed preliminary experiments on a co-authorship network (DBLP) derived from DBLP (http://dblp.uni-trier.de), that consists of $404,892$ nodes and $1,422,263$ edges. Since the DBLP dataset is not inherently uncertain, we are using a method to convert the graph to uncertain by examining the similarity between the neighborhood

Table 1. Properties of the $[k]$-core decomposition on the DBLP graph.

Correlation	*Partial*	*Full*
τ_B	0.66	0.12
r	0.79	−0.02

of two nodes – towards computing the probability of being linked. In particular, we have applied the *Jaccard similarity coefficient* to quantify the similarity of two nodes, following two different approaches. The first one, denoted by *Partial*, is adding a probability of existence only to current edges of the graph. However, it does not allow to consider more pairs of nodes, than the already existing edges in the graph. The second approach, denoted by *Full*, is computing the Jaccard similarity coefficient for all pairs of nodes. To overcome the complexity of examining all possible pairs, we restrict our interest to pairs that have at least one neighbor in common. Then, we compute the core decomposition on both the original graph and the uncertain one, and examine the correlation among them using the Kendall rank correlation coefficient τ_B (which measures how much the ranking output is the same in both cases), and Pearson's correlation r (which measures how much the core numbers are linearly related).

Table 1 depicts the results. As we observe, computing the similarity only for existing edges retains the structural properties of the decomposition in both cases. Nevertheless, notice that in this case the original graph can easily be recovered from the uncertain one (for example, when obfuscating the graph for privacy preserving reasons making it uncertain, this approach is not possible). However, computing the probabilities for all pair of nodes (*Full*), erases all the structural information of cores (both correlation measures are close to zero).

Currently, we are working towards examining practical applications of the proposed $[k]$-core decomposition in uncertain graphs. For example, in the DBLP graph, it would be interesting to conduct an exploratory study comparing the authors (i.e., nodes) belonging to the maximal core subgraph extracted by the algorithms on the deterministic and uncertain graphs respectively. Moreover, we plan to examine the performance of the high core number nodes detected by the proposed decomposition, in the task of influence maximization.

References

1. Batagelj V., Zaversnik M.: An $\mathcal{O}(m)$ algorithm for cores decomposition of networks. arXiv (2003)
2. Bonchi, F., Gullo, F., Kaltenbrunner, A., Volkovich, Y.: Core decomposition of uncertain graphs. The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (KDD), pp. 1316–1325 (2014)
3. Parchas, P., Gullo, F., Papadias, D., Bonchi, F.: Uncertain graph processing through representative instances. ACM Trans. Database Syst., 40(3):20:1–20:39 (2015)
4. Seidman, S. B.: Network Structure and Minimum Degree. Social Networks, 5:269–287 (1983)