IMPROVING NETWORK FAIRNESS WITH EDGE ADDITION

MIHIR PATEL

Abstract. Assuming information, opportunites, and resources will spread across edges in a social network, individual nodes can be more advantaged than others simply due to their position in the graph. Specifically, nodes with higher degree and/or shorter distances to all other nodes will have greater access to, and control over the spread of information across the network. This work formulates several NP-hard optimization problems which quantify advantage gaps between nodes in a social network and proposes network-editing approximation algorithms that reduce this gap. At a high level, we hope to find the connections that should be added to a graph so that nodes possess equal advantage.

Contents

Date: 2024-2-11.

This document is a senior thesis submitted to the Department of Mathematics and Statistics at Haverford College in partial fulfillment of the requirements for a major in Mathematics.

1. Introduction

A graph is a mathematical object defined by a set of vertices/nodes V and edges/links $E \subseteq V \times V$. Graphs are especially useful for modelling real world environments, behaviors, and interactions — as long as there are distinct entities (nodes) which are related under some criteria (edges). For example, a graph could model an airline's network, where airports are represented as nodes, and edges connect airports which have a direct flight between them. This helps travellers find efficient paths between airports and allows airlines to pinpoint airports which have high traffic, but maybe not many facilities. Graphs are also of great practical importance in computer science, and can be used to store large amounts of data efficiently. If each node stores some amount of data, certain edge dispersions across the graph can allow quicker access to data than other more common data structures.

However, we are most interested in *social networks*, graphs which describe interpersonal relations. A typical example is LinkedIn, where nodes represent users and edges connect users who follow each other. Inherent to social networks is the concept of *network fairness* [2, 5, 8]— if LinkedIn users share resources and job opportunities with their followers, who you follow is directly related to how much information you recieve, and how soon this information reaches you. In a social network, certain nodes have more advantage than others simply due to the edge structure across the network.

For example, in Figure 1, the four nodes in the center are generally more advantaged than the four nodes in the periphery. They have more direct connections to other nodes and they are on average closer to all other nodes. If a random node were to post a job opportunity, it is more likely that these nodes in the middle will hear about it before nodes in the periphery. In fact, no matter where this job opportunity "starts", a node in the center will

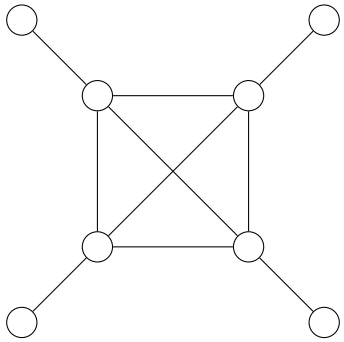

FIGURE 1. In this social network, the four central nodes are more advantaged than the four peripheral nodes as they have greater access to the spread of information across the network.

always hear about a job opportunity before any other node in the periphery. This example validates the idea that some nodes are more advantaged than others, but note that quantifying *how much* more advantaged some nodes are than others is entirely based on specific definitions.

A significant body of research focuses on maximizing the advantage and information access of nodes. Specifically, [2, 3] study the problem of BROADCAST IMPROVEMENT, where the goal is to add edges to a graph to maximize the graph's broadcast, the minimum probability over all vertex pairs u, v that information starting at u will reach v. This notion is similar to the diameter of a graph, except edges each have a probability of transmitting information, i.e. "working". However, less attention has been given to *equalizing* information access across different parts of a network, which is the focus of our work.

At a high level, the proposed work aims to add a budgeted number of edges to a graph to make two vertices equally important in the graph. Formally, given a graph and a budget k, our goal is to augment the graph by adding at most k edges to make the ratio between the closeness centrality of two designated vertices as close to 1 as possible. The closeness centrality of a vertex v is defined as the sum of shortest path distances from v to all other vertices, making it a useful measure of how efficiently a node can access information in the shortest-path metric — if a node is further away from more vertices (higher closeness centrality), it is less likely to receive information quickly, assuming information can arise from any part of the network.

To tackle this problem, we draw insights from existing work on edge augmentation in shortest-path settings. Previous research has studied strategies to maximize the centrality of a single node or a group of nodes [6, 12], although this is still a notably different problem from ours, which seeks to equalize

closeness centralities. We also study Diameter Minimization [1, 4, 7, 11], which seeks to add a limited number of edges to minimize a graph's diameter, the longest shortest path between any two vertices in a graph. These existing approaches provide theoretical guarantees and algorithmic strategies which inform our work, connecting network fairnesss with more classical graph optimization problems and techniques.

1.1. **Problem Statement.** The shortest path length between two vertices $u, v \in V$ within a graph $G = (V, E)$ is represented as $d_G(u, v)$. A graph $G = (V, E)$ augmented with an edge set $S \subseteq V^2 \backslash E$ is denoted as $G + S = (V, E \cup S)$. Finally, we define the closeness centrality of a vertex $v \in V$ in a graph $G = (V, E)$ as $c_G(v) = \sum_{u \in V} d_G(u, v)$, the sum of the shortest paths to each other vertex in the graph. Note that having a smaller closeness centrality value means a node is closer to more vertices in the graph, i.e. more important in the graph.

Given a graph G and vertices a, b, we want to find the k edges which will make the ratio of their closeness centralities as close to 1 as possible. We use a min/max statement on the centralities of a, b to require the smaller value to be in the numerator. As a result, the ratio can never be greater than 1, and the task of maximizing this ratio is the same as making the ratio as close to 1 as possible.

Closeness Ratio Improvement

Input:	A graph $G = (V, E)$, vertices $a, b \in V$, $k \in \mathbb{N}$.
Problem:	Find a set of edges T of size at most k which maximizes $\frac{\min(c_{G+T}(a), c_{G+T}(b))}{\max(c_{G+T}(a), c_{G+T}(b))}$.

Why have we chosen closeness centrality as our measure of node importance, when many other such measures exist? Primarily, edge additions cannot increase/worsen the closeness centrality of any vertex in the graph. Paths can only get shorter with edge additions, not longer. This is different than betweenness centrality, where adding an edge to increase the betweenness centrality of a vertex can inadvertenly decrease the betweenness centrality of another vertex.

Closeness centrality also simplifies the concept of being close to *all vertices* in a graph — a vertex needs to be reasonably close to all other vertices to achieve a low centrality value. Furthermore, if a vertex is close to most of the vertices in the graph but very far from a few, this vertex's centrality score can blow up just the same as if it was far from everything. This agrees with our assumptions about information flow in a network, new information can arise from anywhere and we want to maximize a vertex's ability to receive this information as soon as possible.

Why have we chosen to use the ratio of a, b to quantify how close their centralities are? Could we not take the absolute value of their difference, and try to minimize that value? In short, ratio scales better than difference, especially for smaller centrality values. For arithmetic simplicity, we do not normalize our closeness centrality, but centralities are often constrained between 0 and 1, with 1 representing the most central a vertex could be. This normalization could be achieved by simply considering closeness centrality as the reciprocal of its current formulation. With the normalization in mind, it is reasonable to expect in large graphs that many vertices will have extremely small centrality values. If, for example, two nodes have centralities 0.02 and 0.01 respectively, a difference metric would tell us these centralities differ by 0.01 (and it would thus assert their centralities are very close). However, a ratio metric would assert that their centralities are in fact very different, as they differ by a factor of 2. The societal and mathematical reasons for choosing ratio instead of difference are thus very similar — ratio measures better capture small differences that may be common for this problem [13].

1.2. **Our Contributions.** For Closeness Ratio Improvement, we show that achieving any closeness ratio $\tau \in \left(\frac{1}{2}, 1\right]$ is NP-hard (Section 3). In Section 4, we then present a simple algorithm that always achieves a closeness ratio of at least $\frac{1}{2}$. Furthermore, if $\max(c_G(a), c_G(b)) > cn$, where n is the number of vertices in the graph and $c \in \mathbb{N}$, this algorithm achieves a closeness ratio of at least $\frac{c}{c+1}$. As the best possible closeness ratio is 1, this also implies that our algorithm for achieving a closeness ratio of $\frac{1}{2}$ is a $\frac{1}{2}$-approximation for our problem. Finally, in search of an algorithm that achieves better than a $\frac{1}{2}$-approximation, we present several examples in Section 5 of common strategies which *do not* work.

2. Diameter Minimization

With the goal of ultimately developing algorithms and proving hardness of Closeness Ratio Improvement, we begin with an in-depth look at algorithms which add edges to a graph to reduce its diameter. Compiling several papers, we present results/proofs of NP-hardness and a 4-approximation for this problem. To formally state the problem, Bounded Cardinality Minimal Diameter (which we will refer to as Diameter Minimization), recall that the diameter of a graph G is the longest shortest path between any two vertices in the graph, and denoted as D_G. Specifically, $D_G = \max_{u,v \in V} d_G(u, v)$.

BOUNDED CARDINALITY MINIMAL DIAMETER (DIAMETER MINIMIZATION)

Input:	A graph $G = (V, E)$ and a positive integer $k \in \mathbb{N}$.
Problem:	Find a set of edges $S \subseteq V^2 \backslash E$ of size at most k such that $D_{G+S} = \max_{u,v \in V} d_{G+S}(u, v)$ is minimized.

Why is this relevant for CLOSENESS RATIO IMPROVEMENT*?* Within a network-editing framework, asking which edges best reduce the diameter of a graph is a very natural question, and consequently this problem is well-researched. Although improving the closeness centrality of a vertex (or set of vertices) is a different problem than reducing the diameter over an entire graph, conceptually they both seek to make the graph more compact, minimizing the number of "long" paths. Understanding DIAMETER MINIMIZATION is thus useful in providing algorithmic intuition for our problem — a reasonable first step would be to see *how well* strategies for minimizing diameter perform in improving the closeness centrality of a vertex.

This paper assumes familiarity with theory of NP-hardness and approximation algorithms. To quickly summarize, a problem is NP-hard if it belongs to a class of problems for which no polynomial (efficient) algorithm is known. To prove a problem is NP-hard, we must thus show an equivalency between this problem and a known NP-hard problem. An approximation algorithm is a polynomial-time algorithm that approximately solves an NP-hard problem (since we do not know how to perfectly solve NP-hard problems). To say an algorithm is a c-approximation for an NP-hard problem means that whatever the best possible solution on a problem instance is, this algorithm never gets worse than c times that solution. For example, if the best possible diameter with k edges additions on an instance of DIAMETER MINIMIZATION is 6, a 4-approximation guarantees a diameter no worse than 24.

Although many diameter minimization strategies exist, we focus on approximation algorithms which use the k-center problem (formalized below). At a high level, these algorithms identify "central" vertices and connect all of them, guaranteeing all vertices in the graph are somewhat close to each other. In the ensuing sections, we present the NP-hardness of DIAMETER MINIMIZATION [1], then show Gonzalez's 2-approximation for k-Center [9] and its use in Li's 4-approximation for DIAMETER MINIMIZATION [11]. Tighter analysis of Li's algorithm has shown this strategy in fact achieves a 2-approximation [4].

2.1. **NP-Hardness.** We now present an NP-hardness proof of DIAMETER MINIMIZATION by reducing from Set Cover [1]. This proof is not original to this thesis, but is instead a full presentation of an existing result. We include

it in detail as a simpler case – it introduces the strategy we will later use to prove the hardness of our problem, providing a useful starting point for the reader. To maintain consistency with our later proofs, we have slightly modified some details and terminology from Adriaens' original paper. While many hardness proofs for Diameter Minimization exist, we have chosen to analyze this one as it reduces from a well-established NP-hard problem, Set Cover [10].

Set Cover

Input:	A universe $U = \{e_1, e_2, \ldots, e_n\}$, a collection of subsets $S_1, S_2, \ldots, S_m \subseteq U$, and a positive integer $k \in \mathbb{N}$
Problem:	Is there a set of k subsets such that their union equals U?

In order to properly show hardness, Adriaens [1] also formulates Diameter Minimization as a decision problem.

Diameter Minimization (Decision Variant)

Input:	A graph $G = (V, E)$, and positive integers $k, \delta \in \mathbb{N}$.
Problem:	Is there a set of edges $S \subseteq V^2 \backslash E$ of size at most k such that $D_{G+S} \leq \delta$?

Theorem 2.1 ([1])**.** *Diameter Minimization is NP-Hard.*

Proof. Given an instance of Set Cover with universe $U = \{e_1, e_2, \ldots, e_n\}$, a collection of subsets $S_1, S_2, \ldots, S_m \subseteq U$, and a positive integer $k \in \mathbb{N}$, construct a decision instance of Diameter Minimization as follows. Create a clique of vertices $\{x_1, x_2, \ldots, x_{2k+1}\}$, and connect all of them to the vertex a, (we refer to the vertices in the clique as the set X). Connect a to the vertex b, and for each set S_i, create vertex s_i and connect it to b. Also, create n vertices $e_1, e_2, \ldots, e_n$, one for each element in the universe. Next, connect s_i to each vertex that represents an element contained within S_i. Finally, for any two distinct sets S_i, S_j, create the edge $s_i s_j$ (creating a clique of all the set vertices). This is now an instance of Diameter Minimization, $(G = (V, E), k)$. The construction of G is depicted in Figure 2.

Claim 2.2. *There exists a k-sized set cover of U if and only if there is a set T of at most k edges such that $D_{G+T} \leq 3$.*

Before any edge additions, the diameter of G is 4. The only distances of length 4 are distances between vertices in X and element vertices. Thus to

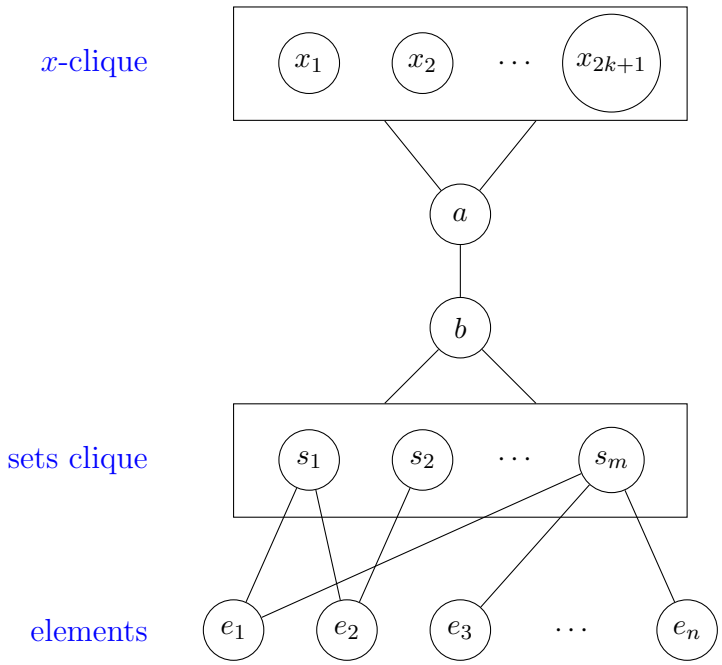

FIGURE 2. Construction of a DIAMETER MINIMIZATION instance from Set Cover (vertices within a box form a clique).

decide if $D_{G+T} \leq 3$, only consider if all of these distances are reduced to length 3.

k-sized set cover $\implies$ diameter is less than or equal to 3. If there is a set cover of size k, for each set S_i in the cover, construct the edge as_i. Now every element vertex must be adjacent to a set vertex in the cover (by definition of a cover), and then each of these set vertices has a newly constructed edge to a, which is directly connected to each vertex in X. This path has length 3, implying $D_{G+T} \leq 3$, as desired.

No k-sized set cover $\implies$ diameter is greater than 3. If there is no set cover of size k, there must be some element vertex with no new edges incident on it or on any set vertex adjacent to it. To see why this is true, suppose every element vertex *did* have a new edge incident on it or on one of the set vertices containing it. For any element vertex that is the endpoint of one of the k new edges, reconstruct this edge to go to one of the sets containing this element instead. Now k edges incident only on set vertices have been identified such that every element is the neighbor of at least one such set vertex, and this can be used to derive a k-sized set cover (a contradiction).

So consider this specific element vertex (call it e) with no new edges incident on it or any set vertex adjacent to it. There must also exist one vertex $x \in X \subset V$ without any new edges incident on it ($2k+1$ vertices in the clique, k edges to add). Now trace the path from e to x. The path from e to any other vertex must start with following an edge from e to the set level, at which point it arrives at some vertex s representing a set that contains e, which also has no shortcut edges incident on it. From there, s cannot reach x in 2 moves, as no matter what vertex it goes to, there will not be an edge (including shortcut edges) which goes directly to x. This is true for all set vertices adjacent to e, implying $d_{G+T}(e, x) > 3$ and thus $D_{G+T} > 3$. □

With a simple reduction from Set Cover, Adriaens shows that the *decision variant* of Diameter Minimization is NP-hard. This shows that the optimization variant (the problem of greater focus) is NP-hard as well — if you can solve the optimization version in polynomial time then you can also solve the decision version in polynomial time, which is conjectured to be impossible. Showing that Diameter Minimization is NP-hard re-positions research goals for the problem, as trying to design an optimal algorithm is now equivalent to trying to prove P=NP. We thus present approximation results which rely on Gonzalez's approximation for k-Center.

2.2. **Gonzalez's 2-Approximation for k-Center.** Before looking at the 4-Approximation for Diameter Minimization, we introduce its subroutine, the k-center problem for graphs, which seeks to identify k center vertices which minimize the maximum distance any vertex is from its closest center. A real-world example of this problem is a supermarket chain deciding where to open new stores to ensure that the maximum distance any customer has to travel to their nearest supermarket is as small as possible. After defining the problem formally, we present Gonzalez's simple greedy 2-approximation for this problem [9]. That is, given a problem instance of k-Center, this algorithm will get a k-Center distance at most 2 times larger than the minimum possible k-Center distance for this problem instance.

k-Center (Graph Variant)

Input:	A graph $G = (V, E)$ and a positive integer $k \in \mathbb{N}$.
Problem:	Find a set of vertices $C \subseteq V$ such that $\lvert C\rvert \leq k$ and $\max_{v \in V} \min_{c \in C} d_G(v, c)$ is minimized.

Use R_C to denote the k-Center distance achieved by some set of centers C. That is, given a graph $G = (V, E)$, $k \in \mathbb{N}$, and $C \subseteq V$ such that $|C| \leq k$, $R_C = \max_{v \in V} \min_{c \in C} d_G(v, c)$. Use C^* to denote the optimal set of centers. Specifically, define $C^* = \arg\min_{C \subseteq V, |C| \leq k} R_C$. Note that the following theorem is not original work to this thesis, although certain notation

has been modified from Gonzalez's original paper for greater consistency with the rest of the paper.

Theorem 2.3 ([9])**.** *Given $G = (V, E)$ and $k \in \mathbb{N}$, there exists a polynomial-time algorithm which produces $C \subseteq V$ such that $|C| \leq k$ and $R_C \leq 2R_{C^*}$.*

Proof. Instantiate the set of centers C with an arbitrary vertex. Then iteratively add the vertex which is farthest away from its closest center in C, until C contains k vertices.

This algorithm has polynomial time complexity (its complexity is not exponential in terms of k or n). Assume access to a distance oracle that can compute vertex distances in $O(1)$ (vertex distances can be computed with a traversal algorithm in polynomial time, so this assumption will not take this algorithm out of polynomial time). In a given iteration, compare the distance of n vertices to at most k centers, selecting the vertex that is farthest from any center. This takes $O(kn)$, and this is needed for k iterations, meaning this approach takes $O(k^2n)$ time overall.

The proof that this achieves an $R_C \leq 2R_{C^*}$ is deferred to Theorem 2.3 in the Appendix. □

2.3. **Li's 4-Approximation for Diameter Minimization.** Equipped with *Gonzalez*, we now present Li's 4-approximation for Diameter Minimization, which uses *Gonzalez* to identify approximately good centers for the graph, connecting them with edge additions. All notation (including the Diameter Minimization problem definition) remains consistent from the previous sections. The set of edges which optimally minimizes the diameter of G is denoted as S^*. That is, $S^* = \arg\min_{S \subseteq V^2 \setminus E, |S| \leq k} D_{G+S}$. Again, the following theorem/proof is not original work to this thesis, and certain notation has been changed from Li's original paper to maintain consistency with the rest of this paper.

Theorem 2.4 ([11])**.** *Given $G = (V, E)$ and $k \in \mathbb{N}$, there exists a polynomial-time algorithm which produces $S \subseteq V^2 \setminus E$ such that $|S| \leq k$ and $D_{G+S} \leq 4D_{G+S^*} + 2$.*

Proof. Given an instance of Diameter Minimization with graph $G = (V, E)$ and $k \in \mathbb{N}$, run *Gonzalez* to find an approximate solution on G for $k+1$ centers. Return the set of edges which connect all centers to one center, at most k edges in total. Note that this remains within the budget of k for Diameter Minimization, even though *Gonzalez* was used as a subroutine with $k + 1$ centers.

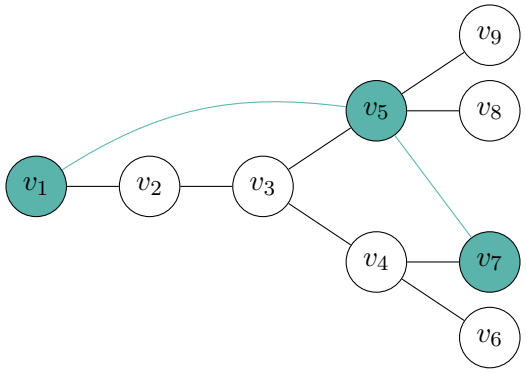

FIGURE 3. Example run of Li's 4-approximation for DIAMETER MINIMIZATION. Blue vertices indicate outputted centers from *Gonzalez*, blue edges are proposed edge additions from *Li*.

This algorithm requires running *Gonzalez*, which is $O(kn)$, and then connecting returned vertices to each other (if they don't already exist in the graph). Depending on how information about the graph is stored, querying if edges exist in the graph already could take at most $O(n^2)$, but certainly this algorithm is polynomial-time.

The proof that $D_{G+S} \leq 4D_{G+S^*} + 2$ is deferred to Theorem 2.4 in the Appendix. □

2.3.1. *Example Run.* We now provide an example run of Li's algorithm for DIAMETER MINIMIZATION on Figure 3 with $k = 2$. We will refer to the graph in Figure 3 as G. First, we run $Gonzalez(G, k+1)$ as a subroutine (meaning we are looking for 3 centers). We will arbitrarily choose a first center, say v_5. From there, 3 vertices maximize our k-center distance function, i.e. v_1, v_7, v_6 are all distance 3 from v_5. We break ties arbitarily, say we choose v_7. Now only one vertex is distance 3 from its closest center, v_1, so that will be our third and final center. Now, given the set of centers $\{v_1, v_5, v_7\}$, we arbitrarily choose one of these centers (say v_5) and connect all other centers to this one, as long as that connection does not already exist in G. This outputs the set of edges $\{v_1v_5, v_5v_7\}$, and as we have shown, the diameter achieved by augmenting G with this edge set cannot be greater than 4 times the diameter achieved by the optimal $k = 2$ solution.

In summary, DIAMETER MINIMIZATION seeks to add edges to a graph to minimize its diameter. This problem is NP-hard and we specifically look at approximations which identify central vertices, and connect them. In particular, these strategies can achieve a 2-approximation (although we have only

provided the simpler analysis which proves a 4-approximation). Understanding these algorithmic strategies may provide a launching point for designing algorithms for Closeness Ratio Improvement.

3. Hardness of Closeness Ratio Improvement

Having gone through similar existing results for Diameter Minimization in detail, we return to our problem, Closeness Ratio Improvement. All of the ensuing content is original work to this thesis. In this section, we will use a similar constrution as in [1] to establish the NP-hardness of CRI, over a range of target closeness ratios τ.

Theorem 3.1. *Closeness Ratio Improvement is NP-Hard for* $\tau = 1$.

This theorem simply states that it is NP-hard to make the closeness ratio of two vertices equal to 1, i.e. make the closeness centralities of two vertices equal.

Proof. Given an instance of Set Cover, construct a decision instance of Closeness Ratio Improvement as follows. Vertices a, b, c form a clique. For each of the n elements from the universe, create a corresponding vertex $e_1, e_2, \ldots, e_n$. For each set S_i, create vertex s_i and connect it to c. Connect s_i to each vertex that represents an element contained within S_i. Finally, create an independent set of $X = n + k$ vertices each connected to a. We refer to the size of this independent set as X for later constructions when it will not remain as $n + k$. The construction of G is depicted in Figure 4.

We claim that there exists a set cover of U of size at most k if and only if there is a set T of at most k edges such that $\frac{\min(c_{G+T}(a), c_{G+T}(b))}{\max(c_{G+T}(a), c_{G+T}(b))} = 1$. Note that in our construction, a is initially more central than b. As we cannot increase the closeness centrality of a vertex by adding edges, the task of maximizing their closeness is initially analagous to decreasing b's centrality.

Suppose that there is a set cover of U of size at most k. For each set S_i in the cover, construct the edge bs_i. Now every element vertex must be adjacent to a set vertex in the cover, and then each of these set vertices has a newly constructed edge to b. Thus each element vertex is distance 2 from b, and k set vertices are distance 1 away from b. Now $c_{G+T}(b) = 2(n+k)+1(2)+1(k)+2(m-k)+2(n) = 4n+2m+k+2$. Similarly, $c_{G+T}(a) = 1(n+k)+1(2)+2(m)+3(n) = 4n+2m+k+2$. Then $\frac{\min(c_{G+T}(a), c_{G+T}(b))}{\max(c_{G+T}(a), c_{G+T}(b))} = 1$, as desired.

Suppose that no k sets cover all of U. As previously stated, in order to maximize the closeness ratio, we must decrease b's closeness centrality as much

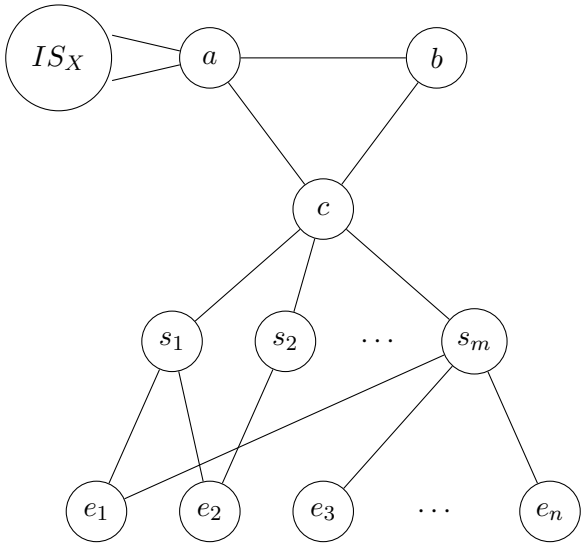

FIGURE 4. Construction of a CLOSENESS RATIO IMPROVEMENT instance with vertices a, b from SET COVER.

as possible with k edges. We now present a helpful claim which identifies the best way for b to improve its closeness centrality, allowing us to bound $c_{G+T}(b)$ below for any set of k edges T.

Lemma 3.2. *Let $G = (V, E)$ be a graph of this construction derived from a set cover instance of m sets, n elements and $k \in \mathbb{N}$. If there is no set cover of size k, then there is a set $T \subseteq V^2 \backslash E$ of at most k edges, all incident on b and some set vertex, whose addition to the graph optimally improves the closeness centrality of b.*

The proof of this fact is deferred to Lemma 3.2 in the Appendix. In other words, the best way b can reduce its closeness centrality is by connecting itself to set vertices. This method can make k set vertices distance 1 from b, and at most $n-1$ element vertices distance 2 from b. However, there must remain at least one element vertex with no newly-added edge incident on it or any vertex of a set containing it, as there is no set cover. Any such element vertex will be distance 3 from b. Then $c_{G+T}(b) \geq 2(n+k)+2+2m-k+2(n-1)+3(1) = 4n+2m+k+3$, and consequently, $\frac{\min(c_{G+T}(a), c_{G+T}(b))}{\max(c_{G+T}(a), c_{G+T}(b))} \leq \frac{4n+2m+k+2}{4n+2m+k+3} < 1$. □

We have shown that it is NP-hard to achieve a closeness centrality ratio of 1, but are smaller ratios achievable in polynomial time? By manipulating the size of the independent set X connected to a (for $\tau = 1$, it was $n+k$ vertices) we can in fact prove a much stronger hardness result.

Theorem 3.3. *Closeness Ratio Improvement is NP-Hard for* $\tau \in (\frac{1}{2}, 1)$.

We will go through the construction of a Closeness Ratio Improvement instance from Set Cover, but the analysis of yes and no cases is deferred to the appendix.

Proof. Fix an arbitrary $\tau \in (\frac{1}{2}, 1)$. Consider an instance of Set Cover with m sets, n elements, and $k \in \mathbb{N}$ which satisfies $\frac{2m+4n+k}{1+2m+4n+k} \geq \tau^2$. This fraction converges to 1 as we increase m, n, k, and $\tau^2 < 1$ when $\tau \in (\frac{1}{2}, 1)$, so we should always be able to find m, n, k that satisfy this inequality. Use the same construction of G described in the proof of Theorem 3.1 and depicted in Figure 4. However, now let X (the number of vertices in the independent set attached to $a \in V$) be an integer in the interval $(\frac{2+2m+3n-\tau(2+2m-k+2n+1)}{2\tau-1}, \frac{2+2m+3n-\tau(2+2m-k+2n)}{2\tau-1}]$. We know such an integer always exists by Lemma B.3.

We claim that there exists a k-sized set cover of U if and only if there is a set T of at most k edges such that $\frac{\min(c_{G+T}(a), c_{G+T}(b))}{\max(c_{G+T}(a), c_{G+T}(b))} \geq \tau$. The analysis of the forward and reverse directions are fairly similar to the proof of Theorem 3.1, with some more complicated algebra to manage the more complicated independent set sizes. Due to the length and similarity to previous content, the remainder of this proof is deferred to Theorem 3.3 in the Appendix. □

In summary, we have shown that for all $\tau \in (\frac{1}{2}, 1]$, it is NP-hard to solve Closeness Ratio Improvement. This then motivates the question, which approximations are achievable for target ratios in this range?

4. $\frac{1}{2}$-Approximation for Closeness Ratio Improvement

In the following section, we present a simple observation about improving the closeness ratio of a, b, and how it can be used to achieve an approximation algorithm for Closeness Ratio Improvement. Again, these results are all original work for this problem. To restate the problem, we are given a graph G, vertices a, b and an edge budget k. In the decision variant of this problem, we are given a target ratio of τ, but generally we would like to make the ratio of a and b's closeness centrality in the augmented graph as close to 1 as possible. Observe that if a and b are connected, we can bound our closeness centrality ratio below.

4.1. Bounding the Closeness Centrality Ratio Below When $ab \in E$.

Specifically, if a and b are connected, for any $v \in V$, we can leverage the fact that $|d_G(a, v) - d_G(b, v)| \leq 1$. Loosely, whatever path a can take to v, b can go to a directly first and then use the same path to v (and vice versa). We now present a more formal statement.

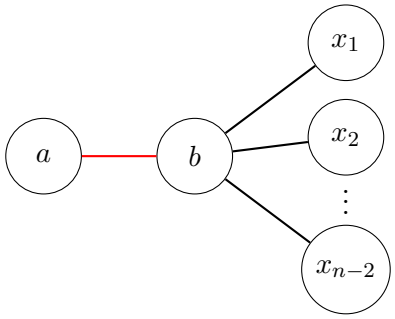

FIGURE 5. The worst possible closeness ratio when a and b are connected.

Lemma 4.1. *If a and b are connected, and without loss of generality, a is more central than b, and if a's centrality is greater than or equal to cn (where $n = |V|$, $c \in \mathbb{N}$), the closeness centrality ratio of a, b is at least $\frac{c}{c+1}$.*

Proof. We suppose that a and b are connected, a is more central than b, and $c_G(a) \geq cn$. This implies that the closeness centrality of b is no more than $c_G(a) + n(1)$. For any vertex v that b needs to get to, it can use the edge ab and then whatever path a uses to get to v. That is, $d_G(b, v) \leq 1 + d_G(a, v)$. If we sum this over all n vertices, we get the desired claim about closeness centralities. Putting this together with the fact that $c_G(a) \geq cn$, and that $\frac{c_G(a)}{c_G(a)+n}$ is an increasing function of $c_G(a)$:

$$\frac{\min(c_G(a), c_G(b))}{\max(c_G(a), c_G(b))} = \frac{c_G(a)}{c_G(b)} \geq \frac{c_G(a)}{c_G(a) + n} \geq \frac{cn}{cn + n} = \frac{c}{c+1}$$

. □

But what is the worst case of this analysis? The best possible centrality that a vertex can have is n (it is one away from all n vertices). In this case, $c = 1$, and then the ratio we achieve is at least $\frac{1}{2}$ (see Figure 5). For this example to actually form a ratio of $\frac{1}{2}$ and not better, the entire graph must be directly connected to b, and a's only neighbor is b itself. In summary, connecting ab bounds below the closeness ratio by $\frac{1}{2}$ for arbitrary graphs, but this relies on the assumption that the best centrality in the graph is around n, i.e. connected to all other vertices. In real-world networks with many more nodes, this is rarely the case. The best closeness centralities will usually be much larger, implying this strategy actually achieves a much better closeness ratio as $\frac{c}{c+1} \to 1$.

4.2. **A Simple Approximation Algorithm.** Now consider the algorithm, given $k \geq 1$ edges, which simply adds the edge ab (if it does not already exist

in the graph). As we just argued, our closeness ratio is now at least $\frac{1}{2}$. This agrees very well with our hardness results, as we showed getting a target ratio greater than $\frac{1}{2}$ in all cases is NP-hard, but getting a target ratio of exactly $\frac{1}{2}$ is easy (in fact, it only requires one edge). Furthermore, this strategy is a $\frac{1}{2}$-approximation for CRI. This follows from the observation that, simply by our definition of closeness ratio, a ratio better than 1 is not possible. By guaranteeing a ratio of $\frac{1}{2}$, we also guarantee that our algorithm produces no worse than $\frac{1}{2}$ of the optimal ratio.

This strategy is also *no better* than a $\frac{1}{2}$-approximation. That is, there is a certain G, k where this strategy gets a ratio of $\frac{1}{2}$ and some optimal algorithm can get a ratio of 1. Consider Figure 5 again, this time with $k = n - 2$. Our algorithm will want to connect ab (if it's already there, it won't do anything), and then it will stop. As we've previously argued this gets a ratio of $\frac{1}{2}$. However, an optimal algorithm could connect a to the $k = n - 2$ vertices connected to b. Now both a and b are one away from all other vertices, meaning the ratio of their closeness centralities is 1.

To summarize, adding the edge ab is a simple algorithm that requires only $k \geq 1$. It necessarily achieves a closeness ratio of $\frac{1}{2}$ (though often times much more in larger graphs), and this also means it achieves an approximation ratio of at least $\frac{1}{2}$, as the best possible ratio is 1. We also show that this analysis is tight, meaning this strategy is no better than a $\frac{1}{2}$-approximation.

5. Intuition-Building Examples

Having established a trivial $\frac{1}{2}$-approximation, it is also worthwhile to explore other algorithmic strategies and analyze how well they perform. We now present two examples of algorithmic strategies which do not form approximations. In fact, they perform arbitrarily badly compared to an optimal solution. Identifying these types of examples will help inform later algorithm development (knowing what will not work is almost as useful as knowing what will).

5.1. **Single-Node Greedy.** A possible strategy could be to take the less central vertex between a, b, and then greedily improve that vertex. To see why this does not work, assume $a \in V$ has a higher closeness centrality value/is less central than $b \in V$. Consider the graph displayed in Figure 6. The dashed edges e_1 and e_2 are prospective edges, not yet in the graph. The larger nodes IS_X and IS_Y represent independent sets of X nodes and Y nodes respectively. The squiggly lines represent paths of length longer than 1, and the distance of these paths is denoted next to the edge. To start, we compute that

$$c_G(a) = (d + d + 1 + 1)(Y) + (d + 1)(X) = (2d + 2)Y + (d + 1)X$$

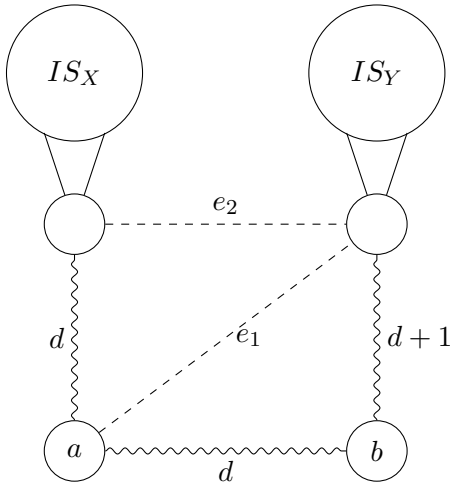

FIGURE 6. Greedily improving the less central vertex's closeness can lead to an arbitrarily bad solution.

$$c_G(b) = (d+1+1)(Y) + (d+d+1)X = (d+2)Y + (2d+1)X$$

Observe that we are only including the distance from a, b to the large independent sets of size X and Y in our computations. We leave the values of X and Y arbitrary, but certainly we can increase them to the point that they dominate all other terms in closeness centrality calculation, and hence we ignore all intermediate vertices. We will also use the practice in later examples to simply calculations.

So if we construct Y sufficiently larger than X, a is less central. With $k = 1$, the greedy edge addition a will make to maximize its own closeness will be e_1, as now a is only distance 2 away from the larger set of Y. In this case:

$$c_{G+e_1}(a) = 2Y + (d+1)X$$

$$c_{G+e_1}(b) = (d+2)(Y) + (2d+1)X$$

Now a is much more central, and for large enough Y, $\frac{c_{G+e_1}(a)}{c_{G+e_1}(b)} \approx \frac{2}{d+2}$. Thus we can make this quantity arbitrarily small, as it is dependent on d. However, there is an edge addition for $k = 1$ which can do much better, e_2.

$$c_{G+e_2}(a) = (d+1)Y + (d+1)X$$

$$c_{G+e_2}(b) = (d+2)Y + (d+3)X$$

Still, a is more central, but for large enough Y, $\frac{c_{G+e_2}(a)}{c_{G+e_2}(b)} \approx \frac{d+1}{d+2}$. Now, we can make this quantity as close to 1 as we want (which is what we hope to do in the context of this problem).

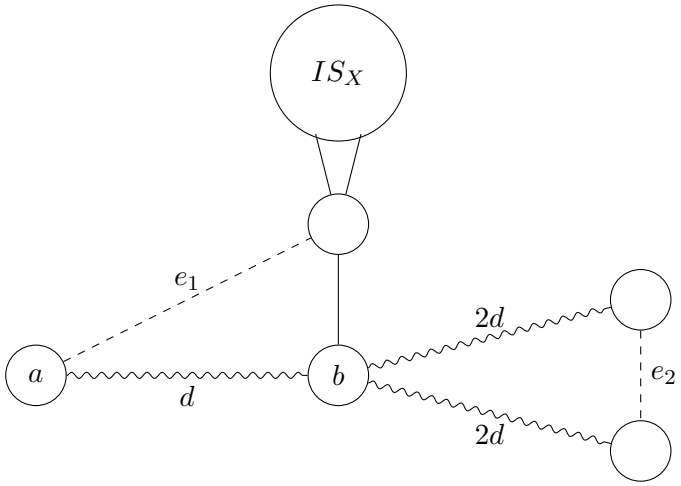

FIGURE 7. Adding the edge which optimally minimizes the diameter of a graph can lead to an arbitrarily bad solution for CRI.

In short, we can choose large enough Y and d in this construction such that if we greedily try to maximize a's centrality, we get a ratio of centralities that is arbitrarily small. However, there exists an edge addition that makes this quantity arbitrarily close to 1 for appropriate choices of X, Y and d, meaning this greedy choice is an arbitrarily bad solution with respect to the optimal.

5.2. **Diameter-Minimizing Edge Addition.** In an earlier section, we established the similarity between DIAMETER MINIMIZATION and CLOSENESS RATIO IMPROVEMENT. How well do Diameter Minimization strategies perform on our problem? If we just add the edge(s) that minimizes the diameter of the graph, how well can this do?

Consider Figure 7. Before adding any edges, G has diameter $4d$ (the distance from the two nodes peripheral to b to each other). The closeness of b is around $2X$ (the size of the independent set dominates other vertices). The closeness of a is then $(d+2)X$, giving us a ratio of $\frac{2}{d+2}$. If we add the best diameter minimizing edge, e_2 (which gets us a diameter of $3d$), we still have a ratio of $\frac{2}{d+2}$. However, if we add the edge e_1, now a and b are equidistant from the independent set, giving us ratio $\approx \frac{2}{2} = 1$. Thus, we can keep increasing d to show that adding the diameter minimizing edge performs arbitrarily badly compared to an optimal solution.

This example also validates the novelty of this work — while similar research exists, the same strategies used to solve those problems will not necessarily carry over.

6. Conclusion

6.1. **Summary.** Graphs serve as a powerful mathematical model for real-world social networks, allowing us to represent disparities in information access using well-known graph metrics such as closeness centrality. With this framework, we aim to add a budgeted number of edges to a graph to make the closeness centrality of two given vertices as equal as possible, providing them with similar importance in the network. To approach this problem, we first examine the related network-editing problem of Diameter Minimization, in case it may inform our solution. We then prove that our problem, Closeness Ratio Improvement, is NP-hard for a broad range of values, but we also show that simply adding an edge between the two vertices guarantees a closeness (and thus approximation) ratio of at least $\frac{1}{2}$. Finally, we explore several common algorithmic strategies in search of a better approximation algorithm, demonstrating that they perform poorly.

6.2. **Future Work.** Certainly, we have not provided a tight analysis of Closeness Ratio Improvement. Is there an approximation algorithm that can guarantee a ratio better than $\frac{1}{2}$? What is the best possible approximation algorithm? We know that a 1-approximation is not possible, but is a $\frac{9}{10}$ possible? These unanswered questions, trying to bridge the gap between best-known and best-possible approximations, would be the main direction of future work.

Recall that the initial motivation for this problem was to develop strategies for providing an entire network with equal information access. The two-vertex case that this paper focuses on does not quite achieve that (it only equalizes two vertices). The hope is that strategies from this problem could provide intuition for a similar problem which focuses on the entire graph.

To provide a more formal problem statement, we seek to maximize the worst possible ratio in a graph between all vertices. Note that connecting ab cannot guarantee any constant factor ratio (as before).

All-Pairs Closeness Ratio Improvement

Input:	A graph $G = (V, E)$, vertices $a, b \in V$, $k \in \mathbb{N}$.
Problem:	Find the set of edges T of size at most k such that $\frac{\min_{u \in V} c_{G+T}(u)}{\max_{v \in V} c_{G+T}(v)}$ is maximized.

This is a much more difficult problem mathematically, but (if solved) would provide a better guarantee that an entire network is "fair".

Appendix A. Deferred Diameter Minimization Proofs

This section contains deferred content from the literature survey of Diameter Minimization, and its content is not original to this thesis. In particular, it contains the proof from [9] proving that the greedy algorithm for k-center achieves a 2-approximation. It also contains two proofs from [11] that ultimately show their approach for Diameter Minimization achieves a 4-approximation.

Theorem 2.3 ([9])**.** *Given $G = (V, E)$ and $k \in \mathbb{N}$, there exists a polynomial-time algorithm which produces $C \subseteq V$ such that $|C| \leq k$ and $R_C \leq 2R_{C^*}$.*

Proof. Given a graph $G = (V, E)$ and $k \in \mathbb{N}$, consider the polynomial-time algorithm *Gonzalez*, which outputs a set of centers $C \subseteq V$ such that $|C| \leq k$ and $R_C \leq 2R_{C^*}$. The first two claims follow immediately from the construction of this algorithm — C is exclusively selected from V in an iterative fashion until $|C| = k$.

Now to show that $R_C \leq 2R_{C^*}$, at the termination of *Gonzalez*, define $x \in V$ as the vertex which realizes R_C. The set $C \cup \{x\}$ contains $k + 1$ vertices, while the optimal solution C^* only contains k vertices, implying two vertices in $C \cup \{x\}$ must be closest to the same center in C^* (by the Pigeonhole Principle). Call these two vertices u and v, and the center that they share $c \in C^*$. *Gonzalez* greedily adds vertices to C which maximize R_C, implying that $d_G(u, v) \geq R_C$. If this were not the case, x (which is R_C away from its closest center) would have been added into C before either u or v.

The triangle inequality then says that either $d_G(u, c) \geq \frac{R_C}{2}$ or $d_G(v, c) \geq \frac{R_C}{2}$. If neither were true, there would be a path between u and v of length less than R_C, which has been established as not possible. In any case, there is a vertex which is at least $\frac{R_C}{2}$ away from its center in the optimal solution, implying $R_{C^*} \geq \frac{R_C}{2}$ and thus $R_C \leq 2R_{C^*}$, as desired. □

The following content is slighlty modified from [11] and shows that their strategy achieves a 4-approximation. It starts with a helpful lemma that compares the magnitude of k-center solutions to Diameter Minimization solutions, and then establishes the bound in the following theorem.

Lemma A.2 ([11])**.** *For any graph $G = (V, E)$ and $k \in \mathbb{N}$, $R_{C^*} \leq D_{G+S^*}$, where C^* is the optimal solution for $(k + 1)$-center and S^* is the optimal solution for Diameter Minimization with k edges.*

Proof. Given a graph $G = (V, E)$, $k \in \mathbb{N}$, and the optimal solution of k edges for Diameter Minimization S^*, construct a solution C in the following manner for $(k+1)$-center such that $R_C \leq D_{G+S^*}$. This implies that $R_{C^*} \leq D_{G+S^*}$, where R_{C^*} is the radius achieved by the optimal set of centers C^* and D_{G+S^*} is the diameter achieved by the optimal set of edges S^*.

First, instantiate C with an arbitrary $x \in V$. Then for each edge $(uv) \in S^*$, add whichever vertex is further from x to the set of centers. There were k edges in S^*, and at most one center was added to C for each edge, and C was instantiated with one vertex, thus $|C| \leq k+1$.

Now if it can be shown that for any $v \in V$, the distance in G to its closest center $c \in C$ is smaller than the distance to x in $G + S^*$, this would imply that $R_C \leq d_{G+S^*}(v, x) \leq D_{G+S^*}$. Consider the shortest path from v to x in $G + S^*$. There are two cases: either this path uses an edge in S^* or it does not. In the first case, call the first new edge it encounters $(ab) \in S^*$ and, without loss of generality, $d(v, a) \leq d(v, b)$. Then by construction of C, a is a center in C (it is the endpoint of an edge in S^* farther from x). So the path from v to a center a contains no edges in S^* and it is also a subpath of the shortest path from v to x in $G + S^*$. If the path from v to x contains no edges in S^*, then x itself is a center, and the distance in G between x and its closest center is the same as the distance in $G + S^*$ from v to x. In either case, the distance from v to its closest center in G is less than or equal to the distance from v to x in $G + S^*$. As this is for arbitrary $v \in V$, $R_C \leq D_{G+S^*}$, implying $R_{C^*} \leq D_{G+S^*}$, as desired. □

Theorem 2.4 ([11])**.** *Given $G = (V, E)$ and $k \in \mathbb{N}$, there exists a polynomial-time algorithm which produces $S \subseteq V^2 \backslash E$ such that $|S| \leq k$ and $D_{G+S} \leq 4D_{G+S^*} + 2$.*

Proof. Given a graph $G = (V, E)$ and $k \in \mathbb{N}$, consider the polynomial-time algorithm *Li*, which outputs a set of edges $S \subseteq V^2 \backslash E$ such that $|S| \leq k$ and $D_{G+S} \leq 4D_{G+S^*} + 2$. The first two claims follow immediately from the construction of the algorithm — S can only be edges which do not already exist in G and S connects k distinct centers all to one center (at most k edges).

Define R_C as the $(k+1)$-center distance achieved by the set of centers C on G using *Gonzalez*. Then for any $u, v \in V$, both u and v cannot be more than R_C away from their respective centers c_u and c_v. *Li* then connects all centers to one center, so c_u and c_v cannot be more than 2 apart in $G + S$. Thus $d_{G+S}(u, v) \leq 2R_C + 2$ for any $u, v \in V$, implying $D_{G+S} \leq 2R_C + 2$. And as *Gonzalez* can be used to achieve a 2-approximation for the k-center problem, $D_{G+S} \leq 2(2R_{C^*}) + 2$, and then by Lemma A.2, $D_{G+S} \leq 2(2D_{G+S^*}) + 2 = 4D_{G+S^*} + 2$, as desired. □

Appendix B. Deferred Proofs of NP-Hardness of CRI

The following section contains deferred proofs that are used in our NP-hardness argument, and are all original content to this thesis. The following claim and lemma argue that, in our construction, there is always a set of k edges all incident on b and some set vertex which optimally improves the closeness centrality of b.

Claim B.1. *Given $G = (V, E)$, $b \in V$ and $k \in \mathbb{N}$, there is always a set $T \subseteq V^2 \backslash E$ of at most k edges, all incident on b, whose addition to the graph optimally improves the closeness centrality of b.*

Proof. Suppose that none of the sets of k edges which optimally improved the closeness centrality of b were all incident on b. Consider one such set S, and we'll create the set of edges T as follows: for each edge in S that *is* incident on b, include it in T. For each edge $xy \in S$ that is not incident on b, suppose without loss of generality that $d_{G+S}(b, x) \leq d_{G+S}(b, y)$, and then include the edge by in T. Now $|T| \leq k$ and all edges in T are incident on b.

For an arbitrary $v \in V$, we consider two cases, and show that in both $d_{G+T}(b, v) \leq d_{G+S}(b, v)$. Either the shortest path from b to v in $G + S$ used an edge in S not incident on b, or it did not. In the first case, let xy be the last edge in S not incident on b used by the shortest path from b to v in $G + S$. Then we know that the edge by exists in $G + T$, by construction. Thus,

$$\begin{aligned} d_{G+T}(b, v) &\leq d_{G+T}(b, y) + d_{G+T}(y, v) = 1 + d_{G+T}(y, v) \\ &\leq d_{G+S}(b, y) + d_{G+S}(y, v) = d_{G+S}(b, v) \end{aligned}$$

If the shortest path from b to v in $G+S$ did not use an edge in S not incident on b, then all the edges on the path will still exist in $G+T$, by construction. Thus $d_{G+T}(b, v) \leq d_{G+S}(b, v)$.

So for all vertices v, $d_{G+T}(b, v) \leq d_{G+S}(b, v)$, implying that $c_{G+T}(b) \leq c_{G+S}(b)$. But this contradicts the fact that none of the sets which optimally improve the closeness centrality of b are all incident on b. □

Lemma B.2. *Let $G = (V, E)$ be a graph of the construction provided in Figure 4 derived from a set cover instance of m sets, n elements and $k \in \mathbb{N}$. If there is no set cover of size k, then there is a set $T \subseteq V^2 \backslash E$ of at most k edges, all incident on b and some set vertex, whose addition to the graph optimally improves the closeness centrality of b.*

Proof. From Claim B.1, we know that there must exist a set S of k edges which optimally improves the closeness centrality of b such that all edges are

incident on b. Suppose towards a contradiction that some of the edges in S are not incident on a set vertex.

For each edge in S not incident on b and a set vertex, we thus consider two cases: either the edge is incident a vertex x_i in the independent set, or it is incident on an element vertex e_i. In both cases, we'll show that we can replace this edge with an edge from b to a set vertex which reduces the closeness centrality by at least as much as this edge when added to the graph. Note that we do not consider incidence on a or c as b is already connected to both of those vertices in G.

In the first case, bx_i reduces the closeness centrality of b by at most 1. It has changed $d_G(x_i, b)$ from 2 to 1, but not made b closer to any other vertices. If we instead connect b to a set vertex s_i, then $d_G(s_i, b)$ has been changed from 2 to 1, meaning this edge reduces the closeness centrality of b by at least as much as the edge bx_i.

In the second case, be_i reduces the closeness centrality of b by at most 2 — it has changed $d_G(e_i, b)$ from 3 to 1, but not made b closer to any other vertices. Recall that there is no set cover of size k for the instance our graph is derived from, and thus there must always exist some element vertex with no edges in S incident on it, or any set containing it. So instead connect b to a set vertex s_j, where $e_j \in S_j$. Then $d_G(s_j, b)$ has been reduced from 2 to 1, and $d_G(e_j, b)$ from 3 to 2, meaning bs_j reduces the closeness centrality of b by 2, at least as much as the edge be_i.

We have thus described a method to derive a set T of k edges all incident on b and some set vertex, such that $c_{G+T}(b) \leq c_{G+S}(b)$, a contradiction. □

For the purpose of proving hardness for $\tau \in (\frac{1}{2}, 1]$, Lemma B.3 argues that there is always an integer in the range we draw independent set sizes from. This shows that our construction is valid (i.e., we are not saying for some values of τ to construct an independent set of 3.7 vertices).

Lemma B.3. *For $m, n, k \in \mathbb{N}$ and $\tau \in \left(\frac{1}{2}, 1\right]$, there exists $X \in \mathbb{N}$ such that $X \in \left(\frac{2+2m+3n-\tau(2+2m-k+2n+1)}{2\tau-1}, \frac{2+2m+3n-\tau(2+2m-k+2n)}{2\tau-1}\right]$.*

Proof. Taking the difference of the bounds of the interval $\left(\frac{2+2m+3n-\tau(2+2m-k+2n+1)}{2\tau-1}, \frac{2+2m+3n-\tau(2+2m-k+2n)}{2\tau-1}\right]$, we get that the length of the interval is $\frac{\tau}{2\tau-1}$, which is greater than or equal to 1 for all $\tau \in \left(\frac{1}{2}, 1\right]$. Therefore, the length of our interval is bounded below by 1, and so there must exist some integer X within this interval. Furthermore, as the numerator and

denominator of the bounds of this interval are positive when $m, n, k \in \mathbb{N}$ and $\tau \in \left(\frac{1}{2}, 1\right]$, $X \in \mathbb{N}$ as desired. □

This is the deferred proof for Closeness Ratio Improvement being NP-hard for $\tau \in (\frac{1}{2}, 1)$.

Theorem 3.3. *Closeness Ratio Improvement is NP-Hard for $\tau \in (\frac{1}{2}, 1)$.*

Proof. (Concluding the argument started earlier, with construction given).

k-sized set cover $\implies \frac{\min(c_{G+T}(a), c_{G+T}(b))}{\max(c_{G+T}(a), c_{G+T}(b))} \geq \tau$. If there is a set cover, follow the same protocol as in the proof of Theorem 3.1 (connect b to the k set vertices representing sets in the cover). In this case, $c_{G+T}(a) = X + 2 + 2m + 3n$ and $c_{G+T}(b) = 2X + 2 + 2m - k + 2n$.

We want to show that the ratio we achieve is greater than or equal to τ, but we must still ensure that our ratio is less than or equal to 1, else we violate our min/max definition of closeness ratio. Note that this was not a conceren when $\tau = 1$, because we showed our ratio when there was a set cover was 1 exactly. So we consider the case where $c_{G+T}(b) \geq c_{G+T}(a)$ and the case where $c_{G+T}(b) < c_{G+T}(a)$, and show that in both cases the ratio of $\frac{\min(c_{G+T}(a), c_{G+T}(b))}{\max(c_{G+T}(a), c_{G+T}(b))} \geq \tau$.

If $c_{G+T}(b) \geq c_{G+T}(a)$, then

$$\frac{\min(c_{G+T}(a), c_{G+T}(b))}{\max(c_{G+T}(a), c_{G+T}(b))} = \frac{X + 2 + 2m + 3n}{2X + 2 + 2m - k + 2n}$$

$$\geq \frac{(\frac{2+2m+3n-\tau(2+2m-k+2n)}{2\tau-1}) + 2 + 2m + 3n}{2(\frac{2+2m+3n-\tau(2+2m-k+2n)}{2\tau-1}) + 2 + 2m - k + 2n} \geq \tau$$

Alternatively, if $c_{G+T}(b) < c_{G+T}(a)$, then

$$\frac{\min(c_{G+T}(a), c_{G+T}(b))}{\max(c_{G+T}(a), c_{G+T}(b))} = \frac{2X + 2 + 2m - k + 2n}{X + 2 + 2m + 3n}$$

$$> \frac{2(\frac{2+2m+3n-\tau(2+2m-k+2n+1)}{2\tau-1}) + 2 + 2m - k + 2n}{(\frac{2+2m+3n-\tau(2+2m-k+2n+1)}{2\tau-1}) + 2 + 2m + 3n} = \frac{(2 - 2\tau) + 2m + 4n + k}{\tau + 2m\tau + 4n\tau + k\tau}$$

As $\tau \in (\frac{1}{2}, 1)$, $2 - 2\tau > 0$, implying:

$$\frac{(2 - 2\tau) + 2m + 4n + k}{\tau + 2m\tau + 4n\tau + k\tau} > \frac{2m + 4n + k}{\tau + 2m\tau + 4n\tau + k\tau} = \frac{1}{\tau}(\frac{2m + 4n + k}{1 + 2m + 4n + k})$$

We chose our Set Cover instance such that $\frac{2m+4n+k}{1+2m+4n+k} \geq \tau^2$. Thus we have that:

$$\frac{1}{\tau}(\frac{2m+4n+k}{1+2m+4n+k}) \geq \frac{1}{\tau}(\tau^2) = \tau$$

Thus in either case, if there is a set cover of U, we can add k edges to G to get a closeness ratio greater than or equal to τ.

No k-sized set cover $\implies \frac{\min(c_{G+T}(a),c_{G+T}(b))}{\max(c_{G+T}(a),c_{G+T}(b))} < \tau$. If there is no k-sized set cover, we use the same analysis from the proof of Theorem 3.1 to argue that the best way b can reduce its closeness centrality is by connecting itself to set vertices. Furthermore, this method makes k set vertices distance 1 from b, and thus $n-1$ element vertices distance 2 from b, but as we've previously argued, there must be some element which remains distance 3 from b. Then the closeness ratio of a, b is:

$$\frac{\min(c_{G+T}(a), c_{G+T}(b))}{\max(c_{G+T}(a), c_{G+T}(b))} = \frac{X+2+2m+3n}{2X+2+2m-k+2n+1}$$

$$< \frac{(\frac{2+2m+3n-\tau(2+2m-k+2n+1)}{2\tau-1}) + 2 + 2m + 3n}{2(\frac{2+2m+3n-\tau(2+2m-k+2n+1)}{2\tau-1}) + 2 + 2m - k + 2n + 1} = \tau$$

□

References

[1] Florian Adriaens and Aristides Gionis. "Diameter Minimization by Shortcutting with Degree Constraints". In: *2022 IEEE International Conference on Data Mining (ICDM)*. IEEE, Nov. 2022, pp. 843–848. DOI: 10.1109/icdm54844.2022.00095.

[2] Ashkan Bashardoust et al. "Reducing Access Disparities in Networks using Edge Augmentation". In: *FAccT '23* (2023), pp. 1635–1651.

[3] Aditya Bhaskara et al. "Optimizing Information Access in Networks via Edge Augmentation". In: (2024). arXiv: 2407.02624 [cs.DS].

[4] Davide Bilò, Luciano Gualà, and Guido Proietti. "Improved approximability and non-approximability results for graph diameter decreasing problems". In: *Theoretical Computer Science* 417 (2012), pp. 12–22.

[5] danah boyd, Karen Levy, and Alice Marwick. "The Networked Nature of Algorithmic Discrimination". In: *Data and Discrimination: Collected Essays* (2014), pp. 43–57.

[6] Pierluigi Crescenzi et al. "Greedily Improving Our Own Closeness Centrality in a Network". In: *ACM Transcations on Knowledge Discovery from Data* 11.1 (2016), pp. 1–32.

[7] Erik Demaine and Morteza Zadimoghaddam. “Minimizing the Diameter of a Network using Shortcut Edges”. In: *Algorithm Theory - SWAT 2010* 6139 (2010), pp. 420–431.

[8] Benjamin Fish et al. “Gaps in Information Access in Social Networks”. In: (2019). DOI: 10.1145/3308558.3313680.

[9] Teofilo F. Gonzalez. “Clustering to minimize the maximum intercluster distance”. In: *Theoretical Computer Science* 38 (1985), pp. 293–306.

[10] Richard Karp. “Reducibility among Combinatorial Problems”. In: *Complexity of Computer Computations* (1972), pp. 85–103.

[11] Chung-Lun Li, S. Thomas McCormick, and David Simchi-Levi. “On the Minimum-Cardinality-Bounded-Diameter and the Bounded-Cardinality-Minimum-Diameter Edge Addition Problems”. In: *Operations Research Letters* 11 (5 1991), pp. 303–308.

[12] Sourav Medya et al. “Group Centrality Maximization via Network Design”. In: 2018, pp. 126–134.

[13] Min-Hsuan Yeh et al. “Analyzing the Relationship Between Difference and Ratio-Based Fairness Metrics”. In: *The 2024 ACM Conference on Fairness, Accountability and Transparency* (2024), pp. 518–528.