Hardness Results for Community Detection

Apoorva Bhagwat, advised by Avrim Blum

School of Computer Science, Carnegie Mellon University

Introduction

Graph clustering is an important problem in many domains. At a high level, given a graph, we want to find sets of related vertices in the graph. This objective can be captured in several ways, so several formulations of the clustering problem exist (such as k-medians). However, until recently, most formulations of clustering did not allow clusters to overlap. Overlapping clusters are of interest in community detection, which motivates the following definition of an (α, β)-cluster [1] in a graph.

Given a graph $G = (V, E)$ (where each vertex has a self-loop), an (α, β)-cluster is a set $S \subseteq V$ that is internally dense and externally sparse, i.e.

For all $u \in S$, u has at least $\alpha|S|$ neighbors in S

For all $v \notin S$, v has at most $\beta|S|$ neighbors in S

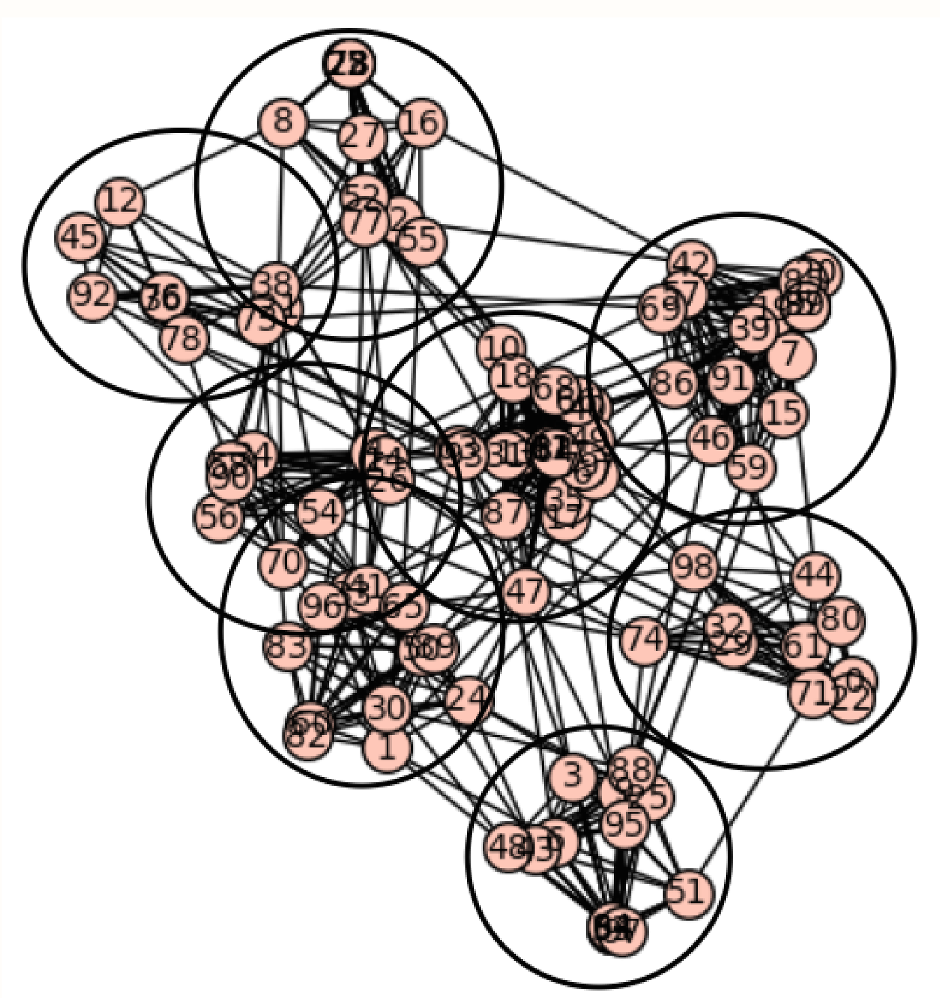

Figure 1: A random graph with overlapping communities

Previous work

Mishra et al [1] introduced this notion and gave an algorithm that finds all (α, β)-clusters with an additional constraint (the 'ρ-champion constraint') in polynomial time. Balcan et al [2] showed instances where the graph has quasi-polynomially many (α, β)-clusters, showed that finding even one cluster is as hard as the hidden clique problem, and gave a quasi-polynomial time algorithm to find all such clusters (through a related, but different notion of (Θ, α, β)-self-determined communities).

Key Result

It is natural to relax the external sparsity condition and analyze the problem. We formulate the problem as follows : given a graph G and a number k, does G have a $\frac{3}{4}$-cluster of size k? We show that this makes the problem NP-hard. We reduce the balanced biclique problem to this problem. The balanced biclique problem is formulated as follows : given a bipartite graph and a size parameter j, is there a copy of $K_{j,j}$ in the graph?

The reduction is as follows : Given a bipartite graph H and a size parameter j, we simply add a clique of size $2j - 1$ to H and fully connect all original vertices in H to the clique. This transformation has the property that the new graph has a $\frac{3}{4}$-cluster of size $4j - 1$ iff H had a balanced biclique of size $2j$.

This reduction generalizes to fractions of the form $\frac{p+1}{p+2}$: we simply add a clique of size $pj - 1$, and make the new size parameter $(p + 2)j - 1$.

Future directions

One immediate hope is to generalize this reduction to any fixed fraction (as opposed to fractions of a specific form). Another possible direction (suggested by Prof Venkat Guruswami) is to approximate (or prove hardness thereof) the density fraction : for example, given a graph G and a size parameter k, is there an induced subgraph of size k with min-degree $\frac{3}{4}$, or do all such subgraphs of size k have min-degree at most $\frac{3}{4} - \epsilon$?

Another direction that we explored was the following : Balcan et al's algorithm finds (α, β)-clusters when α and β are constants and node degrees are proportional to cluster sizes. This can only find dense clusters, and they also show that sparser clusters may be super-polynomial in number. Can we find sparse clusters in polynomial time if we assume that the observed edges are randomly sampled from a denser latent graph?

Acknowledgements

I thank my advisor Prof Blum for his continuous guidance and Prof Venkat Guruswami for the core idea of the reduction.

References

[1] Mishra, Nina, Robert Schreiber, Isabelle Stanton, and Robert E. Tarjan. "Clustering social networks." In International Workshop on Algorithms and Models for the Web-Graph, pp. 56-67. Springer Berlin Heidelberg, 2007.

[2] Balcan, Maria-Florina, Christian Borgs, Mark Braverman, Jennifer Chayes, and Shang-Hua Teng. "Finding endogenously formed communities." In Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 767-783. Society for Industrial and Applied Mathematics, 2013.