CS 703 Project Proposal - Revised

Xiating Ouyang

1 Motivation

Program synthesis aims at constructing a program which satisfies all the required specifications, and one common form of specification is the input-output examples. One trivial construction is to utilize the if-then-else statements to capture all input cases and return the corresponding output. However, this trivial approach is not scalable when all given examples conform to one program, while the number of cases captured by the synthesized program is proportional to the size of the examples. Another approach is to synthesize all possible programs over every subset of the examples, select the minimum number of the synthesized programs that can cover all examples and finally combine the programs using the branching statements. Despite this approach potentially explores exponential many subsets, in practice however, it is efficient in certain circumstances since the input-output examples follow certain distributions. We hence wish to study different underlying distribution of the input-output examples and explain the efficiency of program synthesis on certain distributions.

Given a set of n input-output examples, we may construct a graph G as follows: introduce a vertex for each input-output example, and connect two vertices if there exists a program consistent with both examples. A necessary condition for the existence of a program consistent on k examples is that their corresponding vertices form a clique in G. Note that the edges in the graph may represent distinct programs. The largest number of examples that a program can be consistent on is hence no greater than the size of the maximum clique in the graph G. In addition, if the the maximum clique has $O(\log n)$ vertices, then the synthesis program only needs to explore all $O(n)$ subsets of these examples to search for consistent programs. Therefore, we focus on deriving an appropriate probablistic graph model for the input-output examples and evaluate the expectation of the maximum clique size to provide an explanation of the efficiency of our synthesis algorithm.

2 Preliminaries

All graphs discussed in this proposal are undirected and simple. A set of vertices in a graph is a *clique* if there is an edge between each pair of vertices within the set. A clique is *maximal* if it is not properly contained in another clique. A *maximum clique* is a clique with the maximum number of vertices. Note that a graph may have multiple maximum cliques. Given a graph G, the size of the maximum clique of G is called the *clique number*, denoted by $\omega(G)$.

3 Problem definition

The problem this project attempts to address can be formulated as:

> Let $\mathcal{G}$ be a probablistic graph model and a random graph G with $G \sim \mathcal{G}$. Compute $\mathbf{E}[\omega(G)]$.

Multiple random graph models have already been carefully studied. One classical example is the Erdős-Rényi model [3] in which each pair of vertices in a graph with n vertices has equal probability p of being adjacent. For example, when $p = 1/2$, the clique number of G is at most $(2 + \varepsilon) \log_2(n) + 1$ with high probability, for all $\varepsilon > 0$ [4]. Power Law graphs characterize the networks in which the number of vertices with a certain degree k is propotional to $1/k^c$ for some constant c. This model is studied in [1], but an estimation of the size of the maximum clique is not present in the work.

4 Action plan

We propose to investigate the power law graph model and derive a bound for the size of the maximum cliques in the power law graphs for different parameter c. Our expectation is that for smaller c the

maximum clique size should be larger, as more vertices have large degrees and hence more likely larger cliques may be present. However, this still needs to be supported by proofs and experiments.

In addition to the power law graph model, we plan to find another appropriate random graph model for our synthesis setting. Hence we propose to research on the existing random graph models with interesting properties[2], and study the input-output examples from real data to extract some succint models.

5 Deliverables and milestones

The milestones for the projects are as follows with deliverables highlighted in **bold**.

- Oct 21: Finish implementing the **Power Law graph model**.
- Oct 23: Finish reviewing the literature[1].
- Oct 26: Finish constructing **new plausible models**.
- Nov 12: Finish implementing and experiments on the new model if possible.
- Nov 18: **Proof sketch** of maximum clique size for power law graph model and new models if available.
- Dec 2: Finish **report** write-up and **slides** preparation.

References

[1] William Aiello, Fan Chung, and Linyuan Lu. A random graph model for massive graphs. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 171–180. ACM, 2000.

[2] David Easley and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[3] Paul Erdos and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[4] Daniel A. Spielman. Random graphs: Markov's inequality, October 2007.

[1] A not so measurable milestone, but still critical.

Appendix: Progress report 1

Implementations of the Erdos-Renyi graph model and power law graph model have been completed and they are available at https://github.com/Lunaticalized/VSAandGraph. In addition to the graph models, a greedy algorithm for computing the maximal cliques in a graph has been implemented and tested on the graph models.

The literature review extends to the references that the Aiello et al. paper includes. The current technical obstacle is to understand the key techniques used in proving the bounds on the component size in the power law graph model. Since the maximum clique size is always bounded above by the size of the connected component that contains it, it could be helpful to try to bound the size of the connected component. And the Aiello et al. paper has already given a $O(\log n)$ bound on the component size, and now it remains to see if that technique can be adopted to prove the bound for other graph models.

However, the current understanding of how the input-output examples are distributed is still limited, and hence the construction of new graph models that fits the input-output example setting is postponed.

s Some additional checkpoints and experiment plans are listed below:

- Nov 20: Construction of new graph model [if not able to, start experimenting on the existing models]
- Nov 25: Finish