Graph Algorithms with Strong and Weak Signals

Eli Friedman and Shahrzad Haddadan

This work was carried out while Eli Friedman was a participant in the 2024 DIMACS REU program at Rutgers University, supported by NSF grant CNS-2150186

June 4, 2024
Strong and Weak Signals

- Two methods to find the similarity of two text documents:
  - Compare the text embeddings produced by word2vec, which is more noisy but cheap to compute
  - Use a large ML model, which is more accurate but expensive to compute

Figure: The Illustrated Word2vec:
https://jalammar.github.io/illustrated-word2vec/
Let $G = (V, E)$ be the true similarity graph, with $(u, v) \in E$ if $u$ is similar to $v$, and $\Gamma(v)$ as the neighborhood of $v$

Strong oracle $O_S(e)$ outputs whether edge $e \in E$

For a noise factor of $\gamma$, the weak oracle $O_\gamma^W(v)$ outputs a subset of $V$ such that $\Gamma(v) \subseteq O_\gamma^W(v)$ and $|O_\gamma^W(v)| \leq (1 + \gamma)|\Gamma(v)|$

The KwikBucks paper uses limited queries to a strong oracle in combination with unlimited queries to a weak oracle in order to solve the correlation clustering problem [Silwal et al., 2023]
Densest Subgraph

- Given an undirected graph $G = (V, E)$, find the set $S$ such that $\frac{|E(S)|}{|S|}$ is maximized
- Real-world applications in online community detection and throughout computational biology

Figure: Finding Dense Subgraphs: https://www.cs.princeton.edu/~zdvir/apx11slides/charikar-slides
Densest Subgraph Algorithm

- Charikar introduced a greedy algorithm that results in a 2-approximation [Charikar, 2000]
  - Iteratively remove the vertex with minimal degree in the remaining graph
- Among other followup papers, Bahmani et al. modify algorithm for large-scale and streaming settings [Bahmani et al., 2012]
  - Each pass, removes all vertices with degree lower than \((2 + 2\varepsilon)\rho(S)\), where \(\rho(S)\) is the density of the remaining graph
- Can we use a weak oracle \(O^\gamma_W\) to solve this problem given a limited budget of queries to the fully accurate \(O_S\)? What other graph theory algorithms could this strong and weak signal setting extend to?
