Mining Maximal Cliques from an Uncertain Graph

Arko Provo Mukherjee [#1], Pan Xu [*2], Srikanta Tirthapura [#3]

[#] *Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA*
[1] arko@iastate.edu
[3] snt@iastate.edu

[*] *Department of Computer Science, University of Maryland, College Park, MD, USA*
[2] panxu@cs.umd.edu

***Abstract*—We consider mining dense substructures (maximal cliques) from an uncertain graph, which is a probability distribution on a set of deterministic graphs. For parameter $0 < \alpha < 1$, we consider the notion of an α-maximal clique in an uncertain graph. We present matching upper and lower bounds on the number of α-maximal cliques possible within a (uncertain) graph. We present an algorithm to enumerate α-maximal cliques whose worst-case runtime is near-optimal, and an experimental evaluation showing the practical utility of the algorithm.**

I. Introduction

Large datasets often contain information that is uncertain in nature. For example, given people A and B, it may not be possible to definitively assert a relation of the form "A knows B" using available information. Our confidence in such relations are commonly quantified using probability, and we say that the relation exists with a probability of p, for some value p determined from the available information. In this work, we focus on *uncertain graphs*, where our knowledge is represented as a graph, and there is uncertainty in the presence of each edge in the graph. Uncertain graphs have been used extensively in modeling, for example, in communication networks [14], [6], [24], social networks [1], [16], [25], [30], [28], [7], protein interaction networks [3], [4], [41], and regulatory networks in biological systems [20].

Identification of dense substructures within a graph is a fundamental task, with numerous applications in data mining, including in clustering and community detection in social and biological networks [37], the study of the co-expression of genes under stress [42], integrating different types of genome mapping data [17]. Perhaps the most elementary dense substructure in a graph, also probably the most commonly used, is a clique, a completely connected subgraph. We are typically interested in a *maximal clique*, which is a clique that is not contained within any other clique. Enumerating all maximal cliques from a graph is one of the most basic problems in graph mining, and has been applied in many settings, including in finding overlapping communities from social networks [37], [33], [5], [39], finding overlapping multiple protein complexes [13], analysis of email networks [38] and other problems in bioinformatics [18], [15], [46].

While the notion of a dense substructure and methods for enumerating dense substructures are well understood in a deterministic graph, the same is not true in the case of an uncertain graph. This is an important open problem today, given that many datasets increasingly incorporate data that is noisy and uncertain in nature. Uncertainty can result from a lack of data. For example, in constructing a social network from data collected through sensors, some communications between individuals maybe missed, or maybe anonymized [1]. In some cases, relationships themselves are probabilistic in nature; for example, the relation of one person influencing another in a social network [10]. In biological networks such as protein–protein interaction networks, it is known that there are frequent errors in finding interactions and our knowledge is best modeled probabilistically [3].

In this work, we consider the analog of a maximal clique in an uncertain graph. Intuitively, a clique in an uncertain graph is a set of vertices that has a high probability of being a completely connected subgraph. In other words, when we sample from the uncertain graph, this set is likely to form a (deterministic) clique. Finding such sets of vertices enables us to unearth robust communities within an uncertain graph, for example, a group of proteins such that it is likely that each protein interacts with each other protein. We present a systematic study of the problem of identifying cliques within an uncertain graph.

A. Our Contributions

First, we present a precise definition of a maximal clique in an uncertain graph, leading to the notion of an α-maximal clique, for parameter $0 < \alpha \leq 1$. A set of vertices U in an uncertain graph is an α-maximal clique if U is a clique with probability at least α, and there does not exist a vertex set U' such that $U \subset U'$ and U' is a clique with probability at least α. When $\alpha = 1$, the above definition reduces to the well understood notion of a maximal clique in a deterministic graph.

Number of Maximal Cliques: We first consider a basic question on maximal cliques in an uncertain graph: *how many α-maximal cliques can be present within an uncertain*

graph? For deterministic graphs, this question was first considered by Moon and Moser [35] in 1965, who presented matching upper and lower bounds for the largest number of maximal cliques within a graph; on a graph with n vertices, the largest possible number of maximal cliques is $3^{\frac{n}{3}}$[1]. For the case of uncertain graphs, we present the first matching upper and lower bounds for the largest number of α-maximal cliques in a graph on n vertices. We show that for any $0 < \alpha < 1$, the maximum number of α-maximal cliques possible in an uncertain graph is $\binom{n}{\lfloor n/2 \rfloor}$, i.e. there is an uncertain graph on n vertices with $\binom{n}{\lfloor n/2 \rfloor}$ uncertain maximal cliques and no uncertain graph on n vertices can have more than $\binom{n}{\lfloor n/2 \rfloor}$ α-maximal cliques.

Algorithm for Enumerating Maximal Cliques: We present a novel algorithm, *MULE* (*M*aximal *U*ncertain c*L*ique *E*numeration), for enumerating all α-maximal cliques within an uncertain graph. MULE is based on a depth-first-search of the graph, combined with optimizations for limiting exploration of the search space, and a fast way to check for maximality based on an incremental computation of clique probabilities. We present a theoretical analysis showing that the worst-case runtime of MULE is $O(n \cdot 2^n)$, where n is the number of vertices. This is nearly the best possible dependence on n, since our analysis of the number of maximal cliques shows that the size of the output can be as much as $O(\sqrt{n} \cdot 2^n)$. Such worst-case behavior occurs only in graphs that are very dense; for typical graphs, we can expect the runtime of MULE to be far better, as we show in our experimental evaluation. We also present an extension of MULE to efficiently enumerate only large maximal cliques.

Note that the worst–case runtime of our algorithm is not the same as an exhaustive search. The cost of checking whether an uncertain clique is maximal or not can be as large as $\Theta(n^2)$. Considering that there are 2^n subsets of vertices of the graph, exhaustive search has a worst-case runtime of $O\left(n^2 \cdot 2^n\right)$, which is worse than our algorithm by a factor of $O(n)$.

Experimental Evaluation: We present an experimental evaluation of MULE using synthetic as well as real-world uncertain graphs. Our evaluation shows that MULE is practical and can enumerate maximal cliques in an uncertain graph with tens of thousands of vertices, more than hundred thousand edges and more than two million α-maximal cliques. Interestingly, the observed runtime of this algorithm is proportional to the size of the output. The real-world graphs included a protein–protein interaction network, and a collaboration network inferred from DBLP.

B. Related Work

There has been much recent work in the database and data mining communities on mining from uncertain graphs, including shortest paths [45], nearest neighbors [40], clustering [27], enumerating frequent and reliable subgraphs [19], [49], [21], [47], [31], [26], and distance-constrained reachability [22]. Our problem of enumerating dense substructures is different from the problems mentioned above. In particular, the problem of finding reliable subgraphs is one of finding subgraphs that are connected with a high probability. However, these individual subgraphs may be sparse. In contrast, we are interested in finding subgraphs that are not just connected, but also fully connected with a high probability. The most closely related work to ours is on mining cliques from an uncertain graph by Zou et. al [48]. Our work is different from theirs in significant ways as elaborated below.

- While we focus on enumerating all α-maximal cliques in a graph, they focus on a different problem, that of enumerating the k cliques with the highest probability of existence.
- We present bounds on the number of such cliques that could exist, while by definition, their problem requires them to output no more than k cliques.
- We provide a runtime complexity analysis of our algorithm and show that it is near optimal. No runtime complexity analysis was provided for the algorithm presented in [48].
- We also provide an algorithm to enumerate only large maximal uncertain cliques.

There is substantial prior work on maximal clique enumeration from a deterministic graph. A popular algorithm for maximal clique enumeration problem is the Bron-Kerbosch algorithm [8], also based on depth-first-search. Tomita et al. [43] improved the depth-first-search approach through a better strategy for pivot selection; their resulting algorithm runs in time $O(3^{\frac{n}{3}})$, which is worst-case optimal, due to the bound on the number of maximal cliques possible [35]. Further work on enumeration of maximal cliques includes [9], [12], [34], [44], [11], [23], [32].

Our algorithm uses the general structure of search presented in [8], [43]. However, unlike the case of a deterministic maximal clique where it is easy to incrementally maintain the set of vertices that can be added to the clique, for an uncertain graph, this is more complex, since we need to be aware of the change in clique probabilities. Recomputing these can be expensive, and our algorithms reduce this cost through an incremental computation. Our runtime analysis and correctness proof need to take this into account, and do not follow from the analysis in [8] or [43].

Roadmap. We present a problem definition in Section II, bounds on the number of α-maximal cliques in Section III, an algorithm to enumerate all α-maximal cliques in Section IV, followed by experimental results in Section V.

[1]This assumes that 3 divides n. If not, the expressions are slightly different

II. Problem Definition

An uncertain graph is a probability distribution over a set of deterministic graphs. We deal with undirected simple graphs, i.e. there are no self-loops or multiple edges. An uncertain graph is a triple $\mathcal{G} = (V, E, p)$, where V is a set of vertices, $E \subseteq V \times V$ is a set of (possible) edges, and $p : E \rightarrow (0, 1]$ is a function that assigns a probability of existence $p(e)$ to each edge $e \in E$. As in prior work on uncertain graphs, we assume that the existence of different edges are mutually independent events.

Let $n = |V|$ and $m = |E|$. Note that $\mathcal{G}$ is a distribution over 2^m deterministic graphs, each of which is a subgraph of the undirected graph (V, E). This set of possible deterministic graphs is called the set of "possible graphs" of the uncertain graph $\mathcal{G}$, and is denoted by $D(\mathcal{G})$. Note that in order to sample from an uncertain graph $\mathcal{G}$, it is sufficient to sample each edge $e \in E$ independently with a probability $p(e)$.

In an uncertain graph $\mathcal{G} = (V, E, p)$, two vertices u and v are said to be adjacent if there exists an edge $\{u, v\}$ in E. Let the neighborhood of vertex u, denoted $\Gamma(u)$, be the set of all vertices that are adjacent to u in $\mathcal{G}$. The next two definitions are standard, and apply not to uncertain graphs, but to deterministic graphs.

Definition 1. *A set of vertices $C \subseteq V$ is a clique in a graph $G = (V, E)$, if every pair of vertices in C is connected by an edge in E.*

Definition 2. *A set of vertices $M \subseteq V$ is a maximal clique in a graph $G = (V, E)$, if (1) M is a clique in G and (2) There is no vertex $v \in V \setminus M$ such that $M \cup \{v\}$ is a clique in G.*

Definition 3. *In an uncertain graph $\mathcal{G}$, for a set of vertices $C \subseteq V$, the clique probability of C, denoted by $clq(C, \mathcal{G})$, is defined as the probability that in a graph sampled from $\mathcal{G}$, C is a clique. For parameter $0 \leq \alpha \leq 1$, C is called an α-clique if $clq(C, \mathcal{G}) \geq \alpha$.*

For any set of vertices $C \subseteq V$, let E_C denote the set of edges $\{e = \{u, v\} | e \in E, u, v \in C \text{ and } u \neq v\}$, i.e. the set of edges connecting vertices in C.

Observation 1. *For any set of vertices $C \subseteq V$ in $\mathcal{G} = (V, E, p)$, such that C is a clique in $G = (V, E)$, $clq(C, \mathcal{G}) = \prod_{e \in E_C} p(e)$.*

Proof: Let G be a graph sampled from $\mathcal{G}$. The set C will be a clique in G iff every edge in E_C is present in G. Since the events of selecting different edges are independent of each other, the observation follows. ∎

Definition 4. *Given an uncertain graph $\mathcal{G} = (V, E, p)$, and a parameter $0 \leq \alpha \leq 1$, a set $M \subseteq V$ is defined as an α-maximal clique if (1) M is an α-clique in $\mathcal{G}$, and (2) There is no vertex $v \in (V \setminus M)$ such that $M \cup \{v\}$ is an α-clique in $\mathcal{G}$.*

Definition 5. *The* **Maximal Clique Enumeration** *problem in an Uncertain Graph $\mathcal{G}$ is to enumerate all vertex sets $M \subseteq V$ such that M is an α-maximal clique in $\mathcal{G}$.*

The following two observations follow directly from Observation 1.

Observation 2. *For any two vertex sets A, B in $\mathcal{G}$, if $B \subset A$ then, $clq(B, \mathcal{G}) \geq clq(A, \mathcal{G})$.*

Observation 3. *Let C be an α-clique in $\mathcal{G}$. Then for all $e \in E_C$ we have $p(e) \geq \alpha$.*

III. Number of Maximal Cliques

The maximum number of maximal cliques in a deterministic graph on n vertices is known exactly due to a result by Moon and Moser [35]. If $n \mod 3 = 0$, this number is $3^{\frac{n}{3}}$. If $n \mod 3 = 1$, then it is $4 \cdot 3^{\frac{n-4}{3}}$, and if $n \mod 3 = 2$, then it is $2 \cdot 3^{\frac{n-2}{3}}$. The graphs that have the maximum number of maximal cliques are known as Moon-Moser graphs.

For uncertain cliques, no such bound was known so far. In this section, we establish a bound on the maximum number of α-maximal cliques in an uncertain graph. For $0 < \alpha < 1$, let $f(n, \alpha)$ be the maximum number of α-maximal cliques in any uncertain graph with n nodes, without any assumption about the assignments of edge probabilities. The following theorem is the main result of this section.

Theorem 1. *Let $n \geq 2$, and $0 < \alpha < 1$. Then: $f(n, \alpha) = \binom{n}{\lfloor n/2 \rfloor}$*

Proof: We can easily verify that the theorem holds for $n = 2$. for $n \geq 3$, let $g(n) = \binom{n}{\lfloor n/2 \rfloor}$. We show $f(n, \alpha)$ is at least $g(n)$ in Lemma 1, and then show that $f(n, \alpha)$ is no more than $g(n)$ in Lemma 2. ∎

Lemma 1. *For any $n \geq 3$, and any $\alpha, 0 < \alpha < 1$, there exists an uncertain graph $\mathcal{G} = (V, E, p)$ with n nodes which has $g(n)$ α-maximal cliques.*

Proof: First, we assume that n is even. Consider $\mathcal{G} = (V, E, p)$, where $E = V \times V$. Let $\kappa = \binom{n/2}{2}$. For each $e \in E$, let $p(e) = q$ where $q^\kappa = \alpha$. We have $0 < q < 1$ since $0 < \alpha < 1$. Let S be an arbitrary subset of V such that $|S| = n/2$. We can verify that S is an α-maximal clique since (1) the probability that S is a clique is $q^\kappa = \alpha$ and (2) for any set $S' \supsetneq S, S' \subseteq V$, the probability that S' is a clique is at most $qq^\kappa = q\alpha < \alpha$. We can also observe that for any subset $S \subseteq V$, S cannot be an α-maximal clique if $|S| < n/2$ or $|S| > n/2$. Thus we conclude that a subset $S \subseteq V$ is an α-maximal clique iff $|S| = n/2$ which implies that the total number of α-maximal cliques in $\mathcal{G}$ is $\binom{n}{n/2}$. A similar proof applies when n is odd. ∎

Note that our construction in the Lemma above employs the condition that $n \geq 3$ and $0 < \alpha < 1$. When $\alpha = 1$, the upper bound is from the result of Moon and Moser for deterministic graphs, and in this case $f(n, \alpha) = 3^{\frac{n}{3}}$ and

is smaller than $g(n)$. Next we present a useful definition required for proving the next Lemma.

Definition 6. *A collection of sets $\mathscr{C}$ is said to be non-redundant if for any pair $S_1, S_2 \in \mathscr{C}$, $S_1 \neq S_2$, we have $S_1 \nsubseteq S_2$ and $S_2 \nsubseteq S_1$.*

Lemma 2. *$g(n)$ is an upper bound on $f(n,\alpha)$.*

Proof: Let $\mathscr{C}^{\alpha}(\mathscr{G})$ be the collection of all α-maximal cliques in $\mathscr{G}$. Note that by the definition of α-maximal cliques, any α-maximal clique S in $\mathscr{G}$ can not be a proper subset of any other α-maximal clique in $\mathscr{G}$. Thus from Definition 6, for any uncertain graph $\mathscr{G}$, $\mathscr{C}^{\alpha}(\mathscr{G})$ is a non-redundant collection. Hence, it is clear that the largest number of α-maximal cliques in $\mathscr{G}$ should be upper bounded by the size of a largest non-redundant collection of subsets of V.

Let $\mathscr{C}$ be the collection of all subsets of V. Based on $\mathscr{C}$, we construct such an undirected graph $\widehat{G} = (\mathscr{C}, \widehat{E})$ where for any two nodes $S_1 \in \mathscr{C}, S_2 \in \mathscr{C}$, there is an edge connecting S_1 and S_2 iff $S_1 \subseteq S_2$ or $S_2 \subseteq S_1$. It can be verified that a sub-collection $\mathscr{C}' \subseteq \mathscr{C}$ is a non-redundant iff $\mathscr{C}'$ is an independent set in $\widehat{G}$. In Lemma 3, we show that $g(n)$ is the size of a largest independent set of $\widehat{G}$, which implies that $g(n)$ is an upper bound for the number of α-maximal cliques in $\mathscr{G}$. ■

Let $\mathscr{C}^*$ be a largest independent set in $\widehat{G}$. Also, let $\mathscr{C}_k \subseteq \mathscr{C}, 0 \leq k \leq n$ be the collection of subsets of V with the size of k. Observe that for each $0 \leq k \leq n$, $\mathscr{C}_k$ is an independent set of $\widehat{G}$. Also let $L(n)$ and $U(n)$ be respectively the minimum and maximum size of sets in $\mathscr{C}^*$. We can show that $L(n)$ and $U(n)$ can be bounded as shown in Lemma 4 and Lemma 5 respectively.

Lemma 3. *For any $n \geq 3$, $|\mathscr{C}^*| = g(n)$.*

Proof: We first consider the case when n is even. By Lemmas 4 and 5, we know $n/2 \leq L(n) \leq U(n) \leq n/2$. Thus we have $L(n) = U(n) = n/2$ which implies $\mathscr{C}^* = \mathscr{C}^*_{n/2}$. Recall that $\mathscr{C}_k \subseteq \mathscr{C}, 0 \leq k \leq n$ is the collection of subsets of V with the size of k.

We have (1) $\mathscr{C}^* = \mathscr{C}^*_{n/2} \subseteq \mathscr{C}_{n/2}$ and (2) $|\mathscr{C}^*| \geq |\mathscr{C}_{n/2}|$ since $\mathscr{C}^*$ is a largest independent set of $\widehat{G}$. Thus we conclude $\mathscr{C}^* = \mathscr{C}_{n/2}$ which has the size of $\binom{n}{(n/2)} = g(n)$.

We next consider the case when n is odd. From Lemmas 4 and 5, we know $(n-1)/2 \leq L(n) \leq U(n) \leq (n+1)/2$. Thus we have $\mathscr{C}^* = \mathscr{C}^*_{(n-1)/2} \bigcup \mathscr{C}^*_{(n+1)/2}$. For notation convenience, we set $n_1 = (n-1)/2, n_2 = (n+1)/2$. Let $\widehat{G}(\mathscr{C}_{n_1}, \mathscr{C}_{n_2})$ be the subgraph of $\widehat{G}$ induced by $\mathscr{C}_{n_1} \cup \mathscr{C}_{n_2}$. We can view $\widehat{G}(\mathscr{C}_{n_1}, \mathscr{C}_{n_2})$ as a bipartite graph with two disjoint vertex sets $\mathscr{C}_{n_1}$ and $\mathscr{C}_{n_2}$ respectively. Observe that $\mathscr{C}^*_{n_1} \subseteq \mathscr{C}_{n_1}$ and $\mathscr{C}^*_{n_2} \subseteq \mathscr{C}_{n_2}$. Let $\widehat{E}(\mathscr{C}^*_{n_1})$ be the set of edges induced by $\mathscr{C}^*_{n_1}$ in $\widehat{G}(\mathscr{C}_{n_1}, \mathscr{C}_{n_2})$. Since $\mathscr{C}^*$ is an independent set of $\widehat{G}$, none of the edges in $\widehat{E}(\mathscr{C}^*_{n_1})$ will have an end in a node of $\mathscr{C}^*_{n_2}$, i.e, all the edges of $\widehat{E}(\mathscr{C}^*_{n_1})$ should have an end falling in $\mathscr{C}_{n_2} \setminus \mathscr{C}^*_{n_2}$. Note that in $\widehat{G}(\mathscr{C}_{n_1}, \mathscr{C}_{n_2})$, all nodes have a degree of n_2. Thus we have:

$$|\widehat{E}(\mathscr{C}^*_{n_1})| = |\mathscr{C}^*_{n_1}| * n_2 \leq |\mathscr{C}_{n_2} \setminus \mathscr{C}^*_{n_2}| * n_2 = (|\mathscr{C}_{n_2}| - |\mathscr{C}^*_{n_2}|) * n_2$$

from which we obtain $|\mathscr{C}^*| = |\mathscr{C}^*_{n_1}| + |\mathscr{C}^*_{n_2}| \leq |\mathscr{C}_{n_2}| = \binom{n}{n_2}$. Note that $\mathscr{C}_{n_2}$ itself is an independent set of $\widehat{G}$ with size $\binom{n}{n_2}$. Thus we conclude that $|\mathscr{C}^*| = \binom{n}{n_2} = g(n)$. ■

Lemma 4. $L(n) \geq \lfloor n/2 \rfloor$

Proof: Let us assume n is an even number. We prove by contradiction as follows. Suppose $L(n) = \ell \leq n/2 - 1$. Let $\mathscr{C}^*_k \subseteq \mathscr{C}^*, L(n) \leq k \leq U(n)$ be the collection of all sets in $\mathscr{C}^*$ which has the size of k, i.e, $\mathscr{C}^*_k = \{S \in \mathscr{C}^* || S| = k\}$. In the following we construct a new collection $\mathscr{C}_{new} \subseteq \mathscr{C}$ which proves to be an independent set in $\widehat{G}$ with the size being strictly larger than $\mathscr{C}^*$. For each $S \in \mathscr{C}^*_{\ell}$, we add to $\mathscr{C}^*$ all subsets of V which has the form as $S \cup \{i\}$ where $i \in V \setminus S$ and remove S from $\mathscr{C}^*$ meanwhile. Let $\mathscr{C}_{new}$ be the collection obtained after we process the same route for all $S \in \mathscr{C}^*_{\ell}$. Mathematically, we have: $\mathscr{C}_{new} = \mathscr{C}_1 \bigcup \mathscr{C}_2$ where $\mathscr{C}_1 = \bigcup_{S \in \mathscr{C}^*_{\ell}} \bigcup_{i \in V \setminus S} \{S \cup \{i\}\}, \mathscr{C}_2 = \mathscr{C}^* \setminus \mathscr{C}^*_{\ell}$. First we show $\mathscr{C}_{new}$ is an independent set of $\widehat{G}$. Arbitrarily choose two distinct sets, say $S_1 \in \mathscr{C}_{new}, S_2 \in \mathscr{C}_{new}, S_1 \neq S_2$. We check all the possible cases one by one:

- $S_1 \in \mathscr{C}_1, S_2 \in \mathscr{C}_1$. We observe that $|S_1| = |S_2| = \ell + 1$ and $S_1 \neq S_2$. Thus no inclusion relation could exist between S_1 and S_2.
- $S_1 \in \mathscr{C}_2, S_2 \in \mathscr{C}_2$. In this case no inclusion relation can exist between S_1 and S_2 since $\mathscr{C}_2$ is an independent set of $\widehat{G}$.
- $S_1 \in \mathscr{C}_1, S_2 \in \mathscr{C}_2$. Since $\mathscr{C}^*_{\ell}$ is the collection of sets in $\mathscr{C}^*$ which has the smallest size ℓ, we get that $|S_2| \geq \ell + 1 = |S_1|$. Therefore there is only one possible inclusion relation existing here, that is $S_1 \subset S_2$. Suppose $S_1 = S'_1 \cup \{i_1\} \subset S_2$ for some $S'_1 \in \mathscr{C}^*_{\ell}$. Thus we get that $S'_1 \subset S_2$ which implies $\mathscr{C}^*$ is not an independent set of $\widehat{G}$. Hence we conclude that no inclusion relation could exist between S_1 and S_2.

Summarizing the analysis above, we get that no inclusion relation could exist between S_1 and S_2 which yields $\mathscr{C}_{new}$ is an independent set of $\widehat{G}$.

Now we prove that $|\mathscr{C}_{new}| > |\mathscr{C}^*|$. Observe that $\mathscr{C}_1$ and $\mathscr{C}_2$ are disjoint from each other; otherwise $\mathscr{C}^*$ is not an independent set. So we have $|\mathscr{C}_{new}| = |\mathscr{C}_1| + |\mathscr{C}_2|$. Note that $|\mathscr{C}^*| = |\mathscr{C}^*_{\ell}| + |\mathscr{C}_2|$ since $\mathscr{C}^*$ is the union of the two disjoint parts $\mathscr{C}^*_{\ell}$ and $\mathscr{C}_2$. Therefore $|\mathscr{C}_{new}| > |\mathscr{C}^*|$ is equivalent to $|\mathscr{C}_1| > |\mathscr{C}^*_{\ell}|$. Let $\widehat{G}(\mathscr{C}^*_{\ell}, \mathscr{C}_1)$ be the induced subgraph graph of $\widehat{G}$ by $\mathscr{C}^*_{\ell} \bigcup \mathscr{C}_1$. Note that $\widehat{G}(\mathscr{C}^*_{\ell}, \mathscr{C}_1)$ can be viewed as a bipartite graph where the two disjoint vertex sets are $\mathscr{C}^*_{\ell}$ and $\mathscr{C}_1$ respectively. In $\widehat{G}(\mathscr{C}^*_{\ell}, \mathscr{C}_1)$ we observe that (1) for each node $S_1 \in \mathscr{C}^*_{\ell}$, its degree $d(S_1) = n - \ell$; (2) for each node $S_2 \in \mathscr{C}_1$, its degree $d(S_2) \leq \ell + 1$. Thus we get that

$|\widetilde{E}| = |\mathscr{C}^*_\ell|(n-\ell) \leq |\mathscr{C}_1|(\ell+1)$. According to our assumption we have $\ell \leq n/2-1$. Thus we have
$|\mathscr{C}^*_\ell|/|\mathscr{C}_1| \leq (\ell+1)/(n-\ell) \leq (n/2)/(n/2+1) < 1$, yielding $|\mathscr{C}^*_\ell| < |\mathscr{C}_1|$ which is equivalent to $|\mathscr{C}^*| < |\mathscr{C}_{new}|$.

So far we have successfully constructed a new collection $\mathscr{C}_{new} \subseteq \mathscr{C}$ such that (1) it is an independent set of $\widehat{G}$ and (2) $|\mathscr{C}_{new}| > |\mathscr{C}^*|$. That contradicts with the fact that $\mathscr{C}^*$ is a largest independent set of $\widehat{G}$. Thus our assumption $\ell \leq n/2-1$ does not hold, which yields $\ell \geq n/2$. For the case when n is odd, we can process essentially the same analysis as above and get $\ell \geq (n-1)/2$. ■

Lemma 5. $U(n) \leq \lceil n/2 \rceil$

Proof: Let us assume n is an even number. Based on $\mathscr{C}^*$, we construct a dual collection $\mathscr{C}^*_{dual}$ as follows: Initialize $\mathscr{C}^*_{dual}$ as an empty collection. For each $S \in \mathscr{C}^*$, we add $V \setminus S$ into $\mathscr{C}^*_{dual}$. Mathematically, we have: $\mathscr{C}^*_{dual} = \bigcup_{S \in \mathscr{C}^*} \{V \setminus S\}$. First we show $\mathscr{C}^*_{dual}$ is an independent set of $\widehat{G}$. Arbitrarily choose two distinct sets, say $V \setminus S_1 \in \mathscr{C}^*_{dual}, V \setminus S_2 \in \mathscr{C}^*_{dual}$, where $S_1 \in \mathscr{C}^*, S_2 \in \mathscr{C}^*, S_1 \neq S_2$. Note that

$$V \setminus S_1 \subset V \setminus S_2 \Leftrightarrow S_1 \supset S_2, V \setminus S_2 \subset V \setminus S_1 \Leftrightarrow S_2 \supset S_1$$

Thus we have that no inclusion relation could exist between $V \setminus S_1$ and $V \setminus S_2$ since no inclusion relation exists between S_1 and S_2 resulting from the fact that $\mathscr{C}^*$ is an independent set of $\widehat{G}$. So we get $\mathscr{C}^*_{dual}$ is an independent set as well.

We can verify that $|\mathscr{C}^*_{dual}| = |\mathscr{C}^*|$. Therefore we can conclude $\mathscr{C}^*_{dual}$ is a largest independent set of $\widehat{G}$. By Lemma 4, we get to know the minimum size of sets in $\mathscr{C}^*_{dual}$ should be at least $n/2$, which yields the maximum size of of sets in $\mathscr{C}^*$ should be at most $n/2$. For the case when n is odd, we can analyze essentially the same as above. ■

IV. Enumeration Algorithm

In this section, we present **MULE** (Maximal Uncertain cLique Enumeration), an algorithm for enumerating all α-maximal cliques in an uncertain graph $\mathscr{G}$, followed by a proof of correctness and an analysis of the runtime. We assume that $\mathscr{G}$ has no edges e such that $p(e) < \alpha$. If there are any such edges, they can be pruned away without losing any α-maximal cliques, using Observation 3. Let the vertex identifiers in $\mathscr{G}$ be $1, 2, \ldots, n$. For clique C, let $\max(C)$ denote the largest vertex in C. For ease of notation, let $\max(\emptyset) = 0$, and let $clq(\emptyset, \mathscr{G}) = 1$.

Intuition: We first describe a basic approach to enumeration using depth-first-search (DFS) with backtracking. The algorithm starts with a set of vertices C (initialized to an empty set) that is an α-clique and incrementally adds vertices to C, while retaining the property of C being an α-clique, until we can add no more vertices to C. At this point, we have an α-maximal clique. Upon finding a clique that is α-maximal, the algorithm backtracks to explore other possible vertices that can be used to extend C, until all possible search paths have been explored. To avoid exploring the same set C more than once, we add vertices in increasing order of the vertex id. For instance, if C was currently the vertex set $\{1,3,4\}$, we do not consider adding vertex 2 to C, since the resulting clique $\{1,2,3,4\}$ will also be reached by the search path by adding vertices $1,2,3,4$ in that order.

MULE improves over the above basic DFS approach in the following ways. First, given a current α-clique C, the set of vertices that can be added to extend C includes only those vertices that are already connected to every vertex within C. Instead of considering every vertex that is greater than $\max(C)$, it is more efficient to track these vertices as the recursive algorithm progresses – this will save the effort of needing to check if a new vertex v can actually be used to extend C. This leads us to incrementally track vertices that can still be used to extend C.

Second, note that not all vertices that extend C into a clique preserve the property of C being an α-clique. In particular, adding a new vertex v to C decreases the clique probability of C by a factor equal to the product of the edge probabilities between v and every vertex in C. So, in considering vertex v for addition to C, we need to compute the factor by which the clique probability will fall. This computation can itself take $\Theta(n)$ time since the size of C can be $\Theta(n)$, and there can be $\Theta(n)$ edges to consider in adding v. A key insight is to reduce this time to $O(1)$ by incrementally maintaining this factor for each vertex v still under consideration. The recursive subproblem contains, in addition to current clique C, a set I consisting of pairs (u,r) such that $u > \max(C)$, u can extend C into an α-clique, and adding u will multiply the clique probability of C by a factor of r. This set I is incrementally maintained and supplied to further recursive calls.

Finally, there is the cost of checking maximality. Suppose that at a juncture in the algorithm we found that I was empty, i.e. there are no more vertices greater than $\max(C)$ that can extend C into an α-clique. This does not yet mean that C is an α-maximal clique, since it is possible there are vertices less than $\max(C)$, but not in C, which can extend C to an α-maximal clique (note that such an α-maximal clique will be found through a different search path). This means that we have to run another check to see if C is an α-maximal clique. Note that even checking if a set of vertices C is an α-maximal clique can be a $\Theta(n^2)$ operation, since there can be as many as $\Theta(n)$ vertices to be potentially added to C, and $\Theta(n^2)$ edge interactions to be considered. We reduce the time for searching such vertices by maintaining the set X of vertices that can extend C, but will be explored in a different search path. By incrementally maintaining probabilities with vertices in I and X, we can reduce the time for checking maximality of C to $\Theta(n)$.

MULE incorporates the above ideas and is described in Algorithm 1.

Algorithm 1: MULE($\mathcal{G}, \alpha$)

Input: $\mathcal{G}$ is the input uncertain graph
Input: $\alpha, 0 < \alpha < 1$ is the user provided probability threshold

$\hat{I} \leftarrow \emptyset$
forall the $u \in V$ **do**
  $\hat{I} \leftarrow \hat{I} \cup \{(u,1)\}$
Enum-Uncertain-MC($\emptyset$, 1 ,$\hat{I}$, $\emptyset$)

Algorithm 2: Enum-Uncertain-MC(C, q, I, X)

Input: We assume $\mathcal{G}$ and α are available as immutable global variables
Input: C is the current Uncertain Clique being processed
Input: $q = clq(C, \mathcal{G})$, maintained incrementally
Input: I is a set of all tuples (u,r), such that $\forall (u,r) \in I$, $u > max(C)$, and $clq(C \cup \{u\}, \mathcal{G}) = q \cdot r \geq \alpha$, i.e. $C \cup \{u\}$ is an α-clique in $\mathcal{G}$
Input: X is a set of all tuples (v,s), such that $\forall (v,s) \in X$, $v \notin C$, $v < max(C)$, and $clq(C \cup \{v\}, \mathcal{G}) = q \cdot s \geq \alpha$, i.e. $C \cup \{v\}$ is an α-clique in $\mathcal{G}$

if $I = \emptyset$ *and* $X = \emptyset$ **then**
  Output C as α-maximal clique
  return
forall the $(u,r) \in I$ *considered in increasing order of u* **do**
  $C' \leftarrow C \cup \{u\}$ // Note $m = max(C') = u$
  $q' \leftarrow q \cdot r$ // $clq(C \cup \{v\}, \mathcal{G})$
  $I' \leftarrow GenerateI(C', q', I)$
  $X' \leftarrow GenerateX(C', q', X)$
  Enum-Uncertain-MC(C', q', I', X')
  $X \leftarrow X \cup \{(u,r)\}$

Algorithm 3: GenerateI(C', q', I)

Input: We assume $\mathcal{G}$ and α are available as immutable global variables

$m \leftarrow max(C')$, $I' \leftarrow \emptyset$, $S \leftarrow \emptyset$
forall the $(u,r) \in I$ **do**
  $S \leftarrow S \cup \{u\}$
$S \leftarrow S \cap \{\Gamma(m)\}$
forall the $(u,r) \in I$ **do**
  if $u > m$ *and* $u \in S$ **then**
    $clq(C' \cup \{u\}, \mathcal{G}) \leftarrow q' \cdot r \cdot p(\{u,m\})$
    if $(clq(C' \cup \{u\}, \mathcal{G})) \geq \alpha$ **then**
      $u' \leftarrow u$
      $r' \leftarrow r \cdot p(\{u,m\})$
      $I' \leftarrow I' \cup \{(u', r')\}$
return I'

Algorithm 4: GenerateX(C', q', X)

Input: We assume $\mathcal{G}$ and α are available as immutable global variables

$m \leftarrow max(C')$, $X' \leftarrow \emptyset$, $S \leftarrow \emptyset$
forall the $(v,s) \in I$ **do**
  $S \leftarrow S \cup \{v\}$
$S \leftarrow S \cap \{\Gamma(m)\}$
forall the $(v,s) \in X$ **do**
  if $v \in S$ **then**
    $clq(C' \cup \{v\}, \mathcal{G}) \leftarrow q' \cdot s \cdot p(\{v,m\})$
    if $(clq(C' \cup \{v\}, \mathcal{G}) \geq \alpha$ **then**
      $v' \leftarrow v$
      $s' \leftarrow s \cdot p(\{v,m\})$
      $X' \leftarrow X' \cup \{(v', s')\}$
return X'

A. Proof of Correctness

In this section we prove the correctness of MULE. Many of the proofs are omitted due to lack of space and can be found in the Technical Report [36].

Theorem 2. *MULE (Algorithm 1) enumerates all α-maximal cliques from an input uncertain graph $\mathcal{G}$.*

Proof: To prove the theorem we need to show the following. First, if C is a clique emitted by Algorithm 1, then C must be an α-maximal clique. Next, if C is an α-maximal clique, then it will be emitted by Algorithm 1. We prove them in Lemmas 8 and 9 respectively. ■

Before proving Lemmas 8 and 9, we prove some properties of Algorithm 2.

Lemma 6. *When Algorithm 2 is called with C' in line 9, I' is a set of all tuples $(u'r')$, where $u' \in V$ and $0 < r' \leq 1$, such that, $\forall (u', r') \in I'$, $u' > max(C')$, and $clq(C' \cup \{u'\}, \mathcal{G}) = q' \cdot r' \geq \alpha$, i.e. $C' \cup \{u'\}$ is an α-clique in $\mathcal{G}$.*

The following observation follows from Lemma 6.

Observation 4. *The input C to Algorithm 2 is an α-clique.*

Lemma 7. *When Algorithm 2 is called with C' in line 9, X' is a set of all tuples (v', s'), where $v' \in V$ and $0 < s' \leq 1$, such that, $\forall (v', s') \in X'$, we have $v' \notin C'$, $v' < \max(C')$, and $(clq(C' \cup \{v'\}, \mathcal{G}) = q' \cdot s') \geq \alpha$, i.e. $C' \cup \{v'\}$ is an α-clique in $\mathcal{G}$.*

Lemma 8. *Let C be a clique emitted by Algorithm 2. Then C is an α-maximal clique.*

Proof: Algorithm 2 emits C in Line 2. From Observation 4, we know that C is an α-clique. We need to show that

C is α-maximal. We use proof by contradiction. Suppose C is non-maximal. This means that there exists a vertex $u \in V$, such that $C \cup \{u\}$ is an α-clique. We know that $I = \emptyset$ when C is emitted. From Lemma 6, we know that there exists no vertex $u \in V$ such that $u > max(C)$ that can extend C. Again, we know that $X = \emptyset$ when C is emitted. Thus from Lemma 7, we know that there exists no vertex $v \in V$ such that $v < \max(C)$ that can extend C. This is a contradiction and hence C is an α-maximal clique. ■

Lemma 9. *Let C be an α-maximal clique in $\mathcal{G}$. Then C is emitted by Algorithm 2.*

B. Runtime Complexity

Theorem 3. *The runtime of MULE (Algorithm 1) on an input graph of n vertices is $O(n \cdot 2^n)$.*

Proof: MULE initializes variables and calls to Algorithm 2, hence we analyze the runtime of Algorithm 2. An execution of the recursive Algorithm 2 can be viewed as a search tree as follows. Each call to Enum-Uncertain-MC is a node of this search tree. The first call to the method is the root node. A node in this search tree is either an internal node that makes one or more recursive calls, or a leaf node that does not make further recursive calls. To analyze the runtime of Algorithm 2, we consider the time spent at internal nodes as well as leaf nodes.

The runtime at each leaf node is $O(1)$. For a leaf node, the parameter $I = \emptyset$, and there are no further recursive calls. This implies that either C is α-maximal ($X = \emptyset$) and is emitted in line 2 or it is non-maximal ($X \neq \emptyset$) but cannot be extended by the loop in line 4 as $I = \emptyset$. Checking the sizes of I and X takes constant time.

We next consider the time taken at each internal node. Instead of adding up the times at different internal nodes, we equivalently add up the cost of the different edges in the search tree. At each internal node, the cost of making a recursive call can be analyzed as follows. Line 5 takes $O(n)$ time as we add all vertices in C to C' and also u. Line 6 takes constant time. Lines 7 and 8 take $O(n)$ time (Lemmas 10 and 11 respectively). Note that lines 5 to 8 can get executed only once in between the two calls. Thus total runtime for each edge of the search tree is $O(n)$.

Note that the total number of calls made to the method method Enum-Uncertain-MC is no more than the possible number of unique subsets of V, which is $O(2^n)$. We see that for internal nodes, time complexity is $O(n)$ and for leaf nodes it is $O(1)$. Hence the time complexity of Algorithm 2 is $O(n \cdot 2^n)$. ■

Thus now we need to prove that lines 7 and 8 take $O(n)$ time. This implies that time complexity of Algorithms 3 and 4 is $O(n)$. We prove the same in Lemmas 10 and 11 respectively. Due to the lack of space, the proofs can be found in the Report [36].

Lemma 10. *The runtime of Algorithm 3 is $O(n)$.*

Lemma 11. *The runtime of Algorithm 4 is $O(n)$.*

Observation 5. *The worst-case runtime of any algorithm that can output all maximal cliques of an uncertain graph on n vertices is $\Omega(\sqrt{n} \cdot 2^n)$.*

Proof: From Theorem 1, we know that the number of maximal uncertain cliques can be as much as $\binom{n}{\lfloor n/2 \rfloor} = \Theta\left(\frac{2^n}{\sqrt{n}}\right)$ (using Stirling's Approximation). Since the size of each uncertain clique can be $\Theta(n)$, the total output size can be $\Omega(\sqrt{n} \cdot 2^n)$, which is a lower bound on the runtime of any algorithm. ■

Lemma 12. *The worst-case runtime of MULE on an n vertex graph is within a $O(\sqrt{n})$ factor of the runtime of an optimal algorithm for Maximal Clique Enumeration on an uncertain graph.*

Proof: The proof follows from Theorem 3 and Observation 5. ■

C. Enumerating Only Large Maximal Cliques

For a typical input graph, many maximal cliques are small, and may not be interesting to the user. Hence it is helpful to have an algorithm that can enumerate only large maximal cliques efficiently, rather than enumerate all maximal cliques. We now describe an algorithm that enumerates every α-maximal clique with more than t vertices, where t is an user provided parameter.

As a first step, we prune the input uncertain graph $\mathcal{G} = (V, E, p)$ by employing techniques described by Modani and Dey [34]. We apply the "Shared Neighborhood Filtering" where edges are recursively checked and removed as follows. First drop all edges $\{u, v\} \in E$, such that $|\Gamma(u) \cap \Gamma(v)| < (t-2)$. Next drop every vertex $v \in V$, that doesn't satisfy the following condition. For vertex $v \in V$, there must exist at least $(t-1)$ vertices in $\Gamma(v)$, such that for $u \in \Gamma(v)$, $|\Gamma(u) \cap \Gamma(v)| < (t-2)$. Let $\mathcal{G}'$ denote the graph resulting from $\mathcal{G}$ after the pruning step.

Algorithm 5 runs on the pruned uncertain graph $\mathcal{G}'$ to enumerate only large maximal cliques. The recursive method in Algorithm 6 differs from Algorithm 2 as follows. Before each recursive call to method Enum-Uncertain-MC-Large (Algorithm 6), the algorithm checks if the sum of the sizes of the current working clique C' and the candidate vertex set I' are greater than the size threshold t. If not, the recursive method is not called. This optimization leads to a substantial pruning of the search space and hence a reduction in runtime.

Lemma 13. *Given an input graph $\mathcal{G}$, LARGE–MULE (Algorithm 5) enumerates every α-maximal clique with more than t vertices.*

V. Experimental Results

We report the results of an experimental evaluation of our algorithm. We implemented the algorithm using Java. We

Algorithm 5: LARGE–MULE($\mathcal{G}, \alpha, t$)

Input: $\mathcal{G}'$ is the input uncertain graph post pruning
Input: $\alpha, 0 < \alpha < 1$ is the user provided probability threshold
Input: $t, t \geq 2$ is the user provided size threshold

$\hat{I} \leftarrow \emptyset$
forall the $u \in V$ **do**
  $\hat{I} \leftarrow \hat{I} \cup \{(u, 1)\}$
Enum-Uncertain-MC-Large($\emptyset$, 1 ,$\hat{I}$, $\emptyset$,t)

Algorithm 6: Enum-Uncertain-MC-Large(C, q, I, X, t)

Input: C is the current Uncertain Clique being processed
Input: q is pre-computed $clq(C, \mathcal{G})$
Input: I is a set of tuples (u, r), such that $\forall (u, r) \in I$, $u > \max(C)$, and $clq(C \cup \{u\}, \mathcal{G}) = q \cdot r \geq \alpha$, i.e. $C \cup \{u\}$ is an α-clique in $\mathcal{G}$
Input: X is a set of tuples (v, s), such that $\forall (v, s) \in X$, $v \notin C$, $v < \max(C)$, and $clq(C \cup \{v\}, \mathcal{G}) = q \cdot s \geq \alpha$, i.e. $C \cup \{v\}$ is an α-clique in $\mathcal{G}$
Input: t is the user provided size threshold

if $I = \emptyset$ *and* $X = \emptyset$ **then**
  Output C as α-maximal clique
  return
forall the $u, r \in I$ *taken in lexicographical ordering of* u **do**
  $C' \leftarrow C \cup \{u\}$ // Note $m = \max(C') = u$
  $q' \leftarrow q \cdot r$ // $clq(C \cup \{v\}, \mathcal{G})$
  $I' \leftarrow GenerateI(C', q', I)$
  if $|C'| + |I'| < t$ **then**
    continue
  $X' \leftarrow GenerateX(C', q', X)$
  Enum-Uncertain-MC-Large(C', q', I', X', t)
  $X \leftarrow X \cup \{(u, r)\}$

ran all experiments on a system with a 3.19 GHz Intel(R) Core(TM) i5 processor and 4 GB of RAM, with heap space configured at 1.5GB.

Input Data:: Details of the input graphs that we used are shown in Table I.

The first set of graphs consists of real world uncertain graphs shared by authors of [49] and [26]. These include a protein-protein interaction (PPI) network of a Fruit Fly obtained by integrating data from the BioGRID [2] database with that form the STRING [3] database, and the DBLP [4] dataset from authors of [26], which is an uncertain network predicting future co-authorship. The PPI network is an uncertain graph where each vertex represents a protein and two vertices are connected by an edge with a probability representing the likelihood of interaction between the the two proteins. The DBLP network represents co-authorship in academic articles. Each vertex in this network represents an author. Two vertices are connected by an edge with a probability that depends on the "strength" of their co-authorship, which is computed as $1 - e^{-c/10}$, where c is the number of papers co–authored.

The second set of graphs was obtained from the Stanford Large Network Collection [29], and includes graphs representing Internet p2p networks, collaboration networks, and an online social network. The p2p-Gnutella graphs represent peer to peer file sharing networks, where each vertex in the graph represents a computer and the edges represent the communication among them. The p2p-Gnutella04, p2p-Gnutella08 and p2p-Gnutella09 graphs represent communications occurring on 4th, 8th and 9th of August, 2002 respectively. The ca-GrQc graph represents the collaboration network among scientist working on General Relativity and Quantum Cosmology. Each vertex in the graph is a scientist and two vertices are connected by an edge if the corresponding scientists have co-authored a paper. Finally the wiki-vote graph represents the voting that occurs while selecting a new wikipedia administrator. Each vertex is either a wikipedia admin or wikipedia user and the edges represent the votes that each admin / user casts in favor of a candidate. The candidate is also a wikipedia user and hence is represented by a vertex in the graph. For all these graphs, the uncertain graphs were created from these deterministic graphs by assigning edge probabilities uniformly at random. Hence these can be considered as semi–synthetic uncertain graphs.

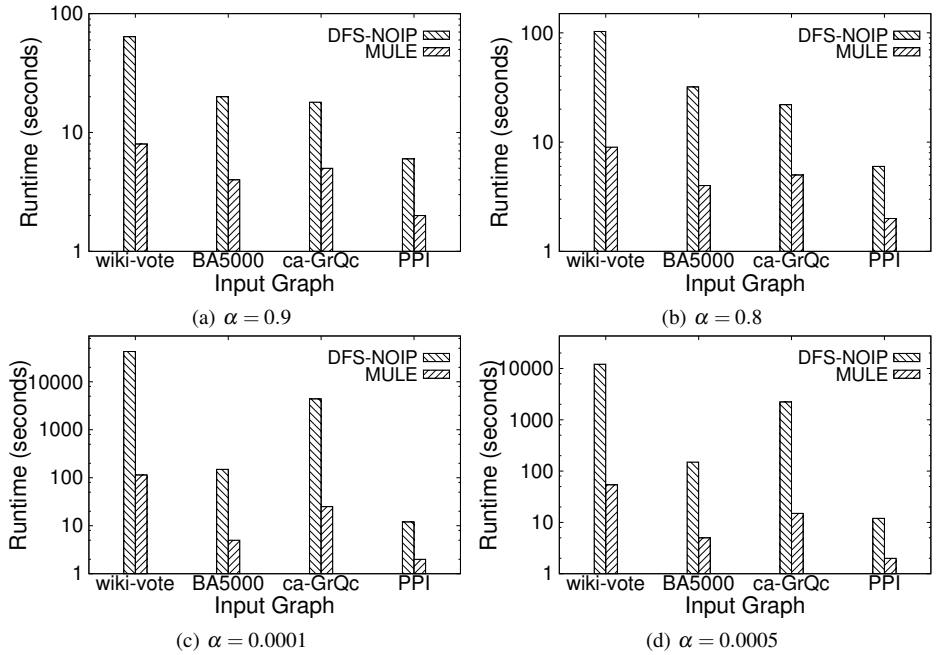
The third set of input graphs was synthetically generated using the Barabási−Albert model for random graphs [2]. Then the edges were assigned probabilities uniformly at random from $[0, 1]$.

Comparison with other approaches. We compare our algorithm with another algorithm based on depth-first-search, which we call DFS-NOIP (DFS with NO Incremental Probability Computation), described in Algorithm 7. This algorithm also performs a depth first search to enumerate all α–maximal cliques but does not compute the probabilities incrementally like MULE does. Figure 1 compares the performance of MULE with DFS–NOIP. The results show that MULE performs much better than DFS–NOIP. For instance, for the graph wiki–vote with $\alpha = 0.9$ DFS–NOIP took 64 seconds while MULE took only 8 secs. The relative performance results hold true over a wide range of input graphs and values of α, including synthetic and real-world graphs, and small and large values of α. For $\alpha = 0.0001$, MULE took only 25 secs to enumerate all maximal cliques in ca-GrQc, while DFS–NOIP took over 4400 secs. On the

[2] http://thebiogrid.org/
[3] http://string-db.org/
[4] http://dblp.uni-trier.de/

Table I: Input Graphs

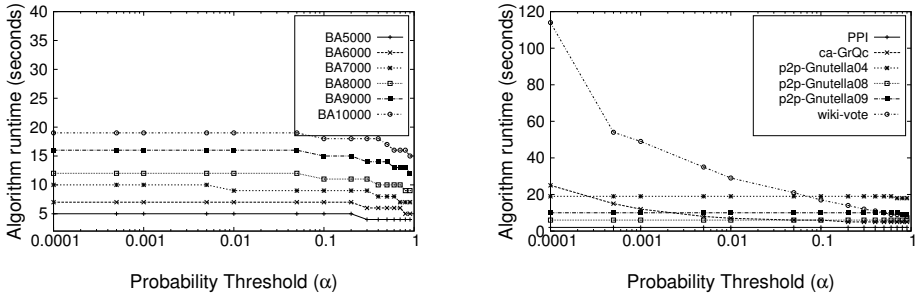
Input Graph	Category	Description	# Vertices	# Edges
Fruit-Fly	Protein Protein Interaction network	PPI for Fruit Fly from STRING Database	3751	3692
DBLP10	Social network	Collaboration network from DBLP	684911	2284991
p2p-Gnutella08	Internet peer-to-peer networks	Gnutella network August 8 2002	6301	20777
p2p-Gnutella04	Internet peer-to-peer networks	Gnutella network August 4 2003	10879	39994
p2p-Gnutella09	Internet peer-to-peer networks	Gnutella network August 9 2003	8114	26013
ca-GrQc	Collaboration networks	Arxiv General Relativity	5242	28980
wiki-vote	Social networks	wikipedia who-votes-whom network	7118	103689
BA5000	Barabási−Albert random graphs	Random graph with 5K vertices	5000	50032
BA6000	Barabási−Albert random graphs	Random graph with 6K vertices	6000	60129
BA7000	Barabási−Albert random graphs	Random graph with 7K vertices	7000	70204
BA8000	Barabási−Albert random graphs	Random graph with 8K vertices	8000	80185
BA9000	Barabási−Albert random graphs	Random graph with 9K vertices	9000	90418
BA10000	Barabási−Albert random graphs	Random graph with 10K vertices	10000	99194



(a) $\alpha = 0.9$ (b) $\alpha = 0.8$

(c) $\alpha = 0.0001$ (d) $\alpha = 0.0005$

Figure 1: Comparison of Simple and Optimized Depth First Search approaches. The Y–Axis is in log–scale.

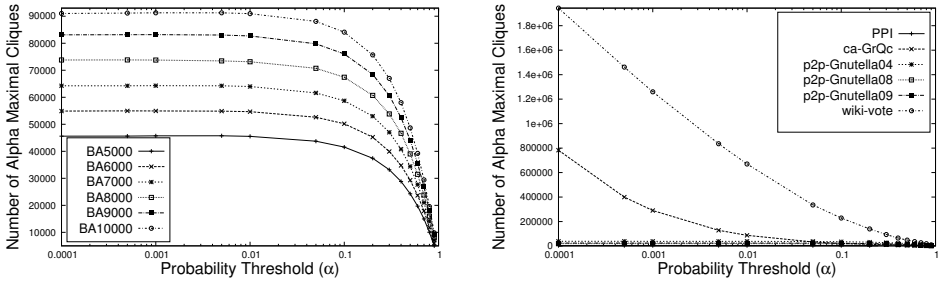


(a) Random Graphs (b) Semi–synthetic and Real Graphs

Figure 2: Runtime vs Alpha (α). The X–Axis is in log–scale

wiki–vote input graph with probability threshold 0.9, MULE took 8 seconds while DFS–NOIP took 64 seconds. For the same graph, with probability threshold 0.0001, MULE took 114 secs, while DFS–NOIP took more than 11 hours.

Dependence on α. We measured the runtime of enumeration as well as the output size, (the number of α-maximal cliques that were output) for different values of α and for the various input graphs described above. The dependence of the runtime on α is shown in Figure 2, and the number of cliques as a function of α is shown in Figure 3. We note that as α increases, the number of maximal cliques, and the time of enumeration both drop sharply. The decrease in runtime



(a) Random Graphs (b) Semi–synthetic and Real Graphs

Figure 3: No of α-maximal cliques vs Alpha (α). The X–Axis is in log–scale

Algorithm 7: DFS–NOIP(C,I)

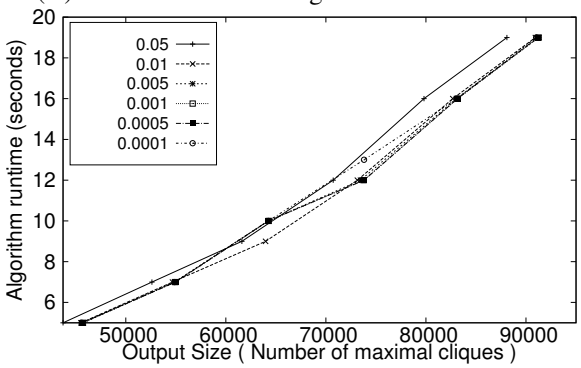
```
1  I_copy ← I
2  forall the u ∈ I_copy do
3      if u ≤ max(C) OR clq(C ∪ {u}) < α then
4          I ← I \ {u}
5  if I = ∅ then
6      if C is an α-maximal clique then
7          Output C as α-maximal clique
8          return
9  forall the v ∈ I do
10     C' ← C ∪ {v}
11     if C' is an α-maximal clique then
12         Output C' as α-maximal clique
13     else
14         I' ← I ∩ Γ(v)
15         DFS–NOIP(C',I')
```

is because with a larger value of α, the algorithm is able to prune search paths aggressively early in the enumeration.

We note that the number of α-maximal cliques does not have to always decrease as α increases. Sometimes it is possible that the number of α-maximal cliques increases with α. This is because as α increases, a large maximal clique may split into many smaller maximal cliques. However, these differences are negligible, and are not visible in the plots.

Dependence on Size of Output. Figure 4 shows the change in runtime with respect to the number of α-maximal cliques enumerated, for the randomly generated graphs. It can be seen that the runtime of the algorithm is almost proportional to the number of maximal cliques in the output. This shows that the algorithm runtime scales well with the number of α-maximal cliques in output. This comparison was not done for real world or semi–synthetic graphs as these graphs have different structural properties, hence different sizes of maximal cliques and thus there is no meaningful way to interpret the results.

Enumerating Large Maximal Cliques. Figures 5 and 6



(a) Random Graphs

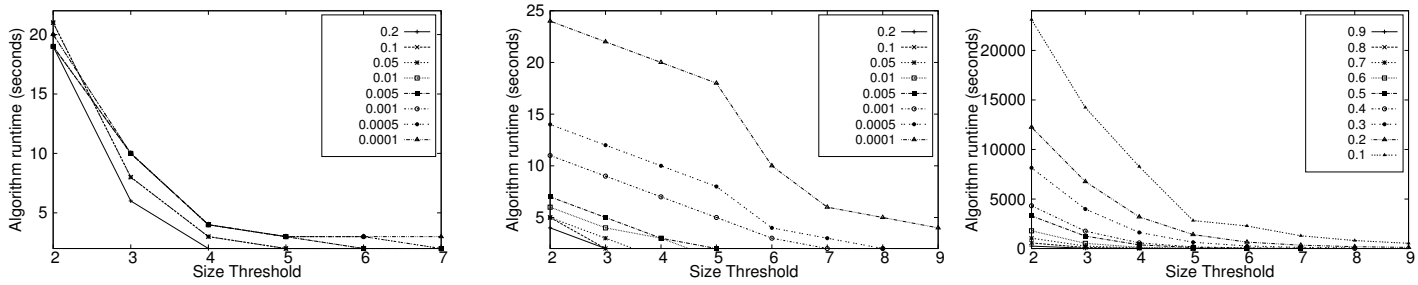
Figure 4: Runtime vs Output Size

show the runtime of LARGE–MULE (Algorithm 5) and the output size respectively as a function of t, the minimum size of an α-maximal clique that is output. As t increases, both runtime and output size decrease substantially. For instance, MULE takes 76797 seconds to enumerate all uncertain maximal cliques from the DBLP dataset (for probability threshold 0.9). However, LARGE–MULE takes only 32 seconds when $t = 3$. Similarly, for input graph ca-GrQc and $\alpha = 0.0001$, MULE takes 125 seconds, while LARGE–MULE takes 10 seconds when $t = 6$ and 6 seconds when $t = 7$.

VI. Conclusion

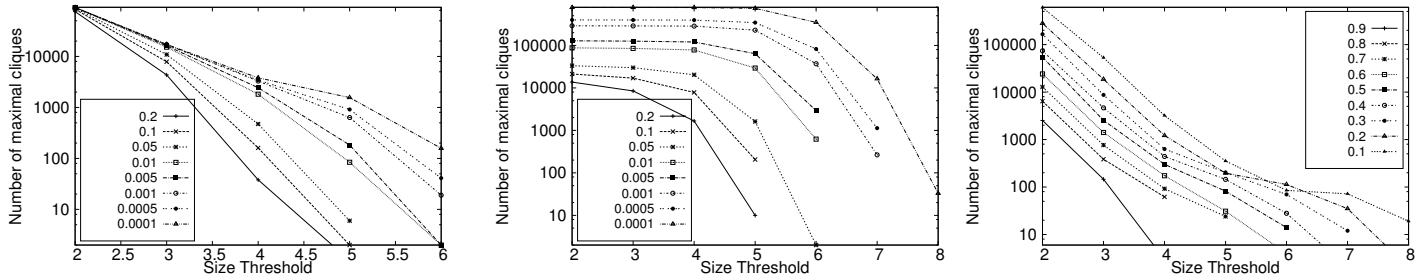
We present a systematic study of the enumeration of maximal cliques from an uncertain graph, starting from a precise definition of the notion of an α-maximal clique, followed by a proof showing that the maximum number of α-maximal cliques in a graph on n vertices is exactly $\binom{n}{\lfloor n/2 \rfloor}$, for $0 < \alpha < 1$. We present a novel algorithm, MULE, for enumerating the set of all α-maximal cliques from a graph, and an analysis showing that the worst-case runtime of this algorithm is $O(n \cdot 2^n)$. We present an experimental evaluation of MULE showing its performance, and an extension for faster enumeration of large maximal cliques.

An interesting open problem is to design an algorithm for enumerating maximal cliques from an uncertain graph whose time complexity is worst-case optimal, $O(\sqrt{n} \cdot 2^n)$. Finally, there are various dense substructures that can be found in a network. Some examples include bicliques, quasi–cliques and k-cores. Finding these dense substructures in the context



(a) BA10000 (b) ca-GrQc (c) DBLP

Figure 5: Runtime vs Size threshold of enumerated uncertain maximal cliques



(a) BA10000 (b) ca-GrQc (c) DBLP

Figure 6: Number of α-maximal cliques vs threshold on minimum size of uncertain maximal clique

of uncertain graphs can be an important future direction of work.

REFERENCES

[1] Eytan Adar and Christopher Re. Managing uncertainty in social networks. *IEEE Data Engineering Bulletin*, 30(2):15–22, 2007.

[2] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, Jan 2002.

[3] Saurabh Asthana, Oliver D. King, Francis D. Gibbons, and Frederick P. Roth. Predicting protein complex membership using probabilistic network reliability. *Genome Research*, 14:1170–1175, 2004.

[4] J.S. Bader, A. Chaudhuri, J.M. Rothberg, and J. Chant. Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology*, 22(1):78–85, 2004.

[5] H.Russell Bernard, Peter D. Killworth, and Lee Sailer. Informant accuracy in social network data iv: a comparison of clique-level structure in behavioral and cognitive network data. *Social Networks*, 2(3):191 – 218, 19791980.

[6] Sanjit Biswas and Robert Morris. Exor: opportunistic multi-hop routing for wireless networks. *ACM SIGCOMM Computer Communication Review*, 35(4):133–144, August 2005.

[7] Paolo Boldi, Francesco Bonchi, Aristides Gionis, and Tamir Tassa. Injecting uncertainty in graphs for identity obfuscation. *Proceedings of the VLDB Endowment*, 5(11):1376–1387, 2012.

[8] Coen Bron and Joep Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *Communications of ACM*, 16(9):575–577, September 1973.

[9] F. Cazals and C. Karande. A note on the problem of reporting maximal cliques. *Theoretical Computer Science*, 407(1-3):564 – 568, 2008.

[10] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 199–208, New York, NY, USA, 2009. ACM.

[11] Norishige Chiba and Takao Nishizeki. Arboricity and subgraph listing algorithms. *SIAM Journal on Computing*, 14:210–223, February 1985.

[12] David Eppstein and Darren Strash. Listing all maximal cliques in large sparse real-world graphs. In Panos Pardalos and Steffen Rebennack, editors, *Experimental Algorithms*, volume 6630 of *Lecture Notes in Computer Science*, pages 364–375. Springer Berlin / Heidelberg, 2011.

[13] Gavin AC et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141 – 147, 2002.

[14] J. Ghosh, H.Q. Ngo, Seokhoon Yoon, and Chunming Qiao. On a routing problem within probabilistic graphs and its application to intermittently connected networks. In *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, pages 1721–1729, 2007.

[15] Helen M Grindley, Peter J Artymiuk, David W Rice, and Peter Willett. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *Journal of Molecular Biology*, 229(3):707–721, 1993.

[16] R. Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th International conference on World Wide Web*, WWW '04, pages 403–412, New York, NY, USA, 2004. ACM.

[17] Eric Harley and Anthony Bonner. Uniform integration of genome mapping data using intersection graphs. *Bioinformatics*, 17(6):487–494, 2001.

[18] Eric Harley, Anthony Bonner, and Nathan Goodman. Uniform integration of genome mapping data using intersection graphs. *Bioinformatics*, 17(6):487–494, 2001.

[19] Petteri Hintsanen and Hannu Toivonen. Finding reliable subgraphs from large probabilistic graphs. *Data Mining and*

Knowledge Discovery, 17(1):3–23, August 2008.

[20] Rui Jiang, Zhidong Tu, Ting Chen, and Fengzhu Sun. Network motif identification in stochastic networks. *Proceedings of the National Academy of Sciences*, 103(25):9404–9409, 2006.

[21] Ruoming Jin, Lin Liu, and Charu C. Aggarwal. Discovering highly reliable subgraphs in uncertain graphs. In *Proceedings of the 17th ACM SIGKDD International conference on Knowledge discovery and data mining*, KDD '11, pages 992–1000, New York, NY, USA, 2011. ACM.

[22] Ruoming Jin, Lin Liu, Bolin Ding, and Haixun Wang. Distance-constraint reachability computation in uncertain graphs. *Proceedings of the VLDB Endowment*, 4(9):551–562, June 2011.

[23] David S. Johnson, Mihalis Yannakakis, and Christos H. Papadimitriou. On generating all maximal independent sets. *Information Processing Letters*, 27(3):119 – 123, 1988.

[24] Haruko Kawahigashi, Y. Terashima, N. Miyauchi, and T. Nakakawaji. Modeling ad hoc sensor networks using random graph theory. In *Second IEEE Consumer Communications and Networking Conference*, pages 104–109, 2005.

[25] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International conference on Knowledge discovery and data mining*, KDD '03, pages 137–146, New York, NY, USA, 2003. ACM.

[26] Arijit Khan, Francesco Bonchi, Aristides Gionis, and Francesco Gullo. Fast reliability search in uncertain graphs. In *Proceedings of the 16th International Conference on Extending Database Technology*, EDBT '14, pages 535–546, New York, NY, USA, 2014. ACM.

[27] George Kollios, Michalis Potamias, and Evimaria Terzi. Clustering large probabilistic graphs. *IEEE Transactions on Knowledge and Data Engineering*, 25(2):325–336, 2013.

[28] Ugur Kuter and Jennifer Golbeck. Using probabilistic confidence models for trust inference in web-based social networks. *ACM Transactions on Internet Technology*, 10(2):8:1–8:23, June 2010.

[29] J. Leskovec. Stanford large network dataset collection.

[30] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the 12th International conference on Information and knowledge management*, CIKM '03, pages 556–559, New York, NY, USA, 2003. ACM.

[31] Lin Liu, Ruoming Jin, C. Aggarwal, and Yelong Shen. Reliable clustering on uncertain graphs. In *IEEE 12th International Conference on Data Mining (ICDM)*, pages 459–468, 2012.

[32] Kazuhisa Makino and Takeaki Uno. New algorithms for enumerating all maximal cliques. In Torben Hagerup and Jyrki Katajainen, editors, *Algorithm Theory - SWAT 2004*, volume 3111 of *Lecture Notes in Computer Science*, pages 260–272. Springer Berlin / Heidelberg, 2004.

[33] Julian Mcauley and Jure Leskovec. Discovering social circles in ego networks. *ACM Transactions on Knowledge Discovery from Data*, 8(1):4:1–4:28, February 2014.

[34] Natwar Modani and Kuntal Dey. Large maximal cliques enumeration in sparse graphs. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 1377–1378, New York, NY, USA, 2008. ACM.

[35] J. Moon and L. Moser. On cliques in graphs. *Israel Journal of Mathematics*, 3:23–28, 1965.

[36] Arko Provo Mukherjee, Pan Xu, and Srikanta Tirthapura. Mining maximal cliques from an uncertain graph. http://arxiv.org/pdf/1310.6780.pdf.

[37] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814 – 818, 2005.

[38] N. Pathak, S. Mane, and J. Srivastava. Who thinks who knows who? socio-cognitive analysis of email networks. In *Sixth International Conference on Data Mining*, pages 466–477, 2006.

[39] Jeffrey Pattillo, Nataly Youssef, and Sergiy Butenko. Clique relaxation models in social network analysis. In *Handbook of Optimization in Complex Networks*, Springer Optimization and Its Applications, pages 143–162. Springer New York, 2012.

[40] Michalis Potamias, Francesco Bonchi, Aristides Gionis, and George Kollios. k-nearest neighbors in uncertain graphs. *Proceedings of the VLDB Endowment*, 3(1-2):997–1008, September 2010.

[41] Daniel R Rhodes, Scott A Tomlins, Sooryanarayana Varambally, Vasudeva Mahavisno, Terrence Barrette, Shanker Kalyana-Sundaram, Debashis Ghosh, Akhilesh Pandey, and Arul M Chinnaiyan. Probabilistic model of the human protein-protein interaction network. *Nature Biotechnology*, 23(8):951–959, 2005.

[42] Oleg Rokhlenko, Ydo Wexler, and Zohar Yakhini. Similarities and differences of gene expression in yeast stress conditions. *Bioinformatics*, 23(2):184–190, 2007.

[43] Etsuji Tomita, Akira Tanaka, and Haruhisa Takahashi. The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical Computer Science*, 363:28–42, October 2006.

[44] Shuji Tsukiyama, Mikio Ide, Hiromu Ariyoshi, and Isao Shirakawa. A new algorithm for generating all the maximal independent sets. *SIAM Journal on Computing*, 6(3):505–517, 1977.

[45] Ye Yuan, Lei Chen, and Guoren Wang. Efficiently answering probability threshold-based shortest path queries over uncertain graphs. In *Database Systems for Advanced Applications*, volume 5981 of *Lecture Notes in Computer Science*, pages 155–170. Springer Berlin Heidelberg, 2010.

[46] Bing Zhang, Byung-Hoon Park, Tatiana Karpinets, and Nagiza F Samatova. From pull-down data to protein interaction networks and complexes with biological relevance. *Bioinformatics*, 24(7):979–986, 2008.

[47] Zhaonian Zou, Hong Gao, and Jianzhong Li. 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, pages 633–642, New York, NY, USA, 2010. ACM.

[48] Zhaonian Zou, Jianzhong Li, Hong Gao, and Shuo Zhang. Finding top-k maximal cliques in an uncertain graph. In *IEEE 26th International Conference on Data Engineering (ICDE)*, pages 649–652, 2010.

[49] Zhaonian Zou, Jianzhong Li, Hong Gao, and Shuo Zhang. Mining frequent subgraph patterns from uncertain graph data. *IEEE TKDE*, 22(9):1203–1218, 2010.