Minimizing Congestion for Balanced Dominators

Yosuke Mizutani
yos@cs.utah.edu
University of Utah
Salt Lake City, Utah, USA

Annie Staker
annie.staker@utah.edu
University of Utah
Salt Lake City, Utah, USA

Blair D. Sullivan
sullivan@cs.utah.edu
University of Utah
Salt Lake City, Utah, USA

ABSTRACT

A primary challenge in metagenomics is reconstructing individual microbial genomes from the mixture of short fragments created by shotgun sequencing [3]. Recent work of Brown et al. [1], implemented in spacegraphcats[1], leverages the sparsity of the assembly graph to find r-dominating sets which enable rapid approximate queries through a dominator-centric graph partition. Their approach relies on finding an r-dominating set (using Dvorak's approximation algorithm for sparse graphs [2]), then partitioning the assembly graph into bounded-radius *pieces* by assigning each vertex to one of its closest dominators. The process is repeated on the piece graph to form a hierarchy of dominating sets which enables effective navigation and categorization of the data. In this work, we consider two problems related to reducing uncertainty and improving scalability in this setting.

First, we observe that nodes with multiple closest dominators necessitate arbitrary tie-breaking in the existing pipeline. As such, we propose finding *sparse* dominating sets which minimize this effect via a new *congestion* parameter—the average number of dominators appearing in an arbitrary vertex neighborhood. We prove minimizing congestion (formulated as Minimum Congestion r-Dominating Set) is NP-hard, and give an $O(\sqrt{\Delta^r})$ approximation algorithm, where Δ is the max degree. We compare this with the $O(r \log \Delta)$ standard approximation algorithm for finding a (smallest) r-Dominating Set. We note that sparse dominating sets have no explicit size restriction (as shown in Figure 1), and discuss tradeoffs between solution size and congestion.

To improve scalability, the graph should be partitioned into uniformly sized pieces, subject to placing vertices with a closest dominator. This leads to *balanced neighborhood partitioning*: given an r-dominating set, find a partition into connected subgraphs with optimal uniformity so that each vertex is co-assigned with some closest dominator. Using variance of piece sizes to measure uniformity, we show this problem is NP-hard if and only if r is greater than 1. We design and analyze several algorithms, including a polynomial-time approach which is exact when $r = 1$ (and heuristic otherwise).

We complement our theoretical results with computational experiments on a corpus of real-world networks showing sparse dominating sets lead to more balanced neighborhood partitionings. Further, on the metagenome HuSB1, our approach maintains high query containment and similarity while reducing piece size variance.

ACKNOWLEDGMENTS

This work was supported by grant GBMF4560 to Blair D. Sullivan from the Gordon and Betty Moore Foundation.

REFERENCES

[1] C. Titus Brown, Dominik Moritz, Michael P. O'Brien, Felix Reidl, Taylor Reiter, and Blair D. Sullivan. 2020. Exploring neighborhoods in large metagenome assembly graphs using spacegraphcats reveals hidden sequence diversity. *Genome Biology* 21, 1 (06 Jul 2020), 164.

[2] Zdeněk Dvořák. 2013. Constant-factor approximation of the domination number in sparse graphs. *European Journal of Combinatorics* 34, 5 (2013), 833–840.

[3] Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata. 2017. Shotgun metagenomics, from sampling to analysis. *Nature biotechnology* 35, 9 (2017), 833–844.

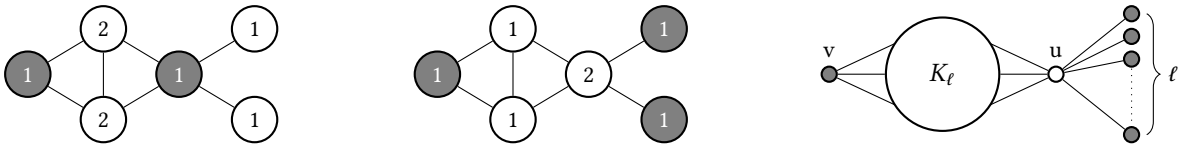

Figure 1: Examples of the difference between solutions to Minimum Dominating Set and Minimum Congestion r-Dominating Set. Dominators are shaded in gray, and vertices are labelled with their congestion. A minimum dominating set (left, size 2) has average congestion $8/6$, whereas a size 3 dominating set (center) achieves average congestion $7/6$. At right, the dominating set $\{v, u\}$ has minimum size, but the set shaded in gray has lower congestion. As ℓ grows, the size of this dominating set can be arbitrarily large.

[1] https://github.com/spacegraphcats/spacegraphcats